Spring 2023

A Multivariate K-Means Cluster Analysis of Historical Pharmaceutical Research and Development Expenditure Efficiency's Relationship to Forward Earnings and Sales Multiples

Nicholas Leonard Anduze
Bard College

A Multivariate K-Means Cluster Analysis of Historical Pharmaceutical Research and Development Expenditure Efficiency's Relationship to Forward Earnings and Sales Multiples

Senior Project Submitted to
The Division of Social Studies
of Bard College

by

Nicholas Anduze

Annandale-on-Hudson, New York
May 2023

Acknowledgements

I want to extend a special thanks to the world-class Economics department at Bard. The school has done an incredible job recruiting and retaining a phenomenal team of researchers and educators. To them, I am forever grateful.

In addition, I want to thank my immensely supportive family and friends that made my time at Bard particularly special.

Last, I'd like to thank my SPROJ advisor, Taun Toay, for working closely with me over the previous two semesters on this paper. Despite my attempts to take this project in twenty different directions, his guidance helped unify this paper into one cohesive narrative. His patience and thoughtfulness throughout this process have been invaluable. Thank you.

Table of Contents

Abstract

In most sectors, estimating the economic impact of specific events is a laborious and imprecise task. This exercise requires triangulating end-market demand, propensity to consume, and the opportunity costs consumers incur when selecting one competitor's good or service over another to determine the optimal assortment of capital and labor to supply a market profitably (Pindyck, 245). In these competitive sectors, consumers set prices, and firms act as price-takers focusing on improving their operations to eke out a profit. Although intellectually stimulating, this analysis may prove fickle if consumer preferences suddenly shift from blue widgets to red ones. This analog, however, is not transferable for firms operating in the Life Sciences industry. Unsurprisingly, the demand for a therapeutic that allows a patient to avoid the high mortality rate of a condition's prevailing standard of care is inelastic. The economic rewards for a firm that can discover, patent, develop, clinically validate, and commercialize its product is undoubtedly desirable (Temin, 436). However, there is no such thing as a free lunch – life science companies often generate no revenue for years while undergoing rigorous clinical trials in animal models and humans to prove efficacy to the FDA. An FDA approval not only grants access to lucrative, inelastic end-markets with the potential to generate windfall revenues and profits that can offset development costs but the ability to re-invest excess cash flow into promising clinical trials (Spitz, 5). Despite steadily increasing research and development costs, the FDA approves fewer therapeutics each year (Scannel, 192). From both a capital allocation and healthcare-outcomes perspective, the deteriorating efficacy of R&D is concerning. If the economic profit and thus present value of life science companies are mostly determinable from binary FDA decisions, an understanding of how market participants value more efficient capital allocators is invaluable. That is precisely this paper's focus.

In the international arena, only a few market participants wield the rare authority to ascribe value to essential consumer assets. The Food and Drug Administration (FDA) is one of these select few - established seven years prior to the Federal Reserve in 1906, the FDA has spent the last 116 years approving over 5,600 drugs for widespread commercialization. Despite employing a small staff, the FDA reviews thousands of New Drug (NDA) and Biologic License Applications (BLA) each year - extending marketing and distribution rights to only a select few. For a handful of manufacturers, an FDA approval can grant access to captive, inelastic end-markets with the potential to generate monopolistic revenues and profits. FDA approval is incredibly lucrative for manufacturers possessing intellectual property protection due to their ability to offset high fixed development costs and re-invest excess cash flow into promising clinical trials.

Literature Review

History of Therapeutics

Modern pharmaceutical manufacturing began 3500 years ago with the use of willow bark and leaves to treat inflammation, migraines, and fevers (Montinari, Maria Rosa, et al. 1). First prescribed by Sumerian and Egyptian physicians, evidence of willow's usage can be found across continents and communities. Originating from the willow tree, its primary healing mechanisms derive from the bark's Salicylic acid (Ibid, et al. 2). Despite the willow tree's enduring popularity, its affordability remained an issue. For this reason, throughout the 1800s, several German Biochemists sought to engineer a stable chemical substance that could be affordably mass-produced. In 1897, after several decades of trials, a Bayer Chemist acetalized a Phosphyl from Sacilyic acid, thus generating the stable substance now known as Aspirin (Ibid, et al. 4). Two years later, Germany approved Aspirin to treat back and joint pains. The USA followed suit in

1900, and four years later, Aspirin was the most widely used medication globally – a position it retains to this day.

Despite widespread industrial innovation throughout the 19th century, the manufacturing process for therapeutics in America remained largely untouched. Aspirin's discovery, development, commercialization, and overall success in the early 1900s solidified Germany as the world's leading exporter of cutting-edge pharmaceutical research (Daemmrich, 63). In contrast, most medications in the US were still blended at apothecaries or local family-owned healers. For this reason, American reliance on German pharmaceutical exports remained durable until war-time blockades in 1914 forced domestic Chemists to engineer replacements for critical compounds like Aspirin, anti-microbial salvarsan, and Barbital (Conroy, 47). The first world war highlighted the importance of domestic pharmaceutical manufacturing, so legislators passed the US Alien Property Custodian Law, allowing producers to commercialize German-patented therapeutics in the US (Daemmrich, 66).

Between 1930 and 1950, the rise of manufacturing practices would reduce the roles of on-site pharmacists and elevate physicians as authoritative figures in patients' healthcare lives. In 1930, the percentage of prescriptions requiring compounding on-site by a pharmacist stood at 75%. By 1960, this figure fell to only 4% (Ibid 64). During World War II, the US government directed funds toward pharmaceutical manufacturing to assist the war effort. As a result, former competitors cooperated to mass-test, manufacture, and distribute novel therapeutics like penicillin to soldiers (Whayne, 170). After the war, the price of penicillin fell, encouraging researchers to engineer solutions that could utilize the same proven fermentation process as penicillin for other conditions. Vertical integration of life science operators provided leverage in implementing advances from medicinal chemistry toward developing therapeutics targeting a

range of indications in virology. Additionally, due to the American Psychiatric Associations' efforts in classifying psychoneurotic disorders, the pharmaceutical industry began targeting less severe mental conditions like schizophrenia and anxiety (Daemmrich, 5). This wave of innovation in a narrow set of indications was phenomenal for consumers, expanding options to include branded and generic therapeutics at different prices. As a result, it became common practice amongst pharmacists to substitute patients' generic prescriptions with a more expensive branded alternative to reduce the inventory costs of holding several therapeutics (Ibid, 67). In response, by 1959, over 44 states established an anti-substitution provision restricting pharmacists from dispensing patients anything other than what a physician had prescribed (McCarey, 105).

In the 1960s, the FDA strengthened its grip over drug approval by increasingly emphasizing the importance of developing therapeutics in sterile and safe laboratories (Daemmrich, 68). As a result, maintaining good manufacturing and laboratory standards became commonplace for life science firms seeking approval for their therapeutics. Regardless, issues still arose – famously, in 1982, seven people perished after consuming Johnson & Johnson's extra strength Tylenol, resulting in their removal from shelves for nearly six months (Adubato, 13). Upon Tylenol's return, J&J added additional features to the bottle that made it challenging to open (Daemmrich, 68). The following year, congress responded by passing the Anti-Tampering act, making it a federal crime to meddle with packaged consumer goods (Ibid 23).

The advent of recombinant DNA in the 1970s allowed microbiologists to produce large quantities of cellular proteins (Galambos, Louis, and Jeffrey L. Sturchio, 257). The importance of rDNA should not be understated – it's allowed researchers to better understand the genetic profile of encoded proteins and, in life sciences, its ability to influence the translation process by

isolating and combining strands of DNA to produce proteins for therapeutic usage. Examples include insulin for treating diabetes, interferon in Hepatitis B and C, and erythropoietin for anemia (Ibid, 262). Interestingly, advances in recombinant DNA were not deployed by the classic, vertically integrated stalwarts of the 1950s and 60s. Characteristically, small firms commonly spun out from research institutions and labs led the way. By the early 1970s, vertically integrated manufacturers had spent decades accumulating expertise in microbial biochemistry and enzyme inhibition (Ibid, 255). These efforts culminated with sizeable capital outlays for several promising clinical drugs. For example, Smith Kline & French's (Now GlaxoSmithKline) Tagamet would become the first H-2 antagonist anti-ulcer drug in this period (Ibid, 255). Additionally, Glaxo's Zantac and Squibb's Captopril, targeting ulcers and preventing increases in blood pressure, would be other notable small-molecule advancements (Ibid, 255). For this reason, pivoting toward the new paradigm would prove challenging for prominent players. Instead of troubling themselves with identifying academics, employing, and building out even larger R&D teams, entrenched players pivoted towards making equity investments into promising biotechnology trials. The primary financing structures were in and out-licensing, which allowed entrenched players to acquire the patent rights to promising clinical drugs to develop and commercialize in specific geographies (Crama, 1539). Additionally, traditional joint ventures and non-dilutive and dilutive financing structures became commonplace. These financing structures remain ideal for small players not generating revenue because they provide financing and credibility. Entrenched operators gain proximity to cutting-edge research and upside exposure if the therapeutic proves successful.

Drug Discovery & Development

The few therapeutics that receive FDA approval must undergo a lengthy process that includes initial discovery, development, and clinical trials. All in all, this timeline typically lasts between five and thirteen years (Campbell 90). Additionally, the cost of this process has steadily increased each year, with dwindling efficacy rates in garnering FDA approval. Recent estimates suggest the cost of developing a therapeutic product range between 350 million and 2.8 billion (Ibid 93). Discovering a new therapeutic molecule typically takes between one and three years and requires several scientific disciplines to be effective. First, experts in biochemistry and genetics identify and validate potential compounds that may generate a favorable response against possible indications.

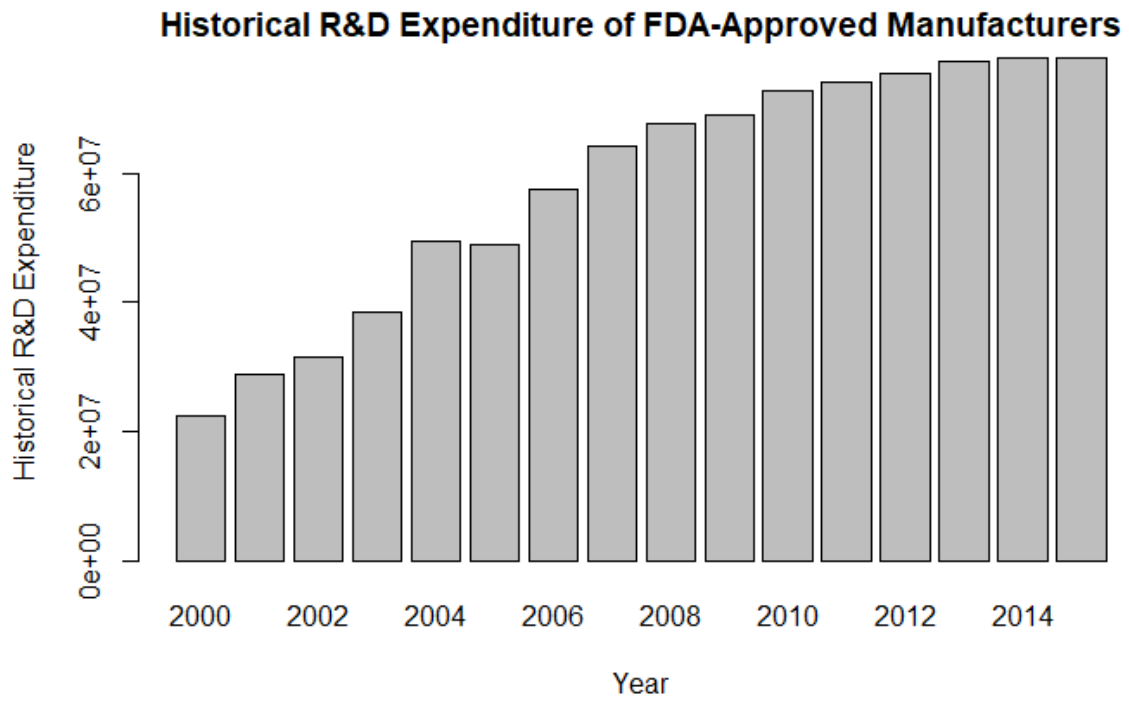

Figure A. Historical research and development expenditure of publicly traded FDA-approved manufacturers included within this paper's dataset.

From here, researchers utilize medicinal and analytical chemistry to screen potential compounds against targets in an in-vitro setting (Ibid). Next, researchers perform initial analysis of the molecule's impact on target cells, and the corresponding physiologic effects across the body, to test for toxicity in vitro and vivo settings (Ibid). The remaining molecules that make it through the preceding three steps enter pre-clinical development, which typically takes one to two years (Ibid). Researchers utilize animals to perform additional toxicity tests before moving into live humans (Ibid). Additionally, a replicable manufacturing process for their potential molecules is designed (Ibid). The formulation of a manufacturing process is the most variable component of development - novel molecules typically require a unique approach, while common molecules typically possess standardized pathways (Buckley, Kevin, and Alan G Ryder 1,086). For example, protein products like Monoclonal Antibodies (MAbs) are common due to their straightforward production process and high known purity (Shukla, Abhinav A., et al 171). Lastly, development concludes with submitting an Investigational New Drug (IND) application to the FDA (Campbell 97). To prove efficacy, firms typically pursue three phases (four in special cases) of clinical trials.

Food & Drug Administration (FDA) Approval Process

Historically, clinical trials last between three to eight years and serve as intermediaries between initial drug discovery and commercialization (Campbell 97). Clinical trials allow researchers to demonstrate their therapeutics' efficacy (or, unfortunately, in many cases, the lack thereof) in various sample sizes and study structures. Among life science observers, a randomized, double-blinded, placebo-controlled study is considered the golden standard for producing reliable data (Campbell 105). There are numerous other combinations – patient-

blinded, not randomized, sham-procedure, and controlled, not utilizing a placebo, to name a few (Ibid).

Before starting a clinical trial, applicants must submit an Investigational New Drug application or IND. Categorically, INDs are divided into substances seeking approval for research or commercial usage. Commercial INDs are categorized into emergency, treatment, and investigator INDs.

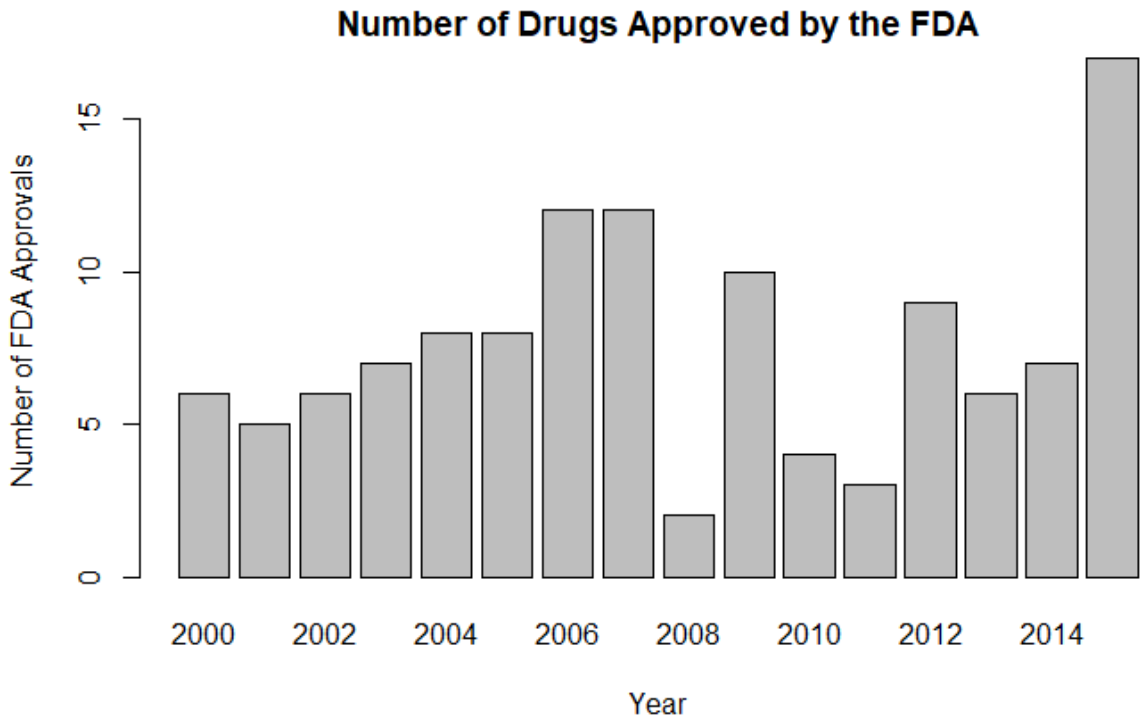

Figure B. Historical FDA approvals for publicly traded FDA-approved manufacturers included within this paper's dataset.

Note: *In 2015, 10 of the 18 drugs approved in the dataset were from Novartis*

The first clinical phase typically lasts between one to two years and administers a dosing regimen to between fifty and one hundred healthy patients (Campbell, 91). The importance of the first trial is to identify any potential side effects the molecule may have in a small sample of live humans (Campbell, 91). For this reason, patients are administered increasingly higher doses until

side effects appear. This is the critical endpoint - once side effects appear; clinicians determine that the previous dose is the maximum tolerated dosage in patients. In Phase one trials, only one dosage is typically used. The results of this study are shared with the FDA, which either approves or denies another round of trials.

In Phase two, the candidate therapeutic is administered to a broader sample size (500 – 1,000 patients) that contains the target indication (Campbell, 91). Like Phase one, a focus on side effects is important – however, the critical endpoint here is to determine the less common side effects present in patients with the target indication. Additionally, exploring therapeutic efficacy in reducing indication symptoms against a placebo or a current treatment standard is measured. Phase two typically lasts between one and three years. Similarly to the previous trial, advancement to phase three requires approval by the FDA. However, additional toxicity and safety data is incorporated into the results.

If approved, therapeutics move on to a phase three clinical trial, which often lasts between three to six years. Again, the sample size expands (1,000 – 5,000 patients) and contains only patients with the target indication (Campbell, 91). Here, safety, toxicity, and efficacy data points gleaned from previous trials are confirmed. In special cases, the FDA requires researchers to complete an additional phase four clinical trial to gather additional data (Ibid). For this reason, life science companies will often pre-emptively begin a phase four study immediately after submitting an NDA or BLA and before receiving a decision from the FDA.

After receiving approval from the FDA, life science companies can market their drugs directly to consumers by providing free samples, coupons, and direct advertisements (Campbell 172). Before a full FDA approval, firms are only permitted to advertise to consumers via

"unbranded" ads that do not explicitly state a product's name but implore individuals to "ask your doctor about [fill in the blank drug]."

Methodology

Hypothesis Statements

This paper seeks to identify if historically, superior research and development expenditure efficiency has a statistically significant impact on the forward earnings multiples public market investors award pharmaceutical manufacturers. Additionally, we're curious if "experience" (defined as the number of FDA approvals a particular manufacturer receives between a five-year period) has an impact on forward earnings multiples. The means of quantitatively assessing both of these questions are discussed in further detail below.

Historical Analyses

Multiple Imputation

This paper's computation of R&D efficiency requires an accurate account of historical drug approvals different manufacturers have received from the FDA. Fortunately, most historical clinical-trial data is free and publicly accessible via sources like Drugs@FDA and Clinicaltrials.gov. For this reason, information on drug indications, mechanisms of action, study structures, and endpoints are readily identifiable for most clinical trials. Unfortunately, the consistency of reporting can sometimes differ between researchers, resulting in challenges in identifying comparable data points. Although clinical-trial researchers must undergo a rigorous approval process and submit numerous disclosure filings, some data is lost. Additionally, previous literature has shown that researchers rarely return to updating their profiles after a trial fails, which only confounds the comparison between failed and successful trials. Traditionally,

this missingness issue led researchers like DiMasi and Hermann to utilize listwise deletion in their variable selection process, which had the unintended consequence of introducing bias into their results. In response, other academics like Lo and Sia incorporated multiple imputation techniques to fill out otherwise incomplete datasets. Although this paper performs simple variable removals, the author believes this does not introduce hypothesis-disruptive bias because this study's deletions are strictly related to maintaining variable standards (discussed further in the data collection and organization process section), not data inaccessibility.

Experience

Similar to this paper, other authors have sought to determine whether more "experienced" pharmaceutical manufacturers can benefit from achieving economies of scale. Notably, in 2001, Cockburn and Henderson examined research and development expenditure data from ten firms (smaller than this paper's analysis) and concluded that although scale was significant within specific therapeutic categories, in aggregate, scale was insignificant in producing an FDA approval. In the results section, this paper utilizes its materially larger dataset and study period to address this assertion (Danzon, 322).

Eroom's Law

In 2012, while observing the last few decades of pharmaceutical R&D efficiency, Scannel noted that approximately every nine years the pace of drug-innovation appeared to decline in lockstep with a continually accelerating cost to develop new therapeutics. As a nod to Gordon Moore's infamous "Moore's law," Scannel coined the term, "*Eroom's Law*" to describe his observation. Numerous other researchers like Hall (2016), Norman (2017), Halim (2019) and most recently, Miller (2023) have addressed the topic – primarily to comment on its existence or offer potential policy solutions to ameliorate the issue. Considering this is a widely discussed

topic in the pharmaceutical industry it makes this paper's analysis particularly interesting, as its findings can contextualize

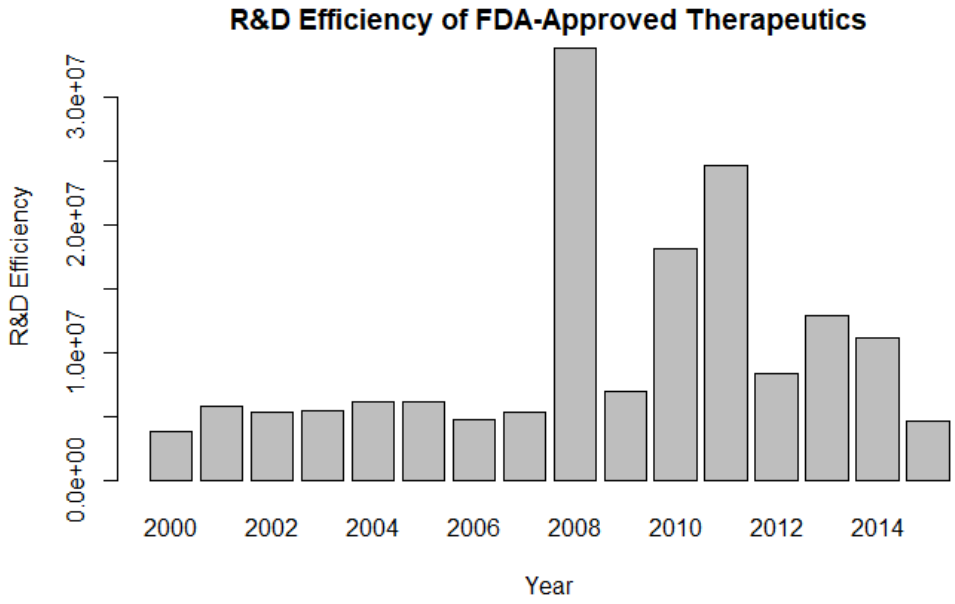

Figure C. Research and development expenditure efficiency between 2000 and 2015 of publicly traded manufacturers included within this paper's dataset

Techniques Utilized

Data Collection and Organization Process

The dataset utilized in this study was the result of combining two smaller datasets containing historical financials and valuation metrics with company-specific drug approval variables. The approved therapeutics utilized in this analysis were sourced from Drug Bank Online (hereafter referred to as DBO), a proprietary platform that aggregates relevant industry data. Dimensionality reduction was essential here since the initial DBO dataset far exceeded this study's scope. The initial DBO dataset included roughly 180 variables such as drug mechanism of action, pharmacodynamics, pricing, indications, sequences, reactions, and numerous others. Since this study only sought to analyze the historical relationship between R&D efficiency and

forward earnings multiples, the dataset was reduced to only include variables directly related to this purpose. Specifically, the refined DBO dataset utilized in this study included five variables – branded drug and generic names, date of marketing approval, date of market withdrawal, and the manufacturer's name. Branded and generic drug names were necessary variables because it allowed us to remove potential duplicates from the dataset and associate specific therapeutics with manufacturers which would be essential once linking the two datasets. Date of marketing approval was another essential variable as it was necessary to determine when specific manufacturers' therapeutics received FDA approval to market their drugs to the public. This approval date would later be cross-referenced against historical financials of the corresponding period. Additionally, date of market withdrawal was included as a variable to account for drugs like Refludan or Enbrel which were deemed unsafe after receiving FDA approval. Lastly, manufacturer's names were essential components of the dataset as it could be associated with specific therapeutics, approval and withdrawal dates, and historical R&D and valuation figures.

After refining the DBO dataset variables, further minor organization was necessary – duplicates were prevalent throughout the therapeutic dataset and were subsequently mass-removed. The final meaningful edit included removing all listed manufacturers that were not at one point publicly traded companies or data was difficult to source. A handful of manufacturers impacted by this received FDA approvals for their therapeutics but were later acquired by a larger entity like Pfizer or Merck. Regardless, the resulting dataset included 425 publicly traded manufacturers and 128 approved drugs between 2000 and 2015.

Historical multiples were collected utilizing Refinitiv, a software program commonly used by investment professionals to stream historical and live financial data. The multiples selected for this study include forward enterprise value to sales (EV/Sales) and forward price to

earnings (P/E). Considering a meaningful number of operators in our dataset were single-therapeutic manufacturers with products in clinical trial, the inclusion of solely a forward earnings metric (P/E) was deemed unsatisfactory as a measure of investor enthusiasm. In the same vein, we also chose to only include observations with positive earnings multiple as these are typically what investors closely observe. This decision shrank the P/E observations in our dataset from 6,493 to 826 – or a 90% reduction. Similarly, total forward EV/Sales observations shrank, albeit less dramatically than P/E (65% reduction from 6,517 to only 2,249). Nonetheless, the decision to constrain our analysis to only publicly traded pharmaceutical manufacturers and solely consider observations that correspond with positive earnings multiples, introduces potential survivorship bias to our analysis, which we address at the end of this section.

Forward multiples were determined more appropriate for this study than last twelve months multiples, which are inherently backwards looking. Although the past certainly influences investor expectations, so too do consensus estimates of the future. In short, this paper assumes that relying on historical forward multiples is a better litmus test for historical investor expectations. For ease of analysis, annual averages of each company's daily market-close forward earnings and sales multiple were utilized.

Single and Multivariable Regressions

This paper utilizes single and multivariable regressions to quantitatively assess if research and development expenditure efficiency has an effect on consensus forward earnings and sales multiples. In short, regressions are a statistical technique utilized to model the relationship between two or more variables. More precisely, regressions help researchers quantitatively identify the effect one or more independent variables (X_1) has on the value of a dependent variable (Y_1). In our analysis, R&D efficiency is the independent variable, and our forward

multiples represent the dependent variable. We also include additional variables in our extended analysis of each dataset. The following two equations are characteristic of how linear regressions are typically expressed within academic literature (Hoffman, 1).

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i \quad \{\text{or use } \beta_0 \text{ for } \alpha\} \tag{1}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2 / (n-1)} \tag{2}$$

Our primary goal in utilizing regressions is to estimate the slope and intercept of a line that best fits the data, so that we can use this line to make predictions about future values of Y_1 based on new values of X_1. The slope of the line (β_1) represents the change in Y_1 for a one-unit increase in X_1, while the intercept (β_0) represents the predicted value of Y_1 when X_1 is equal to zero.

Our regression analysis involves utilizing a technique called ordinary least squares (OLS), which quantifies the deviation between a model's observed and predicted values (Ibid, 2). Mathematically, this deviation (referred to as the Sum of Squared Errors or Residuals) can be expressed as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

The SSE (alternatively, SSR) equation squares the differences between each Y_1, which represents an observed value, and *Y hat*, which is the regression model's corresponding predicted value. This equation is important because it emphasizes the strength, or lack thereof, of a particular regression model's ability to predict or explain the relationship between several variables.

$$T = \sqrt{n}y_n/\sigma \tag{4}$$

$$p = 1 - \Phi(t), \tag{5}$$

To measure our regression's significance, we observe each of our model's resulting p-values (*p*), as shown in equation five (Hung, 11). The p-value operates under the assumption that the null hypothesis is correct and thus quantifies the likelihood of observing a predicted test-statistic (*T*) within the observed data. For this reason, if the p-value is below a previously defined significance threshold (typically 0.05), we can safely reject the null hypothesis.

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (\mathbf{y} - \mathbf{1}\bar{y})'(\mathbf{y} - \mathbf{1}\bar{y}), \tag{6}$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{7}$$

In addition to our model's p-value and F-statistics, we also utilize R^2 as an indication of our regression's relative prediction power. In short, R^2 is a useful shorthand for determining what proportion of our dependent variable's variance (*SST*) can be explained by the independent variable (Helland, 62). Equation six highlights how the dependent variable's variance is computed – the model simply subtracts the dependent's mean from each observed value. Next, in equation seven, the model's variance, SSE, is divided by the dependent variable's observed variance to compute R^2.

K-Means Cluster Analysis

Although our analysis deals with a limited range of variables, our examination is of one dataset segmented between manufacturers that received an FDA approval and those who did not. Unfortunately, due to the inclusion of FDA approvals in our R&D efficiency equation's denominator, it is impossible for us to compute this ratio for manufacturers who did not receive

an approval during our study period. To circumvent this issue and still provide meaningful comparison between each segment, we've opted to utilize a simple unsupervised clustering technique to highlight any apparent differences between each category.

K-mean clustering, or simply clustering analysis, is a widely used data mining tool typically deployed by researchers to assist in organizing large unlabeled or labelled datasets (Ikotun, 3). Although single and multivariable regressions are fantastic at determining what relationship two or more groups may possess, it requires a priori variable selection, whereas cluster analysis allows researchers to determine which categories are most important. Additionally, the regression model's core assumption is that our variables possess a linear relationship – clustering on the other hand, allows us (with the assistance of useful visuals) to identify non-linear relationships that deserve additional analysis.

As the name suggests, the first parameter in a K-means cluster analysis is the identification of the appropriate quantity of clusters to segment the dataset (Hu, 2). This can be mathematically expressed as follows:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

Here, dataset *X* is segmented into *K* clusters, *C,* so as to minimize the sum square error of each cluster (Ikotun, 3). In short, the goal in clustering is the selection of data points with high intra-cluster similarity, whilst boasting low inter-cluster similarity. Stated differently, clustering analysis aims to identify groups of data points that are rich in commonalities, whilst ensuring that data points in other clusters are as dissimilar as possible. For this reason, a useful analysis including a K-means algorithm aims to reduce within cluster sum of squares (WSS) while segmenting the

data into as few clusters as possible, as shown in Figure D. In this paper's analysis, this will be especially useful in organizing our three datasets into smaller, more comparable groups.

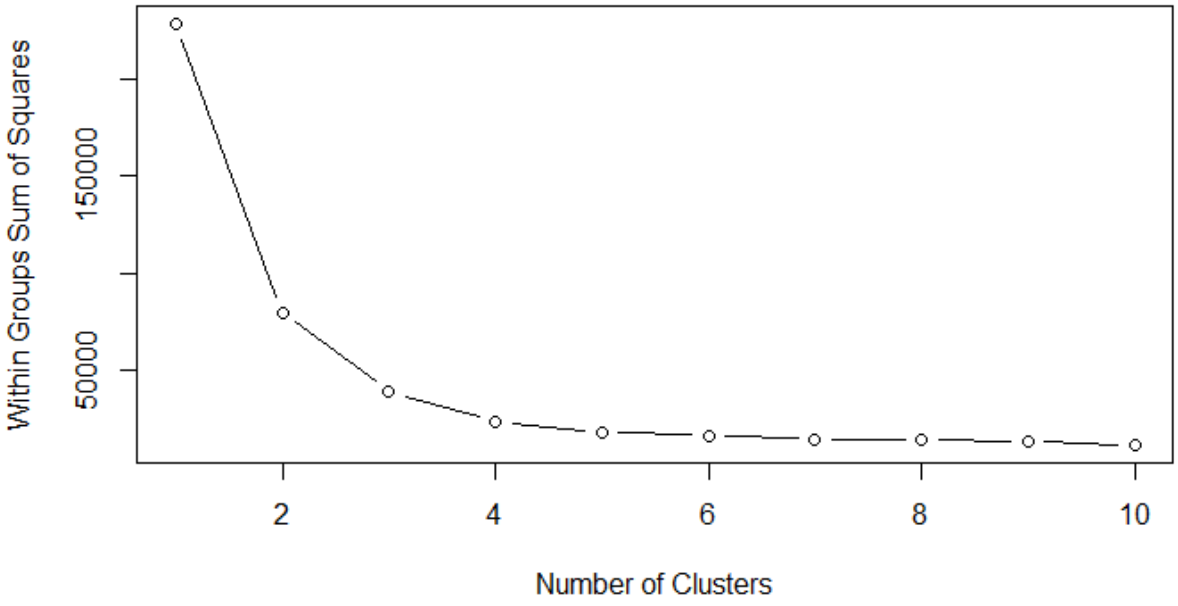

Figure D. Within group sum of square (WSS) and cluster minimization. Colloquially referred to as the "elbow method" of identifying the ideal quantity of clusters in a dataset.

Boxplots and Histograms

Lastly, in addition to deploying traditional single and multivariable regression analyses and creating categories to compare our datasets, this paper also utilizes visual aids such as histograms and boxplots. Recall, our dataset includes both well-known, prominent manufacturers like Pfizer and Johnson & Johnson, but also small firms like Acer Therapeutics and Acura Pharmaceuticals. To address this discrepancy, we further categorized firms by small (1), medium (2), and large (3) scale manufacturers. The decision to do this was predicated on our findings that firm size (by market capitalization) has a statistically significant impact on the scale and

efficiency of research and development expenditure. Thus, it was posited that if firms of comparable sizes typically trade at similar earnings multiples, perhaps the relationship between R&D efficiency and forward multiples were similar for these firms.

Equations

Research & Development Expenditure Efficiency

This paper computes research and development efficiency by dividing a manufacturer's total R&D expense during a specified period by the number of FDA approvals correspondingly received.

$$Research\ and\ Development\ Efficiency = \frac{R\&D\ Expense}{FDA\ Approvals}$$

Unfortunately, we weren't able to identify more granular breakdowns of R&D's devotion towards particular internal projects (such a task would likely prove arduous or even impossible, considering these are multi-billion-dollar capital commitments made by entities with numerous stakeholders) and thus have to rely on the accuracy of publicly disclosed SEC filings. However, assuming on aggregate that most Analysts focus on publicly available figures to make their estimations, we believe utilizing GAAP (Generally Accepted Accounting Principles) research and development expense is a close-enough proxy for computing the relative efficiency of dollars deployed towards clinical trials and in extension, FDA approvals.

Forward Price / Earnings

The computation of Price / Earnings is a two-fold process involving the union of the prevailing market price with estimated future earnings per share accounting figure. According to the academic literature, equity valuations (and thus price) reflect the cumulative opinion of market participants on the quantity and speed at which a particular company may generate future

cash flows, discounted to the present day (Chen, 845). This relationship can be modelled mathematically as follows:

$$P_t = \sum_{k=1}^{T} \frac{FE_{t+k}(1-b_{t+k})}{(1+q_t)^k} + \frac{FE_{t+T+1}}{q_t(1+q_t)^T} = f\left(c^t, q_t\right), \quad (1)$$

Here, P_t represents the prevailing market price, FE_{t+k} is the implied future consensus earnings estimate, k is the number of years forecasted, $1\text{-}b_{t+k}$ is the assumed payout ratio, and q_1 is the weighted average cost of capital, or discount rate applied to future cash flows.

Most importantly in the above computation of price is the application of a discount rate, which we've included the derivation for below (Mejia-Pelaez, 55).

$$E_{t-1} = \frac{E_t + CFE_t - (Ku_t - Kd_t).D_{t-1} + (Ku_t - \psi_t).V_{t-1}^{TS}}{1 + Ku_t} \quad (2)$$

$$V_{t-1} = \frac{V_t + FCF_t + TS_t + (Ku_t - \psi_t).V_{t-1}^{TS}}{1 + Ku_t} \quad (3)$$

$$V_N^{TV\_L} = \frac{FCF_{N+1}}{(Ku - g).\varphi} \quad (4)$$

Equations two and three are critical components of this analysis because the weighted average cost of debt and equity capital applied to the consensus-estimated future stream of cash flows determines the prevailing market price. Functionally, this formula asserts that in a rising discount rate environment, all else equal, equity prices should depreciate and vice versa. The other components in WACC, the cost of equity and debt, have separate equations. From left to right in equation two, *E* represents equity value, cash flow to equity is identified as *CFE,* the unleveled

cost of equity is *Ku*, the cost of debt is *Kd*, *D* is the market value of debt, V is the firm's market value, and free cash flow is *FCF*. However, to keep this section brief, we only identified the formula for calculating the terminal value (equation four) as another critical component of price.

Estimating future earnings per share (EPS) is comparatively simple - in theory at least.

$$Net\ Income\ or\ \pi(q) = R(q) - C(q) \tag{5}$$

$$\text{Earnings per Share} = \frac{\text{Net Income}}{\text{Total Shares Outstanding}} \tag{6}$$

Calculating future EPS requires estimating a firm's residual earnings after devoting revenues to direct, operational, and non-operational expenses. Alternatively, consensus estimates are often easily obtainable for most sizeable publicly traded companies by sell-side equity research departments or a data provider like Bloomberg or Refinitiv. Either way, net income is subsequently divided by a firm's total outstanding shares (typically diluted shares) to arrive at earnings per share. Lastly, the forward price/earnings multiple is derived by dividing the prevailing market price calculated in formulas one through four by EPS.

$$Price\ to\ Earnings\ Multiple = \frac{Market\ Price}{Earnings\ per\ Share} \tag{7}$$

Forward Enterprise Value / Sales

Like calculating forward P/E multiples, EV/Sales is a union of a firm's enterprise value with an estimated future revenue accounting figure. The enterprise value calculation combines a firm's net debt with market capitalization.

$$Market\ Capitalization = Market\ Price\ x\ Total\ Shares\ Outstanding \tag{8}$$

This involves simply multiplying the prevailing market price yielded from equation one against the total outstanding shares.

$$Net\ Debt\ =\ Total\ Debt - Total\ Cash \quad (9)$$

$$Enterprise\ Value = Net\ Debt + Market\ Capitalization \quad (10)$$

Lastly, the firm's total cash and cash equivalents are netted from outstanding debt holdings to compute net debt, which yields enterprise value after being added to market capitalization.

The revenue portion of the forward EV/Sales equation derives from market participants' consensus expectations on a firm's future output potential (*Q*), multiplied by a price (*P*) (Pindyck, 284).

$$Revenue\ =\ Q * P \quad (11)$$

Again, the calculation of our forward multiple is obtained by dividing the current market-determined figure (enterprise value) by a forward accounting metric (independently estimated or consensus-reported forward revenue).

$$Enterprise\ Value\ to\ Sales\ Multiple\ =\ \frac{Enterprise\ Value}{Revenue} \quad (12)$$

Limitations

Survivorship Bias

As noted earlier, although this paper utilizes R&D efficiency to determine if a relationship exists with forward EV/Sales and P/E multiples, this substantially reduced our original dataset to only include operators who generate revenue (or are forecasted to do so in the near future) or produce a positive EPS figure. Most publicly traded life science firms in our original dataset did not have a previously approved drug, so most operators neither generated

revenue nor positive EPS. For this reason, we divided our analysis into three segments – one focused solely on manufacturers that received FDA approvals, another on those who did not, and lastly, a consolidated analysis including both. We believe that by utilizing this approach, our study meaningfully addresses any accusation of survivorship bias, as we will share findings from each component of our dataset and meaningfully compare the results.

Mergers and Acquisitions

For uniformity, we calculate R&D efficiency by utilizing historical research and development expenditure figures publicly disclosed in routine quarterly and annual SEC filings. The downside here was that manufacturers who were not at one point publicly traded companies were removed from the dataset. This raises a potential issue with data collection because numerous privately-operated manufacturers received FDA approvals during our study period but were later (or during the study period) acquired by a larger, publicly traded entity in our dataset like Pfizer or Merck. This issue is most acute for private manufacturers who were acquired a year or two prior to receiving an FDA approval but were nonetheless credited in our dataset as having developed the therapeutic. Since our dataset did not discriminate between parent and absorbed companies' approvals (and we deleted private companies), some of our more acquisitive manufacturers may be unfairly penalized for their additional R&D expense. Future analysis should consider how to account for these acquired therapeutics more precisely.

Explicit Research & Development Breakdown

Additionally, we could not identify a more granular breakdown of pharmaceutical manufacturers' expenditure toward developing particular therapeutics. Similar to the issue with M&A, our analysis' emphasis on uniformity and reliance on publicly disclosed SEC filings constrained our ability to credit cases where a particular manufacturer may outline in an investor

presentation (or other investor materials) explicitly what R&D was funding. Recall, assuming perfect execution from initial research to FDA approval, on average, is a ten plus year process. For this reason, we believe R&D expense is effectively a long-term capital outlay, which is a necessary component of pharmaceutical manufacturing businesses and can thus be utilized as a proxy for computing the relative efficiency of dollars deployed towards clinical trials and in extension, FDA approvals.

Additional Factors

Lastly, this paper recognizes numerous quantitative and qualitative factors unrelated to research and development expenditures could possess a more straightforward explanation for forward multiples at the micro level. For example, a manufacturer trading at a low forward P/E multiple relative to peers and historical norms could reflect consensus opinions on an ongoing lawsuit (i.e., Johnson and Johnson's infamous Talcum tort lawsuits) or perhaps the upcoming termination of a lucrative patent. Future analysis should attempt to incorporate these additional factors.

Results

In short, between 2000 and 2015, the data suggests that the efficiency of publicly traded pharmaceutical manufacturer's research and development expenditure does have an impact on forward P/E multiples, but not EV/Sales. However, as will become increasingly apparent throughout this section, the relationship that R&D efficiency possesses with forward P/E multiples appears miniscule at best. For this reason, the majority of this paper's discussion will center on exhaustively isolating each variable across all three of our datasets and analyzing their mutual relationships to gain a better understanding of our initial hypotheses.

Before launching into the primary hypothesis results, we will briefly address our experimentation with lags. As noted in the previous section, this paper views R&D as an essential long-term capital outlay – however, identifying the objectively correct period to quantify and compare its efficiency between operators is effectively an impossible task with the number of manufacturers in our dataset. For this reason, we implemented lags to see if the study results would materially differ. In short, it did not. There was a marginal improvement in the exhibited statistical relationship between forward P/E and EV/Sales when R&D efficiency was lagged by two years, however, considering our dataset was already quite limited after the initial selection process, the inclusion of a lag further reduced our observation count. For this reason, the paper only presents results that do not incorporate a lag on either variable.

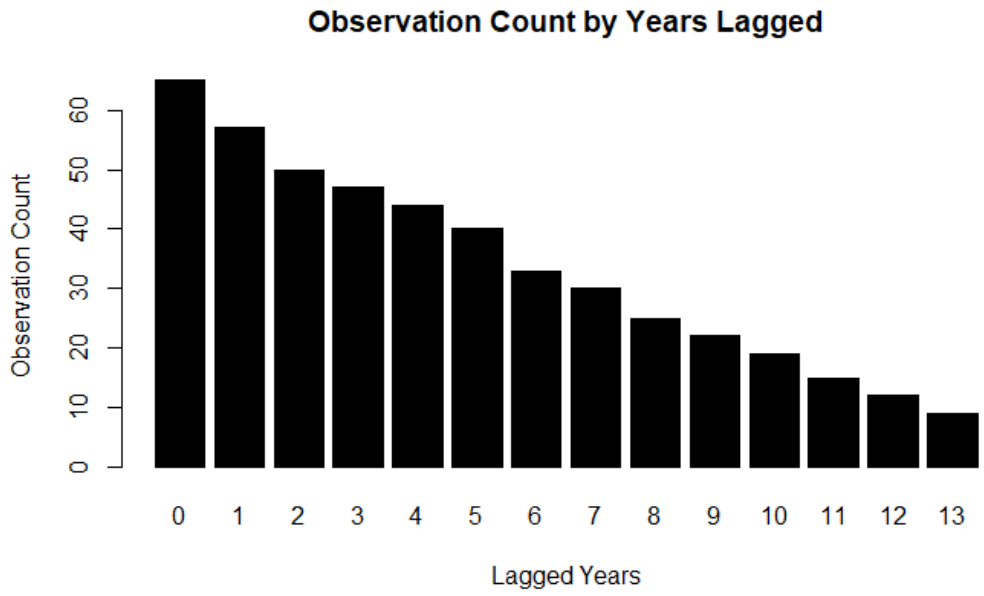

Figure 1. Observation count by number of years lagged

Single Variable Regressions

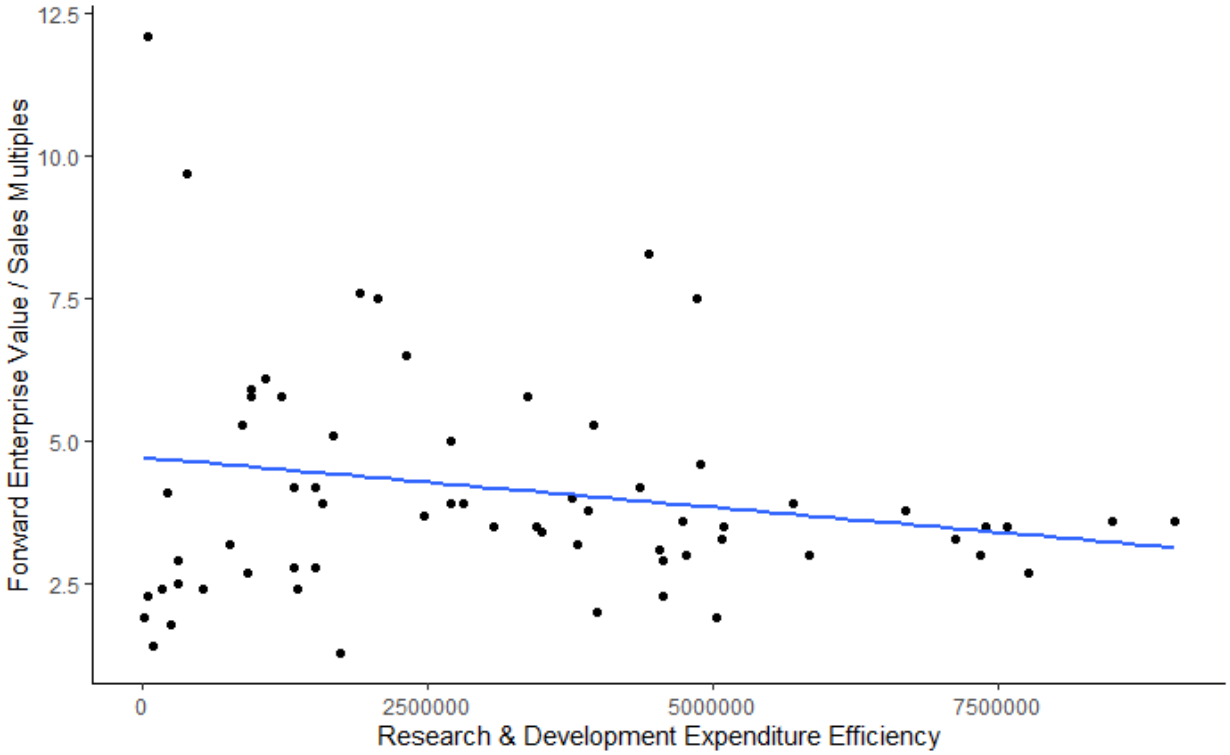

Figure 2. The effect of research and development expenditure efficiency on forward Enterprise Value / Sales (*F*~(1,63)~ = 2.448; *P* = 0.1227). The regression line is explained by Y = 4.722X - 1.751e-07. The Adjusted R-Squared is 0.022.

Interestingly, the relationship between research and development expenditure efficiency and forward enterprise value to sales multiples is insignificant. Unfortunately, the lack of a relationship is quite apparent in figure one – as R&D efficiency deteriorates, the forward sales multiple applied to these manufacturers appears to be completely random. To clarify, recall that an improvement in R&D presents as a nominal reduction in our efficiency figure since the FDA approval input rests in the equation's denominator. Thus, as a manufacturer receives more approvals, the nominal R&D efficiency figure should decline if less incremental expense is required to obtain a marginal approval. In figure one, its apparent that whether or not efficiencies

in R&D are achieved, it has no bearing on the forward EV/Sales multiple a particular manufacturer received during our study period.

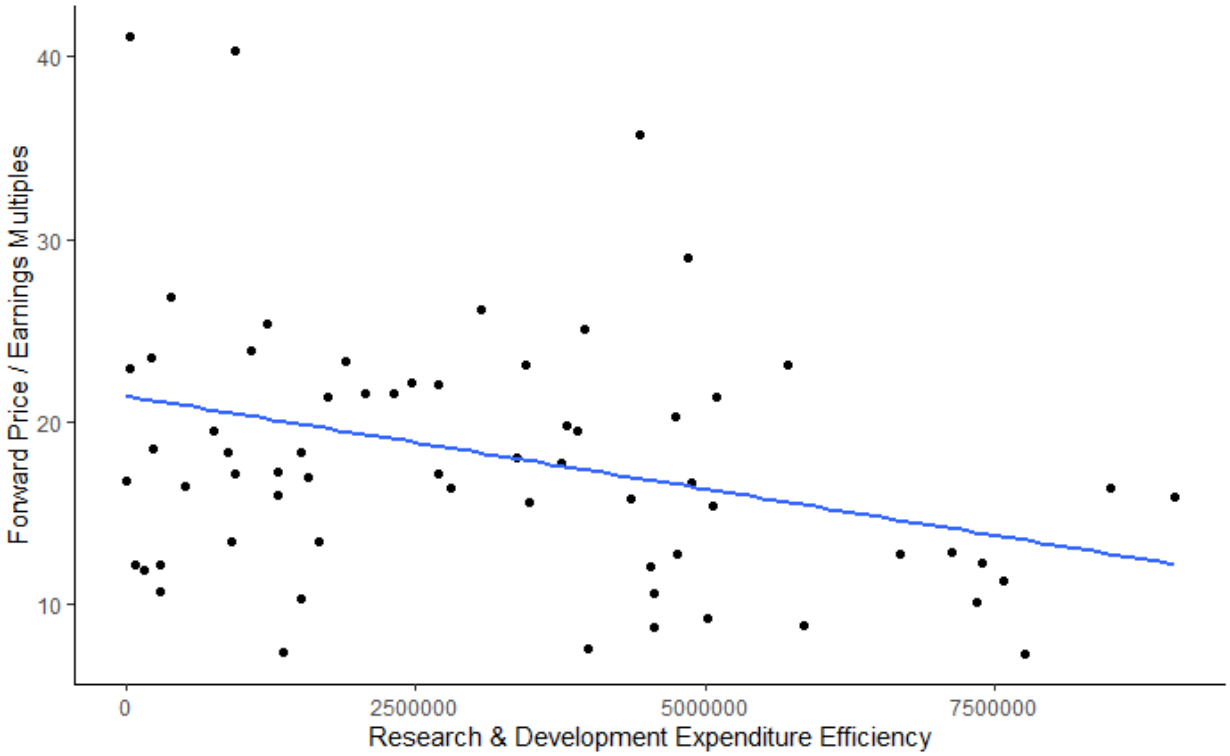

Figure 3. The effect of research and development expenditure efficiency on forward Price / Earnings (*F*~(1,63)~ = 2.448; *P* = 0.007938). The regression line is explained by Y = 2.141e+01X - 1.018e-06. The Adjusted R-Squared is 0.092.

In contrast to forward EV/Sales, our findings suggest that research and development expenditure efficiency did have a statistically significant impact on forward P/E multiples. This discovery was somewhat surprising at first glance because, as noted earlier, many of the manufacturers in our initial dataset were single-therapeutic operators that did not generate positive accounting net income. However, as we show later, the average firm within our approved dataset was materially larger than those within the unapproved or consolidated dataset. For this reason, we believe it makes sense that public market investors prioritize earnings-related

metrics over a sales-oriented one. Although the adjusted R-squared is small, our p-value is below the 0.05 threshold – thus, in combination with positive coefficients, suggests that operators who exhibit superior R&D efficiency receive correspondingly higher P/E multiples and vice versa.

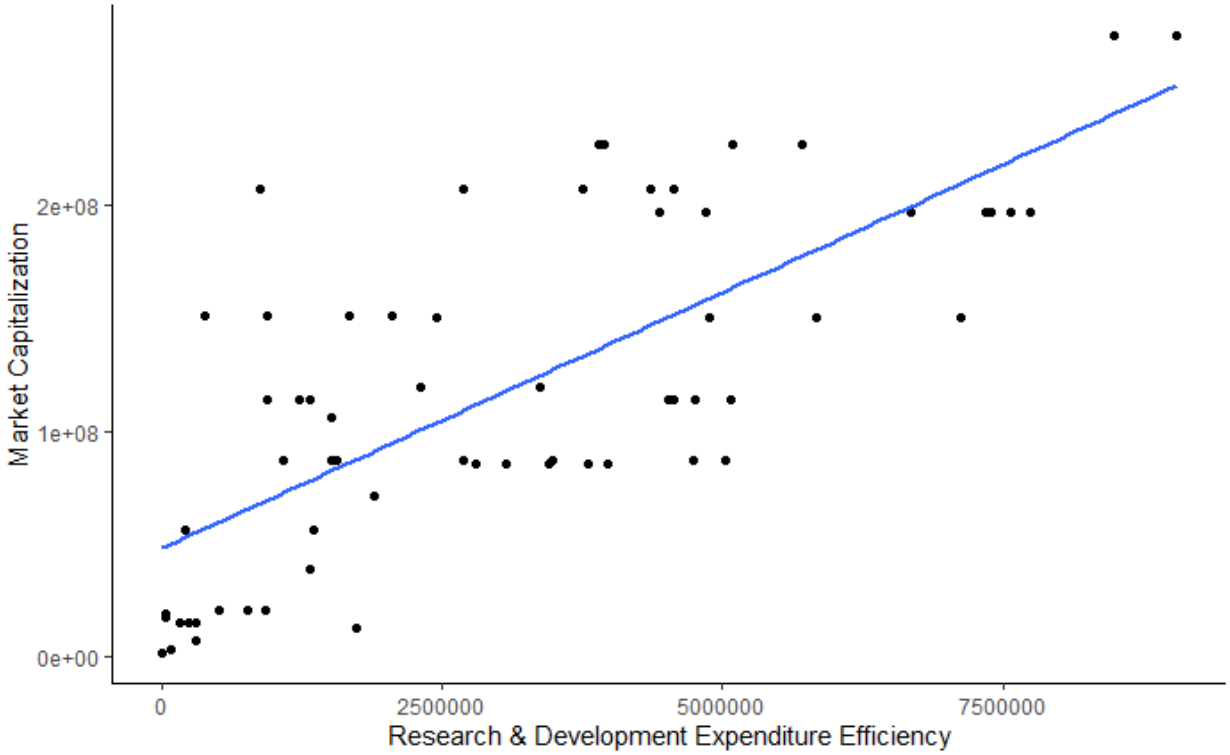

Figure 4. The effect of research and development expenditure efficiency on market capitalization (*F*~(1,63)~ = 75.479; *P* = 2.252e-12). The regression line is explained by Y = 4.809e+07X + 2.263e+01. The Adjusted R-Squared is 0.537.

Figure four contextualizes our previous findings in a helpful manner while also addressing other academic's questions surrounding the benefits of scale economies in utilizing R&D more efficiently to garner approvals. Although it's not a given that firms with larger market capitalizations necessarily have "economies of scale," it's fair to assert that more of these manufacturers are scaled operators. Thus, it's interesting that one of the strongest statistical relationships exhibited in this study actually shows that lower R&D efficiency is characteristic of

larger firms. In other words, firms with the most efficient deployment of R&D dollars are typically smaller manufacturers – thus invalidating the persistent theory within pharmaceutical academic literature that size and scale are potentially a benefit.

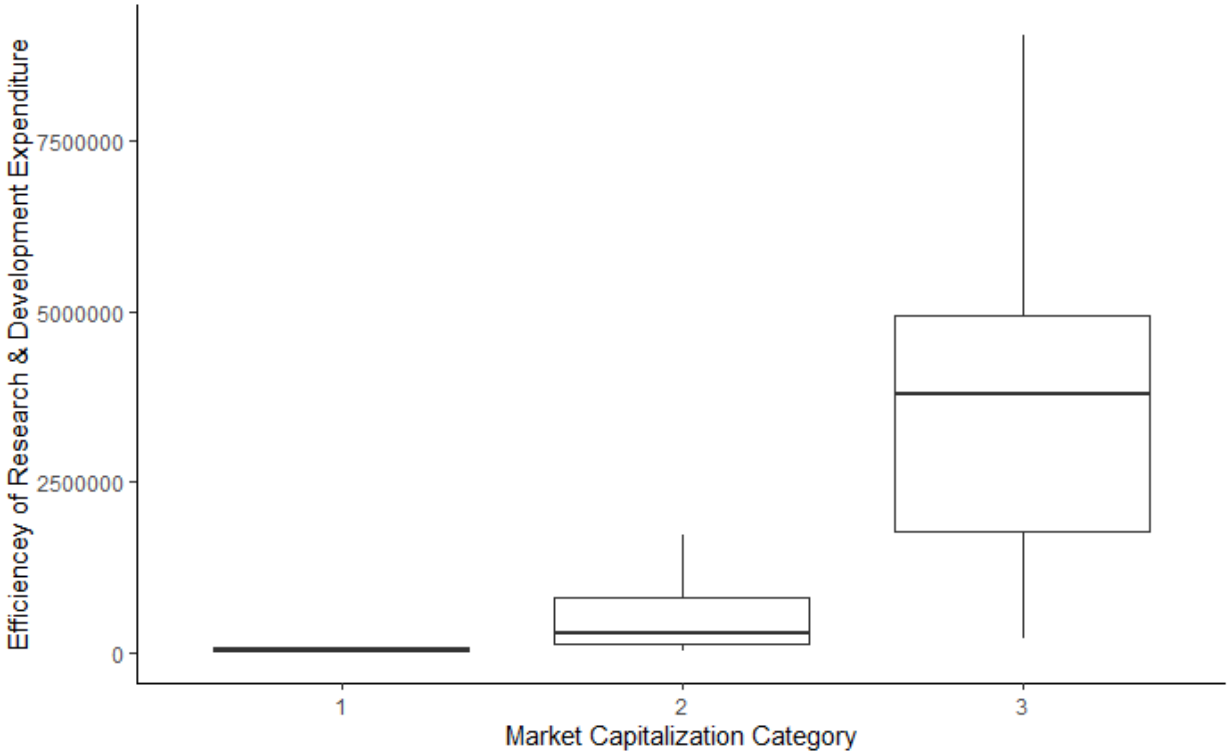

Figure 5. The effect of research and development expenditure efficiency on Size Factor (*F*~(1,63)~ = 32.493; *P* = 2.373e-06). The regression line is explained by Y = 2.301e+00X + 1.319e-07. The Adjusted R-Squared is 0.3298.

Identifying that smaller manufacturers were typically the most efficient R&D allocators was a fascinating discovery – so, as discussed earlier, we further segmented the approved dataset into three categories by their respective market capitalization size. Again, the relationship exhibited in figure four persisted, albeit with a slightly smaller adjusted R-squared (0.32 vs. 0.53 when not explicitly segmented by size). Nonetheless, our findings appear to suggest that

economies of scale or size, although an enviable position, does not guarantee management is more efficient at allocating R&D dollars towards projects likely to garner an FDA approval.

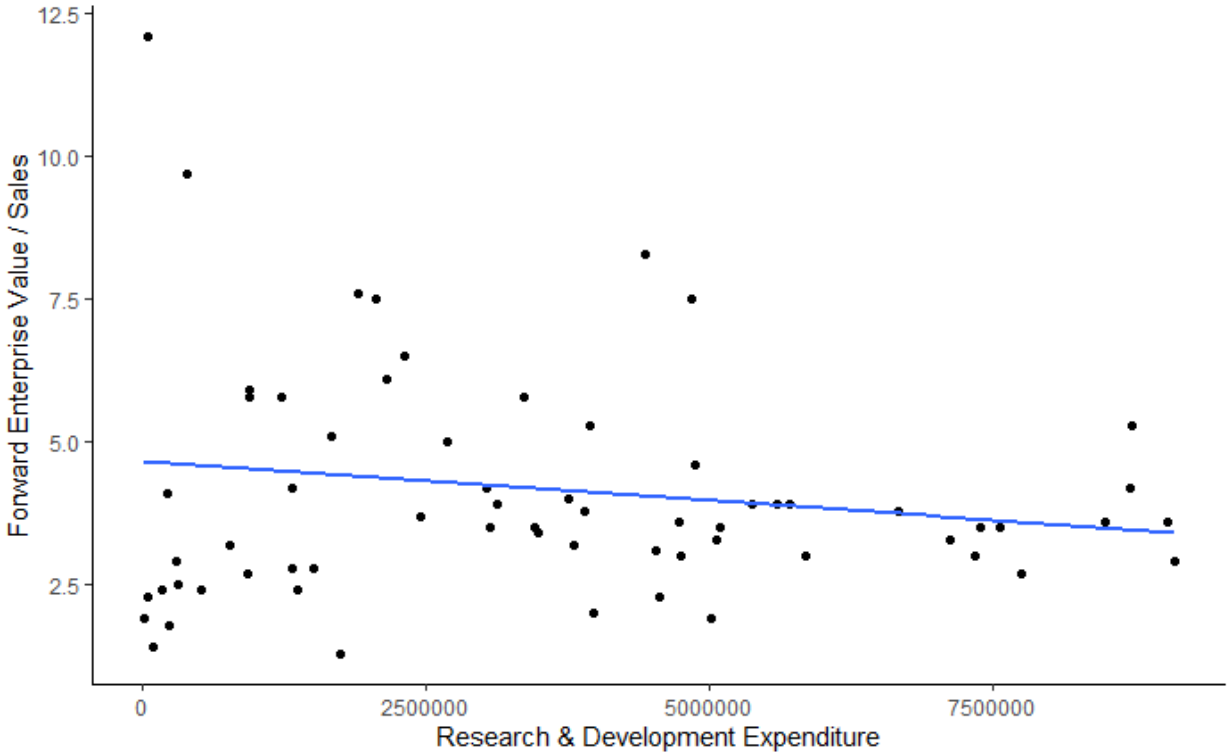

Figure 6. The effect of research and development expenditure on forward enterprise value / sales (*F*~(1,63)~ = 1.8027; *P* = 0.1842). The regression line is explained by Y = 4.659e+00X - 1.367e-07. The Adjusted R-Squared is 0.0123.

Due to the lackluster results that were immediately apparent in both figures two and three, we were curious which of our variables exhibited a stronger relationship with the response variables – research and development expenditure or FDA approvals? Similar to this paper's earlier findings, we found that research and development expense had a statistically insignificant impact on forward enterprise value to sales multiples for pharmaceutical manufacturers. The reasoning for this is likely multi-factorial – however, this paper believes some of the discrepancy is due to public-market investors potentially relying less on forward EV/Sales multiples for

larger firms, which, as noted earlier, characterizes most of the manufacturers within the approved dataset.

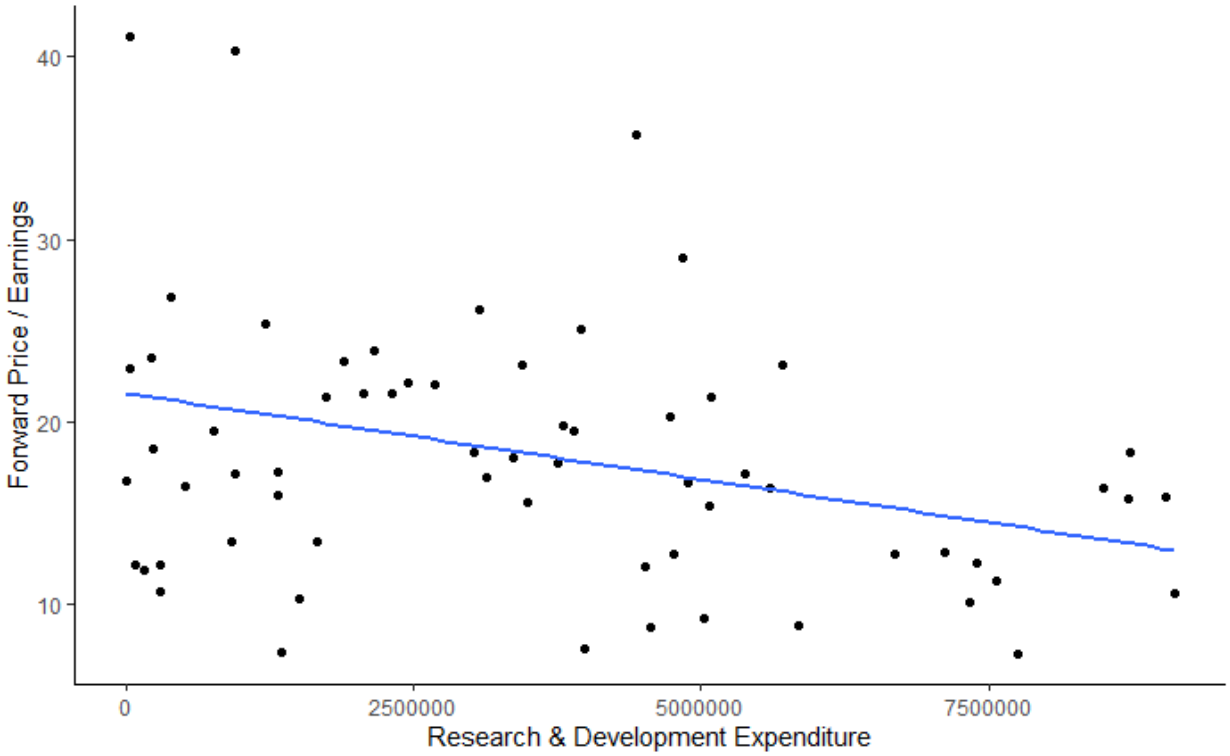

Figure 7. The effect of research and development expenditure on forward price / earnings (*F*~(1,63)~ = 1.8027; *P* = 0.0066). The regression line is explained by Y = 2.156e+01X - 9.412e-07. The Adjusted R-Squared is 0.09714.

Similar to the discrepancy identified previously between R&D efficiency and forward EV/Sales and P/E – nominal R&D expense has a statistically more significant impact on forward P/E than EV/Sales. Again, this paper believes this is due to public-market investors' focus on utilizing earnings-related metrics for larger companies, which is characteristic of the manufacturers within the approved dataset. However, it's important to note that although the relationships between R&D efficiency and expense possess a statistically more significant

impact on forward P/E than EV/Sales, it's nonetheless miniscule – the adjusted R-squared is only 0.097.

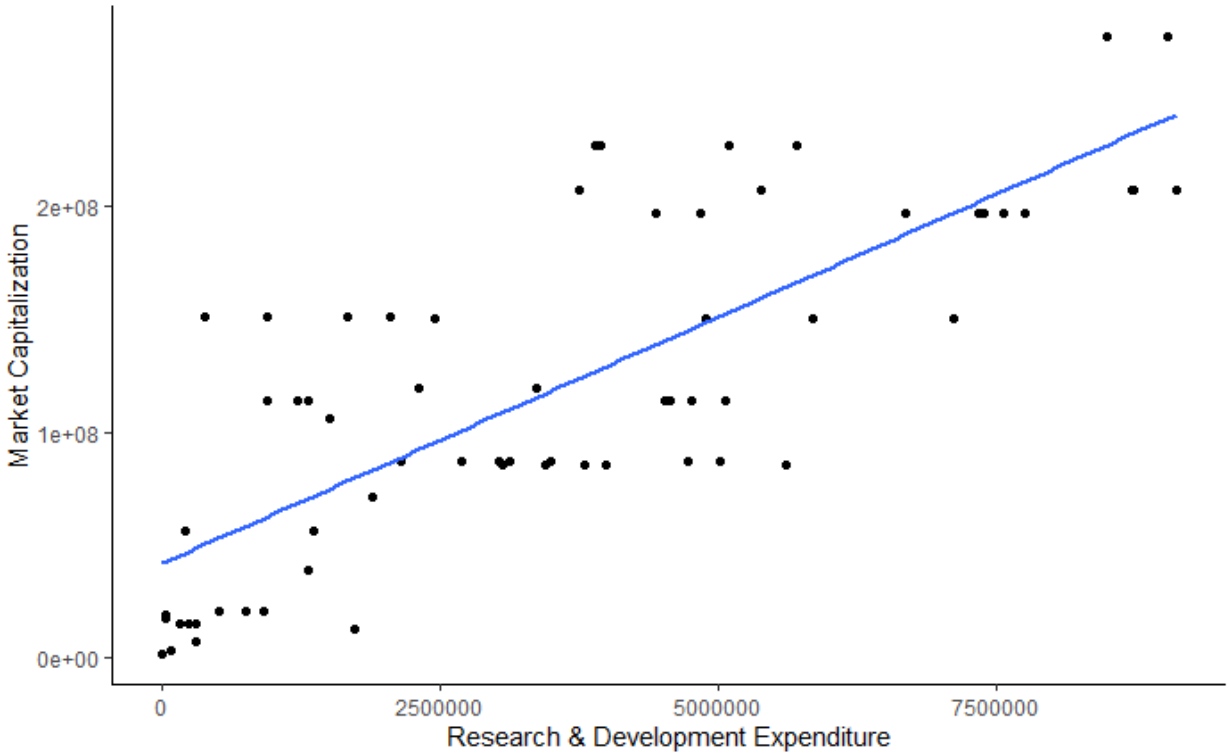

Figure 8. The effect of research and development expenditure on market capitalization (*F*~(1,63)~ = 101.4; *P* = 9.548e-15). The regression line is explained by Y = 4.184e+07X + 2.179e+01. The Adjusted R-Squared is 0.6107.

As anticipated, the scale of pharmaceutical research and development expenditure appears to have a material impact on market capitalization. In other words, larger firms possess the resources, either due to cash flow generated from internal operations, or ample access to equity and debt markets, to allocate a nominally larger quantity of capital towards research and development efforts in comparison to smaller peers. This is of course not a novel discovery, but it extends this paper's discussion about economies of scale – although larger firms can and do

dedicate more resources towards developing promising therapeutics, they were not more capable than smaller peers at selecting projects most likely to obtain an FDA approval. Figure nine

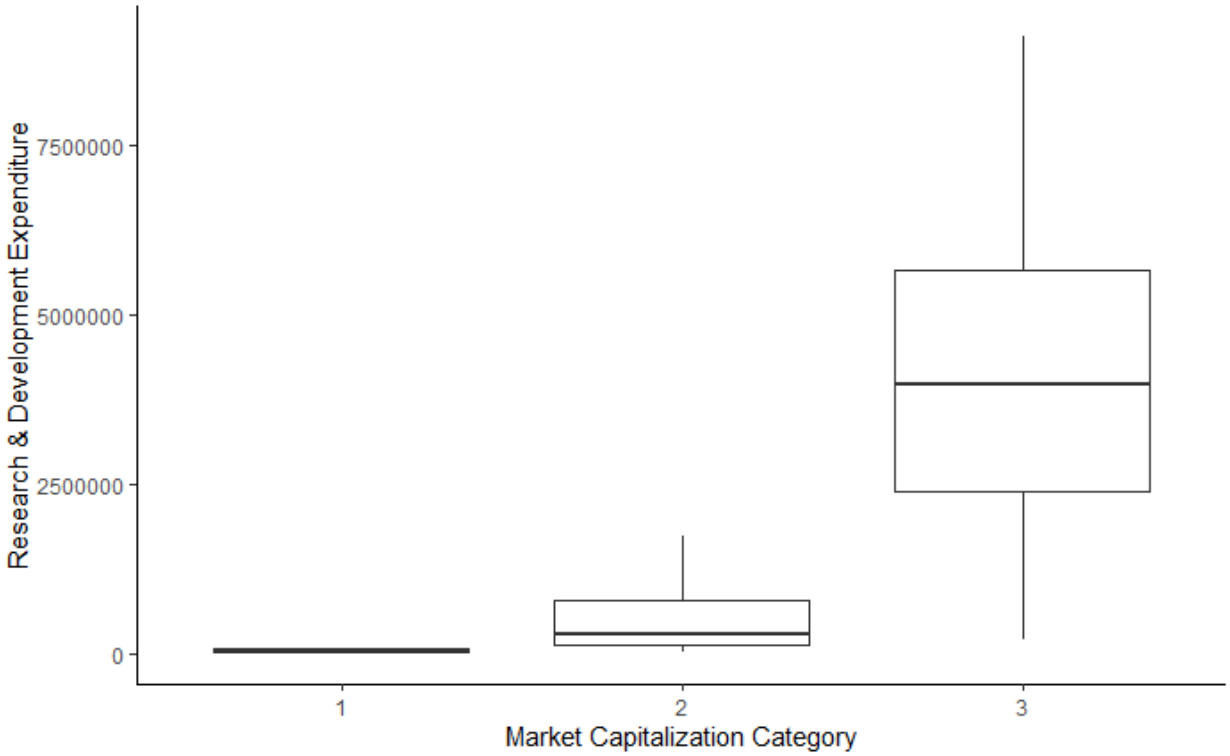

Figure 9. The effect of research and development expenditure on Size Factor (*F*~(1,63)~ = 37.172; *P* < 7.257e-08). The regression line is explained by Y = 2.273e+00X + 1.246e-07. The Adjusted R-Squared is 0.3611.

serves as a useful benchmark for emphasizing the nominal difference in expenditure dedicated towards R&D efforts at larger pharmaceutical firms in comparison to smaller peers. Frankly, the difference is magnitudes. In short, although a relatively linear relationship exists between a firm's market capitalization and research and development expenditure, this correlation appears to break down with FDA approvals.

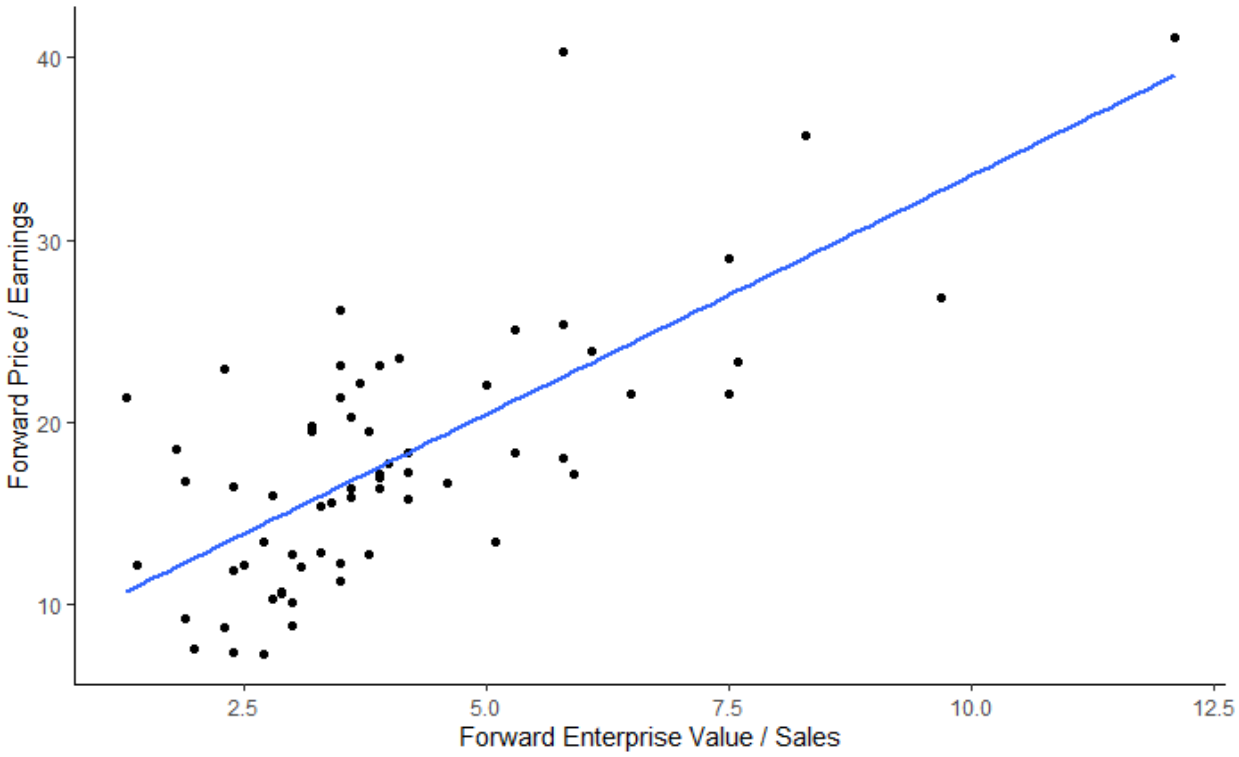

Figure 10. The effect of forward enterprise value / sales on forward price / earnings (*F*~(1,63)~ = 86.39; *P* < 2.008e-13). The regression line is explained by Y = 7.3239X + 2.6186. The Adjusted R-Squared is 0.5716.

Before moving into the multivariable regressions, this paper also sought to contextualize what relationship, if any, existed between historical forward P/E and EV/Sales multiples for pharmaceutical manufacturers within our approved dataset. Unsurprisingly, the relationship was statistically significant – in fact, it would have been odd if there was a material discrepancy here, as both metrics utilized in this analysis rely on historical forward consensus accounting earnings figures. Similarly, it would have been odd if the adjusted R-squared was 1 – this is because mechanically, EV/Sales and P/E do not always move in the same direction. For example, consider a situation where an Analyst forecasts revenue to grow, but envisions margins deteriorating due to rising input costs or merger-related expenses – assuming price, market

capitalization, enterprise value, and shares outstanding were held constant, the price / earnings multiple would remain the same, but enterprise value to sales would decline due to the elevated topline. Although this paper did not examine each observation individually and determine the factors influencing different firms' sales or earnings multiples during the study period, it's important to note the strong likelihood that numerous influences outside of this analysis are responsible for impacting the results.

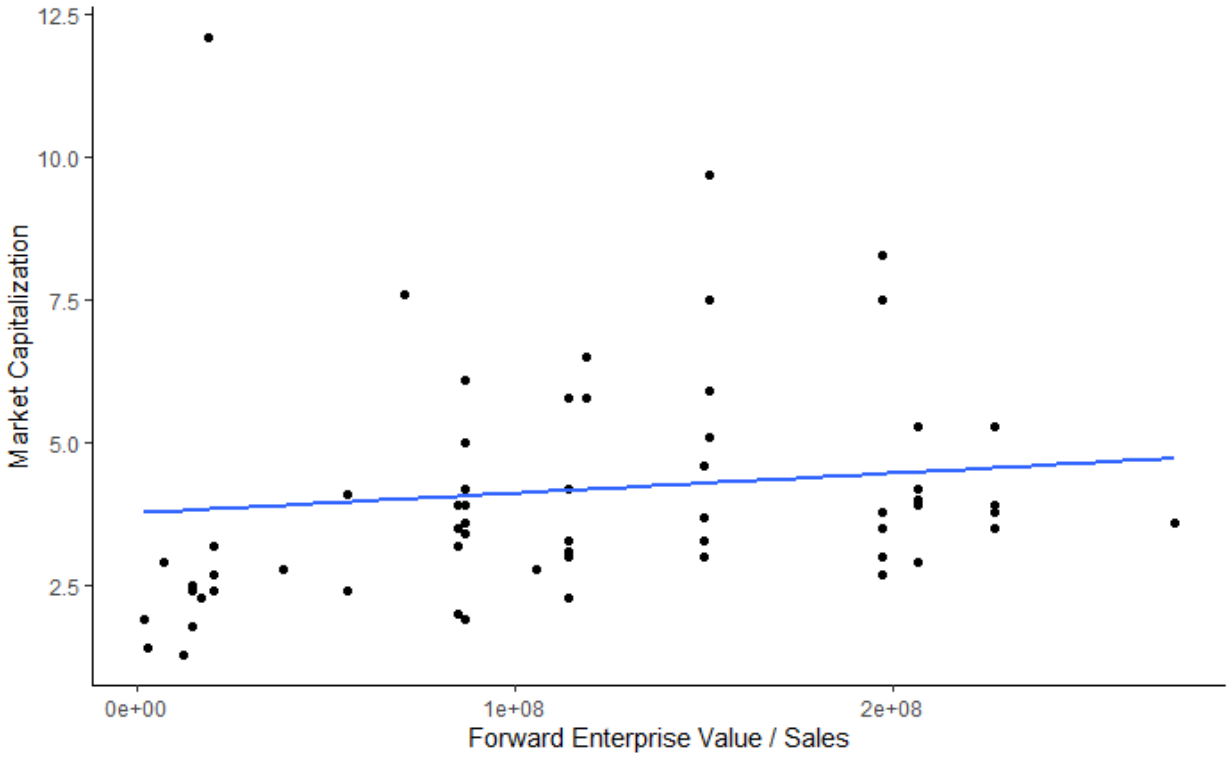

Figure 11. The effect of forward enterprise value to sales on market capitalization (*F*~(1,63)~ = 0.891; *P* = 0.3488). The regression line is explained by Y = 101183036X + 3998133. The Adjusted R-Squared is -0.0017.

Peering further into the drivers underlying pharmaceutical manufacturers' forward multiples, figure eleven suggests that forward enterprise value to sales multiples do not

meaningfully differ between manufacturers of varying sizes. This observation conforms with additional regression analyses performed on each category separately. In boxplot format,

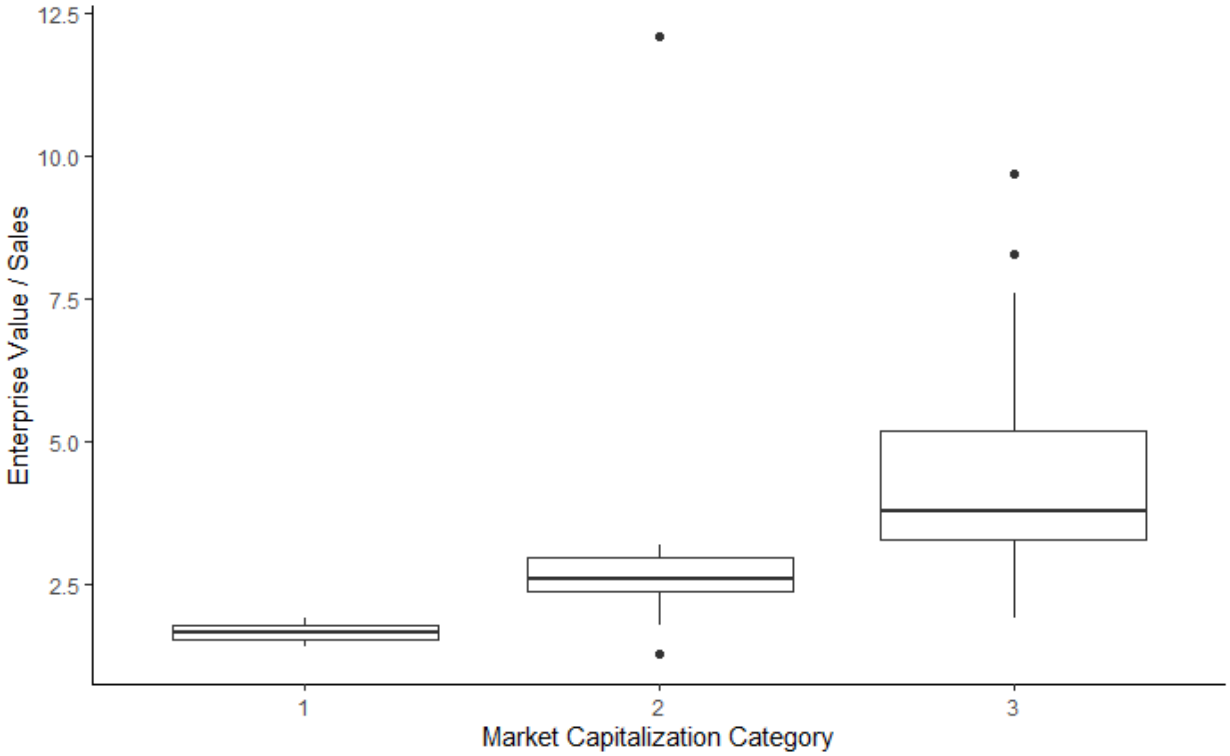

Figure 12. The effect of market capitalization category on forward enterprise value / sales multiples (*F*~(1,63)~ = 2.4088 ; *P* = 0.1237). The regression line is explained by Y = 2.50737x + 0.04791. The Adjusted R-Squared is 0.02154.

its clear a few extreme outliers within the mid and large-scale firm categories meaningful skew their respective segment mean enterprise value to sales multiples. A similar result is also observed with forward price to earnings multiples – market capitalization category has no significant impact on the forward earnings multiple a particular manufacturer receives. Again, although a few outliers skew the mean price to earnings multiple amongst mid and large manufacturers, our regression analysis suggests that no relationship existed between size and forward multiples within our approved dataset.

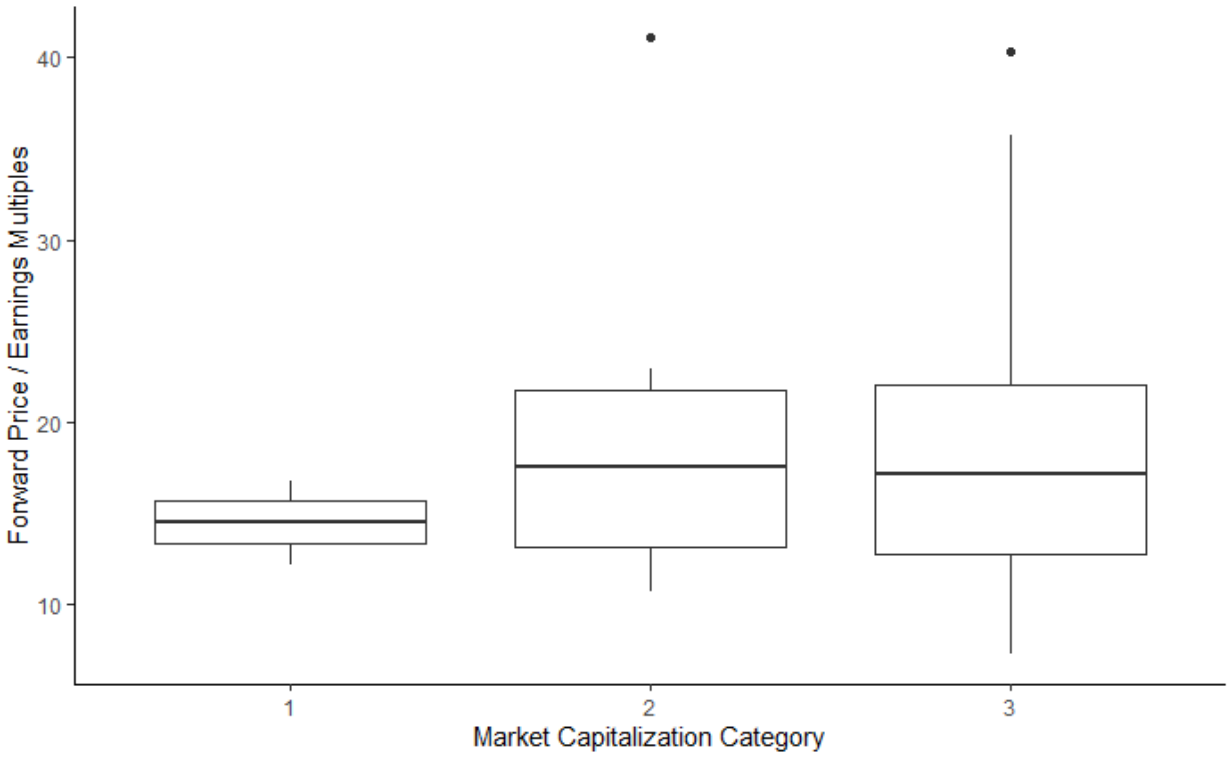

Figure 13. The effect of size factor on forward price / earnings multiples (*F*~(1,63)~ = 0.1243; *P* = 0. 7256). The regression line is explained by Y = 101183036X + 3998133. The Adjusted R-Squared is -0.0138.

Multivariable Linear Regressions

The previous discussion in this paper's results section showcased a handful of the simple single variable regressions conducted to both answer the primary hypothesis, but also gain a high-level understanding about the relationship between different variables within our approved manufacturers dataset. The next section will expand on this examination by combining several variables to determine which are most significant in explaining the historical variation in forward enterprise value to sales and price to earnings multiples within the approved manufacturer dataset.

Multivariable Linear Regression Results

	Dependent variable:
	EV_Sales
RD_efficiency	-0.0000
	(0.0000)
Mkt_Cap	0.00
	(0.00)
factor(Size_Factor)2	-6.46
	(4.04)
factor(Size_Factor)3	2.67
	(2.47)
RD_efficiency:Mkt_Cap	0.00
	(0.00)
RD_efficiency:factor(Size_Factor)2	0.0000
	(0.0000)
RD_efficiency:factor(Size_Factor)3	0.0000
	(0.0000)
Mkt_Cap:factor(Size_Factor)2	0.0000***
	(0.0000)
RD_efficiency:Mkt_Cap:factor(Size_Factor)2	-0.00**
	(0.00)
RD_efficiency:Mkt_Cap:factor(Size_Factor)3	
Constant	1.93
	(2.05)
Observations	65
R2	0.39
Adjusted R2	**0.30**
Residual Std. Error	1.85 (df = 55)

Figure 14. The effect of research & development expenditure efficiency, market capitalization, and size, on forward enterprise value / sales multiples (*F*~(9,55)~ = 3.983; *P* = 0.0005). The Adjusted R-Squared is 0.2955.

A model consisting of R&D efficiency, market capitalization, and firm size as independent variables, had a statistically significant and meaningful impact on forward enterprise value to sales multiples. A number of additional combinations were conducted, with varying degrees of significance, so with the aim of improving the model, we chose to sequentially remove insignificant variables.

Multivariable Linear Regression Results

	Dependent variable:
	EV_Sales
RD_efficiency	-0.0000
	(0.0000)
factor(Size_Factor)2	3.71
	(2.39)
factor(Size_Factor)3	3.57
	(2.31)
RD_efficiency:factor(Size_Factor)2	0.0000
	(0.0000)
RD_efficiency:factor(Size_Factor)3	0.0000
	(0.0000)
Constant	1.95
	(2.24)
Observations	65
R2	0.23
Adjusted R2	**0.16**
Residual Std. Error	2.02 (df = 59)

Figure 15. The effect of research & development expenditure efficiency, market capitalization, and size, on forward enterprise value / sales multiples (*F*~(1,59)~ = 4.25; *P* = 0.002286). The Adjusted R-Squared is 0.16.

Multivariable Linear Regression Results

	Dependent variable:
	EV_Sales
RD_efficiency	-0.0000*
	(0.0000)
Mkt_Cap	0.0000***
	(0.00)
RD_efficiency:Mkt_Cap	-0.00
	(0.00)
Constant	3.85***
	(0.61)
Observations	65
R2	0.19
Adjusted R2	**0.15**
Residual Std. Error	2.04 (df = 61)

Figure 16. The effect of research & development expenditure efficiency, market capitalization, and size, on forward enterprise value / sales multiples (*F*~(1,61)~ = 4.687; *P* = 0.005202). The Adjusted R-Squared is 0.1473.

To maintain brevity (and the reader's patience) , we've sped along the included outputs here so as to only highlight the model's strength before and after the final variable deletion for both forward enterprise value to sales and price to earnings. Unsurprisingly, for both forward

Multivariable Linear Regression Results

	Dependent variable:
	PE
RD_efficiency	-0.0000
	(0.0001)
EV_Sales	8.25**
	(3.34)
factor(Size_Factor)2	-0.94
	(83.71)
factor(Size_Factor)3	-11.62
	(9.15)
Mkt_Cap	0.0000*
	(0.0000)
RD_efficiency:EV_Sales	-0.0000
	(0.0000)
RD_efficiency:factor(Size_Factor)2	0.0001
	(0.0001)
RD_efficiency:factor(Size_Factor)3	0.0000
	(0.0001)
EV_Sales:factor(Size_Factor)2	-7.21
	(29.85)
RD_efficiency:Mkt_Cap	-0.00*
	(0.00)
EV_Sales:Mkt_Cap	-0.0000**
	(0.0000)
factor(Size_Factor)2:Mkt_Cap	0.0000
	(0.0000)
RD_efficiency:EV_Sales:factor(Size_Factor)2	-0.0000
	(0.0000)
RD_efficiency:EV_Sales:Mkt_Cap	0.00**
	(0.00)
RD_efficiency:factor(Size_Factor)2:Mkt_Cap	-0.00
	(0.00)
EV_Sales:factor(Size_Factor)2:Mkt_Cap	0.0000
	(0.0000)
RD_efficiency:EV_Sales:factor(Size_Factor)2:Mkt_Cap	0.00
	(0.00)
Observations	65
R2	0.74
Adjusted R2	**0.64**
Residual Std. Error	4.56 (df = 47)

Figure 17. The effect of research & development expenditure efficiency, market capitalization, factored by size on forward price / earnings multiples (*F*~(1,54)~ = 2.389; *P* = 0. 1.366e-08). The Adjusted R-Squared is 0.64.

Multivariable Linear Regression Results	
Dependent variable:	
	PE
RD_efficiency	-0.0001
	(0.0001)
factor(Size_Factor)2	6.16
	(8.58)
factor(Size_Factor)3	5.31
	(8.29)
RD_efficiency:factor(Size_Factor)2	0.0001
	(0.0001)
RD_efficiency:factor(Size_Factor)3	0.0001
	(0.0001)
Constant	17.28**
	(8.05)
Observations	65
R2	0.16
Adjusted R2	**0.09**
Residual Std. Error	7.26 (df = 59)

Figure 18. The effect of research & development expenditure efficiency and market capitalization size factor on forward price / earnings multiples (*F*~(1,59)~ = 2.245; *P* = 0.06149). The Adjusted R-Squared is 0.08866.

Multivariable Linear Regression Results	
Dependent variable:	
	PE
RD_efficiency	-0.0000**
	(0.0000)
Mkt_Cap	0.0000
	(0.0000)
RD_efficiency:Mkt_Cap	0.00
	(0.00)
Constant	20.42***
	(2.12)
Observations	65
R2	0.18
Adjusted R2	**0.14**
Residual Std. Error	7.05 (df = 61)

Figure 19. The effect of research & development expenditure efficiency and market capitalization on forward price / earnings multiples (*F*~(1,61)~ = 4.47; *P* = 0.006672). The Adjusted R-Squared is 0.1399.

EV/Sales and P/E, the lowest p-value and highest adjusted r-squared values arose when utilizing all the variables in our dataset. This of course is simply an example of over-fitting the model and communicates little to readers and researcher alike – for this reason, we continued removing variables and interactions that were insignificant to arrive at a more useful answer. Figures sixteen and nineteen highlight that the two variables with the strongest relationship with forward EV/Sales and P/E is research and development expenditure efficiency and market capitalization.

K-Means Cluster Analysis

Approved Drug Dataset

With an understanding that several variables within our approved drug dataset have a statistically meaningful impact on forward EV/Sales and P/E multiples, the next portion of this discussion focuses on utilizing K-means cluster analysis to algorithmically segment our approved, unapproved, and consolidated groups to glean a better understanding on our data's constitution.

In the single-variable regression discussion we found that R&D expense did not have a statistically meaningful impact on forward EV/Sales - however, this result only suggests that these two variables don't share a linear relationship, not that certain pockets of the dataset don't.

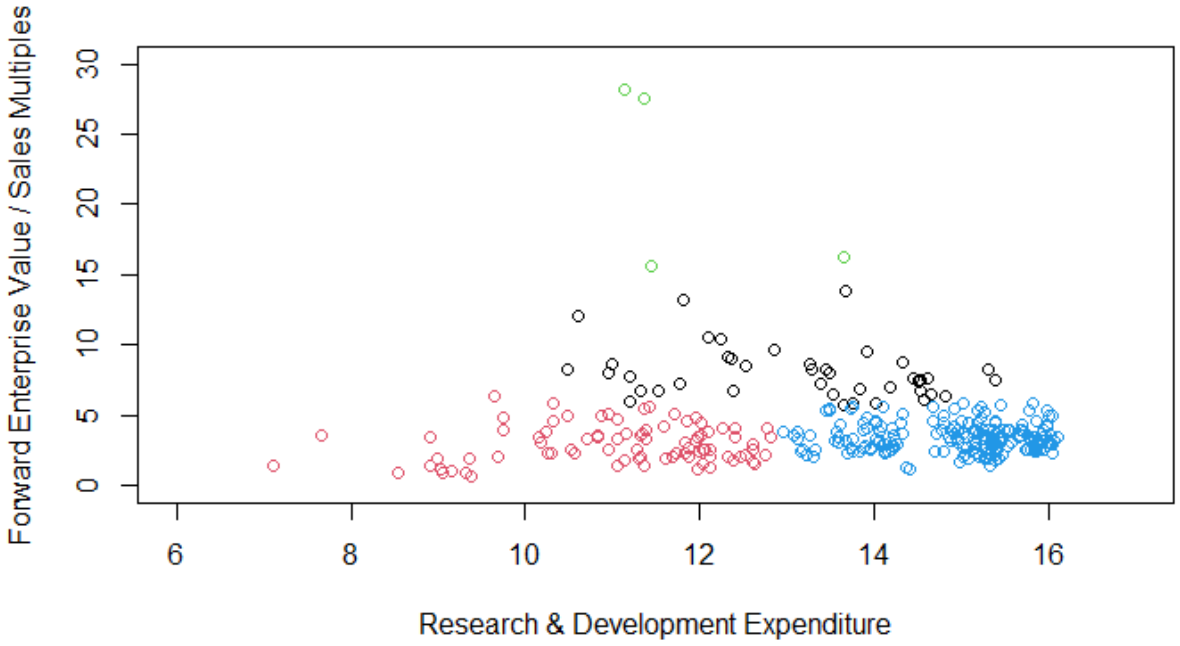

Figure 20. Research and development expense and forward enterprise value to sales multiples

For this reason, deploying a clustering technique is helpful. Utilizing the cluster-algorithms discussed in the methodology section we calculated that the relationship between logged research and development and forward enterprise value to sales multiples for approved manufacturers could be categorized into four separate clusters, as seen in figure twenty. The red cluster in the bottom-left represents firms that spend less on R&D and correspondingly are valued at multiples between 1 – 5x revenue. In contrast, the blue cluster represents the highest nominal R&D spenders who nonetheless trade for similar multiples as their lowest R&D peers. Lastly, the black and green clusters allocate in a similar fashion to both low and high R&D firms, yet trade at

premium multiples in comparison.

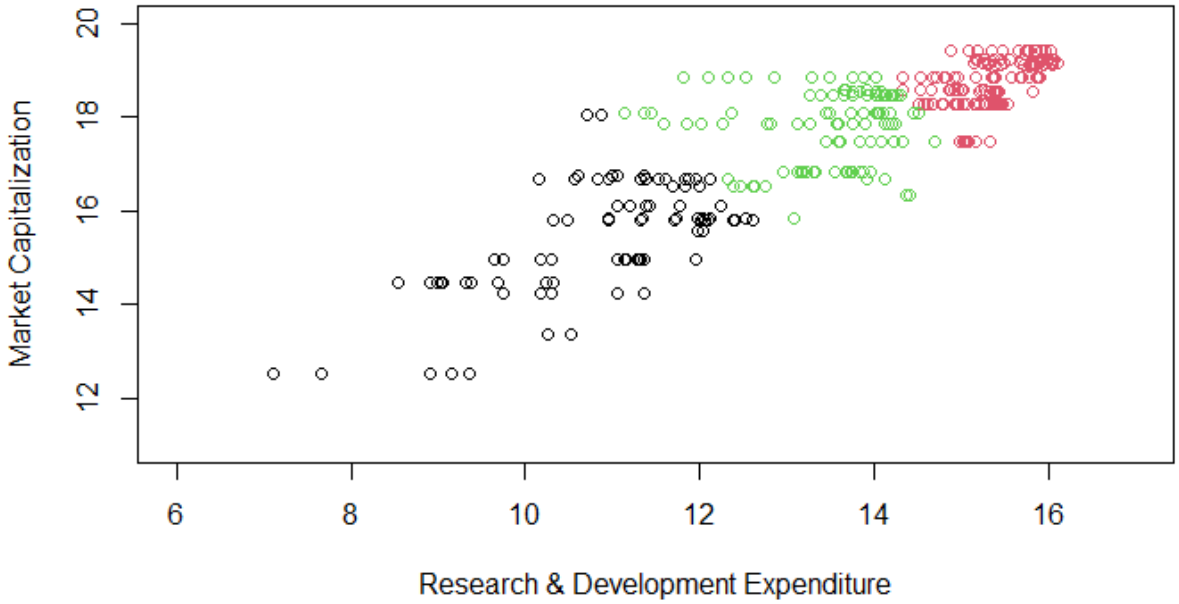

Figure 21. Research and development expense and market capitalization

For variables that are known to possess a linear relationship, clustering the data is still useful in further confirming the characteristics of each segment align with the initial conclusion. For example, in figure twenty-one, each cluster (black, green, and red) exhibits the expected relationship with minimal noticeable outliers. In terms of market capitalization, the smallest firms commit the least towards R&D (black), while the middle (green) and large-scale (red) manufacturers dedicate the most.

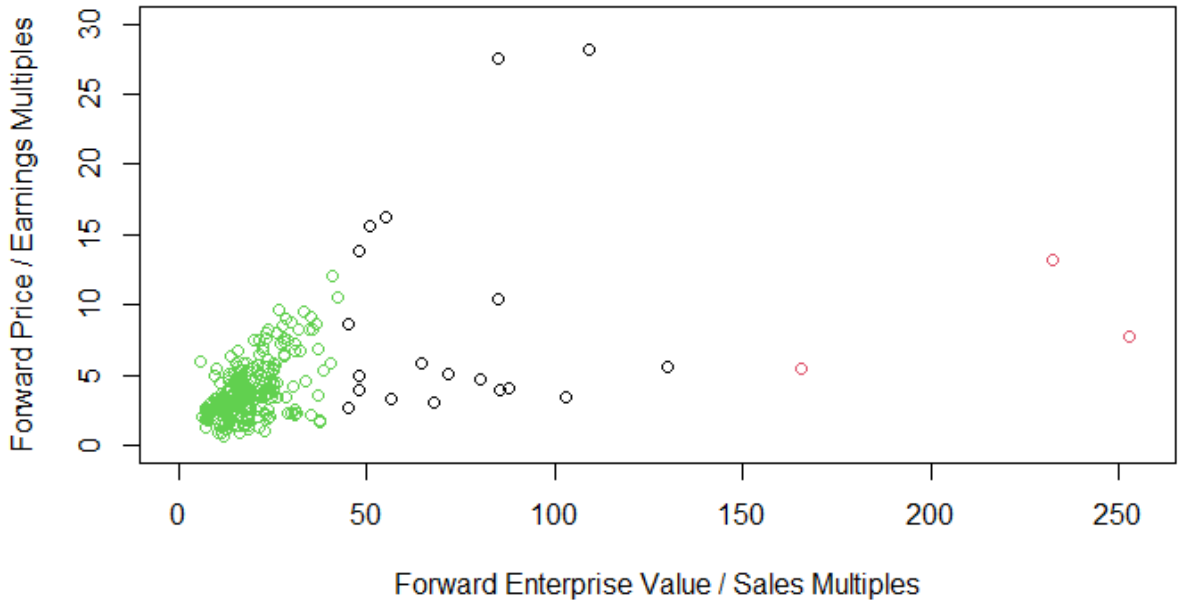

Figure 22. Forward enterprise value / sales and price / earnings multiples

Another relatively linear relationship reinforced with the usage of clustering was forward enterprise value to sales and price to earnings. As expected, most observations rest in the bottom left (green), however, the two outlier categories (black and red) identify datapoints where a manufacturer's EV/Sales multiple expands, yet P/E remains similar to peers. Although further diligence was not conducted into the outlier scenarios in groups black and red, this paper suspects those datapoints represent situations similar to the hypothetical ones outlined in the single-variable discussion.

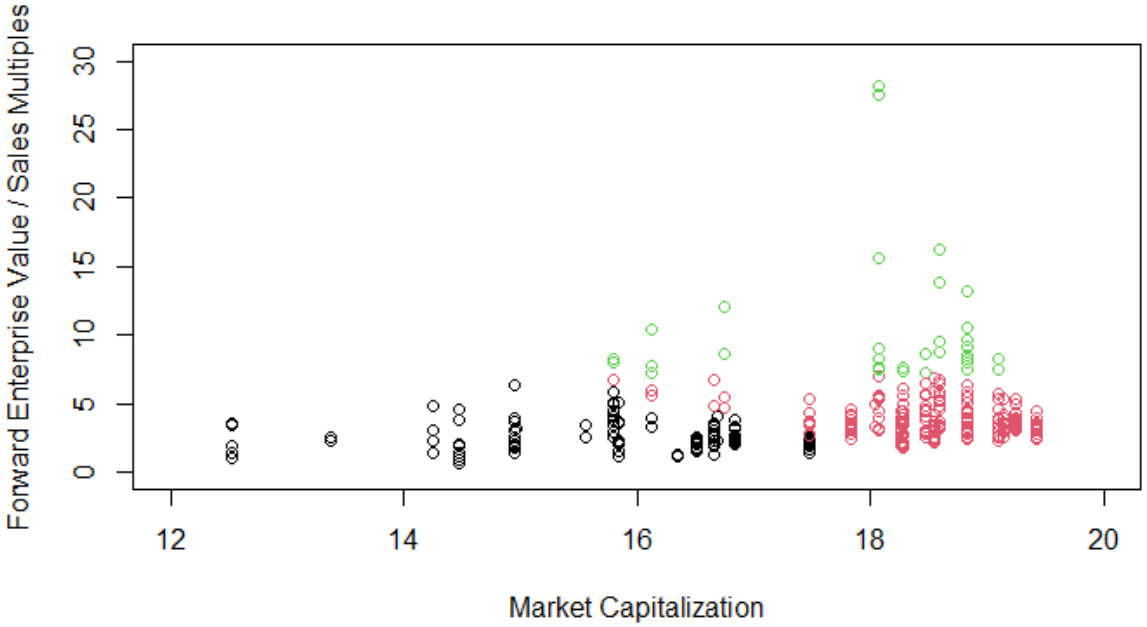

Figure 23. Logged market capitalization and forward enterprise value / sales multiples

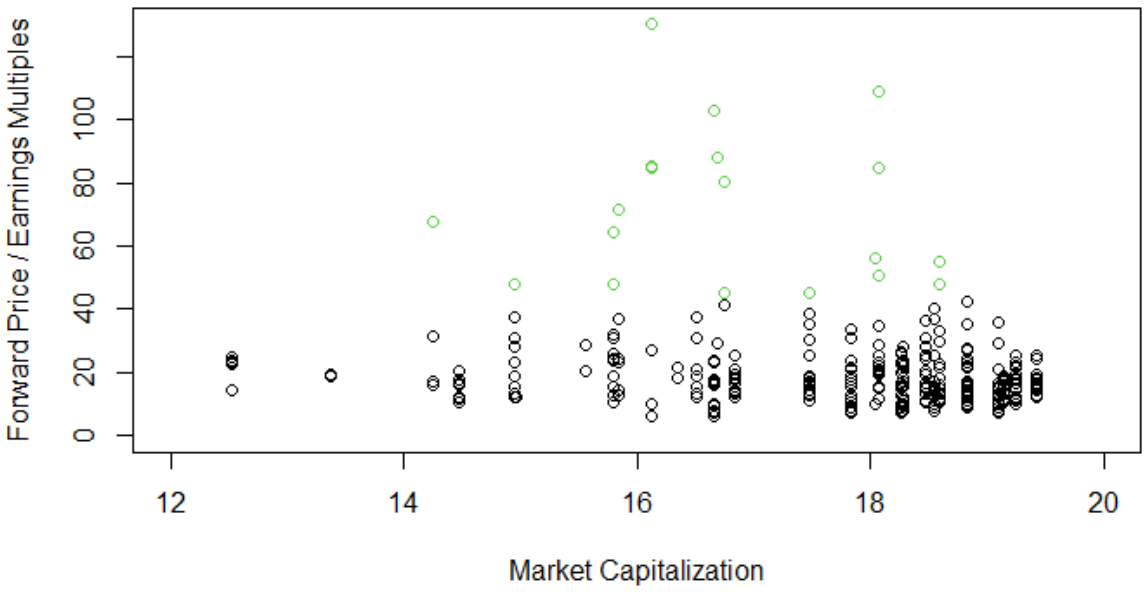

Figure 24. Logged market capitalization and forward price / earnings multiples

Figures twenty-three and twenty-four, similar to figure twenty, are effective at contextualizing previously non-linear groups of data. Although our regression identified market capitalization as having an insignificant impact on forward enterprise value to sales multiples, clustering the data further into three categories yields useful insights. Immediately apparent is the fact that both small (black) and large (red) firms similarly trade at between 1 – 5x revenue. Less apparent at first glance, however, is that the outliers (green), which are characteristically large manufacturers, trade at premium multiples to their adjacent peers. A similar relationship is observed in figure twenty-four – the outliers (green) are both mid and large-scale manufacturers that trade at premium earnings multiples in comparison to peers. Additional interesting analysis would be to determine why these operators trade differently from peers – perhaps some consensus view on operational excellence or R&D efficiency?

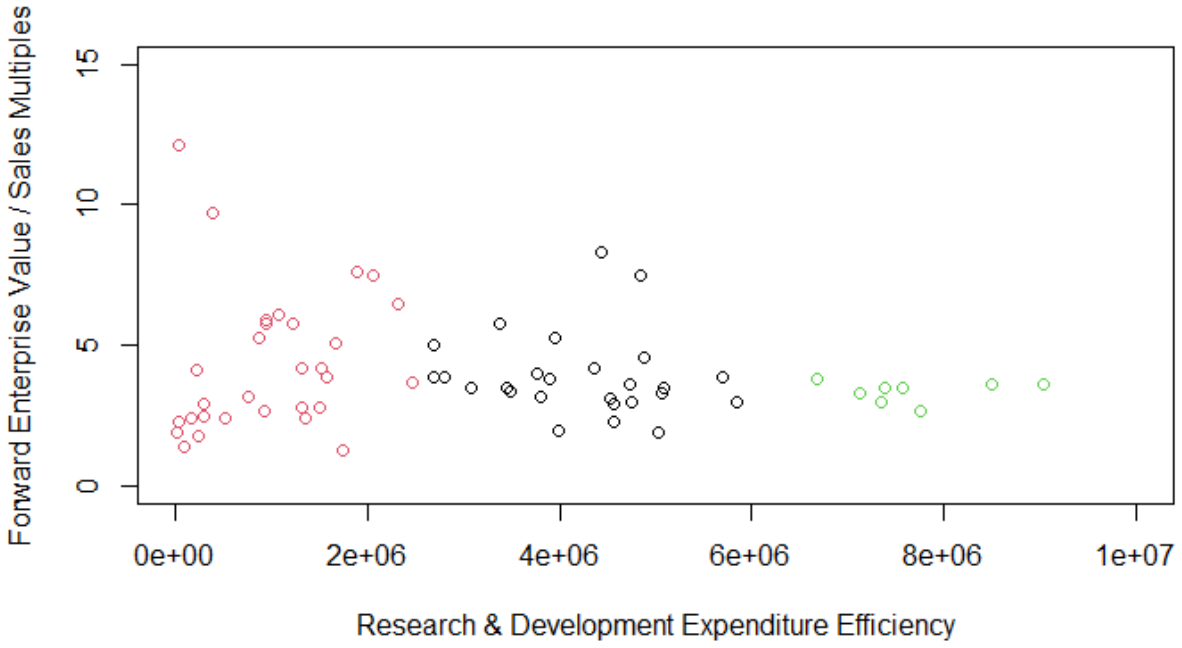

Figure 25. Research and development efficiency and forward enterprise value / sales multiples

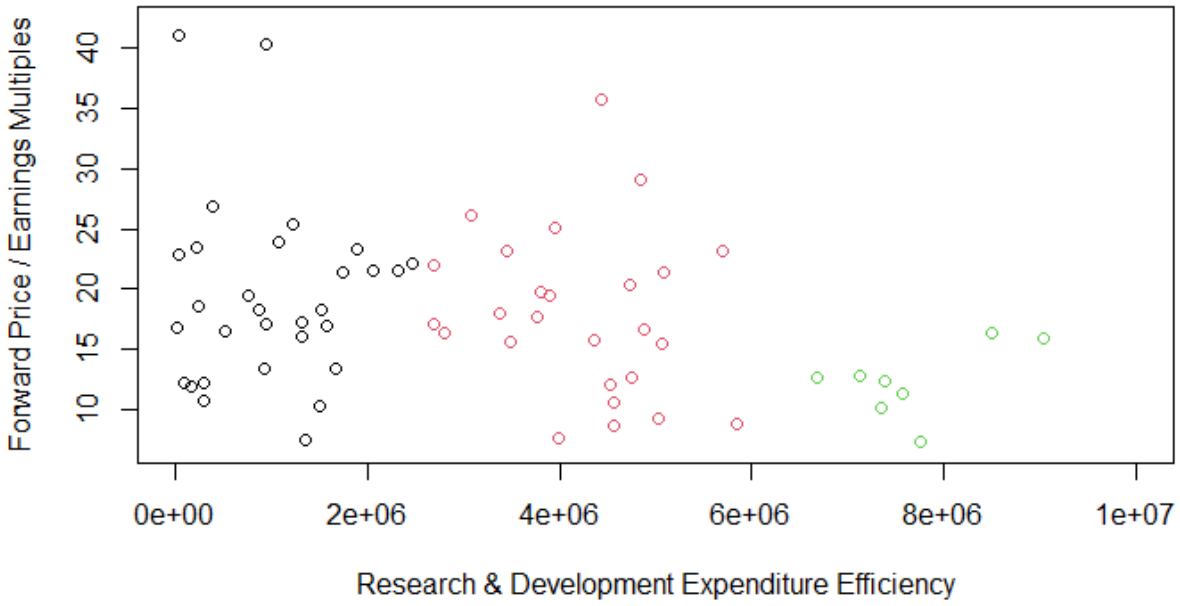

Figure 26. Research and development efficiency and forward price / earnings multiples

This brings us back to this paper's initial question – does research and development expenditure efficiency have a statistically significant impact on forward enterprise value to sales or prices to earnings multiples? A single-variable regression determined that in aggregate, R&D efficiency does not have an impact on forward EV/Sales multiples. Nonetheless, in figures twenty-five and six, we identified three clusters which make additional analysis fruitful. The most efficient R&D allocators (red) in our dataset trade for between one to six times revenue, whereas those in the middle (black) and end of the pack (green) range between three to five and four to five times revenue. In comparison, the least efficient R&D allocators in figure twenty-six (green), trade for the lowest multiples (10 – 12 times earnings), meanwhile, the middle and top quintile allocators, on average, trade at a premium.

Unapproved Drug Dataset

In the next section, this paper continues utilizing k-means clustering to identify interesting relationships between our selected variables. However, the focus now rests primarily on the 379 firms (from our initial dataset including 425 manufacturers) that did not receive an FDA approval during the fifteen-year study period.

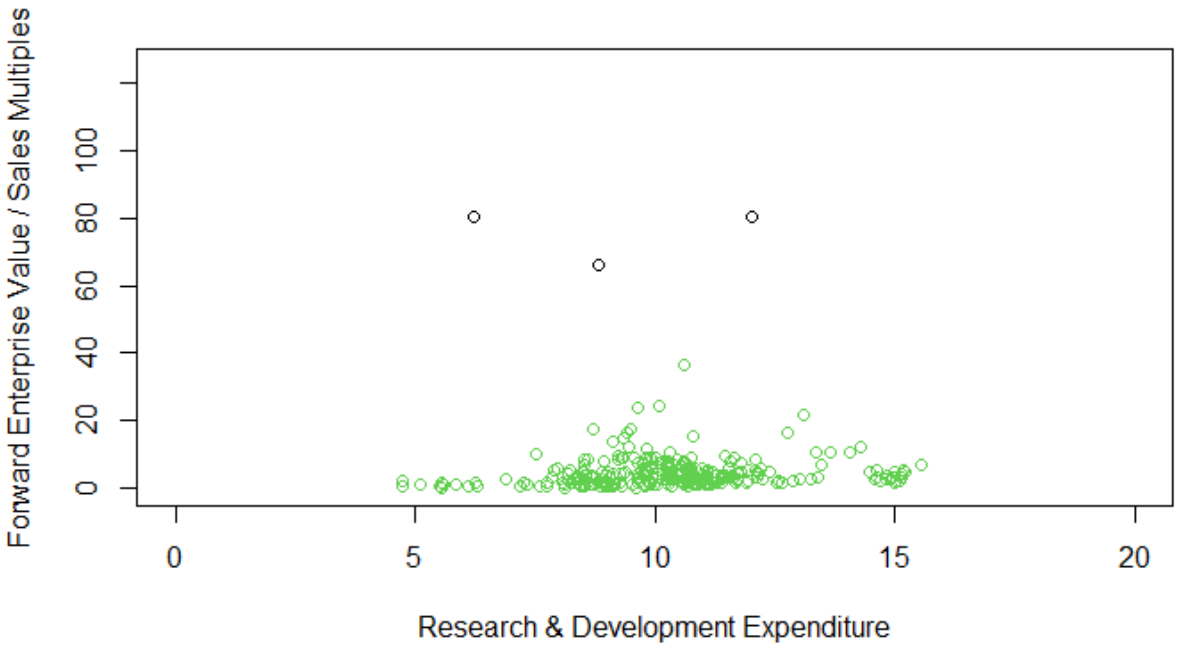

Figure 27. Logged research and development and forward enterprise value / sales multiples

The most notable difference while utilizing clustering techniques on the unapproved dataset is the lack of inter-cluster dissimilarity. In other words, although mathematically any dataset can be segmented into any number of sections, the observations in our unapproved dataset are characteristically quite similar. This point will be made even more apparent to the reader in the histogram analysis – however, figure twenty-seven, is also a good example. Here, we note that the distribution of R&D and forward EV/Sales multiples appear to coalesce near the

center. The same is true for R&D and forward P/E multiples – both also possess a few outliers, identified as a separate cluster that trade at premium multiples.

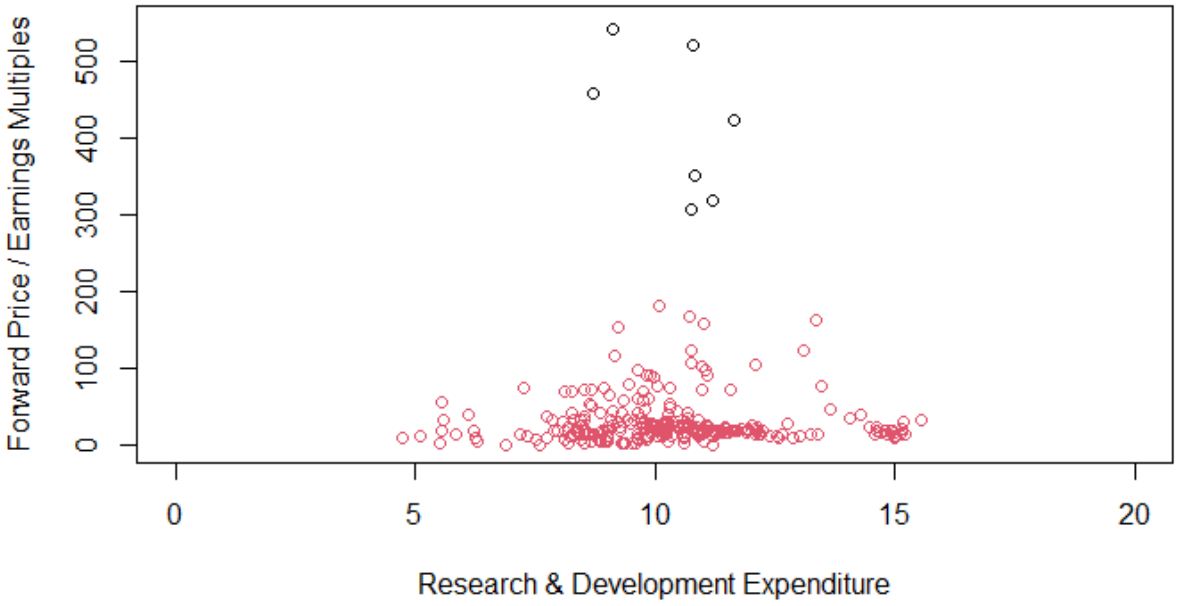

Figure 28. Logged research and development expenditure and forward price / earnings multiples

Despite the distribution of R&D expenditure differing amongst unapproved manufacturers, these firms still retain a few similar characteristics to their successful peers. Namely, larger firms, more so than smaller ones, dedicate nominally more dollars towards internal and external research and development efforts, as is showcased in figure twenty-nine.

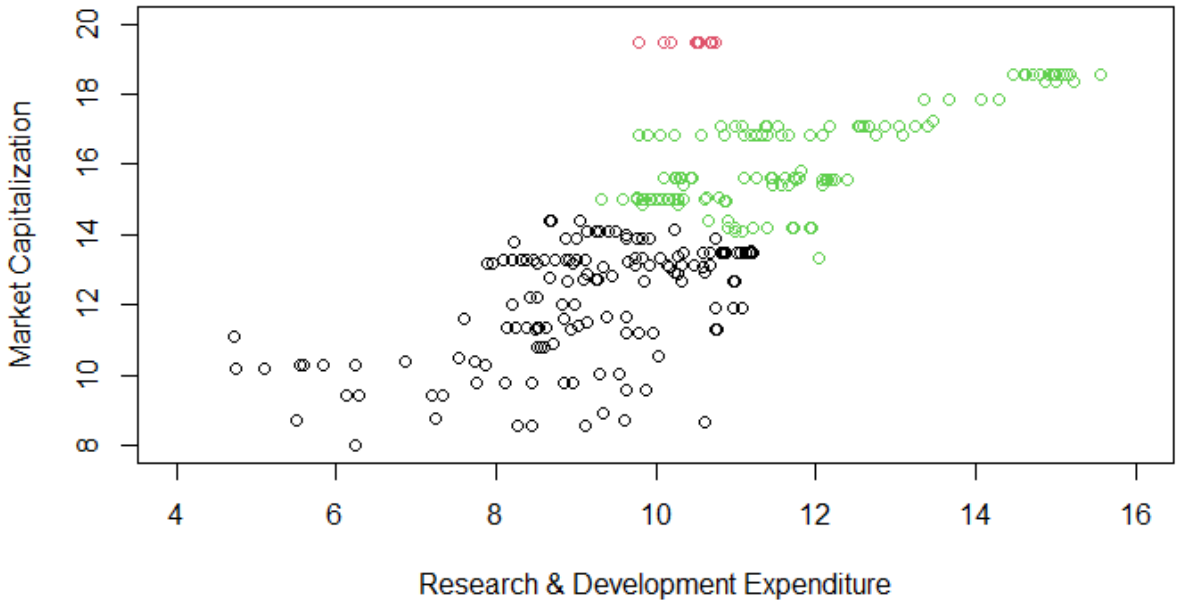

Figure 29. Logged research and development expenditure and market capitalization

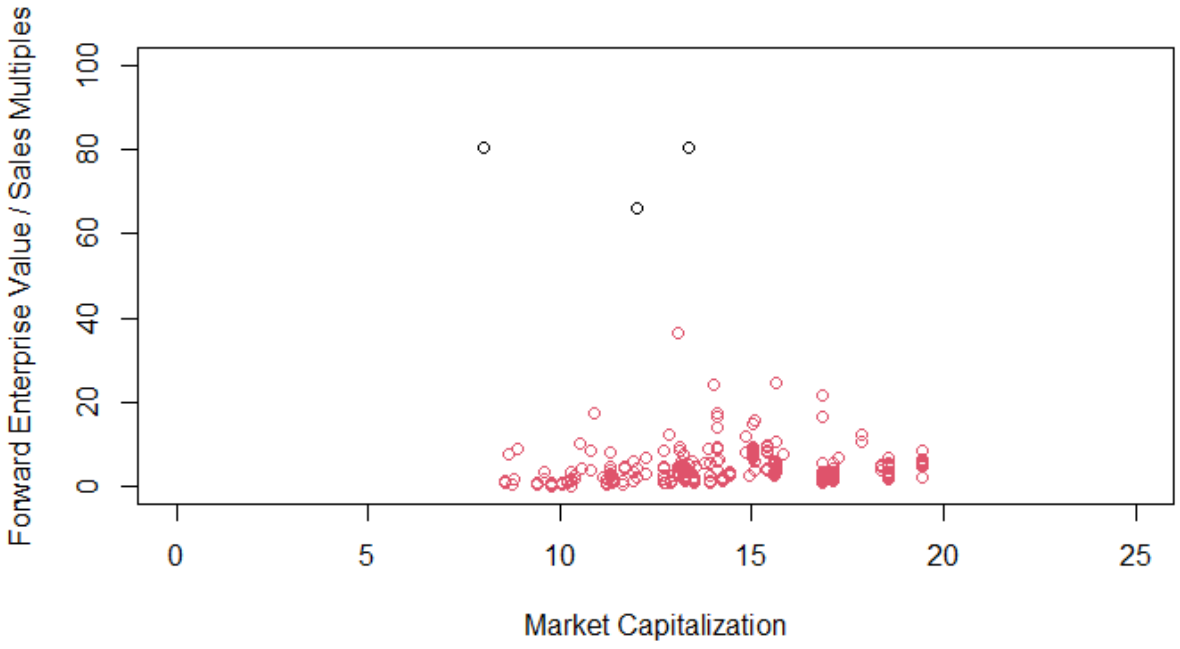

Figure 30. Logged market capitalization and forward enterprise value to sales multiples

However, unlike the approved manufacturers, a pocket of outliers populates the top-center region (black). These firms appear to spend similarly to middle, and some low R&D peers yet boast a larger market capitalization. Considering all of these firms represent manufacturers that did not gain an FDA approval during this paper's fifteen-year study period (a select few received approvals afterwards), it's interesting to note that public-market investors were more enthusiastic about these manufacturers' research and development efforts.

Similar to R&D, the distribution of market capitalizations for unapproved manufacturers appears to coalesce near the center of figure thirty. Again, due to low dissimilarity amongst datapoints, the k-means algorithm categorizes most of our observations in one cluster, with another cluster identifying the extreme outliers with similar market capitalizations, but vastly different forward enterprise value to sales multiples.

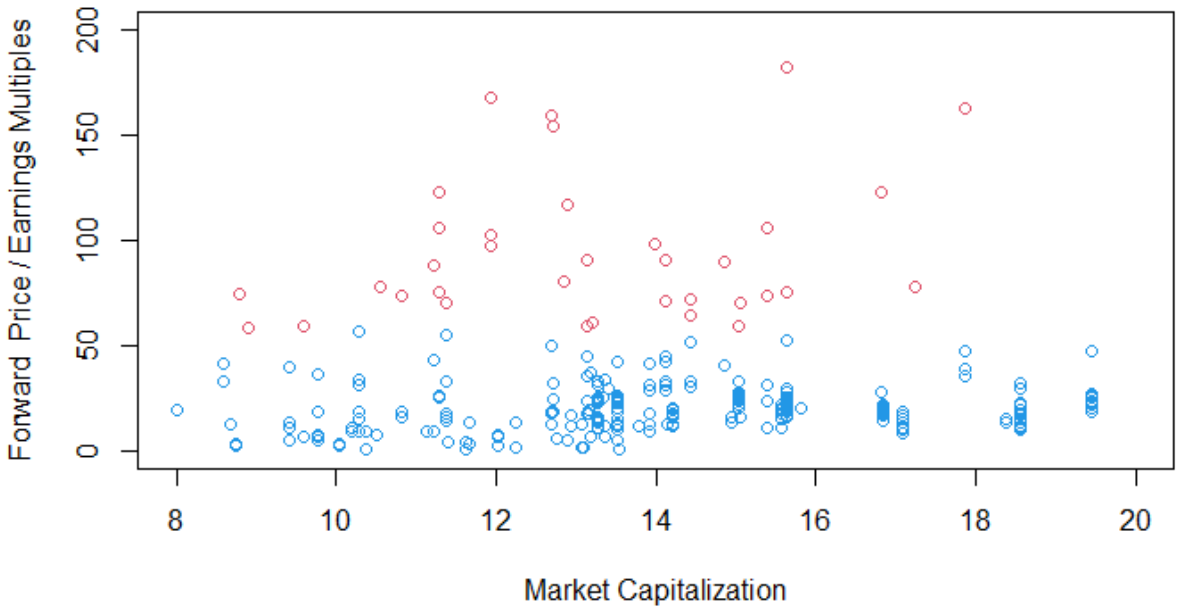

Figure 31. Logged market capitalization and forward price / earnings multiples

Figure 32. Forward enterprise value / sales and price / earnings multiples

Unlike the relationship between market capitalization and forward enterprise value to sales, the clustering algorithm handles forward price to earnings quite well. For example, in figure thirty-one it's apparent that two clusters actually exist – notably, small to large-scale manufacturers trading anywhere between one and fifty-times earnings and those trading for a premium. Although it's difficult to rationalize placing operators trading at ten and fifty-times earnings in the same category, a potential counterargument could be that this group primarily consists of clinical-stage biopharmaceutical firms, which, on average trade at premium multiples relative to most other sectors.

Complete Dataset – Approved and Unapproved Drugs

Lastly, the next section in the K-means cluster analysis portion of the results discussion involves segmenting, analyzing, and comparing the entire dataset against both the unapproved

Figure 33. Research and development expenditure and forward enterprise value / sales multiples

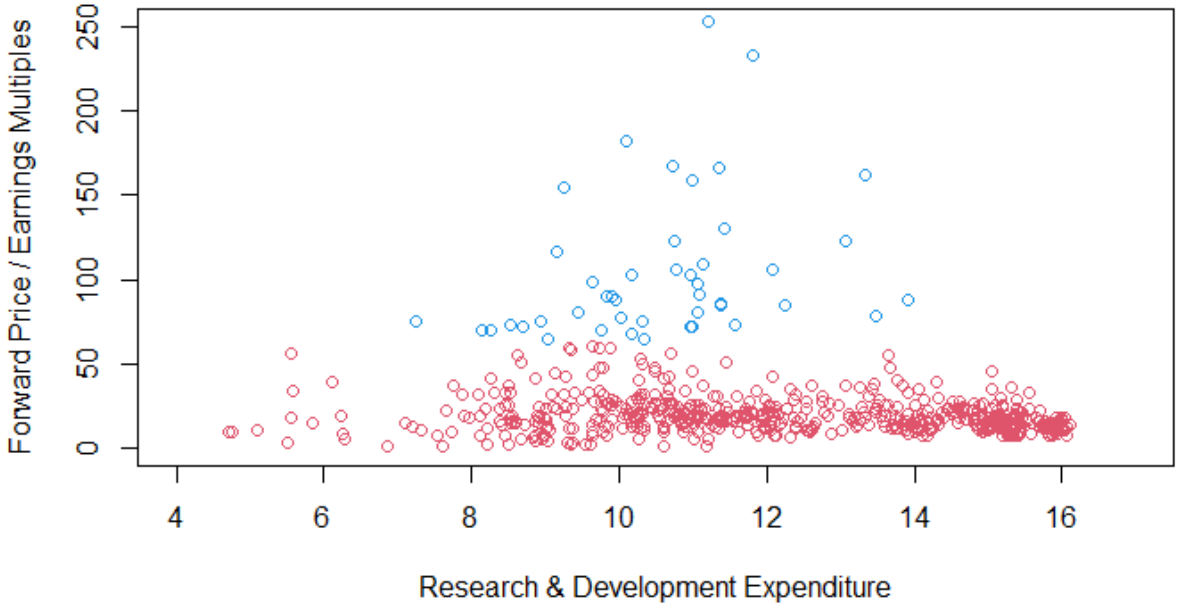

Figure 34. Research and development expenditure and forward price to earnings multiples

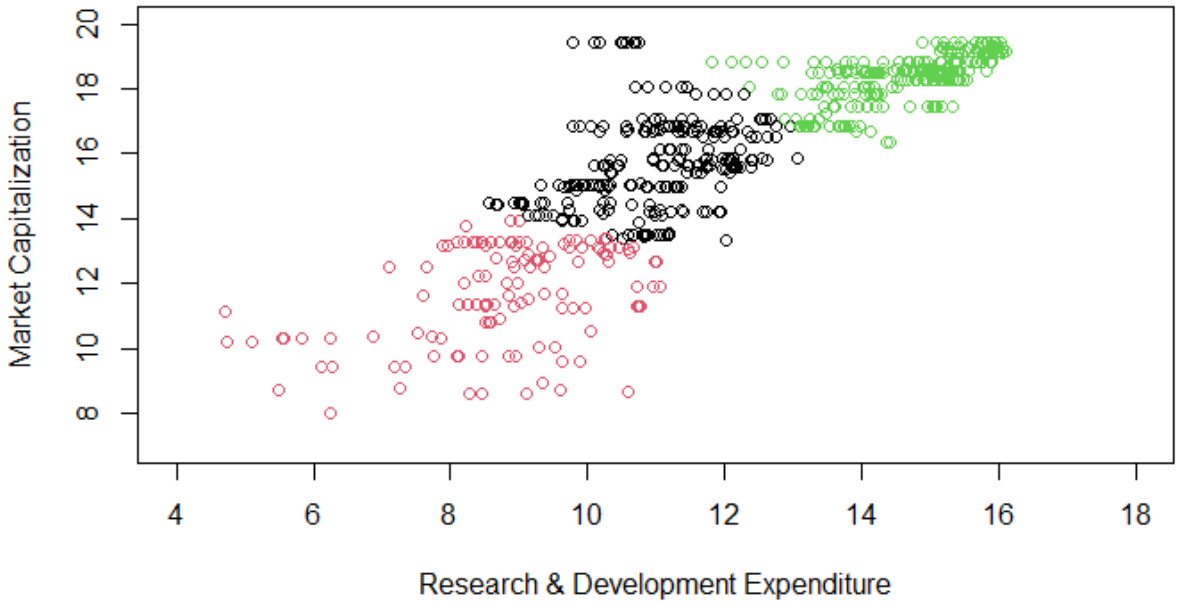

Figure 35. Research and development expenditure and market capitalization

and approved datasets. Within the consolidated drug data frame, the impact of research and development expenditure on forward enterprise value to sales multiples is insignificant. Additionally, clustering similar data points is of little assistance here since there is limited inter-cluster dissimilarity, which disrupts the algorithm's ability to compute distinct centroids – this is apparent in figure thirty-three. However, in figure thirty-four, forward price to earnings has marginally superior cluster separation – firms highlighted in blue represent outliers trading for premium multiples whilst expending similar amounts of resources on R&D as peers within the red cluster.

As expected, research and development expenditure across each market capitalization category remained clearly defined and consistent with the initial conclusion presented in the approved drug dataset. Similarly, forward enterprise value to sales, price to earnings

Figure 36. Forward enterprise value / sales and price / earnings multiples

Figure 37. Market capitalization and forward enterprise value / sales multiples

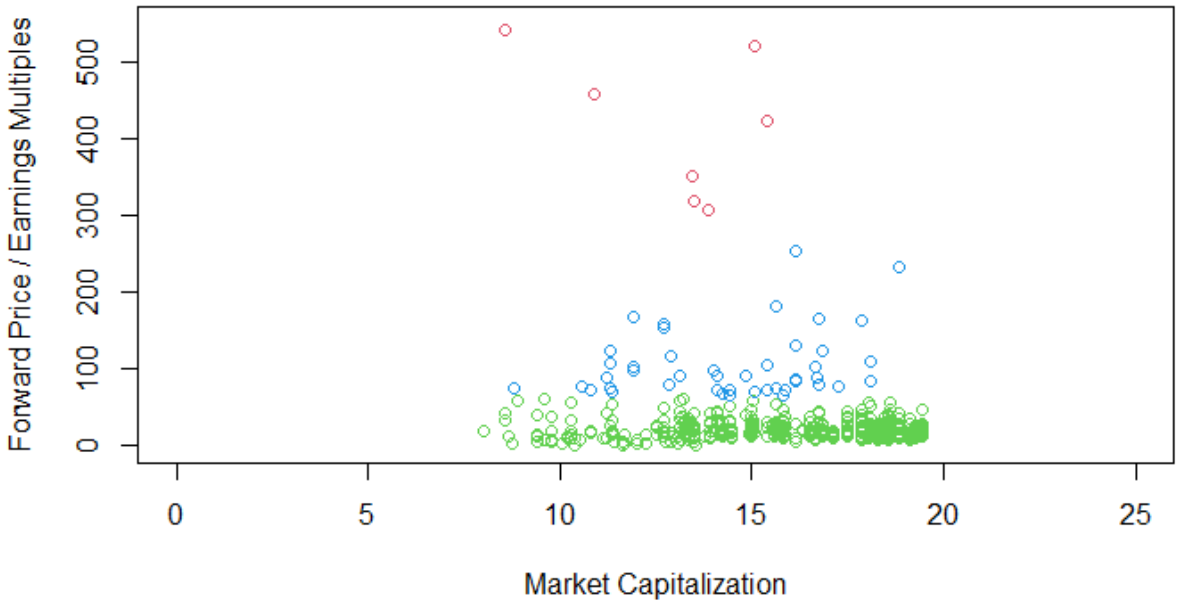

Figure 38. Market capitalization and forward price / earnings multiples

multiples, and market capitalization each demonstrated distinctive clusters in figures thirty-six and eight. However, upon comparison with enterprise value to sales in figure thirty-seven, market capitalization's clusters are less poignant and likely reflect a dataset not ideal for generating centroids.

Distribution Analysis - Histograms

Lastly, before addressing our second hypothesis, we will close by briefly comparing the distributions of each variable utilized to answer our primary hypothesis. For the most part, the approved manufacturers appear to visually differ from the overall dataset, which makes sense considering this segment consists of the most successful publicly traded pharmaceutical firms within the fifteen-year study period. Most notably, the distribution of research and development

Approved Drug Manufacturers – Variable Distributions

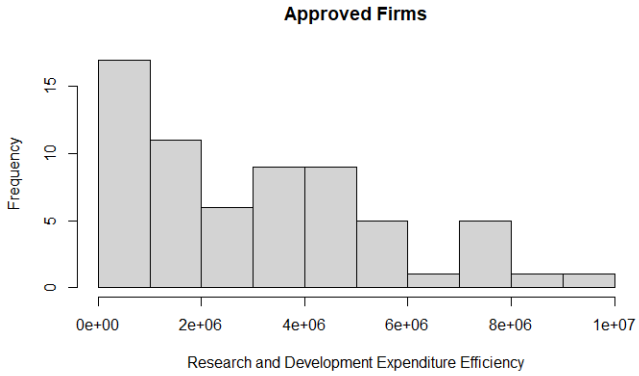

Figure 39. R&D efficiency distribution

Figure 42. Forward P/E distribution

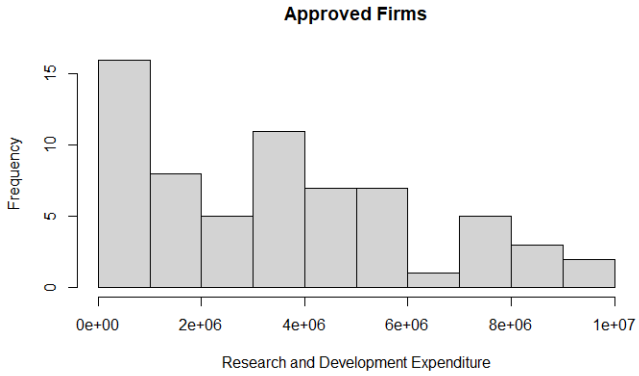

Figure 40. R&D distribution

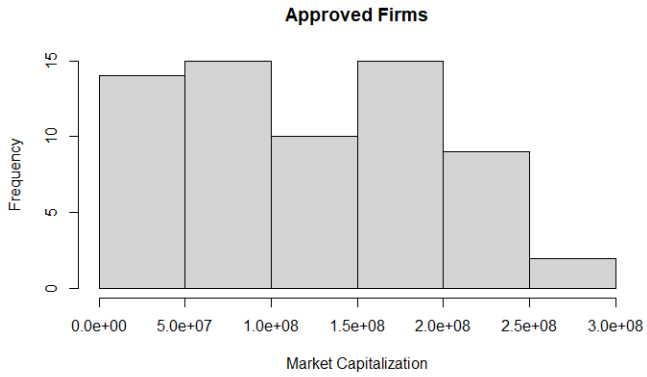

Figure 43. Firm size distribution

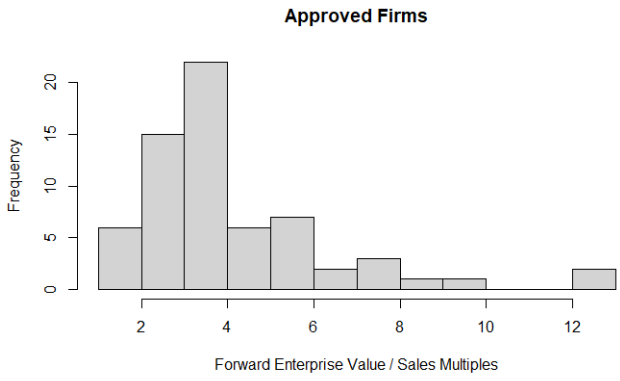

Figure 41. Forward EV/ Sales distribution

Unapproved and Consolidated Drug Manufacturers – Variable Distributions

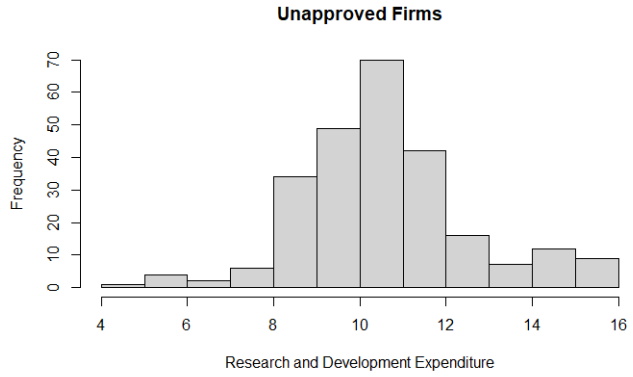

Figure 44. R&D distribution

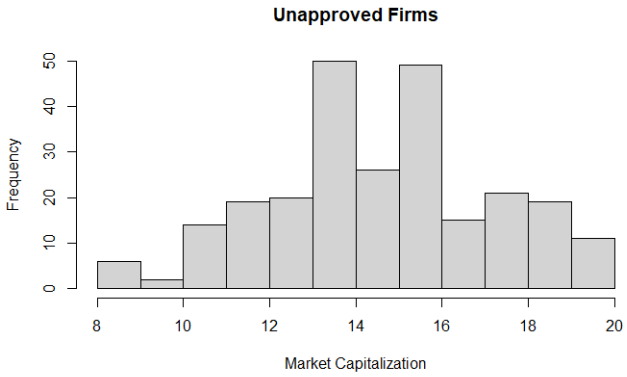

Figure 47. Firm size distribution

Figure 45. Forward EV/Sales distribution

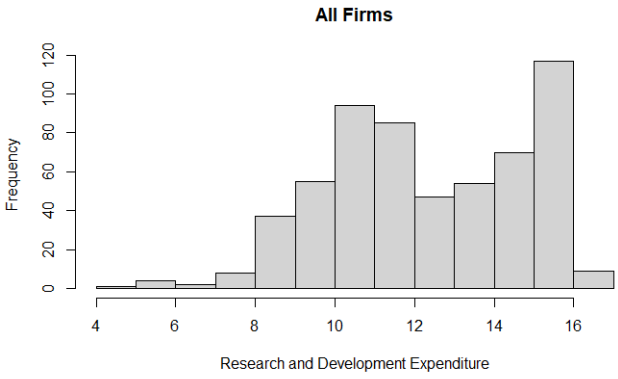

Figure 48. R&D distribution

Figure 46. Forward Price/Earnings distribution

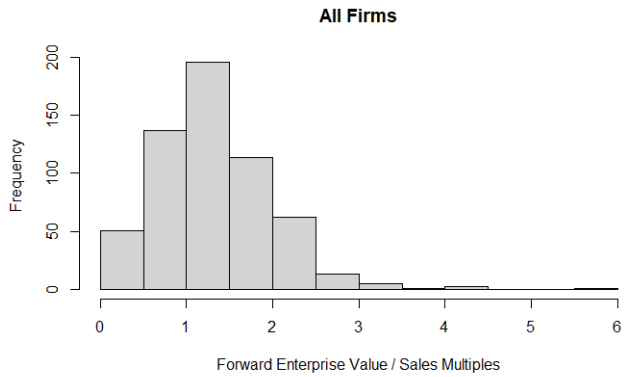

Figure 49. Forward EV/Sales distribution

Figure 50. Forward Price/Earnings distribution

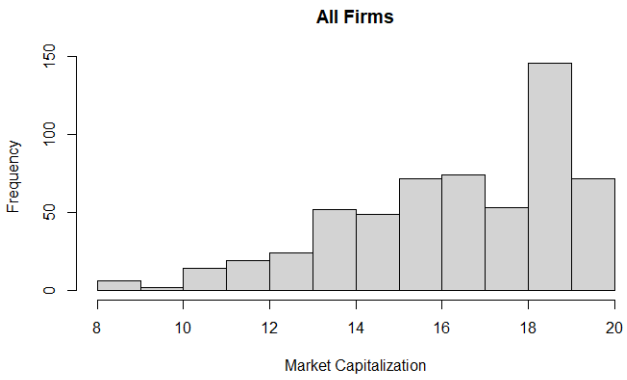

Figure 51. Firm size distribution

expenditure for approved manufacturers (figure forty) skews heavily to the far right, which contrasts the normal distribution exhibited by the overall (figure forty-eight) and unapproved (figure forty-four) firms. In a similar vein, firm-size skews to the larger-end amongst the approved manufacturers (figure forty-three), which makes sense when considering R&D efficiency (figure thirty-nine) also skews nominally to the higher end as well (recall, this implies lower efficiency). EV/Sales and P/E multiples within the approved dataset (figures forty-one and two) present as normal distributions, which surprisingly aligns with manufacturers in the overall (figure forty-nine and fifty) and unapproved (figures forty-five and six) groups on P/E while differing on EV/Sales.

Secondary Hypothesis

This paper also sought to conclude if "experience," as defined by the number of FDA approvals a particular manufacturer receives within a pre-defined terminal, has a statistically significant impact on forward earnings and or sales multiples. Functionally, this was a relatively simple exercise to conduct once in possession of all the relevant data – we identified "experienced" firms as manufacturers that received an FDA approval between 2000 and 2005 and performed simple single variable regressions for both forward multiples. The boxplots in

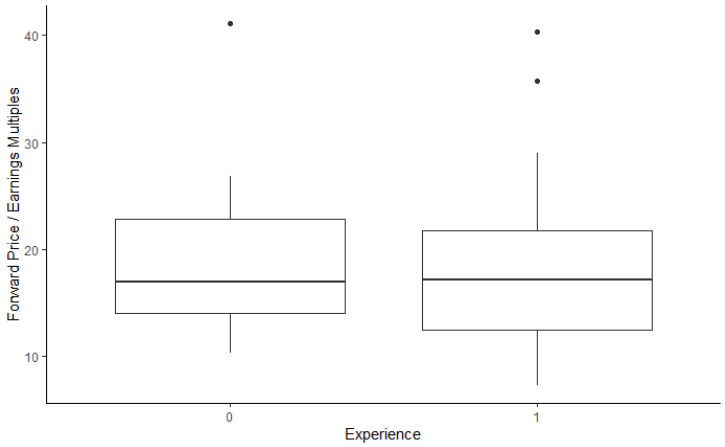

Figure 52. The effect of experience for approved manufacturers on forward price / earnings multiple (*F*~(1,63)~ = 1.646; *P* = 0.2043). The Adjusted R-Squared is 0.009.

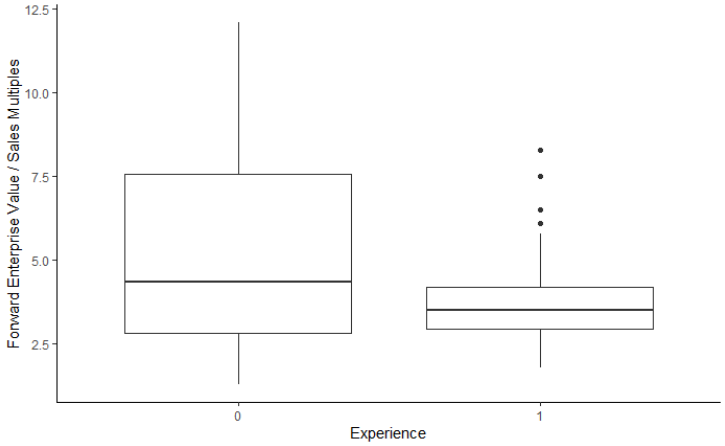

Figure 53. The effect of experience for approved manufacturers on forward Enterprise Value / Sales multiple (*F*~(1,63)~ = 7.38; *P* = 0.0085). The Adjusted R-Squared is 0.090.

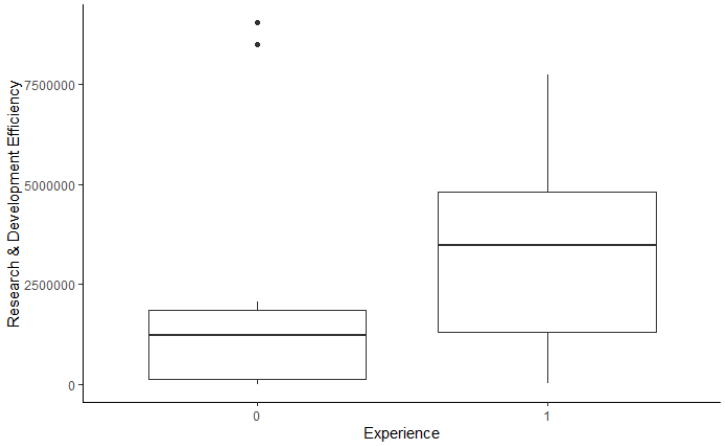

Figure 54. The effect of experience for approved manufacturers on research and development expenditure efficiency (*F*~(1,63)~ = 3.576; *P* = 0.06324). The Adjusted R-Squared is 0.0386.

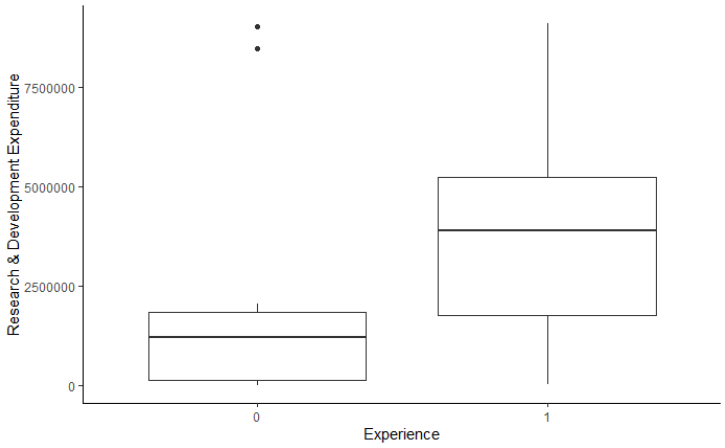

Figure 55. The effect of experience for approved manufacturers on research and development expenditure (*F*~(1,63)~ = 5.759; *P* = 0.01938). The Adjusted R-Squared is 0.0692.

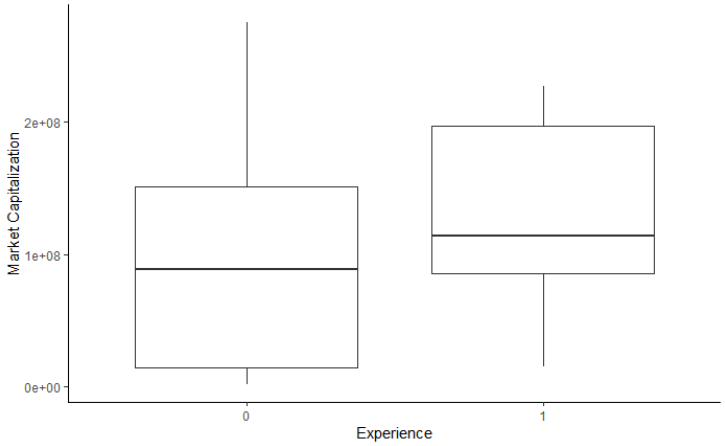

Figure 56. The effect of experience for approved manufacturers on market capitalization (*F*~(1,63)~ = 1.068; *P* = 0.3053). The Adjusted R-Squared is 0.001.

figures fifty-two and three visually summarize our results and the remaining figures showcase how experience interacts with other variables. In short, experience does not seem to have a statistically significant impact on either forward earnings or sales multiples. However, interestingly, experience does appear to have a minor impact on the quantity and efficiency of dollars deployed into internal research and development projects.

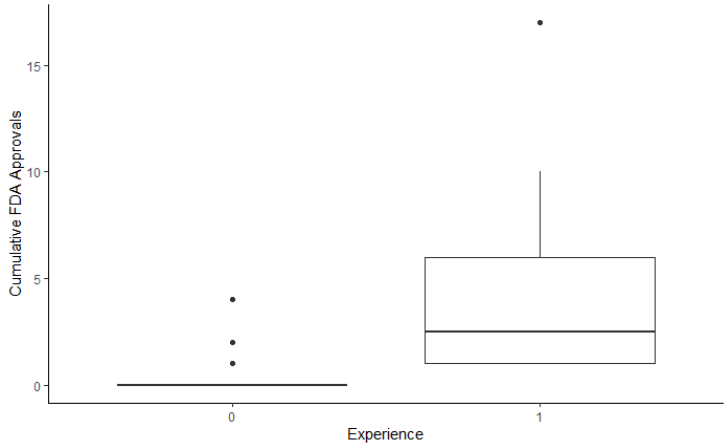

Figure 65. The effect of experience on cumulative FDA approvals (*F*~(1,423)~ = 366.9 ; *P* = 2.2e-16). The Adjusted R-Squared is 0.4632

Conclusion

In summary, this paper sought to determine if the efficiency of pharmaceutical research and development expenditure has a statistically significant impact on forward earnings and sales multiples. Additionally, we sought to better understand the relationship of the various variables within our dataset and determine if experience also has a meaningful affect. In conclusion, this paper found that research and development expenditure efficiency has a miniscule statistical impact on forward price to earnings multiples and none at all on enterprise value to sales. Similarly, the relative experience of a particular firm does not have a statistical impact on the respective forward earnings or sales multiple said manufacturer receives.

Appendix

Additional Boxplot Analysis

Unapproved Drugs Dataset

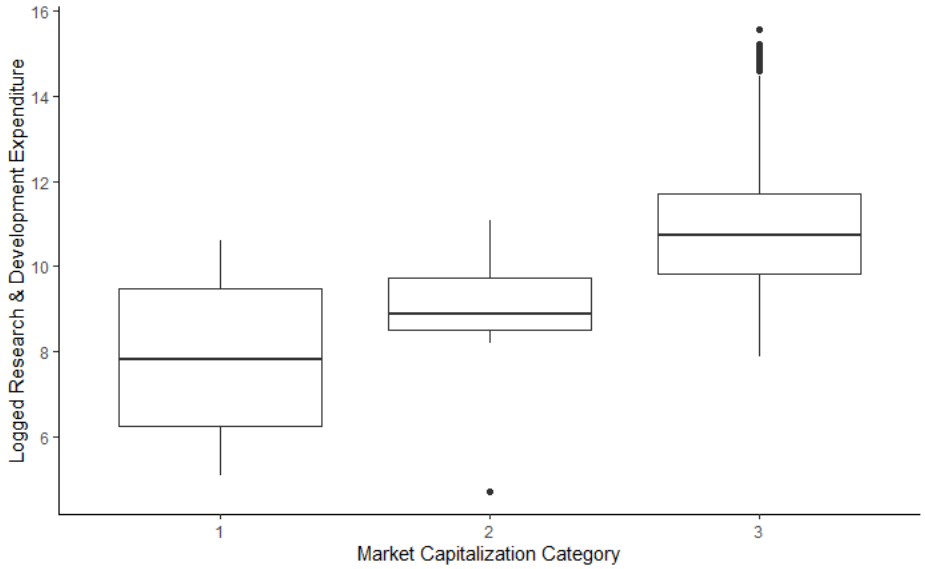

Figure 57. The effect of market capitalization category (size factor) for unapproved manufacturers on research and development expenditure (*F*~(1,249)~ = 41.07; *P* = 3.84e-16). The Adjusted R-Squared is 0.242.

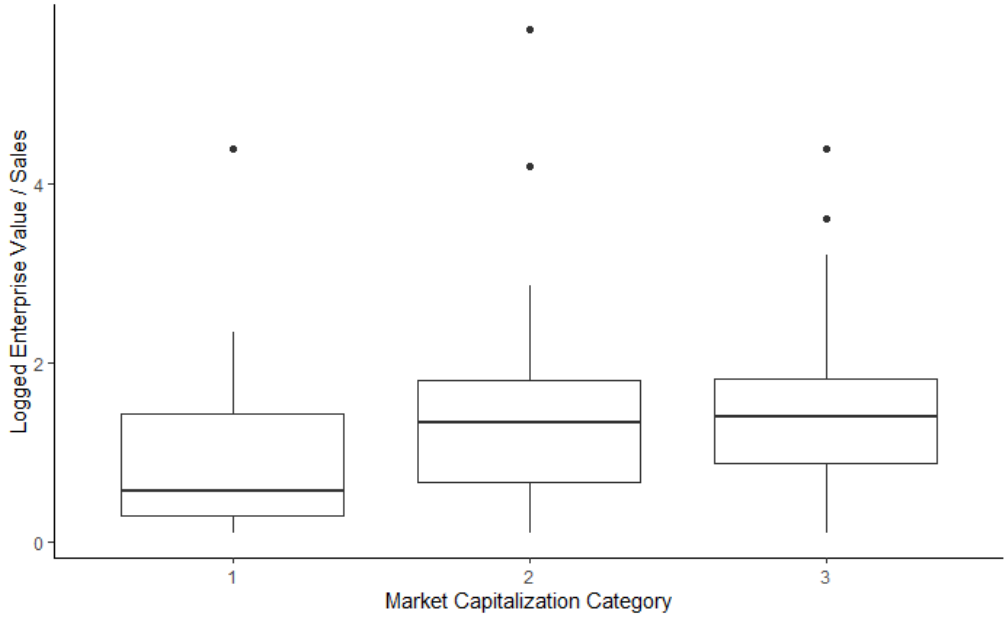

Figure 58. The effect of market capitalization category (size factor) for unapproved manufacturers on forward enterprise value / sales (*F*~(1,249)~ = 1.713; *P* = 0.1824). The Adjusted R-Squared is 0.005.

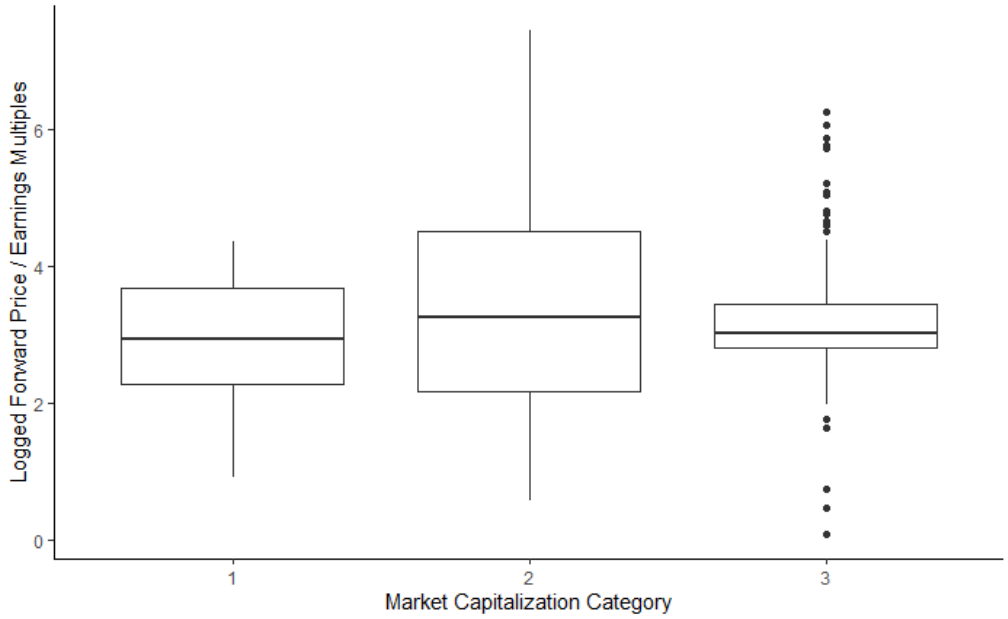

Figure 59. The effect of market capitalization category (size factor) for unapproved manufacturers on forward price / earnings multiples (*F*~(1,249)~ = 0.7904; *P* = 0.4548). The Adjusted R-Squared is -0.001.

Total Drug Dataset

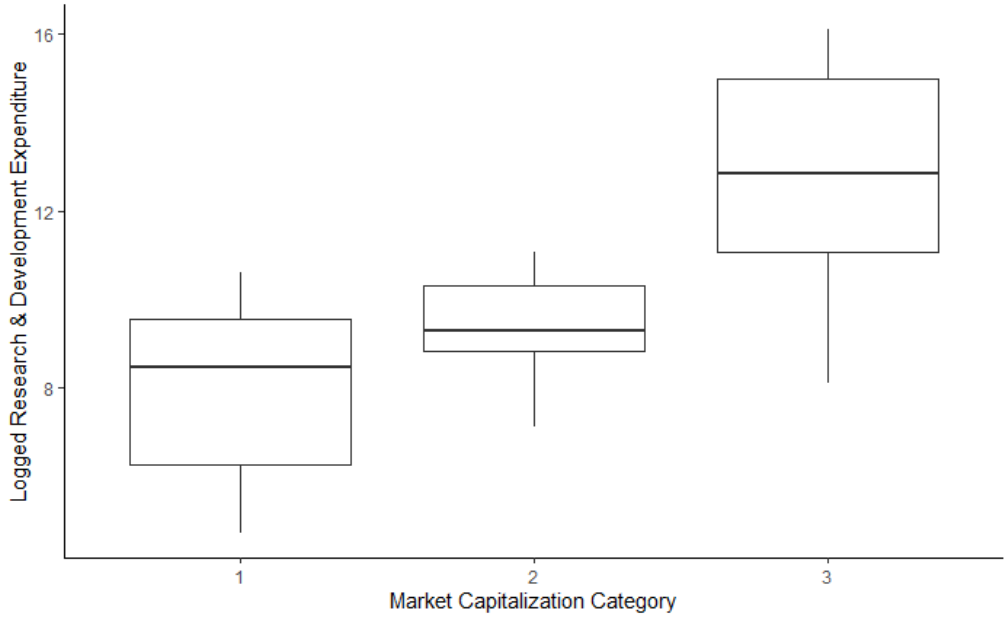

Figure 60. The effect of market capitalization category (size factor) for all manufacturers on research and development expenditure (*F*~(2,580)~ = 1.702; *P* = 0.1833). The Adjusted R-Squared is 0.002.

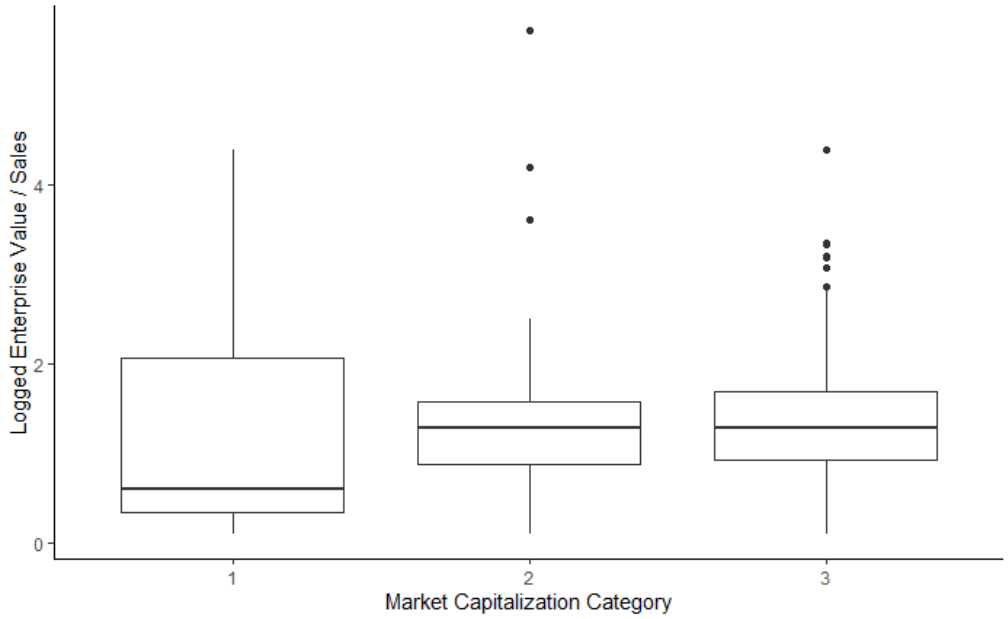

Figure 61. The effect of market capitalization category (size factor) for all manufacturers on forward enterprise value / sales multiples (*F*~(2,580)~ = 5.73; *P* = 0.0034). The Adjusted R-Squared is 0.016.

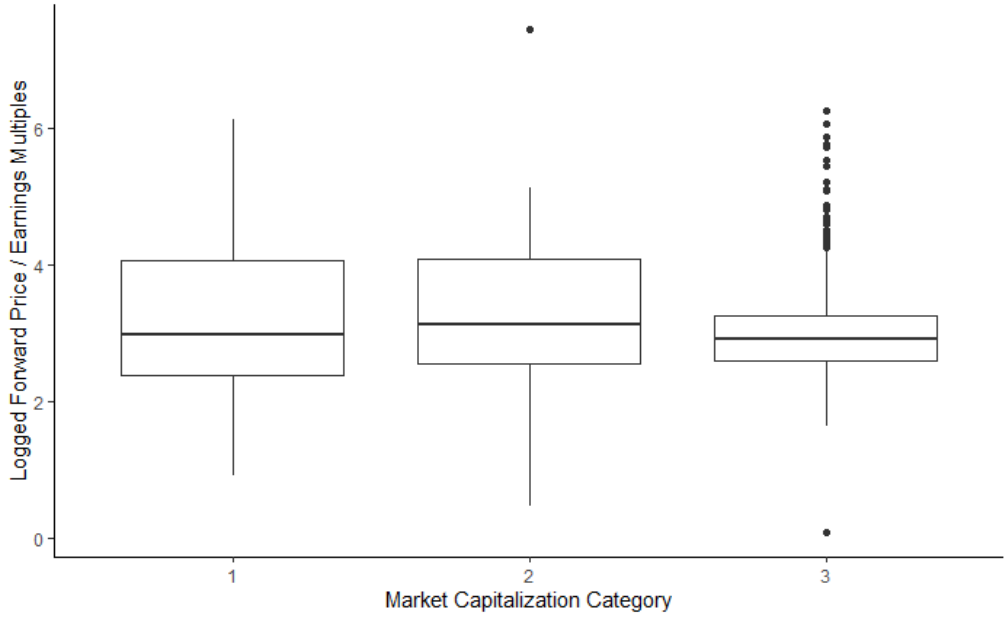

Figure 62. The effect of market capitalization category (size factor) for all manufacturers on forward price / earnings multiples (*F*~(2,580)~ = 1.702; *P* = 0.1833). The Adjusted R-Squared is 0.002.

R Studio Project Code

````
---
title: "SPROJ"
author: "Nick"
date: "3/6/2023"
output: html_document
---

# R Startup - Installing Packages
```{r}
install.packages("ggplot2")
install.packages("readr")
install.packages("dplyr")
install.packages("mice", dependencies = TRUE)
install.packages("caret")
install.packages("ggpmisc")
```

# R Startup - Loading Packages
```{r}
library(ggpmisc)
library(ggpubr)
library(caret)
library(mice)
library(dplyr)
library(ggplot2)
library(readr)
```

# Dataset Importing - Historical Approved Drug Data + Converting Characters into Numerics
```{r}
Drug_Data <- read.csv("C:/Users/nicho/Desktop/SPROJ/Data/R Studio/Datasets/Drug_Data.csv")
Drug_Data <- Drug_Data %>%
  mutate_at(vars(RD_Spend, EV_Sales, FDA_Approvals, PE, Year, Size_Factor, Mkt_Cap), as.numeric)
```

# Dataset Importing - Historical Unapproved Drug Data + Converting Characters into Numerics
```{r}
Unapproved_Drug_Data <- read.csv("C:/Users/nicho/Desktop/SPROJ/Data/R Studio/Datasets/Unapproved_Drug_Data.csv")
Unapproved_Drug_Data <- Unapproved_Drug_Data %>%
  mutate_at(vars(RD_Spend, FDA_Approvals, EV_Sales, PE, Year, Size_Factor, Mkt_Cap), as.numeric)
```

# Dataset Importing - Historical Unapproved + Approved Drug Data + Converting Characters into Numerics
```{r}
Total_Drug_Data <- read.csv("C:/Users/nicho/Desktop/SPROJ/Data/R Studio/Datasets/Approved_Unapproved_Drug_Data.csv")
Total_Drug_Data <- Total_Drug_Data %>%
  mutate_at(vars(RD_Spend, FDA_Approvals, EV_Sales, PE, Year, Size_Factor, Success, Mkt_Cap), as.numeric)

head(Total_Drug_Data)
```

# Dataset Importing - Summary Approval Data
```{r}
Approval_Visuals <- read.csv("C:/Users/nicho/Desktop/SPROJ/Data/R Studio/Datasets/Approval_Visuals.csv")
```

# Creating Functions - Master Function; Computes Lagged R&D Efficiency alongside Forward Price/Earnings and EV/ Sales Multiples
```{r}
compute_rd_lag <- function(data, lag_years) {
  result <- data.frame()
  for (year in unique(data$Year)) {
    for (company in unique(data$Company)) {
      if (!is.na(max(data$Year))) {
        if ((year + lag_years) <= max(data$Year)) {  # check if there are enough years of data available
          company_data <- subset(data, Mkt_Cap == Mkt_Cap & Experience == Experience & Size_Factor == Size_Factor & PE == PE & EV_Sales == EV_Sales & Company == company & Year == year)
          if (nrow(company_data) > 0) { # check if company_data is not empty
            rd <- company_data$RD_Spend[1] # use the first row of company_data to get R&D expense for the current year
            future_approvals_year <- year + lag_years
            approvals <- subset(data, Company == company & Year == future_approvals_year)$FDA_Approvals # use the row that's lag_years years ahead to get FDA approvals for the lagged year
            pe <- subset(data, Company == company & Year == future_approvals_year)$PE
            EV_Sales <- subset(data, Company == company & Year == future_approvals_year)$EV_Sales
            RD_Spend <- subset(data, Company == company & Year == future_approvals_year)$RD_Spend
            if (length(approvals) > 0 && approvals != 0) { # filter out zero approvals and empty values
              efficiency <- rd / approvals
              Size_Factor <- median(company_data$Size_Factor)
              Mkt_Cap <- median(company_data$Mkt_Cap)
              Experience <- median(company_data$Experience)
              result <- rbind(result, data.frame(Company = company, Experience = Experience, Year = year, RD_efficiency = efficiency, Size_Factor = Size_Factor, PE = pe, EV_Sales = EV_Sales,
RD_Spend = RD_Spend, Mkt_Cap = Mkt_Cap))
            }
          }
        }
      } else {
        print("Error: Missing values in data$Year column.")
        return(NULL)
      }
    }
  }
  result <- na.omit(result) # remove rows that contain "NA"
  return(result)
}
```

# Statistical Analysis
**Computing R&D Efficiency - Lagged**
```{r}
lag_test_0 <- compute_rd_lag(Drug_Data, 0)
lag_test_1 <- compute_rd_lag(Drug_Data, 1)
lag_test_2 <- compute_rd_lag(Drug_Data, 2)
lag_test_3 <- compute_rd_lag(Drug_Data, 3)
lag_test_4 <- compute_rd_lag(Drug_Data, 4)
lag_test_5 <- compute_rd_lag(Drug_Data, 5)
lag_test_6 <- compute_rd_lag(Drug_Data, 6)
lag_test_7 <- compute_rd_lag(Drug_Data, 7)
lag_test_8 <- compute_rd_lag(Drug_Data, 8)
lag_test_9 <- compute_rd_lag(Drug_Data, 9)
lag_test_10 <- compute_rd_lag(Drug_Data, 10)
lag_test_11 <- compute_rd_lag(Drug_Data, 11)
lag_test_12 <- compute_rd_lag(Drug_Data, 12)
lag_test_13 <- compute_rd_lag(Drug_Data, 13)
```

**Number of Observations per Lag**
```{r}
lag_0 <- nrow(lag_test_0)
lag_1 <- nrow(lag_test_1)
lag_2 <- nrow(lag_test_2)
lag_3 <- nrow(lag_test_3)
lag_4 <- nrow(lag_test_4)
lag_5 <- nrow(lag_test_5)
lag_6 <- nrow(lag_test_6)
lag_7 <- nrow(lag_test_7)
lag_8 <- nrow(lag_test_8)
lag_9 <- nrow(lag_test_9)
lag_10 <- nrow(lag_test_10)
lag_11 <- nrow(lag_test_11)
lag_12 <- nrow(lag_test_12)
lag_13 <- nrow(lag_test_13)
```

# Statistical Analysis: 1 - 6 Year Lagged Variables (P/E to R&D Efficiency)
```{r}
# 0 Year Lag
lag_pe_rd_model <- lm(lag_test_0$PE ~ lag_test_0$RD_efficiency)
````
````

```
summary(lag_pe_rd_model)

# 1 Year Lag
lag_pe_rd_model_1 <- lm(lag_test_1$PE ~ lag_test_1$RD_efficiency)
summary(lag_pe_rd_model_1)

# 2 Year Lag
lag_pe_rd_model_2 <- lm(lag_test_2$PE ~ lag_test_2$RD_efficiency)
summary(lag_pe_rd_model_2)

# 3 Year Lag
lag_pe_rd_model_3 <- lm(lag_test_3$PE ~ lag_test_3$RD_efficiency)
summary(lag_pe_rd_model_3)

# 4 Year Lag
lag_pe_rd_model_4 <- lm(lag_test_4$PE ~ lag_test_4$RD_efficiency)
summary(lag_pe_rd_model_4)

# 5 Year Lag
lag_pe_rd_model_5 <- lm(lag_test_5$PE ~ lag_test_5$RD_efficiency)
summary(lag_pe_rd_model_5)

# 6 Year Lag
lag_pe_rd_model_6 <- lm(lag_test_6$PE ~ lag_test_6$RD_efficiency)
summary(lag_pe_rd_model_6)
```

Statistical Analysis: 1 - 6 Year Lagged Variables (EV/Sales to R&D Efficiency)
```{r}
# 0 Year Lag
lag_EVs_rd_model <- lm(lag_test_0$EV_Sales ~ lag_test_0$RD_efficiency)
summary(lag_EVs_rd_model)

# 1 Year Lag
lag_EVs_rd_model_1 <- lm(lag_test_1$EV_Sales ~ lag_test_1$RD_efficiency)
summary(lag_EVs_rd_model_1)

# 2 Year Lag
lag_EVs_rd_model_2 <- lm(lag_test_2$EV_Sales ~ lag_test_2$RD_efficiency)
summary(lag_EVs_rd_model_2)

# 3 Year Lag
lag_EVs_rd_model_3 <- lm(lag_test_3$EV_Sales ~ lag_test_3$RD_efficiency)
summary(lag_EVs_rd_model_3)

# 4 Year Lag
lag_EVs_rd_model_4 <- lm(lag_test_4$EV_Sales ~ lag_test_4$RD_efficiency)
summary(lag_EVs_rd_model_4)

# 5 Year Lag
lag_EVs_rd_model_5 <- lm(lag_test_5$EV_Sales ~ lag_test_5$RD_efficiency)
summary(lag_EVs_rd_model_5)

# 6 Year Lag
lag_EVs_rd_model_6 <- lm(lag_test_6$EV_Sales ~ lag_test_6$RD_efficiency)
summary(lag_EVs_rd_model_6)
```

Statistical Analysis - Additional Examination

Process Step - Creating Lagged Variables
```{r}
RD_Efficiency <- lag_test_0$RD_efficiency
Size_Factor <- lag_test_0$Size_Factor
Mkt_Cap <- lag_test_0$Mkt_Cap
RD_Spend <- lag_test_0$RD_Spend
PE <- lag_test_0$PE
EV_Sales <- lag_test_0$EV_Sales
```

Statistical Analysis - R&D Efficiency vs. Size Factor
```{r}
RDE_S_model <- lm(RD_Efficiency ~ factor(Size_Factor))
summary(RDE_S_model)
```
Note: Adjusted R-squared: 0.32, p-value: 3.503e-06, F-statistic: 15.6, DF of 60

Statistical Analysis - R&D Spend vs. Size Factor
```{r}
lag_model_rd <- lm(RD_Spend ~ factor(Size_Factor))
summary(lag_model_rd)
```
Note: Adjusted R-squared: 0.39, p-value: 8.204e-08, F-statistic: 21.47, DF of 62

Data Visualization - Single Linear Regressions

Primary Data Visualization Analysis

Linear Regressions

  R&D Efficiency vs. Forward Earnings Multiples

Statistical Analysis - Linear Regression - R&D Efficiency vs. Forward EV/Sales
```{r}
RDe_EVs <- lm(data=lag_test_0, EV_Sales~RD_Efficiency)
summary(RDe_EVs)
anova(RDe_EVs)
```

Linear Regression - R&D Efficiency vs. Forward EV/Sales
```{r}
Fig8 <- ggplot(lag_test_0, aes(x=RD_efficiency, y=EV_Sales)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Research & Development Expenditure Efficiency")+ylab("Forward Enterprise Value / Sales Multiples")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig8
```
Figure 1. The effect of research and development expenditure efficiency on forward Enterprise Value / Sales (*F*~(1,63)~ = 2.448; *P* < 0.1227). The regression line is explained by Y = 4.722X - 1.751e-07. The Adjusted R-Squared is 0.022.

Statistical Analysis - Linear Regression - R&D Efficiency vs. Forward Price/Earnings
```{r}
RDe_PE <- lm(data=lag_test_0, PE~RD_Efficiency)
summary(RDe_PE)
anova(RDe_PE)
```

Linear Regression - R&D Efficiency vs. Forward Price/Earnings
```{r}
Fig9 <- ggplot(lag_test_0, aes(x=RD_efficiency, y=PE)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Research & Development Expenditure Efficiency")+ylab("Forward Price / Earnings Multiples")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig9
```
Figure 2. The effect of research and development expenditure efficiency on forward Price / Earnings (*F*~(1,63)~ = 2.448; *P* < 0.007938). The regression line is explained by Y = 2.141e+01X - 1.018e-06. The Adjusted R-Squared is 0.092.

  R&D Efficiency vs. Market Capitalization

Statistical Analysis - Linear Regression - R&D Efficiency vs. Market Capitalization

```{r}
RDe_MktCap <- lm(data=lag_test_0, Mkt_Cap~RD_Efficiency)
summary(RDe_MktCap)
anova(RDe_MktCap)
```

Linear Regression - R&D Efficiency vs. Market Capitalization
```{r}
Fig10 <- ggplot(lag_test_0, aes(x=RD_efficiency, y=Mkt_Cap)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Research & Development Expenditure Efficiency")+ylab("Market Capitalization")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig10
```
Figure 3. The effect of research and development expenditure efficiency on market capitalization (*F*~(1,63)~ = 75.479; *P* < 2.252). The regression line is explained by Y = 4.809e+07X + 2.263e+01. The Adjusted R-Squared is 0.537.

Statistical Analysis - Linear Regression - R&D Efficiency vs. Size Factor
```{r}
RDe_Size <- lm(data=lag_test_0, Size_Factor~RD_Efficiency)
summary(RDe_Size)
anova(RDe_Size)
```

Linear Regression - R&D Efficiency vs. Size Factor
```{r}
Fig11 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=RD_efficiency)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Efficiencey of Research & Development Expenditure") +
        theme_classic()
Fig11
```
Figure 4. The effect of research and development expenditure efficiency on Size Factor (*F*~(1,63)~ = 32.493; *P* < 2.373e-06). The regression line is explained by Y = 2.301e+00X + 1.319e-07. The Adjusted R-Squared is 0.3298.

R&D Efficiency vs. R&D Spend Analysis

Statistical Analysis - Linear Regression - R&D Efficiency vs. R&D Spend
```{r}
RDe_Spend <- lm(data=lag_test_0, RD_Spend~RD_Efficiency)
summary(RDe_Spend)
anova(RDe_Spend)
```

Linear Regression - R&D Efficiency vs. R&D Spend
```{r}
Fig12 <- ggplot(lag_test_0, aes(x=RD_efficiency, y=RD_Spend)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Research & Development Expenditure Efficiency")+ylab("Research & Development Expenditure")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig12
```
Figure 5. The effect of research and development expenditure efficiency on research and development expenditure (*F*~(1,63)~ = 199.91; *P* < 2.2e-16). The regression line is explained by Y = 5.190e+05X + 9.635e-01. The Adjusted R-Squared is 0.7566.

R&D Spend vs. Forward Earnings Multiples

Statistical Analysis - Linear Regression - R&D Spend vs. Enterprise Value / Sales
```{r}
RDs_EVs <- lm(data=lag_test_0, EV_Sales~RD_Spend)
summary(RDs_EVs)
anova(RDs_EVs)
```

Linear Regression - R&D Spend vs. Enterprise Value / Sales
```{r}
Fig13 <- ggplot(lag_test_0, aes(x=RD_Spend, y=EV_Sales)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Research & Development Expenditure")+ylab("Forward Enterprise Value / Sales")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig13
```
Figure 6. The effect of research and development expenditure on forward enterprise value / sales (*F*~(1,63)~ = 1.8027; *P* < 0.1842). The regression line is explained by Y = 4.659e+00X - 1.367e-07. The Adjusted R-Squared is 0.0123.

Statistical Analysis - Linear Regression - R&D Spend vs. Price / Earnings
```{r}
RDs_PE <- lm(data=lag_test_0, PE~RD_Spend)
summary(RDs_PE)
anova(RDs_PE)
```

Linear Regression - R&D Spend vs. Price / Earnings
```{r}
Fig14 <- ggplot(lag_test_0, aes(x=RD_Spend, y=PE)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Research & Development Expenditure")+ylab("Forward Price / Earnings")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig14
```
Figure 7. The effect of research and development expenditure on forward price / earnings (*F*~(1,63)~ = 1.8027; *P* < 0.00662). The regression line is explained by Y = 2.156e+01X - 9.412e-07. The Adjusted R-Squared is 0.09714.

Statistical Analysis - Linear Regression - R&D Spend vs. Market Capitalization
```{r}
RDs_MktCap <- lm(data=lag_test_0, Mkt_Cap~RD_Spend)
summary(RDs_MktCap)
anova(RDs_MktCap)
```

Linear Regression - R&D Spend vs. Market Capitalization
```{r}
Fig15 <- ggplot(lag_test_0, aes(x=RD_Spend, y=Mkt_Cap)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Research & Development Expenditure")+ylab("Market Capitalization")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig15
```
Figure 8. The effect of research and development expenditure on market capitalization (*F*~(1,63)~ = 101.4; *P* < 9.548e-15). The regression line is explained by Y = 4.184e+07X + 2.179e+01. The Adjusted R-Squared is 0.6107.

Statistical Analysis - Linear Regression - R&D Spend vs. Size Factor
```{r}
RDs_Size <- lm(data=lag_test_0, Size_Factor~RD_Spend)
summary(RDs_Size)
anova(RDs_Size)
```

Linear Regression - R&D Spend vs. Size Factor
```{r}
Fig16 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=RD_Spend)) +
        geom_boxplot() +
```

```
        xlab("Market Capitalization Category") +
        ylab("Research & Development Expenditure") +
        theme_classic()
Fig16
```
Figure 9. The effect of research and development expenditure on Size Factor (*F*~(1,63)~ = 37.172; *P* < 7.257e-08). The regression line is explained by Y = 2.273e+00X + 1.246e-07. The Adjusted R-Squared is 0.3611.

Forward Enterprise Value / Sales vs. Forward Earnings Multiples

Statistical Analysis - Linear Regression - Forward Enterprise Value / Sales vs. Forward Price / Earnings
```{r}
EVs_PE <- lm(data=lag_test_0, PE~EV_Sales)
summary(EVs_PE)
anova(EVs_PE)
```

Linear Regression - Forward Enterprise Value / Sales vs. Forward Price / Earnings
```{r}
Fig17 <- ggplot(lag_test_0, aes(x=EV_Sales, y=PE)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Forward Enterprise Value / Sales")+ylab("Forward Price / Earnings")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig17
```
Figure 10. The effect of forward enterprise value / sales on forward price / earnings (*F*~(1,63)~ = 86.39; *P* < 2.008e-13). The regression line is explained by Y = 7.3239X + 2.6186. The Adjusted R-Squared is 0.5716.

Statistical Analysis - Linear Regression - Forward Enterprise Value / Sales vs. Market Capitalization
```{r}
EVs_Mktcap <- lm(data=lag_test_0, EV_Sales~Mkt_Cap)
summary(EVs_Mktcap)
anova(EVs_Mktcap)
```

Linear Regression - Forward Enterprise Value / Sales vs. Market Capitalization
```{r}
Fig18 <- ggplot(lag_test_0, aes(x=Mkt_Cap, y=EV_Sales)) +
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  theme_classic()+
  xlab("Forward Enterprise Value / Sales")+ylab("Market Capitalization")+
  theme_classic() +
  theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
Fig18
```
Figure 11. The effect of research and development expenditure on Size Factor (*F*~(1,63)~ = 0.891; *P* < 0.3488). The regression line is explained by Y = 101183036X + 3998133. The Adjusted R-Squared is 0.3488.

Statistical Analysis - Linear Regression - Forward Enterprise Value / Sales vs. Size Factor
```{r}
EVs_Size <- lm(data=lag_test_0, Size_Factor~EV_Sales)
summary(EVs_Size)
anova(EVs_Size)
```

Linear Regression - Forward Enterprise Value / Sales vs. Size Factor
```{r}
Fig19 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=EV_Sales)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Enterprise Value / Sales") +
        theme_classic()
Fig19
```
Figure 12. The effect of Forward Enterprise Value / Sales Multiples on Size Factor (*F*~(1,63)~ = 2.4088 ; *P* < 0.1257). The regression line is explained by Y = 2.50737x + 0.04791. The Adjusted R-Squared is 0.02154.

Forward Price/Earnings vs. Size Factor

Statistical Analysis - Linear Regression - Forward Price / Earnings vs. Size Factor
```{r}
PE_Size <- lm(data=lag_test_0, Size_Factor~PE)
summary(PE_Size)
anova(PE_Size)
```

Linear Regression - Forward Price / Earnings vs. Size Factor
```{r}
Fig20 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=PE)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Forward Price / Earnings Multiples") +
        theme_classic()
Fig20
```
Figure 13. .

Multivariable Linear Regressions

Statistical Analysis - EV / Sales - Multi-variable Linear Regression Analysis

```{r}
install.packages("stargazer")
library(stargazer)
```

Statistical Analysis - EV / Sales Model 1: Model Simplification (Short Version)

Statistical Analysis - EV / Sales Model 1: Model Simplification - R&D Efficiency Version
```{r}
Model_1 <- lm(EV_Sales~RD_efficiency*Mkt_Cap*factor(Size_Factor), data=lag_test_0)
summary(Model_1)
stargazer(Model_1, type = "text", title = "Multivariable Linear Regression Results",
  font.size = "small", header = FALSE, digits = 2, se = NULL,
  omit.stat = "f", column.sep.width = "0.3")
```
Figure 14. The effect of research & development expenditure efficiency, market capitalization, and size, on forward enterprise value / sales multiples (*F*~(9,55)~ = 3.983; *P* = 0.0005). The Adjusted R-Squared is 0.2955..

Statistical Analysis - EV / Sales Model 1x: Model Simplification - R&D Efficiency Version, less Market Capitalization
```{r}
Model_1x <- lm(EV_Sales~RD_efficiency*factor(Size_Factor), data=lag_test_0)
summary(Model_1x)
stargazer(Model_1x, type = "text", title = "Multivariable Linear Regression Results",
  font.size = "small", header = FALSE, digits = 2, se = NULL,
  omit.stat = "f", column.sep.width = "0.3")
```
Figure 15. The effect of research & development expenditure efficiency, factored by market capitalization size on forward enterprise value / sales multiples (*F*~(1,59)~ = 4.25; *P* < 0.002286). The Adjusted R-Squared is 0.2025.

Statistical Analysis - EV / Sales Model 1xx: Model Simplification - R&D Efficiency Version, less Size Factor, only Market Capitalization
```{r}
Model_1xx <- lm(EV_Sales~RD_efficiency*Mkt_Cap, data=lag_test_0)
summary(Model_1xx)
stargazer(Model_1xx, type = "text", title = "Multivariable Linear Regression Results",
  font.size = "small", header = FALSE, digits = 2, se = NULL,
  omit.stat = "f", column.sep.width = "0.3")
```
Figure 16. The effect of research & development expenditure efficiency and market capitalization on forward enterprise value / sales multiples (*F*~(1,61)~ = 4.687; *P* < 0.005202). The Adjusted R-Squared is 0.1473.

```{r}
Model_2xx <- lm(EV_Sales~RD_efficiency+Mkt_Cap, data=lag_test_0)
```

```
summary(Model_1xx)
stargazer(Model_1xx, type = "text", title = "Multivariable Linear Regression Results",
  font.size = "small", header = FALSE, digits = 2, se = NULL,
  omit.stat = "f", column.sep.width = "0.3")
```

Statistical Analysis - Price / Earnings - Multi-variable Linear Regression Analysis

Statistical Analysis - Price / Earnings Model 1b: Model Simplification (Short Version)

Statistical Analysis - Price / Earnings Model 1b: Model Simplification - R&D Efficiency Version
```{r}
Model_1b <- lm(PE~RD_efficiency*Mkt_Cap*factor(Size_Factor), data=lag_test_0)
summary(Model_1b)
stargazer(Model_1b, type = "text", title = "Multivariable Linear Regression Results",
  font.size = "small", header = FALSE, digits = 2, se = NULL,
  omit.stat = "f", column.sep.width = "0.3")
```
Figure 17. The effect of research & development expenditure efficiency, market capitalization, factored by Size on forward price / earnings multiples (*F*~(1,54)~ = 2.389; *P* < 0.01988). The Adjusted R-Squared is 0.1784.

Statistical Analysis - Price / Earnings Model 1bx: Model Simplification - R&D Efficiency Version, less Market Capitalization
```{r}
Model_1bx <- lm(PE~RD_efficiency*factor(Size_Factor), data=lag_test_0)
summary(Model_1bx)
stargazer(Model_1bx, type = "text", title = "Multivariable Linear Regression Results",
  font.size = "small", header = FALSE, digits = 2, se = NULL,
  omit.stat = "f", column.sep.width = "0.3")
```
Figure 18. The effect of research & development expenditure efficiency, factored by market capitalization size on forward price / earnings multiples (*F*~(1,59)~ = 3.543; *P* < 0.007221). The Adjusted R-Squared is 0.1657.

Statistical Analysis - Price / Earnings Model 1bxx: Model Simplification - R&D Efficiency Version, less Size Factor, only Market Capitalization
```{r}
Model_1bxx <- lm(PE~RD_efficiency*Mkt_Cap, data=lag_test_0)
summary(Model_1bxx)
stargazer(Model_1bxx, type = "text", title = "Multivariable Linear Regression Results",
  font.size = "small", header = FALSE, digits = 2, se = NULL,
  omit.stat = "f", column.sep.width = "0.3")
```
Figure 19. The effect of research & development expenditure efficiency and market capitalization on forward price / earnings multiples (*F*~(1,61)~ = 4.47; *P* < 0.006672). The Adjusted R-Squared is 0.1399.

Statistical Analysis - Price / Earnings Model 1b3x: Multi-linear Regression - R&D Efficiency, Fwd. EV/Sales, Market Capitalization, Size Factor
```{r}
Model_1b3x <- lm(PE~RD_efficiency*EV_Sales*factor(Size_Factor)*Mkt_Cap, data=lag_test_0)
summary(Model_1b3x)
stargazer(Model_1b3x, type = "text", title = "Multivariable Linear Regression Results",
  font.size = "small", header = FALSE, digits = 2, se = NULL,
  omit.stat = "f", column.sep.width = "0.3")
```
Figure 20.

K-Means Clusters Analysis

Statistical Analysis - K-Means Cluster Analysis - Approved Drugs Dataset

Statistucal Analysis - Installing Cluster Packages and Loading

Statistical Analysis - K-cluster Package Installation and Load-out
```{r}
install.packages("cluster")
library(cluster)
```

Statistical Analysis - K-Means Cluster Analysis - Refining Approved Drug Data Dataset
```{r}
Approved_Drug_Data_Start <- Drug_Data[, sapply(Drug_Data, is.numeric)]
Approved_Drug_Data_no_EVE_1 <- Approved_Drug_Data_Start[, -3]
Approved_Drug_Data_no_Yr_1 <- Approved_Drug_Data_no_EVE_1[, -5]
Approved_Drug_Data <- Approved_Drug_Data_no_Yr_1[, -6]
```

Statistical Analysis - K-Means Clusters - Log Transforming Market Capitalization
```{r}
Log_Approved_DD <- data.frame(EV_Sales = Approved_Drug_Data$EV_Sales)
Log_Approved_DD$FDA_Approvals <- Approved_Drug_Data$FDA_Approvals
Log_Approved_DD$Mkt_Cap_Log <- log(Approved_Drug_Data$Mkt_Cap)
Log_Approved_DD$RD_Spend <- log(Approved_Drug_Data$RD_Spend)
Log_Approved_DD$PE <- Approved_Drug_Data$PE
Log_Approved_DD <- na.omit(Log_Approved_DD)
Log_Approved_DD
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) for my "Approved Drug Data" Dataframe
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(Log_Approved_DD, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of Clusters", ylab = "Within Groups Sum of Squares")
```
Note: Elbow seems ideal @ 3 or 4

Statistical Analysis - Assigning # of Clusters for my "Approved Drug Data" Dataframe
```{r}
k <- 3
```

Statistical Analysis - K-cluster Mean Analysis - Approved Drug Data Dataframe
```{r}
kmeans_Approved_DD <- kmeans(Log_Approved_DD[, 1:3], k, nstart = 10, iter.max = 100)
```

Statistical Analysis - K-cluster Mean Analysis Summary - Approved Drug Data Dataframe
```{r}
summary(kmeans_Approved_DD)
```

Data Visualization - K-cluster Mean Analysis - Approved Drug Data Dataframe

Importing Average Approved Drug Dataset
```{r}
Drug_Data_Average<- read.csv("C:/Users/nicho/Desktop/SPROJ/Data/R Studio/Datasets/Drug_Data_Average.csv")
```

Statistical Analysis - K-Means Cluster Analysis - Refining Average Approved Drug Data Dataset
```{r}
Aveage_Approved_Drug_Data <- Drug_Data_Average[, sapply(Drug_Data_Average, is.numeric)]
Average_Approved_Drug_Data_no_EVE_1 <- Aveage_Approved_Drug_Data[, -3]
Average_Approved_Drug_Data_KC_F <- Average_Approved_Drug_Data_no_EVE_1[, -6]
Avg_Approved_DD <- Average_Approved_Drug_Data_KC_F
```

Statistical Analysis - K-Means Clusters - Log Transforming Market Capitalization
```{r}
Log_Avg_Approved_DD <- data.frame(EV_Sales = Avg_Approved_DD$Avg_EV_Sales)
Log_Avg_Approved_DD$FDA_Approvals <- Avg_Approved_DD$FDA_Approvals
Log_Avg_Approved_DD$Avg_PE <- Avg_Approved_DD$Avg_PE
Log_Avg_Approved_DD$Mkt_Cap_Log <- log(Avg_Approved_DD$Mkt_Cap)
Log_Avg_Approved_DD$Avg_RD_Spend_Log <- log(Avg_Approved_DD$Avg_RD_Spend)
Log_Avg_Approved_DD <- na.omit(Log_Avg_Approved_DD)
```

Log_Avg_Approved_DD
```



**Data Visualization - K-cluster Mean Analysis - FDA Approvals & Average EV/Sales**
```{r}
plot(Log_Avg_Approved_DD[, 2], Log_Avg_Approved_DD[, 1], col = fit$cluster,
     xlab = "Cumulative FDA Approvals", ylab = "Enterprise Value / Sales",
     xlim = c(0, 25), ylim = c(0, 25))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 2)
```

Data Visualization - K-cluster Mean Analysis - Cumulative FDA Approvals & Average P/E
```{r}
plot(Log_Avg_Approved_DD[, 2], Log_Avg_Approved_DD[, 3], col = fit$cluster,
     xlab = "Cumulative FDA Approvals", ylab = "Forward Price / Earnings Multiples",
     xlim = c(0, 20), ylim = c(0, 50))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 2)
```
Creating Cumulative FDA Approvals & Market Capitalization Dataframe
```{r}
FDA_Mktcap_df <- data.frame(Mkt_Cap = Log_Avg_Approved_DD$Mkt_Cap_Log)
FDA_Mktcap_df$FDA <- Log_Avg_Approved_DD$FDA_Approvals
FDA_Mktcap_df <- na.omit(FDA_Mktcap_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Cumulative FDA Approvals & Market Capitalization
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(FDA_Mktcap_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 4-5 clusters ideal

Data Visualization - K-cluster Mean Analysis - Cumulative FDA Approvals & Market Capitalization
```{r}
fit <- kmeans(FDA_Mktcap_df, 4)

plot(FDA_Mktcap_df[, 1], FDA_Mktcap_df[, 2], col = fit$cluster,
     xlab = "Cumulative FDA Approvals", ylab = "Market Capitalization",
     xlim = c(11, 21), ylim = c(0, 50))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 1.5)
```

Creating R&D Spend & Forward EV/Sales Dataframe
```{r}
RD_EVs_df <- data.frame(RD_Spend = Log_Approved_DD$RD_Spend)
RD_EVs_df$EV_Sales <- Log_Approved_DD$EV_Sales
RD_EVs_df <- na.omit(RD_EVs_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Forward Enterprise Value / Sales Multiple**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(RD_EVs_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 4 - 5 clusters ideal

Data Visualization - K-cluster Mean Analysis - Research & Development Expenditure & Forward Enterprise Value / Sales Multiple
```{r}
fit <- kmeans(RD_EVs_df, 4)

plot(Log_Approved_DD[, 4], Log_Approved_DD[, 1], col = fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Forward Enterprise Value / Sales Multiples",
     xlim = c(6, 17), ylim = c(0, 30))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 20

Creating R&D Spend & Market Capitalization Dataframe
```{r}
RD_Mktcap_df <- data.frame(RD_Spend = Log_Approved_DD$RD_Spend)
RD_Mktcap_df$Mkt_Cap <- Log_Approved_DD$Mkt_Cap_Log
RD_Mktcap_df <- na.omit(RD_Mktcap_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Market Capitalization**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(RD_Mktcap_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Research & Development Expenditure & Market Capitalization
```{r}
fit <- kmeans(RD_Mktcap_df, 3)

plot(RD_Mktcap_df[, 1], RD_Mktcap_df[, 2], col = fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Market Capitalization",
     xlim = c(6, 17), ylim = c(11, 20))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```

Figure 21

Creating P/E & EV/Sales Dataframe
```{r}
PE_EVs_df <- data.frame(EV_Sales = Log_Approved_DD$EV_Sales)
PE_EVs_df$PE <- Log_Approved_DD$PE
PE_EVs_df <- na.omit(PE_EVs_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - P/E & EV/Sales**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(PE_EVs_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```

Note: 3 - 4 clusters ideal

**Data Visualization - K-cluster Mean Analysis - Forward Price / Earnings & Enterprise Value / Sales Multiples **
```

```{r}
PE_EVs_fit <- kmeans(PE_EVs_df, 3)

plot(PE_EVs_df[, 2], PE_EVs_df[, 1], col = PE_EVs_fit$cluster,
     xlab = "Forward Enterprise Value / Sales Multiples", ylab = "Forward Price / Earnings Multiples",
     xlim = c(0, 255), ylim = c(0, 30))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```

Figure 22.

Creating Mkt Cap & EV/Sales Dataframe
```{r}
EVS_Mktcap <- data.frame(EV_Sales = Log_Approved_DD$EV_Sales)
EVS_Mktcap$Mkt_Cap <- Log_Approved_DD$Mkt_Cap_Log
EVS_Mktcap <- na.omit(EVS_Mktcap)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Mkt Cap & EV/Sales**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(EVS_Mktcap, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```

Note: 3-4 clusters ideal

Data Visualization - K-cluster Mean Analysis - EV/Sales & Mkt. Cap
```{r}
EVs_Mktcap_fit <- kmeans(EVS_Mktcap, 3)

plot(EVS_Mktcap[, 2], EVS_Mktcap[, 1], col = EVs_Mktcap_fit$cluster,
     xlab = "Market Capitalization", ylab = "Forward Enterprise Value / Sales Multiples",
     xlim = c(12, 20), ylim = c(0, 30))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 23.

Creating Mkt Cap & P/E Dataframe
```{r}
PE_Mktcap <- data.frame(PE = Log_Approved_DD$PE)
PE_Mktcap$Mkt_Cap <- Log_Approved_DD$Mkt_Cap_Log
PE_Mktcap <- na.omit(PE_Mktcap)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Mkt Cap & P/E**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(PE_Mktcap, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3-4 clusters ideal

Data Visualization - K-cluster Mean Analysis - PE & Mkt. Cap
```{r}
PE_Mktcap_fit <- kmeans(PE_Mktcap, 3)

plot(PE_Mktcap[, 2], PE_Mktcap[, 1], col = PE_Mktcap_fit$cluster,
     xlab = "Market Capitalization", ylab = "Forward Price / Earnings Multiples",
     xlim = c(12, 20), ylim = c(0, 130))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```

Figure. 24

Creating R&D Spend & R&D Effiency Dataframe
```{r}
RDs_RDe_df <- data.frame(RD_Spend = lag_test_0$RD_Spend)
RDs_RDe_df$RD_Efficiency <- lag_test_0$RD_efficiency
RDs_RDe_df <- na.omit(RDs_RDe_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Spend vs. R&D Efficiency**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(RDs_RDe_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```

Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - R&D Spend vs. R&D Efficiency
```{r}
RDs_RDe_df_fit <- kmeans(RDs_RDe_df, 3)

plot(RDs_RDe_df[, 1], RDs_RDe_df[, 2], col = RDs_RDe_df_fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Research & Development Expenditure Efficiency",
     xlim = c(8000, 10000000), ylim = c(8000, 10000000))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure. 25

Creating EV_Sales & R&D Effiency Dataframe
```{r}
RDe_EV_Sales_df <- data.frame(EV_Sales = lag_test_0$EV_Sales)
RDe_EV_Sales_df$RD_Efficiency <- lag_test_0$RD_efficiency
RDe_EV_Sales_df <- na.omit(RDe_EV_Sales_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - EV/Sales vs. R&D Efficiency**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(RDe_EV_Sales_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - R&D Spend vs. R&D Efficiency
```{r}
RDe_EV_Sales_df_fit <- kmeans(RDe_EV_Sales_df, 3)

plot(RDe_EV_Sales_df[, 2], RDe_EV_Sales_df[, 1], col = RDe_EV_Sales_df_fit$cluster,
     xlab = "Research & Development Expenditure Efficiency", ylab = "Forward Enterprise Value / Sales Multiples",
     xlim = c(8000, 10000000), ylim = c(0, 15))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure. 26

Creating P/E & R&D Effiency Dataframe
```{r}
RDe_PE_df <- data.frame(PE = lag_test_0$PE)
RDe_PE_df$RD_Efficiency <- lag_test_0$RD_efficiency
RDe_PE_df <- na.omit(RDe_PE_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - PE vs. R&D Efficiency**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(RDe_PE_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - PE vs. R&D Efficiency
```{r}
RDe_PE_df_fit <- kmeans(RDe_PE_df, 3)

plot(RDe_PE_df[, 2], RDe_PE_df[, 1], col = RDe_PE_df_fit$cluster,
     xlab = "Research & Development Expenditure Efficiency", ylab = "Forward Price / Earnings Multiples",
     xlim = c(8000, 10000000), ylim = c(7, 42))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure. 27

Creating Mkt Cap & R&D Effiency Dataframe
```{r}
RDe_Mkt_Cap_df <- data.frame(Mkt_Cap = lag_test_0$Mkt_Cap)
RDe_Mkt_Cap_df$RD_Efficiency <- lag_test_0$RD_efficiency
RDe_Mkt_Cap_df <- na.omit(RDe_Mkt_Cap_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Mkt Cap vs. R&D Efficiency**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(RDe_Mkt_Cap_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Mkt Cap vs. R&D Efficiency
```{r}
RDe_Mkt_Cap_df_fit <- kmeans(RDe_Mkt_Cap_df, 3)

plot(RDe_Mkt_Cap_df[, 2], RDe_Mkt_Cap_df[, 1], col = RDe_Mkt_Cap_df_fit$cluster,
     xlab = "Research & Development Expenditure Efficiency", ylab = "Market Capitalization",
     xlim = c(8000, 10000000), ylim = c(1800000, 275000000))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure. 28

Creating Market Cap & Forward P/E Multiple Dataframe
```{r}
MktCap_PE_df <- data.frame(Mkt_Cap = Log_Approved_DD$Mkt_Cap_Log)
MktCap_PE_df$PE <- Log_Approved_DD$PE
MktCap_PE_df <- na.omit(MktCap_PE_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Market Cap & Forward Price / Earnings Multiple**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(MktCap_PE_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Market Cap & Forward Price / Earnings Multiple
```{r}
MktCap_PE_df_fit <- kmeans(MktCap_PE_df, 3)

plot(MktCap_PE_df[, 1], MktCap_PE_df[, 2], col = MktCap_PE_df_fit$cluster,
     xlab = "Market Capitalization", ylab = "Forward Price / Earnings Multiples",
     xlim = c(11, 20), ylim = c(5, 260))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 2)
```

Statistical Analysis - Enterprise Value / Sales, R&D Efficiency, and Market Capitalization Dataframe
```{r}
numeric_lag <- lag_test_0[, sapply(lag_test_0, is.numeric)]
numeric_lag_noY <- numeric_lag[, -1]
numeric_lag_noSF <- numeric_lag_noY[, -2]
numeric_lag_noPE <- numeric_lag_noSF[, -2]
numeric_lag_EVs_m1 <- numeric_lag_noPE[, -3]
```

```{r}
# create new dataframe with original EV_Sales variable
df_log <- data.frame(EV_Sales = numeric_lag_EVs_m1$EV_Sales)

# add log-transformed RD_efficiency variable
df_log$RD_efficiency_log <- log(numeric_lag_EVs_m1$RD_efficiency)

df_log$EV_Sales <- log(numeric_lag_EVs_m1$EV_Sales)

# add log-transformed Mkt_Cap variable
df_log$Mkt_Cap <- numeric_lag_EVs_m1$Mkt_Cap
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) for Enterprise Value / Sales, R&D Efficiency, and Market Capitalization Dataframe
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(df_log, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: Elbow seems ideal @ 3 or 4

Statistical Analysis - Assigning # of Clusters - EV/Sales, R&D Efficiency, and Mkt. Cap Dataframe
```{r}
k <- 3
```

Statistical Analysis - K-cluster Mean Analysis - EV/Sales, R&D Efficiency, and Mkt. Cap Dataframe
```{r}
kmeans_fit_EVs_m1 <- kmeans(df_log[, 1:3], k, nstart = 10, iter.max = 100)
```

Statistical Analysis - K-cluster Mean Analysis Summary - EV/Sales, R&D Efficiency, and Mkt. Cap Dataframe
```{r}
summary(kmeans_fit_EVs_m1)
```

Data Visualization - K-cluster Mean Analysis - EV/Sales, R&D Efficiency, and Mkt. Cap Dataframe
```{r}
# Set scaling factor for point size
size_factor <- 50000000

# Calculate size vector based on market capitalization
size_vector <- df_log[, 3] / size_factor

# Create scatterplot of data points, colored by cluster with flipped axes
plot(df_log[, 2], df_log[, 1], col = fit$cluster,
     xlab = "Logged Research & Development Efficiency", ylab = "Logged Forward Enterprise Value / Sales Multiple",
     xlim = c(8, 18), ylim = c(0, 3))

# Add cluster centers to the plot
points(fit$centers[, 2], fit$centers[, 1], pch = 20, cex = 2, col = 1:k)
```

Statistical Analysis - K-cluster Mean Analysis - Forward Enterprise Value / Sales and Price / Earnings Multiples; Cleaning Dataframe
```{r}
numeric_lag <- lag_test_0[, sapply(lag_test_0, is.numeric)]
numeric_lag_noY <- numeric_lag[, -1]
numeric_lag_noSF <- numeric_lag_noY[, -2]

# Price / Earnings K-Cluster DataFrame
numeric_lag_PE <- numeric_lag_noSF[, -3]

# Enterprise Value / Sales K-Cluster DataFrame
numeric_lag_EVs <- numeric_lag_noSF[, -2]
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) for P/E Data
```{r}
# Create a vector of WSS values for different numbers of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(numeric_lag_PE, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plot the WSS values against the number of clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: Elbow seems ideal @ 3

Statistical Analysis - Assigning # of Clusters
```{r}
k <- 4
```

```{r}
kmeans_fit <- kmeans(numeric_lag_PE, k)
```

Statistical Analysis - K-cluster Mean Analysis Summary
```{r}
summary(kmeans_fit)
```
Data Visualization - K-cluster Mean Analysis
```{r}
k <- 4
fit <- kmeans(numeric_lag_PE[, 1:4], k)

# Create scatterplot of data points, colored by cluster
plot(numeric_lag_PE[, 1], numeric_lag_PE[, 2], col = fit$cluster)

# Add cluster centers to the plot
points(fit$centers[, 1], fit$centers[, 2], pch = 20, cex = 2, col = 1:k)

```


Statistical Analysis - Unapproved Drug Dataset

K-Cluster Mean Analysis

Statistical Analysis - K-Means Cluster Analysis - Refining Unapproved Drug Data Dataset
```{r}
Unapproved_Drug_Data_Start <- Unapproved_Drug_Data[, sapply(Unapproved_Drug_Data, is.numeric)]
Unapproved_Drug_Data_no_Yr_2 <- Unapproved_Drug_Data_Start[, -5]
Unapproved_DD <- Unapproved_Drug_Data_no_Yr_2[, -6]
Unapproved_DD <- Unapproved_DD[Unapproved_DD$Mkt_Cap != 0, ]
Unapproved_DD <- Unapproved_DD[Unapproved_DD$RD_Spend != 0 & Unapproved_DD$RD_Spend != -150, ]
Unapproved_DD <- Unapproved_DD[Unapproved_DD$PE > 0 & Unapproved_DD$EV_Sales > 0, ]
Unapproved_DD <- na.omit(Unapproved_DD)
```

Statistical Analysis - K-Means Clusters - Log Transforming Market Capitalization & R&D Expenditure
```{r}
Log_Unapproved_DD <- data.frame(EV_Sales = Unapproved_DD$EV_Sales)
Log_Unapproved_DD$FDA_Approvals <- Unapproved_DD$FDA_Approvals
Log_Unapproved_DD$Mkt_Cap_Log <- log(Unapproved_DD$Mkt_Cap)
Log_Unapproved_DD$RD_Spend <- log(Unapproved_DD$RD_Spend)
Log_Unapproved_DD$PE <- Unapproved_DD$PE
Log_Unapproved_DD <- na.omit(Log_Unapproved_DD)
Log_Unapproved_DD
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) for my "Unapproved Drug Data" Dataframe
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(Log_Unapproved_DD, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: Elbow seems ideal @ 3 or 4

Statistical Analysis - Assigning # of Clusters for my "Unapproved Drug Data" Dataframe
```{r}
k <- 3
```

Statistical Analysis - K-cluster Mean Analysis - Unapproved Drug Data Dataframe
```{r}
kmeans_Unapproved_DD <- kmeans(Log_Unapproved_DD[, 1:3], k, nstart = 10, iter.max = 100)
```

Statistical Analysis - K-cluster Mean Analysis Summary - Unapproved Drug Data Dataframe
```{r}
summary(kmeans_Unapproved_DD)
```

Data Visualization - K-cluster Mean Analysis - Unapproved Drug Data Dataframe

Creating R&D Spend & Forward EV/Sales Dataframe
```{r}
URD_EVs_df <- data.frame(RD_Spend = Log_Unapproved_DD$RD_Spend)
URD_EVs_df$EV_Sales <- Log_Unapproved_DD$EV_Sales
URD_EVs_df <- na.omit(URD_EVs_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Forward Enterprise Value / Sales Multiple**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(URD_EVs_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 2 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Unapproved Dataset - Research & Development Expenditure & Forward Enterprise Value / Sales Multiple
```{r}
URD_EVs_df_fit <- kmeans(URD_EVs_df, 3)

plot(URD_EVs_df[, 1], URD_EVs_df[, 2], col = URD_EVs_df_fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Forward Enterprise Value / Sales Multiples",
     xlim = c(0, 20), ylim = c(0, 125))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure. 29

Creating R&D Spend & Forward PE Dataframe
```{r}
URD_PE_df <- data.frame(RD_Spend = Log_Unapproved_DD$RD_Spend)
URD_PE_df$PE <- Log_Unapproved_DD$PE
URD_PE_df <- na.omit(URD_PE_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Forward Price / Earnings Multiple**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(URD_PE_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 2 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Unapproved Dataset - Research & Development Expenditure & Forward Enterprise Value / Sales Multiple
```{r}
URD_PE_df_fit <- kmeans(URD_PE_df, 3)

plot(URD_PE_df[, 1], URD_PE_df[, 2], col = URD_PE_df_fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Forward Price / Earnings Multiples",
     xlim = c(0, 20), ylim = c(0, 550))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure. 30

Creating R&D Spend & Market Capitalization Dataframe - Unapproved Drug Data
```{r}
URD_Mktcap_df <- data.frame(RD_Spend = Log_Unapproved_DD$RD_Spend)
URD_Mktcap_df$Mkt_Cap <- Log_Unapproved_DD$Mkt_Cap_Log
URD_Mktcap_df <- na.omit(URD_Mktcap_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Unapproved R&D Expenditure & Market Capitalization**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(URD_Mktcap_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Unapproved Research & Development Expenditure & Market Capitalization
```{r}
URD_Mktcap_df_fit <- kmeans(URD_Mktcap_df, 3)

plot(URD_Mktcap_df[, 1], URD_Mktcap_df[, 2], col = URD_Mktcap_df_fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Market Capitalization",
     xlim = c(4, 16), ylim = c(8, 25))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 31.

Creating P/E & EV/Sales Dataframe - Unapproved Drug Dataset
```{r}
UPE_EVs_df <- data.frame(EV_Sales = Log_Unapproved_DD$EV_Sales)
UPE_EVs_df$PE <- Log_Unapproved_DD$PE
UPE_EVs_df <- na.omit(UPE_EVs_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - P/E & EV/Sales**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(UPE_EVs_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

**Data Visualization - K-cluster Mean Analysis - Forward Price / Earnings & Enterprise Value / Sales Multiples **
```{r}
UPE_EVs_df_fit <- kmeans(UPE_EVs_df, 3)

plot(UPE_EVs_df[, 1], UPE_EVs_df[, 2], col = UPE_EVs_df_fit$cluster,
     xlab = "Forward Enterprise Value / Sales Multiples", ylab = "Forward Price / Earnings Multiples",
     xlim = c(0, 40), ylim = c(0, 550))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 32.

Creating Mkt Cap & EV/Sales Dataframe - Unapproved Drug Dataset
```{r}
UEVS_Mktcap <- data.frame(EV_Sales = Log_Unapproved_DD$EV_Sales)
UEVS_Mktcap$Mkt_Cap <- Log_Unapproved_DD$Mkt_Cap_Log
UEVS_Mktcap <- na.omit(UEVS_Mktcap)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Unapproved Drug Dataset - Mkt Cap & EV/Sales**
```{r}
```

```
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(UEVS_Mktcap, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```

Note: 3-4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Unapproved Drug Dataset - EV/Sales & Mkt. Cap
```{r}
EVs_Mktcap_fit <- kmeans(UEVS_Mktcap, 3)

plot(UEVS_Mktcap[, 2], UEVS_Mktcap[, 1], col = EVs_Mktcap_fit$cluster,
     xlab = "Market Capitalization", ylab = "Forward Enterprise Value / Sales Multiples",
     xlim = c(0, 25), ylim = c(0, 100))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 33.

```{r}
# Creating Mkt Cap & P/E Dataframe - Unapproved Drugs Dataset
UPE_Mktcap <- data.frame(PE = Log_Unapproved_DD$PE)
UPE_Mktcap$Mkt_Cap <- Log_Unapproved_DD$Mkt_Cap_Log
UPE_Mktcap <- na.omit(UPE_Mktcap)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Unapproved Drugs Dataset - Mkt Cap & P/E
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(UPE_Mktcap, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3-4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Price / Earnings & Mkt. Cap - Unapproved Drugs Dataset
```{r}
UPE_Mktcap_fit <- kmeans(UPE_Mktcap, 4)

plot(UPE_Mktcap[, 2], UPE_Mktcap[, 1], col = UPE_Mktcap_fit$cluster,
     xlab = "Market Capitalization", ylab = "Forward  Price / Earnings Multiples",
     xlim = c(8, 20), ylim = c(0, 200))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 34.


```
# End of Unapproved Drug Dataset Statistical Analysis



# Complete Dataset - Approved + Unapproved Drugs Statistical Analysis

# Statistical Analysis - K-cluster Mean Analysis - Complete Drug Dataset

# Statistical Analysis - K-Means Cluster Analysis - Refining Complete Drug Data Dataset
```{r}
Total_Drug_Data_Start <- Total_Drug_Data[, sapply(Total_Drug_Data, is.numeric)]
Total_Drug_Data_no_Yr_2 <- Total_Drug_Data_Start[, -5]
Total_DD <- Total_Drug_Data_no_Yr_2[, -6]
Total_DD <- Total_DD[Total_DD$Mkt_Cap != 0, ]
Total_DD <- Total_DD[Total_DD$RD_Spend != 0 & Total_DD$RD_Spend != -150, ]
Total_DD <- Total_DD[Total_DD$PE > 0 & Total_DD$EV_Sales > 0, ]
Total_DD <- na.omit(Total_DD)
```

Statistical Analysis - K-Means Clusters - Log Transforming Market Capitalization & R&D Expenditure
```{r}
Log_Total_DD <- data.frame(EV_Sales = Total_DD$EV_Sales)
Log_Total_DD$FDA_Approvals <- Total_DD$FDA_Approvals
Log_Total_DD$Success <- Total_DD$Success
Log_Total_DD$Mkt_Cap_Log <- log(Total_DD$Mkt_Cap)
Log_Total_DD$RD_Spend <- log(Total_DD$RD_Spend)
Log_Total_DD$PE <- Total_DD$PE
Log_Total_DD <- na.omit(Log_Total_DD)
Log_Total_DD
```

Creating R&D Spend & Forward EV/Sales Dataframe - Total Drug Data Dataset
```{r}
TRD_EVs_df <- data.frame(RD_Spend = Log_Total_DD$RD_Spend)
TRD_EVs_df$EV_Sales <- Log_Total_DD$EV_Sales
TRD_EVs_df <- na.omit(TRD_EVs_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - R&D Expenditure & Forward Enterprise Value / Sales Multiple**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(TRD_EVs_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Total Drug Dataset - Research & Development Expenditure & Forward Enterprise Value / Sales Multiple
```{r}
TRD_EVs_df_fit <- kmeans(TRD_EVs_df, 4)

plot(TRD_EVs_df[, 1], TRD_EVs_df[, 2], col = TRD_EVs_df_fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Forward Enterprise Value / Sales Multiples",
     xlim = c(4, 17), ylim = c(0, 40))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 35.

Creating R&D Spend & Forward Price/Earnings Dataframe - Total Drug Data Dataset
```{r}
TRD_PE_df <- data.frame(RD_Spend = Log_Total_DD$RD_Spend)
TRD_PE_df$PE <- Log_Total_DD$PE
TRD_PE_df <- na.omit(TRD_PE_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - R&D Expenditure & Forward Price / Earnings Multiple**
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(TRD_PE_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal
```

Data Visualization - K-cluster Mean Analysis - Total Drug Dataset - Research & Development Expenditure & Forward Price / Earnings Multiples
```{r}
TRD_PE_df_fit <- kmeans(TRD_PE_df, 4)

plot(TRD_PE_df[, 1], TRD_PE_df[, 2], col = TRD_PE_df_fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Forward Price / Earnings Multiples",
     xlim = c(4, 17), ylim = c(0, 250))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 36.

Creating R&D Spend & Market Capitalization Dataframe - Total Drug Dataset
```{r}
TRD_Mktcap_df <- data.frame(RD_Spend = Log_Total_DD$RD_Spend)
TRD_Mktcap_df$Mkt_Cap <- Log_Total_DD$Mkt_Cap_Log
TRD_Mktcap_df <- na.omit(TRD_Mktcap_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - R&D Expenditure & Market Capitalization
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(TRD_Mktcap_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 2 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Research & Development Expenditure & Market Capitalization - Total Drug Dataset
```{r}
TRD_Mktcap_df_fit <- kmeans(TRD_Mktcap_df, 3)

plot(TRD_Mktcap_df[, 1], TRD_Mktcap_df[, 2], col = TRD_Mktcap_df_fit$cluster,
     xlab = "Research & Development Expenditure", ylab = "Market Capitalization",
     xlim = c(4, 18), ylim = c(7, 20))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 37.

Creating P/E & EV/Sales Dataframe - Total Drug Dataset
```{r}
TPE_EVs_df <- data.frame(EV_Sales = Log_Total_DD$EV_Sales)
TPE_EVs_df$PE <- Log_Total_DD$PE
TPE_EVs_df <- na.omit(TPE_EVs_df)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - P/E & EV/Sales
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(TPE_EVs_df, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3 - 4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Forward Price / Earnings & Enterprise Value / Sales Multiples - Total Drug Dataset
```{r}
TPE_EVs_df_fit <- kmeans(TPE_EVs_df, 4)

plot(TPE_EVs_df[, 1], TPE_EVs_df[, 2], col = TPE_EVs_df_fit$cluster,
     xlab = "Forward Enterprise Value / Sales Multiples", ylab = "Forward Price / Earnings Multiples",
     xlim = c(0, 40), ylim = c(0, 250))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 38.

Creating Mkt Cap & EV/Sales Dataframe - Total Drug Dataset
```{r}
TEVS_Mktcap <- data.frame(EV_Sales = Log_Total_DD$EV_Sales)
TEVS_Mktcap$Mkt_Cap <- Log_Total_DD$Mkt_Cap_Log
TEVS_Mktcap <- na.omit(TEVS_Mktcap)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Mkt Cap & EV/Sales - Total Drug Dataset
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(TEVS_Mktcap, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3-4 clusters ideal

Data Visualization - K-cluster Mean Analysis - EV/Sales & Mkt. Cap
```{r}
TEVS_Mktcap_fit <- kmeans(TEVS_Mktcap, 4)

plot(TEVS_Mktcap[, 2], TEVS_Mktcap[, 1], col = TEVS_Mktcap_fit$cluster,
     xlab = "Market Capitalization", ylab = "Forward Enterprise Value / Sales Multiples",
     xlim = c(0, 25), ylim = c(0, 100))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```
Figure 39.

Creating Mkt Cap & P/E Dataframe - Total Drug Dataset
```{r}
TPE_Mktcap <- data.frame(PE = Log_Total_DD$PE)
TPE_Mktcap$Mkt_Cap <- Log_Total_DD$Mkt_Cap_Log
TPE_Mktcap <- na.omit(TPE_Mktcap)
```

Statistical Analysis - K-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Mkt Cap & P/E - Total Drug Dataset
```{r}
# Vector of WSS Values for Different # of clusters
wss <- vector("numeric", length = 10) # set the maximum number of clusters you want to consider
for (i in 1:10) {
  k <- kmeans(TPE_Mktcap, centers = i, nstart = 10)
  wss[i] <- k$tot.withinss
}

# Plotting  WSS values Against # of Clusters
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
Note: 3-4 clusters ideal

Data Visualization - K-cluster Mean Analysis - Forward Price / Earnings Multiples & Mkt. Cap - Total Drug Dataset
```{r}
TPE_Mktcap_fit <- kmeans(TPE_Mktcap, 4)

plot(TPE_Mktcap[, 2], TPE_Mktcap[, 1], col = TPE_Mktcap_fit$cluster,
     xlab = "Market Capitalization", ylab = "Forward Price / Earnings Multiples",
     xlim = c(0, 25), ylim = c(0,550))

points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```

Figure 40.

End of Total Drug Dataset K-cluster Mean Analysis

Statistical Analysis - Boxplots

Approved Drug Dataset

Statistical Analysis - Refining Cumulative FDA Approvals and Size Factor Dataframe
```{r}
FDA_SF_df <- data.frame(Size_Factor = Aveage_Approved_Drug_Data$Size_Factor)
FDA_SF_df$FDA <- Aveage_Approved_Drug_Data$FDA_Approvals
FDA_SF_df <- FDA_SF_df[FDA_SF_df$Size_Factor != 0 & FDA_SF_df$Size_Factor != -150, ]
FDA_SF_df <- na.omit(FDA_SF_df)
```

Cumulative FDA Approvals vs. Size Factor
```{r}
FigFDASF <- ggplot(data=FDA_SF_df, aes(x=factor(Size_Factor), y=FDA)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Cumulative Number of FDA Approvals") +
        theme_classic()
FigFDASF
```
Figure 41

R&D Efficiency vs. Size Factor
```{r}
Fig11 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=RD_efficiency)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Efficiencey of Research & Development Expenditure") +
        theme_classic()
Fig11
```
Figure 1. The effect of research and development expenditure efficiency on forward Enterprise Value / Sales (*F*~(1,63)~ = 2.448; *P* < 0.1227). The regression line is explained by Y = 4.722X - 1.751e-07. The Adjusted R-Squared is 0.022.

R&D Spend vs. Size Factor
```{r}
Fig16 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=RD_Spend)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Research & Development Expenditure") +
        theme_classic()
Fig16
```
Figure 9. The effect of research and development expenditure on Size Factor (*F*~(1,63)~ = 37.172; *P* < 7.257e-08). The regression line is explained by Y = 2.273e+00X + 1.246e-07. The Adjusted R-Squared is 0.3611.

Enterprise Value / Sales vs. Size Factor
```{r}
Fig21 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=EV_Sales)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Enterprise Value / Sales") +
        theme_classic()
Fig21
```
Figure 12. The effect of Forward Enterprise Value / Sales Multiples on Size Factor (*F*~(1,63)~ = 2.4088 ; *P* < 0.1257). The regression line is explained by Y = 2.50737x + 0.04791. The Adjusted R-Squared is 0.02154.

Forward Price / Earnings vs. Size Factor
```{r}
Fig20 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=PE)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Forward Price / Earnings Multiples") +
        theme_classic()
Fig20
```
Figure 13. The effect of size factor on forward price / earnings multiples (*F*~(1,63)~ = 0.891; *P* < 0.3488). The regression line is explained by Y = 101183036X + 3998133. The Adjusted R-Squared is 0.3488.

R&D Efficiency vs. Size Factor
```{r}
Fig11 <- ggplot(data=lag_test_0, aes(x=factor(Size_Factor), y=RD_efficiency)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Efficiencey of Research & Development Expenditure") +
        theme_classic()
Fig11
```


Statistical Analysis - Boxplots

Unapproved Drug Dataset

Refining Unapproved Drug Dataset Dataframe
```{r}
Unapproved_Drug_Data_Start_2 <- Unapproved_Drug_Data[, sapply(Unapproved_Drug_Data, is.numeric)]
Unapproved_Drug_Data_no_Yr_3 <- Unapproved_Drug_Data_Start_2[, -5]
Unapproved_DD_BP <- Unapproved_Drug_Data_no_Yr_3[Unapproved_Drug_Data_no_Yr_3$Mkt_Cap != 0, ]
Unapproved_DD_BP <- Unapproved_DD_BP[Unapproved_DD_BP$RD_Spend != 0 & Unapproved_DD_BP$RD_Spend != -150, ]
Unapproved_DD_BP$Log_RD_Spend <- log(Unapproved_DD_BP$RD_Spend)
Unapproved_DD_BP$Log_EV_Sales <- log(Unapproved_DD_BP$EV_Sales)
Unapproved_DD_BP <- Unapproved_DD_BP[Unapproved_DD_BP$Log_EV_Sales != 0 & Unapproved_DD_BP$Log_EV_Sales > 0, ]
Unapproved_DD_BP$Log_PE <- log(Unapproved_DD_BP$PE)
Unapproved_DD_BP <- Unapproved_DD_BP[Unapproved_DD_BP$PE > 0 & Unapproved_DD_BP$EV_Sales > 0, ]
Unapproved_DD_BP <- Unapproved_DD_BP[Unapproved_DD_BP$Log_PE > 0 & Unapproved_DD_BP$Log_PE > 0, ]
Unapproved_DD_BP <- Unapproved_DD_BP[Unapproved_DD_BP$Log_PE > 0 & Unapproved_DD_BP$Log_PE > 0, ]
Unapproved_DD_BP$Log_Mkt_Cap <- log(Unapproved_DD_BP$Mkt_Cap)
Unapproved_DD_BP <- Unapproved_DD_BP[Unapproved_DD_BP$Log_Mkt_Cap != 0 & Unapproved_DD_BP$Log_Mkt_Cap > 0, ]
Unapproved_DD_BP <- na.omit(Unapproved_DD_BP)
```


Logged R&D Spend vs. Size Factor - Unapproved Drug Dataset
```{r}
RDSSF_model <- lm(Unapproved_DD_BP$Log_RD_Spend ~ factor(Unapproved_DD_BP$Size_Factor))
summary(RDSSF_model)
```
Note: Adjusted R-squared: 0.32, p-value: 3.503e-06, F-statistic: 15.6, DF of 60

Logged R&D Spend vs. Size Factor - Unapproved Drug Dataset
```{r}
Fig22 <- ggplot(data=Unapproved_DD_BP, aes(x=factor(Size_Factor), y=Log_RD_Spend)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Logged Research & Development Expenditure") +
        theme_classic()
Fig22
```
Figure 41.The effect of market capitalization category (size factor) on research and development expenditure (*F*~(1,63)~ = 41.07; *P* = 3.84e-16). The Adjusted R-Squared is 0.242

EV/Sales vs. Size Factor - Unapproved Drug Dataset
```{r}
RDSEVs_model <- lm(Unapproved_DD_BP$Log_EV_Sales ~ factor(Unapproved_DD_BP$Size_Factor))
summary(RDSEVs_model)
```

```

**Unapproved Drug Dataset - Enterprise Value / Sales vs. Size Factor**
```{r}
Fig23 <- ggplot(data=Unapproved_DD_BP, aes(x=factor(Size_Factor), y=Log_EV_Sales)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Logged Enterprise Value / Sales") +
        theme_classic()
Fig23
```
Figure 42.The effect of market capitalization category (size factor) for unapproved manufacturers on forward enterprise value / sales (*F*~(1,249)~ = 1.713; *P* = 0.1824). The Adjusted R-Squared is 0.005.

P/E vs. Size Factor - Unapproved Drug Dataset
```{r}
PESF_model <- lm(Unapproved_DD_BP$Log_PE ~ factor(Unapproved_DD_BP$Size_Factor))
summary(PESF_model)
```

Logged Forward Price / Earnings vs. Size Factor
```{r}
Fig24 <- ggplot(data=Unapproved_DD_BP, aes(x=factor(Size_Factor), y=Log_PE)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Logged Forward Price / Earnings Multiples") +
        theme_classic()
Fig24
```
Figure 43.The effect of market capitalization category (size factor) for unapproved manufacturers on forward price / earnings multiples (*F*~(1,249)~ = 0.7904; *P* = 0.4548). The Adjusted R-Squared is -0.0016.

Statistical Analysis - Boxplots

Total Drug Dataset

Refining Total Drug Dataset Dataframe
```{r}
Total_Drug_Data_Start_2 <- Total_Drug_Data[, sapply(Total_Drug_Data, is.numeric)]
Total_Drug_Data_no_Yr_3 <- Total_Drug_Data_Start_2[, -5]
Total_DD_BP <- Total_Drug_Data_no_Yr_3[Total_Drug_Data_no_Yr_3$Mkt_Cap != 0, ]
Total_DD_BP <- Total_DD_BP[Total_DD_BP$RD_Spend != 0 & Total_DD_BP$RD_Spend != -150, ]
Total_DD_BP$Log_RD_Spend <- log(Total_DD_BP$RD_Spend)
Total_DD_BP$Log_EV_Sales <- log(Total_DD_BP$EV_Sales)
Total_DD_BP <- Total_DD_BP[Total_DD_BP$Log_EV_Sales != 0 & Total_DD_BP$Log_EV_Sales > 0, ]
Total_DD_BP$Log_PE <- log(Total_DD_BP$PE)
Total_DD_BP <- Total_DD_BP[Total_DD_BP$PE > 0 & Total_DD_BP$EV_Sales > 0, ]
Total_DD_BP <- Total_DD_BP[Total_DD_BP$Log_PE > 0 & Total_DD_BP$Log_PE > 0, ]
Total_DD_BP$Log_Mkt_Cap <- log(Total_DD_BP$Mkt_Cap)
Total_DD_BP <- Total_DD_BP[Total_DD_BP$Log_Mkt_Cap != 0 & Total_DD_BP$Log_Mkt_Cap > 0, ]
Total_DD_BP <- na.omit(Total_DD_BP)
```

Logged R&D Spend vs. Size Factor - Total Drug Dataset
```{r}
TRDSF_model <- lm(Total_DD_BP$Log_PE ~ factor(Total_DD_BP$Size_Factor))
summary(TRDSF_model)
```

Logged R&D Spend vs. Size Factor - Total Drug Dataset
```{r}
Fig25 <- ggplot(data=Total_DD_BP, aes(x=factor(Size_Factor), y=Log_RD_Spend)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Logged Research & Development Expenditure") +
        theme_classic()
Fig25
```
Figure 44.The effect of market capitalization category (size factor) for all manufacturers on research and development expenditure (*F*~(2,580)~ = 1.702; *P* = 0.1833). The Adjusted R-Squared is 0.002406.

Logged EV_Sales vs. Size Factor - Total Drug Dataset
```{r}
TEVsSF_model <- lm(Total_DD_BP$EV_Sales ~ factor(Total_DD_BP$Size_Factor))
summary(TEVsSF_model)
```

Total Drug Dataset - Enterprise Value / Sales vs. Size Factor
```{r}
Fig26 <- ggplot(data=Total_DD_BP, aes(x=factor(Size_Factor), y=Log_EV_Sales)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Logged Enterprise Value / Sales") +
        theme_classic()
Fig26
```
Figure 45.The effect of market capitalization category (size factor) for all manufacturers on forward enterprise value / sales multiples (*F*~(2,580)~ = 5.73; *P* = 0.0034). The Adjusted R-Squared is 0.016.

Logged PE vs. Size Factor - Total Drug Dataset
```{r}
TPESF_model <- lm(Total_DD_BP$Log_PE ~ factor(Total_DD_BP$Size_Factor))
summary(TPESF_model)
```

Logged Forward Price / Earnings vs. Size Factor
```{r}
Fig27 <- ggplot(data=Total_DD_BP, aes(x=factor(Size_Factor), y=Log_PE)) +
        geom_boxplot() +
        xlab("Market Capitalization Category") +
        ylab("Logged Forward Price / Earnings Multiples") +
        theme_classic()
Fig27
```
Figure 46.The effect of market capitalization category (size factor) for all manufacturers on forward enterprise value / sales multiples (*F*~(2,580)~ = 1.702; *P* = 0.1833). The Adjusted R-Squared is 0.0024.

Statistical Analysis - Histograms

Approved Drug Dataset

R&D Efficiency Distribution
```{r}
lag_test_0$RD_efficiency_log <- log(lag_test_0$RD_efficiency)

hist(lag_test_0$RD_efficiency_log, xlab = "Research and Development Expenditure Efficiency", main = "Approved Firms")
```
Figure 47. Distribution of research and development expenditure efficiency for approved firms

R&D Expenditure Distribution
```{r}
lag_test_0$RD_Spend_log <- log(lag_test_0$RD_Spend)

hist(lag_test_0$RD_Spend_log, xlab = "Research and Development Expenditure", main = "Approved Firms")
```
Figure 48. Distribution of research and development expenditure for approved firms

EV/Sales Multiples Distribution
```{r}
lag_test_0$EV_Sales_log <- log(lag_test_0$EV_Sales)
```
```

```{r}
hist(lag_test_0$EV_Sales_log, xlab = "Forward Enterprise Value / Sales Multiples", main = "Approved Firms")
```
Figure 49. Distribution of Forward Enterprise Value / Sales Multiples for approved firms

P/E Multiples Distribution
```{r}
lag_test_0$PE_log <- log(lag_test_0$PE)
hist(lag_test_0$PE_log, xlab = "Forward Price / Earnings Multiples", main = "Approved Firms")
```
Figure 50. Distribution of Forward Price / Earnings Multiples for approved firms

Market Capitalization Distribution
```{r}
lag_test_0$Mkt_Cap_log <- log(lag_test_0$Mkt_Cap)
hist(lag_test_0$Mkt_Cap_log, xlab = "Market Capitalization", main = "Approved Firms")
```
Figure 51. Distribution of Market Capitalization for approved firms

Statistical Analysis - Histograms

Unapproved Drug Dataset

R&D Expenditure Distribution
```{r}
hist(Unapproved_DD_BP$Log_RD_Spend, xlab = "Research and Development Expenditure", main = "Unapproved Firms")
```
Figure 52. Distribution of Research and Development Expenditure for unapproved firms

EV/Sales Multiples Distribution
```{r}
hist(Unapproved_DD_BP$Log_EV_Sales, xlab = "Forward Enterprise Value / Sales Multiples", main = "Unapproved Firms")
```
Figure 53. Distribution of Forward Enterprise Value / Sales Multiples for unapproved firms

P/E Multiples Distribution
```{r}
hist(Unapproved_DD_BP$Log_PE, xlab = "Forward Price / Earnings Multiples", main = "Unapproved Firms")
```
Figure 54. Distribution of Forward Price / Earnings Multiples for unapproved firms

Market Capitalization Distribution
```{r}
hist(Unapproved_DD_BP$Log_Mkt_Cap, xlab = "Market Capitalization", main = "Unapproved Firms")
```
Figure 55. Distribution of Market Capitalization for unapproved firms

Statistical Analysis - Histograms

Total Drug Dataset

R&D Expenditure Distribution
```{r}
hist(Total_DD_BP$Log_RD_Spend, xlab = "Research and Development Expenditure", main = "All Firms")
```
Figure 56. Distribution of Research and Development Expenditure for all firms

EV/Sales Multiples Distribution
```{r}
hist(Total_DD_BP$Log_EV_Sales, xlab = "Forward Enterprise Value / Sales Multiples", main = "All Firms")
```
Figure 57. Distribution of Forward Enterprise Value / Sales Multiples for all firms

P/E Multiples Distribution
```{r}
hist(Total_DD_BP$Log_PE, xlab = "Forward Price / Earnings Multiples", main = "All Firms")
```
Figure 58. Distribution of Forward Price / Earnings Multiples for all firms

Market Capitalization Distribution
```{r}
hist(Total_DD_BP$Log_Mkt_Cap, xlab = "Market Capitalization", main = "All Firms")
```
Figure 59. Distribution of Market Capitalization for all firms

Additional Statistical Analysis

Statistical Analysis - Experience on Forward Multiples and Other Metrics

PE vs. Experience - Approved Drug Dataset
```{r}
ExPE_model <- lm(lag_test_0$PE ~ factor(lag_test_0$Experience))
summary(ExPE_model)
```

Forward P/E Multiples vs. Experience
```{r}
Fig28 <- ggplot(data=lag_test_0, aes(x=factor(Experience), y=PE)) +
        geom_boxplot() +
        xlab("Experience") +
        ylab("Forward Price / Earnings Multiples") +
        theme_classic()
Fig28
```
Figure 60. The effect of experience for approved manufacturers on forward price / earnings multiple (*F*~(1,63)~ = 1.646; *P* = 0.2043). The Adjusted R-Squared is 0.0099.

EV_Sales vs. Experience - Approved Drug Dataset
```{r}
ExEVs_model <- lm(lag_test_0$EV_Sales ~ factor(lag_test_0$Experience))
summary(ExEVs_model)
```

Forward EV/Sales Multiples vs. Experience
```{r}
Fig29 <- ggplot(data=lag_test_0, aes(x=factor(Experience), y=EV_Sales)) +
        geom_boxplot() +
        xlab("Experience") +
        ylab("Forward Enterprise Value / Sales Multiples") +
        theme_classic()
Fig29
```
Figure 61. The effect of experience for approved manufacturers on forward Enterprise Value / Sales multiple (*F*~(1,63)~ = 7.38; *P* = 0.0085). The Adjusted R-Squared is 0.0906.

RD_Efficiency vs. Experience - Approved Drug Dataset
```{r}
ExRDe_model <- lm(lag_test_0$RD_efficiency ~ factor(lag_test_0$Experience))
summary(ExRDe_model)
```

Research & Development Efficiency vs. Experience
```{r}
Fig30 <- ggplot(data=lag_test_0, aes(x=factor(Experience), y=RD_Efficiency)) +
        geom_boxplot() +
        xlab("Experience") +
        ylab("Research & Development Efficiency") +
        theme_classic()
Fig30
```
Figure 62. The effect of experience for approved manufacturers on research and development expenditure efficiency (*F*~(1,63)~ = 3.576; *P* = 0.06324). The Adjusted R-Squared is 0.03869.

RD_Spend vs. Experience - Approved Drug Dataset
```{r}
ExRDs_model <- lm(lag_test_0$RD_Spend ~ factor(lag_test_0$Experience))
summary(ExRDs_model)
```

Research & Development Expenditure vs. Experience
```{r}
Fig31 <- ggplot(data=lag_test_0, aes(x=factor(Experience), y=RD_Spend)) +
        geom_boxplot() +
        xlab("Experience") +
        ylab("Research & Development Expenditure") +
        theme_classic()
Fig31
```

Figure 63. The effect of experience for approved manufacturers on research and development expenditure efficiency (*F*~(1,63)~ = 5.759; *P* = 0.01938). The Adjusted R-Squared is 0.06921.

Mkt_Cap vs. Experience - Approved Drug Dataset
```{r}
ExMktcap_model <- lm(lag_test_0$Mkt_Cap ~ factor(lag_test_0$Experience))
summary(ExMktcap_model)
```

Market Capitalization vs. Experience
```{r}
Fig32 <- ggplot(data=lag_test_0, aes(x=factor(Experience), y=Mkt_Cap)) +
        geom_boxplot() +
        xlab("Experience") +
        ylab("Market Capitalization") +
        theme_classic()
Fig32
```
Figure 64. The effect of experience for approved manufacturers on research and development expenditure efficiency (*F*~(1,63)~ = 1.068; *P* = 0.3053). The Adjusted R-Squared is 0.0010.

FDA Approvals vs. Experience - FDA Approvals and Experience Drug Dataset
```{r}
FDA_Approvals <- read.csv("C:/Users/nicho/Desktop/SPROJ/Data/R Studio/Datasets/FDA Approvals and Experience.csv")
```

FDA Approvals vs. Experience - FDA Approvals and Experience Drug Dataset
```{r}
Experience_Test <- lm(FDA_Approvals$Approvals ~ factor(FDA_Approvals$Experience))
summary(Experience_Test)
```

FDA Approvals vs. Experience - FDA Approvals and Experience Drug Dataset
```{r}
Fig33 <- ggplot(data=Experience_Test, aes(x=factor(FDA_Approvals$Experience), y=FDA_Approvals$Approvals)) +
        geom_boxplot() +
        xlab("Experience") +
        ylab("Cumulative FDA Approvals") +
        theme_classic()
Fig33
```
Figure 65. The effect of experience on cumulative FDA approvals (*F*~(1,423)~ = 366.9 ; *P* = 2.2e-16). The Adjusted R-Squared is 0.4632.

End of Additional Analysis

Bibliography

Arthur Daemmrich. “Pharmaceutical Manufacturing in America: A Brief History.” *Pharmacy in History*, vol. 59, no. 3, 2017, p. 63, 10.26506/pharmhist.59.3.0063. Accessed 29 Sept. 2019.

Beinse, Guillaume, et al. “Prediction of Drug Approval after Phase I Clinical Trials in Oncology: RESOLVED2.” *JCO Clinical Cancer Informatics*, no. 3, Dec. 2019, pp. 1–10, https://doi.org/10.1200/cci.19.00023. Accessed 2 April 2023.

Buckley, Kevin, and Alan G Ryder. “Applications of Raman Spectroscopy in Biopharmaceutical Manufacturing: A Short Review.” *Applied Spectroscopy*, vol. 71, no. 6, 2017, pp. 1085–1116, www.ncbi.nlm.nih.gov/pubmed/28534676, 10.1177/0003702817703270. Accessed 24 Oct. 2019.

Campbell, John. “Understanding Pharma: The Professional’s Guide to How Pharmaceutical and Biotech Companies Really Work.” New Jersey: Syneos Health, 2014. Third Edition.

Chen, Long, et al. “What Drives Stock Price Movements?” *The Review of Financial Studies*, vol. 26, no. 4, 2013, pp. 841–876, www.jstor.org/stable/pdf/23355383.pdf?refreqid=excelsior%3Ab3426aec2b3e53a6da11739c7d704da2&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&origin=&initiator=. Accessed 2 April 2023.

Conroy, Mary Schaeffer. “Russian-American Pharmaceutical Relations, 1900-1945.” *Pharmacy in History*, vol. 46, no. 4, 2004, pp. 143–166, www.jstor.org/stable/pdf/41112230.pdf?refreqid=fastly-default%3A858f54b6f662ce779bdc56e6f81c946a&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&origin=&initiator=search-results. Accessed 2 April 2023.

Crama, Pascale, et al. “Milestone Payments or Royalties? Contract Design for R&D Licensing.” *Operations Research*, vol. 56, no. 6, 2008, pp. 1539–1552, www.jstor.org/stable/pdf/25580905.pdf?refreqid=fastly-default%3A5566632bc07316f79404af7e9ddc36ad&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&origin=&initiator=search-results. Accessed 2 April 2023.

Danzon, Patricia M., et al. “Productivity in Pharmaceutical–Biotechnology R&D: The Role of Experience and Alliances.” *Journal of Health Economics*, vol. 24, no. 2, Mar. 2005, pp. 317–339, https://doi.org/10.1016/j.jhealeco.2004.09.006. Accessed 27 Apr. 2020.

DiMasi, JA, et al. “A Tool for Predicting Regulatory Approval after Phase II Testing of New Oncology Compounds.” *Clinical Pharmacology & Therapeutics*, vol. 98, no. 5, 24 Sept. 2015, pp. 506–513, https://doi.org/10.1002/cpt.194. Accessed 10 Apr. 2023.

Galambos, Louis, and Jeffrey L. Sturchio. “Pharmaceutical Firms and the Transition to Biotechnology: A Study in Strategic Innovation.” *Business History Review*, vol. 72, no. 2, 1998, pp. 250–278, 10.2307/3116278.

Goffin, John, et al. “Objective Responses in Patients with Malignant Melanoma or Renal Cell Cancer in Early Clinical Studies Do Not Predict Regulatory Approval.” *Clinical Cancer Research*, vol. 11, no. 16, 15 Aug. 2005, pp. 5928–5934, https://doi.org/10.1158/1078-0432.ccr-05-0130. Accessed 11 Oct. 2022.

Halim, Abdel B. “Chapter 1 - Pharmaceutical Crisis.” *ScienceDirect*, Academic Press, 1 Jan. 2019, www.sciencedirect.com/science/article/abs/pii/B9780128161210000015. Accessed 2 April 2023.

Hall, Jeremy, et al. “The Paradox of Sustainable Innovation: The “Eroom” Effect (Moore’s Law Backwards).” *Journal of Cleaner Production*, vol. 172, Jan. 2018, pp. 3487–3497, https://doi.org/10.1016/j.jclepro.2017.07.162. Accessed 5 Mar. 2020.

Hamilton, James D. *Time Series Analysis*. *JSTOR*, Princeton University Press, 1994, www.jstor.org/stable/pdf/j.ctv14jx6sm.11.pdf?refreqid=excelsior%3A57865c02f8f5a5849b0f287f20648b1a&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&origin=&initiator=&acceptTC=1. Accessed 2 April 2023.

Helland, Inge S. “On the Interpretation and Use of R 2 in Regression Analysis.” *Biometrics*, vol. 43, no. 1, Mar. 1987, p. 61, https://doi.org/10.2307/2531949. Accessed 30 Oct. 2021.

Hoffmann, John. “Chapter Title: Review of Linear Regression Models Book Title: Regression Models for Categorical, Count, and Related Variables Book Subtitle: An Applied Approach.” *University of California Press*, https://doi.org/10.1525/j.ctv1wxrfr.5. Accessed 2 April 2023.

Hu, Haize, et al. “An Effective and Adaptable K-Means Algorithm for Cluster Analysis.” *Pattern Recognition*, Feb. 2023, p. 109404, https://doi.org/10.1016/j.patcog.2023.109404. Accessed 20 Feb. 2023.

Hung, H. M. James, et al. “The Behavior of the P-Value When the Alternative Hypothesis Is True.” *Biometrics*, vol. 53, no. 1, Mar. 1997, p. 11, https://doi.org/10.2307/2533093. Accessed 22 June 2020.

Ikotun, Abiodun M., et al. “K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data.” *Information Sciences*, Dec. 2022, https://doi.org/10.1016/j.ins.2022.11.139. Accessed 7 Dec. 2022.

Khanal, Ohnmar, and Abraham M. Lenhoff. “Developments and Opportunities in Continuous Biopharmaceutical Manufacturing.” *MAbs*, vol. 13, no. 1, 1 Jan. 2021, p. 1903664, 10.1080/19420862.2021.1903664.

Lo, Andrew W., et al. “Machine Learning with Statistical Imputation for Predicting Drug Approvals.” *Harvard Data Science Review*, vol. 1, no. 1, 23 June 2019, hdsr.mitpress.mit.edu/pub/ct67j043, 10.1162/99608f92.5c5f0525. Accessed 19 Feb. 2020.

McCAREY, MICHAEL C. “Generic Substitution Policy.” *Food, Drug, Cosmetic Law Journal*, vol. 34, no. 2, 1979, pp. 103–107, www.jstor.org/stable/pdf/26658162.pdf?refreqid=fastly-default%3Acebf0c624dbc3adca20fad15eac63d0e&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&origin=&initiator=search-results. Accessed 2 April 2023.

Mejía-Peláez, Felipe, and Ignacio Vélez-Pareja. “Analytical Solution to the Circularity Problem in the Discounted Cash Flow Valuation Framework.” *Innovar: Revista de Ciencias Administrativas Y Sociales*, vol. 21, no. 42, 2011, pp. 55–68, www.jstor.org/stable/pdf/23744040.pdf?refreqid=excelsior%3A48a750908acdd1cb86b2bb6e68f9e45b&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&origin=&initiator=. Accessed 2 April 2023.

Miller, Susan, et al. “Chapter 2 - Backgrounder—Part 2: Drug Discovery: Research, Discovery, and Development—Art and Science.” *ScienceDirect*, Woodhead Publishing, 1 Jan. 2023, www.sciencedirect.com/science/article/abs/pii/B9780128243046000109. Accessed 2 April 2023.

---. “Chapter 2 - Backgrounder—Part 2: Drug Discovery: Research, Discovery, and Development—Art and Science.” *ScienceDirect*, Woodhead Publishing, 1 Jan. 2023, www.sciencedirect.com/science/article/abs/pii/B9780128243046000109. Accessed 2 April 2023.

Montinari, Maria Rosa, et al. “The First 3500 years of Aspirin History from Its Roots – a Concise Summary.” *Vascular Pharmacology*, vol. 113, Feb. 2019, pp. 1–8, 10.1016/j.vph.2018.10.008. Accessed 24 July 2019.

Pindyck, Robert S, and Daniel L Rubinfeld. *Microeconomics*. Pearson, 2001.

Scannell, Jack W., et al. “Diagnosing the Decline in Pharmaceutical R&D Efficiency.” *Nature Reviews Drug Discovery*, vol. 11, no. 3, Mar. 2012, pp. 191–200, www.nature.com/articles/nrd3681, https://doi.org/10.1038/nrd3681.

Shukla, Abhinav A., et al. “Evolving Trends in MAb Production Processes.” *Bioengineering & Translational Medicine*, vol. 2, no. 1, Mar. 2017, pp. 58–69, www.onlinelibrary.wiley.com/doi/10.1002/btm2.10061, 10.1002/btm2.10061. Accessed 11 Oct. 2019.

Skandrani, Hamida, and Malek Sghaier. “The Dark Side of the Pharmaceutical Industry.” *Marketing Intelligence & Planning*, vol. 34, no. 7, 3 Oct. 2016, pp. 905–926, https://doi.org/10.1108/mip-06-2015-0123.

Slingerland, Annabelle S., et al. “Then and Now: Hypes and Hopes of Regenerative Medicine.” *Trends in Biotechnology*, vol. 31, no. 3, 1 Mar. 2013, pp. 121–123, pubmed.ncbi.nlm.nih.gov/23280408/, 10.1016/j.tibtech.2012.12.001. Accessed 27 Feb. 2022.

Spitz, Janet, and Mark Wickham. “Pharmaceutical High Profits: The Value of R&D, or Oligopolistic Rents?” *American Journal of Economics and Sociology*, vol. 71, no. 1, 2012, pp. 1–36, www.jstor.org/stable/pdf/23245176.pdf?refreqid=fastly-default%3A5bf64fe67f03d6490200dc4714251b32&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&origin=&initiator=search-results&acceptTC=1. Accessed 2 April 2023.

Temin, Peter. “Technology, Regulation, and Market Structure in the Modern Pharmaceutical Industry.” *The Bell Journal of Economics*, vol. 10, no. 2, 1979, p. 429, https://doi.org/10.2307/3003345. Accessed 4 June 2020.

Van Norman, Gail A. “Overcoming the Declining Trends in Innovation and Investment in Cardiovascular Therapeutics.” *JACC: Basic to Translational Science*, vol. 2, no. 5, Oct. 2017, pp. 613–625, https://doi.org/10.1016/j.jacbts.2017.09.002. Accessed 22 Nov. 2019.

Wang, Yifei, and Shengyong Yang. “Multispecific Drugs: The Fourth Wave of Biopharmaceutical Innovation.” *Signal Transduction and Targeted Therapy*, vol. 5, no. 1, 4 June 2020, 10.1038/s41392-020-0201-3. Accessed 2 Mar. 2021.

Whayne, Tom F. “The History of Preventive Medicine in World War II.” *Public Health Reports (1896-1970)*, vol. 74, no. 2, 1959, pp. 170–174, www.jstor.org/stable/4590405?seq=1#metadata_info_tab_contents, https://doi.org/10.2307/4590405. Accessed 9 Oct. 2020.

Zarin, Deborah A., et al. “Trial Reporting in ClinicalTrials.gov — the Final Rule.” *New England Journal of Medicine*, vol. 375, no. 20, 17 Nov. 2016, pp. 1998–2004, https://doi.org/10.1056/nejmsr1611785. Accessed 16 Mar. 2022.