Introduction to Data Science

CISD41 – Fall 2019

Danny Lam
Julio Gutierrez
Yannick Brossel

New York City – Census Data

Introduction

For our Final Project, we will be analyzing the census data collected of New York City. This dataset (https://www.kaggle.com/muonneutrino/new-york-city-census-data#nyc_census_tracts.csv) contains data from the United States Census Bureau and revised by the 2015 American Community Survey 5-year Estimates. Each census tract block includes demographics and economic characteristics of nearby NYC neighborhoods. We will use Python and its many libraries to run exploratory data analysis, data cleaning, and data visualization to take raw statistical data and convert them into useful information we can use to show certain trends.

Data Cleaning

To begin with, this dataset has a total of 2167 records or census tract with a few columns containing null values. This is already a great set to work with because the majority of the data is full complete, however, we would like to work with as clean of data as possible.

The first thing we did was remove any unwanted columns. Since there were still many columns showing null values within margin, we dropped any row where half or more of the data is missing. With only 4 columns with null values left, we decided to fill those null values with the mean of each of the 4 columns and their respective county. We then converted the 'County' column from an object to a category type for later use with the creation of plots. Success! We have now confirmed that all of the columns share an equal number of rows and all have non-null values.

Another challenge we faced was that many of the columns were in percentages. This presents an issue for statistical testing because the population per block are not equally distributed, thus, taking averages of percentages will give weight for both small and largely populated blocks. Using simply mathematics, we ran a function that converted the percentages into a decimal value by dividing the percentage by 100 and multiplying it by the total population. We truncated a decimal value since we cannot have a fraction of a person count.

Demographics

Now that our data is ready for data analysis, let's consider our population. Here, we asked ourselves a few questions to answer:

- **How many people reside in each county?**
    - The total population of NYC is 8.4 million
    - Bronx has 1.4 million
    - Kings has 2.6 million
    - NY has 1.6 million
    - Queens has 2.3 million
    - Richmond has 0.5 million
- **Are there more males or females on average?**
    - On average, there are more women (4.4 million) than men (4 million).
- **Which counties have the most and least number of blocks?**
    - Kings has the most at 750 blocks
    - Richmond has the least at 108 blocks
- **Which counties are most congested and which are the least congested?**
    - New York is the most congested with an average of 5,800 per block
    - Kings and Queens are the least congested with an average of 3,500 per block
- **Is our Population normally distributed?**
    - The average of each block is approximately 4,000 and it is skewed to the right
    - Since the p-value is 4.53e-208, and is less than 0.05, it rejects the null hypothesis and concludes that the total population is not normally distributed (or not Gaussian).

Wealth and Ethnic Disparities

We then considered deeper topics such as what are the wealth and ethnic disparities among counties, if any? Thus, we created a second pivot table with a new set of questions:

- **What are the distributions of ethnicities per county?**
    - Asian = 1.1 million total
        - There are the most Asians in Queens and least in both the Bronx and Richmond
    - Black = 1.8 million total
        - There are the most Blacks in Kings and least in Richmond
    - Hispanic = 2.4 million total
        - There are the most Hispanics in the Bronx and least in Richmond
    - Native = 15 thousand total
        - There are the most Natives in Kings and Queens and least in Richmond
    - White = 2.7 million total
        - There are the most Whites in Kings and NY and least in the Bronx
- **What are the disparities in median household income and income per capita per county?**
    - New York, Richmond, and Queens demonstrate median household incomes greater than the national average ($60,000).
    - Bronx and Kings shows values below the national average median household income.
    - New York demonstrate Income Per Capita greater than the national average ($51,000).
        - We can infer that New York County residents are on average paid more than other counties with one individual working per household.
    - All remaining counties showing values below the national average median household income.
        - Richmond and Queens typically will on average have two people working per household.
    - The ECDF we ran shows a median IncomePerCap well-above $50,000 for New York and significantly less than $50,000 for the remaining counties.
- **What are the disparities of poverty and child poverty per county?**
    - Kings, Bronx, and Queens county shows a staggering amount of poverty.
    - Kings, Bronx, and Queens county also show a staggering amount of child poverty. These values coincide with their poverty levels.
    - We also ran a Pearson's correlation test between poverty and child poverty.
        - We concluded that there is strong evidence of 0.88 to reject the null hypothesis and suggest that poverty and child poverty are dependent.
- **What are some, if any, correlations regarding income?**
    - Poverty and ChildPoverty (r = .88)
    - Income and IncomePerCap (r = .81)
    - Poverty and Income (r = -.71)
    - Poverty IncomePerCap (r = -.49)

- ChildPoverty and Income (r = -.69)
- #ChildPoverty and IncomePerCap (r = -.50)

Indicators of Wealth

Previously we learned that there were wealth disparities present between the counties of New York City, with New York County being the most affluent and the Bronx being the least affluent. Let us see if there are any other indicators that will continue to demonstrate such wealth disparities. For these pivot tables, we will use the total amount of people employed instead of the total population because we are analyzing people that work. Using the total population for measurement would give us an inaccurate analysis of economic characteristics because not everyone is employed like children or stay-at-home wives or husbands. Let's go ahead an answer these questions:

- **What are the average and total values of job sectors employed for each county?**
  - Professional:
    - New York demonstrates the highest total of "professional" workers
    - The Bronx has the least amount "professional" workers.
  - Service:
    - Kings and Queens has a comparable lead of "Service" workers.
    - Richmond has the least amount of "service" workers.
    - The Bronx shows a significantly greater number of average and total "Service" workers than New York
  - Office:
    - New York shows the highest "Office" workers on average per block.
    - Kings and Queens shows having a comparable lead on total "Office" workers.
    - New York demonstrates approximately 50% more "Office" workers on average and total than the Bronx
  - Construction:
    - Richmond shows the highest "Construction" workers on average per block.
    - Kings and Queens shows having a significant lead on total "Construction" workers.
    - The Bronx has approximately twice the number of construction workers on average and total than New York.
  - Production:
    - Bronx and Queens demonstrate having a significant lead on the average "Production" workers per block
    - Kings and Queens shows have a significant lead on total "Production" workers.
    - The Bronx demonstrates approximately 50% more "Production" workers on average and total than the New York

    - Does this fit our previous data that New York is wealthier and the Bronx is least wealthy? Probably, yes.

- **What are the average and total values of means of transportation used for each county?**
    - Drive:
        - New York County has the least total and average "Drivers"
            - Probably because NY is heavily congested to drive through
        - Richmond has the highest average "Drivers".
        - Queens has the greatest total "Drivers".
    - Carpool:
        - The trend mirrors that of the "Drivers".
    - Transit:
        - The greatest average number of "Transit" commuters are from New York.
            - The subway system in NY is massive
        - Kings has the greatest total "Transit" commuters.
    - Walk:
        - New York has the greatest total and average for "Walking" to work.
        - Perhaps many jobs are local to New York County.
    - OtherTransp:
        - New York has the greatest total and average for "Other" means of transportation to work.
            - Biking and skating to work is really popular in NY
    - WorkAtHome:
        - New York has the greatest total and average for "Working at home".
            - Probably because NY has more professional, corporate, and tech type jobs and are able to have the opportunity to work remotely.
    - Does this fit our previous data that New York is wealthier and the Bronx is least wealthy? Most likely, yes.
- **Is there any a correlation between income and the average commute time to work?**
    - New York has the least average commute time by 10 minutes
    - The correlation between Income and MeanCommute time to work is -0.36
        - As income increases, commute times tends to decrease.
    - Income is negatively correlated with income and as previously shown, New York has the highest Income and also shows the least average commute time.
- **Draw a conclusion if possible…. Are the job types and transportation type suspected of higher social status present in greater amounts for the more affluent county (New York) and diminished in less affluent counties (Bronx)? In short, did we find additional indicators that New York county is wealthier whereas the Bronx is financially deficient?**
    - We have found additional indicators supporting that New York is wealthier than the other counties and the Bronx is less wealthy than the other counties.

- **Summary #1:** New York County has a propensity to house more "Professional" and "Office" among all counties. The Bronx houses significantly greater "Service", "Construction", and "Production" workers than New York with leading average per block in "Service" and "Production".
- **Summary #2:** New York County has a propensity to house more "WorkingAtHome", "Walk", and "Transit" commuters among all counties. The Bronx shows comparable commuting activity compared to the other the majority of counties for a given category.

Searching for our own Indicator

Previously we discovered data supporting that New York county is wealthier and the Bronx is less than wealthy than the other counties within New York City. Several of these inferences could be accepted on prima facie alone given one's experiences that develop into a "common sense", so to speak. Big Data allows for evidence-based decisions to be made by meticulously parsing and analyzing information. We will use our present skills to either support or refute a hypothesis for which the group could not unanimously agree. We asked ourselves…

- **Do the proportions of work category also have a significant correlation with the county's social economic status? In other words, will these categories show more "private" employees in New York versus more "public" employees in the Bronx?**
- We *hypothesize* that New York will house a proportion of “Private” employees, while the Bronx will house a greater proportion of “Public” employees. This is because "Private" positions are generally compensated with a greater income than that of "Public" positions.
- After running data analysis and a few z-tests of all the different job types by county, we have *concluded* that:
    - New York demonstrates comparable levels of total and average "Private Work" compared to the Bronx with data visualizations. (Hypothesis not supported)
    - However, according to hypothesis testing there is signficant evidence to conclude that New York has a mean "PrivateWork" value greater than 81 meanwhile the Bronx was smaller than 81.
        - This leads me to believe that there is a statistical difference with New York having a greater number of "Private" employees than the Bronx.
    - The Bronx shows a higher 50% average and sum of "PublicWork" compared to New York according to data visualization. (Hypothesis supported)
    - Hypothesis testing indicates signficant evidence to conclude that the Bronx has a mean "PublicWork" value greater than 11 meanwhile New York has a smaller value than 11.
        - This leads me to believe that there is a statistical difference with the Bronx having a greater number of "Public" employees than New York.
- In summary, New York shows a greater wealth than the Bronx and it can be supported by a number of indicators including income, poverty, mean commute times, proportions of job types help, proportions of commuting times for the residents, and including proportions of job sectors within the counties.