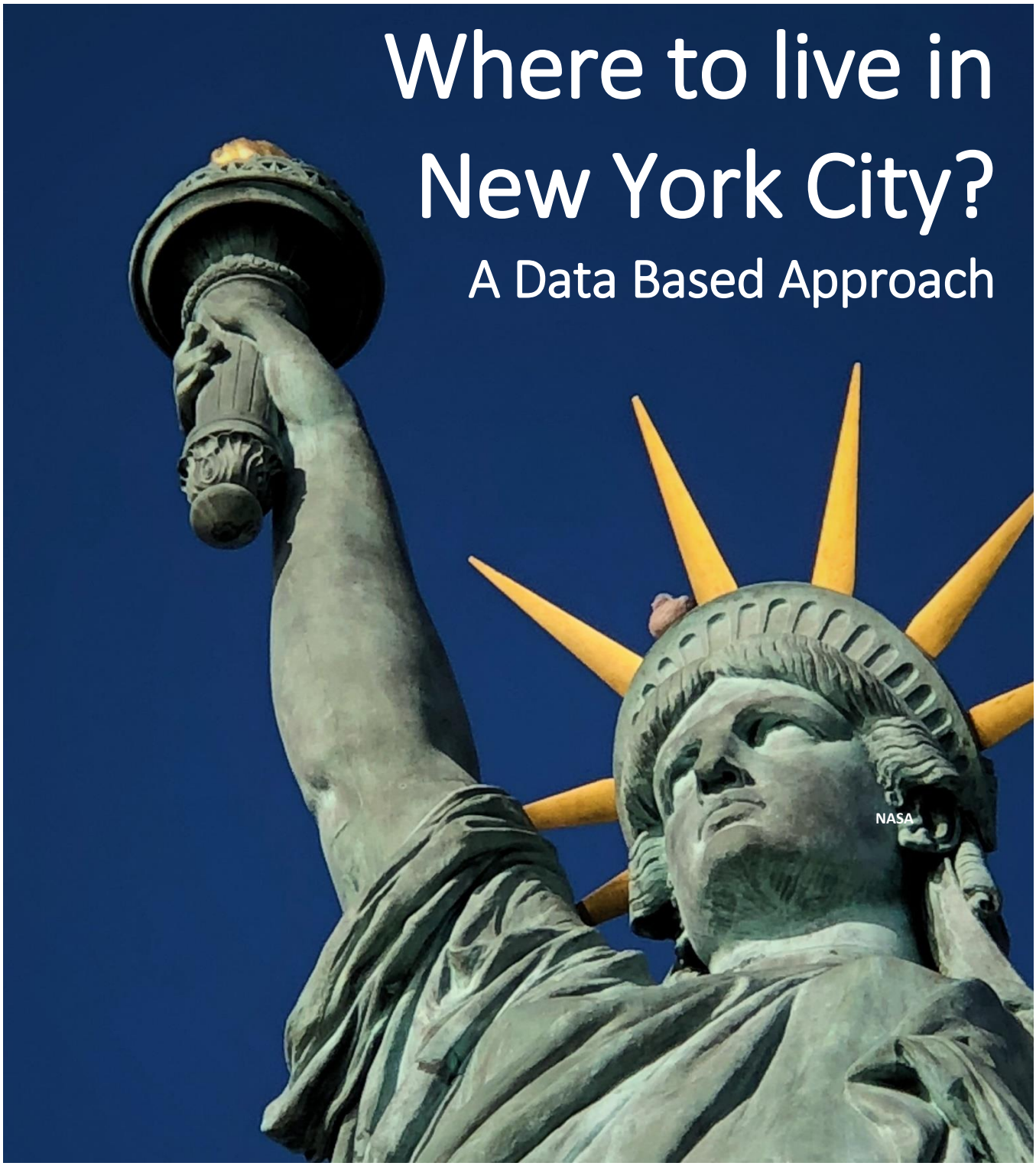
# Where to live in New York City?

## A Data Based Approach




Ankur Arya| 5th April 2019| https://www.linkedin.com/in/ankur-arya

# ABSTRACT

Data science to employed to search places which are best to live in New York city. The search is done by using real and recent data obtained from NYC Open Data, Zillow, Google and Foursquare. Types of data include residential property prices, crime reported to NYPD, real traffic-based commute time, restaurant cuisines, shopping complexes, parks and education institutions. The data is extracted, compiled and manipulated for analysis and visualization. Insights gained from data is used to compare boroughs and neighborhoods of the city. Using examples of short list criteria, best neighborhoods are chosen. The work flow allows future integration with richer data and complex selection methods.

# TABLE OF CONTENTS

# 1 INTRODUCTION

New York City is described as cultural, financial and media capital of the world. It is the most densely populated metropolitan cities in the United States with over 8 million people [1]. New York City is divided into 5 boroughs - Manhattan, Queens, Bronx, Brooklyn and Staten Island. Additionally, it has 59 community districts and hundreds of neighborhoods recognized by the Department of City Planning [2].

NY city attracts people from within and outside US borders. The city has some of the most expensive residential areas and finding a neighborhood becomes even more challenging with additional factors to consider like crime, commute time to work, and schools or university.

This raises the question - **Where to live in New York City?** This work attempts to answer this question pertinent to many new and existing NY city dwellers, where data science is employed to find neighborhoods best to live in the city. Preference and requirements differ by individual. Certain example criteria are used to arrive at the answer. However, the data and methodology can be fit in differing scenarios.

Maps and datasets from NYC Open Data [3] are used extensively in this work. APIs from Zillow [4], Google [5] and Foursquare [6] enable rich, realistic and recent data. Most basic machine learning algorithms are used to process the dataset. Statistical inferences and informative charts are used to analyze the data and obtain results. Jupyter notebook [7] using Python contains the code and framework for the work.

# 2 DATA

Maps and datasets of NY city neighborhood centroids, boundaries, and crimes reported to NYPD in 2018 from NYC Open Data platform are used in this work. Residential property prices are extracted using Zillow API, and commute time to places of interest like work or airport is estimated using Google's Distance API matrix. Lastly, neighborhoods are explored for restaurants, shopping complexes and universities using Foursquare API.

Data extraction is followed by an examination of values, removal of undesired features and values, and creating any calculations if necessary. In several cases, the data set thus obtained is summarized by value or merged with datasets. Tables, histograms, and maps are used to understand and validate the dataset. If deemed necessary in a later task in the Methodology section, the data query or preprocessing is adjusted.

In this section, each type of data used is introduced with processes involving extraction, processing, compilation, and summarization if needed.

## 2.1 CRIME DATA FROM NYC OPEN DATA

Complains reported to NYPD are available in NYC Open Data. Imported crime data set is inspected, cleaned for irrelevant features and erroneous values, and manipulated for meaningful information like comparing crimes in boroughs and neighborhoods on NY city.

### 2.1.1 Import and Compile Crime Dataset

Crime dataset is downloaded from NYC Open Data [8]. Features of data which are relevant for analysis - crime complaints id (*CMPLNT_NUM*), crime category, crime description, borough name and geographical coordinate (*latitude* & *longitude*) are retained. The crime data is further filtered for any missing or undefined values and saved as a csv file for data exploration and analysis to follow. The crime dataset, shown in table 1, contains about 450,000 entries which amounts to 140Mb of data storage space. Randomly sampled 10,000 entries are chosen from the dataset and visualized on map of New York city to ensure good geospatial distribution.

*Table 1: Filtered Crime Dataset obtained from NYC Open Data (first 5 rows are shown)*

| index | CMPLNT_NUM | LAW_CAT_CD | OFNS_DESC | BORO_NM | Latitude | Longitude |
|---|---|---|---|---|---|---|
| 0 | 118899213 | MISDEMEANOR | PETIT LARCENY | STATEN ISLAND | 40.53 | -74.2037 |
| 1 | 237656719 | VIOLATION | HARRASSMENT 2 | STATEN ISLAND | 40.5735 | -74.1068 |
| 2 | 192377396 | FELONY | BURGLARY | BROOKLYN | 40.5814 | -73.9645 |
| 3 | 547742823 | VIOLATION | HARRASSMENT 2 | MANHATTAN | 40.8046 | -73.9323 |
| 4 | 417880460 | MISDEMEANOR | SEX CRIMES | BROOKLYN | 40.6806 | -73.9743 |

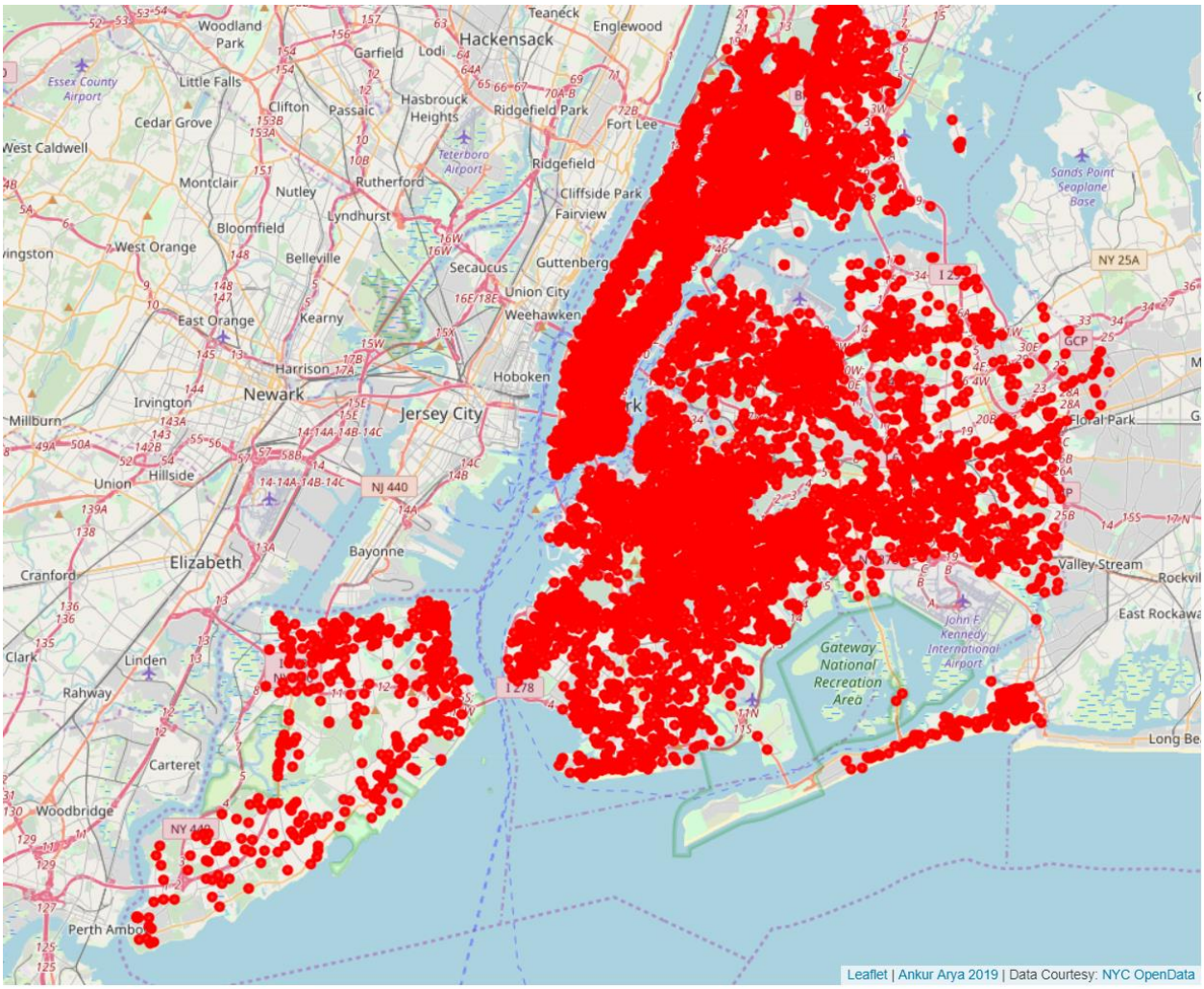



*Figure 1: Map of New York City showing randomly sampled crime sites (red dots) from NYC Open Data.*

### 2.1.2 Geospatial Distribution of Crime Sites

To ensure complete geographical coverage of crime sites over New York city, folium package is used to plot crime sites on open street map layer of the city. As seen in figure 1, there appears to be enough coverage of randomly sampled crime sites. These sites shall be classified to nearest neighborhood in later section, then used for analysis.

## 2.2 Residential Property Prices in NY City using Zillow API

NY city has regions which are among the most expensive places to live. Yet, over 8 million people reside here. NY city residential property prices are obtained from Zillow. The dataset is processed and analyzed to draw some interesting information about boroughs and neighborhoods of the city. Most expensive and cheapest neighborhoods are listed by boroughs and geospatial distribution is observed on map of NY city.

### 2.2.1 Extraction of Property Values from Zillow

Several functions are created to facilitate extraction and compilation of property sites obtained by HTTP requests to Zillow API [4]. For a successful query using Zillow API, at least street address and zip code are required. Two open sources were found useful.

- Open address data from openaddress.io [9]
- Open street address from NYC open data [10]

Data from [9] is used for property price extraction from Zillow; example of dataset is shown in table 2.

*Table 2: Open address data from openaddress.io containing addresses on New York city (top 5 entries shown)*

| | LON | LAT | NUMBER | STREET | UNIT | CITY | DISTRICT | REGION | POSTCODE | ID | HASH |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -73.979408 | 40.756086 | 560 | 5 AVE | NaN | NaN | NaN | NaN | 10036.0 | 1030290.0 | 0b9cfa59b5973cb8 |
| **1** | -73.994266 | 40.701937 | 25 | COLUMBIA HTS | NaN | NaN | NaN | NaN | 11201.0 | 3002257.0 | 6c8667827b8bb8cd |
| **2** | -73.918882 | 40.712779 | 1903 | FLUSHING AVE | NaN | NaN | NaN | NaN | 11385.0 | 89223.0 | efefc767cb0e0bc1 |
| **3** | -73.986376 | 40.778571 | 205 | W END AVE | NaN | NaN | NaN | NaN | 10023.0 | 1027498.0 | 36c6da0fae7cc8f4 |
| **4** | -73.912696 | 40.884183 | 3205 | ARLINGTON AVE | NaN | NaN | NaN | NaN | 10463.0 | 5178668.0 | bb534b03c465b374 |

Only valid and meaningful results from HTTP requests are saved as xml files, then property details are parsed from the xml files and merged into common dataframe. This method uses lot of memory space and takes several hours to execute due to huge number of queries. The compilation of results by this method can be viewed in table 3.

*Table 3: Residential property dataset obtained from Zillow with all features. (top 5 entries shown)*

| index | amount | bathrooms | finishedSqFt | latitude | link | longitude | lotSizeSqFt | state | ... | totalRooms | useCode |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2115985.0 | NaN | 2400.0 | 40.813168 | https://www.zillow.com/homedetails/12-Convent-... | -73.953323 | 940.0 | NY | ... | NaN | Duplex |
| 1 | NaN | NaN | 80742.0 | 40.820216 | https://www.zillow.com/homedetails/2351-12th-A... | -73.958509 | 6490.0 | NY | ... | NaN | SingleFamily |
| 2 | 155050.0 | NaN | NaN | 40.814280 | https://www.zillow.com/homedetails/36-Convent-... | -73.953401 | NaN | NY | ... | NaN | Cooperative |
| 3 | 136157.0 | NaN | NaN | 40.814280 | https://www.zillow.com/homedetails/36-Convent-... | -73.953401 | NaN | NY | ... | NaN | Cooperative |
| 4 | 360414.0 | 1.0 | 500.0 | 40.814022 | https://www.zillow.com/homedetails/33-Convent-... | -73.952795 | NaN | NY | ... | NaN | MultiFamily2To4 |

From property price dataset, the categories of properties in *useCode* column referring to irrelevant entries like '*vacant land*', '*unknown*' are removed. Useful features in the dataset are price (*amount*), geographical coordinates (*latitude* and *longitude*) and borough. The property details thus obtained are about ~ 150,000 in number amounting to ~ 36 Mb of size, which is saved as csv file and imported when needed for exploration and analysis.

### 2.2.2 Geospatial Distribution of Property Sites

Like crime data, the property data is plotted on NY city map to check the spread (geospatial distribution) over the city area. The coverage of sites, seen in figure 2, looks enough for analysis.

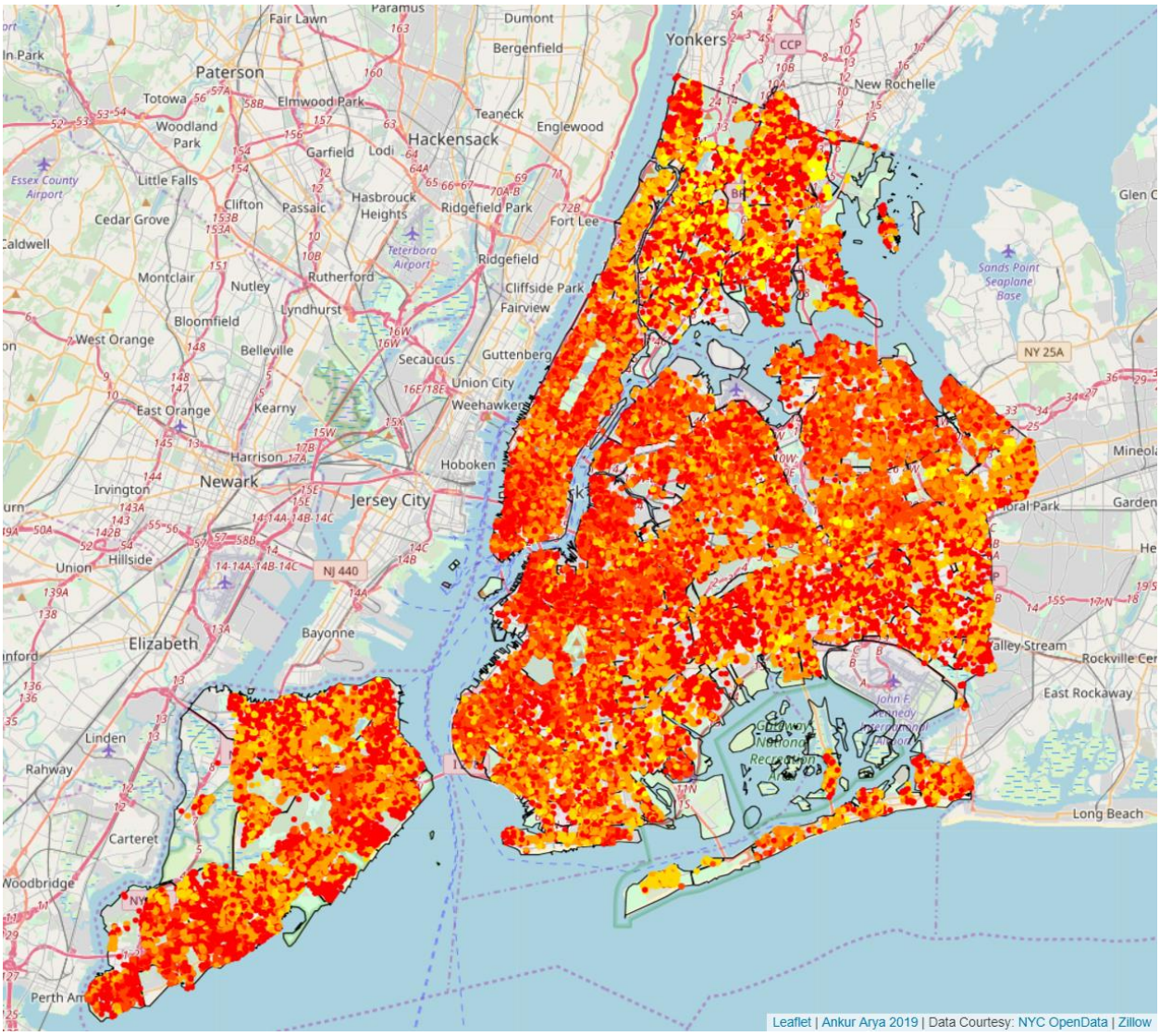



*Figure 2: Distribution of residential property sites highlighted by price in NY city. Sites are colored based on the price, yellow being lowest and red is highest. So, regions where red dots are dense resemble expensive properties. As an example, financial district (south Manhattan) and along perimeter of central park (north central Manhattan) have large clusters of red dots implying expensive properties.*

## 2.3 NEIGHBORHOODS OF NY CITY FROM NYC OPEN DATA

NY City has 5 boroughs and hundreds of neighborhoods. Neighborhoods will be used to identify regions to live in the city. Crime and property data collected need to be associated to neighborhoods, so the pin point location or centroid of neighborhoods is needed. The list of recognized neighborhood centroids recognized by identified by Department of City Planning is available in NYC Open Data as a map layer file [11].

### 2.3.1 Import and Compile Neighborhood Centroid List

The layer file is downloaded from the NYC Open Data and imported as csv and used to extract geographical coordinates of neighborhood centroids. Necessary features like names of neighborhood and borough and latitude & longitude are compiled in a dataset.

### 2.3.2 Visualize Neighborhood Centroids on Map

To ensure neighborhood centroid dataset correctness, the centroids are plotted on New York city map in figure 3.

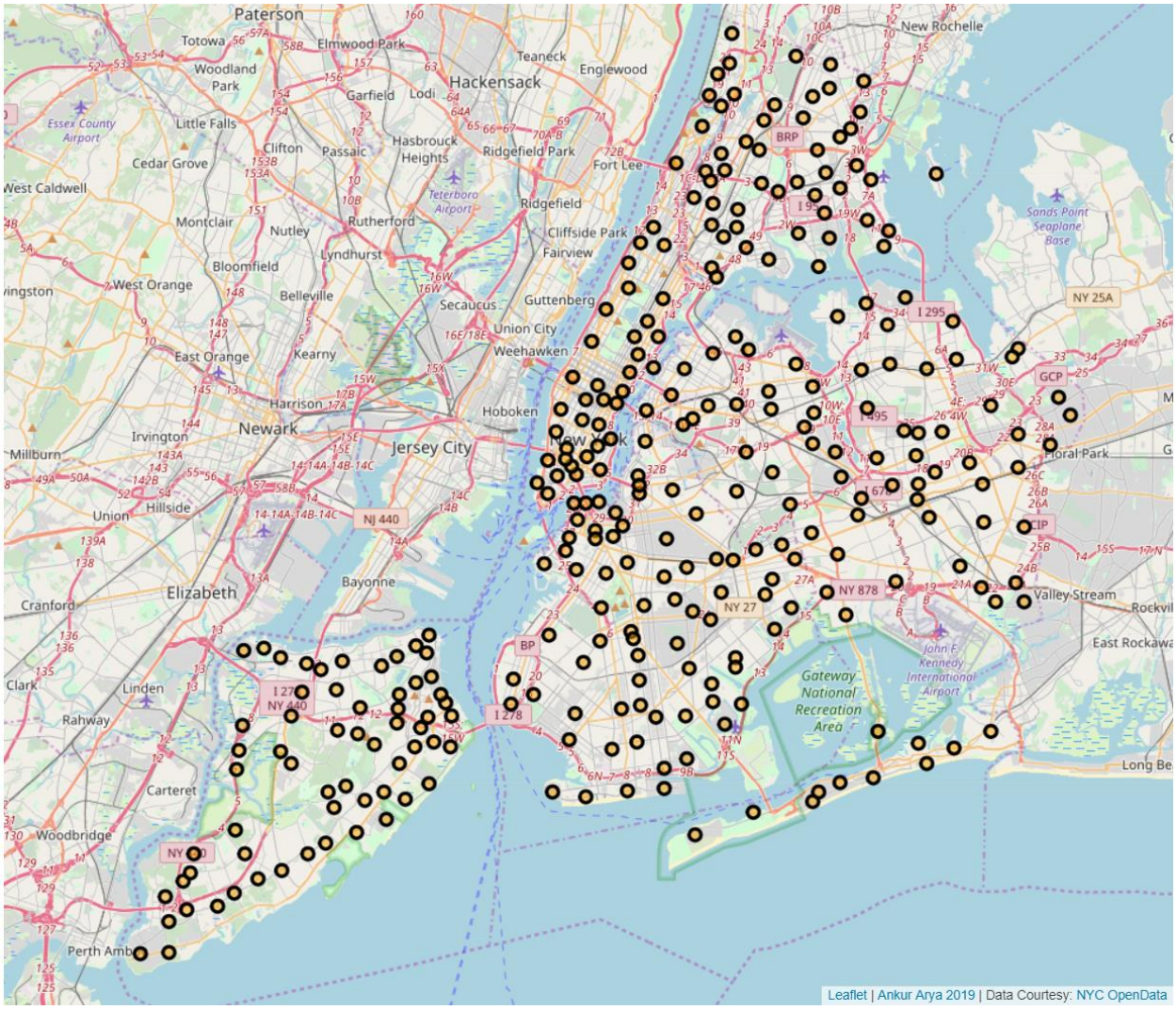



*Figure 3. Neighborhood centroids obtained from NYC Open Data seen on map of New York city*

## 2.4 Commute Time to Work & Travel from Google API

Midtown Manhattan is largest central business district in the world and central portion of borough of Manhattan [12]. John F. Kennedy International (JFK) Airport is the primary international airport serving New York city and busiest international air passenger gateway into North America [13].

For this work, Midtown Manhattan is assumed as a focal point for work and JFK airport is assumed as frequently used airport for travel. Then the choice of neighborhood should include time to commute to these places. Google Distance Matrix API [5] is used to get time to commute by car or by transit like subway or bus to Midtown Manhattan and JFK airport.

### 2.4.1 Obtain Drive & Transit Time using Google API

Commute time is obtained from all neighborhoods as origin, destination being Midtown Manhattan and JFK airport and both modes of transportation, car and transit like subway/bus. The geographic coordinates of neighborhood centroids available from dataset compiled in last section are used in query. As per Google's Distance Matrix API documentation [5], the commute time by driving assumes average historical traffic in default query. Few entries from the compiled result are in table 4.

*Table 4: Example of dataset containing commute time in minutes for each neighborhood.*

| | Name | Borough | Latitude | Longitude | Manhattan_drive_time | Manhattan_subway_time | JFK_airport_drive_time |
|---|---|---|---|---|---|---|---|
| 0 | Wakefield | BRONX | 40.894705 | -73.847201 | 40.100000 | 71.616667 | 36.416667 |
| 1 | Co-op City | BRONX | 40.874294 | -73.829939 | 40.800000 | 94.333333 | 35.566667 |
| 2 | Eastchester | BRONX | 40.887556 | -73.827806 | 36.950000 | 84.533333 | 31.733333 |
| 3 | Fieldston | BRONX | 40.895437 | -73.905643 | 31.516667 | 61.766667 | 39.450000 |
| 4 | Riverdale | BRONX | 40.890834 | -73.912585 | 30.083333 | 60.216667 | 40.316667 |

## 2.5 Venues in Neighborhoods using FourSquare API

Neighborhoods are also explored for venues of most common interests like restaurants, shopping complex or malls, parks and college or university. This is performed by HTTP requests to Foursquare API [6] which returns json file containing venues belonging to queried category. Venues details are parsed from json results. Since the neighborhood centroids are only couple of miles from each other, radius of 1 mile is used to search venues. It should be kept in mind that number of venues per neighborhood is be capped to 50, due to limit set in query.

### 2.5.1 Get Restaurants and Cuisines Venues

Search query 'restaurant' is used to get venues tagged as restaurants. Some of the results thus obtained have venue categories like Food, Restaurant, etc., which do not represent cuisine or meaningful category. Only results with meaningful venue categories are retained for each neighborhood, as seen in table 5.

*Table 5: Example of compiled results from venue search query 'Restaurant'*

| | Borough | Neighborhood | Venue_Category | Venue_Latitude | Venue_Longitude | Venue_Name |
|---|---|---|---|---|---|---|
| 0 | BRONX | Wakefield | Caribbean | 40.899767 | -73.857135 | Big Daddy's Caribbean Taste Restaurant |
| 1 | BRONX | Wakefield | Chinese | 40.904359 | -73.849795 | Red Flower Chinese Restaurant |
| 2 | BRONX | Wakefield | Caribbean | 40.899768 | -73.857184 | Kaieteur Restaurant & Bakery |

The food character of neighborhood from restaurants is defined by taking top five highest counts of restaurant categories or cuisine. This helps with identification of dominant restaurant cuisine or category in neighborhood. As an example, table 6 shows Wakefield in Bronx is most Caribbean restaurants.

*Table 6: Example of summary of top 5 restaurant cuisine or category*

| | Borough | Neighborhood | Venue_Category | Total_Count |
|---|---|---|---|---|
| 0 | BRONX | Wakefield | Caribbean | 12 |
| 1 | BRONX | Wakefield | Chinese | 5 |
| 2 | BRONX | Wakefield | American | 2 |
| 3 | BRONX | Wakefield | Asian | 2 |
| 4 | BRONX | Wakefield | Latin American | 2 |

### 2.5.2 Get Shopping Centers and Mall Venues

Search query 'shopping mall' is used to get venues tagged as shopping or mall. Some of the results thus obtained have venue categories like Office, Electronics, Plaza, etc., which do not strictly represent a big shopping complex or mall. Only results with meaningful venue categories as 'mall' are used for summarizing total counts per neighborhood. See Table 7 for an example.

*Table 7: Summary example of retained results from query 'Shopping Mall' and venue category as 'Mall'*

| | Borough | Neighborhood | Venue_Category | Total_Count |
|---|---|---|---|---|
| 0 | BRONX | Co-op City | Mall | 4 |
| 1 | BRONX | Eastchester | Mall | 2 |
| 2 | BRONX | Fieldston | Mall | 1 |
| 3 | BRONX | Riverdale | Mall | 1 |
| 4 | MANHATTAN | Marble Hill | Mall | 1 |
| 5 | BRONX | Baychester | Mall | 3 |

### 2.5.3 Get Parks and State Parks Venues

Use search query 'parks' to get parks around a neighborhood, and relevant results where venue category is 'Park' or 'State Park' are retained. Table 8 shows example of summary, with total counts of parks or state parks per neighborhood.

*Table 8: Summary example of retained results from query 'Park' and venue category as 'Park' or 'State Park'*

| | Borough | Neighborhood | Venue_Category | Total_Count |
|---|---|---|---|---|
| 0 | BRONX | Wakefield | Park | 3 |
| 1 | BRONX | Co-op City | Park | 5 |
| 2 | BRONX | Eastchester | Park | 4 |
| 3 | BRONX | Fieldston | Park | 14 |
| 4 | BRONX | Riverdale | Park | 16 |
| 5 | BRONX | Kingsbridge | Park | 17 |

### 2.5.4 Get Colleges and Universities Venues

Colleges and universities are searched using query ‘colleges’, and relevant results have venue category as 'University' or 'Community College' or ‘General College and Education’ and ‘College and Education’. Table 9 shows example of summary, with total counts of colleges per neighborhood.

*Table 9: Summary example of retained results from query ‘colleges’ and relevant venue categories*

| | Borough | Neighborhood | Venue_Category | Total_Count |
|---|---|---|---|---|
| 0 | Arlington | STATEN ISLAND | College/University | 1 |
| 1 | Arrochar | STATEN ISLAND | College/University | 1 |
| 2 | Arverne | QUEENS | College/University | 1 |
| 3 | Astoria | QUEENS | College/University | 2 |
| 4 | Astoria Heights | QUEENS | College/University | 1 |

Searched venues results for parks, malls and colleges are concatenated into single dataset for later analysis and comparison of neighborhoods.

# 3 METHODOLOGY

After the extraction and compilation of desired datasets, each type of dataset is further explored to ascertain sufficiency and processed for analysis and results. Crime and residential property sites need to be associated to neighborhoods. Following two approaches were considered.

1. Neighborhood boundary layer [14] available at NYC Open Data. Upon examination, it is found that some neighborhood centroids like on the boundaries in the layer. So, the neighbor centroids and neighborhood tabulation boundary layer cannot be used together.
2. Define a new boundary layer which partitions the sites such that, each site lies within the boundary of nearest neighborhood centroid. This intuitive classification of crime and residential property sites to neighborhood centroid is done using nearest neighbor algorithm of machine learning.

Venues like restaurants and colleges are already classified because they are searched using radius around the neighborhood. Likewise, data of commute times which originated from list of neighborhood centroids do not need classification. Data statistics like distribution and centering is used to compare and rank boroughs and neighborhoods of NY city. Heatmaps are created to visualize geospatial distribution of crime and property prices of neighborhoods in the city.

Criteria on price, crime rate, commute time is set to shortlist neighborhoods while studying any underlying trade-offs. Down selected neighborhoods are further explored using venues. This methodology of selection is highly dependent on values of criteria. Same methodology can be easily employed in other scenarios or with new additional datasets.

## 3.1 CRIME DATA ANALYSIS

Original crime data obtained from NYC Open data platform has about 450,000 entries, a random subset of 10,000 entries with enough spread over NY city is used. At first, the sites are classified to

neighborhood centroid using nearest neighbor search. The statistics of crime data is studied and used for comparing boroughs and neighborhoods of the city. Heatmap is generated to check out the crime distribution in city map.

### 3.1.1 Classification of Data to Nearest Neighborhood

Using nearest neighbor search, each crime site is grouped to nearest neighborhood centroid. This classification is done by evaluation of euclidian distance between a site to all neighborhood centroids, and the site is classified to neighborhood with smallest distance. For consistency, it is enforced that crime sites and neighbor centroids are in the same borough, this is critical for sites on the borders of boroughs. Use of nearest neighbor algorithm then predefines the boundaries around centroids partitioning the space to nearest neighbor. The boundary layer is derived from the neighbor centroid list [11] and borough boundary layer [14] from NYC Open Data. Voronoi polygon and dissolve methods were implemented using QGIS [15] to form the layer. Results are shown in figure 4.

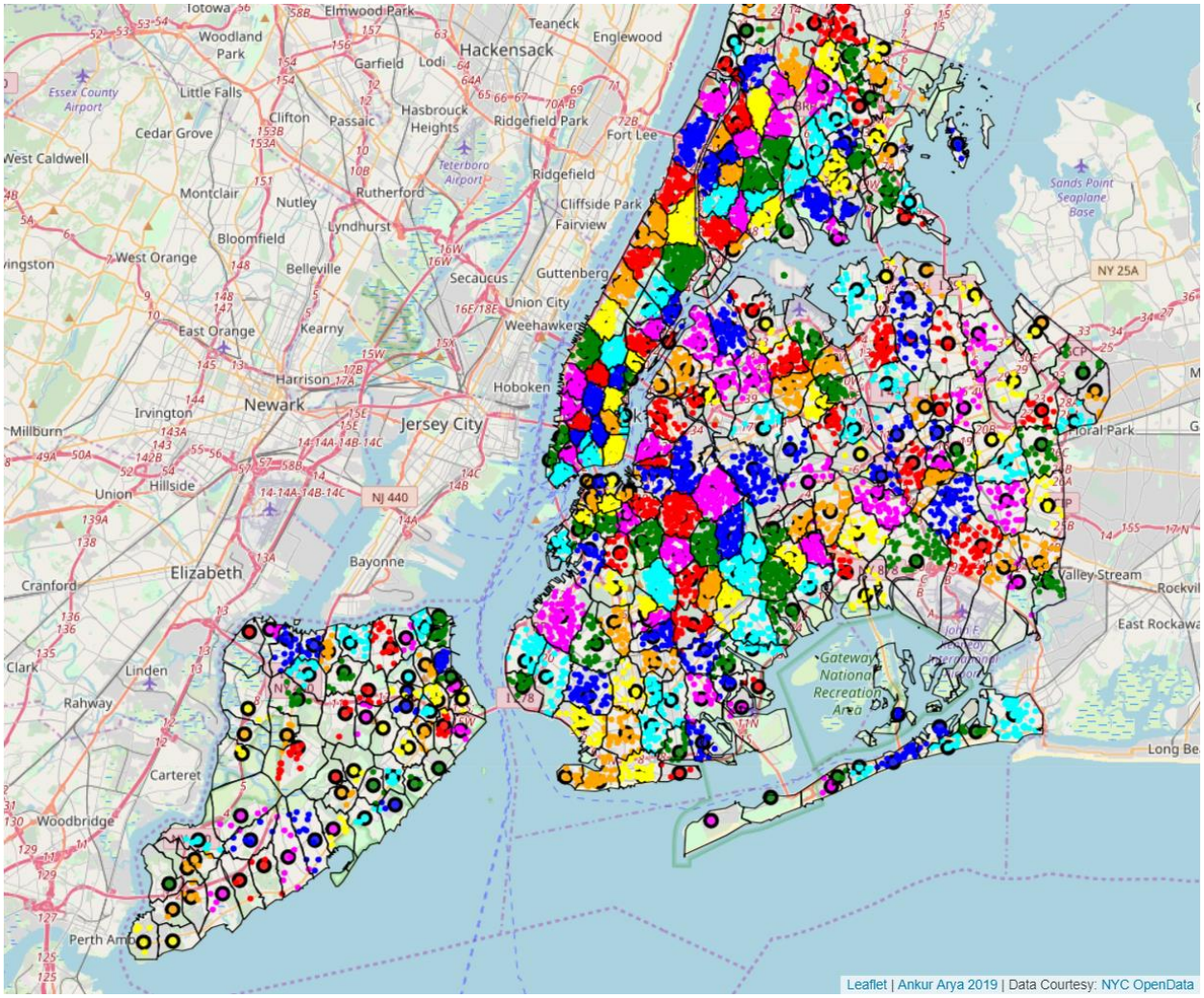



*Figure 4: NY city map shows crime sites classified to nearest neighborhood in same color. Neighborhood centroids are larger circles with black boundary. Derived boundary layer representing equi-distant regions are added to show boundaries of neighborhood.*

### 3.1.2 Crime Statistics

Crime data is analyzed for useful statistical information on city's boroughs and neighborhoods. The results from analysis are useful to understand the data and help define criteria to shortlist the neighborhoods.

Comparison of boroughs by total number of crimes in figure 5, shows Brooklyn has highest (~29%) reported crimes in NY city, closely followed by Manhattan (~24%). Staten Island (~ 4%) is least infested with crime. Using this data, the boroughs cannot be regarded unsafe to live, because there may be safe neighborhoods within a borough.

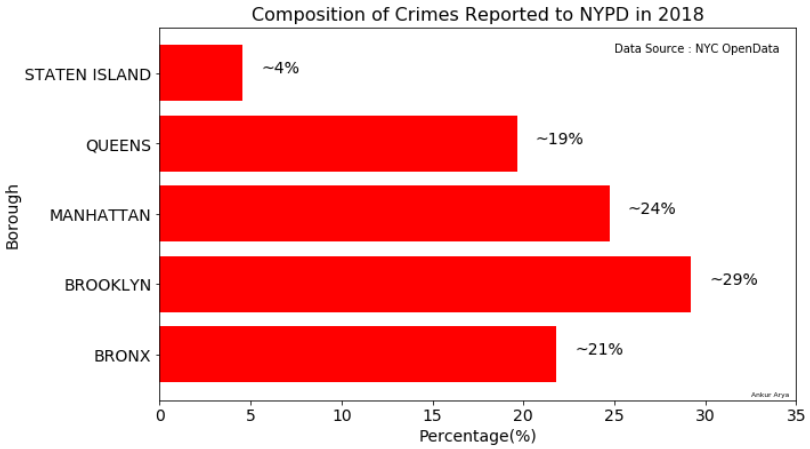



*Figure 5: Composition of crimes reported to NYPD in 2018 by boroughs*

Distribution of crime data summarized as total count by neighborhood is presented in figure 6. Total crimes reported in 2018 in each neighborhood range from zero to 250 and the median is 22 crimes per neighborhood. There are several neighborhoods, roughly 100 with total crimes in 2018 less than 20, this can be useful criteria to shortlist safe neighborhoods.

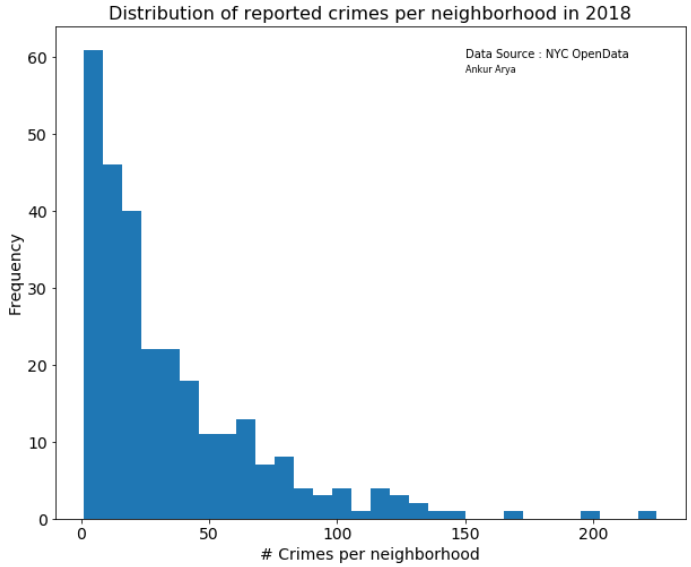



*Figure 6: Distribution of total number of crimes per neighborhood in 2018*

Type of crimes can give additional insight if neighborhood is unsafe for living. In figure 7, top 10 reported crimes are presented with its constituent boroughs. Among the top most crimes reported to NYPD in 2018, grand and petit larceny are biggest form of crimes in Manhattan and Bronx's top most crimes are felony assault, assault related offenses and drug related. Other boroughs appear to have almost all types of crimes. For selection of neighborhood, the types of crimes are not differentiated.

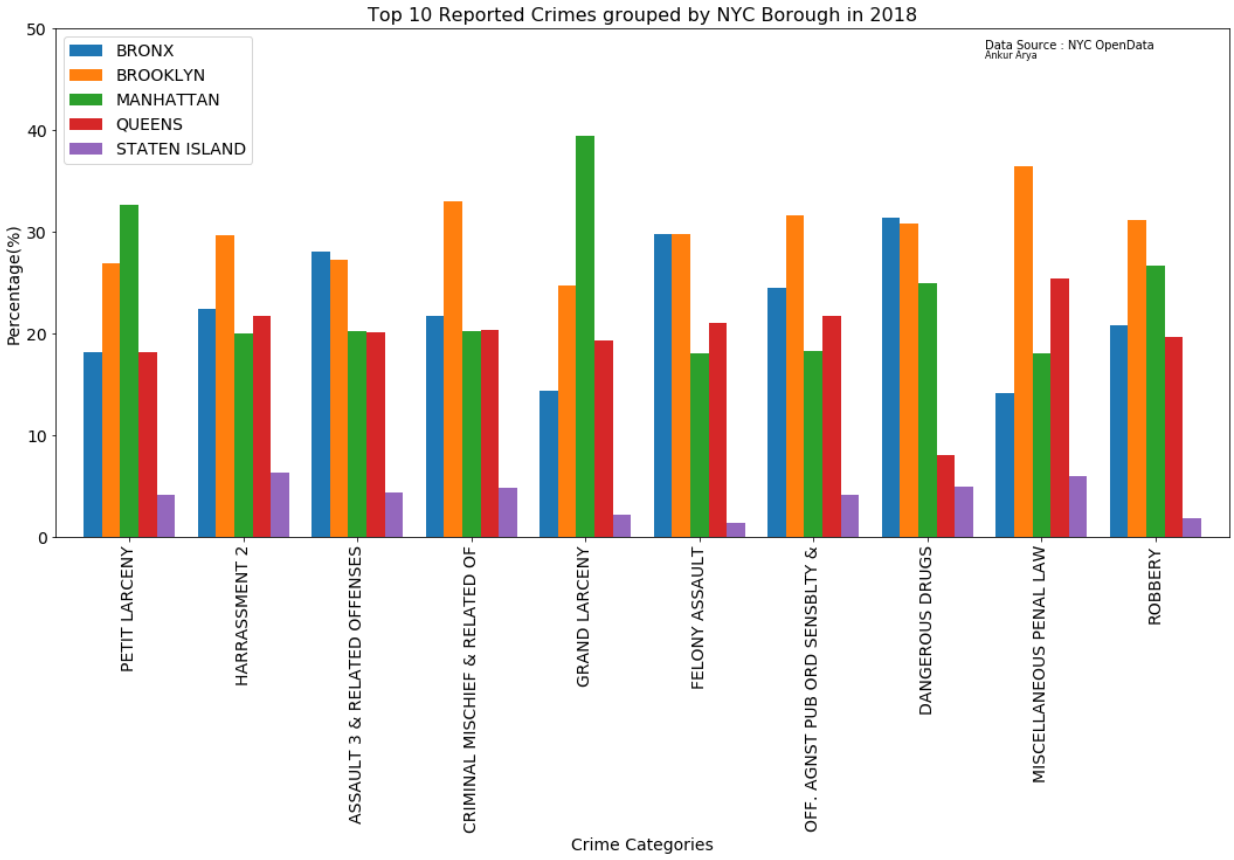



*Figure 7: Top 10 crimes in 2018 grouped by NY city boroughs*

### 3.1.3 Heatmap of Crime by Neighborhood

Since a borough can have both safe and most crime prone neighborhoods, it is helpful to understand the geospatial distribution of crime infested neighborhoods. This can help define additional criteria for searching safer neighborhoods to live. An example of selecting safe neighborhood could be such that the desired neighborhood should be at least two neighborhoods away from worst neighborhood. Although this adds richness to the selection criteria, this complicated criterion will not be used.

A heatmap of crime by neighborhood can help visualize if any clustering is observed in safe or unsafe neighborhoods. Folium choropleth is used to create heatmaps of total crime reported in 2018 by neighborhood in figure 8. Neighborhoods with high number of reported crimes are observed in north eastern Brooklyn, midtown south Manhattan, northern Manhattan and southern Bronx. Low crime neighborhoods could be from less populated regions like north western Staten Island.

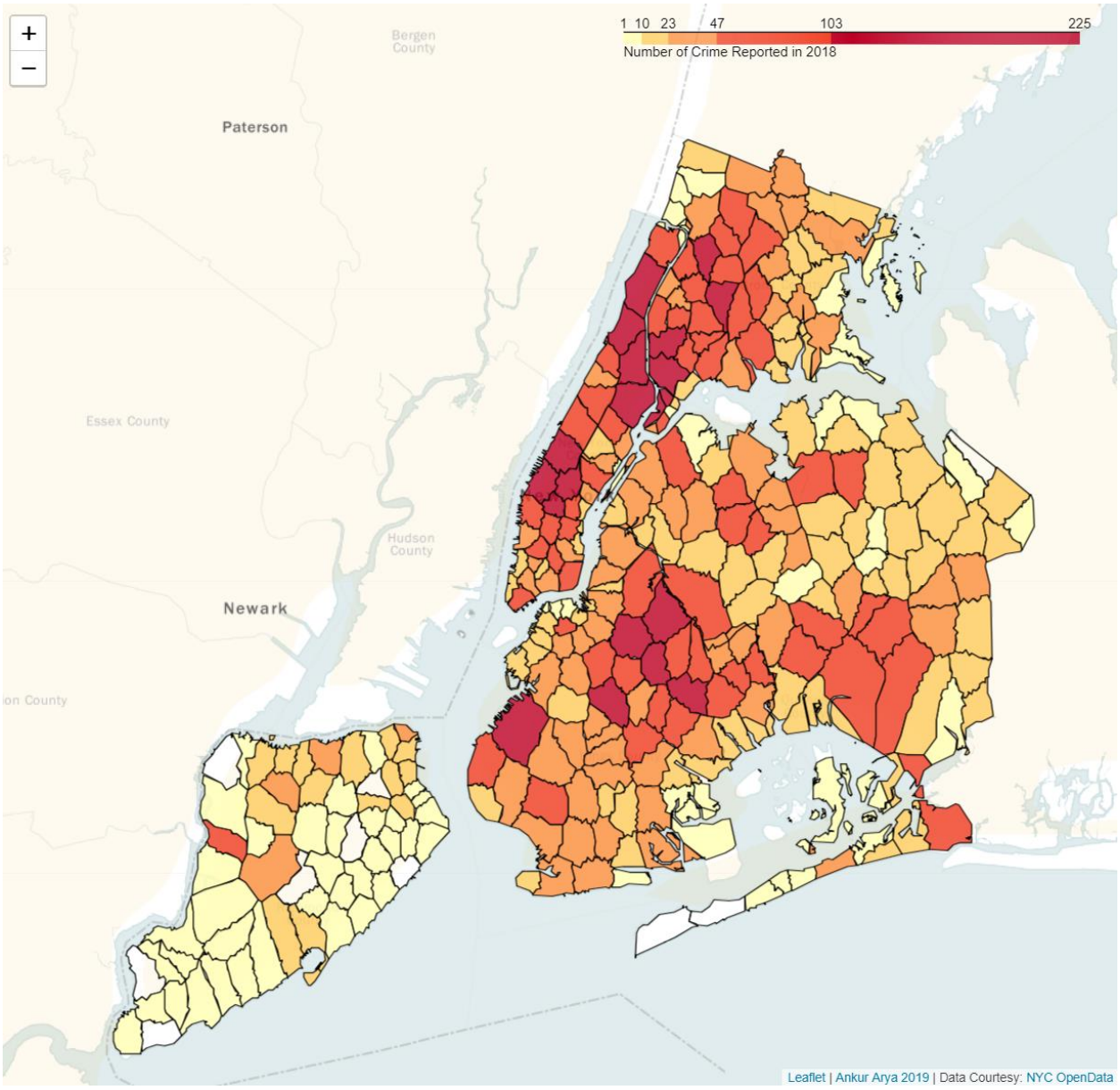



*Figure 8: Heatmap of total crimes reported in 2018 by neighborhoods on NY city.*

## 3.2 Residential Property Prices Data Analysis

About 150,000 residential property details in NY city have been extracted using Zillow API. The sites are classified to neighborhood centroid using nearest neighbor search. The statistics of crime data is studied and used for comparing boroughs and neighborhoods of the city. Heatmap is generated to visualize variation of price on city map.

### 3.2.1 Classification of Data to Nearest Neighborhood

Like crime data, each property site is grouped to nearest neighborhood centroid using nearest neighbor search. This classification is done by evaluation of euclidian distance between a site to all neighborhood

centroids, and the site is then classified to neighborhood with nearest distance. This classification is shown on map in figure 9.

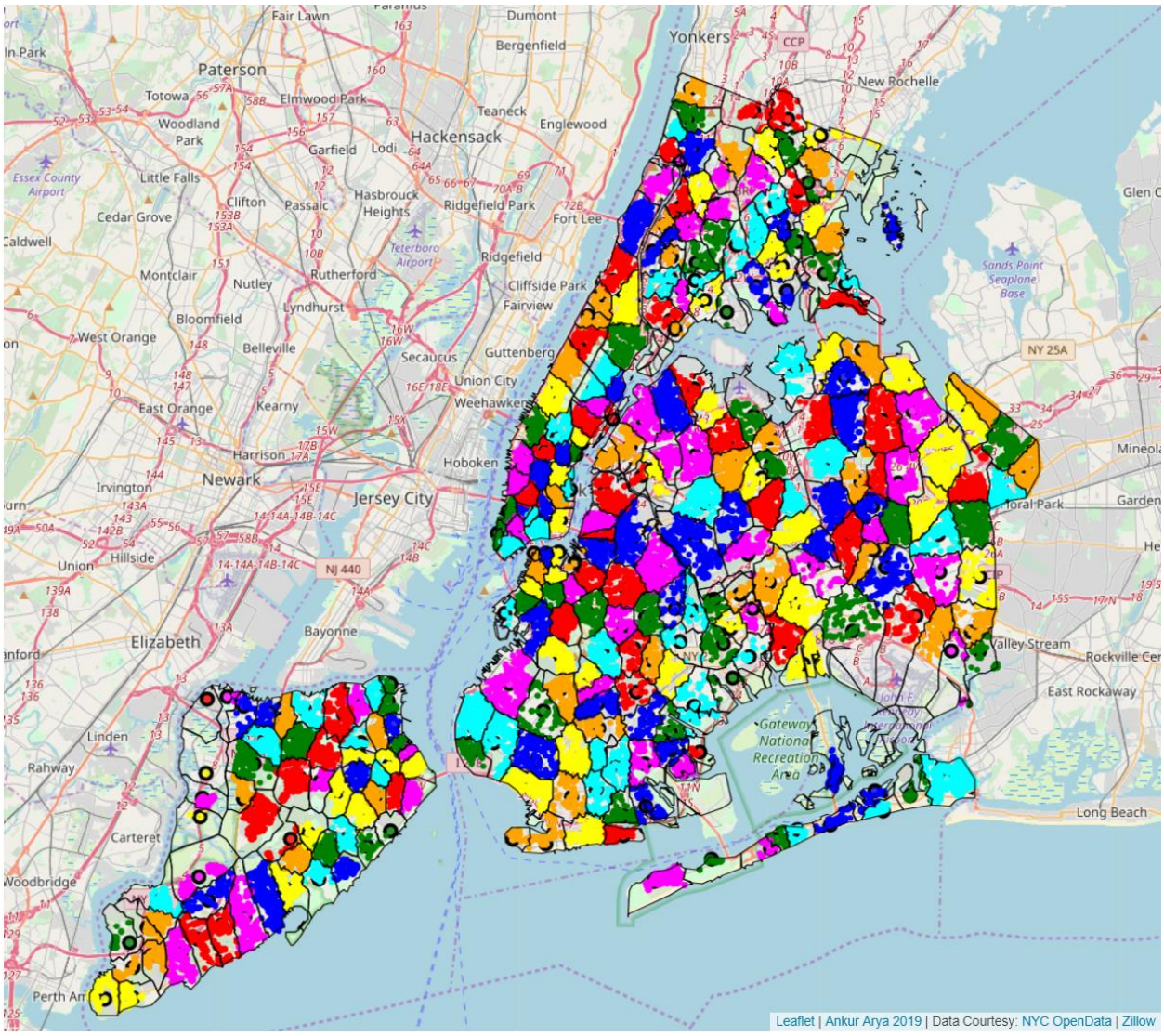



*Figure 9: NY city map shows residential property sites classified to nearest neighborhood highlighted in same color. Neighborhood centroids are larger circles with black boundary. Derived boundary layer representing equi-distant regions are added to show boundaries of neighborhood.*

### 3.2.2 Prices Statistics

Residential property price data imported from Zillow is analyzed for useful statistical information on city's boroughs and neighborhoods. The results from analysis are useful to understand the data and help define criteria to shortlist the neighborhoods.

It is noteworthy that the data includes all kinds of residential property like apartments, duplex and houses. For selection of neighborhood, no distinction is made between the type of residential property. Figure 10 shows the histogram of prices of all residential property sites in NY city. The distribution of

prices is almost log normal and spread over 2 orders of magnitude from $100,000 up to $35 million, with median price close to $800,000.

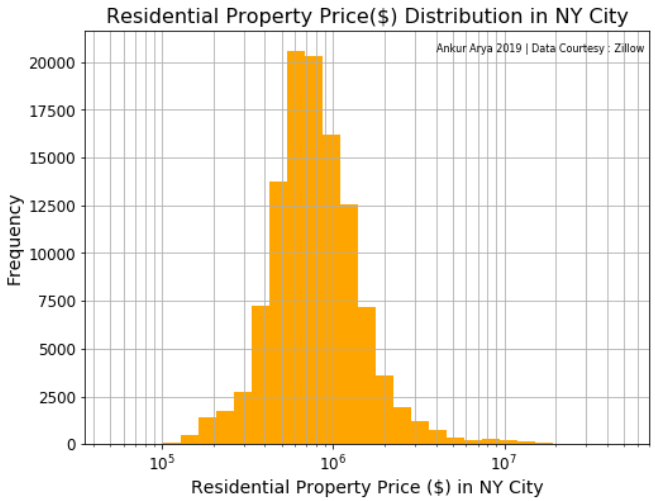



*Figure 10: Residential property price distribution in NY city*

The residential property price dataset can be used to compare boroughs and neighborhoods. Price medians, distribution and tails are characteristic of the region's development. Figure 11 shows price distribution for each borough of NY city. The distribution of prices of residential property in NY city is log normal for most boroughs. Manhattan's price has a tail extending towards expensive side. Prices in Bronx show bimodal behavior, it is shown in later section that lower price mode has higher crime. Median price of Brooklyn and Manhattan are highest and close to $1 million dollars and is twice that of median price of Bronx.

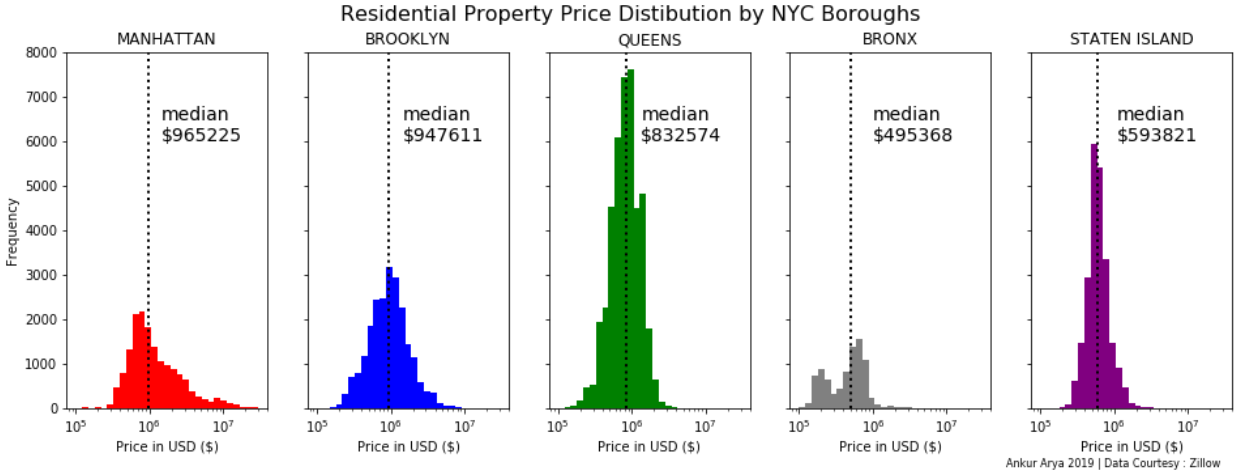



*Figure 11: Residential property price distribution by boroughs of NY city*

Using the property price dataset, most expensive neighborhoods can be recognized in each borough. This helps choose price range criteria for shortlisting neighborhoods. Median of property price per neighborhood is chosen to compare and rank neighborhoods.

Figure 12 shows top 5 most expensive neighborhoods by boroughs of NY city. Most expensive neighborhoods in Manhattan are Tribeca and Soho, with median price over whopping $35 million dollars. Fulton Ferry in Brooklyn is close to Brooklyn bridge connecting southern Manhattan with Brooklyn, properties here also price around $35 million dollars. Most expensive neighborhoods in Queens, Bronx and Staten Island are between $10-15 million dollars.

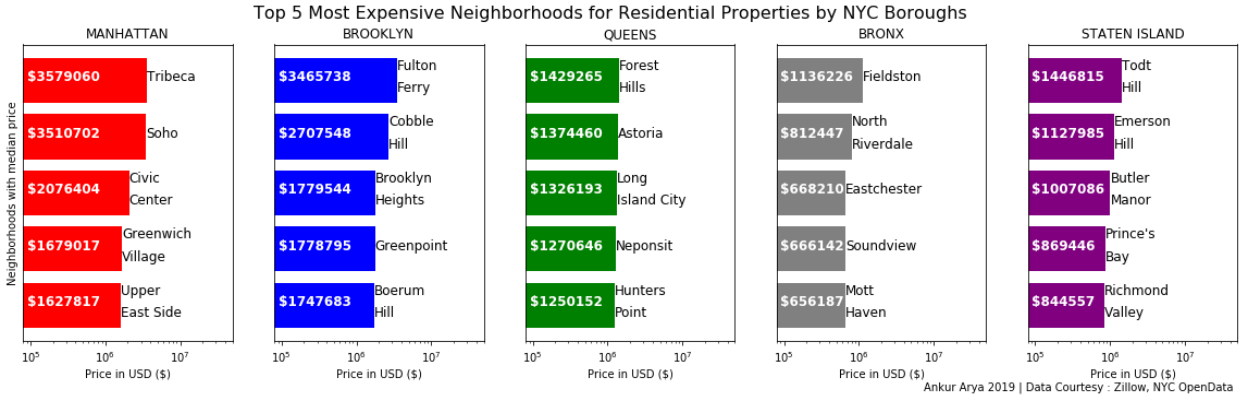



*Figure 12: Top 5 most expensive neighborhoods for residential properties by boroughs of NY city.*

Likewise, as seen in figure 13, cheapest neighborhood in Manhattan is *Inwood*, in Brooklyn is *Midwood*, and both have median price near $450,000. In other boroughs, cheapest prices are close to $300,000 in *Roxbury* in Queens and *Port Ivory* in Staten Island. Lowest of all in NY city is at *Co-op City* in Bronx with $182,250. As mentioned before, the dataset is ensemble of different kinds of residents like apartments, duplex, houses and for simplicity they are not differentiated.


*Figure 13: Top 5 most expensive neighborhoods for residential properties by NYC boroughs.*

### 3.2.3 Heatmap OF Residential property Price by Neighborhood

Like crime per neighborhood, it is helpful to understand geospatial distribution of expensive and cheap neighborhoods. This can help define additional criteria for searching for neighborhoods to live. For

example, a cheaper neighborhood can be selected adjacent to expensive neighborhood in anticipation of increase in future property price valuation. This adds prospect of good investment to the selection criteria, but this type of criterion will be avoided in this work.

A heatmap of residential property prices of NY city is shown in figure 14. From the map, it's evident that neighborhoods in south Manhattan and in proximity the neighborhoods of Brooklyn and Queens are most expensive. One of the likely reasons is lesser commute time to south Manhattan which is the business hub. In next section relation between commute time and the prices will be investigated.

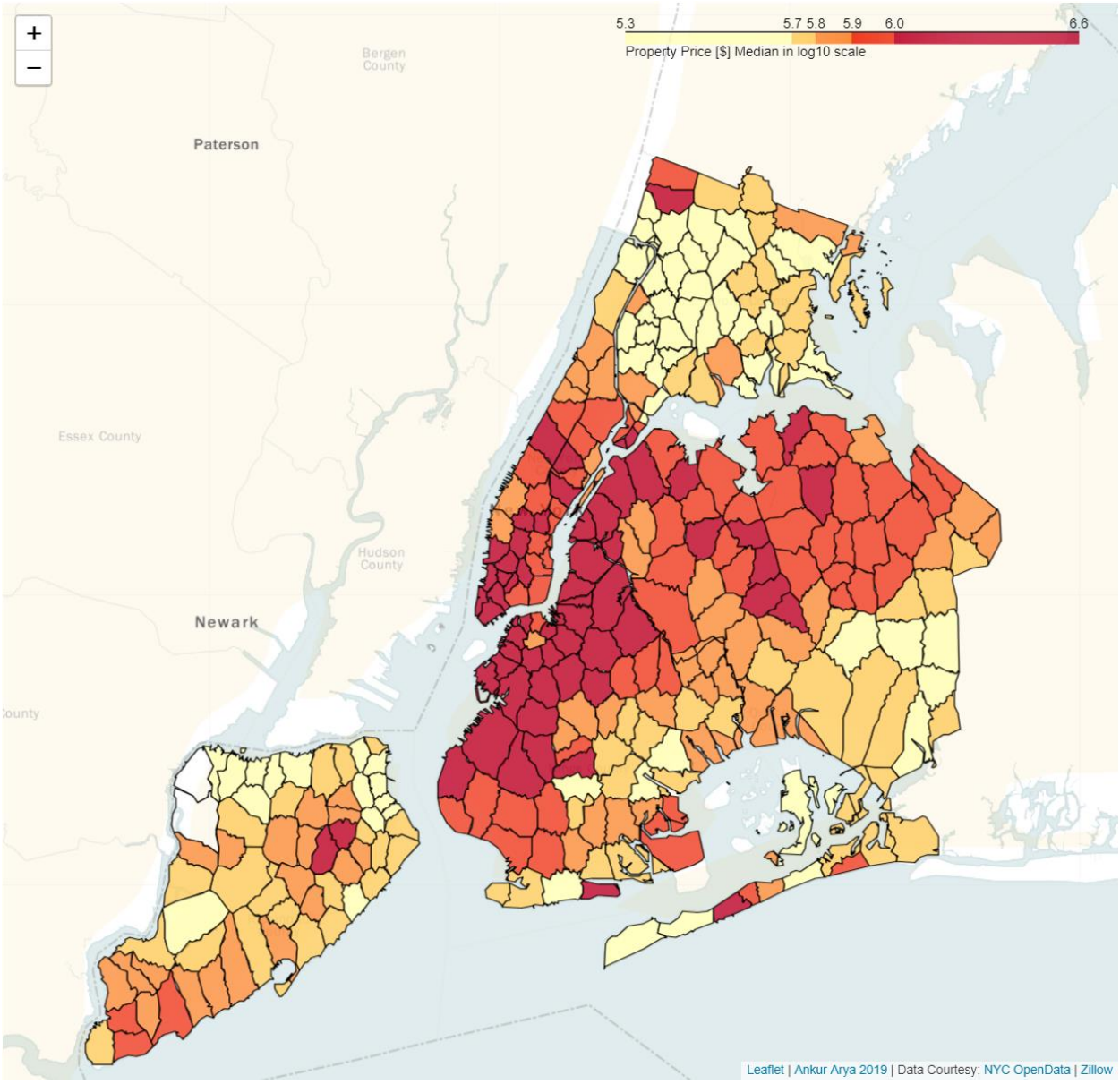



*Figure 14: Heatmap of residential property prices based on median by neighborhoods on NY city.*

## 3.3 Residential Property Prices Trade-Offs

Assume that neighborhoods in NY City with higher residential property prices would have lower crimes or must be closer to business hub in Manhattan. Such a relationship between property price and crime or commute time then creates trade-off situation. In a strong trade-off situation, a lower priced neighborhood is then expected to have higher crime or longer commute time and its highly unlikely to have lower prices and lower crime or shorter commute time all in a single neighborhood. In this section, possibility of trade-off and criteria of upper limit of prices, crime and commute time to shortlist neighborhoods are examined. For this purpose, the datasets of total crime per neighborhood, property price median by neighborhood and commute time from neighborhood centroids are merged for analysis.

### 3.3.1 Tradeoff between Property Prices and Crime

Selection of neighborhoods based on property price and crime reported in 2018 requires to check relation between the two variables, so that appropriate filter on values can be set to shortlist neighborhoods. In figure 15, since not much of correlation is evident, there is no apparent tradeoff between property prices and crime. However, in Bronx and Staten Island, few sites of lowest priced neighborhoods have relatively higher number of crimes reported.

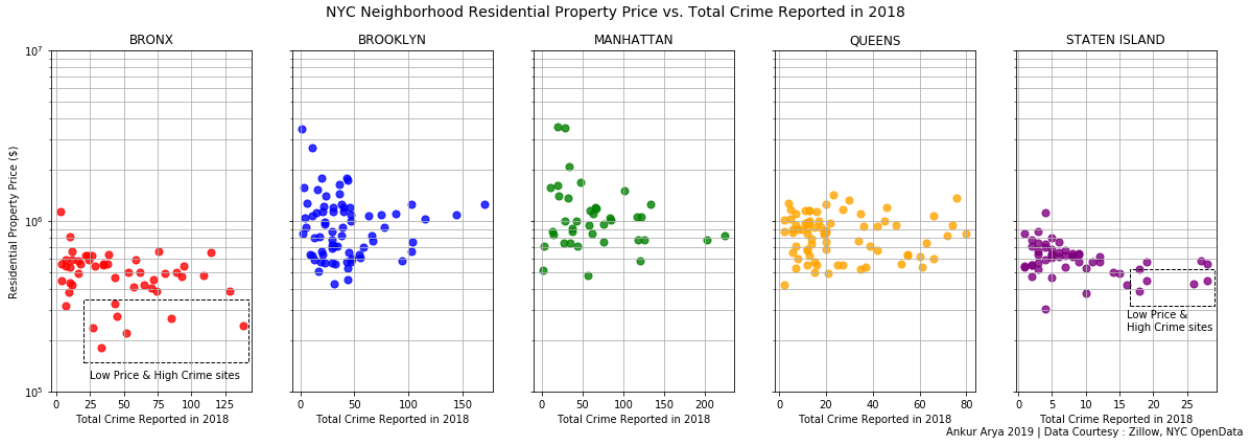



*Figure 15: NYC neighborhood property price median vs. Total crime reported to NYPD in 2018 by boroughs.*

The scatter plots in figure 15 are useful to commit price and crime limits dependent on boroughs to shortlist neighborhoods. Instead, single price and crime limits will be used for shortlisting neighborhoods for all boroughs.

Considering criteria of property price lower than $1.25 million dollars and total crimes less than 20, as seen in figure 16, there are plenty of neighborhoods in NY city to choose from. Almost all neighborhoods in Staten Island qualify this criterion and can be seen within the dotted box. Other constraints like commute time is considered in next section.

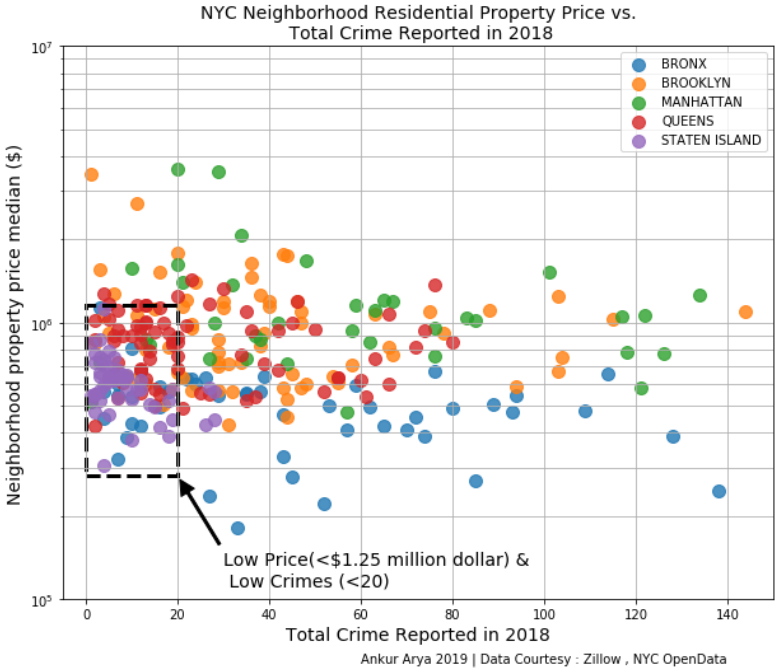



*Figure 16: NYC neighborhood property price median vs. Total crime reported to NYPD in 2018.*

### 3.3.2 Tradeoff between Property Prices and Commute Time

Relation between residential property price and commute time to midtown Manhattan and JFK airport in examined. So that appropriate filter on values can be set to shortlist best neighborhood to live in NY city.

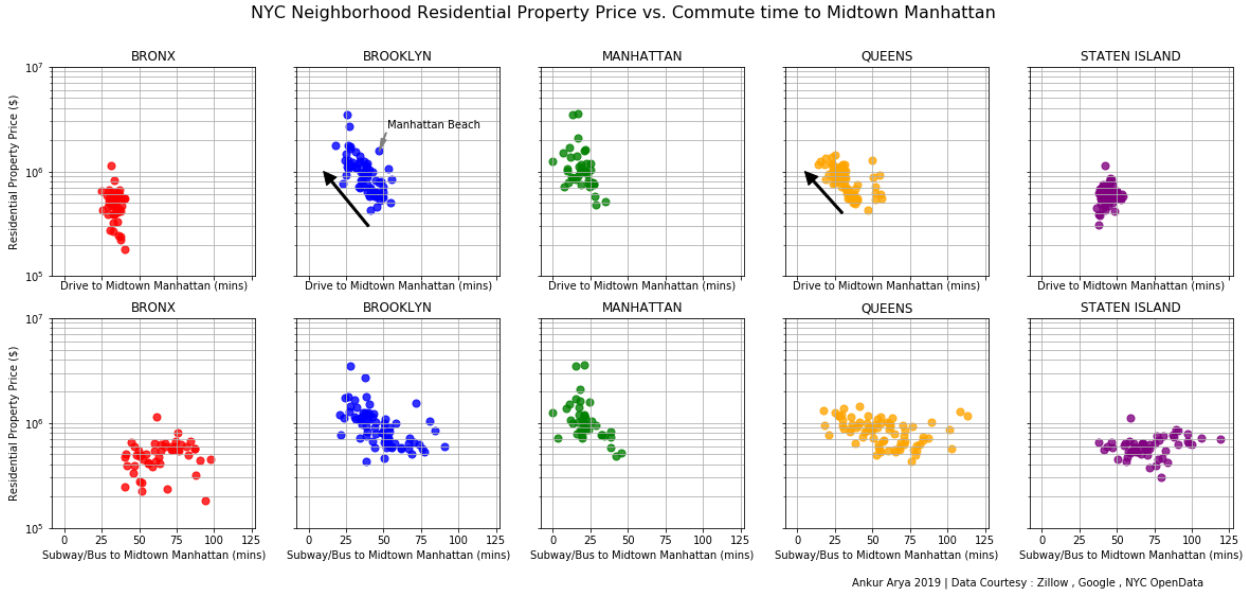



*Figure 17: NYC neighborhood residential property price median vs. Commute time to Midtown Manhattan*

Plots of residential property price median per neighborhood against commute time to midtown Manhattan shown in figure 17, imply positive correlation in Queens and Brooklyn. So, there is trade-off between property prices and commute time in Queens and Brooklyn. Lower priced residential properties are farther from business hub of Manhattan.

Some neighborhoods do not fall in the trend, which could be because the property price is influenced by factors other than proximity to business hub in Manhattan. For example, in Brooklyn*, Manhattan Beach* is 45 mins by driving and 65 mins by transit (bus/subway) from Manhattan, yet median property price is close to $1.5 million dollars. Building a price model for neighborhood is interesting topic but the focus is to determine neighborhoods to live in NY city.

Repeating criteria of property price lower than $1.25 million dollars and now considering a short commute time of less than 30 mins by driving, there are several neighborhoods in NY city to choose from as observed in figure 18. There is no trade-off observed in the neighborhoods lying inside the dotted box depicting the criteria. However, almost all neighborhoods in Staten Island does not qualify this criterion and lies outside the dotted box. Recollect, Staten Island neighborhoods did qualify low price and low crime criteria seen earlier. The neighborhoods which qualify all three criteria of low price, low crime and low commute time will be fewer. Shortlisting of neighborhoods based on these criteria is done step by step in next section.

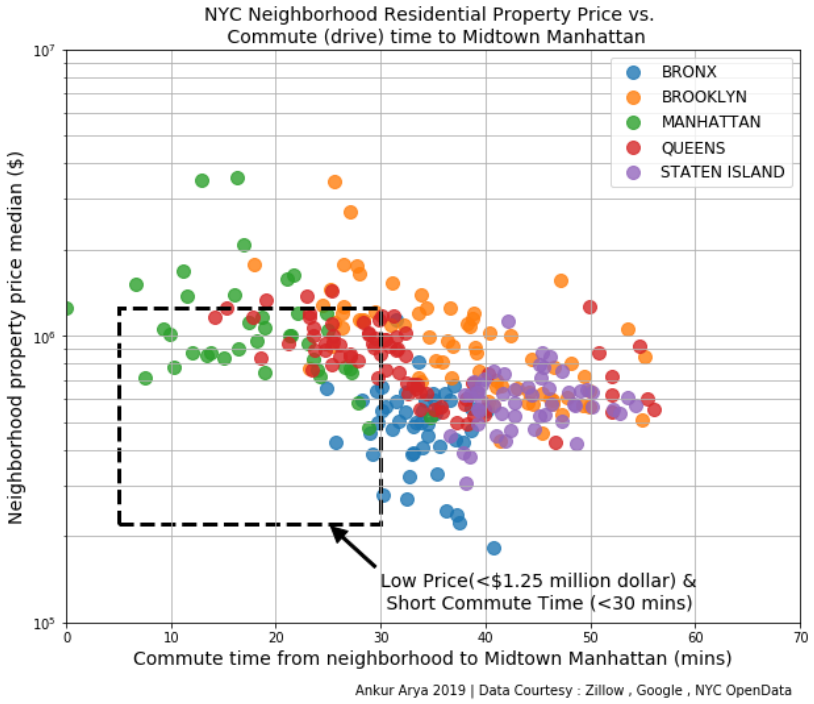



*Figure 18: NYC neighborhood residential property price median vs. Commute time to Midtown Manhattan*

## 3.4 SHORTLISTING OF NEIGHBORHOODS TO LIVE IN NY CITY

Criteria based on upper limits on property price, total crime and commute time will be used to short list neighborhoods. Choice of neighborhood to live in city can be using factors not reflected in the dataset, like ratings of school, proximity to beach, etc. To limit the scope of the work, property price, crime and commute time are used as primary factors for shortlisting neighborhoods. In later section, shortlisted neighborhoods are further down selected based on variety and number of venues of leisure, education and recreation.

Following preferential criteria in order are assumed:

- Price Upper Limit of $1.25 million dollars.
- Driving time to Manhattan and JFK airport within 30 mins.
- Total crime reported in 2018 less than 10

### 3.4.1 Downselect based on Property Price

At first, the datasets include total crime per neighborhood and residential property price median of each neighborhood and commute times by neighborhoods of NY city are merged. This dataset has about 287 entries, first few rows of the dataframe are shown in table 10.

*Table 10: Merged dataset used for shortlisting. First 5 rows are shown.*

| | Neighborhood | Borough | Latitude | Longitude | Residential Property Price median ($) | Drive to Midtown Manhattan (mins) | Subway to Midtown Manhattan (mins) | Drive to JFK Airport (mins) | Total Crime Reported 2018 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Baychester | BRONX | 40.866858 | -73.835798 | 537981 | 35 | 76 | 30 | 10 |
| 1 | Bedford Park | BRONX | 40.870185 | -73.885512 | 410292 | 35 | 63 | 33 | 57 |
| 2 | Belmont | BRONX | 40.857277 | -73.888452 | 493852 | 34 | 51 | 32 | 80 |
| 3 | Castle Hill | BRONX | 40.819014 | -73.848027 | 496198 | 33 | 82 | 29 | 16 |
| 4 | City Island | BRONX | 40.847247 | -73.786488 | 561559 | 38 | 75 | 33 | 4 |

After filtering of neighborhood based on upper limit on property price of $1.25 million dollars, about 260 entries remain.

### 3.4.2 Downselect based on Commute Time

Further shortlisting of neighborhoods is done by driving commute time to both Midtown Manhattan and JFK airport less than 30 mins. Although commuting through transits like subway is preferred by lot of people, using subway as filtering criteria will exclude regions not near a subway.

Table 11 shows first 5 rows of dataset after shortlisting for price and commute time to Midtown Manhattan and JFK airport less than 30 mins. At this point, only 34 entries or neighborhood are left.

*Table 11: Dataset after shortlisting for upper limit on property price $1.25 million dollars and drive commute time less than 30 mins. (only first 10 entries shown)*

| | Neighborhood | Borough | Latitude | Longitude | Residential Property Price median ($) | Drive to Midtown Manhattan (mins) | Subway to Midtown Manhattan (mins) | Drive to JFK Airport (mins) | Total Crime Reported 2018 |
|---|---|---|---|---|---|---|---|---|---|
| **20** | Longwood | BRONX | 40.815099 | -73.895788 | 595731 | 28 | 46 | 29 | 59 |
| **25** | Mott Haven | BRONX | 40.806239 | -73.916100 | 656187 | 24 | 45 | 28 | 114 |
| **35** | Port Morris | BRONX | 40.801664 | -73.913221 | 424923 | 25 | 60 | 29 | 12 |
| **45** | West Farms | BRONX | 40.839475 | -73.877745 | 498512 | 29 | 48 | 28 | 62 |
| **145** | Murray Hill | MANHATTAN | 40.748303 | -73.978332 | 711365 | 7 | 3 | 28 | 44 |
| **152** | Tudor City | MANHATTAN | 40.746917 | -73.971219 | 873520 | 12 | 23 | 28 | 13 |
| **161** | Astoria Heights | QUEENS | 40.770317 | -73.894680 | 1158351 | 23 | 113 | 24 | 11 |
| **165** | Beechhurst | QUEENS | 40.792781 | -73.804365 | 1026172 | 28 | 89 | 22 | 5 |
| **170** | Briarwood | QUEENS | 40.710935 | -73.811748 | 712692 | 29 | 50 | 15 | 39 |
| **176** | Downtown Flushing | QUEENS | 40.761164 | -73.829368 | 820455 | 27 | 34 | 21 | 72 |

### 3.4.3 Downselect based on Crimes

Only 34 neighborhoods are remaining after screening for property price and commute time. In previous section, upper limit of 10 total crimes was used for windowing of dataset. With severe reduction in data from 287 neighborhoods to 34 neighborhoods, the criteria for safe neighborhood must be revisited.

Figure 19 shows the distribution of total crime per neighborhood of shortlisted neighborhoods. An arbitrary choice of 12 total crimes reported in 2018 is chosen, which should retain six safest neighborhoods.

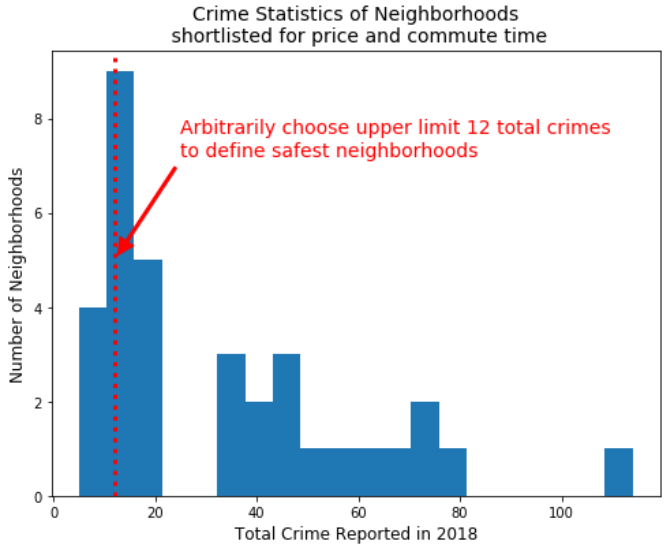



*Figure 19: Distribution of total crime in 2018 per neighborhood shortlisted for property price and commute time.*

The final shortlist result with only 6 neighborhoods listed in table 12. Interestingly, all of neighborhoods belong to Queens borough. As mentioned earlier, different criteria can be chosen to shortlist neighborhoods.

*Table 12: Final list of neighborhoods of NY city shortlisted based on price, crime and commute time.*

| | Neighborhood | Borough | Latitude | Longitude | Residential Property Price median ($) | Drive to Midtown Manhattan (mins) | Subway to Midtown Manhattan (mins) | Drive to JFK Airport (mins) | Total Crime Reported 2018 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Beechhurst | QUEENS | 40.792781 | -73.804365 | 1026172 | 28 | 89 | 22 | 5 |
| 1 | Utopia | QUEENS | 40.733500 | -73.796717 | 943645 | 29 | 70 | 23 | 7 |
| 2 | Malba | QUEENS | 40.790602 | -73.826678 | 1110734 | 28 | 63 | 22 | 7 |
| 3 | Steinway | QUEENS | 40.775923 | -73.902290 | 895391 | 24 | 39 | 25 | 8 |
| 4 | Astoria Heights | QUEENS | 40.770317 | -73.894680 | 1158351 | 23 | 113 | 24 | 11 |
| 5 | Oakland Gardens | QUEENS | 40.745619 | -73.754950 | 975464 | 29 | 52 | 20 | 11 |

### 3.4.4 Visualize the Shortlisted Neighborhoods on Map

Map of final shortlisted neighborhoods is shown in figure 20. Comparison with crime heatmap in figure 8 confirms that these neighborhoods do not lie within clusters of high crime neighborhoods.


*Figure 20: Map showing centroid of final short-listed neighborhoods in NY city with 1-mile radius circle.*

## 3.5 Final Selection from Shortlisted Neighborhood

Shortlisted neighborhoods are now compared based on variety and count of most common desired venues like restaurants, shopping centers, park and colleges. These venues were searched using Foursquare API and within 1-mile radius of the neighborhood centroid. This method of comparing neighborhoods by venues has subjectivity associated with it. Nonetheless, the methodology defined can be adjusted for different needs.

### 3.5.1 Select using Restaurant & Cuisine

Subset of data containing summary of restaurants cuisine or category of shortlisted neighborhoods is obtained. The tree map shown in figure 21 shows composition of top 5 restaurant cuisine or category.

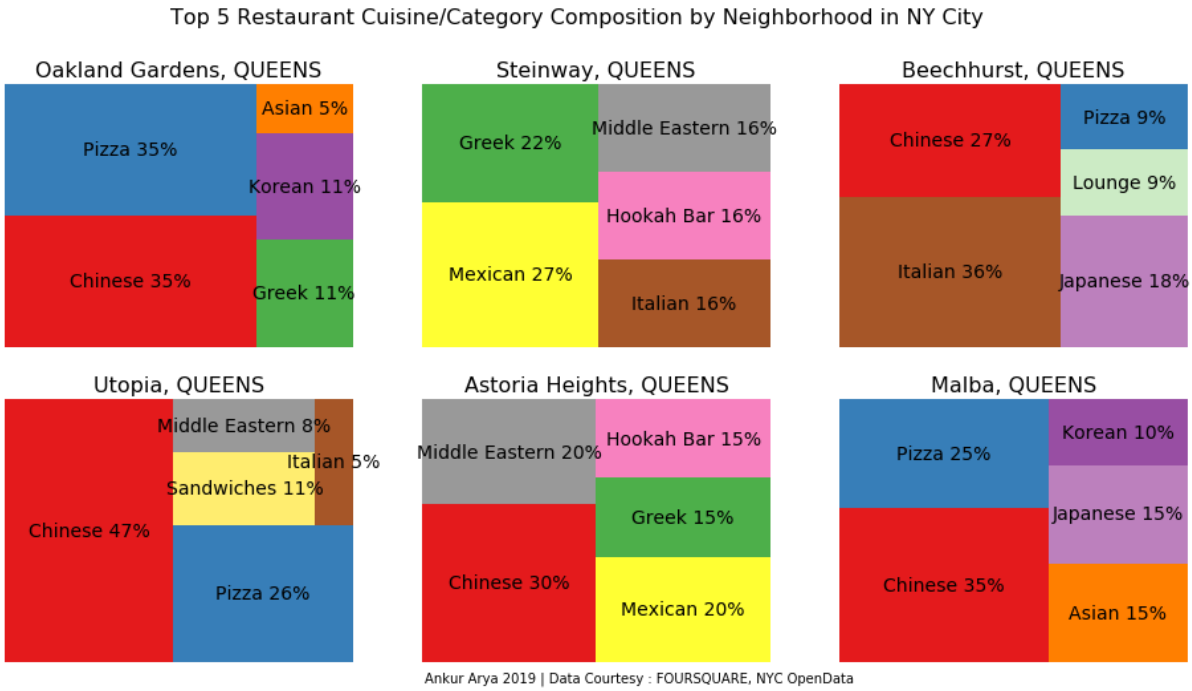



*Figure 21: Top 5 restaurant cuisine / category composition of neighborhoods in NY city.*

- Oakland Gardens and Utopia have large portion of Chinese and Pizza restaurants.
- Steinway and Astoria Heights have good presence of Mexican, Greek and Middle Eastern (includes Kosher) cuisines.
- Malba and Beechhurst are dominantly Asian (Chinese, Japanese, Korean).

Steinway and Astoria Heights are located close to each other and them combined have a wide variety of international cuisines.

### 3.5.2 Select using Shopping/Parks/College

Besides restaurant cuisine or categories, shortlisted neighborhoods are compared using the dataset of venues like big shopping complexes/malls, parks and colleges/universities. The dataset contains total counts of each of these venues per neighborhood. Subset of data is obtained containing summaries of only shortlisted neighborhoods. Figure 22 shows total number of venues for shortlisted neighborhoods.

- All neighborhoods have ample number of parks.
- Oakland Gardens is the only neighborhood devoid of shopping centers or malls.
- Malba and Utopia lack presence of college or universities in their vicinity.

Astoria Heights, Beechhurst and Steinway all three kinds of venues - parks, shopping centers/malls and colleges/universities.

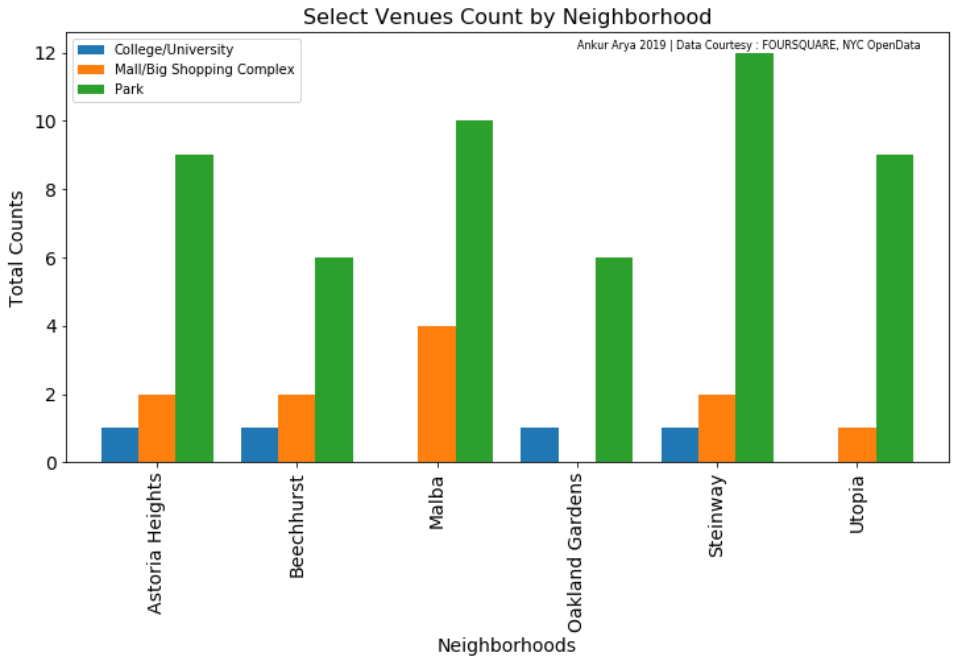



*Figure 22: Total count of select venues by shortlisted neighborhoods in NY city.*

# 4 RESULTS

It must be reminded that neighborhoods being referred in the article correspond to neighborhood centroids recognized by Department of City Planning of New York city. Review median property prices, total crime reported in 2018 and commute times for shortlisted neighborhoods in table 13 below.

*Table 13: Review of shortlisted neighborhoods of NY city*

| Neighborhood | Borough | Residential Property Price median ($) | Drive to Midtown Manhattan (mins) | Subway to Midtown Manhattan (mins) | Drive to JFK Airport (mins) | Total Crime Reported 2018 |
|---|---|---|---|---|---|---|
| Beechhurst | QUEENS | 1026172 | 28 | 89 | 22 | 5 |
| Utopia | QUEENS | 943645 | 29 | 70 | 23 | 7 |
| Malba | QUEENS | 1110734 | 28 | 63 | 22 | 7 |
| Steinway | QUEENS | 895391 | 24 | 39 | 25 | 8 |
| Astoria Heights | QUEENS | 1158351 | 23 | 113 | 24 | 11 |
| Oakland Gardens | QUEENS | 975464 | 29 | 52 | 20 | 11 |

- Steinway is the cheapest for residential properties and close to Manhattan by driving or transit like subway. Astoria Heights is close to Steinway and closest to Manhattan and JFK airport.

Review of comparison of restaurant cuisine / category (figure 21) and other venues like park, malls and colleges (figure 22) of shortlisted neighborhood centroids.

- Steinway and Astoria Heights combined have a big diversity of cuisine.
- Astoria Heights, Beechhurst and Steinway all three kinds of venues - parks, malls and colleges.

Combination of above results leads to two neighborhood centroids located within same neighborhood of Astoria.

(1) (Ditmars) Steinway [16], lowest residential property price $895391 and shorter commute to Midtown Manhattan both by driving and subway makes it attractive. It has the highest number of parks with presence of colleges and malls in the vicinity. Steinway also has a good blend of international cuisines - Greek, Middle Eastern, Mexican and Italian.

(2) Astoria Heights [17], which is very close to Steinway, the residential property price median is $1,158,351 and closest to both Midtown Manhattan and JKF airport. Like Steinway, it has ample numbed of parks, malls and colleges. It has wide variety of world cuisine - Chinese, Middle Eastern, Mexican and Greek.

Both the neighborhood centroids appear equally attractive. With lower property prices and shorter commute by subway and drive to Midtown Manhattan, **(Ditmars) Steinway stands out as an excellent place to live in NY city.**

# 5 Discussion

This work is successfully able to narrow down search of best neighborhoods from 300 choices to a few chosen from real and recent data. Data obtained from NYC Open Data, Zillow, Google, and Foursquare is judiciously used. The methodology and results use certain criteria to shortlist neighborhoods, a similar methodology can be applied with different criteria.

To limit the scope, only certain kinds of datasets were used - residential property prices, total crime reported in 2018, commute time to Midtown Manhattan and JFK Airport, places of interest like restaurants, parks, shopping malls, and colleges. Several other factors could be considered like school ratings, proximity to beach or religious center, number of sex offenders, demographics, the density of population, noise level, close to friends and family residents, proximity to spouse's work. Some of these additional factors can be added to the dataset in the future.

Simple statistics from dataset were used like median of all kinds of property prices and the total sum of crimes per neighborhood. Other measures like the spread of prices (interquartile range or standard deviation), type of property (apartments or houses), the category of crime (larceny or drug or sexual assault) can be added using the same dataset. The criteria for shortlisting based on median price, total crime and commute times determined after examination of tradeoffs between price and commute time can be tailored to individuals need.

Confusing categories in restaurant cuisines (like Asian vs. Chinese) and missing features like brands or outlets within Malls, size of parks, type of Universities (Law or Medicine) can create ambiguity and more subjectivity in using venues as a rigid criterion for comparison of neighborhoods. Instead, it is used as qualitative aid in choosing neighborhoods. The work could be further expanded to search for hyperlocal areas like streets or blocks to live in NY city.

# 6 CONCLUSION

(Ditmars) Steinway is determined as the best place to live in New York City. A careful selection was done using real and recent datasets like residential property prices, total crime reported to NYPD in 2018 and commute time to Midtown Manhattan and JFK airport. Further exploration using data of restaurant cuisines, parks, shopping complexes, and colleges is done to finalize the selection. Applying similar methodology and different shortlist criteria is expected to give different results. The framework presented in this work can be used to add richer data and complex selection criteria.

# 7 ACKNOWLEDGEMENT

Author would like to thank the reader for taking time to go through the article. It must be noted that this article uses one of several approaches to search for a desirable place to live in NY city. The work would not be possible without availability of real and recent data from NYC Open Data, Zillow, Google and Foursquare.

# 8 REFERENCES

[1] "New York City", Wikipedia, https://en.wikipedia.org/wiki/New_York_City

[2] "Map/Data", *NYC Planning,* City of New York, https://www1.nyc.gov/site/planning/data-maps/maps-geography.page

[3] *NYC Open Data*, City of New York. https://opendata.cityofnewyork.us/

[4] "GetRegionChildren API", *Zillow API Network,* Zillow, https://www.zillow.com/howto/api/GetRegionChildren.htm

[5] "Distance Matrix API", *Google Maps Platform*, Google, https://developers.google.com/maps/documentation/distance-matrix/intro

[6] "Search for Venues", *Foursquare Developer,* Foursquare, https://developer.foursquare.com/docs/api/venues/search

[7] "Where to Live in New York City", GitHub

[8] "NYPD Complaint Data Current (Year To Date)", *NYC Open Data,* City of New York, https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data

[9] Open Addresses, http://results.openaddresses.io/ [zip file]

[10] "NYC Address Points", *NYC Open Data,* City of New York, https://data.cityofnewyork.us/City-Government/NYC-Address-Points/g6pj-hd8k

[11] “Neighborhood Names GIS”, *NYC Open Data*, City of New York, https://data.cityofnewyork.us/City-Government/Neighborhood-Names-GIS/99bc-9p23

[12] “Midtown Manhattan”, Wikipedia, https://en.wikipedia.org/wiki/Midtown_Manhattan

[13] “John F. Kennedy International Airport”, Wikipedia, https://en.wikipedia.org/wiki/John_F._Kennedy_International_Airport

[14] “Neighborhood Tabulation Areas”, *NYC Open Data*, City of New York, https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq

[15] “QGIS Algorithm Provider”, *Documentation QGIS Testing,* QGIS, https://docs.qgis.org/2.8/en/docs/user_manual/processing_algs/qgis/index.html

[16] “Ditmars Steinway”, *Astoria Queens*, Wikipedia, https://en.wikipedia.org/wiki/Astoria,_Queens#Ditmars

[17] “Astoria Heights”, *Astoria Queens,* Wikipedia, https://en.wikipedia.org/wiki/Astoria,_Queens#Astoria_Heights