Database and Analytical Programming Report: Research on Residing Conditions of New York City

Raj Shrikant Sonawane
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21155054@student.ncirl.ie

Viplav Vijay Gadewar
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21164274@student.ncirl.ie

Kapil Lalwani
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21123292@student.ncirl.ie

Rohan Sanjay Kore
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x19214413@student.ncirl.ie

***Abstract*—Living fulfillment is one of the most important and fundamental aspects of human needs. Property, commute, livelihood, and a sense of security are just a few of the considerations that anyone, whether relocating to a new city or currently living there, will make. Some factors have a direct or indirect impact on customers considering real estate purchases. When deciding whether or not to relocate to this area, one must evaluate a variety of factors, including the distribution of natural vegetation around the city. Trees are the only natural element that has been demonstrated to be useful in the midst of the concrete jungle. We looked at some of the factors that influence living circumstances in New York City in this post.**

***Index Terms*—Living conditions, Property, Trees, Crime, Python, Database, Visualization**

I. Introduction

When a person moves to a new location for job, school, or company, they strive to settle in by considering a number of factors. The sort of housing available and the neighborhood in which it is located are both crucial elements in determining whether or not a person can live in a certain region. Buyers, tenants, and other entities make decisions based on a variety of factors, such as how convenient it is to commute within the city, whether the residing area is safe or unsafe, crime rates in the city, housing prices in relation to the current market, living conditions, and even environmental factors such as trees, rivers, parks, and animal habitats. The main purpose of this research is to think about these things and come to a decision. We chose New York as our major example because it is one of the most densely populated cities in the United States, with people arriving for a variety of reasons and attempting to establish themselves. Buying a home in New York City is a difficult undertaking, especially when considering the city's high prices. Many people consider pricing, as well as a variety of other factors when making a home purchase. Taking the safety as a top priority, researchers agree that crime diminishes safety and disturbs social order, both of which have an impact on living conditions. Taking nature into consideration, we examined trees in New York for this study. These trees are not only city emblems, but they also have significant ecological benefits. Mature trees, like these heritage trees, assist to cool the air, eliminate pollution, collect rain, and preserving energy[1]. We have used this data to analyze various factors and trends that impact property sales and prices.

II. Related Works

A. Dataset 1: Trees in New York

Living in close proximity to greenery, often known as "green area" or "vegetation indices," has been linked to a number of health benefits. The two most essential components of urban vegetation, trees and grass, may have different effects on people's health, we predicted. While accounting for air pollution increased the beneficial associations between trees and self-reported health, it only somewhat negated the effects of park adjustments. Larger levels of surrounding greenness [2], higher percentages of green space near one's home [3], and objective assessments of the quantity and quality of vegetation in a community have all been associated to better self-reported health, according to a number of studies. [4]

B. Dataset 2: NYC Shooting Incident Data

In comparison to other comparable countries, the United States has a greater rate of gun related crime. Around 21% of gun owners have children, and the top motivator for criminals is self-defence. Every year, the number of gun owners in the United States rises. [6] A study was conducted in 2019 with the goal of understanding the rise in gun violence and the restrictions in place to limit the rate of shooting-related crime, as well as the time of year when criminal activity peaks. Also included are recommendations for the current trend in the rate of committed crime in the United States.[5] This might aid home-buyers in making informed decisions about where they want to live.

C. Dataset 3: NYC Property Valuation

Real estate valuation is an important part of a variety of activities such as financing, real estate, investigating investments, and calculating taxes. However, the most common use of real estate valuation is to determine the asking or purchase price of real estate. It can be difficult to evaluate a property because each property has unique characteristics such as Location, property size, floor plan, equipment. The value of real estate is influenced by general market fundamentals such as supply and demand in a particular region. Individual properties need

to be valued using one of many different methods to determine market value. [15] The Treasury assigns market value to all properties in New York City. In 2019, a paper was published which aimed at examining the increasing number of factors that a client considers while buying property in todays world and it was found that many minor factors have started affecting the price of properties in urban areas and the paradigm is slowly shifting to other features as the lifestyle of people is changing. [14]

D. Dataset 4: NYC Property Sales

Prices in the City surged during the most recent surge, particularly in Manhattan and the outer boroughs bordering Manhattan. According to a research undertaken by the Furman Center [10] to assess trends in NYC property prices, prices have risen by 124.2 percent in the city as a whole, and by up to 500 percent in Manhattan. Despite pricing instability in the past, the City's real estate market has been in good shape for the past 30 years. Household income was found to be a poor predictor of future neighborhood price changes. As a result, we'll try to figure out what other factors might affect people buying property in a given location.

III. Methodology

Knowledge Discovery in Database (KDD) was utilized to get useful insights from this research, where data was picked, then transformed and cleaned before being saved in the database. For Visualizing purposes, the saved data was retrieved once again. The research method is depicted in Fig. 1. We used Python for the entire project course because it

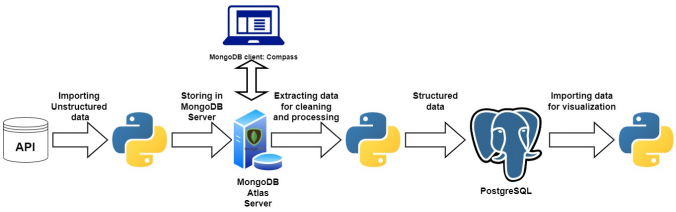

Fig. 1. Research process

provides a large number of libraries that assist data gathering, exploratory data analysis, visualization, and database connections. Because our retrieved data was in JSON format, MongoDB was the better approach that could store JSON data. MongoDB is an open-source, cross-platform document-based database designed for quick application development and scaling. This NoSQL database was designed by MongoDB Inc. We used MongoDB Atlas, a cloud-based database, for this project, and MongoDB Compass was used to visually see all of the databases and collections. We later put our data in PostgreSQL, an open-source object-relational database system, after retrieving it from MongoDB and sanitizing it. On a per-individual basis, we retrieved data from PostgreSQL after normalizing it as needed and then used visualization to generate replies to study questions based on the findings.

A. Dataset Description

1) Dataset 1: Trees in New York: The data was gathered from the open-source website NYC OpenData, and it consisted of real data values from a study of the number of trees conducted in New York City. In order to build patterns and make judgments, it was also necessary to get real data from reliable sources. The website provided a number of data retrieval methods, one of which was to use the SODA API to obtain JSON-formatted replies. The dataset has 1000 records with a total of 42 variables, such as:

- Tree ID, Block ID, and data creation for unique specification of the tree.
- Trees conditions, their identification in Latin, and Common name.
- Types of problems faced by trees and their components.
- The geographic locations of each tree help in locating each tree's data that has been recorded.

2) Dataset 2: NYC Shooting Incident Data: The NYC Open Data website [12] included information on the ages, genders, and other characteristics of victims and perpetrators involved in gunshot occurrences in New York. [7] It shows the number of fatalities by borough as well as the number of lethal occurrences. The crime rate in New York in 2006 is depicted in this graph. For the sake of analysis, I selected 3000 rows. As a result, this data collection was utilized to investigate New York's violence.

3) Dataset 3: NYC Property Valuation: Property Valuation: - This dataset contains information about the housing property valuation in different borough/county spread over distinct blocks. [13] The dataset includes varied properties of a house such as land area, type of property, tax class applicable, property valuation. The dataset is fetched through API consisting of 13 thousand 500 records consisting of 46 columns and is fetched from open-source NYC website with public API. [13]

4) Dataset 4: NYC Property Sales: We used the SODA API to get NYC property sales data from the NYC Open Data website. This dataset contains information about every building or building unit (apartment, condo, etc.) sold in the New York City real estate market. The location, address, type, market price, and sale date of building total sales are all included in this information. For better visualisations and understanding of the data, we chose 1000000 rows. The association between various columns may be seen in Fig 2 below.

B. DATA CLEANING AND HANDLING:

1) Dataset 1: Trees in New York: The dataset was retrieved in JSON format via the NYC open-sourcesource website's API. Because the original data recorded only included a maximum of 29 null values, accounting for only 2.9% of the total dataset, deleting rather than managing was the best option. We deleted a few columns from the dataset because they were not important to the research effort. The total number of variables in the final dataset was 28. The dataset

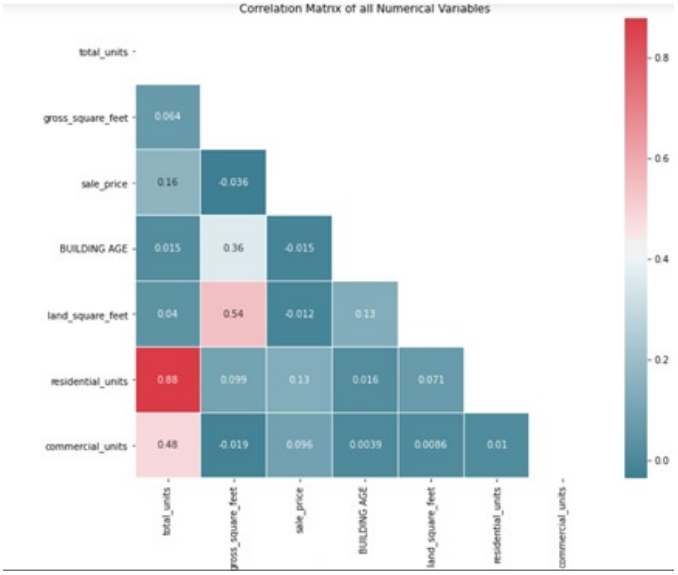

Fig. 2. Correlation Matrix

retained its datatype integrity while stored and fetched by MongoDB and PostgreSQL Database, hence no transformation was necessary.

2) Dataset 2: NYC Shooting Incident Data: The data was retrieved using the Socrata API with a 3000 row restriction. The data collection had 59.03 percent null values, thus the columns were eliminated. Unwanted columns that were not used throughout the analysis were removed [8]. The data is now free of null values. Integer, float, date-time, and category are the new data types. Data is stored in MongoDB and PostgreSQL databases and then retrieved for visualization.

3) Dataset 3: NYC Property Valuation: To begin with, the columns that are unrelated in property valuation such as phone extension, easement level are excluded from the dataset. All the null values were removed or handled with mean/median techniques. Zero values in the columns were converted to NotANumbers and were handled. Overall, 23 columns were found to be useful. Following is the handling process followed:

- Mean value of columns like stories of the building, area of land, land extension was applied.
- The data is also cleaned through database level by performing inner joins over the tables
- In PostgreSQL, "Property Valuation" Datasets with 1N and 2N where the data is entirely atomic and non-key attributes are completely functional and dependent on the primary key have been normalized.

4) Dataset 4: NYC Property Sales: We cleaned this dataset by deleting NaN values, removing unrelated columns, dividing down date columns into month and year, and adjusting data types.

C. DATA ANALYSIS AND VISUALISATION:

1) Dataset 1: Trees in New York: As part of the research, we have formed a few questions below on the distribution of trees spread across the city of New York. In Fig. 3, we can see which trees can be located in New York City. According to the analysis, Gleditsia triacanthos var. inermis trees are abundant when compared to all others.

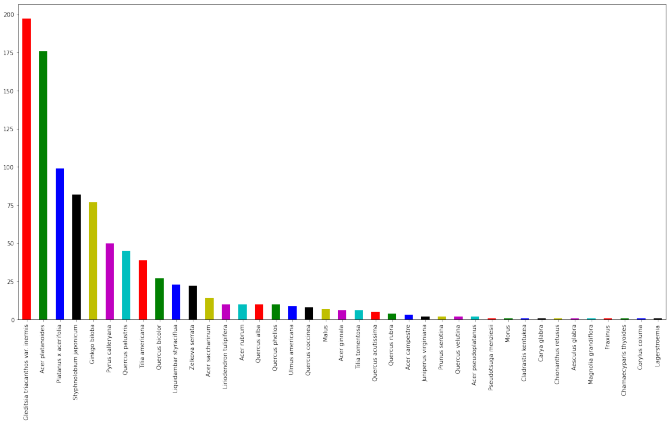

Fig. 3. Trees in New York (Latin Names)

76 percent of trees are in good health, 19 percent are in fair health, and the remaining 6% are in bad health. We can see the health of the trees in each borough of New York in Fig. 4, using Python's countplot. Brooklyn is first because it has the greatest number of trees in good condition, whereas Staten Island has the least amount of trees but none that are in poor condition.

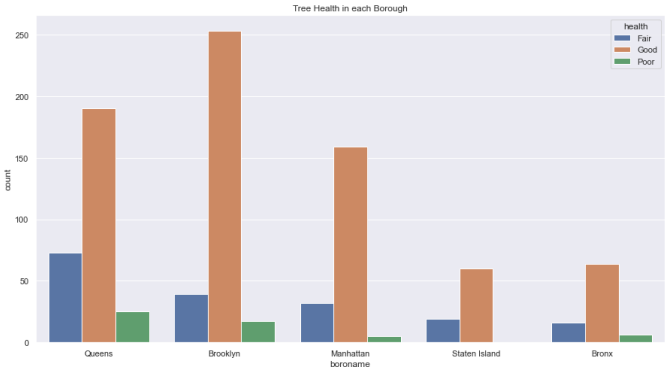

Fig. 4. Tree's heath in each NYC Borough

To visualize the tree and type of tree, in Fig. 5, using python's Folium libraries we have plotted every tree against given coordinate details in the dataset. The dataset also listed the problems that are caused to trees.

Using python's Scatter Mapbox we can visualize which tree in New York is facing certain kinds of problems as depicted in Fig 6. Using other visualization techniques we can conclude which tree problems are more in common.

2) Dataset 2: NYC Shooting Incident Data: The pie chart in Fig. 7, depicts the number of events per borough. We may deduce from the pie graphic that the majority of the crimes occurred in Brooklyn. The murder flag is depicted on the map Fig. 8, according to latitude and longitude. We may deduce

Fig. 5. Plotting Trees in NYC (Common Name)

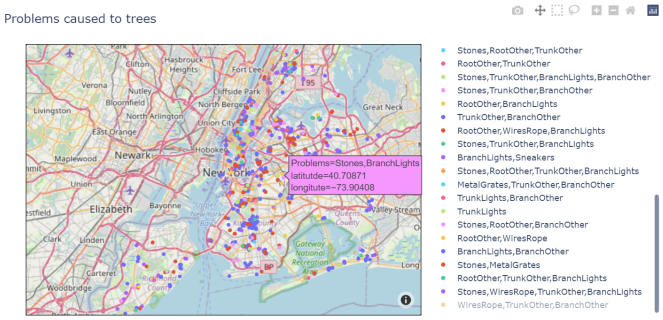

Fig. 6. Problems caused to Trees

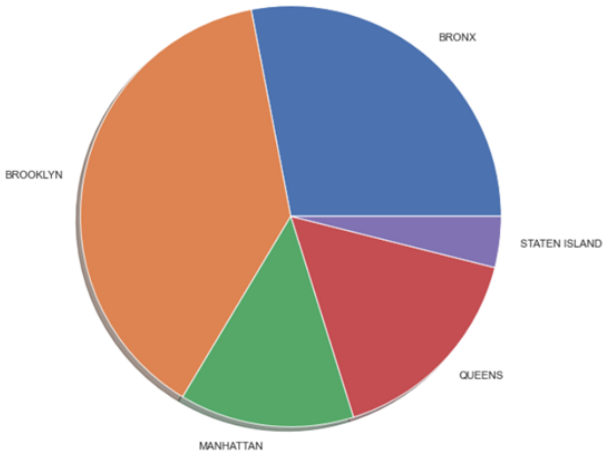

Fig. 7. Area-wise Shooting Incidents

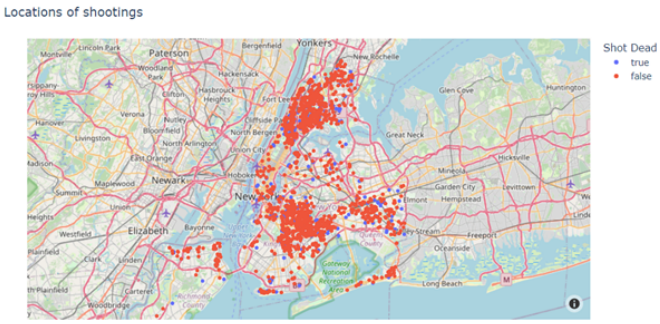

Fig. 8. Incidents involving shootings that resulted in death

from this that most of the victims were not killed. The murder flag is depicted on the map according to latitude and longitude. We may deduce from this that most of the victims were not killed.

In the bar graph shown in Fig. 9, the victims' age categories (18, 18- 24, 25-44, 45-64, 65+, Unknown) and the number of incidents are represented. The age groups 25-44 had the highest number of victims, as seen in the bar graph.

3) Dataset 3: NYC Property Valuation: The sunburst chart in Fig. 10, shows which streets among different boroughs have some of the costliest properties as well as those having cheapest properties. From the above distribution we can conclude that the property valuation is maximum for the areas in Manhattan. In Fig.11, the barplot is implemented for zip codes having top 20 property valuations. It can be seen that the maximum property valuation amongst the top 20 zip codes is at 10458. Fig. 12 is a scatter plot between "Property Block" and its "Valuation" which has sufficient data to conclude that most number of the houses lie in the property block 3700 or neighboring blocks.

Fig. 13, Shows the distribution of the percentage properties in each of the borough. It is evident that approximately 77% of properties are equally divided among "Queens" and "Staten is".

4) Dataset 4: NYC Property Sales: The most popular neighborhood names are depicted in the word cloud illustrated in Fig. 14 below. A word cloud is a textual data visualization that allows anyone to see the most frequently occurring terms within a body of text in a single glance. Word clouds are frequently used to examine, transmit, and evaluate data. There was no discernible trend in sales price over the course of the year studied. However, we can see in Fig. 15, below that sales at the start of the year are higher than sales at the end of the year. We can see in Fig. 16, the more expensive properties, as expected, are located near Manhattan. The sale prices in Queens and the Bronx are comparable. Staten Island has the most affordable housing options. Similarly commercial and residential units are priced highest in Manhattan.

IV. Results and Evaluation

To see if there is an indirect or direct impact, we combined New York's Property Valuation analysis with the city's Trees. As shown in Fig. 17, Manhattan has the highest valuation price, and it is the only city with a large number of trees. Manhattan is the smallest borough, although it has the highest property worth and the most trees. As a result, we can deduce that having natural vegetation nearby may have an impact on the value

Fig. 9. Highest number of Age group victimized

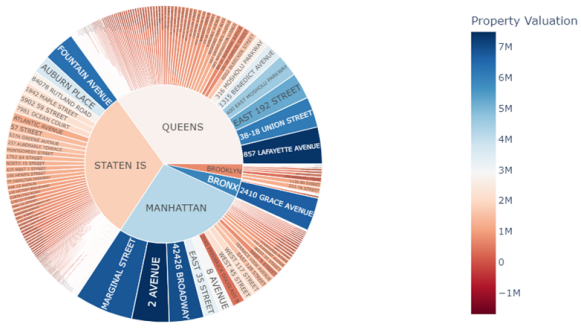

Fig. 10. Distribution of Property Valuation sum as per Street Address

of a property. So, if a person is looking for a good environment, they must pay a premium when choosing a location. The Fig. 18, was created by combining Dataset 2: NYC shooting incidents with Dataset 4: NYC property sales. We can see that shooting incidents are highest in Brooklyn, but sales are also very high. Staten Island has the fewest reported shootings, but sales there are lower than in other boroughs. As a result formed in Fig. 18, we can interpret that shooting violence has no direct impact on property sales.

V. CONCLUSION

In the project our aim was to perform analysis on impact of factors an individual might consider while searching a property in NYC like green vegetation, violence, sales and prices. Distinguish visualizations were showcased on above parameters throughout NYC. According to data, Manhattan is the most green and safe location, with the fewest reported crimes. In addition, the highest valuation and moderate sales were recorded, confirming Manhattan as the most favourable and popular location for an individual looking for property.

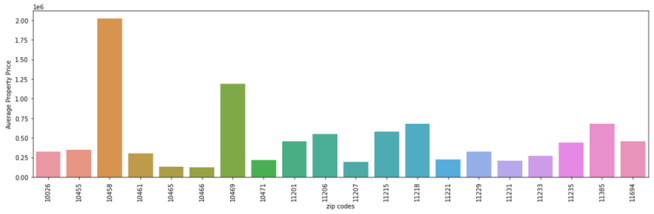

Fig. 11. Top 20 zip codes property valuations

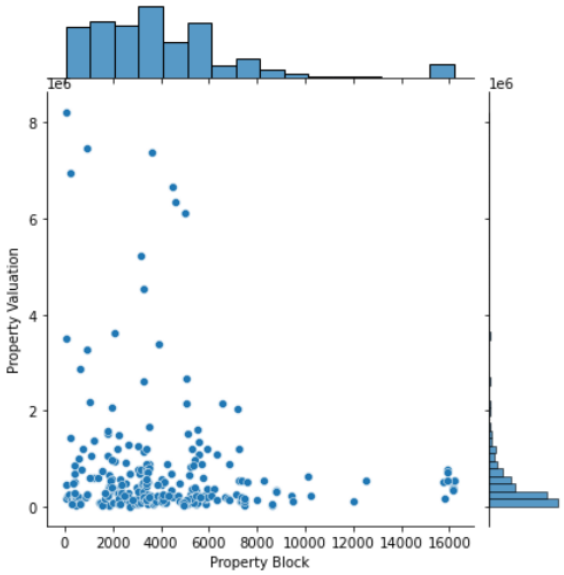

Fig. 12. Property Block vs Valuation

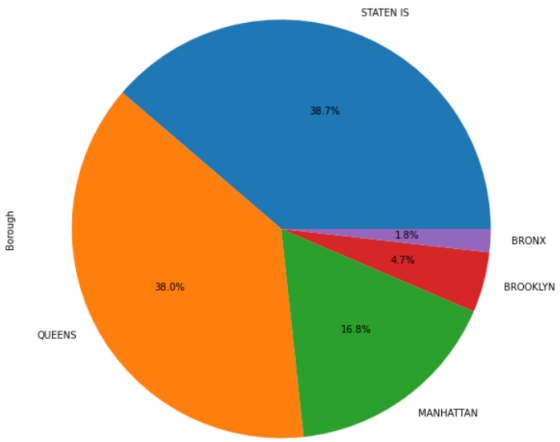

Fig. 13. Borough wise property percentage

Fig. 14. Popular Neighbourhood

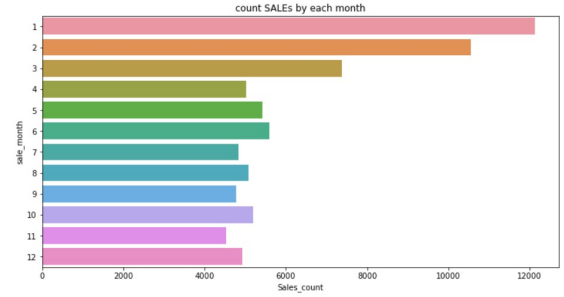

Fig. 15. Sales count by each Month

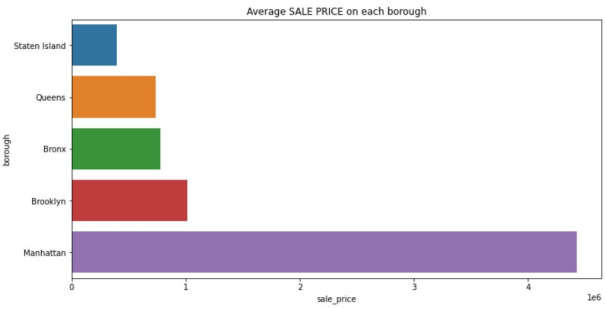

Fig. 16. Sales count by each Month

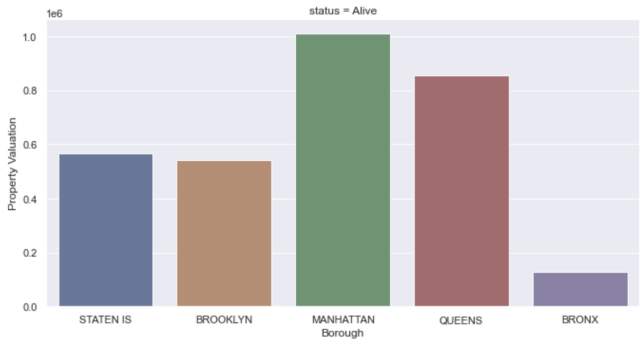

Fig. 17. Effects of Trees on Property Valuation

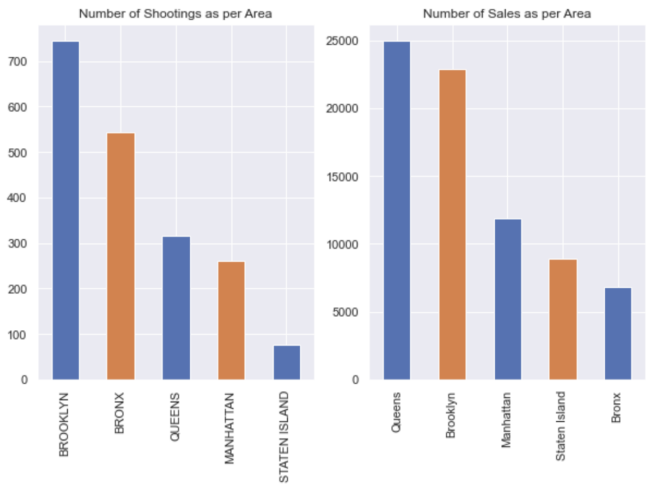

Fig. 18. Borough Wise Crime and Property Sales

REFERENCES

[1] NYC Parks Recreation,(2022), Retrieved from Great Trees of New York City website.https://www.nycgovparks.org/facilities/great-trees

[2] Triguero-Mas, M., Dadvand, P., Cirach, M., Martínez, D., Medina, A., Mompart, A., ... Nieuwenhuijsen, M. J. (2015). Natural outdoor environments and mental and physical health: relationships and mechanisms. Environment international, 77, 35-41.

[3] Maas, J.; Verheij, R.A.; Groenewegen, P.P.; de Vries, S.; Spreeuwenberg, P. Green space, urbanity, and health: How strong is the relation? J. Epidemiol. Community Health 2006, 60, 587–592.

[4] Reid, C.E.; Clougherty, J.E.; Shmool, J.L.C.; Kubzansky, L.D. Is All Urban Green Space the Same? A Comparison of the Health Benefits of Trees and Grass in New York City. Int. J. Environ. Res. Public Health 2017, 14, 1411. https://doi.org/10.3390/ijerph14111411

[5] Prediction of Crime in Neighbourhoods of New York City using Spatial Data Analysis —Abrar A. Almuhanna; Marwa M. Alrehili; Samah H. Alsubhi; Liyakathunisa Syed — 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA) — Year: 2021 — Conference Paper — Publisher: IEEE

[6] CriPAV: Street-Level Crime Patterns Analysis and Visualization — Germain Garcia-ZANABRIA; Marcos M. M. Raimundo; Jorge Poco; Marcelo Batista Nery; Claudio T. Silva; Sergio Franca Adorno de Abreu; Luis Gustavo Nonato — IEEE Transactions on Visualization and Computer Graphics — Year: 2021 — Early Access Article — Publisher: IEEE

[7] A Relationship Between Fines and Violent Crimes — Samuel Smith; Kedar Gangopadhyay; Simranjyot Singh Gill; Suzanne McIntosh — 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService) — Year: 2018 — Conference Paper — Publisher: IEEE

[8] Data visualization of crimes in a city using machine learning — Prabhat Sharma; Shreyansh Khater; Shilpi Sharma — 2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH) — Year: 2020 — Conference Paper — Publisher: IEEE

[9] Data Source for NYC Shooting Incident https://data.cityofnewyork.us/resource/833y-fsy8.json

[10] The furman center for real estate urban policy. Trends in New York City housing price appreciation (2008).https://furmancenter.org/files/Trends_in_NYC_Housing_Price_f

[11] Data Source for NYC Shooting Incident. https://data.cityofnewyork.us/resource/833y-fsy8json

[12] Data Source for NYC Property Sales. https://data.cityofnewyork.us/resource/w2pb-icbujson

[13] Data Source for NYC Property Valuation. https://data.cityofnewyork.us/City-Government/Property-Valuation-and-Assessment-Data/yjxr-fw8i

[14] Using Machine Learning to Forecast Residential Property Prices in Overcoming the Property Overhang Issue Lim Wan Yee;Nur Azaliah Abu Bakar;Noor Hafizah Hassan;Norziha Megat Mohd

[15] Property Tax, Supply and Demand Elasticity and Housing Price Weida Kuang;Hua Zhou