Mapping Urban Dynamics: Crime, Property, Subway Ridership, and Business in New York City

Amitkumar Patel
New York University
ap7986@nyu.edu

Kamalesh Neerasa
New York University
kn2359@nyu.edu

Stanley Jovel
New York University
lsj3272@nyu.edu

Viraj Parikh
New York University
vp2359@nyu.edu

***Abstract*—Motivated by the imperative need for enhanced safety measures and informed decision-making in New York City, this project embarks on a comprehensive investigation into the intricate relationship between crime rates and various facets of city neighborhoods. Through an amalgamation of NYPD Complaint Data, Property Valuation and Assessment Data, and Google Local Data, we aim to unravel how crime rates impact property values and business perceptions. By leveraging cutting-edge technologies such as Apache Zeppelin, Hive, Presto, Spark SQL, and Tableau, our study meticulously dissects these complex relationships. The analysis aims to reveal invaluable insights that offer actionable guidance for pedestrians, homeowners, policymakers, and urban planners alike.**

Pedestrians stand to benefit from access to enhanced information regarding safer routes and areas, empowering them to navigate the city with confidence. Homeowners will gain comprehensive insights into property values, facilitating informed decision-making within the real estate market. Furthermore, policymakers will be armed with data-driven strategies to enhance public safety, optimize infrastructure development, and foster sustainable economic growth. Through these insights, stakeholders can collaboratively work towards creating safer, more resilient, and prosperous communities across New York City.

I. Introduction

Crime rates are a significant concern for residents for any city. They affect not only public safety but also the overall character and economic well-being of a neighborhood. This project aims to gain a comprehensive understanding of how crime rates influence various aspects of New York City neighborhoods. We will achieve this by using big data analysis and combining several publicly available datasets.

First, our project aims to uncover the relationship between crime rates and property values by analyzing NYPD Complaint Data and Property Valuation and Assessment Data; this information could be valuable for potential homeowners and investors, as well as policymakers who are interested in revitalizing neighborhoods. In addition to property values, we aim to explore how crime rates affect local businesses. Using the Google Local Data Dataset to assess whether crime rates correlate with negative business reviews. This information can be helpful for business owners in determining the best locations and for city officials in developing strategies to support local businesses. Finally, the project will examine the relationship between crime rates and subway ridership using MTA Subway Stations data and ridership data. This analysis can help us understand how crime affects public transportation usage and inform strategies for improving public safety in subway stations and surrounding areas.

Overall, our project will provide valuable insights into the complex relationship between crime rates and various aspects of New York City neighborhoods. The findings can be used to inform urban planning decisions, public safety initiatives, and business development strategies.

II. Data Sources

The *NYPD Complaint Data* documents entries into the NYPD 911 System. It contains data regarding the type of incidents reported (one per row) such as burglary alarms, disputes, accidents, assaults, etc. which can help assess the prominence of such events in different city regions.

The *Google Local Data* contains review information on Google map (ratings, text, images, etc.), business metadata (address, geographical info, descriptions, category information, price, open hours, and MISC info), and links (relative businesses) up to Sep 2021 in the United States.

The *NYC Property Valuation and Assessment Data* contains detailed information on property valuation and assessment data for tax classes 1, 2, 3, and 4 in New York City. It includes market-assessed values, tax values, exemption values, and other related metrics necessary for calculating property tax, determining eligibility for exemptions, and assessing property valuations. The dataset is maintained by the Department of Finance and is updated annually. Using this dataset, our aim is to assess whether and how crime incidents influence property values across different neighborhoods.

The *MTA Subway Hourly Ridership Dataset* contains detailed information on the hourly ridership of passengers in the subway system throughout the city of New York. It comprises information on the time of travel, station details, number of ridership in that station within the hour, payment means, and other related metrics that can help analyze trends in ridership in New York over the past couple of years.

TABLE I
DATASET SIZE

Dataset	Size
NYPD Complaint Data	3.03 GB
Google Local Data	11.3 GB
NYC Property Valuation and Assessment Data [2]	1.15 GB
MTA Subway Hourly Ridership Data	8.8 GB

III. IMPLEMENTATION DETAILS

A. Data Ingestion

All datasets were exported from their respective data sources in the CSV format. It was transferred to the dataproc cluster from local using `gcloud compute scp <dataset_local_path> <dataset_remote_path>` and then ingested into HDFS using `hadoop fs –put <dataset_csv> <hdfs_path>`.

B. Data Pre-processing

This section describes the common pre-processing steps applied to all datasets. Spark jobs and SparkSQL were utilized for all pre-processing computations. The raw dataset file serves as the input, and the process generates output in the form of comma-separated values (CSV) containing extracted attributes. The only attribute maintained across all datasets is Zip Code. For missing Zip codes, the value is populated using latitude and longitude information (described in the next section). Records lacking both latitude, longitude, and Zip code are discarded.

C. Mapping Lat-Long to Zip Codes

The main units of analysis for our project are the zip code/postcode and borough. Some of our datasets have a zip code however the Subway Ridership dataset includes only borough information in the records. For those that contain empty, null or no zipcodes, we have created a comprehensive mapping of geo-locations (latitude and longitude) to the zip code and borough using open-source US Govt. Census Data. The missing zip code values are filled using the one which has the nearest haversine distance d given by:

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)$$
$$c = 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$$
$$d = R \cdot c$$

where ϕ_1, ϕ_2 are the latitudes of the two points, R is the radius of the Earth (mean radius = 6,371 kilometers or 3,959 miles), and d is the straight-line distance between the two points (along the surface of the sphere). This is referred to as the *reference dataset* in this report. It is broadcasted as a Broadcast Variable on each node for faster access.

D. Mapping Borough, Block and Lot (BBL) information to Zip Codes

In analyzing the Property Valuation dataset, a notable issue arose wherein a subset of records lacked valid zip code information. Given the absence of latitude and longitude coordinates per record, addressing this challenge demanded a novel approach. Leveraging the Borough-Block-Lot (BBL) data present in relevant records, we interfaced with the NYC Department of Finance's Property Information Portal [3], a repository offering geospatial data based on BBL inputs.

Utilizing this geospatial data, we inferred missing zip code values and populated them within a dataframe. Subsequently, through a data fusion process, this enriched dataframe was merged with the Property Valuation Dataset, effectively rectifying zip code inconsistencies. This meticulous methodology significantly enhanced the precision and integrity of our analytical insights.

E. Data Profiling

Data cleaning was performed using Zeppelin notebooks. The general process involves:

- **Data Loading**: loading of raw data in a suitable format (e.g., DataFrame).
- **Data Preprocessing:**
  - Removing unnecessary or irrelevant columns.
  - Standardization of data formats (for example, date and time).
  - Deriving additional features, if applicable (e.g., zip codes from addresses).
  - Filtering data based on specific criteria (e.g. location).
  - Handling missing values (e.g., imputation, removal).
  - Validating data integrity (e.g., checking for invalid values).
- **Data Profiling**: Analyzing the cleaned data to understand its characteristics, such as:
  - Final record count.
  - Minimum and maximum values for numeric fields.
  - Distribution of categorical variables.
  - Identifying outliers or anomalies.
- **Persisting the cleaned data**: Saving the cleaned and profiled data to a shared HDFS directory for further analysis.

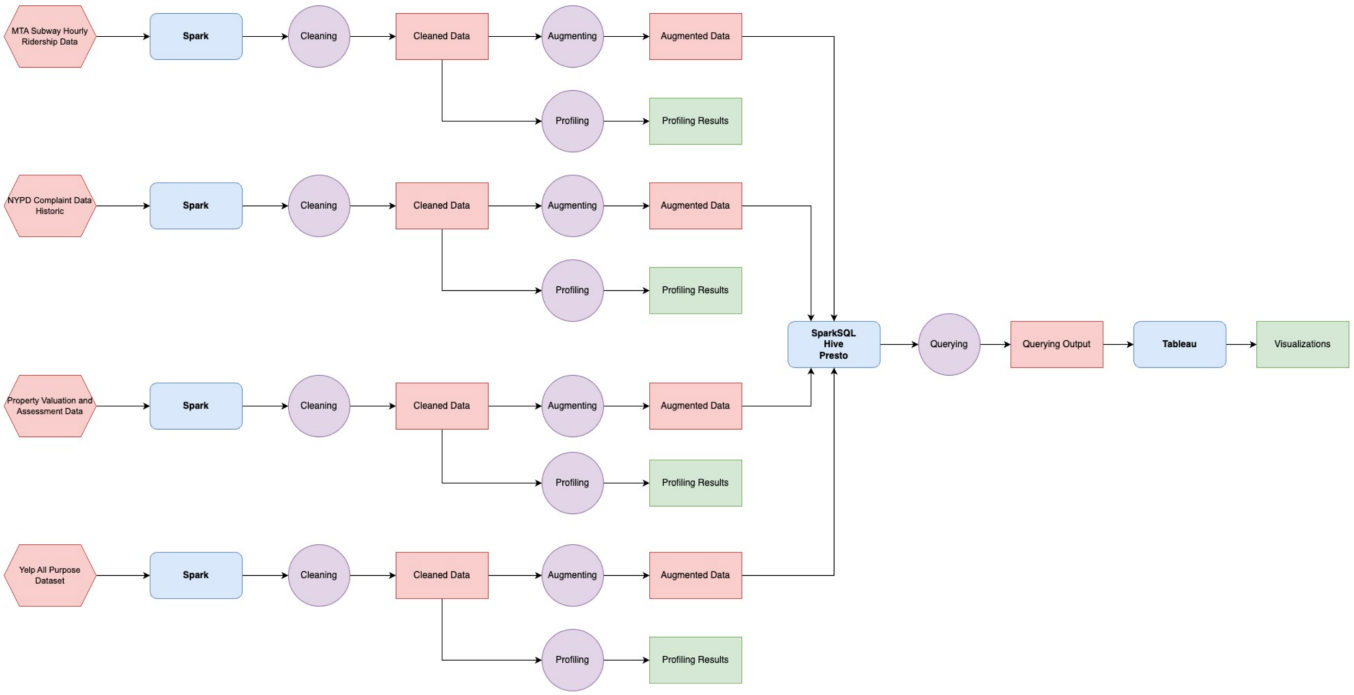

Fig. 1. Design Diagram

F. Combining Datasets

The sanitized datasets were merged based on zip codes to generate statistical insights on a zip code level from our data. Given our objective to assess the influence of crime complaints on urban dynamics such as Business Reviews, Subway Ridership, and Property Valuation, each of these datasets was integrated with the Crime Complaints dataset to derive comprehensive statistics at the overall, annual, and borough levels.

G. Querying using Hive

For any analytics project, visualization is a key aspect. We chose to leverage Tableau's powerful visualization features to achieve our objectives. However, to load our datasets from HDFS directly into Tableau, we made use of Hive and Presto. Hive was used to create the Metastore and external tables. We then loaded our data into these tables, and using Presto drivers, queried our datasets into Tableau for visualization purposes.

IV. Results and Analysis

Our analysis of the relationship between crime rates and various aspects of New York City neighborhoods yielded several significant findings.

A. Is crime selectively impacting certain businesses?

Our investigation into urban crime dynamics underscores a significant correlation between criminal incidents and businesses with high foot traffic and cash transactions (see Figure 2). Notably, establishments such as restaurants, salons, coffee shops, bars, and grocery stores emerge as focal points for criminal activities. The bustling nature of these businesses, coupled with frequent cash exchanges and valuable inventory, renders them particularly susceptible to theft, vandalism, and other illicit behaviors. Moreover, the social dynamics inherent in these settings, including crowded environments and varied patronage, exacerbate security challenges. Recognizing these patterns is crucial for formulating effective crime prevention strategies tailored to the specific needs of these businesses and the communities they serve, ultimately contributing to safer urban environments.

B. Does crime influence customer sentiment towards local businesses?

Our analysis of crime data and Google reviews revealed a Pearson's correlation of 0.032, signaling a lack of correlation between total complaints and average restaurant reviews in most areas as shown in Figure 11. However, neighborhoods like The Bronx and Brooklyn have a negative correlation, on the contrary, tourist hotspots such as Astoria, Jackson Heights, Williamsburg, Dumbo, Lower Manhattan, Chelsea, Clinton, and Midtown [4] showed a positive correlation, suggesting that the high volume of tourists may mask the impact of crime on local reviews, compare green areas on Figure 11 versus dark blue areas in Figure 12.

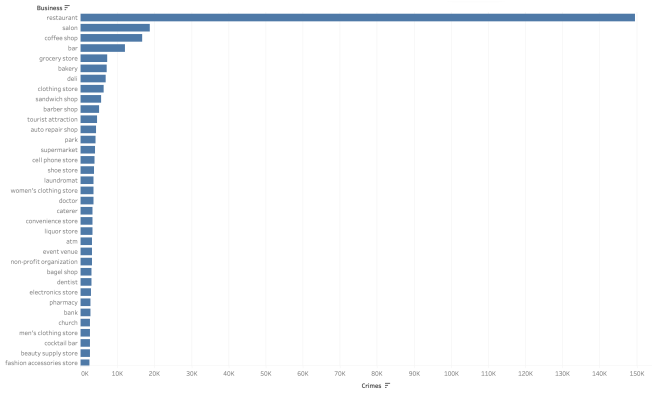

Fig. 2. Distribution of number of crimes per business category

C. Do crime complaints influence Subway Ridership in New York City?

While trying to find whether there is any relationship between Subway Ridership frequency and complaints based on regions, we observed a strong positive correlation. This initially surprised us, however, we soon came to terms with an important factor - greater ridership also indicates populous areas, and the frequency of crimes also drastically increases. This is why Manhattan where industries like Business, Tourism, etc attract huge counts of people is bound to witness higher ridership and increased crimes.

Contrastingly, when we shifted focus to areas within Brooklyn and Queens, we noticed that areas like Williamsburg, Astoria, and Downtown Brooklyn have relatively higher subway ridership as compared to other areas in Brooklyn and Queens, but lower complaints registered (Fig 3). This most certainly implied that these are safer areas and subway lines to travel in. Conversely, neighborhoods like Crown Heights and the Bronx, along with subway lines such as the Blue (A), Red, and Green (2, 3, 4, 5) lines into Brooklyn, suggest lower foot traffic and higher complaint counts, thus implying relatively unsafe localities to commute or live in.

We then dived further down into this analysis by increasing the granularity in terms of ridership and complaints over different times of the day. For this purpose we divided 24 hours into three parts essentially: Night (12 AM-6 AM), Morning - Noon (7 AM - 3 PM), and Evening - Night (4 PM - 12 AM), and achieved interesting results.

What we observed is that however there is a stark contrast in ridership levels between Morning and Night, the complaints level in both these parts of the day remain consistent as seen in Figs(4, 5, and 6). This suggests that while ridership may not be directly correlated with complaints in a region, it is certainly associated with a greater foot count, and with a greater number of people, the frequency of crimes is bound to increase.

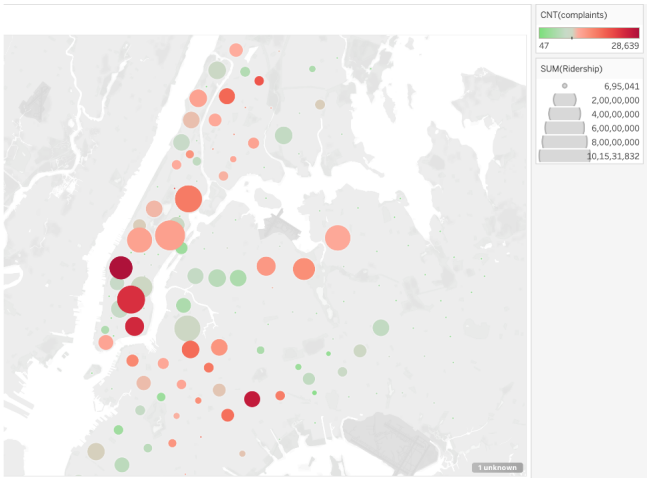

Fig. 3. Geographical Distribution of Subway Ridership and Complaints

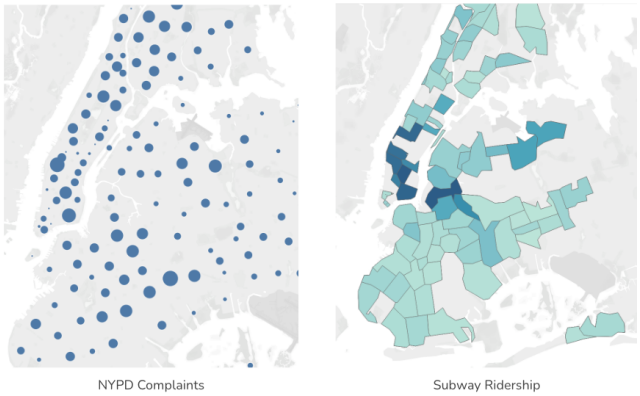

Fig. 4. Geographical Distribution of Subway Ridership and Complaints: 12 AM - 6 AM

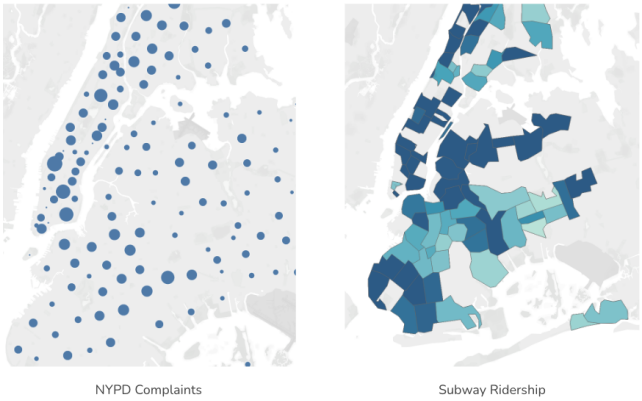

Fig. 5. Geographical Distribution of Subway Ridership and Complaints: 7 AM - 3 PM

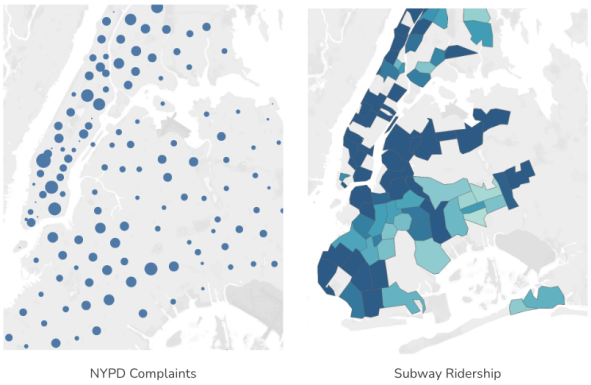

Fig. 6. Geographical Distribution of Subway Ridership and Complaints: 4 PM - 12 AM

Our analysis revealed a strong correlation between ridership and complaints. This correlation is validated by the understanding that higher ridership often correlates with increased foot traffic, subsequently impacting crime rates. We also made use of Pearson's correlation coefficient which suggested a strong positive correlation between the two variables in consideration as shown in Table II.

TABLE II
SUBWAY RIDERSHIP AND COMPLAINTS CORRELATION COEFFICIENT

Pearson's Correlation
0.39042712333790075

D. Does crime influence property valuation in New York City?

In our examination of the Property Valuation dataset, our analysis unveils intricate spatial dynamics regarding property valuations across New York City's zip code regions. Notably, midtown Manhattan emerges as a hub of heightened property valuation metrics, displaying the highest average total property valuation (Fig.7) and land valuation per zip code (Fig.8). These findings align with the area's recognized status as a locus of premium real estate assets.

Furthermore, our investigation identifies distinct valuation patterns in areas such as the vicinity of JFK Airport and coastal zones like Breeze Point and Roxury. These observations underscore the heterogeneous economic influences shaping property valuations across diverse geographical contexts within the city.

Expanding our scrutiny to Manhattan's borough-wide landscape, we discern a more uniform distribution of average land values per zip code, contrasting sharply with the pronounced divergences observed in other parts of New York City. Noteworthy among these disparities are the elevated complaint rates documented in lower Manhattan, select zones of North Brooklyn, and lower Queens—a trend inversely correlated

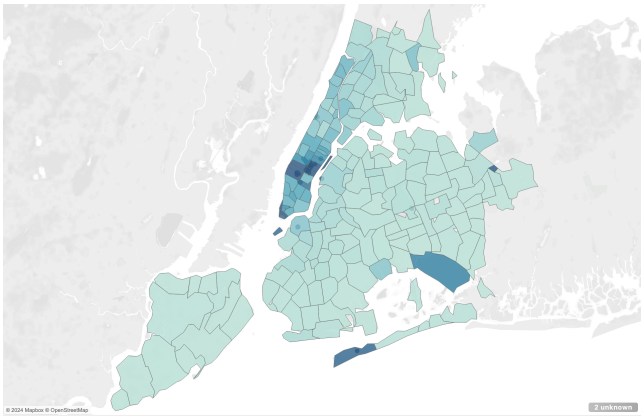

Fig. 7. Geographical Distribution of Average Property Total Valuation for Year 2022

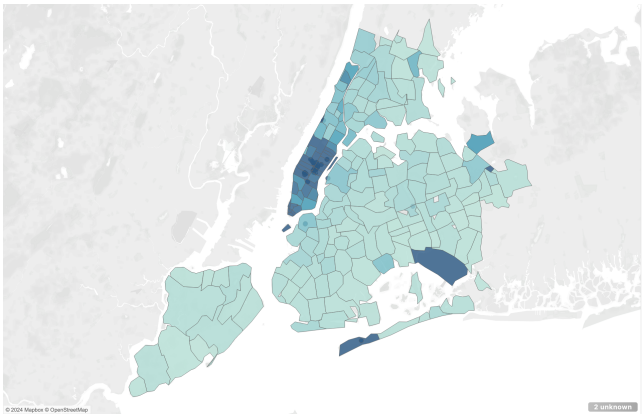

Fig. 8. Geographical Distribution of Average Property Land Valuation for Year 2022

with property valuations in these regions. However, our temporal analysis of property values and complaint volumes reveals a nuanced narrative. Despite localized fluctuations, the year-over-year Pearson's correlation coefficient between property values and complaints for the years 2020 through 2022 as seen in Table.III consistently yields small negative values, indicative of a tenuous and inconclusive correlation. This underscores the multifaceted interplay between property valuations and socio-economic dynamics, necessitating further inquiry into underlying causal mechanisms.

TABLE III
CITYWIDE AND YEARWISE CORRELATION COEFFICIENT

Year	Total Value vs Compl.	Land Value vs Compl.
2020	-0.13695946886941152	-0.12992674825249084
2021	-0.11774119261128072	-0.11880280017388564
2022	-0.10296746891451626	-0.10983337705914004

In addition to our spatial analysis, we conducted a borough-wise and yearly examination of correlation coefficients, revealing consistent trends across all boroughs except Staten Island. Notably, Staten Island exhibited a distinctive pattern

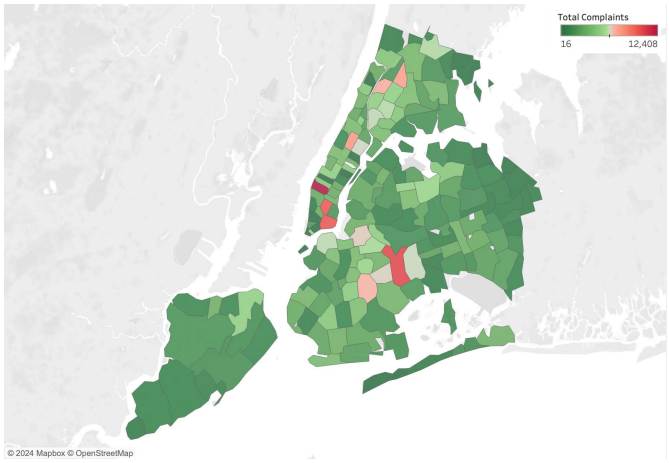

Fig. 9. Total Crime Complaints Per Zip Code for Year 2022

TABLE IV
YEARWISE CORRELATION COEFFICIENT FOR STATEN ISLAND BOROUGH

Year	Total Value vs Compl.	Land Value vs Compl.
2020	0.7075383831627579	0.5845798707084664
2021	0.7552447576713144	0.6779665947897546
2022	0.24476347383559444	0.22846937764494685

characterized by a positive and robust correlation, ranging from (0.70, 0.58) in 2020, peaking at (0.75,0.67) in 2021, and subsequently attenuating to a weaker correlation of (0.24,0.22) (see Table IV).

We hypothesize that these anomalous findings in Staten Island may stem from unique socio-economic and demographic characteristics specific to the borough. For instance, Staten Island's distinct suburban environment, relatively lower population density, and different market dynamics compared to other boroughs may contribute to a stronger correlation between property values and complaint volumes. Additionally, factors such as community cohesion, law enforcement practices, and neighborhood development initiatives could influence residents' reporting behaviors and perceptions of safety, thus impacting the observed correlation between property values and complaints. Further research into these contextual factors is warranted to elucidate the underlying mechanisms driving the observed correlations in Staten Island.

Furthermore, we investigated the annual trend of average property values citywide in relation to total complaints per zip code. As presented in Fig.10, our analysis reveals a notable divergence: while the aggregate average complaints per zip code witnessed an upward trajectory from 2020 to 2022, the overall average actual land value across New York City remained relatively stable, experiencing a marginal decrease. Concurrently, the total average actual property value across the city exhibited a decline from 2020 to 2021, with a partial

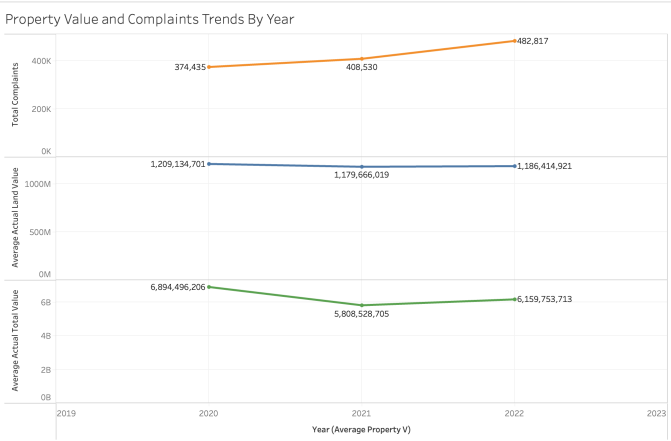

Fig. 10. Year-wise trends of Average Property Values and Complaints

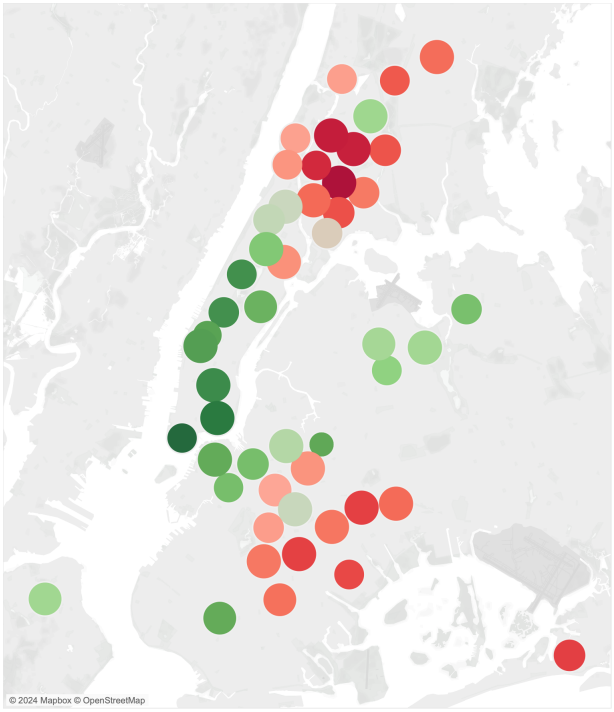

Fig. 11. Restaurant reviews in top 100 most dangerous regions. Larger circles denote more crime, green denotes good reviews

recovery observed in 2022, albeit not reaching pre-pandemic levels. This trend is posited to have arisen in response to the COVID-19 pandemic, which precipitated a decline in rents and property prices due to diminished demand. This pattern also underscores the weak correlation between crime complaints and property prices, suggesting that while crime complaints increased over the observed period, property values remained relatively stable or experienced only minor fluctuations.

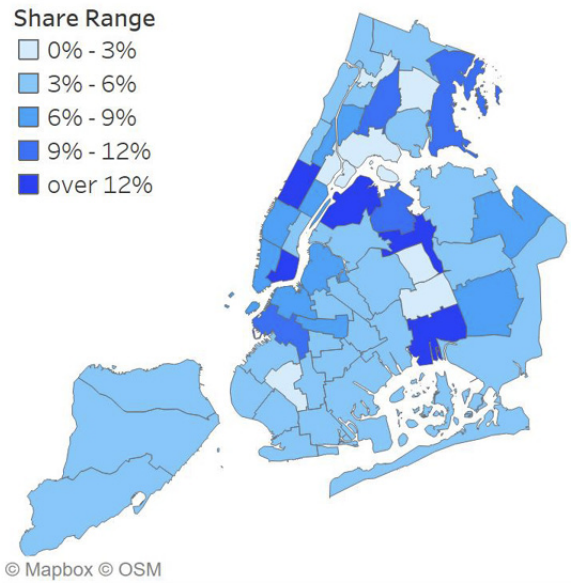

Fig. 12. Tourism Share of Employment by Neighborhood. Office of the New York State Comptroller

V. Challenges

A. Refining Business Analysis by Category

The sheer volume of business subcategories within the Google Reviews dataset (over 3,000) presented a significant challenge. While we explored leveraging Large Language Models (LLMs) for automated clustering, the immense data size proved incompatible with readily available models, resulting in unreliable category groupings.

B. Refining Business Analysis by Category

The task of categorizing businesses within the Google Reviews dataset, which comprises over 3,000 subcategories, presented a formidable challenge. Despite exploring the potential of Large Language Models (LLMs) for automated clustering, the sheer magnitude of the dataset rendered conventional models ineffective, leading to unreliable category groupings.

C. Data Consistency Concerns: Addressing Date Variability

Initially, inconsistent date formats across datasets posed a significant obstacle. However, upon deeper analysis, we realized that the immediate impact of crime reports might not manifest instantaneously in surrounding areas; rather, it could exert a gradual influence over time. This insight enabled us to make informed decisions based on the available data, mitigating the impact of date discrepancies.

D. Managing Large Datasets: A Case Study of Subway Ridership

The sizeable nature of certain datasets, such as Subway Ridership, posed formidable challenges in data sourcing. Attempts to retrieve data from online sources were met with recurrent export errors, attributable to the substantial dataset size overwhelming the data source server.

Through iterative experimentation, we discerned that the sheer volume of data was the root cause of these challenges. To circumvent this issue, we employed the query tool provided by the data source to selectively filter data based on specific years and relevant columns of interest. This strategic approach enabled us to obtain a more manageable dataset size, facilitating seamless extraction from the online source and facilitating further analysis.

VI. Summary and Conclusion

This study delved into the intricate dynamics between crime rates, business performance, subway ridership, and property values in New York City (NYC), revealing nuanced insights with profound implications across multiple domains.

Our analysis unveiled a discernible correlation between certain business types, such as restaurants and shops, and elevated crime rates. These establishments, characterized by high foot traffic and cash transactions, are more susceptible to criminal activities. This understanding equips businesses with valuable insights to devise targeted strategies aimed at mitigating crime risks and bolstering security measures. Interestingly, despite experiencing higher crime rates, tourist hotspots in NYC consistently maintained positive reviews. This phenomenon suggests that the allure of these destinations often overshadows concerns regarding crime, thereby masking its impact on visitor perceptions and experiences.

While subway ridership itself does not exhibit a direct correlation with crime rates, areas with high ridership tend to demonstrate lower complaint volumes. This observation hints at the presence of enhanced safety measures or favorable perceptions of safety among commuters, contributing to a safer public transit environment.Moreover, our analysis revealed a weak negative correlation between crime rates and property values in NYC. Despite the presence of crime, the enduring demand for properties in the city mitigates its impact on property values to some extent. This insight provides valuable guidance for homeowners, investors, and policymakers in navigating the complex interplay between crime and property values.

These findings offer actionable insights for various stakeholders in NYC. Businesses can adopt targeted security measures to mitigate crime risks. Policymakers can utilize this knowledge to enhance public safety initiatives in specific areas. Homeowners gain a better understanding of how crime rates might influence property values. Finally, urban planners can leverage this information to develop safer and more resilient communities, fostering a more positive overall environment for residents and visitors alike.

VII. Acknowledgement

We are grateful to New York University for providing us with a free student license for Tableau, which significantly facilitated our data visualization tasks. Additionally, we extend our sincere thanks to the NYU HPC team for ensuring the Dataproc cluster remained operational throughout our project, enabling us to efficiently ingest and process large datasets. We are also grateful to the organizations that generously open-sourced their data for general reference. Finally, we express our deep appreciation to Professor Yang Tang for his invaluable guidance and for introducing us to a diverse range of cutting-edge concepts and technologies throughout this course. His knowledge and insights have been instrumental in our learning and research journey.

References

[1] NYPD Complaint Data Historic, *Data Source:* https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[2] NYC Property Valuation and Assessment Data Tax Classes 1,2,3,4, *Data Source:* https://data.cityofnewyork.us/City-Government/Property-Valuation-and-Assessment-Data-Tax-Classes/8y4t-faws/about_data

[3] NYC Department of Finance Property Information Portal *Data Source:* https://propertyinformationportal.nyc.gov/

[4] DiNapoli, Thomas P. "The Tourism Industry in New York City Report." New York State Comptroller's Office, April 2021. *Data Source:* https://www.osc.ny.gov/files/reports/osdc/pdf/report-2-2022.pdf.

[5] NYU Furman Center - Has Falling Crime Driven New York City's Real Estate Boom? *Data Source:* https://furmancenter.org/files/publications/Has_Fallen_Crime.pdf

[6] NYC Crime Map *Data Source:* https://www.arcgis.com/apps/instant/sidebar/index.html?appid=8153f961507040de8dbf9a53145f18c4

[7] MTA Ridership Returns to New York *Data Source:* https://comptroller.nyc.gov/reports/riders-return/.

[8] MTA (Subway and Bus) Ridership Reports *Data Source:* https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2023#:~:text=The%20subway%20has%20a%20daily,238%20local%20bus%20routes.