New York County Housing Market Analysis

Induni Sandapiumi Nawarathna Pitiyage[1], Sara Campolattano[2]
[1,2] University of Milan Bicocca, CdLM Data Science
Matricola[1]: 906451, Matricola[2]: 906453

Abstract

This project, *New York County Housing Market Analysis*, provides an in-depth study of housing trends by leveraging data-driven techniques. The analysis is structured into four main phases: data acquisition, optimization, storage and analysis. Initially, housing data is collected through APIs and web scrapping techniques. The acquired data undergoes quality assessment, profiling, integration and enrichment to optimize its management and reliability. The processed data is stored efficiently for further analysis.

Contents

1. Introduction

The real estate market in New York County is dynamic and influenced by various economic, social and infrastructural factors. Understanding housing trends, price distributions and property availability is crucial for investors, policymakers and potential homeowners.

The study is divided into four main sections. First, data acquisition is conducted using APIs and web scraping to gather real estate listings and relevant housing information. Next, the data management phase ensures data quality through profiling, integration and enrichment to enhance usability and accuracy. The MongoDB storage is used to handle the data for the implementation of efficient queries and analysis.

Finally, data analysis focuses on key queries include calculating the average price per square foot, determining the total number of available properties per neighbourhood, identifying properties closest to amenities and highlighting the highest-listed individual housing units. The findings provide valuable insight into the New York County housing market.

2. Data Acquisition

In the process of data acquisition phase, two commonly used techniques API (Application Programming Interface) integration and web scrapping are involved to provide the comprehensive view of the housing market in New York County.

2.1 API

API integration was used efficiently to retrieve structured data from trusted and accessible Sources. The New York Housing market dataset available on Kaggle platform was accessed via Kaggle's API. The original dataset contains information on 4801 properties and 17 attributes:

- BROKERTITLE - Title of the broker.
- TYPE - Type of the house.
- PRICE -Price of the house.
- BEDS - Number of beds.
- BATH - Number of baths.
- PROPERTYSQFT - Square footage of the property.
- ADDRESS - Complete address of the house.
- STATE - State of the house.
- MAIN_ADDRESS - Main address information.
- ADMINISTRATIVE_AREA_LEVEL_2 - Administrative area level 2 information.
- LOCALITY - Locality
- SUBLOCALTY - Sub-locality
- STREET_NAME - Name of the Street.

- LONG_NAME - Complete name of the street.
- FORMATTED ADDRESS – Formatted house.
- LATITUDE – Latitude of the property location.
- LONGITUDE – Longitude of the property location.

Additionally, OpenStreetMap (OSM) was used through its API to acquire its precise geographic coordinates (longitude, latitude) of the essential amenities:

- Schools
- Public stop positions (bus-stops, train-stops)
- Hospitals

2.2 Web Scraping

Web Scraping was employed to get insights further about the neighbourhood of the listed house locations, data is extracted from this website: https://www.walkscore.com/score/.

Extracted data contains information on walk score, transit score and bike score which corresponds to neighbourhood of the houses that listed in the dataset obtain from Kaggle platform.

- Walk Score – A metric that measures the walkability of a specific location which evaluates how convenient to access everyday amenities.
- Transit score - Evaluates the availability and usefulness of public transportation.
- Bike score - measures how bike-friendly the area is.

3. Optimizing Data Management:

Effective data management plays a major role in making informed decisions in any organization. This process begins with the assessing data quality to ensure accuracy, consistency and completeness. Data profiling plays a critical role in this phase by providing a detailed understanding of the data structure, content and potential issues, allowing to identify and address anomalies or inconsistencies on early stage.

Once the data quality and profiling established, data integration enables seamless consolidation of disparate datasets to create unified view across a system. To further enhance the value of the data, data enrichment adds contextual information, transforming raw data to gain actionable insights of to maximize the data utility and supports strategic decision-making.

3.1 Data Quality and Data Profiling Assessment.

To ensure the dataset reliability, several data quality checks were conducted for the dataset obtained by the Kaggle platform. First, duplicate records were identified and

quantified based on the *BROKERTITLE* and the *MAIN_ADDRESS*, highlighting cases where identical entries were present. In the initial dataset, the original data shape consisted of 4,801 entries and 17 columns. This dataset is filtered to include only properties located in New York County, the data shape reduced to 1,295 entries and 18 variables. Within this subset, 64 duplicate rows were identified, representing cases where identical records were appeared multiple times. These duplicate entries were removed from the dataset to ensure data integrity, resulting in a clean dataset with 1,231 rows and 17 columns.

Geospatial coordinates (*LATITUDE* and *LONGITUDE*) were validated to ensure they fell within the bounds of New York County (40.4774, 40.9176) for latitude and (-74.2591, -73.7002) for longitude. After this validation, all 1,231 rows were confirmed to have valid geospatial coordinates, with no records classified as invalid.

Additionally, fields such as *BEDS* and *BATH* were examined for inconsistent or invalid values including negative numbers or unrealistic ranges, which could indicate data entry errors. Decimal values in these fields were also standardize to integers for uniformity.

Furthermore, the completeness of address-related fields such as *ADDRESS* and *LOCALITY* was assessed to identify missing or incomplete information, ensuring the dataset provide a comprehensive representation of property details.

A compressive data profiling process was conducted to understand dataset's structure, quality and distribution. Missing values were not identified across various columns, ensuring no critical information gaps remained. A detailed variable overview was performed using the summary statistics, providing insights into variability for the numerical fields and data distribution analysis was performed in *4.1 Data Analysis section* to provide a detailed understanding of the dataset, aiding in decision-making and further analysis.

3.2 Data Integration and Data Enrichment.

To enhance the dataset's value and analytical potential, a data integration and enrichment process was carried out. Multiple datasets were combined, including the data scraped from various web pages providing walk scores, transit scores and bike scores were integrated together based on the neighbourhood location. Since the extracted data from the webpage were based on the location found on the dataset obtained through the Kaggle API, were merged to ensure its consistency. Additionally, geographic data fetched form OpenStreetMap (OSM), including the coordinates of the amenities: schools, hospitals and public transportation stop locations are merged into one dataset.

The data enrichment process added valuable attributes to the housing dataset, such as walk scores, transit scores and bike scores providing insights into accessibility and liveability. Furthermore, the enriched data included proximity features, calculating the

distance from the listed housing location to nearby schools, hospitals and public transportation stations or stops. This integrated and enriched dataset enables deeper analysis and feature creation, supporting comprehensive evaluations of the housing locations.

After completing the data quality and data profiling assessment, the data integration and data enrichment phase, the resulting dataset composed with 1,182 rows and including the following variables:

- BROKERTITLE - Title of the broker.
- TYPE - Type of the house.
- STATE – State in which the property belongs to.
- Neighbourhood – Neighbourhood of the property location.
- Location – Location of the property
- Loc_City - City in which the property is located.
- Loc_ZIP – Postal ZIP code of the property location
- LATITUDE – Latitude of the property location.
- LONGITUDE – Longitude of the property location.
- PRICE – Price of the property.
- BEDS – Number of bedrooms of the property.
- BATH – Number of the bathrooms of the property.
- PROPERTYSQFT – Square-feet of the property.
- Walk Score – Walking score of neighbourhoods of the property locations.
- Transit Score – Transit score of neighbourhoods of the property locations.
- Bike Score – Bike score of neighbourhoods of the property locations.
- Dist_transportation_meters – Minimum distance to the nearest public transportation stop or station from the property location.
- Dist_hospital_meters - Minimum distance to the nearest hospital from the property location.
- Dist_school_meters – Minimum distance to nearest school from the property location.

These variables provide a detailed view of the real estate listings, integrating the geographical coordinates, property features, accessibility metrics and the proximity to essential services such as transportation, hospitals and schools.

4. Data Storage

To handle the diversity and the scale of data effectively, a NoSQL database, MongoDB is selected for storage due to its flexibility in managing unstructured and semi-structured data. This structure ensures seamless integration and querying of data, enabling efficient retrieval and analysis for task involving geographical insights, property evaluations and accessibility assessments.

4.1 Data Analysis

In real estate analysis, data analysis helps determine property valuation, identify factors influencing prices (e.g: location, size, amenities) and assess market trends. The descriptive statistics (*Table 1.1*) and correlation analysis (*Figure 1.1*) provide insights into the distribution of variables in the dataset.

	PRICE	BEDS	BATH	PROPERTY SqFt	Walk Score	Transit Score	Bike Score	Dist Transportation meters	Dist Hospital meters	Dist School meters
count	1182	1182	1182	1182	1182	1176	1182	1182	1182	1182
Mean	5.6e+06	2.9	2.46	2357.622	95.582	99.359	82.99	145.044	1123.39	0.002
Std	6.3e+07	2.5	1.97	1864.468	12.513	3.353	11.10	105.026	827.98	0.0014
Min	2.5e+03	1	0	230	19	48	39	9.865	59.13	0.0001
25%	6.8e+05	2	1	1375.25	98	100	82	72.728	555.72	0.0015
50%	1.4e+06	3	2	2184.20	99	100	84	123.118	900.73	0.0023
75%	3.8e+06	3	3	2184.20	100	100	90	196.094	1440.70	0.0033
max	2.1e+09	32	20	24000	100	100	99	1841.295	6389.13	0.0144

Table 1.1: Descriptive Statistics.

The *table 1.1* shows that the average properties available around New York County, price approximate around $5.6 million, but with the high standard deviation of $62.9 million, indicating the presence of extreme outliers with maximum price reaching $2.1 billion. Properties typically have 3 bedrooms and 2 bathrooms with sizes averaging 2357.6 sqft.

Furthermore, it is evident by the strong correlations (*Figure 1.1*) depicted below that the larger properties tend to have more bedrooms and bathrooms (correlation of 0.79 between beds and baths, 0.70 between bath and property size). Typically, larger properties tend to have higher prices, though correlation is moderate, this may probably be due to the visual scenery (such as skyline and water views, park-facing units etc.) and skylight and natural units (eg: high-floor apartments receive better natural light and tend to be expensive than the lower floor units.).

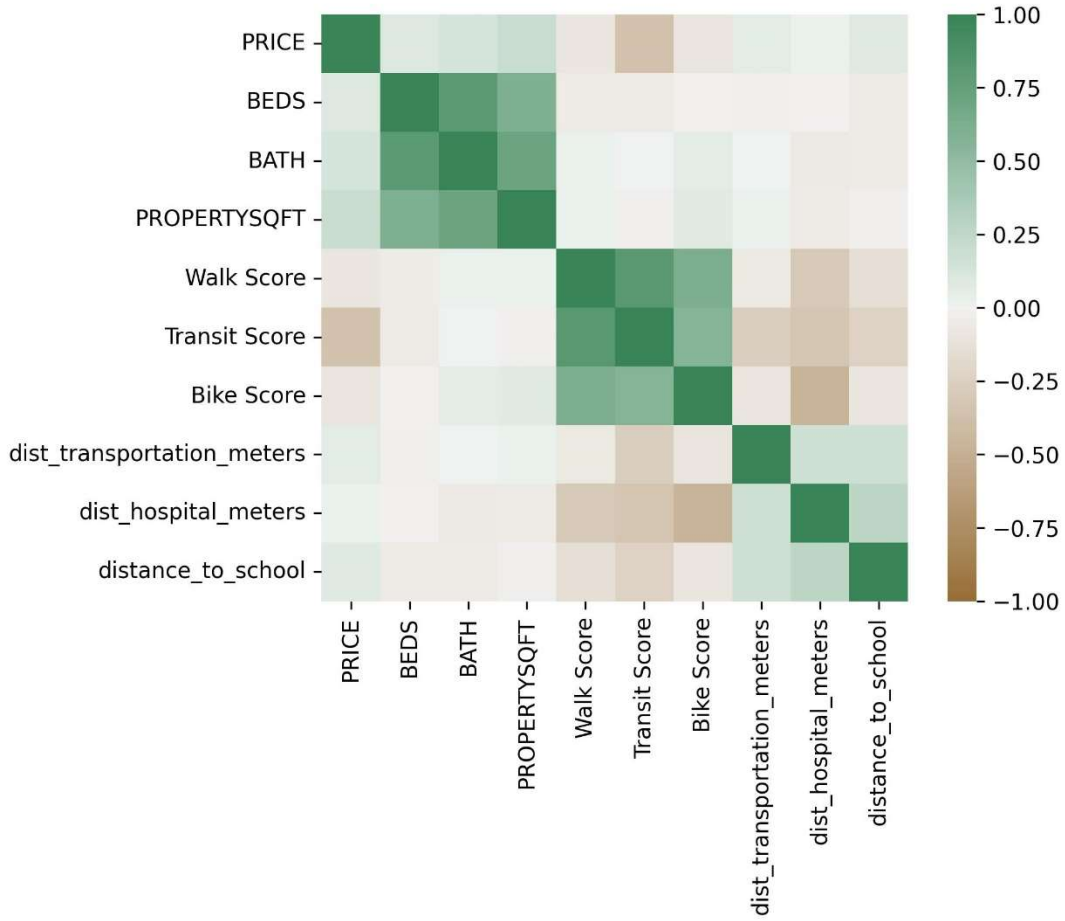

Figure 1.1: Correlation Heatmap.

Interestingly, we observe negative correlation between property prices and the accessibility scores (Transit score, Bike score and Walk score) revealing following dynamics:

- Wealthier area prefers private transport: Affluent neighbourhoods may prioritize car ownership and have fewer transit options.
- Proximity to noisy transit hubs: Areas with high transit accessibility often include high density, lower income housing or commercial zones, which may not be attractive for the high-end buyers.

4.2 Queries

To gain deeper insights into the spatial distribution of the property prices and density, the following analysis is conducted using MongoDB queries.

4.2.1 Calculating the number of properties, average price and property SqFt available for each neighbourhood.

To analyse the property trends across different neighbourhoods, we calculated the average price per property, average property square footage and the total number of available properties respective each New York County neighbourhoods. *Table 1.2* followed by the *figure 1.2* give insights about the average property prices across the neighbourhoods.

Neighbourhood	Number of Properties	Average Price (in millions)	Average Property SqFt
SoHo	61	8.09	3306.0
Tribeca	16	8.07	3132.0
Hell's Kitchen	72	6.67	2318.0
Upper East Side	157	5.58	2651.0
Chelsea	71	5.63	2663.0
East Village	11	5.26	4393.0
West Village	42	4.21	2353.0
Garment District	3	3.32	2065.0
Manhattan	300	3.25	2192.0
Upper West Side	71	2.97	2443.0
Greenwich Village	32	2.78	2039.0
Murray Hill	49	2.65	2085.0
Two Bridges	20	2.23	2387.0
Midtown East	99	1.99	1998.0
Financial District	19	1.98	1532.0
East Harlem	25	1.73	1984.0
Battery Park City	6	1.61	1328.0
Randalls-Wards Island	16	1.58	2786.0
West Harlem	15	1.52	2972.0
New York	23	1.31	1609.0
Central Harlem	36	1.05	2130.0
Roosevelt Island	3	0.83	1696.0
Long Island City	1	0.82	930.0
Travis – Chelsea	1	0.70	1538.0
Washington Heights	19	0.67	1877.0
Inwood	9	0.53	1806.0
Kingsbridge	3	0.46	1936.0
Gravesend	1	0.19	600.0

Table 1.2: calculated table for average price per property, average property square footage and the total number of available properties respective each New York County neighbourhoods.

Table 1.2 reveals significant difference in real estate pricing and availability. SoHo and Tribeca top the list with highest average property prices, at $8.09 and $8.07 million, respectively with the spacious average property sizes exceeding 3,100 sqft. Hell's Kitchen, Upper East Side and Chelsa price ranges from $5 million to $6.67 offering properties in-between 2,300 to 2,663sqft. Notably, East Village

comprises with the largest average property size (4,393 sqft) despite a relatively moderate price of $5.26 million.

Figure 1.2: Average Neighbourhood locations with property prices.

On more affordable end, neighbourhoods like Central Harlem ($1.05 million), Roosevelt Island, Chelsa, Washington Heights, Inwood offer lower cost properties with sizes between 1,800 sqft to 2,130 sqft. Kingsbridge and the Gravesend have the lowest prices, at $0.46 and $0.19 but also features the smallest availability and sizes. Manhatton comprises lot of properties and averages $3.25 million per property with an average size of 2,192 sqft, reflecting the diverse range of pricing and property size.

4.2.2 Adding a new field: PRICE_PER_SqFt.

Introducing a new metric, PRICE_PER_SqFt calculated from the existing variables (PRICE/PROPERTYSQFT) provides a clearer comparison of property values across different neighbourhoods in New York City. *Table 1.3* followed by the *figure 1.3*

depict below shows the retrieved average price per square feet for the respective neighbourhood.

Neighbourhood	Number of Properties	Average price per sqft
Tribeca	16	2243.0
SoHo	61	2130.0
Upper East Side	157	2020.0
Chelsea	71	1860.0
Hell's Kitchen	72	1745.0
Garment District	3	1656.0
West Village	42	1656.0
Greenwich Village	32	1399.0
East Village	11	1299.0
Manhattan	300	1298.0
Upper West Side	71	1186.0
Financial District	19	1181.0
Battery Park City	6	1151.0
Murray Hill	49	1097.0
Two Bridges	20	1039.0
Midtown East	99	924.0
East Harlem	25	902.0
Long Island City	1	887.0
New York	23	821.0
Randalls-Wards Island	16	589.0
Roosevelt Island	3	563.0
Central Harlem	36	533.0
West Harlem	15	462.0
Travis – Chelsea	1	454.0
Washington Heights	19	409.0
Inwood	9	340.0
Gravesend	1	316.0
Kingsbridge	3	272.0

Table 1.3: Retrieved table for average price per square feet for the respective neighbourhood.

Tribeca leads with the highest average price per square foot at $2,243 followed closely by SoHo at $2,130 and the Upper East Side at $2,020 reflecting their high-end real estate market.

In contrast, more affordable neighbourhoods like Washington Heights ($409 per sqft), and Inwood ($340 per sqft) offer low-cost alternatives. Gravesend, at $316 per sqft, and Kingsbridge ($272 peer sqft) highlighting further budget friendly options.

This metric is essential for investors and buyers seeking the best value for their money, as it standardizes the property regardless of the total size, making direct comparisons across the neighbourhoods.

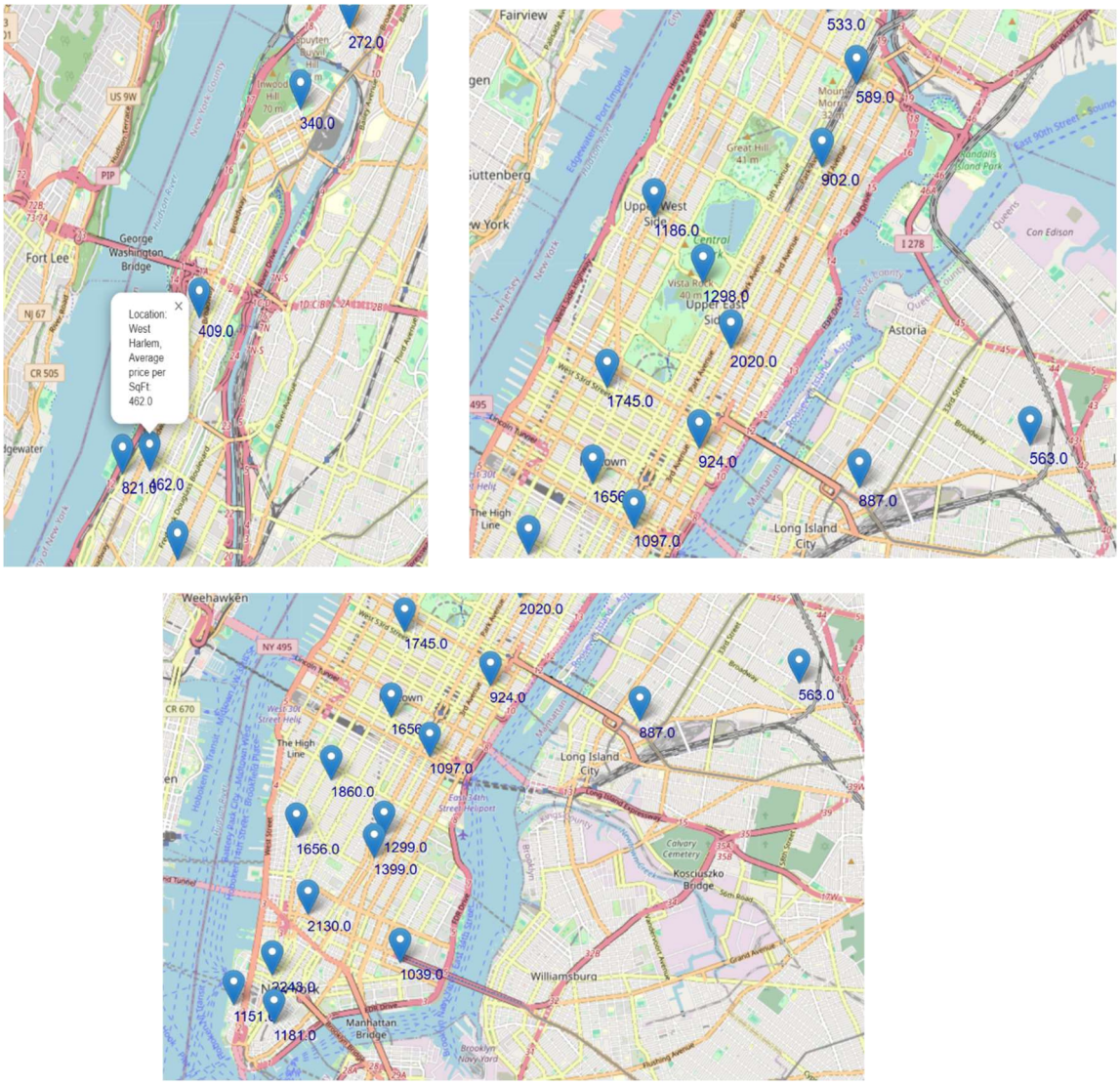

Figure 1.2: Average Neighbourhood locations with average price per square feet.

4.2.3 Finding the properties closer to Amenities.

Analysing the proximity to key amenities (such as transportation hubs, schools and hospitals), reveals crucial insights for buyers and investors seeking convenience and accessibility.

A total of 866 individual properties are located within the 200 meters of public transportation stations/stops and boast a transit and walk score greater than 70, highlighting their strong connectivity. Among these individual properties Manhattan leads with 214 such properties, emphasizing its dense urban structure. *Figure 1.5, plot (a)* summarizes the number of available properties and the neighbourhood in which it belongs to.

Proximity to school is a key factor for families and investors looking for properties in well-connected neighbourhoods. A total of 1,071 individual properties are located within 200 meters of a school while maintaining walk, bike and transit score above 70. Manhattan tops the list with 270 such properties, making it a

attractive choice for families, followed by the other neighbourhoods such as Upper East Side (157 properties), Midtown East (99 properties) etc. (summarizes in the *figure 1.5, plot (b)*)

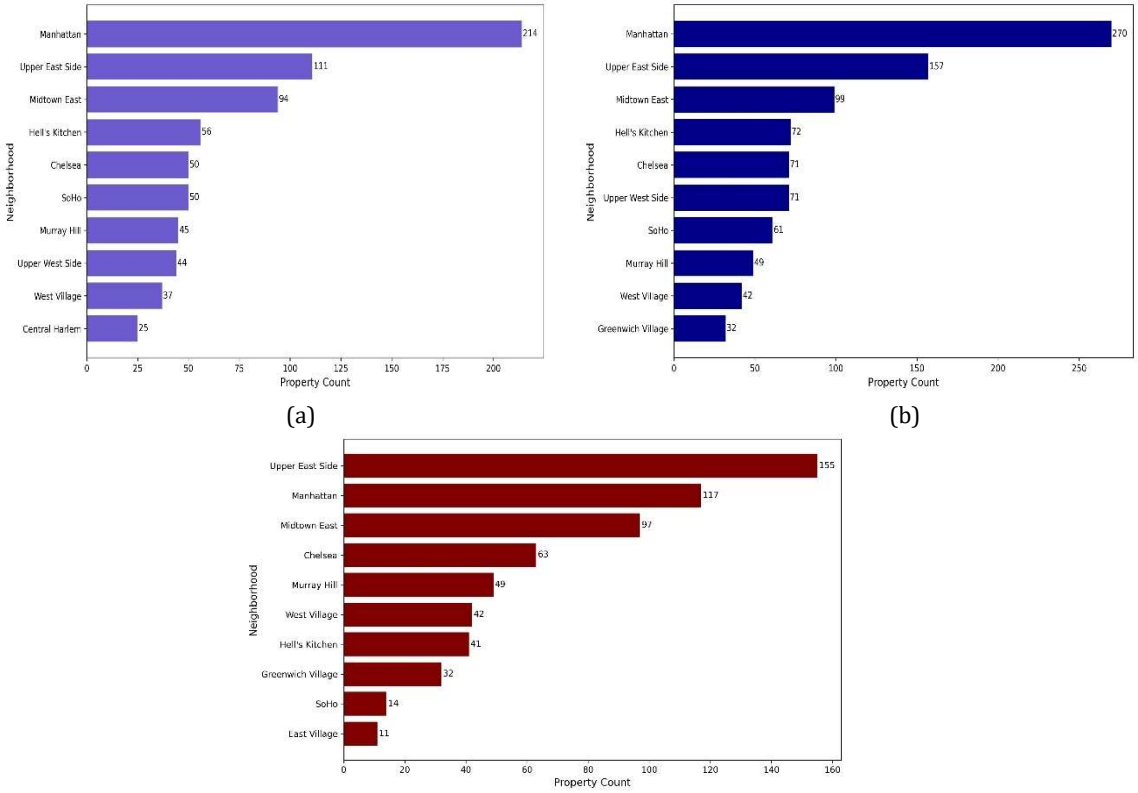

Figure 1.5: (a) Number of listed properties by the neighbourhood within 200m away from the public stations/stops. (b) Number of listed properties by the neighbourhood within 200m away from the schools. (c) Number of listed properties by the neighbourhood within 1000m away from the hospitals.

Access to healthcare facilities is a crucial factor for many homebuyers in urban areas. A total of 638 individual properties are located within 1000 meters of a hospitals maintaining a transit or walk score above 70, ensuring easy access to medical services. As depicted *in figure 1.5, plot (c),* Upper East Side leads with 155 individual properties, followed by Manhattan (117 properties), Midtown East (97) , making prime location for those prioritizing proximity to healthcare.
These findings emphasize the importance of the locations for individuals seeking properties convenient access to their day-to-day needs.

4.2.4 Highest Individual Housing listed.

New York City's luxury real estate market is home to some of the most prestigious and high-priced properties in the world. *Table 1.4* shows the most extravagant listings, with the highest priced housing listed in the dataset.

Broker Title	Neighbourhood Location	Price in Millions ($)
Broker by Serhant	Hell's Kitchen	195
Broker by Compass	Upper East Side	60
Brokered by Douglas Elliman - 575 Madison Ave	Hell's Kitchen	56
Brokered by Douglas Elliman - 575 Madison Ave	Manhattan	55
Brokered by Nest Seekers International, Midtown	SoHo	50
Brokered by Sotheby's International Realty	Upper East Side	44
Brokered by Douglas Elliman - New Development	Tribeca	40
Brokered by Douglas Elliman - 575 Madison Ave	Upper East Side	36
Brokered by Corcoran West Side	Chelsea	34
Brokered by Corcoran East Side	Upper East Side	32
Brokered by Reserve	SoHo	32
Brokered by Corcoran Chelsea/Flatiron	Hell’s Kitchen	31
Brokered by Sotheby's International Realty	Chelsea	29.95
Brokered by Christie’s Int. Estate Group	Murray Hill	29.20
Brokered by Peter Ashe Real Estate	Upper East Side	28.50
Brokered by Compass	SoHo	28
Brokered by The Agency	Chelsea	27.75
Brokered by Garfield, Leslie J. & Co.	West Village	27
Brokered by Douglas Elliman - 575 Madison Ave	Manhattan	27
Brokered by BHHS New York Properties	SoHo	26
Brokered by Serhant	Upper East Side	26
Brokered by Douglas Elliman - 575 Madison Ave	Upper East Side	25.75
Brokered by Brown Harris Stevens - 1926 Broadway	Manhattan	25
Brokered by CORE Group Marketing, LLC	Chelsea	25

Table 1.4: Highest-priced listed property with the respective broker title.

The highest priced property is listed in Hell’s Kitchen for $195 million, brokered by Serhant. Following this Upper East Side boasts a $60 million property listed by Compass. Other notable listings include Hell’s Kitchen ($56 million, Brokered by Douglas Elliman - 575 Madison Ave), Manhattan ($55 million, Brokered by

Douglas Elliman - 575 Madison Ave) and SoHo ($50 million, Brokered by Nest Seekers International, Midtown).

The Upper East Side leads in luxury listings, with 7 properties followed by Chelsea and SoHo, each having 4 luxury properties. *Figure 1.6* depict the luxury individual property locations.

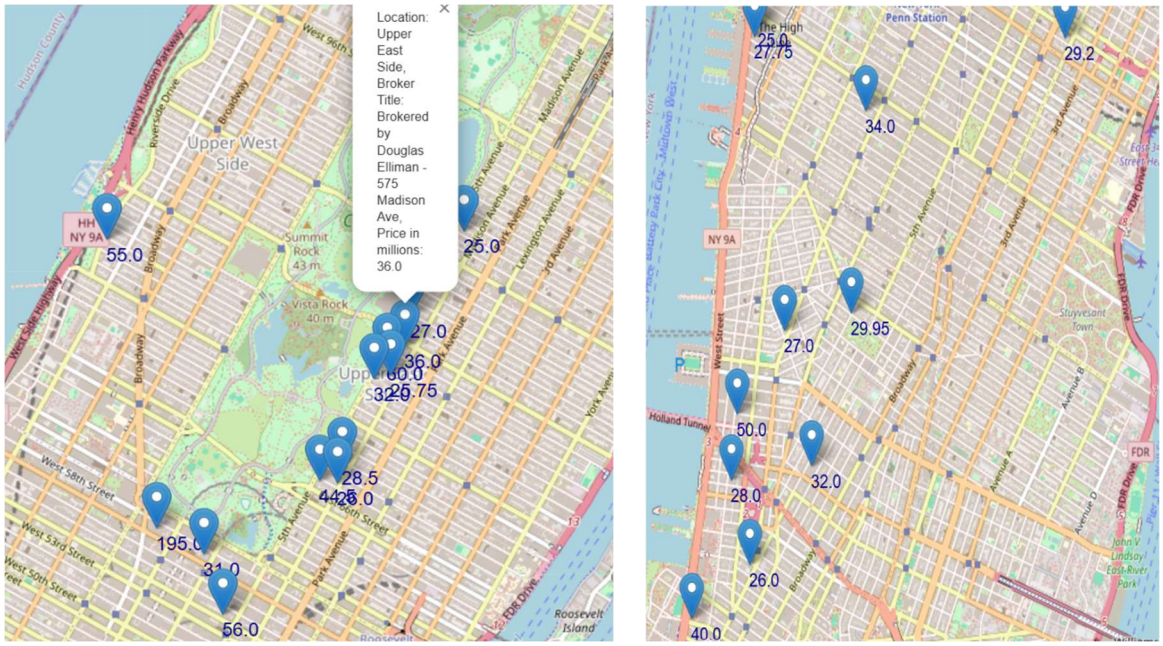

Figure 1.6: Highest- priced property locations.

Among these extravagant listings, many of the most listed apartments are located along the Central Park, offering the residents stunning skyline and park views.

5. Conclusion

This project demonstrates the end-to-end process of acquiring, merging and analysing the real estate data. By leveraging APIs and web scrapping techniques, we efficiently gathered extensive real estate data. Optimization data management through quality assessment, integration and enrichment allowed for accurate and meaningful analysing. Storing the data effectively enabled the execution of queries.

This analysis carried out using queries highlights variations in real estate pricing, property sizes and availability across New York City's neighbourhoods. SoHo and Tribeca stand out as the most expensive areas, offering luxury properties with high price-per square-foot values, while more affordable alternatives like Central Harlem, Washington Heights and Inwood provide budget conscious buyers with viable options. Manhattan, with its diverse range of property prices and sizes, remain a central hub for real estate investment.

Additionally, for investors and buyers, the price-per-square-foot metric serves as an essential tool for standardizing value assessment across the neighbourhood, allowing for more informed decision making.

Beyond pricing, location-based factors play a critical role in property valuation. The accessibility of public transportation, schools and healthcare facilities significantly enhances a property's desirability. Manhattan leads in well-connected properties, reinforcing its appeal for buyers prioritizing convenience and lifestyle.
Furthermore, this analysis also reveals insights about exceptional range of high-end properties, with Central park-facing apartments among most desirable. The Upper East Side stand out with the highest number of luxury insights. Ultimately, this study underscores the importance of location, connectivity and amenities in shaping real estate trends providing valuable insights for both perspective homeowners and real estate investors.

6.Refernces

Data acquisition through Kaggle API :-
https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market

Data acquisition based on web scrapping:- https://www.walkscore.com/score/.

https://drive.google.com/drive/folders/1Kcq1D0hdrQpcMNh7pSlb7RyoESeGa6rO?usp=sharing , at this link all the materials used to develop are available.