From Rural to Urban: Understanding America Through the Census

R. Owen Hartman
Bard College, rh8311@bard.edu
From Rural to Urban:
Understanding America Through the Census

Senior Project submitted to
the Division of Social Studies
and
the Division of Science, Mathematics and Computing
of Bard College

by
R. Owen Hartman

Annandale on Hudson, NY
Ithaca, NY
May 2020
Acknowledgements

Inspired by Bob Kibbee and the History Center of Ithaca’s HistoryForge Project, this first-ever Computer Science and Historical Studies Senior Project would never have been possible without the data provided by the IPUMS project. All data used in the creation of this project was provided by IPUMS.org and I will forever be in debt to the work of Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek.

Without the guidance of Amy Shein, this dyslexic student would never have been able to reach any of the goals that I had set out for myself in college. Amy will forever have a place in my heart for supporting me throughout my time in higher education.

Thank you to Keith O’Hara, Greg Moynahan, Joel Perlmann and Kerri Ann Norton for advising my project and participating in my board.

Thanks to Mary Grace Williams for always keeping me safe and another special thank you to both Omar Cheta and Christian Crouch for forever instilling into me the image of a model historian.

Last but not least, thank you to my friends and family for always supporting me. My parents, Ellen and Rick, my little brother Cameron, my Grandparents, Godparents, Aunts, Uncles and Cousins, my ever-loving and always-loyal schnoodle, Rose-Marie “Rosie” Hartman and my wonderful roommates; Addison Dowell, Will Burghard and Lindsey Williams.
Dedicated to my Grandmother, M. Jayne Hartman who lost her battle against the Coronavirus on Saturday morning, April 25th 2020.
Table of Contents

Introduction:

Du Bois, Hacker and the Case Studies 6

The Census of 1890 12

IPUMS.org 20

Regarding the Accuracy of the Census of 1870 21

Alexandria Ocasio-Cortez and the Reconstruction era 28

Research and Display Methods when Visualizing the Industrial Revolution 30

The Exposition Universelle 33

The Charts 39

Program Analysis 70

Final Thoughts 88

Code Used in Programs 89

Bibliography 105
Following in the footsteps of the work performed by W.E.B. Du Bois and David J. Hacker, the first-ever Senior Project to combine the disciplines of Historical Studies and Computer Science at Bard College, will showcase a variety of infographics created with Python3 programming, utilizing Census data provided by IPUMS.org. All charts and figures showcased were handcrafted unless otherwise specified. During the creation of the infographics, a conscious choice was made to remain consistent with the types of graphs used and save experimentation for the type of data collected. Various forms of statistics are displayed in pie and bar charts with a focus on minimalism and clarity so that they are easy to understand for viewers.

Though Computer Science was invented less than a century from the writing of this paper, historians have always been involved in making graphs and charts in order to display population data. A prime example of this would be the *American Negro Exhibit*, created by W.E.B. Du Bois in 1900 for the *Universelle Exposition* in Paris. W.E.B. Du Bois sent 63 infographics or “plates” across the Atlantic Ocean in order to showcase life for African Americans in the American South through data and statistics. Du Bois, one of the earliest sociologists in the United States, and his team of students had created these infographics with a clear intention of challenging beliefs about life for African Americans which were commonly held by Europeans. Color was used to great effect on each of the plates, a practice rare for infographics during the time they were created.
Featured below is an analysis of the value of property owned by African Americans in the American South created by W.E.B. Du Bois and his team of researchers.

This chart is noticeably minimalist in its design. There is no key or legend, values and data description is included inside of the chart. A dark black circle, located in the center, represents the value of property in 1875. Different colored rings represent different years studied.
Though the title of this piece is “From Rural to Urban”, the actual work done spans far beyond just the movement of Americans from rural spaces to urban ones. The industrial revolution during the turn of the 20th century fundamentally changed the lives of countless Americans in many different ways. Two separate case studies have been created in order to understand and visualize the data collected during this time period.

Case study 1: Americans traveled from rural areas to urban ones. Which cities experienced the most growth, how many immigrants account for the growth of those cities. The first year that more Americans marked off on the census to be living in urban areas rather than rural ones is 1920. Farms and the agriculture industry also began to lose prominence by 1920. Though the agriculture industry remained the most worked in industry it had fallen a long way from where it had originally started. Farms were no longer as sizable a percentage of American households as they had been in the census of 1870. This can be seen inside of the workforce as well with both the agriculture and private household industries seeing a decline by the census of 1930.
An example of **Case Study 1** is shown below:

Urban/Rural Percentages in the Census of 1910 with Home Ownership Subdivisions

- **Urban Owners**: 15.9%
- **Rural Renters**: 25.0%
- **Urban Renters**: 28.4%
- **Rural Owners**: 30.7%
Case Study 2: How did the American workforce change during the industrial revolution?

Agriculture is the most worked in industry recorded throughout all the censuses from 1870 to 1930. Other industries such as construction were prominent throughout the time period. Private household work decreased steadily from 1870 to 1930, Americans were doing less work inside of households for single families and doing more work inside of factories. Railways and railroads began to play a large role in the American economy as the country moved westward and required transportation of people and materials back and forth. Industrial America moved many workers from all over the country from work on farms such as agriculture and private households to more industrial work inside of or in relation to cities.
An example of Case Study 2 is shown below:
When analyzing censuses taken inside of the United States from 1870 to 1930 it is extremely important to recognize that there is a gap in population data from 1880 to 1900. An examination of the ill-fated Census of 1890 follows:

The date is Sunday, March 22nd 1896. Early in the morning a fire breaks out in Marini’s Hall in Washington D.C. Named after a man called Marini, a famed dance instructor, the school opened in 1876.¹ By this time in 1896, the Census Bureau had been occupying the building for years in order to house their census records. At the time of the fire, a number of records from the Census of 1890 were being housed there. The fire started in the basement and was thick with smoke which prevented firefighters from locating and putting out the flames. Fighting through the dense smoke, the firefighters were eventually able to extinguish the flames. This would not be the end however. Later that very same day fire again broke out in Marini’s Hall. This time, the fire department was prepared and were able to contain the flames in much quicker time. Damage to the building was estimated to be around 10,000 dollars. When adjusted for inflation this number exceeds 300,00 dollars.²

Three of the top minds behind the 1890 census were called in to inspect the damage to the records: Carroll D. Wright, commissioner of the census; George S. Donnell, chief of the census division of the Interior Department; and W.C Hunt, head of population statistics of the census. After investigation, they came to the conclusion that;

¹ With the Rambler
² Fire in Marini’s Hall
“the loss to the government was inconsiderable.” The census had survived. The valiant firefighters had saved the day and preserved the valuable census records from destruction.³

In 1896, there was no permanent census bureau, there would not be one until 1902 when the Department of the Interior absorbed the Census Office, transforming it into the Census Bureau.⁴ The fire at Marini’s Hall was a definite shock to the enumerators, as William A. King declared, “I do not consider this building at all suited for its present use. We have had one fire here… and I constantly fear another.”⁵ By 1913 the census files had been relocated to the basement of the newly created Census Bureau’s headquarters located inside of the Department of Commerce. During the winter of 1921, the records were stored securely. Records from 1790 to 1870 (excluding 1840 and 1850) were stored on the fifth floor. This was done in order to make the records available for researchers. Records for 1830, 1840, 1880, 1900 and 1910 were being held in a secure basement vault. The recent 1920 census was located in a different building altogether.⁶ The 1890 census was a different story however. The 1890 census had been, according to the Washington Post, “arranged on pine shelves with only 20-inch aisles between them.”⁷ Placing the census on these oak shelves would have permanent consequences. On January 19th 1921, a fire broke out in the Commerce building. The fire was extinguished, though not without serious damage to the census records. Records of the censuses, which had been stored in the secure

³ Fire in Marini’s Hall  
⁴ Blake 1  
⁵ Dorman 396  
⁶ Dorman 371  
⁷ Dorman 372
vault, sustained significant water damage due to the sprinkler system. This was not enough to lose the data however as researchers were able to re-record any material that was damaged beyond the help of conventional drying methods. A lucky fate not shared by the 1890 census because outside of the vault, on it's oak shelves, the 1890 census had been all but destroyed in the flames.

At the time of the fire it was estimated by the census director that “25 percent of the records had been consumed by the flames, while another 50 percent were damaged by water, smoke and fire.” This would suggest that a significant amount of the records survived. However, by 1932, the records were added to a list of documents to be destroyed pending approval of the Librarian of Congress. In accordance with the “useless paper” law of 1889, this destruction was approved and the records were subsequently destroyed. This occurred sometime in 1934 or 1935. With this action, Congress hammered the final nail on the coffin of the 1890 census. Whatever had been saved from destruction in 1921 had now been destroyed a decade later. No reason was given for the destruction of the records. Ironically, President Hoover laid the cornerstone of the National Archives Building only a day earlier.

This destruction of the 1890 census is a permanent loss to the national historic record. The census had been taken less than three decades after the end of the Civil War. It was also the first to be taken after the formal end of the Reconstruction era in the United States. The final census of the 19th century, 1890 was the first year to field more than a million people in a majority of the states surveyed. The 1890 census was

---

8 Dorman 375
9 Dorman 372
10 Dorman 376
the first to include a separate schedule for family forms. It was also the first to use
Herman Hollerith’s Electrical Tabulation system. This new technology involved
population data being punched into cards and funneled into data tables. At the time
this tabulation method was cutting edge technology and would later become infamous for its
connection to Nazi Germany in the 1930s. After working on the census, Herman
Hollerith would go onto found a company called Tabulating Machine Company.
Following a series of mergers and acquisitions, the Tabulating Machine Company would
become known as “The International Business Machines Corporation” or IBM for short.
The combination of the Electronic Tabulation and the separate schedule for family forms
lead to the census report becoming extremely large.

When asked about the size of the 1890 census Commissioner Wright replied that
it weighed, “more than 300 tons.” This extreme size was part of the reason why the
1890 census was stored outside of the other censuses. At the time of its destruction, the
1890 census was the only census to have separate family forms. This would become
common practice in the United States after 1970. Up until that point however, the 1890
census would be unique in this aspect. The combination of the census’s extreme size
and the resulting need for a separate storage space is further explained in William A.
King’s statement that, “The population schedules of this census alone will make five
times as much as all the schedules of all the censuses preceding and will require an
amount of shelving equal to one row over 7,000 feet long.” While being stored in the

---

1 Blake 2
2 Dorman 359-360
3 Dorman 356
4 Dorman 368
new Census Bureau location there would be many precautions taken in order to prevent fire. This included the prohibition of smoking, fire extinguishers made readily available, and the hiring of “night watchmen” in order to supervise the building. At the time there were already issues arising out of the basement storage space.\textsuperscript{15} These issues can already be seen by 1916 when a report stated that, “the basement storage area lay adjacent to a boiler room; temperatures there could reach ninety degrees.”\textsuperscript{\textsuperscript{15}}

The question must be asked, “What was truly lost during the destruction of the 1890 census?” At the time many census enumerators struggled with this question and encountered difficulty when explaining the importance of taking and preserving census records. “When census clerks and directors tried to explain the importance of preserving past census schedules, politicians and other government officials usually met them with incomprehension.”\textsuperscript{16} The Census of 1890 had a remarkable effect on American life in specifically two ways. Firstly, history classrooms in America were forever changed by its results. Secondly, the 1890 Census was used as the basis for the strict immigration laws passed by Congress in the 1920s. Robert L. Dorman can be quoted in his article \textit{The Creation and Destruction of the 1890 Census} claiming that, “in the minds of frontier mythologists and immigration opponents the 1890 census seemed to offer evidence on the perennial question, ‘What is American?’”\textsuperscript{17} The 1890 census was cited by Fredrick Jackson Turner in his infamous essay \textit{The Significance of the Frontier in American History}. This essay would be the origin of Turner’s Frontier Thesis which Dorman

\textsuperscript{15} Dorman 371
\textsuperscript{16} Dorman 366
\textsuperscript{17} Dorman 351
describes as, “perhaps the single most influential essay in American historiography.”\(^\text{18}\)

By the time of Turner’s death in 1932, 60% of leading history departments in the United States were teaching courses on frontier history based on this thesis.\(^\text{19}\) The census also had a big effect on immigration policy. When the United States Congress passed the Immigration Act of 1924, implementing the National Origins Formula, they cited the 1890 Census as their basis. This immigration plan which effectively banned all immigration from Asia and forever changed the ethnic makeup of the country would be in effect until 1965.\(^\text{20}\) The effects of these strict immigration quotas can still be felt in present day America and were based on data that does not exist anymore.

Beginning with the reconstruction era and ending during the height of the Great Depression, this will be a study into the effectiveness of studying the census using modern technology. Firstly, we will attempt to visualize the effects of the industrial revolution on the American landscape. Census officials released statistics in 1870 stating that, “between a fifth and quarter of the population lived in “urban” areas.”\(^\text{21}\) By 1920, the Census Bureau had reported that over 50 percent of Americans lived in urban areas. The urbanization of America would affect the entire country, fundamentally changing the lives of millions. Secondly, another part of this essay will deal with attempting to recreate the 1890 census. It is quite probable that the information in that census is lost to time. However, what would it take to recreate the census? By

\(^{18}\) Dorman 350
\(^{19}\) Allan G. Bogue
\(^{20}\) Dorman 351
\(^{21}\) Anderson 87
examining the censuses that came before and after 1890, it may be possible to find out what the census in 1890 might have looked like.

During the latter half of the 19th century, the census of the United States of America had radically expanded. Increases occurred in the number of volumes published, the amount of people employed in its creation and the cost per person surveyed. The number of published volumes of data increased from 5 in 1860 and 1870 to 23 in 1880 and 33 in 1890. The size of the staff in Washington tripled between 1860 and 1870. It then tripled in size again for the 1880 census. The cost of each person recorded was 6.3 cents in 1860, this would rise to 15.5 cents per person in 1900.22

Under the leadership of the then 29 year old, Francis Walker, the census would rise in popularity across the United States. Walker’s creation, the census Statistical Atlas, would become extremely popular with its many detailed and colorful maps of the United States. This atlas was able to visualize the country’s demographic statistics with only a quick glance at a page by the reader. Walker popularized some of the most well-known visual indicators of American population change such as population density maps and the concept of the center of population. Walker calculated this “center of the population” for every census from 1790 to 1870. He accomplished this by “representing the theoretical point of balance of the weight of the population on a map.”23 This center of the population consistently moved more and more westward with each census.

It takes 72 years for the full report of a census to be released to the public. Seventy-two is the average lifespan of an American. It is assumed that after 72 years

22 Anderson 87
23 Anderson 94
those who were featured in a census would be living drastically different lives so that it would no longer matter if their information were released to the public. In the meantime however, aggregate data is released to the public.24

24 US Census Bureau
Over 5 million entries in censuses ranging from 1870 to 1930 have been examined by computer programs. For this study, those entries were downloaded from IPUMS.org, an organization which provides “census and survey data from around the world integrated across time and space.” IPUMS originally stood for Integrated Public Use Microdata Series. This is no longer the case as the organization’s work expanded far beyond microdata and some projects have restricted access. One percent samples for each of the censuses were used. A sample allows for researchers to analyze data without fears of over or undercounts in specific areas. The following censuses were obtained: 1870, 1880, 1900, 1910, 1920 and 1930. IPUMS allows users to select which variables to include on the report. Housing information including home ownership and rural vs urban classification was included on the download as well as geographic information such as current county of residence. The graphs created for this project were made using Python and JavaScript. Another programming language, R, had to be used in order to decode the raw .dat file which the data was shipped in. IPUMS provides an .xml syntax file in order to apply labels to the often complex data. For example, ICPSR code “01” corresponds to the state of Connecticut, “02” corresponds to Maine and so on. This extends to job and industry descriptions, for example, industry code 597 corresponds to Sanitary services.
Regarding the Accuracy of the Census of 1870

In his article, *New estimates of Census Coverage in the United States*, J. David Hacker examines undercount rates for the native born white population in the United States. There has always been a popular sentiment that the census has undercounted the population throughout American history. In 1790, the genesis of the census, George Washington remarked that the “real number (of inhabitants) will greatly exceed the official return.” This was due to a belief held by Washington that American citizens were fearful of a population count due to religious reasons as well as a fear of being taxed. This belief of Washington was echoed as well by Thomas Jefferson. To quote Jefferson, “making very small allowances for omission which we know to have been great, we are certainly above 4 millions, probably about 4,100,000.” This was in response to the reported population of the United States being short of 4 million. The fascination surrounding population numbers was due to population size being linked to economic strength, a commonly held belief among politicians during the 18th and 19th century. George Washington was recorded to have boasted that the United States population would be found to exceed 5 million members. A population size that he believed would “astonish Europe” and “add consequence” to the United States.

Census Researcher, David Hacker, examined the particularly controversial census of 1870. The first census to be taken after the Civil War, this census was thought to be ridden with under-enumeration primarily in Southern States. Francis

---

25 Hacker 74
26 Hacker 75
Walker, the Census Director who oversaw the census of 1870, once stated that the census was inaccurate due to the duties being carried out by, “Negroes who could not write or read… Accompanied, perhaps, by some poor white man.” 27

This venomous rhetoric was echoed by Census Geographer Henry Gannett who alleged that the 1870 undercount was the result of poorly trained enumerators whom he believed to be, “ignorant negroes or nonresident carpetbaggers.” 26 Hacker finds that the 10.1 percent net undercount that was estimated by those involved in 1890 “far exceeds the new estimates.” Hacker found that relative to the 1860 Census, the 1870 Census missed 1.1 percent of the southern-born population. Additionally Hacker found when compared to the 1880 Census, the 1870 Census experienced an additional 3.6 percent undercount of southern-born whites. 28

Hacker’s examination utilizes censuses from 1850 to 1930 and he found the reputation of the census of 1870 to be “undeserved” with “slightly higher net undercounts for native-born white males relative to native-born white females.” 29 Hacker also finds that the 1880 census has the most accurate coverage of the native-born white population. Hacker was able to come to these results by utilizing data he had received from IPUMS.org. Hacker would then go on to compare his findings with Ansley Johnson Coale and Melvin Zelnik’s famous 1963 census estimates found in New Estimates of Fertility and Population in the United States. Hacker acknowledges that net undercount estimates are based on fallible evidence at the start of his paper. This is due to birth and death certificates not being established in the United States until the

27 Hacker 78  
28 Hacker 94  
29 Hacker 71
mid-1930s. Lack of proper age certification forces Hacker to exclude immigrants from his survey as their ages cannot be verified.\footnote{30}

IPUMS data was used for this study due to Hacker’s view that IPUMS has, “two major advantages over the published census tabulations used by Coale and Zelnik.”\footnote{31} The first being that native-born whites can be cross-tabulated by age and sex down to the single year and the second being that IPUMS allows for cross-tabulation by place of birth allowing for regional coverage to be achieved. Using the data received from IPUMS, Hacker calculates a, “true” annual series of whites births.”\footnote{28} By first projecting backwards, Hacker is able to find a series of “expected” age distributions. He then proceeds to compare to the enumerated population in order to find estimates of coverage based on age and sex. His results include many figures detailing topics such as age distribution, estimated net underenumeration errors, underenumeration errors in males and females, estimated net undercount, and estimated undercount by age and sex in both the north and the south. Hacker concludes that, when compared to the 1880 census, the 1870 census does in fact undercount the American South. However, Hacker goes on to explain that the reputation of being unreliable that was given to the 1870 census is “undeserved.”\footnote{32}

Hacker bases his results on methods detailed by Coale and Zelnik. Single-year census age distributions of native born white females and another distribution of males are back-projected with estimates of survival to the time of birth. This allows for Hacker

\footnotesize\begin{itemize}
\item \footnote{30} Hacker 71-72
\item \footnote{31} Hacker 82
\item \footnote{32} Hacker 95
\end{itemize}
to obtain several estimates of births for each year. These estimates were then
combined to produce, “one ‘true’ annual series of white births.”

Hacker lays out the four major assumptions that are required when using Coale
and Zelnik’s method. The first assumption is that there was negligible out-migration of
native-born whites from the United States. Though there is no comprehensive source to
document the level of out-migration, Hacker found that, “out-migration of native born
whites appears to be minimal.”

Secondly, Hacker assumes that age was correctly reported on the census. This
assumption by Coale and Zelnik is considered to be more “problematic” by Hacker. This
is due to the ability for an “age-heaping” error to occur in the data. An age-heaping error
is defined as when “an individual misstates his age by rounding it to a preferred age.”
Instead of using a smoothing algorithm, Hacker navigates this issue due to Zelnik
having an observation that a linear trend line could be, “fitted to the proportion falsely
choosing or avoiding a particular age between 1880 and 1950.”

The third assumption that Coale and Zelnik make is that reliable estimates of
mortality are available to back-project births from census age distributions. Hacker
notes that up until recently few estimates of mortality were available. Coale and Zelnik
assumed a “linear decline,” Hacker notes that populations, particularly in the Northeast,
had higher proportions of their population living in urban areas. Recent research by

33 Hacker 86
34 Hacker 83
35 Hacker 83
Michael R. Haine suggests that mortality was “variable without trend for most of the late nineteenth century.”

The final assumption made is that there is a “known and unchanging net undercount” of white women aged 15-29. Hacker constructs new estimates for this demographic by constructing new decennial male and female life tables for each decade between 1790 and 1900. Hacker finds that “In general, the new estimates for the nineteenth century represent lower overall survivorship than Coale and Zelnik assumed.” This leads Hacker to suspect that the back projected birth estimates were slightly higher than Coale and Zelnik’s own back projections. This is due to the lower overall survivorship of women during this time than previously thought. In order to find a “true” series of annual births, Hacker takes the average of his back projections and Coale and Zelnik’s. Hacker then proceeds to forward project this series to each census in order to provide an estimate of net census underenumeration. Hacker’s work results in a series of graphs and tables. His first graphs show net under-enumeration totals in both males and females.

By using IPUMS samples, Hacker is able to cross-tabulate the population by single years of age, sex, and birthplace in order to construct state, regional, and sectional series of births. Hacker proceeds to track net census under enumeration, however this time he divides the tables into northern-born and southern-born individuals. His results suggest that the 1890 Census Office greatly overestimated enumeration errors in the 1870 census count of the South.

---

36 Hacker 83-84
37 Hacker 84-85
38 Hacker 85
David Hacker’s findings paint a drastically different portrait of the Census of 1870 when compared to Coale and Zelnik. His description of the 1870 census’s bad reputation as “undeserved” stands in stark contrast with prevailing thought at the time. The 1890 Census Office’s adjusted estimate of the 1870 population has become the “official figure.” No census in history has been able to account for all members of the population. George Washington was correct when he contended that the “real number of inhabitants will greatly exceed the official return.” However, Hacker has been able to fight back against a specific belief, about a specific census, by using specific data collected by IPUMS. Hacker has fought back against the racist rhetoric employed by the 1890 Census Bureau.

Enumerators who surveyed the South during the reconstruction era took on an impossible task and were then scapegoated by those in positions of power. It is the duty of historians who study the census to fight back against people like Francis A. Walker, who would blame the failings of a post-war census on minority groups rather than taking an honest look at the findings of the enumerators.

There was a massive public outcry after the 1890 Census was released.\textsuperscript{39} The initial reaction of the press was positive, in large part to Hermann Hollerith’s tabulating machine. This would not last. After the numbers were released many in America criticized the growth rate in the census as being much smaller than the reality. Francis A. Walker pointed out that more than 5 million immigrants had immigrated to the United States and that if the growth rate was to be believed then there must have been a

\textsuperscript{39} Anderson 109
significant decrease in the American birth rate. In 1891, Walker had begun to politically align himself with a “virulent, racist immigration restriction movement.” Walker believed that American society must be saved from the “degradation” of Europe. These abhorrent political beliefs were held by the most powerful man working for the census at the time.

Work like the kind Hacker was able to produce is extremely powerful in this era of plentiful information. The Census of 1870 is notable politically due to its role as the first full “headcount” of the United States since the end of the American Civil War. Its role in illuminating the American South in the aftermath of the bloody conflict could have been immense, however the census of 1870 had been widely believed to be inaccurate. Criticisms of this census had their roots in the widespread racism inside of America at the time.

40 Anderson 108
Alexandria Ocasio-Cortez and the Reconstruction era

New York Representative, Alexandria Ocasio-Cortez, while in conversation with award winning author, Ta-Nehisi Coates, once controversially stated that, “Algorithms are still made by human beings, and those algorithms are still pegged to basic human assumptions. They’re just automated assumptions. And if you don’t fix the bias, then you are just automating the bias.” Ocasio-Cortez is correct. While Francis A. Walker did not use a computer algorithm to come to the conclusion that the 1870 Census was poorly enumerated, he did allow his racial biases to cloud his judgement and not consider all the facts. In many ways, the 1870 Census and the resulting pushback against its findings mirror the development of the United States during the reconstruction era.

In 2017, Ta-Nehisi Coates released a collection of essays and titled the work We Were Eight Years in Power. This is a quote by Representative Thomas E. Miller who had been asking why white Southerners hated African Americans after all the good they had done during the Reconstruction Era. Through his work, Hacker has added the Census of 1870 to the list of great accomplishments by African-Americans that had been slandered after the end of the Reconstruction period in the American South. The IPUMS project has given historians all over the world access to the important data stored in the census in a format that is easily accessible to the modern historian.

Estaff, Remezcla
In the past, historians working with Census data had to either work with the original records or rely on other work done by historians that came before. Working with the original records can be time consuming as the often hard to read handwriting of previous enumerators can slow down any sort of large-scale study of a particular census. Relying on work performed by other historians can speed up this process but Hacker has shown us that even some of the most well-respected census historians such as Ansley J. Coale and Melvin Zelnik can be inaccurate in their findings. Another praise-worthy aspect of the IPUMS is that the data is provided in file formats that are easily accessible by computer programs. The data is clean and compact with an emphasis placed on keeping the file sizes low and easily navigable. Historians who wish to use the data recorded by the IPUMS project do not need high-tech top of the line equipment such as Herman Hollerith’s tabulating machine.

All kinds of different devices can work with the data that the IPUMS project can provide. This has removed the barrier to entry that many have faced in the past when attempting to work with data recorded in the census.
Thomas Jefferson and George Washington both agreed, population growth proved economic strength. This thought was passed on to other politicians and notables in the country. Cities would compete against one another during every census year. During the census of 1870, after their population was reported to be smaller than expected, the city of Philadelphia ordered a recount of their city.\textsuperscript{42} The recount would be held in the winter as the city expected more of its residents to be inside during this time. This is an example of the tactics employed by different cities in the United States in order to get a leg up on one another. Data from the census can be unreliable due to this uneven count.

In order to address Case Study 1, which examines the movement of people across the American continent from rural areas to urban during the Industrial Revolution, the question must be asked, “What is rural and what is urban?” At times, spaces have been relabeled with the intention of changing census outcomes. For example during the census of 1870, Indianapolis transformed a rural space on the census into an urban one. “Indianapolis city boosters were dismayed to discover that the 1870 census placed its population far short of a hoped-for 50,000. The recount increased the city’s population 18.5 percent but only because the city annexed land in the fall of 1870 and used the redrawn boundaries for the second enumeration.” Indianapolis clearly transformed the surrounding landscape of its city for population gain.

\textsuperscript{42} Anderson 94
on the census. Spaces that were once considered rural areas had been transformed into urban ones in the eyes of the census.40

Cities grew in size and in scope due the Industrial Revolution’s uplifting of American life. The streets of these cities broadened and houses were made smaller as more and more people flocked to urban spaces at the turn of the century. Rural areas also experienced a changing atmosphere as the rural economy changed. This is especially true in the Southern states as the end of the Reconstruction era coincided with a rise of sharecropping practices in the South. One prominent change to rural life was the move away from farms and ranches.

In 1790, 90 percent of Americans lived on a farm. Presently, only 2 percent of Americans live on farms. The rural population in America had been primarily centered around farms in the past. The changing American landscape was also affected by the high levels of immigration that went on during this time as well. A not insignificant percentage of the new residents in urban spaces were immigrants from other countries. In order to meet the housing needs of all the new arrivals, American cities had to adapt and change. While the transformation was not as severe as what many European cities had to go through, American cities had to redraw their streets, for example.

New York City is a prime example of this transformation. Henry Tappan once said, “he who erects his magnificent palace on Fifth Avenue to-day has only fitted out a future boarding-house and probably occupied the site of a future warehouse.”43 This quote is in reference to New York City and the now famous destruction of the Vanderbilt

43Benjamin 141
Mansion on Fifty-Eighth and Fifth as evidence of the rich people of New York City being displaced during the massive influx of new people into the city due to the Second Industrial Revolution. 44

The urban-bound migrants did not just displace the rich in New York City, many poor neighborhoods were demolished in order to make room for new developments that were deemed necessary for the city. Tappan’s quote implies that anyone who decides to build their home in the middle of a busy city should not expect that this place will survive into the future of the city and all the many inhabitants that will require support from the city. Boarding houses were erected to provide new workers and inhabitants shelter, and warehouses were built to store the products created inside of the city.45

In 1900, the census ran smoothly under the backdrop of the presidential election. In 1901, President Theodore Roosevelt suggested that the Census Bureau become a permanent addition to the government.46 Realizing that his office should focus on convincing the United States Government to make the census department permanent, Chief Statistician Walter Wilcox hired many statisticians in order for them to create work that would testify to the “possibilities for analysis” if a permanent office were to be created. Many of these predominantly male statisticians would go on to become prominent figures and policymakers, including names such as Wesley Mitchell, Allyn Young, Thomas Sewall Adams and W.E.B Du Bois.47 One of these men in particular,
W.E.B Du Bois, recently had his population data research transformed and colorized into charts in order to be displayed at the famous *Exposition Universelle* in Paris.

**The Exposition Universelle**

Du Bois’s graphs were created in the city of Atlanta during 1900. The intent for these graphs was for them to be displayed at the *American Negro Exhibit* at the *Exposition Universelle* which was held in Paris. These hand drawn graphs and charts were put on display first in Paris and then they were sent to many different world fairs across the world. Du Bois’s charts were paired with real images of life as an African American during the turn of the century in America.  

There were two set of infographics created for the *American Negro Exhibit*. The first infographic was part of Du Bois’s *The Georgia Negro: A Social Study*. Out of all the states, Georgia had the largest population of African Americans and Du Bois and his team had been using the diverse population of Georgia in order to demonstrate the progress made by African Americans since the Civil War. Du Bois attempted to establish a claim to global modernity by African American South through these colorful infographics.

The second set of infographics prepared by Du Bois and his team had a larger scope. Instead of simply focusing on Georgia, these infographics were more national and global in terms of scope. This set was titled *A Series of Statistical Charts Illustrating the Condition of Descendants of Former African Slaves Now in Residence in the United States of America*. A long title, this set of infographics rendered statistics that shed light

---

48 Battle-Baptiste and Rusert 23
49 Battle-Baptiste and Rusert 24
on national employment and education, population distributions of African Americans, literacy rates relative to other countries and more. 50

Through the creation of these charts and graphs, Du Bois was able to express his theory of “double consciousness.” Double consciousness was a term used by Du Bois in order to describe “the experience of always seeing oneself through the eyes of another—a psychic alienation and social isolation produced by the ‘peculiar’ condition of being black in America.” 51 Du Bois thought of himself as both American and African American. Through double consciousness, Du Bois was able to find a sort of “second sight” that could be “transformed from a curse into a gift.” 47 The use of the visual sense in the American Negro Exhibit expresses this belief in double consciousness by Du Bois. In addition to the graphs and charts, the American Negro Exhibit would also showcase photo albums depicting life in the American South. Du Bois turned to a visual medium in order to express his and many other South American’s life experiences to a foreign audience which had gathered across the Atlantic. 52

Du Bois created these graphs with the help of his many talented assistants. Du Bois was one of the first professors in the United States to train his students in sociological theory and empirical methodologies. Du Bois and his students were well aware of the influence that Social Darwinist thought would have on the Exposition Universelle. Their graphs visualized data which stood in stark contrast to many trains of thought that were present in Social Darwinist groups in Europe at the time. 53
One example of Du Bois’s graphing work opposing social darwinism were his graphs that showed how the African American population was growing rather than decreasing. A prominent thought held by Social Darwinists at the time was that the population of African Americans in the United States was decreasing and that a population decrease signals that African Americans were inferior to people from other races. Du Bois defeated this thought through the use of visual displays of statistics that were easy to digest for uninformed consumers.  

The *Exposition Universelle* was going to be viewed by more than just academics, many Europeans from all walks of life would be in attendance. Du Bois and his students prepared for this however, and had prepared all of their charts to be coated in color. The use of color by Du Bois and his students to create visual sociological charts was a rarity at the time. The exhibit begins with a map showcasing the movement of the African people across the Atlantic due to the Slave Trade. Du Bois and his students also took careful lengths to introduce Europeans to Georgia as many did not know where that state was on a map.

Francis Walker had faced a similar situation when creating his *Statistical Atlas*. In order for his work to be a success, the *Statistical Atlas* needed to be presented in a format that was digestible by an everyday person. Just like Walker’s *Statistical Atlases* and their centers of population, Du Bois was able to clearly visualize statistics to Americans and people all over the world that the African American in the American South population was thriving and increasing. There was also an attempt to link their

---

54 Battle-Baptiste and Rusert 35  
55 Battle-Baptiste and Rusert 34-36
data to European countries. Du Bois’s exhibit displayed information comparing literacy rates in African American populations and countries like Russia and Hungary. Russia was found to have slightly higher literacy numbers than African Americans in the South whereas Hungary was found to have the same level of literacy in its population as African Americans living in the South.  

A Columbia University professor named Timothy Mitchell once said, “It was not always easy in Paris to tell where the exhibition ended and the world began.” The exhibition in Paris featured many exhibits that would paint people from the African continent as “primitive” or “savage.” The impact that the American Negro Exhibit would have on the European attendees cannot be understated. The exhibit stood in contrast to predominantly racist beliefs that were held by white Europeans at the time.  

A specific method to express Du Bois’s statistical work was cartography. Cartography was used by Du Bois in the American Negro Exhibit due to the role the field played in European society at the time. Europeans did not use cartography simply to help themselves navigate the world. Europeans also turned to cartography in order to chart their conquests. By the time Du Bois was born, Europeans had come into contact with diverse groups of people from all over the world. The creation of these maps and charts, which were based on accounts of European colonizers, had perverted academics and historians in Europe to promote theories of racial difference that were based in geography. By linking racial difference to geography and climate, Europeans

---

56 Battle-Baptiste and Rusert 39
57 Battle-Baptiste and Rusert 39-40
had placed themselves at the top of the global totem pole and began to view themselves as “the vanguard of civilization.” 58

In both sets of infographics, Du Bois and his students “redeployed” these methods of cartography. The very first graphic presented by Du Bois and his students was a map of the world. On this colorful map, two circles have been drawn. Eurasia and Africa are contained in one circle and the Americas reside in the other. Lines were drawn and regions were darkened. The lines represented the routes taken across the Atlantic during the African Slave Trade and the darkened areas represented population numbers. There is also a solitary white star present in the map, representing the state of Georgia.

Cartography continues to be present in the second graphic presented at the exhibit. The second graphic is a state map of the United States. Titled “Relative Negro Population of the States of the United States, this chart provides population data for each of the states. Dark colors and line patterns were chosen for states with the highest population numbers and light colors without line patterns were chosen for states with the lowest population numbers. Red and brown, two of the most striking colors utilized, were chosen to represent states with population numbers that were close to the average. 59

Rural workers moving to urban areas for new career opportunities account for only a portion of the demographic makeup inside of the growing population hubs in the United States. Immigrants accounted for a substantial amount of new growth as well.

58 Battle-Baptiste and Rusert 40-41
59 Battle-Baptiste and Rusert 41
inside of these urban areas. Urban areas were also redefined as time went on. One city’s borders might change drastically from census to census. Indianapolis was already referred to as an example of the practice of adding land to a city’s borders in order to raise the population count. Indianapolis annexed land in surrounding areas after being approved for a recount due to the city government’s dissatisfaction with the initial population count performed by census enumerators. Land which had once been defined as rural was now considered urban.

The American workforce transformed dramatically during this time. The Agriculture industry was in a class of its own in terms of industry in America. Industrialization inside of America can be seen best when examining more industries than just the most commonly worked. The American workforce diversified, new jobs were being created and filled. By 1930, an ordinary American could choose between more careers than in 1870 due to more skilled labor being required. Technological advances had created demand for many new industries and goods.
The Charts

Utilizing the 1% sample of censuses from 1870 to 1930, population data from the Census of 1870 reports 25 percent of Americans living in urban areas.\textsuperscript{60}

\textsuperscript{60}Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek
The census of 1880 has roughly 27 percent of Americans living in urban areas with around 73 percent of Americans living in rural ones.
By 1900, 38 percent of Americans were living in urban areas compared to the 62 percent that resided rural areas.
That percentage of urban dwellers would grow to 45 percent in **1910** with the rural population percentage falling to 55 percent.
The census of 1920 is the first census to feature a majority urban population of 50.3%. While this would soon become a mainstay of American life, this would be the first time in history that the city would be the place where most Americans live.
1930 reported the greatest percentage of Americans living in urban spaces with 56 percent of Americans being reported to live in urban areas;

This percentage stands in stark contrast to the 25 percent of Americans reported to be living in urban areas by the census of 1870. There is a clear shift in American life that is detailed through this simple checkbox on the census.
When examining homeownership percentages across America in both rural and urban areas, it is found that ~25% of the population rented homes in urban areas whereas ~35% of the population owned homes in rural areas in 1900.
When examining homeownership percentages across America in both rural and urban areas, it is found that ~28% of the population rented homes in urban areas whereas ~31% of the population owned homes in rural areas in 1910.
When examining homeownership percentages across America in both rural and urban areas, it is found that ~30% of the population rented homes in urban areas whereas ~27% of the population owned homes in rural areas in 1920.
When examining homeownership percentages across America in both rural and urban areas, it is found that ~31% of the population rented homes in urban areas whereas ~22% of the population owned homes in rural areas in 1930.
Mill work, restaurants and hotels experienced lots of growth inside of the state of Alabama when tracked from 1870 to 1930.
Hospitals, Movie Theaters and clothing industries experienced large growth from 1870 to 1930 inside of the state of California.
Hospitals, Laundromats and the Accounting industry experienced large growth from 1870 to 1930 inside of the state of New York.
Railroads, printing/publishing and the postal industry experienced large growth from 1870 to 1930 inside of the state of Texas.
During the census of 1900, Construction was the most popular industry in Los Angeles.

During the census of 1910, Construction would remain the most popular industry in Los Angeles.
During the census of 1920, Construction continued to be the most popular industry in Los Angeles.

During the census of 1930, Construction maintained its position as the most popular industry in Los Angeles.
During the census of 1900, the Private Household industry was the most populated industry inside of New York City.

During the census of 1910, the Construction industry would move up to become the most populated industry inside of New York City.
During the census of 1920, Apparels and Accessories would become the most popular industry inside of New York City.

During the census of 1930, the Construction industry would return as the most populated industry inside of New York City.
~45% of Los Angeles’s population was either a first or second generation immigrant in 1900.
~48% of Los Angeles's population was either a first or second generation immigrant in 1910.
~47% of Los Angeles’s population was either a first or second generation immigrant in 1920.
~47% of Los Angeles’s population was either a first or second generation immigrant in 1930.
~77% of New York’s population was either a first or second generation immigrant in 1900.
~80% of New York’s population was either a first or second generation immigrant in 1910.
~78% of New York’s population was either a first or second generation immigrant in 1920.
~75% of New York’s population was either a first or second generation immigrant in 1930.
~54% of the United State's population was either a first or second generation immigrant in 1900.
~52% of the United State's population was either a first or second generation immigrant in 1910.
~49% of the United State’s population was either a first or second generation immigrant in 1920.
~44% of the United State's population was either a first or second generation immigrant in 1930.
One of the more experimental charts, this “Least Growth” graph tracks which industries showcased the largest percent decrease from 1910 to 1930. Liquor Stores experience -5% growth during this time due in large part to the prohibition of alcohol that occurred during this time period.
Program Analysis

1. years = [1900, 1910, 1920, 1930]
2. for x in years:
3.    urban, rural = urban_or_rural(x)
4.    graph_it_UR(urban, rural, x)

Featured above is the main method used to construct the urban vs rural pie charts showcased previously. The main method calls upon two methods, graph_it and urban_or_rural. These two functions are used to create charts for each of the year values stored inside of the years array. Graph_it_UR is featured below and is used to construct the pie charts.

1. def graph_it_UR(urban, rural, year, check=True):
2.    labels = ['Urban Population', 'non-Urban Population']
3.    sizes = [urban, rural]
4.    colors = ['#99ff99', '#ffcc99']
5.    explode = (0.1, 0.0)
6.    plt.pie(sizes, colors = colors, labels=labels, autopct='%1.1f%%',
7.             startangle=90, pctdistance=0.85, explode = explode)
8.    centre_circle = plt.Circle((0,0),0.70,fc='white')
9.    fig = plt.gcf()
10.   fig.gca().add_artist(centre_circle)
11.   plt.axis('equal')
12.   plt.title('Percentage of Urban and Non-Urban Populations in the Census of ' + str(year))
13.   if check == True:
14.       plt.show()

These percentages were put into pie charts by each census year. The pie charts were created using Python3 and MatPlotLib. The function featured above, graph_it_UR, utilizes MatPlotLib to create the pie charts.
Two For Loops were used in order to extract census data from the IPUMS extract. These for loops will be described in depth later on in this paper. Obtaining the raw count of the total number of Americans living in rural areas and also obtaining a raw count of the total number of Americans living in urban ones, the program proceeds to create pie charts for each census from 1870 to 1930. Showcased below is the function urban_or_rural which returns the urban and rural population counts for a specific year’s census.

```python
1. def urban_or_rural(year):
2.     yearLoc = indLoc = 0
3. 4.     filename = 'census02.txt'
5.     lines = open(filename).read().splitlines()
6.     line0 = lines[0].split('	')
7.     industryCount = {}
8. 9.     #create list
10. 11.     for x in range(len(line0)):
12.         # print(line0[x].strip())
13.         if line0[x] == '"YEAR"':
14.             yearLoc = x + 1
15.         if line0[x] == '"URBAN"':
16.             otherLoc = x + 1
17. 18.     urban = rural = 0
19.     for x in range(len(lines)):
20.         if x != 0:
21.             currentLine = lines[x].split('	')
22.             if int(currentLine[yearLoc]) <= year:
23.                 if int(currentLine[yearLoc]) == year:
24.                     if currentLine[otherLoc] == "1":
25.                         rural = rural + 1
26.                     elif currentLine[otherLoc] == "2":
27.                         urban = urban + 1
28.                     else:
29.                         continue
30.                 else:
31.                     break
32.     return urban, rural
```

Charts were also created in order to track the industries in which Americans were most commonly employed i.e. Agriculture and Petroleum. Python3, MatPlotLib and the IPUMS data extract XML file were used in order to create these charts. Statistics were
taken on both overall count and percentage of working population. Excluding records
that mark a citizen’s industry as “N/A”, it is found that Agriculture jobs are the most
worked in the United States. Records that mark the Industry question as N/A were taken
out of the total count while calculating percentages.

In order to calculate the percentages and create the graphs, the raw IPUMS data
had to be decoded using the .xml file provided. Due to IPUMS.org not supporting
Python3, a series of custom functions had to be created in order to decode the data by
reading the xml files. Featured below is the code used to navigate and then create the
data labels that were provided by IPUMS.

```python
1. import xml.etree.ElementTree as ET
2. def data_Label(value, ident="IND1950"):
3.     label = str(xml_Reader(value, ident))
4.     if len(label) > 20:
5.         token = label.split()
6.         if len(label) % 2 == 0:
7.             index = int(len(token)/2)
8.             str1 = token[:index]
9.             str2 = token[index:]
10.            str1.append('
')
11.            finalStr = str1 + str2
12.            return ' '.join(finalStr)
13.        else:
14.            index = int(len(token)/2)
15.            str1 = token[:index]
16.            str2 = token[index:]
17.            str1.append('
')
18.            finalStr = str1 + str2
19.            return ' '.join(finalStr)
20.        else:
21.            return label
22.     return label
23. 
24. def xml_Reader(value, ident="IND1950"):
25.     tree = ET.parse(ident+'.xml')
26.     root = tree.getroot()
27.     for var in root:
28.         if var.find("catValu") is not None:
29.             if var.find("catValu").text == str(value):
30.                 return var.find("labl").text
31.             if var.find("labl").text == str(value):
32.                 return var.find("catValu").text
```

72
The 5 most common industries were recorded and then displayed in a series of bar charts.

Examining the most common industries in the United States it is found that Agriculture jobs remain the most worked throughout the country in censuses from 1870 to 1930. Though the Agriculture industry will steadily decrease in percentage, it will remain the most common industry in America through the 1930s. In 1870, 16.1 percent
of Americans worked in Agriculture, that percentage would steadily decrease to 8.7 percent of Americans in 1930.

In the 1800s, a considerable number of Americans worked in private households helping with tasks, this would be the second most populated industry in censuses in the 1800s and earlier censuses in the 1900s with roughly 2.5 percent of Americans working in this field during 1870. That percentage sustains a small drop of .3 from 1870 to 1880 before falling to 2.1 percent in 1900. Private households would sustain a small jump to roughly 2.5 percent in 1910 but this would fall by 1930 to less than 2 percent.

The Construction industry maintained a presence in the top 5 most common industries throughout the census years examined. Construction jobs experienced a rise in prominence during the turn of the century in America. In 1870, 1.5 percent of Americans reported working in the Construction industry. This percentage jumps to
roughly 2.5 percent by 1930, a not-insignificant increase for the field. Another industry that experiences a rise in prominence during the early 1900s is the Railroads and Railways industry.

By 1920, almost 2 percent of Americans worked on Railroads and Railways around the country. This industry would be the second most populated industry in the census of 1920. Educational jobs also experience a rise in prominence. While earlier censuses report less than 1 percent of Americans working in Education, by 1920 Education jobs would be the 5th most reported industry in the census of 1920. By 1930, 1.2 percent of Americans would be working in Educational service roles across the country. By 1910, more than 70 percent of American children were attending school. There was also a rise in private schools across the country during the turn of the
Non-dairy food store jobs are reported once to be in the top 5 most worked in industries. The appearance of the “Food stores, except dairy” industry happens just once on the 1910 census. Nearly 1 percent of Americans reported working in this field at the time. Agriculture’s dominance over the American economy cannot be understated as it remained the only industry to have a double digit percentage of Americans working in the field through the 1930s. While it does maintain its standing over the rest of the American economy, the Agriculture industry does also report the largest drop in
percentage of Americans employed out of any industry reported on the census.

Additional pie charts were created for each census from 1870 to 1930. These pie charts tracked rural and urban percentages in America but subdivides both rural and urban households by whether or not the household is a farm. Pie charts were created the same way as before with Python3 and MatPlotLib. The only difference between this program and the urban and rural program is that additional statistics about farms are counted.
The census of 1870 reported that 41 percent of Americans lived on farms whereas 59 percent of Americans reported to be not living on a farm. Of that 59 percent, 34 percent lived in rural areas and 24 percent lived in urban ones.

1880’s census would report a rise in Americans living on farms. This rise could possibly be a side effect of the normalization of life after the end of the American Civil War. The Civil War had plunged many areas into instability and farms all across the American South had been destroyed. The 1880 statistics reported that 57 percent of Americans did not live on farms, with 31 percent of this number living in rural spaces.
and 26 percent living in urban ones.

1900 would report a significant, 5 percent, decrease in the number of Americans living on farms. During this time, 38 percent of Americans would report to live on farms compared to roughly 61 percent of Americans who did not live on farms. Of this 61
percent, 23 percent lived in rural spaces and 38 percent reported to be nonrural.

The census of 1910 reported farms to make up 33 percent of the American population's homes. 67 percent of Americans reported to not be living on farms with 23
percent living in rural areas and 45 percent reporting to live in urban ones.

1920 would show farms to account for even fewer households in America. Only 30 percent of Americans in 1920 were marked as living on farms. At this time, 50 percent of Americans would record themselves as living in non-farm urban spaces and
another 20 percent would report to live in non-farm rural spaces.
The census of 1930 paints a much different picture than the census of 1870. Rural farms would account for less than 25 percent of the population’s household. Non-farm households would account for 75 percent of the population living in the United States at the time. 56 percent of these non-farm households would be located in urban areas and another 19 percent would report to live in non-farm households located in rural spaces.

The population of people recorded to be living on farms located in urban spaces would make up less than one percent of the population in all census years from 1870 to

---

1930. The census of 1900 would report the largest percent, 0.8 percent, of the American population to be living on farms in urban spaces.

The census data that was used for this project is stored in a tab-delimited text file. The code used to traverse this data in order to record statistics consists primarily of two for loops. These loops remain consistent however there are some slight differences inside of these loops depending on what kind of data is being recorded.

Each line of the text file represents a different record and each tab inside of an individual line represents different data recorded about that record for the census. Each new line, except for the initial line, in the text file represents a different record found in the sample. The initial line describes the data found in all the proceeding lines with each of its tab delimited values describing the data that is found in all of the records. Some data is not recorded for all the censuses, this is due to the censuses asking different questions. If data was not recorded for a specific record then it is given the value N/A.

```python
for x in range(len(line0)):
    if line0[x] == '"YEAR"':
        yearLoc = x + 1
    if line0[x] == '"URBAN"':
        otherLoc = x + 1
```

Featured above is the initial loop. This loop traverses the initial line of the text file. The goal of this first loop is to find the locations of the answers to the questions that are being searched for. Each line of the text file is an individual record and every tab signifies a new question. The location of a question in the first line of the file will be the location of the data recorded for that question for all the lines following in that specific file. The initial loop’s purpose is to find this location value in order to provide accurate statistics.

```python
for x in range(len(lines)):
```
2. if x != 0:
3.     currentLine = lines[x].split('	')
4.     if int(currentLine[yearLoc]) <= year:
5.         if int(currentLine[yearLoc]) == year:
6.             if currentLine[otherLoc] == "1":
7.                 rural = rural + 1
8.             elif currentLine[otherLoc] == "2":
9.                 urban = urban + 1
10.            else:
11.                continue
12.        else:
13.            break

Featured above is an example of the second loop being used to count urban
versus rural population percentages. This loop records and categorizes the census data
that is being searched over. Skipping over the initial line in the text file, the second loop
goes through all records that are of a census year that is less than or equal to the
specific year being searched for. If the loop reaches records that are a part of the target
census year, the loop will then proceed to take whichever statistics are needed to create
the chart. The above example counts whether or not a record is marked as urban or
rural. For statistics that require more than one census year’s information to be recorded,
the second loop runs through the entire text file. The loop checks the census year for
each record in the text and records information on that year depending on whether or
not the census year is one of those included in the search. Conditional if statements are
used to determine which data from which census will be included in the statistical
survey.

In an effort to keep file sizes low, IPUMS data is presented to users almost
always numerically. This is primarily used for data that would logically be saved as a
string such as data which has to do with location. Instead of storing an entire string,
IPUMS instead uses numerical codes which are able to be deciphered by the computer.
In order for the computer to decipher these values, IPUMS provides custom XML files for each request it receives. These custom XML files provide variable labels for all the data provided in the approved requests. For example; each state has been given a numerical “ICP” value. Counties are also given similar values. When examining the text file, users will see a two digit number recorded for each record which represents the state that the individual lives in.

What follows after these loops varies from program to program. There are commonalities between the programs however with all of the programs including code to create the graphs and charts. Python and Matplotlib were used to create charts such as the urban and rural distribution, farm household percentages and industry popularity percentages. In order to obtain the variable labels for each record, the correct XML file is searched and the variable labels are obtained. This allows for the graphs to informatively display the data provided in each of the census records. After obtaining the variable labels, the specific graph is created. Key, title, and axis labels are also applied to the graph.

The programs used to extract the data and create the graphs involve two for loops which span two different lengths, therefore these programs run in O(m+n) time with m being the number of records and n being the amount of information held inside of each record. Depending on what graph or chart is being created this number may change as it is occasionally necessary to take statistics on the data which will cause runtime to increase.
The IPUMS project shares many similarities with Hermann Hollerith’s electric tabulating machines. IPUMS exists in a much different age and confronts different problems than Hollerith’s tabulating machine but at their essence, both machines have similarities with one another. Some similarities include the use of identifier numbers rather than something more conventional like a name, in order to tell individual records apart.

Nothing on the cards included information about the names of the individuals, not too dissimilar to data delivered by IPUMS. Rather, punch cards were given a unique number in order to identify the card. This number could then be used to identify which record the punch card was representing. If it were necessary for a researcher to find an individual record’s name Hollerith and the other designers assumed that researchers would just use the identifier number to find an individual’s name.

IPUMS data extracts do not include names on records rather the data extract chooses to give each record a unique number to let it stand out from the rest of the records. This method of using identifier numbers in order to find specific information about individuals, such as a name, is extremely similar to the one found in Herman Hollerith’s electric tabulating machines. Unlike Hollerith’s machine, record numbers can not be referenced back to an individual so there is no way of finding an individual’s name. By failing to find a way to transfer the name of an individual onto the punch cards, census officials had expanded the error of failing to provide for conventional copies of the census to be preserved.
Final Thoughts

At the time of this paper’s writing, the Covid-19 Pandemic has likely corrupted the 2020 American Census. Individuals recorded on the census may pass away right after. It will be until 2030 that the United States census will be able to record the damage.

The great failing of the American censuses recorded from 1870 to 1930 is the way Native Americans are handled. After the end of the Civil War, the United States of America was involved in a major undeclared war against Native American tribes across the North American continent. Genocide lives at the core of these censuses. The final chart displayed is a bar chart, created by Dartmouth researcher Josh Pearl. Pearl showcases an apparent “rise” in Native American population numbers found inside of United States census data collected from IPUMS.org.

![Native American Population 1850 - 1950](chart.png)

---

62 Pearl, Josh “History 90.01: Topics in Digital History.”
functions.py

from xml_reader import data_label
from collections import Counter
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from least_common import least_common_values, Reverse
import matplotlib.ticker as mtick
import random
from graph_it import graph_it_common_city

def most_common(year):
    yearLoc = indLoc = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('\t')
    industryCount = {}
    #create list
    for x in range(len(line0)):
        #print(line0[x].strip())
        if line0[x] == '"YEAR"':
            yearLoc = x + 1
        if line0[x] == '"IND1950"':
            indLoc = x + 1
    for x in range(len(lines)):
        #print(lines[x])
        if x != 0:
            # increment list location by 1 per currentLine
            currentLine = lines[x].split('\t')
            if int(currentLine[yearLoc]) <= year:
                if int(currentLine[yearLoc]) == year:
                    if currentLine[indLoc] in industryCount:
                        industryCount[currentLine[indLoc]] += 1
                    else:
                        industryCount.update({currentLine[indLoc]:1})
            #print(currentLine)
            else:
                break
    k = Counter(industryCount)
    fiveHigh = k.most_common(6)
    fiveHigh.pop(0)
    totalRec = sum(industryCount.values())-industryCount['0']
    indStats = []
    ylablepercent = []
    strName = []
    for i in fiveHigh:
        strName.append(str(data_label(i[0])))
        indStats.append(round(((i[1]/totalRec)*100),1))
        ylablepercent.append(str(round(((i[1]/totalRec)*100),1)) + '%')
    return indStats, ylablepercent

def most_common_city(year, city):
    yearLoc = indLoc = 0
filename = 'census02.txt'
lines = open(filename).read().splitlines()
line0 = lines[0].split('\\t')
industryCount = {}
# create list

for x in range(len(line0)):
    print(line0[x].strip())
    if line0[x] == '"YEAR"':
        yearLoc = x + 1
    if line0[x] == '"IND1950"':
        indLoc = x + 1
    if line0[x] == '"CITY"':
        cityLoc = x + 1

city_num = int(data_Label(city, "CITY"))

for x in range(len(lines)):
    # print(lines[x])
    if x != 0:
        # increment list location by 1 per currentLine = lines[x].split('\\t')
        if int(currentLine[yearLoc]) <= year:
            if int(currentLine[yearLoc]) == year and int(currentLine[cityLoc]) == city_num:
                if currentLine[indLoc] in industryCount:
                    industryCount[currentLine[indLoc]] += 1
                else:
                    industryCount.update({currentLine[indLoc]:1})
            else:
                break
    else:
        break

k = Counter(industryCount)
fiveHigh = k.most_common(6)
fiveHigh.pop(0)
totalRec = sum(industryCount.values())-industryCount['0']
indStats = []
ylabelpercent = []
strName = []
for i in fiveHigh:
    strName.append(str(data_Label(i[0])))
    indStats.append(round(((i[1]/totalRec)*100),1))
    ylabelpercent.append(str(round(((i[1]/totalRec)*100),1)) + '%')

return indStats, ylabelpercent, strName


def most_common_state(year, state):
    yearLoc = indLoc = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('\\t')
    industryCount = {}
    # create list

    for x in range(len(line0)):
        print(line0[x].strip())
        if line0[x] == '"YEAR"':
            yearLoc = x + 1
        if line0[x] == '"IND1950"':
            indLoc = x + 1
        if line0[x] == '"CITY"':
            cityLoc = x + 1

        city_num = int(data_Label(city, "CITY"))

        for x in range(len(lines)):
            # print(lines[x])
            if x != 0:
                # increment list location by 1 per currentLine = lines[x].split('\\t')
                if int(currentLine[yearLoc]) <= year:
                    if int(currentLine[yearLoc]) == year and int(currentLine[cityLoc]) == city_num:
                        if currentLine[indLoc] in industryCount:
                            industryCount[currentLine[indLoc]] += 1
                        else:
                            industryCount.update({currentLine[indLoc]:1})
                    else:
                        break
            else:
                break

        k = Counter(industryCount)
fiveHigh = k.most_common(6)
fiveHigh.pop(0)
totalRec = sum(industryCount.values())-industryCount['0']
indStats = []
ylabelpercent = []
strName = []
for i in fiveHigh:
    strName.append(str(data_Label(i[0])))
    indStats.append(round(((i[1]/totalRec)*100),1))
    ylabelpercent.append(str(round(((i[1]/totalRec)*100),1)) + '%')

return indStats, ylabelpercent, strName
if line0[x] == '"IND1950"':
    indLoc = x + 1
if line0[x] == '"STATEICP"':
    cityLoc = x + 1
city_num = int(data_Label(city, "STATEICP"))
for x in range(len(lines)):
    # print(lines[x])
    if x != 0:
        # increment list location by 1 per
        currentLine = lines[x].split('"t"
        if int(currentLine[yearLoc]) <= year:
            if int(currentLine[yearLoc]) == year and
                int(currentLine[cityLoc]) == city_num:
                if currentLine[indLoc] in industryCount:
                    industryCount[currentLine[indLoc]] += 1
                else:
                    industryCount.update({currentLine[indLoc]:1})
                    #print(currentLine)
            else:
                break
k = Counter(industryCount)
fiveHigh = k.most_common(6)
fiveHigh.pop(0)
totalRec = sum(industryCount.values())-industryCount['0']
indStats = []
ylabelpercent = []
strName = []
for i in fiveHigh:
    strName.append(str(data_Label(i[0])))
    indStats.append(round(((i[1]/totalRec)*100),1))
    ylabelpercent.append(str(round(((i[1]/totalRec)*100),1)) + '%')
return indStats, ylabelpercent, strName

def least_common_ind(year):
    yearLoc = indLoc = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('"t"
    industryCount = {}
    #create list
    for x in range(len(line0)):
        #print(line0[x].strip())
        if line0[x] == '"YEAR"':
            yearLoc = x + 1
        elif line0[x] == '"IND1950"':
            indLoc = x + 1
    for x in range(len(lines)):
        # print(lines[x])
        if x != 0:
            # increment list location by 1 per
            currentLine = lines[x].split('"t"
            if int(currentLine[yearLoc]) <= year:
                if int(currentLine[yearLoc]) == year:
                    if currentLine[indLoc] in industryCount:
                        industryCount[currentLine[indLoc]] += 1
                        else:
                            industryCount.update({currentLine[indLoc]:1})
                            #print(currentLine)
                        else:
                            break
k = Counter(industryCount)
fiveHigh = k.most_common(6)
fiveHigh.pop(0)
totalRec = sum(industryCount.values())-industryCount['0']
indStats = []
ylabelpercent = []
strName = []
for i in fiveHigh:
    strName.append(str(data_Label(i[0])))
    indStats.append(round(((i[1]/totalRec)*100),1))
    ylabelpercent.append(str(round(((i[1]/totalRec)*100),1)) + '%')
return indStats, ylabelpercent, strName

industryCount[currentLine[indLoc]] += 1
else:
    industryCount.update({currentLine[indLoc]:1})
    #print(currentLine)
else:
    break
k = Counter(industryCount)
fiveLeast = least_common_values(k, 5)
totalRec = sum(industryCount.values())-industryCount['0']
indStats = []
ylabelpercent = []
strName = []
for i in fiveLeast:
    strName.append(str(data_Label(i[0])))
    indStats.append(round(((i[1]/totalRec)*100),1))
    ylabelpercent.append(str(round(((i[1]/totalRec)*100),1)) + '%')
return indStats, ylabelpercent, strName

def urban_or_rural(year):
    yearLoc = indLoc = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('\t')
    industryCount = {}
    #create list
    for x in range(len(line0)):
        # print(line0[x].strip())
        if line0[x] == '"YEAR":'
            yearLoc = x + 1
        elif line0[x] == '"URBAN":'
            otherLoc = x + 1
    urban = rural = 0
    for x in range(len(lines)):
        if x != 0:
            currentLine = lines[x].split('\t')
            if int(currentLine[yearLoc]) <= year:
                if int(currentLine[yearLoc]) == year:
                    if currentLine[otherLoc] == "1":
                        rural = rural + 1
                    elif currentLine[otherLoc] == "2":
                        urban = urban + 1
                else:
                    continue
            else:
                break
    return urban, rural

def urban_or_rural_ownership(year):
    yearLoc = indLoc = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('\t')
    industryCount = {}
    #create list
    for x in range(len(line0)):
# print(line0[x].strip())
if line0[x] == ""YEAR"":
    yearLoc = x + 1
if line0[x] == ""URBAN"":
    urLoc = x + 1
if line0[x] == ""OWNERSHP"":
    ownLoc = x + 1
print(ownLoc)

ownedUR = rentUR = ownedRL = rentRL = 0
for x in range(len(lines)):
    if x != 0:
        currentLine = lines[x].split('\t')
        if currentLine[ownLoc] != 'NA':
            currentYR = int(currentLine[yearLoc])
            currentUR = int(currentLine[urLoc])
            currentOW = int(currentLine[ownLoc])
            if currentYR <= year:
                if currentYR == year:
                    if currentUR == 2:
                        if currentOW != 0:
                            if currentOW == 1:
                                ownedUR += 1
                            elif currentOW == 2:
                                rentUR += 1
                    else:
                        continue
                else:
                    break
            return ownedUR, rentUR, ownedRL, rentRL

def urban_BPL(year):
    yearLoc = indLoc = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('\t')
    industryCount = {}
    #create list
    for x in range(len(line0)):
        if line0[x] == ""YEAR"":
            yearLoc = x + 1
        if line0[x] == ""URBAN"":
            urbanLoc = x + 1
        if line0[x] == ""BPL"":
            bpl = x + 1
        if line0[x] == ""MBPL"":
            mbpl = x + 1
        if line0[x] == ""FBPL"":
            fbpl = x + 1
        immigrant = non_immigrant = 0
        for x in range(len(lines)):
if x != 0:
    # print('>0')
    currentLine = lines[x].split('	')
        # print('!= NA')
        yr = int(currentLine[yearLoc])
        ur = int(currentLine[urbanLoc])
        bp = int(currentLine[bpl])
        mp = int(currentLine[mbpl])
        fp = int(currentLine[fbpl])
        if yr <= year:
            # print('yr < yr')
            if yr == year:
                # print('yr == yr')
                if ur == 2:
                    # print('ur == 2')
                    if 150 <= bp < 999 or (150 <= fp < 999 or 150 <= mp < 999):
                        # print('immigrant')
                        immigrant += 1
                    else:
                        non_immigrant += 1
                        # print('nonimmigrant')
                else:
                    continue
            else:
                break
        else:
            continue
    else:
        continue
return immigrant, non_immigrant

def city_BPL(year, city_name):
    yearLoc = indLoc = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('	')
    industryCount = {}
    #create list
    for x in range(len(line0)):
        # print(line0[x].strip())
        if line0[x] == '"YEAR"':
            yearLoc = x + 1
        if line0[x] == '"URBAN"':
            urbanLoc = x + 1
        if line0[x] == '"BPL"':
            bpl = x + 1
        if line0[x] == '"MBPL"':
            mbpl = x + 1
        if line0[x] == '"FBPL"':
            fbpl = x + 1
        if line0[x] == '"CITY"':
            city = x + 1
    print(city)
    immigrant = non_immigrant = 0
    city_cat = data_Label(city_name, "CITY")
for x in range(len(lines)):
    if x != 0:
        # print('>0')
        currentLine = lines[x].split('"t"
            # print('!= NA')
            yr = int(currentLine[yearLoc])
            ur = int(currentLine[urbanLoc])
            bp = int(currentLine[bpl])
            mp = int(currentLine[mbpl])
            fp = int(currentLine[fbpl])
            cy = int(currentLine[city])
            if yr <= year:
                # print('yr < yr')
                if yr == year:
                    # print('yr == yr')
                    if cy == int(city_cat):
                        # print('ur == 2')
                        if 150 <= bp < 999 or (150 <= fp < 999 or 150 <= mp < 999):
                            # print('immigrant')
                            immigrant += 1
                        else:
                            non_immigrant += 1
                            # print('nonimmigrant')
                    else:
                        continue
                else:
                    break
            else:
                continue
        else:
            break
    else:
        continue
    return immigrant, non_immigrant

def growth_rate(year1, year2):
    yearLoc = indLoc = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('"t"
    industryCount1 = {}
    industryCount2 = {}
    prCount = {}
    #create list
    for x in range(len(line0)):
        #print(line0[x].strip())
        if line0[x] == '"YEAR"
            yearLoc = x + 1
        if line0[x] == '"IND1950"
            indLoc = x + 1
    for x in range(len(lines)):
        # print(lines[x])
        if x != 0:
            # increment list location by 1 per
            currentLine = lines[x].split('"t"
            if int(currentLine[yearLoc]) == year1:
                if currentLine[indLoc] != '0':
if currentLine[indLoc] in industryCount1:
    industryCount1[currentLine[indLoc]] += 1
else:
    industryCount1.update({currentLine[indLoc]:1})
elif int(currentLine[yearLoc]) == year2:
    if currentLine[indLoc] != '0':
        industryCount2[currentLine[indLoc]] += 1
    else:
        industryCount2.update({currentLine[indLoc]:1})
k1 = Counter(industryCount1)
k2 = Counter(industryCount2)
totalrec1 = sum(industryCount1.values())
n = year2 - year1
for x in industryCount1:
    VPresent = VPast = PR = 0
    if x in industryCount2:
        VPresent = industryCount2[str(x)]
        VPast = industryCount1[str(x)]
        PR = ((VPresent-VPast)/VPast * 100)/n
        prCount.update({x:round(PR,1)})
prCountCount = Counter(prCount)
indStats = []
ylabels = []
for i in prCountCount.most_common(5):
    indStats.append(i[1])
ylabels.append(str(data_Label(i[0])))
return indStats, ylabels

def growth_rate_state(year1, year2, state):
    yearLoc = indLoc = stateCheck = stateICP = 0
    filename = 'census02.txt'
    lines = open(filename).read().splitlines()
    line0 = lines[0].split('t')
    industryCount1 = {}
    industryCount2 = {}
    prCount = {}
    #create list
    for x in range(len(line0)):
        if line0[x].strip() == '"YEAR"':
            yearLoc = x + 1
        if line0[x].strip() == '"IND1950"':
            indLoc = x + 1
        if line0[x].strip() == '"STATEICP"':
            stateLoc = x + 1
    while True:
        if str(data_Label(stateCheck, "STATEICP")) == str(state):
            stateICP = stateCheck
            break
        else:
            stateCheck+=1
    for x in range(len(lines)):
        if x != 0:
            # increment list location by 1 per
currentLine = lines[x].split("\t")
if int(currentLine[yearLoc]) == year1:
    if currentLine[indLoc] != '0' and int(currentLine[stateLoc])
        stateICP:
            if currentLine[indLoc] in industryCount1:
                industryCount1[currentLine[indLoc]] += 1
            else:
                industryCount1.update({currentLine[indLoc]:1})

== stateICP:
if currentLine[indLoc] != '0' and int(currentLine[stateLoc])
    stateICP:
        if currentLine[indLoc] in industryCount2:
            industryCount2[currentLine[indLoc]] += 1
        else:
            industryCount2.update({currentLine[indLoc]:1})

k1 = Counter(industryCount1)
k2 = Counter(industryCount2)
totalrec1 = sum(industryCount1.values())
n = year2 - year1
for x in industryCount1:
    VPresent = industryCount2[str(x)]
    VPast = industryCount1[str(x)]
    PR = ((VPresent-VPast)/VPast * 100)/n
    prCount.update({x:round(PR,1)})
prCountCount = Counter(prCount)
indStats = []
ylabels = []
for i in prCountCount.most_common(5):
    indStats.append(i[1])
ylabels.append(str(data_Label(i[0])))

return indStats, ylabels

def growth_rate_least(year1, year2):
yearLoc = indLoc = 0
filename = 'census02.txt'
lines = open(filename).read().splitlines()
line0 = lines[0].split("\t")
industryCount1 = {}
industryCount2 = {}
prCount = {}
for x in range(len(line0)):
    #print(line0[x].strip())
    if line0[x] == "YEAR":
        yearLoc = x + 1
    if line0[x] == "IND1950":
        indLoc = x + 1
for x in range(len(lines)):
    # print(lines[x])
    if x != 0:
        # increment list location by 1 per
        currentLine = lines[x].split("\t")
        if int(currentLine[yearLoc]) == year1:
            if currentLine[indLoc] != '0':
                if currentLine[indLoc] in industryCount1:
industryCount1[currentLine[indLoc]] += 1
else:
    industryCount1.update({currentLine[indLoc]:1})
elif int(currentLine[yearLoc]) == year2:
    if currentLine[indLoc] != '0':
        if currentLine[indLoc] in industryCount2:
            industryCount2[currentLine[indLoc]] += 1
        else:
            industryCount2.update({currentLine[indLoc]:1})

k1 = Counter(industryCount1)
k2 = Counter(industryCount2)
totalrec1 = sum(industryCount1.values())
n = year2 - year1
for x in industryCount1:
    VPresent = VPast = PR = 0
    if x in industryCount2:
        VPresent = industryCount2[str(x)]
        VPast = industryCount1[str(x)]
        PR = ((VPresent - VPast) / VPast * 100) / n
        prCount.update({x: round(PR, 1)})
prCountCount = Counter(prCount)
indStats = []
ylabels = []
for i in least_common_values(prCountCount, 5):
    indStats.append(i[1])
    ylabels.append(str(data_Label(i[0])))
return Reverse(indStats), Reverse(ylabels)
def generate(years, cities, states):
    while True:
        year = random.choice(years)
        num = random.randint(0, 1)
        if num % 2 == 0:
            state = random.choice(states)
        else:
            city = random.choice(cities)
        indStat, ylabel, names = most_common_state(year, state)
        graph_it_common_city(indStat, ylabel, names, year, state)
        else:
            indStat, ylabel, names = most_common_city(year, city)
            graph_it_common_city(indStat, ylabel, names, year, city)

Graph_it.py

1. from collections import Counter
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. import numpy as np
5. from least_common import least_common_values
6. import matplotlib.ticker as mtick
7. from xml_Reader import data_Label
8. def graph_it_growth(indStat, ylabel, x, y, least=True, check=True):
   if least == True:
       least = "Least"
   else:
least = "Most"
plt.rcParams['font.family'] = 'sans-serif'
plt.rcParams['font.sans-serif'] = 'Helvetica'
plt.rcParams['axes.edgecolor'] = '#333F4B'
plt.rcParams['axes.linewidth'] = 0.8
plt.rcParams['xtick.color'] = '#333F4B'
plt.rcParams['ytick.color'] = '#333F4B'
plt.rcParams['text.color'] = '#333F4B'

percentages = pd.Series(indStat, ylabel)
df = pd.DataFrame({'percentage': percentages})
df = df.sort_values(by='percentage')
my_range = list(range(1, len(df.index) + 1))
fig, ax = plt.subplots(figsize=(5, 3.5))
plt.hlines(y=my_range, xmin=0, xmax=df['percentage'], color='#007ACC',
alpha=0.2, linewidth=5)
plt.plot(df['percentage'], my_range, 'o', markersize=5, color='#007ACC',
alpha=0.6)

ax.set_xlabel('Percentage', fontsize=15, fontweight='black', color='#333F4B')
ax.set_ylabel('')
ax.xaxis.set_major_formatter(mtick.PercentFormatter())
ax.tick_params(axis='both', which='major', labelsize=12)
plt.yticks(my_range, df.index)
fig.text(0, 0.96, 'The Five Industries That Experienced \nthe ' + least + ' Growth From ' + str(x) + ' to ' + str(y), fontsize=10, fontweight='black', color='#333F4B')

ax.spines['top'].set_color('none')
ax.spines['right'].set_color('none')
ax.spines['left'].set_smart_bounds(True)
ax.spines['bottom'].set_smart_bounds(True)
ax.spines['bottom'].set_position(('axes', -0.04))
ax.spines['left'].set_position(('axes', -0.015))
if check == True:
    plt.show()

def graph_it_state(indStat, ylabel, x, y, z, check=True):
    plt.rcParams['font.family'] = 'sans-serif'
    plt.rcParams['font.sans-serif'] = 'Helvetica'
    plt.rcParams['axes.edgecolor'] = '#333F4B'
    plt.rcParams['axes.linewidth'] = 0.8
    plt.rcParams['xtick.color'] = '#333F4B'
    plt.rcParams['ytick.color'] = '#333F4B'
    plt.rcParams['text.color'] = '#333F4B'
    percentages = pd.Series(indStat, ylabel)
df = pd.DataFrame({'percentage': percentages})
df = df.sort_values(by='percentage')
my_range = list(range(1, len(df.index) + 1))
fig, ax = plt.subplots(figsize=(5, 3.5))
plt.hlines(y=my_range, xmin=0, xmax=df['percentage'], color='#007ACC',
alpha=0.2, linewidth=5)
plt.plot(df['percentage'], my_range, 'o', markersize=5, color='#007ACC',
alpha=0.6)

ax.set_xlabel('Percentage', fontsize=15, fontweight='black', color='#333F4B')
66.    ax.set_ylabel('')
67.    ax.xaxis.set_major_formatter(mtick.PercentFormatter())
68.    ax.tick_params(axis='both', which='major', labelsize=12)
69.    plt.yticks(my_range, df.index)
70.    fig.text(0, 0.96, 'The Five Industries That Experienced 
               the Most Growth 
               From ' + str(x) + ' to ' + str(y) + ' in ' + str(z), fontsize=10, 
               fontweight='black', color = '#333F4B')
71.    ax.spines['top'].set_color('none')
72.    ax.spines['right'].set_color('none')
73.    ax.spines['left'].set_smart_bounds(True)
74.    ax.spines['bottom'].set_smart_bounds(True)
75.    ax.spines['bottom'].set_position(('axes', -0.04))
76.    ax.spines['left'].set_position(('axes', -0.015))
77.    if check == True:
78.       plt.show()
79.    def graph_it_UR(urban, rural, year, check=True):
80.       labels  = ['Urban Population', 'non-Urban Population']
81.       sizes   = [urban, rural]
82.       colors  = ['#99ff99','#ffcc99']
83.       explode = (0.1,0.0)
84.       plt.pie(sizes, colors = colors, labels=labels, autopct='%1.1f%%',
85.               startangle=90, pctdistance=0.85, explode = explode)
86.       centre_circle = plt.Circle((0,0),0.70,fc='white')
87.       fig = plt.gcf()
88.       fig.gca().add_artist(centre_circle)
89.       plt.axis('equal')
90.       plt.title('Percentage of Urban and Non-Urban Populations in the Census of ' 
               + str(year))
91.       if check == True:
92.          plt.show()
93.    def graph_it_BPL(immigrant, non_immigrant, year, check=True):
94.       labels  = ['Urban Immigrant', 'Urban Native']
95.       sizes   = [immigrant, non_immigrant]
96.       colors  = ['#99ff99','#ffcc99']
97.       explode = (0.1,0.0)
98.       plt.pie(sizes, colors = colors, labels=labels, autopct='%1.1f%%',
99.               startangle=90, pctdistance=0.85, explode = explode)
100.      centre_circle = plt.Circle((0,0),0.70,fc='white')
101.     fig = plt.gcf()
102.     fig.gca().add_artist(centre_circle)
103.    plt.axis('equal')
104.    plt.title('Percentage of the Urban Population That Were 1st or 2nd 
               Generation Immigrants in ' + str(year))
105.    if check == True:
106.       plt.show()
107.    def graph_it_OWN(ownedUR, rentUR, ownedRL, rentRL, year, check=True):
108.       print(ownedUR, rentUR, ownedRL, rentRL)
labels = ['Urban Owners', 'Urban Renters', 'Rural Owners', 'Rural Renters']
sizes = [ownedUR, rentUR, ownedRL, rentRL]
colors = ['#4DB9FF', '#4750E5', '#99ff99', '#ffcc99']
explode = (0.1, 0.1, 0.0, 0.0)
plt.pie(sizes, colors=colors, labels=labels, autopct='%1.1f%%',
        startangle=90, pctdistance=0.85)

centre_circle = plt.Circle((0, 0), 0.70, fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.axis('equal')
plt.title('Urban/Rural Percentages in the Census of ' + str(year) + ' with Home Ownership Subdivisions ')
if check == True:
    plt.show()

def graph_it_city_bpl(immigrant, non_immigrant, year, city, check=True):
    labels = ['Urban Immigrant', 'Urban Native']
sizes = [immigrant, non_immigrant]
colors = ['#99ff99', '#ffcc99']
explode = (0.1, 0.0)
plt.pie(sizes, colors=colors, labels=labels, autopct='%1.1f%%',
        startangle=90, pctdistance=0.85, explode=explode)

centre_circle = plt.Circle((0, 0), 0.70, fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.axis('equal')
plt.title('Percentage of People living in ' + str(city) + ' That Were 1st or 2nd Generation Immigrants in ' + str(year))
if check == True:
    plt.show()

def graph_it_common(indStats, ylabelpercent, index, year, check=True):
    plt.rcParams['font.family'] = 'sans-serif'
    plt.rcParams['font.sans-serif'] = 'Helvetica'
    plt.rcParams['axes.edgecolor'] = '#333F4B'
    plt.rcParams['axes.linewidth'] = 0.8
    plt.rcParams['xtick.color'] = '#333F4B'
    plt.rcParams['ytick.color'] = '#333F4B'
    plt.rcParams['text.color'] = '#333F4B'
    percentages = pd.Series(indStats, index)
    df = pd.DataFrame({'percentage': percentages})
    df = df.sort_values(by='percentage')
    my_range = list(range(1, len(df.index) + 1))
    fig, ax = plt.subplots(figsize=(5, 3.5))
    plt.hlines(y=my_range, xmin=0, xmax=df['percentage'], color='#007ACC',
               alpha=0.2, linewidth=5)
plt.plot(df['percentage'], my_range, "o", markersize=5, color='007ACC', alpha=0.6)
ax.set_xlabel('Percentage', fontsize=15, fontweight='black', color='#333F4B')
ax.set_ylabel('')
ax.tick_params(axis='both', which='major', labelsize=12)
plt.yticks(my_range, df.index)
fig.text(0, 0.96, 'Five Most Common Industries Reported in '+ str(year), fontsize=10, fontweight='black', color='333F4B')
ax.spines['top'].set_color('none')
ax.spines['right'].set_color('none')
ax.spines['left'].set_smart_bounds(True)
ax.spines['bottom'].set_smart_bounds(True)
ax.spines['bottom'].set_position(('axes', -0.04))
ax.spines['left'].set_position(('axes', -0.015))
plt.show()

def graph_it_common_city(indStats, ylabelpercent, index, year, city, check=True):
    plt.rcParams['font.family'] = 'sans-serif'
    plt.rcParams['font.sans-serif'] = 'Helvetica'
    plt.rcParams['axes.edgecolor']='#333F4B'
    plt.rcParams['axes.linewidth']=0.8
    plt.rcParams['xtick.color']='#333F4B'
    plt.rcParams['ytick.color']='#333F4B'
    plt.rcParams['text.color']='333F4B'
    percentages = pd.Series(indStats, index)
    df = pd.DataFrame({'percentage' : percentages})
    df = df.sort_values(by='percentage')
    my_range=list(range(1,len(df.index)+1))
    fig, ax = plt.subplots(figsize=(5,3.5))
    plt.hlines(y=my_range, xmin=0, xmax=df['percentage'], color='#007ACC', alpha=0.2, linewidth=5)
    plt.plot(df['percentage'], my_range, "o", markersize=5, color='007ACC', alpha=0.6)
    ax.set_xlabel('Percentage', fontsize=15, fontweight='black', color='#333F4B')
    ax.set_ylabel('')
    ax.tick_params(axis='both', which='major', labelsize=12)
    plt.yticks(my_range, df.index)
    fig.text(0, 0.96, 'Five Most Common Industries Reported inside of '+ str(city) + ' During the Census of '+ str(year), fontsize=10, fontweight='black', color='333F4B')
    ax.spines['top'].set_color('none')
    ax.spines['right'].set_color('none')
```
import xml.etree.ElementTree as ET

def data_label(value, ident="IND1950"):
    label = str(xml_Reader(value, ident))
    if len(label) > 20:
        token = label.split()
        if len(token) % 2 == 0:
            index = int(len(token)/2)
            str1 = token[:index]
            str2 = token[index:]
            str1.append('
')
            finalStr = str1 + str2
            return ' '.join(finalStr)
        else:
            index = int(len(token)/2)
            str1 = token[:index]
            str2 = token[index:]
            str1.append('
')
            finalStr = str1 + str2
            return ' '.join(finalStr)
    else:
        return label

def xml_Reader(value, ident="IND1950"):  
    tree = ET.parse(ident + '.xml')
    root = tree.getroot()
    for var in root:
        if var.find("catValu") is not None:
            if var.find("catValu").text == str(value):
                return var.find("labl").text
            if var.find("labl").text == str(value):
                return var.find("catValu").text

main.py
```

```
from graph_it import *
from functions import *

cities = ["New York, NY", "Los Angeles, CA"]
years = [1900, 1910, 1920, 1930]
state = ["California", "New York", "Alabama", "Texas"]

for x in years:
    urban, rural = urban_or_rural(x)
    graph_it_UR(urban, rural, x)
```
12. immigrant, non_immigrant = urban_BPL(x)
13. graph_it_BPL(immigrant, non_immigrant, x)
14. ownedUR, rentUR, ownedRL, rentRL = urban_or_rural_ownership(x)
15. graph_it_OWN(ownedUR, rentUR, ownedRL, rentRL, x)
16. for y in cities:
17.     immigrant, non_immigrant = city_BPL(x, y)
18.     graph_it_city_bpl(immigrant, non_immigrant, x, y)
19.
20. for x in years:
21.     for y in cities:
22.         indStat, ylabel, names = most_common_city(x, y)
23.         graph_it_common_city(indStat, ylabel, names, x, y)
24.
25.
26. for x in years:
27.     for y in years:
28.         if x < y:
29.             indStat, ylabels = growth_rate_city(x, y)
30.             graph_it_growth(indStat, ylabels, x, y)
31.
32.
33.
34.
35.
36. for x in years:
37.     for y in years:
38.         if x < y:
39.             for z in state:
40.                 indStat, ylabel = growth_rate_state(x, y, z)
41.                 graph_it_state(indStat, ylabel, x, y, z)
42.
43. generate(years, cities, state)
Bibliography


Pearl, Josh. "History 90.01: Topics in Digital History." History 9001 Topics in Digital History. Dartmouth University, January 25, 2016.


https://www.census.gov/.
