**Data Analysis – NYC Housing Based on Boroughs**

Gabriele DAmelio Sanches and Saul Santana Lopez

Baruch College – Zicklin School of Business

CIS9650: Programming for Analytics

Fatemeh Choopani

May 16, 2023

# Abstract

The purpose of this project is to evaluate housing statistics for the boroughs and areas of New York City to forecast things like income, median rent, poverty rate, crime rate, and other relevant factors. To gain insights into the different things that affect the housing market in New York City, the project will use the NYC housing dataset - extrinsic factorsdata from Kaggle.

*Data Collection:* Extracted the CSV file from Kaggle.

*Data Cleaning:* To verify the accuracy and consistency of the data, it was cleaned and preprocessed. This included filling in blanks, adjusting outliers, and changing data types as necessary.

*Data Transformation:* To have a better format for the analysis of the data, it was filtered to remove unwanted data, aggregated to create summary statistics, and merged where needed.

*Data Visualization:* Creating graphs and data frames helps to visualize and analyze the data in a simple way. Using matplotlib, scatterplots and bar graphs were created to visualize how the columns of the dataset relate to one another, and to analyze the useful information found in the previous steps.

**Data Analysis – NYC Housing Based on Boroughs and Areas**

New York City is known as one of the most expensive places to live in, and even with the insanely high prices, it can be challenging to find housing in NYC. Commute time, boroughs, distance to public transportation, park, the number of housing units available, crime rate, and building age are all factors that affect the overall housing aspect in the city.

Other factors that can be analyzed based on housing data are the median income of the population living in certain boroughs and areas, the population and population density, poverty rate, rental vacancy rate, unemployment rates, and crime rates, as well as racial diversity index, and how much of each race is part of each community (only from a housing point of view).

**Data Cleaning**

For this step, the data in the CSV file format was imported into the jupyter notebook using pandas to read the file. To better visualize the data, we used the Boroughs as our main index and title to analyze our data and number based on each borough of New York City.

The original data contained 55 rows and 35 columns. No duplicated values were checked using the *df.duplicated()* function, and six null values needed to be fixed before starting the data analysis process.

The null values were in the columns ‘unemployment’ and ‘labor\_force\_rate’, and to fix this issue, the mean of the values of each column that had the null value was calculated, and this mean was used to fill the blank values we had, using the *.fillna()* function.

After filling in the missing data, once again using the *.duplicated()* and *.isnull()* functions, the dataset was checked to see if all the values were in place and there was no duplicated data.

Graphical user interface, table

Description automatically generated

The data is content, and each variable has the same data type and range values. There are no apparently outliers in the dataset, but since we are looking at different boroughs and areas and comparing, for example, the median rent, median income, and commute to where might be valued in each column might significantly differ. Still, it does not mean that they are outliers. This implies that, in New York City, all those values can vary depending on the area analyzed.

The last step of data cleaning was validating the chosen data by comparing it to data from other sources. The NYC Department of City Planning - Census Factfinder is a demographic data for NYC, including population counts by age, race, ethnicity, and information on household income, education levels, and more. Utilizing this dataset, it is possible to compare the original Data set (from Kaggle) in population, population density, housing units, rental vacancy, racial diversity, and race. The NYC Department of City Planning supported the NYC Housing data for the project.

Zillow data was used to validate the median rent in the dataset used for the project (Zillow New York Rent). The information in both the Kaggle dataset and Zillow matches; There are some minor differences in rent, but this might be because the data were collected in different years, but the difference is not big enough to invalidate the data.

**Data Transformation**

The practice of transforming initial data into a format better suited for analysis or other subsequent procedures is known as data transformation. It includes modifying, reorganizing, or gathering data to make it more informative and helpful.

In this step, the unnecessary columns for the analysis or do not look like a good fit for the project were dropped, filtering the data. The columns 'born\_in\_new\_york\_state' and 'foreign\_born\_population' were dropped because the analysis and visualization will be made concerning the population in general and not based on where they were born. Also, it did not include the people born in the U.S. but not in New York State, which could create an issue with the accuracy of the data. The columns' public\_housing', 'new\_residential\_permits', 'new\_certificates\_of\_occupancy', 'subsidized\_properties', and 'severe\_crowding\_rate' were dropped because all of those were too specific regarding the building and apartment types. In the analysis, we will want to visualize the occupancy rate, the housing number, and the population number, and not the specifics of it, i.e., if the certificate of occupancy is new or not, if the residential permit is new, etc...

Lastly, the columns, '25+\_with\_bachelors' and '25+\_no\_high\_school\_diploma' were dropped because the data that will be used for the analysis and the visualization is the income and unemployment rate data. It does not matter if those data are related to someone with or without a college degree.

At the end of the process of the dropping table, the table contained 55 rows (the same as the original table) and 26 columns (compared to 35 before). To sort the data and have a more precise visual, the .sort\_values function was used, and the column utilized to sort the data, in descending order, was the median rate column, that way, we could see the least expensive boroughs to rent at from this dataset.

Utilizing summary statistics, measures of central tendency like mean, median, and mode are frequently employed to provide a single value that describes a dataset's core or typical value. With the NYC Housing data set, the median rent was $1731.35, and the mean was $1937.23. The mean is more accurate for finding an average value of how much rent costs in New York City because it takes all the values and divides them by the number of samples used in the data. At the same time, the median gets the" in the middle" value, which is not as accurate to use as the average rent.

With the mean function, the average yearly income by household in the city is $45719.61, which is extremely low considering that the mean rent is close to $2000 monthly. The widespread assumption that New York City is an expensive place to live in and that most people struggle to pay rent is backed up by the statistics found in this analysis.

Another interesting data is the racial diversity index, mean of 0.56, meaning that NYC is a highly diverse place. The poverty rate is only 0.2, which shows that NYC has a significant middle-class and affluent population. To create a clear visualization using Matplotlib, the data were grouped based on boroughs; all the rows were grouped based, and the final values on each are the mean of each column group based on the boroughs.

A screenshot of a computer

Description automatically generated with medium confidence

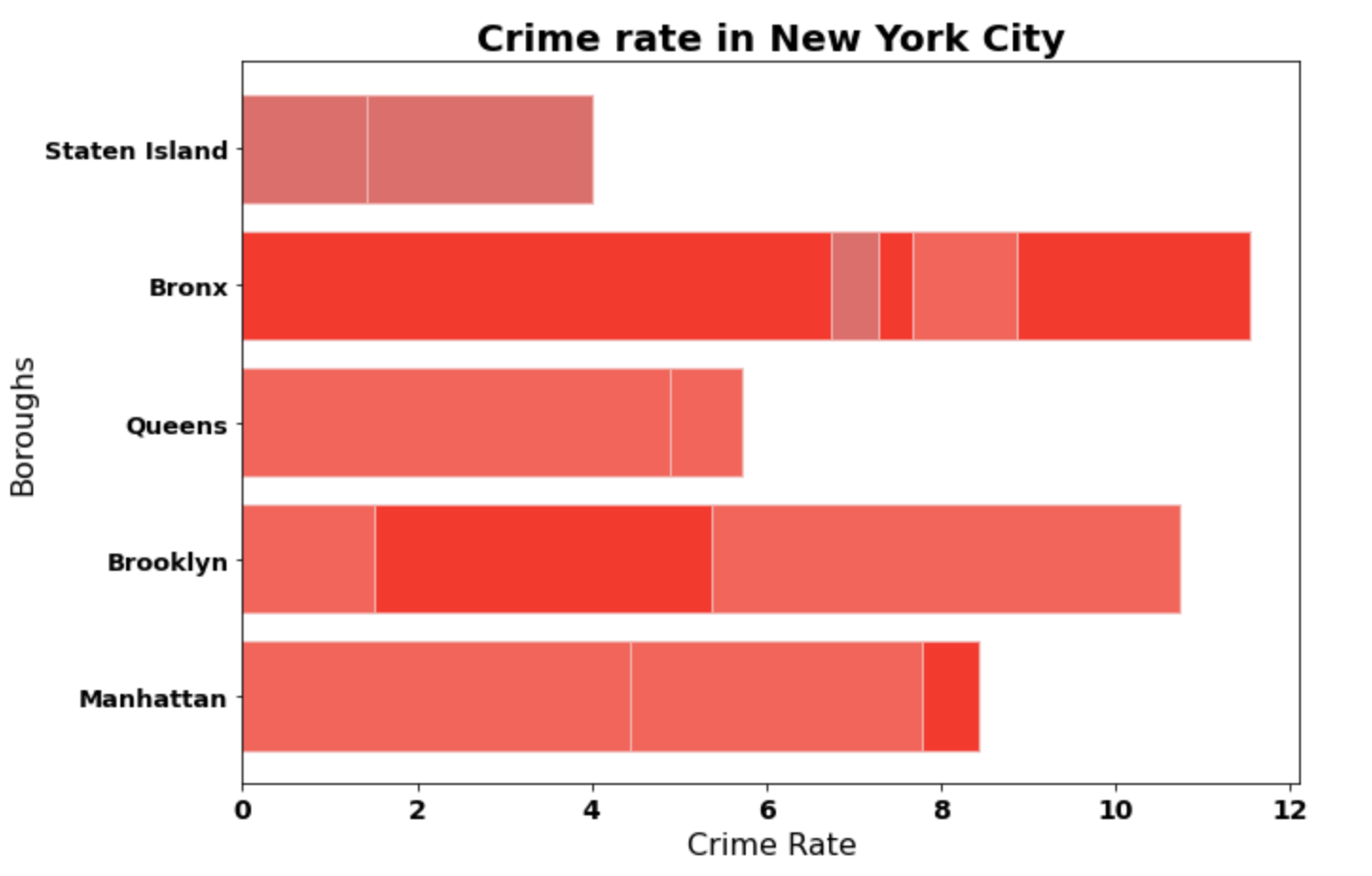
**Data Visualization**

Data from the NYC housing market was utilized to analyze key metrics such as median income, median rent, poverty rate, diversity index, and crime rate across the different boroughs of New York City. In this project step, the matplotlib package in Python was used to visualize the data. During the analysis, we observed that some visualization options, such as histogram, boxplot, line plot, and pie chart, did not help us with the results we wanted for our data visualization. Therefore, we used scatter, bar, and barh plots instead.

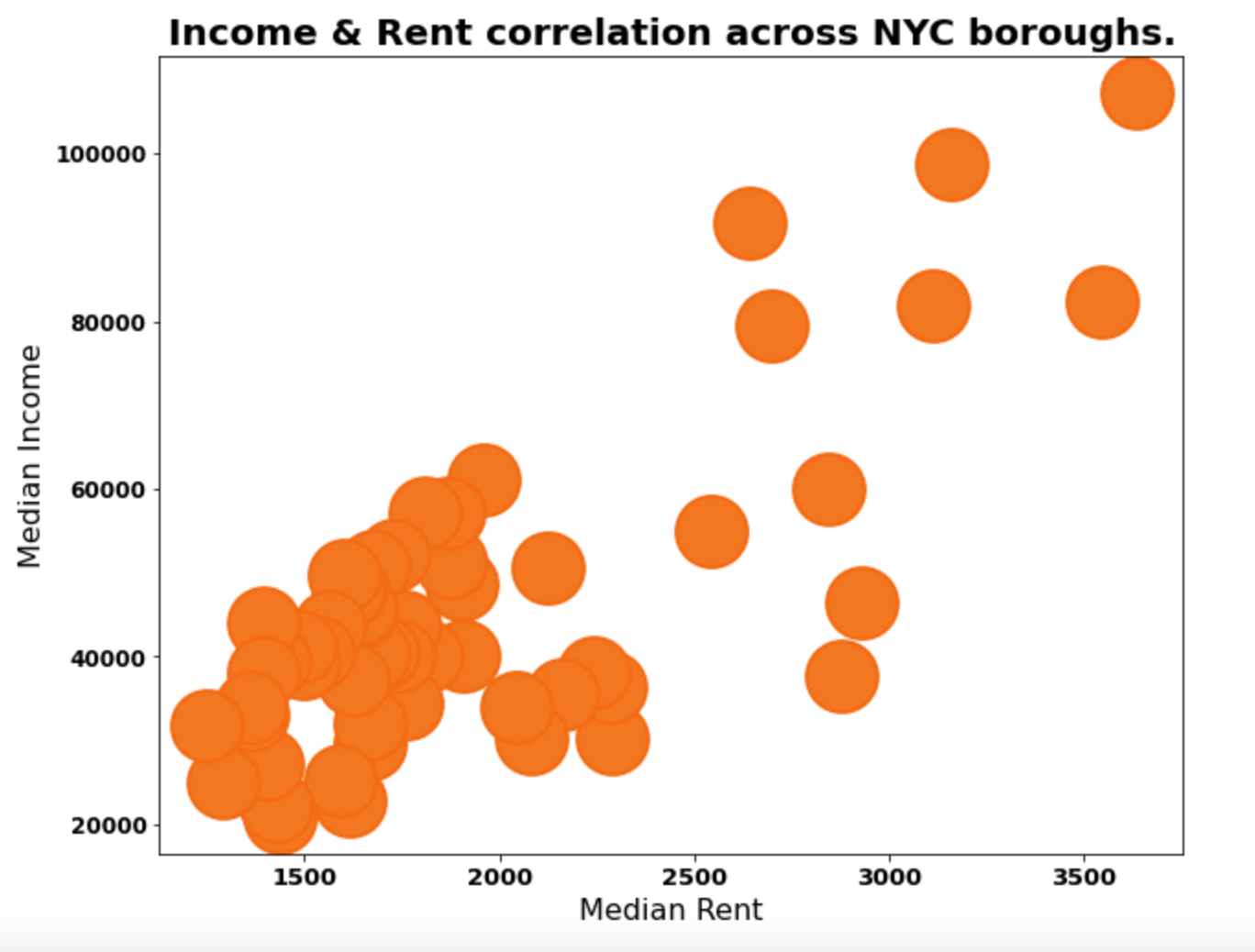
Scatter plots were used to visualize the correlation between two variables, such as the correlation between median income and median rent in different boroughs. Bar plots helped compare different variables across the boroughs, such as the poverty rates across the different boroughs. Barh plots were used to compare the same variables across the boroughs in a horizontal orientation, such as comparing the crime rate in each borough.

The data visualizations revealed significant variations in the median income, median rent, poverty rate, diversity index, and crime rate across the different boroughs. These disparities indicate a need for targeted policies to address the different challenges faced by each borough. For example, areas with high poverty may require more affordable housing options, while areas with high crime rates may need increased police presence.

It is essential to know the crime rates of a borough for several reasons: Safety: High crime rates can indicate that an area may not be safe, which can be a concern for residents and potential homebuyers. This can affect the investment potential of a property, as well as the resale value. Quality of life: Living in an area with high crime rates can harm the quality of life, leading to stress, anxiety, and unease. Community involvement: Understanding the crime rates can also encourage community involvement in crime prevention efforts. Moreover, this is only one of the critical factors that can affect many aspects of housing decisions and the overall well-being of renters.



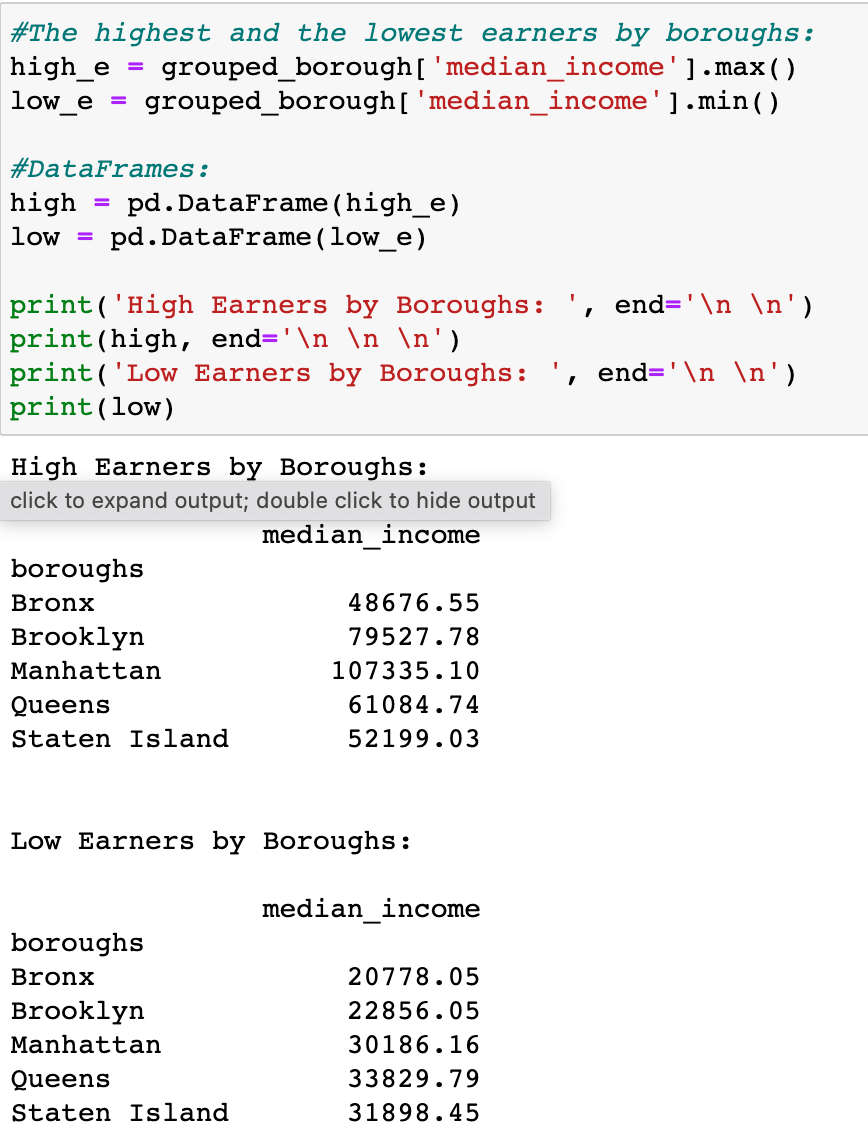
In the case of our scatter plots, we first tried to visualize our rental vacancy rate variable by borough, but this resulted in the borough variable having little influence over the vacancy rate. To show a positive correlation, we decided to show the rent and income correlation, and for a negative correlation, the unemployment and labor force rate. These were general examples of our dataset that define the key factors influencing the individual situation when looking for housing.



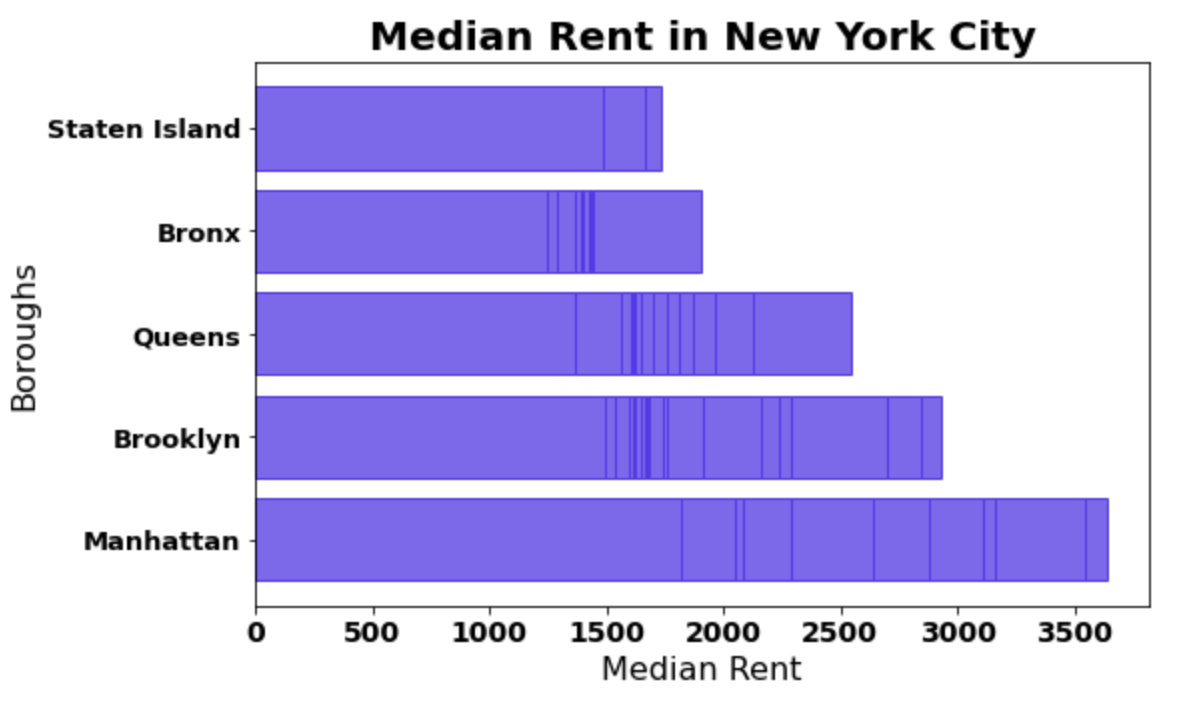
Overall, data visualization techniques helped our group gain insights into the NYC housing market and identify areas that require attention when looking for housing in New York City, providing an overview of the different aspects of the individual's current situations and other factors in the different boroughs. The analysis provides a foundation to make informed decisions about housing, neighborhood selection, and whether there need to be targeted policies to improve access to affordable housing for residents of New York City.

**Data Analysis**

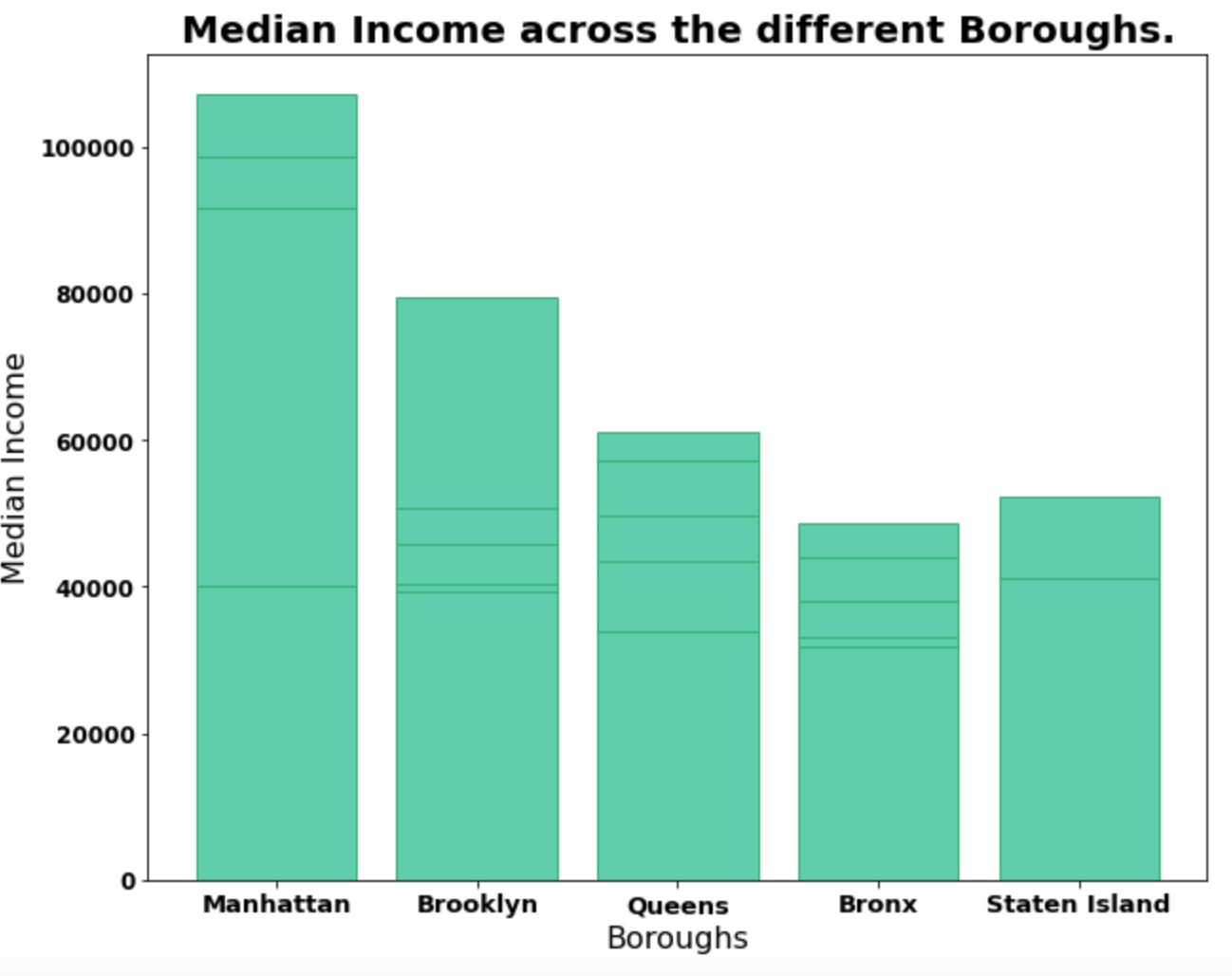
New York was listed at the top of the list of the most expensive cities in the world by the Economist Intelligence Unit (EIU) last year using the Worldwide Cost of Living Index (Rahmanan, 2022). During the analysis of the NYC Housing based on the boroughs dataset, the insights confirm this ranking but also show the differences in prices between boroughs, neighborhoods, and how other factors affect the rate of expensiveness in the city.



The rent is all five boroughs are very high, with the lowest median rent being in $1700 in Staten Island and the highest being $3500 in Manhattan, based on the bar plot created. Looking at only this numbers, it can appear that it is much cheaper and affordable to live in Staten Island rather than in Manhattan. Still, the median income visualization shows that the income is much lower in Staten Island, between $40,000 and $60,000, compared to Manhattan where the median income varies a lot. Still, it includes the highest median income in the dataset, around $100,000, which explains the difference in rent range.

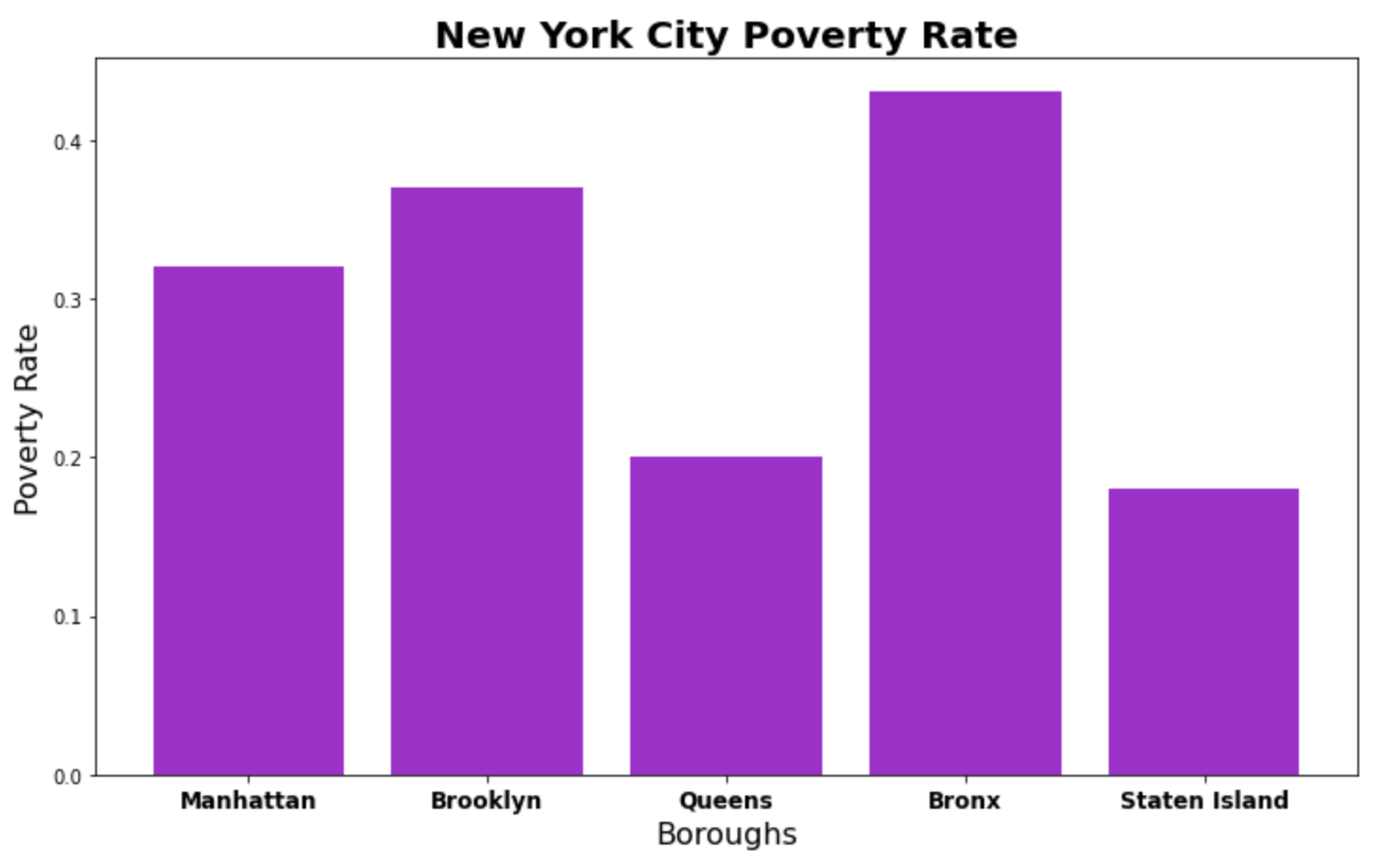


Manhattan shows the most inconsistent and spreader out median income out of all boroughs, half of the dataset for this borough shows that the annual median income is closer of higher than $100,000 and the other half shows that the median annual income is around $40,000, which is a huge difference for the same borough. This is an indication of income inequality in New York City, how there are extremely rich people and people that barely can afford rent living in the same borough.



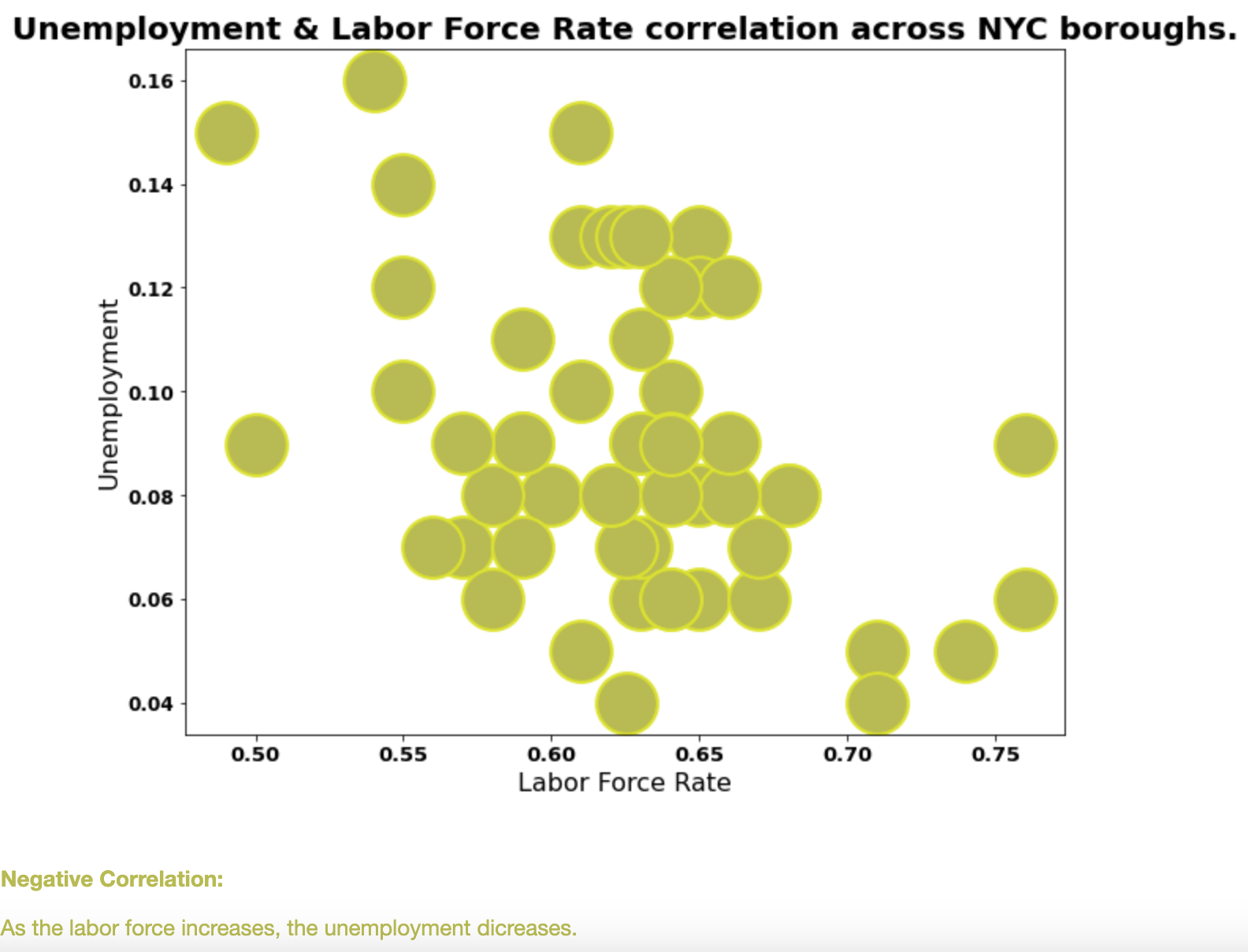
Even with the high rent, it is very hard to find available housing in New York City. Depending on the neighborhood of each boroughs the rental vacancy varies, but Manhattan is the borough in general that has the least rental vacancy rate, followed by Queens and Bronx. Brooklyn also has a low vacancy rate, but it still has a couple more available options than the other three mentioned. Staten Island is the place with the highest vacancy rate, showing that more people want to live and rent in the “heart” of NYC, where public transportation is more accessible, and the commute is easier.

The poverty rate mean in NYC is 0.201, with Queens and Staten Island being the boroughs with the lowest rate and the Bronx the borough with the highest, almost close to 0.5. In the data analysis was found that the unemployment rate in the Bronx is 0.12, the highest unemployment through all the 5 boroughs, explaining why the poverty rate is also the highest.



Our analysis identified a positive correlation between unemployment and the labor force rate in New York City. This relationship indicates that the labor force rate decreases as the unemployment rate decreases. This finding is crucial for understanding the socioeconomic dynamics of the city, as it reveals the impact of employment on housing affordability and the ability of individuals to reside in specific areas.

Recognizing the correlation between unemployment and the labor force rate has significant implications for policymakers and urban planners. It highlights the interdependence between employment opportunities and housing affordability, emphasizing the need for targeted strategies to reduce unemployment and promote inclusive economic growth. By addressing this correlation, policymakers can develop comprehensive initiatives that enhance employment prospects, foster sustainable urban development, and ensure equitable access to housing opportunities.



**Conclusion**

In conclusion, the analysis of the NYC Housing Dataset based on boroughs and areas provides valuable insights into various factors affecting the housing market in New York City. The dataset reveals significant variations in median income, median rent, poverty rate, diversity index, and crime rate across the different boroughs, highlighting the need for targeted policies and interventions.

The analysis confirms that New York City is expensive, with high median rents observed across all boroughs. Manhattan stands out as the most expensive borough, with the highest median rent, while Staten Island has relatively lower rents but also lower median incomes. Income inequality is evident, particularly in Manhattan, where there is a wide disparity between high and low earners.

Additionally, the analysis highlights the challenges of finding available housing in New York City, with low rental vacancy rates observed, especially in Manhattan. The highest poverty rate in the Bronx is accompanied by a high unemployment rate in that borough.

The data visualization techniques employed, such as scatter plots, bar plots, and barh plots, effectively demonstrate the relationships between different variables and provide a clear overview of the housing market dynamics across the boroughs.

Overall, this analysis emphasizes the need for comprehensive strategies to address the housing challenges faced by New York City residents. Policymakers should consider factors such as affordability, income inequality, crime rates, and access to housing in their decision-making processes. By understanding the nuances of each borough and area, targeted policies can be implemented to improve housing access, affordability, and overall quality of life for New York City residents.

**References**

*NYC housing dataset - extrinsic factors*. Kaggle. (2019, November 3). Retrieved April 30, 2023, from <https://www.kaggle.com/datasets/ted8080/nyc-housing-dataset-extrinsic-factors>

*NYC Planning Population factfinder*. NYC Population FactFinder. (n.d.). Retrieved April 30, 2023, from <https://popfactfinder.planning.nyc.gov/#11.67/40.7198/-73.9515>

Rahmanan, A. (2022, December 1). NYC is officially the most expensive city in the world. Time Out New York. <https://www.timeout.com/newyork/news/nyc-is-officially-the-most-expensive-city-in-the-world-120122>

Zillow. *Rental listings in New York NY - 9148 rentals*. Zillow. Retrieved April 30, 2023, from <https://www.zillow.com/new-york-ny/rentals/>