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, ns9240@bard.edu

Follow this and additional works at: https://digitalcommons.bard.edu/senproj_s2023

Part of the Finance Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation
https://digitalcommons.bard.edu/senproj_s2023/297

This Open Access is brought to you for free and open access by the Bard Undergraduate Senior Projects at Bard Digital Commons. It has been accepted for inclusion in Senior Projects Spring 2023 by an authorized administrator of Bard Digital Commons. For more information, please contact digitalcommons@bard.edu.
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 ......................................................................................................................... 7  

Literature Review ........................................................................................................ 7  
  History of Therapeutics ............................................................................................. 8  
  Drug Discovery & Development .............................................................................. 12  
  Food and Drug Administration (FDA) Approval Process .................................. 13  

Methodology ............................................................................................................... 15  

Hypothesis Statements ............................................................................................... 15  

Historical Analyses .................................................................................................... 15  
  Multiple Imputation ................................................................................................. 15  
  Experience Theory .................................................................................................. 16  
  Eroom’s Law ........................................................................................................... 16  

Techniques Utilized ..................................................................................................... 17  
  Data Collection and Organization Process .......................................................... 17  
  Single and Multivariable Regressions .................................................................... 19  
  K-Means Cluster Analysis ....................................................................................... 21  
  Boxplots and Histograms ......................................................................................... 23  

Equations ..................................................................................................................... 24  
  Research and Development Expenditure Efficiency ........................................... 24  
  Forward Price to Earnings Multiple ...................................................................... 24  
  Forward Enterprise Value to Sales Multiple ......................................................... 26  

Limitations ................................................................................................................... 27  
  Survivorship Bias .................................................................................................. 27  
  Mergers and Acquisitions ....................................................................................... 28  
  Explicit Research and Development Breakdown ................................................... 28  
  Additional Factors .................................................................................................. 29  

Results ......................................................................................................................... 29  
  Single Variable Regressions ................................................................................... 31  
  Multi Variable Regressions ..................................................................................... 42  
  K-Means Cluster Analysis ....................................................................................... 47  
  Distributions Analysis .............................................................................................. 63  
  Secondary Hypothesis ............................................................................................. 66  

Conclusion .................................................................................................................... 69  

Appendix ...................................................................................................................... 70  

Bibliography ................................................................................................................ 70
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 Sacilycic 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.

![Number of Drugs Approved by the FDA](image)

**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 ($X_i$) has on the value of a dependent variable ($Y_i$). 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).

\[ y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon_i \]  
\[ \text{or use } \beta_0 \text{ for } \alpha \]  

(1)

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

(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 \( Y_i \) based on new values of \( X_i \). The slope of the line (\( \beta_1 \)) represents the change in \( Y_i \) for a one-unit increase in \( X_i \), while the intercept (\( \beta_0 \)) represents the predicted value of \( Y_i \) when \( X_i \) 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:

\[ \text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  

(3)

The SSE (alternatively, SSR) equation squares the differences between each \( Y_i \), 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.
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.

\[
T = \frac{\sqrt{n}y_n}{\sigma}
\]  

(4)

\[
p = 1 - \Phi(t),
\]  

(5)

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 \( R^2 \).

\[
SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (y - 1\bar{y})'(y - 1\bar{y})
\]  

(6)

\[
R^2 = \frac{SSR}{SST}
\]  

(7)

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

Anduze 22
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:

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

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.

\[
\text{Research and Development Efficiency} = \frac{R&D \ \text{Expense}}{FDA \ \text{Approvals}}
\]

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:

\[
P_t = \sum_{k=1}^{T} \frac{FE_{t+k}(1-b_{t+k})}{(1+q_t)^k} + \frac{FE_{t+T+1}}{q_t(1+q_t)^T}
\]

\[
= f(c', q_t)
\]

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, \( I-b_{t+k} \) is the assumed payout ratio, and \( q_t \) 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).

\[
E_{t-1} = \frac{E_t + CFE_t - (Ku_t - Kd_t)D_{t-1} + (Ku_t - \psi_t)T_{t-1}}{1 + Ku_t}
\]  

\[
V_{t-1} = \frac{V_t + FCF_t + TS_t + (Ku_t - \psi_t)T_{t-1}}{1 + Ku_t}
\]

\[
V_{t+k} = \frac{FCF_{t+k}}{(Ku - g \cdot \phi}
\]

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 $K_u$, the cost of debt is $K_d$, $D$ 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.

\[ Net \ Income \ or \ \pi(q) = R(q) - C(q) \]  

(5)

\[ \text{Earnings per Share} = \frac{\text{Net Income}}{\text{Total Shares Outstanding}} \]  

(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.

\[ \text{Price to Earnings Multiple} = \frac{\text{Market Price}}{\text{Earnings per Share}} \]  

(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.

\[ \text{Market Capitalization} = \text{Market Price} \times \text{Total Shares Outstanding} \]  

(8)
This involves simply multiplying the prevailing market price yielded from equation one against the total outstanding shares.

\[
Net\ Debt = Total\ Debt - Total\ Cash \tag{9}
\]

\[
Enterprise\ Value = Net\ Debt + Market\ Capitalization \tag{10}
\]

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).

\[
Revenue = Q \times P \tag{11}
\]

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).

\[
Enterprise\ Value\ to\ Sales\ Multiple = \frac{Enterprise\ Value}{Revenue} \tag{12}
\]

**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.

![Observation Count by Years Lagged](image)

**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, it’s 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
Anduze

They 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.

Anduze 43
Multivariable Linear Regression Results

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>EV_Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD_efficiency</td>
<td>0.0000</td>
</tr>
<tr>
<td>Mkt_Cap</td>
<td>0.00</td>
</tr>
<tr>
<td>factor(Size_Factor)2</td>
<td>2.67</td>
</tr>
<tr>
<td>factor(Size_Factor)3</td>
<td>0.00</td>
</tr>
<tr>
<td>RD_efficiency:Mkt_Cap</td>
<td>0.00</td>
</tr>
<tr>
<td>RD_efficiency:factor(Size_Factor)2</td>
<td>0.0000</td>
</tr>
<tr>
<td>RD_efficiency:factor(Size_Factor)3</td>
<td>0.0000</td>
</tr>
<tr>
<td>Mkt_Cap:factor(Size_Factor)2</td>
<td>0.0000***</td>
</tr>
<tr>
<td>RD_efficiency:Mkt_Cap:factor(Size_Factor)2</td>
<td>-0.00**</td>
</tr>
<tr>
<td>Constant</td>
<td>1.93</td>
</tr>
</tbody>
</table>

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)~ = 3.98; *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 | factor(Size_Factor)2 | factor(Size_Factor)3 |
   | -0.0000*      | 3.71                | 3.57                |
   |  (0.0000)     | (2.39)              | (2.31)              |

   | RD_efficiency:factor(Size_Factor)2 | RD_efficiency:factor(Size_Factor)3 |
   | 0.0000          | 0.0000              |
   |  (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 | Mkt_Cap | RD_efficiency:Mkt_Cap |
   | -0.0000*      | 0.0000***| -0.00                |
   |  (0.0000)     | (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**

<table>
<thead>
<tr>
<th>Dependent variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PE</strong></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>RD_efficiency</strong></td>
</tr>
<tr>
<td><strong>EV_Sales</strong></td>
</tr>
<tr>
<td><strong>factor(Size_Factor)2</strong></td>
</tr>
<tr>
<td><strong>factor(Size_Factor)3</strong></td>
</tr>
<tr>
<td><strong>Mkt_Cap</strong></td>
</tr>
<tr>
<td><strong>RD_efficiency:EV_Sales</strong></td>
</tr>
<tr>
<td><strong>RD_efficiency:factor(Size_Factor)2</strong></td>
</tr>
<tr>
<td><strong>RD_efficiency:factor(Size_Factor)3</strong></td>
</tr>
<tr>
<td><strong>EV_Sales:factor(Size_Factor)2</strong></td>
</tr>
<tr>
<td><strong>RD_efficiency:Mkt_Cap</strong></td>
</tr>
<tr>
<td><strong>EV_Sales:Mkt_Cap</strong></td>
</tr>
<tr>
<td><strong>factor(Size_Factor)2:Mkt_Cap</strong></td>
</tr>
<tr>
<td><strong>RD_efficiency:EV_Sales:factor(Size_Factor)2</strong></td>
</tr>
<tr>
<td><strong>RD_efficiency:EV_Sales:Mkt_Cap</strong></td>
</tr>
<tr>
<td><strong>RD_efficiency:factor(Size_Factor)2:Mkt_Cap</strong></td>
</tr>
<tr>
<td><strong>EV_Sales:factor(Size_Factor)2:Mkt_Cap</strong></td>
</tr>
<tr>
<td><strong>RD_efficiency:EV_Sales:factor(Size_Factor)2:Mkt_Cap</strong></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
</tr>
<tr>
<td><strong>R2</strong></td>
</tr>
<tr>
<td><strong>Adjusted R2</strong></td>
</tr>
<tr>
<td><strong>Residual Std. Error</strong></td>
</tr>
</tbody>
</table>

**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:

<table>
<thead>
<tr>
<th>Term</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD_efficiency</td>
<td>-0.0001</td>
<td></td>
</tr>
<tr>
<td>factor(Size_Factor)2</td>
<td>6.16</td>
<td>(8.58)</td>
</tr>
<tr>
<td>factor(Size_Factor)3</td>
<td>5.31</td>
<td>(8.29)</td>
</tr>
<tr>
<td>RD_efficiency:factor(Size_Factor)2</td>
<td>0.0001</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>RD_efficiency:factor(Size_Factor)3</td>
<td>0.0001</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>Constant</td>
<td>17.28**</td>
<td>(8.05)</td>
</tr>
</tbody>
</table>

Observations: 65
R^2: 0.16
Adjusted R^2: 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:

<table>
<thead>
<tr>
<th>Term</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD_efficiency</td>
<td>-0.0000**</td>
<td>(0.0000)</td>
</tr>
<tr>
<td>Mkt_Cap</td>
<td>0.0000</td>
<td>(0.0000)</td>
</tr>
<tr>
<td>RD_efficiency:Mkt_Cap</td>
<td>0.00</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Constant</td>
<td>20.42***</td>
<td>(2.12)</td>
</tr>
</tbody>
</table>

Observations: 65
R^2: 0.18
Adjusted R^2: 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.
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 22. Forward enterprise value / sales and price / earnings multiples
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?

![Research and development efficiency and forward enterprise value / sales multiples](image)

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

Anduze 55
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

Anduze 58
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 40. R&D distribution

Figure 41. Forward EV/ Sales distribution

Figure 42. Forward P/E distribution

Figure 43. Firm size distribution
Unapproved and Consolidated Drug Manufacturers – Variable Distributions

**Figure 44.** R&D distribution

**Figure 45.** Forward EV/Sales distribution

**Figure 46.** Forward Price/Earnings distribution

**Figure 47.** Firm size distribution

**Figure 48.** R&D distribution

**Figure 49.** Forward EV/Sales 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 Anduze 66.
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.
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.

**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
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 ($F^{*} = 5.73; P^{*} = 0.0034$). 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 ($F^{*} = 1.702; P^{*} = 0.1833$). The Adjusted R-Squared is 0.002.
## Statistical Analysis: 1

```r
# Number of Observations per Lag
lag_0 < nrow(lag_test_0) + nrow(lag_test_10)
lag_1 < nrow(lag_test_0) + nrow(lag_test_11)
lag_2 < nrow(lag_test_0) + nrow(lag_test_12)
lag_3 < nrow(lag_test_0) + nrow(lag_test_13)
lag_4 < nrow(lag_test_0) + nrow(lag_test_14)
lag_5 < nrow(lag_test_0) + nrow(lag_test_15)
lag_6 < nrow(lag_test_0) + nrow(lag_test_16)
lag_7 < nrow(lag_test_0) + nrow(lag_test_17)
lag_8 < nrow(lag_test_0) + nrow(lag_test_18)
lag_9 < nrow(lag_test_0) + nrow(lag_test_19)
lag_10 < nrow(lag_test_0) + nrow(lag_test_20)
lag_11 < nrow(lag_test_0) + nrow(lag_test_21)
lag_12 < nrow(lag_test_0) + nrow(lag_test_22)
lag_13 < nrow(lag_test_0) + nrow(lag_test_23)
```

## Statistical Analysis: 1 - 6 Year Lagged Variables (P/E to R&D Efficiency)

```r
# 5 Year Lag
lag_pe_rd_model <- lm(lag_test_0 ~ lag_test_1 + lag_test_2 + lag_test_3 + lag_test_4 + lag_test_5 + lag_test_6 + lag_test_7 + lag_test_8 + lag_test_9 + lag_test_10 + lag_test_11 + lag_test_12 + lag_test_13) + na.omit(result) # remove rows that contain "NA" estimate(result)
```
**Statistical Analysis**

Figure 1. The effect of research and development expenditure efficiency on forward Enterprise Value / Sales (*F*~(1,63)~ = 2.448; *P* < 0.007938). The regression line is explained by Y = 2.141e+01X.

**Linear Regression**

```{r}
summary(lag_EVs_rd_model)
```

**Note: Adjusted R-squared**

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 = 1.018e+00X.

**Linear Regression**

```{r}
summary(va(RDe_PE))
```

**Note: Adjusted R-squared**

**Process Step - Creating Lagged Variables**

```{r}
lm(data=lag_test_0, PE~RD_Efficiency)
lm(data=lag_test_0, EV_Sales~RD_Efficiency)
lm(RD_Efficiency ~ factor(Size_Factor))
lm(RD_Spend ~ factor(Size_Factor))
lm(lag_test_0$RD_Spend ~ lag_test_0$Size_Factor)
lm(lag_test_0$Mkt_Cap ~ lag_test_0$Size_Factor)
```

**Additional Examination**

6. The Adjusted R-squared: 0.32, p-value: 3.503e-06, F-statistic: 15.4, DF of 60

7. The Adjusted R-squared: 0.022, p-value: 8.204e-03, F-statistic: 8.204, DF of 60


9. The Adjusted R-squared: 0.092

**Data Visualization - Single Linear Regressions**

**Primary Data Visualization Analysis**

**Linear Regressions**

**R&D Efficiency vs. Forward Price/Earnings**

```{r}
theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
xlab("Research & Development Expenditure Efficiency")+ylab("Forward Price / Earnings Multiples")+
theme_classic()+
geom_smooth(method=lm,se=FALSE)+
geom_point()
```

**R&D Efficiency vs. Forward EV/Sales**

```{r}
theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
theme_classic() +
xlab("Research & Development Expenditure Efficiency")+ylab("Forward Enterprise Value / Sales Multiples")+
theme_classic() +
geom_smooth(method=lm,se=FALSE)+
geom_point()
```

**R&D Efficiency vs. Market Capitalization**

```{r}
theme(axis.line.x = element_line(color = "black"), axis.line.y = element_line(color = "black"))
theme_classic()
```

**Primary Data Visualization Analysis**

**Linear Regressions**

**R&D Spend vs. Size Factor**

```{r}
lm(lag_test_6$EV_Sales ~ lag_test_6$RD_efficiency)
lm(lag_test_5$EV_Sales ~ lag_test_5$RD_efficiency)
lm(lag_test_4$EV_Sales ~ lag_test_4$RD_efficiency)
lm(lag_test_3$EV_Sales ~ lag_test_3$RD_efficiency)
lm(lag_test_2$EV_Sales ~ lag_test_2$RD_efficiency)
lm(lag_test_1$EV_Sales ~ lag_test_1$RD_efficiency)
lm(lag_test_0$EV_Sales ~ lag_test_0$RD_efficiency)
```

**Linear Regressions**

**R&D Spend vs. Size Factor**

```{r}
lm(lag_test_6$PE ~ lag_test_6$RD_efficiency)
lm(lag_test_5$PE ~ lag_test_5$RD_efficiency)
lm(lag_test_4$PE ~ lag_test_4$RD_efficiency)
lm(lag_test_3$PE ~ lag_test_3$RD_efficiency)
lm(lag_test_2$PE ~ lag_test_2$RD_efficiency)
lm(lag_test_1$PE ~ lag_test_1$RD_efficiency)
lm(lag_test_0$PE ~ lag_test_0$RD_efficiency)
```

**Linear Regressions**

**R&D Spend vs. Market Capitalization**

```{r}
lm(lag_test_6$EV_Sales ~ lag_test_6$RD_efficiency)
lm(lag_test_5$EV_Sales ~ lag_test_5$RD_efficiency)
lm(lag_test_4$EV_Sales ~ lag_test_4$RD_efficiency)
lm(lag_test_3$EV_Sales ~ lag_test_3$RD_efficiency)
lm(lag_test_2$EV_Sales ~ lag_test_2$RD_efficiency)
lm(lag_test_1$EV_Sales ~ lag_test_1$RD_efficiency)
```

**Additional Examination**

1. The Adjusted R-squared: 0.39, p-value: 8.204e-03, F-statistic: 8.204, DF of 60

2. The Adjusted R-squared: 0.22, p-value: 3.503e-06, F-statistic: 15.4, DF of 60

3. The Adjusted R-squared: 0.092
The regression line is explained by $Y = 4.184 \times 10^7X + 2.179 \times 10^1$. The regression line is explained by $Y = 2.301 \times 10^0X + 1.319 \times 10^1$. The regression line is explained by $Y = 2.156 \times 10^1X - 9.412 \times 10^{-2}$. The regression line is explained by $Y = 2.064 \times 10^0X + 1.319 \times 10^1$. The regression line is explained by $Y = 4.809 \times 10^7X + 3.123 \times 10^1$. The regression line is explained by $Y = 4.659 \times 10^0X - 3.141 \times 10^2$. The regression line is explained by $Y = 199.91; \ *P* < 2.2 \times 10^{-166}$. The regression line is explained by $Y = 101.4; \ *P* < 9.548 \times 10^{-4}$. The regression line is explained by $Y = 101.4; \ *P* < 9.548 \times 10^{-4}$. The regression line is explained by $Y = 199.91; \ *P* < 2.2 \times 10^{-166}$. The regression line is explained by $Y = 101.4; \ *P* < 9.548 \times 10^{-4}

**Statistical Analysis - Linear Regression - R&D Efficiency vs. Size Factor**

$$ RDs\_Size < \text{return} \text{lm} \text{(data=lag_test_0, Size\_Factor~RD\_Efficiency)} $$

```{r}
summary(RDs\_Size)
```
**Forward Enterprise Value / Sales vs. Size Factor**

The effect of forward enterprise value / sales on size factor (*F*~(1,63)~ = 2.4088; *P* < 0.1257). The regression line is explained by Y = 7.3239X + 2.6186. The Adjusted R-Squared is 0.02154.

**Figure 11.** The effect of research and development expenditure on size factor (*F*~(1,63)~ = 37.172; *P* < 7.257e-07). The regression line is explained by Y = 101183036X + 3998133. The Adjusted R-Squared is 0.5716.

**Statistical Analysis - Linear Regressions - Forward Enterprise Value / Sales vs. Market Capitalization**

**Figure 12.** The effect of forward enterprise value / sales on forward price / earnings (*F*~(1,63)~ = 86.39; *P* < 2.225e-16). The regression line is explained by Y = 3.3239X + 2.6186. The Adjusted R-Squared is 0.9686.

**Figure 13.** The effect of research and development expenditure efficiency a
Summary:

- **Price / Earnings Model 1b: Model Simplification (Short Version)**
  - $\text{Model} \textunderscore \text{class} \leftarrow \text{lm}(\text{PE} \sim \text{RD} \_\text{efficiency} \ast \text{Mkt} \_\text{Capitalization}, \text{data} = \text{lag} \textunderscore \text{test} \_0)$
  - Adjusted $R^2$ is 0.1657.
  - $F$-value is 3.66; $F$ < 0.007221.
  - The Adjusted $R^2$ is 0.1657.

- **Price / Earnings Model 1b3x: Model Simplification (Excluding Size Factor)**
  - $\text{Model} \textunderscore \text{class} \leftarrow \text{lm}(\text{PE} \sim \text{RD} \_\text{efficiency} \ast \text{Mkt} \_\text{Capitalization}, \text{data} = \text{lag} \textunderscore \text{test} \_0)$
  - Adjusted $R^2$ is 0.1399.
  - $F$-value is 3.98; $F$ < 0.007221.
  - The Adjusted $R^2$ is 0.1399.

**Statistical Analysis**

- **Refining Average Approved Drug Data Dataset**
  - `install.packages("cluster")`
  - `kmeans(Log_Approved_DD[,1:3], k, nstart = 10, iter.max = 100)`
  - `summary(kmeans_Approved_DD)`
  - **Note: Elbow seems ideal @ 3 or 4**
  - **Figure 20.**

- **Statistical Analysis**
  - **Figure 19.** The effect of research & development expenditure efficiency, factored by market capitalization size on forward price / earnings multiples ($F^2(-1,59) = 3.663; F^2 < 0.007221$).
  - The Adjusted $R^2$ is 0.1657.
  - **Statistical Analysis**
    - **Figure 18.** The effect of research & development expenditure efficiency, marked by capitalization size on forward price / earnings multiples ($F^2(-1,59) = 3.663; F^2 < 0.007221$).
    - The Adjusted $R^2$ is 0.1657.
  - **Statistical Analysis**
    - **Figure 17.** The effect of research & development expenditure efficiency, market capitalisation, factored by size on forward price / earnings multiples ($F^2(-1,59) = 2.083; F^2 < 0.01986$).
    - The Adjusted $R^2$ is 0.1657.

- **Statistical Analysis**
  - **Figure 16.** The effect of research & development expenditure efficiency and market capitalization on forward price / earnings multiples ($F^2(1,61) = 4.47; F^2 < 0.006672$).
  - The Adjusted $R^2$ is 0.1657.

**Statistical Analysis**

- **Importing Average Approved Drug Dataset**
  - `read.csv("C:/Users/nicho/Desktop/SPROJ/Data/R Studio/Datasets/Drug_Data_Average.csv")`
  - `data.frame(EV_Sales = Approved_Drug_Data$EV_Sales, Mkt_Cap = Approved_Drug_Data$Mkt_Cap)`

- **Statistical Analysis**
  - **Figure 20.**

**Data Visualization**

- **Statistical Analysis**
  - **Figure 19.**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

**Statistical Analysis**

- **Statistical Analysis**
  - **Statistical Analysis**

Anduze 77
**Data Visualization - Evaluate Mean Analysis - Forward Price / Earnings & Average EV/Sales**

```{r}
# Creating R&D Spend & Market Capitalization Dataframe
RD_Mktcap_df = na.omit(RD_Mktcap)
data.frame(RD_Spend = Log_Approved_DD$RD_Spend, Mkt_Cap = Log_Avg_Approved_DD$Mkt_Cap_Log)
RD_Mktcap_df = na.omit(RD_Mktcap_df)
points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 2)
plot(Log_Avg_Approved_DD[, 2], Log_Avg_Approved_DD[, 3], col = fit$cluster, xlab = "Cumulative FDA Approvals", ylab = "Enterprise Value / Sales"
```
```r
RDe_EV_Sales_df$RD_Efficiency <- RDe_EV_Sales_df
```
### Statistical Analysis - k-Means Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Market Cap vs. R&D Efficiency**

- Vector of WSS Values for Different # of clusters
  - `wss = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    wss[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

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

- Note: 3 or 4 clusters ideal

### Data Visualization - k-Means Mean Analysis - Market Cap vs. R&D Efficiency**

- k-means(MktCap_PE_df, centers = 3)
  - `kmeans()` function to identify clusters

- Plotting PE vs. R&D Efficiency
  - `plot(RDe_PE_df[, 2], RDe_PE_df[, 1], col = RDe_PE_df_fit$cluster, xlab = "Research & Development Expenditure Efficiency", ylab = "Market Capitalization")`

- Figure 28

### Statistical Analysis - EV/Sales, R&D Efficiency, and Market Capitalization Dataframe

- Vector of EV/Sales for Different # of clusters
  - `ev_sales = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    ev_sales[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Plotting EV/Sales vs. R&D Efficiency
  - `plot(MktCap_PE_df[, 1], MktCap_PE_df[, 2], col = MktCap_PE_df_fit$cluster, xlab = "Research & Development Expenditure Efficiency", ylab = "Market Capitalization", xlim = c(8000, 10000000), ylim = c(7, 42))`

- Figure 27

### Data Visualization - EV/Sales, R&D Efficiency, and Mkt. Cap Dataframe

- Vector of Market Capitalization for Different # of clusters
  - `mkt_cap = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    mkt_cap[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Plotting Market Capitalization vs. R&D Efficiency
  - `plot(MktCap_PE_df[, 2], MktCap_PE_df[, 1], col = MktCap_PE_df_fit$cluster, xlab = "Market Capitalization", ylab = "Forward Price / Earnings Multiples", xlim = c(11, 20), ylim = c(5, 260))`

- Figure 26

### Creating Market Cap & Forward P/E Multiple Dataframe

- Vector of Market Capitalization for Different # of clusters
  - `mkt_cap = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    mkt_cap[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Creating Market Capitalization Dataframe (Elbow Method)
  - `mkt_cap_df = as.data.frame(MktCap_PE_df)`

- Vector of EV/Sales for Different # of clusters
  - `ev_sales = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    ev_sales[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Creating EV/Sales Dataframe (Elbow Method)
  - `ev_sales_df = as.data.frame(MktCap_PE_df)`

- Vector of R&D Efficiency for Different # of clusters
  - `rd_efficiency = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    rd_efficiency[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Creating R&D Efficiency Dataframe (Elbow Method)
  - `rd_efficiency_df = as.data.frame(MktCap_PE_df)`

### Creating P/E & R&D Efficiency Dataframe

- Vector of EV/Sales for Different # of clusters
  - `ev_sales = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    ev_sales[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Creating P/E & R&D Efficiency Dataframe
  - `ev_sales_df = as.data.frame(MktCap_PE_df)`

- Vector of Market Capitalization for Different # of clusters
  - `mkt_cap = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    mkt_cap[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Creating Market Capitalization Dataframe (Elbow Method)
  - `mkt_cap_df = as.data.frame(MktCap_PE_df)`

- Vector of EV/Sales for Different # of clusters
  - `ev_sales = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    ev_sales[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Creating EV/Sales Dataframe (Elbow Method)
  - `ev_sales_df = as.data.frame(MktCap_PE_df)`

- Vector of R&D Efficiency for Different # of clusters
  - `rd_efficiency = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    rd_efficiency[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

- Creating R&D Efficiency Dataframe (Elbow Method)
  - `rd_efficiency_df = as.data.frame(MktCap_PE_df)`

### Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method)

- Vector of WSS Values for Different # of clusters
  - `wss = vector("numeric", length = 10)` set the maximum number of clusters you want to consider

  ```r
  for (i in 1:10) {
    wss[i] = kmeans(MktCap_PE_df, centers = i, nstart = 10)
  }
  ```

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

- Note: 3 or 4 clusters ideal

### Data Visualization - k-Means Mean Analysis - Market Cap vs. R&D Efficiency**

- k-means(MktCap_PE_df, centers = 3)
  - `kmeans()` function to identify clusters

- Plotting PE vs. R&D Efficiency
  - `plot(RDe_PE_df[, 2], RDe_PE_df[, 1], col = RDe_PE_df_fit$cluster, xlab = "Research & Development Expenditure Efficiency", ylab = "Market Capitalization", xlim = c(8000, 10000000), ylim = c(1800000, 275000000))`

- Figure 28
For statistical analysis, I used the Elbow Method to determine the ideal number of clusters for my "Unapproved Drug Data" Dataframe. Here is the R code I used:

```r
# Statistical Analysis - Elbow Method
kmeans_fit_EVs_m1 <- kmeans(lag_test_0[, sapply(lag_test_0, is.numeric)], vector("numeric", length = 10))
```

I then calculated the within-group sum of squares for different numbers of clusters:

```r
# Plot the WSS values against the number of clusters
plot(k$tot.withinss) # Plot the WSS values against the number of clusters
```

Finally, I ran a k-means clustering algorithm to cluster my data:

```r
kmeans_fit <- kmeans(numeric_lag_PE[, 1:4], k)
```

I then created scatterplots with cluster centers and points colored by cluster:

```r
plot(df_log[, 2], df_log[, 1], col = fit$cluster, size = df_log[, 3] / size_factor)
```

And here is the summary of the k-means fit:

```r
summary(kmeans_fit)
```

I also ran a k-means analysis using the EV/Sales ratio for my "Unapproved Drug Data" Dataframe.

```r
kmeans_fit_EVs_m1 <- kmeans(lag_test_0[, sapply(lag_test_0, is.numeric)], vector("numeric", length = 10))
```

I then calculated the within-group sum of squares for different numbers of clusters:

```r
# Plot the WSS values against the number of clusters
plot(k$tot.withinss) # Plot the WSS values against the number of clusters
```

Finally, I ran a k-means clustering algorithm to cluster my data:

```r
kmeans_fit <- kmeans(numeric_lag_PE[, 1:4], k)
```

I then created scatterplots with cluster centers and points colored by cluster:

```r
plot(df_log[, 2], df_log[, 1], col = fit$cluster, size = df_log[, 3] / size_factor)
```

And here is the summary of the k-means fit:

```r
summary(kmeans_fit)
```

I also ran a k-means analysis using the EV/Sales ratio for my "Unapproved Drug Data" Dataframe.

```r
kmeans_fit_EVs_m1 <- kmeans(lag_test_0[, sapply(lag_test_0, is.numeric)], vector("numeric", length = 10))
```

I then calculated the within-group sum of squares for different numbers of clusters:

```r
# Plot the WSS values against the number of clusters
plot(k$tot.withinss) # Plot the WSS values against the number of clusters
```

Finally, I ran a k-means clustering algorithm to cluster my data:

```r
kmeans_fit <- kmeans(numeric_lag_PE[, 1:4], k)
```

I then created scatterplots with cluster centers and points colored by cluster:

```r
plot(df_log[, 2], df_log[, 1], col = fit$cluster, size = df_log[, 3] / size_factor)
```

And here is the summary of the k-means fit:

```r
summary(kmeans_fit)
```

I also ran a k-means analysis using the EV/Sales ratio for my "Unapproved Drug Data" Dataframe.

```r
kmeans_fit_EVs_m1 <- kmeans(lag_test_0[, sapply(lag_test_0, is.numeric)], vector("numeric", length = 10))
```

I then calculated the within-group sum of squares for different numbers of clusters:

```r
# Plot the WSS values against the number of clusters
plot(k$tot.withinss) # Plot the WSS values against the number of clusters
```

Finally, I ran a k-means clustering algorithm to cluster my data:

```r
kmeans_fit <- kmeans(numeric_lag_PE[, 1:4], k)
```

I then created scatterplots with cluster centers and points colored by cluster:

```r
plot(df_log[, 2], df_log[, 1], col = fit$cluster, size = df_log[, 3] / size_factor)
```

And here is the summary of the k-means fit:

```r
summary(kmeans_fit)
```

I also ran a k-means analysis using the EV/Sales ratio for my "Unapproved Drug Data" Dataframe.

```r
kmeans_fit_EVs_m1 <- kmeans(lag_test_0[, sapply(lag_test_0, is.numeric)], vector("numeric", length = 10))
```

I then calculated the within-group sum of squares for different numbers of clusters:

```r
# Plot the WSS values against the number of clusters
plot(k$tot.withinss) # Plot the WSS values against the number of clusters
```

Finally, I ran a k-means clustering algorithm to cluster my data:

```r
kmeans_fit <- kmeans(numeric_lag_PE[, 1:4], k)
```

I then created scatterplots with cluster centers and points colored by cluster:

```r
plot(df_log[, 2], df_log[, 1], col = fit$cluster, size = df_log[, 3] / size_factor)
```

And here is the summary of the k-means fit:

```r
summary(kmeans_fit)
```

I also ran a k-means analysis using the EV/Sales ratio for my "Unapproved Drug Data" Dataframe.

---

Anduze 81
**Statistical Analysis - 4-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method)**  
- mktcap and ev/sales

```r
points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
plot(URD_Mktcap_df[, 1], URD_Mktcap_df[, 2], col = URD_Mktcap_df_fit$cluster, xlab = "Research & Development Expenditure", ylab = "Market Capitalization", xlim = c(0, 20), ylim = c(0, 150))
points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```

**Figure 30**

- r&d spend and pe

```r
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))
```

**Figure 31**

- r&d spend and ev/sales

```r
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, 550))
```

**Figure 32**

### 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$Mkt_Cap_Log
URD_Mktcap_df$LOG_EV_Sales <- na.omit(Log_Unapproved_DD$EV_Sales)
```

### Statistical Analysis - 4-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Forward Price / Earnings Multiples**

**Note:** 3 - 4 clusters ideal**

**Data Visualization - 4-cluster Mean Analysis - R&D Expenditure & Forward Price / Earnings Multiples**

```r
kmeans(URD_Mktcap_df, centers = i, nstart = 10)
for (i in 1:10) {
  wss <- k$tot.withinss
}
```

**Statistical Analysis - 4-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Forward Price / Earnings Multiples**

**Note:** 3 - 4 clusters ideal**

**Data Visualization - 4-cluster Mean Analysis - R&D Expenditure & Forward Price / Earnings Multiples**

```r
kmeans(URD_PE_df, centers = i, nstart = 10)
for (i in 1:10) {
  wss <- k$tot.withinss
}
```

**Statistical Analysis - 4-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Forward Enterprise Value / Sales Multiples**

**Note:** 3 - 4 clusters ideal**

**Data Visualization - 4-cluster Mean Analysis - R&D Expenditure & Forward Enterprise Value / Sales Multiples**

```r
kmeans(URD_EVs_df, centers = i, nstart = 10)
for (i in 1:10) {
  wss <- k$tot.withinss
}
```

### Creating R&D Spend & Forward PE Dataframe

```r
URD_PE_df <- data.frame(RD_Spend = Log_Unapproved_DD$RD_Spend)
URD_PE_df$PE <- na.omit(Log_Unapproved_DD$PE)
```

### Statistical Analysis - 4-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Forward Price / Earnings Multiples**

**Note:** 3 - 4 clusters ideal**

**Data Visualization - 4-cluster Mean Analysis - R&D Expenditure & Forward Price / Earnings Multiples**

```r
kmeans(URD_PE_df, centers = i, nstart = 10)
for (i in 1:10) {
  wss <- k$tot.withinss
}
```

### 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 <- na.omit(Log_Unapproved_DD$EV_Sales)
```

### Statistical Analysis - 4-cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - R&D Expenditure & Forward Enterprise Value / Sales Multiples**

**Note:** 3 - 4 clusters ideal**

**Data Visualization - 4-cluster Mean Analysis - R&D Expenditure & Forward Enterprise Value / Sales Multiples**

```r
kmeans(URD_EVs_df, centers = i, nstart = 10)
for (i in 1:10) {
  wss <- k$tot.withinss
}
```
**Figure 35.**
```
points(fit$centers[, 2], fit$centers[, 1], pch = 19, cex = 0)
```

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

**Note:** 3 clusters ideal
```
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups sum of squares")
```
**Data Visualization - k-Cluster Mean Analysis - Total Drug Dataset - Research & Development Expenditure & Forward Price / Earnings Multiples**

```{r}
TPE_Mktcap_fit <- kmeans(TPE_Mktcap, 4)
plot(TPE_Mktcap[, 1], TPE_Mktcap[, 2], col = TPE_Mktcap_fit$cluster, 
     xlab = "Research & Development Expenditure", ylab = "Forward Price / Earnings Multiples", 
     xlim = c(0, 25), ylim = c(0, 550))
points(fit$centers[, 1], fit$centers[, 2], pch = 19, cex = 0)
```

**Figure 36.**

# Creating R&D & Market Capitalization Dataframe - Total Drug Dataset

```{r}
TPE_Mktcap_df <- data.frame(TPE_Mktcap) 
TPE_Mktcap_FF <- log(TPE_Mktcap[, 2]) 
TPE_Mktcap_df <- cbind(TPE_Mktcap, TPE_Mktcap_FF)
```

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - R&D Expenditure & Market Capitalization

```{r}
for (i in 1:10) { 
  wss[i] <- kmeans(TPE_Mktcap, centers = i, nstart = 10)
  k$tot.withinss
}
```

**Figure 37.**

**Figure 38.**

**Figure 39.**

**Note: 3-4 clusters ideal**

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - P/E & EV/Sales

```{r}
TRD_PE_df <- kmeans(TRD_PE_df, 4)
plot(TRD_PE_df[, 1], TRD_PE_df[, 2], col = TRD_PE_df_fit$cluster, 
     xlab = "Research & Development Expenditure", ylab = "Market Capitalization", 
     xlim = c(4, 18), ylim = c(7, 20))
points(fit$centers[, 1], fit$centers[, 2], pch = 19, cex = 0)
```

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - P/E & EV/Sales

```{r}
for (i in 1:10) { 
  wss[i] <- kmeans(TPE_EVs_df, centers = i, nstart = 10)
  k$tot.withinss
}
```

**Figure 37.**

**Figure 38.**

**Figure 39.**

**Note: 3-4 clusters ideal**

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - Mkt Cap & P/E

```{r}
TRD_Mktcap_df <- kmeans(TRD_Mktcap_df, 4)
plot(TRD_Mktcap_df[, 1], TRD_Mktcap_df[, 2], col = TRD_Mktcap_df_fit$cluster, 
     xlab = "Research & Development Expenditure", ylab = "Market Capitalization", 
     xlim = c(4, 17), ylim = c(0, 250))
points(fit$centers[, 1], fit$centers[, 2], pch = 19, cex = 0)
```

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - Mkt Cap & P/E

```{r}
for (i in 1:10) { 
  wss[i] <- kmeans(TPE_EVs_df, centers = i, nstart = 10)
  k$tot.withinss
}
```

**Figure 37.**

**Figure 38.**

**Figure 39.**

**Note: 3-4 clusters ideal**

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - P/E & EV/Sales

```{r}
TRD_PE_df <- kmeans(TRD_PE_df, 4)
plot(TRD_PE_df[, 1], TRD_PE_df[, 2], col = TRD_PE_df_fit$cluster, 
     xlab = "Research & Development Expenditure", ylab = "Market Capitalization", 
     xlim = c(4, 17), ylim = c(0, 250))
points(fit$centers[, 1], fit$centers[, 2], pch = 19, cex = 0)
```

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - Mkt Cap & P/E

```{r}
TRD_Mktcap_df <- kmeans(TRD_Mktcap_df, 4)
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[, 1], fit$centers[, 2], pch = 19, cex = 0)
```

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - Mkt Cap & P/E

```{r}
for (i in 1:10) { 
  wss[i] <- kmeans(TPE_EVs_df, centers = i, nstart = 10)
  k$tot.withinss
}
```

**Figure 37.**

**Figure 38.**

**Figure 39.**

**Note: 3-4 clusters ideal**

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - P/E & EV/Sales

```{r}
TRD_PE_df <- kmeans(TRD_PE_df, 4)
plot(TRD_PE_df[, 1], TRD_PE_df[, 2], col = TRD_PE_df_fit$cluster, 
     xlab = "Research & Development Expenditure", ylab = "Market Capitalization", 
     xlim = c(4, 17), ylim = c(0, 250))
points(fit$centers[, 1], fit$centers[, 2], pch = 19, cex = 0)
```

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - Mkt Cap & P/E

```{r}
TRD_Mktcap_df <- kmeans(TRD_Mktcap_df, 4)
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[, 1], fit$centers[, 2], pch = 19, cex = 0)
```

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - Mkt Cap & P/E

```{r}
for (i in 1:10) { 
  wss[i] <- kmeans(TPE_EVs_df, centers = i, nstart = 10)
  k$tot.withinss
}
```

**Figure 37.**

**Figure 38.**

**Figure 39.**

**Note: 3-4 clusters ideal**

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - P/E & EV/Sales

```{r}
TRD_PE_df <- kmeans(TRD_PE_df, 4)
plot(TRD_PE_df[, 1], TRD_PE_df[, 2], col = TRD_PE_df_fit$cluster, 
     xlab = "Research & Development Expenditure", ylab = "Market Capitalization", 
     xlim = c(4, 17), ylim = c(0, 250))
points(fit$centers[, 1], fit$centers[, 2], pch = 19, cex = 0)
```

# Statistical Analysis - k-Cluster Mean Analysis; Identifying Ideal # of Clusters (Elbow Method) - Total Drug Dataset - Mkt Cap & P/E

```{r}
TRD_Mktcap_df <- kmeans(TRD_Mktcap_df, 4)
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[, 1], fit$centers[, 2], pch = 19, cex = 0)
```
The regression line is explained by \( Y = 101183036X + 3998133 \).

The Adjusted R-Squared is 0.6734.

**Figure 41.** The effect of market capitalization category (size factor) on research and development expenditure (*F*~(1,63)~ = 4.722; *P* < 0.0348). The regression line is explained by \( Y = 4.722X + 3998133 \).

**Logged R&D Spend vs. Size Factor**

**Note:** Adjusted R-Squared = 0.6734; F-statistic: 15.6, DF of 60

**Forward Price / Earnings vs. Size Factor**

Squared is 0.3488.

**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 \( Y = 2.173 + 1.751e^{-10}X \).

**Enterprise Value / Sales vs. Size Factor**

Squared is 0.02154.

**Figure 12.** The effect of Forward Enterprise Value / Sales on Size Factor (*F*~(1,63)~ = 2.4088 ; *P* < 0.1257). The regression line is explained by \( Y = 131805059 + 3998133 \).

**R&D Efficiency vs. Size Factor**

Squared is 0.022.

**Figure 40.** The effect of market capitalization category (size factor) on research and development expenditure efficiency on forward Enterprise Value / Sales (*F*~(1,63)~ = 37.172; *P* < 7.257e-07). The regression line is explained by \( Y = 6.722 + 1.763e-07X \). The Adjusted R-Squared is 0.242.

# Statistical Analysis - Boxplots

**Unapproved Drug Dataset - Statistical Analysis**

- **Defining Cumulative FDA Approvals and Size Factor Dataset**
- **Fitting a Linear Model**
- **Summary of the Linear Model**
- **Refining Unapproved Drug Dataset**

**Statistical Analysis - Approved Drug Dataset**

- **Defining Cumulative FDA Approvals and Size Factor Dataset**
- **Fitting a Linear Model**
- **Summary of the Linear Model**

# End of Total Drug Dataset - Cluster Mean Analysis
**Statistical Analysis - Histograms**

### Total Drug Dataset

**Defining Total Drug Dataset DataFrame**

```r
Total_Drug_Data = as.data.frame(sapply(Total_Drug_Data[,is.numeric], log))
```

**Statistical Analysis - Boxplots**

- **Logged Forward Price vs. Size Factor**
- **Logged Enterprise Value vs. Size Factor**
- **Logged Research & Development Expenditure vs. Size Factor**

### Approved Drug Dataset

**Efficiency Distribution**

```r
hist(Summary_EV_Sales / Sales, main = "Efficiency Distribution" , main = "Efficiency Distribution")
```

**Expenditure Distribution**

```r
hist(Summary_Rndrt_mul / Earnings, main = "Expenditure Distribution" , main = "Expenditure Distribution")
```

### Unapproved Drug Dataset

**Efficiency Distribution**

```r
hist(Summary_EV_Sales / Sales, main = "Efficiency Distribution" , main = "Efficiency Distribution")
```

**Expenditure Distribution**

```r
hist(Summary_Rndrt_mul / Earnings, main = "Expenditure Distribution" , main = "Expenditure Distribution")
```

# Statistical Analysis - Boxplots

### Total Drug Dataset

**Defining Total Drug Dataset DataFrame**

```r
Total_Drug_Data$Log_Mkt_Cap = log(Total_Drug_Data$Mkt_Cap)
Total_Drug_Data$Log_PE = log(Total_Drug_Data$PE)
Total_Drug_Data$Log_EV_Sales = log(Total_Drug_Data$EV_Sales)
Total_Drug_Data$Log_RD_Spend = log(Total_Drug_Data$RD_Spend)
```

**Statistical Analysis - Boxplots**

- **Logged Mkt Cap vs. Size Factor**
- **Logged PE vs. Size Factor**
- **Logged EV Sales vs. Size Factor**
- **Logged RD Spend vs. Size Factor**

# Approved Drug Dataset

**Efficiency Distribution**

```r
hist(Summary_EV_Sales / Sales, main = "Efficiency Distribution" , main = "Efficiency Distribution")
```

**Expenditure Distribution**

```r
hist(Summary_Rndrt_mul / Earnings, main = "Expenditure Distribution" , main = "Expenditure Distribution")
```

# Unapproved Drug Dataset

**Efficiency Distribution**

```r
hist(Summary_EV_Sales / Sales, main = "Efficiency Distribution" , main = "Efficiency Distribution")
```

**Expenditure Distribution**

```r
hist(Summary_Rndrt_mul / Earnings, main = "Expenditure Distribution" , main = "Expenditure Distribution")
```

---

**Anduze 86**
```{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**

```{r}
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**

```{r}
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}
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**

```{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**

```{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**

```{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}
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**

```{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**

```{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**

```{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)
```

```{r}
ExEVs_model <- lm(lag_test_0$EV_Sales ~ factor(lag_test_0$Experience))
summary(ExEVs_model)
```

**RD Efficiency vs. Experience - Approved Drug Dataset**

```{r}
ExRDe_model <- lm(lag_test_0$RD_efficiency ~ factor(lag_test_0$Experience))
summary(ExRDe_model)
```

**PE vs. Experience - Approved Drug Dataset**

```{r}
KmPE_model <- glm(lag_test_0$PE ~ factor(lag_test_0$Experience))
summary(KmPE_model)
```

**EV/Sales vs. Experience - Approved Drug Dataset**

```{r}
KmEVs_model <- glm(lag_test_0$EV_Sales ~ factor(lag_test_0$Experience))
summary(KmEVs_model)
```

**Research & Development Efficiency vs. Experience**

```{r}
KmRDe_model <- glm(lag_test_0$RD_efficiency ~ factor(lag_test_0$Experience))
summary(KmRDe_model)
```

**Forward EV/Sales vs. Experience**

```{r}
Fig28 <- ggplot(data=lag_test_0, aes(x=factor(Experience), y=EV_Sales)) + geom_boxplot() + xlab("Experience") + ylab("Forward Enterprise Value / Sales Multiples") + theme_classic()
```

**Figure 60. The effect of experience for approved manufacturers on forward Enterprise Value / Sales multiple (*F*~(1,63)~ = 1.646; *P* = 0.2043). The Adjusted R-squared is 0.0099.**

**Forward P/E Multiples vs. Experience**

```{r}
Fig29 <- ggplot(data=lag_test_0, aes(x=factor(Experience), y=PE)) + geom_boxplot() + xlab("Experience") + ylab("Forward Price / Earnings Multiples") + theme_classic()
```

**Figure 61. The effect of experience for approved manufacturers on forward Price / Earnings multiple (*F*~(1,63)~ = 7.38; *P* = 0.0085). The Adjusted R-squared is 0.0906.**

**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()
```

**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.**

Anduze 87
**RD Spend vs. Experience**

```{r}
ExRDs_model <- lm(lag_test_0$RD_Spend ~ factor(lag_test_0$Experience))
summary(ExRDs_model)
```

**Research & Development Expenditure vs. Experience**

```{r}
ExRDs_model <- lm(lag_test_0$RD_Spend ~ factor(lag_test_0$Experience))
summary(ExRDs_model)
```

**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**

```{r}
ExMktcap_model <- lm(lag_test_0$Mkt_Cap ~ factor(lag_test_0$Experience))
summary(ExMktcap_model)
```

**Market Capitalization vs. Experience**

```{r}
ExMktcap_model <- lm(lag_test_0$Mkt_Cap ~ factor(lag_test_0$Experience))
summary(ExMktcap_model)
```

**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**

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

**FDA Approvals vs. Experience**

```{r}
Experience_Test <- lm(FDA_Approvals$Approvals ~ factor(FDA_Approvals$Experience))
summary(Experience_Test)
```

**FDA Approvals vs. Experience**

```{r}
Experience_Test <- lm(FDA_Approvals$Approvals ~ factor(FDA_Approvals$Experience))
summary(Experience_Test)
```

**Figure 65. The effect of experience on cumulative FDA approvals (F{~(1,629)} = 366.9; *F* = 2.2e-16). The Adjusted R-Squared is 0.4632.**

# End of Additional Analysis
Bibliography


Anduze 90


Anduze 91


Anduze 93


