Bao Ngoc Dinh

Final Project: The Battle of the Neighborhoods

Data Acquisition

- We need the following data:
- NYC Crime data: https://data.cityofnewyork.us/Public-Safety/Citywide-Crime-Statistics/c5dk-m6ea
- geojson of NYC: provided by the course
- Foursquare NYC data

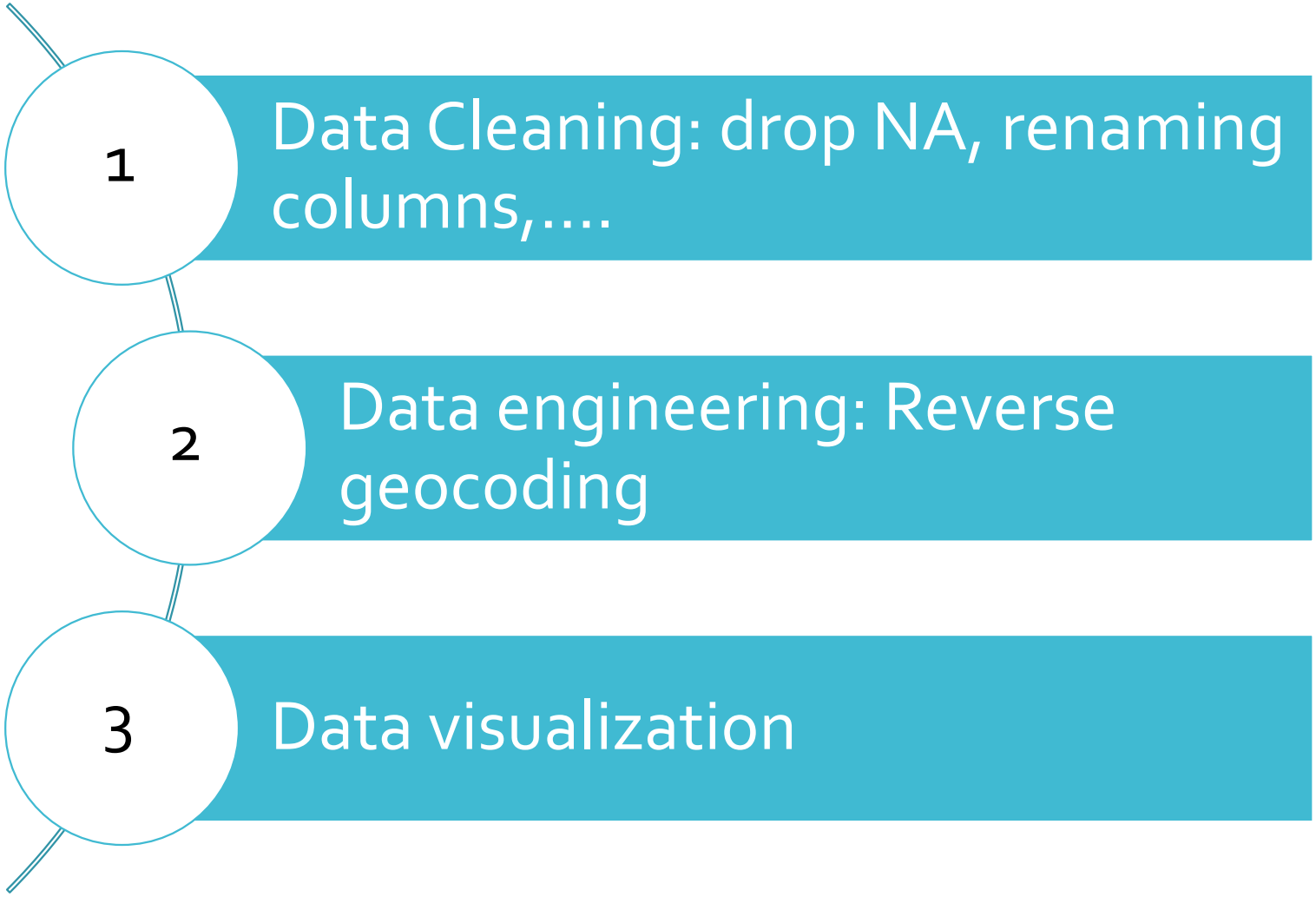
Project Description

A big city in New York can be scary. We created this file to compare NYC neighborhoods by boroughs and neighborhoods. This guide would be perfect for new people who just moved to the city and wants to know which neighborhood fits them in terms of popular venues and low crime rate.

What are some safe neighborhoods in Manhattan? Which ones aren't?

Visualization of the neighborhoods and crime stats?

Crime Dataframe

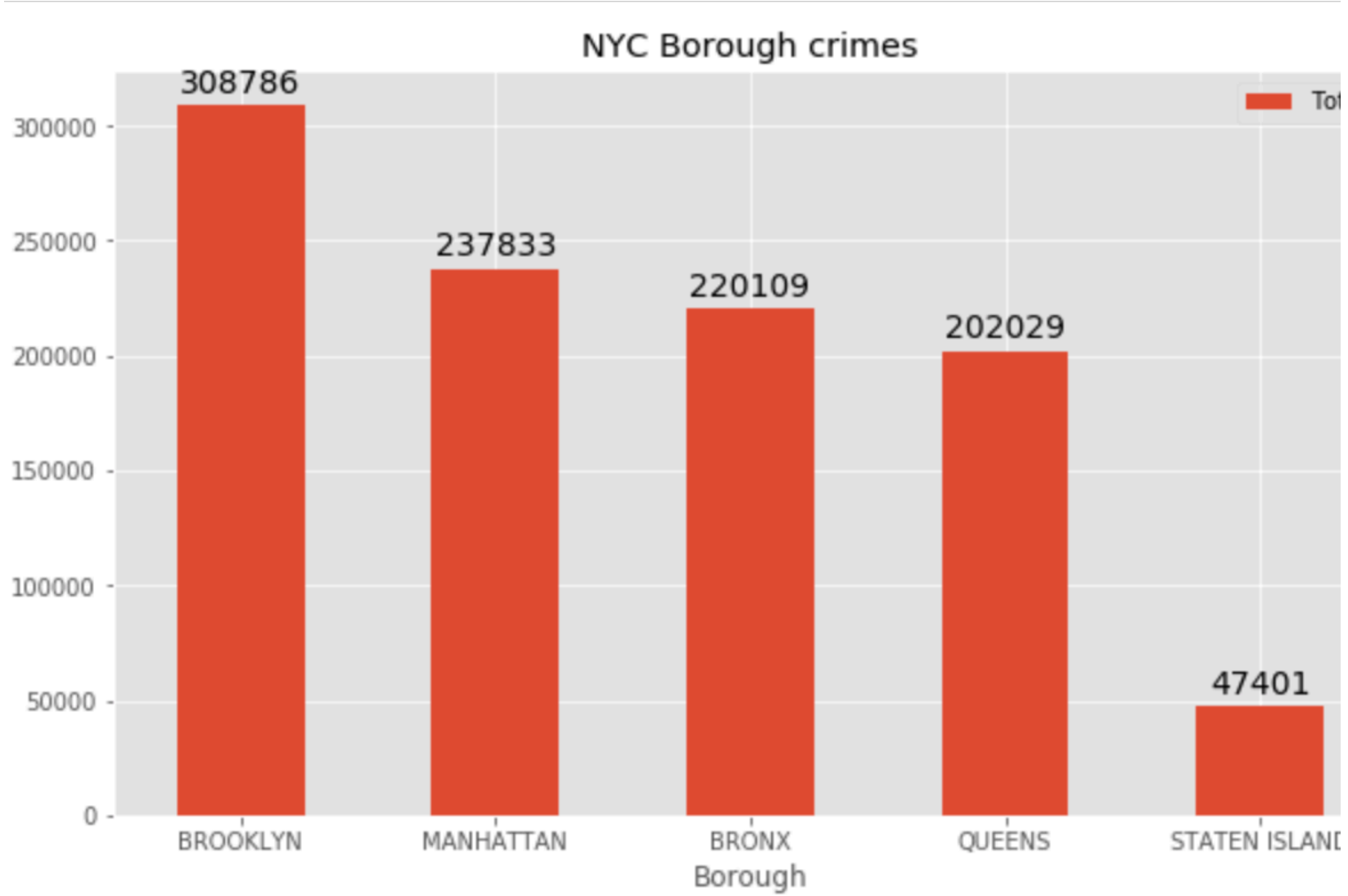


	Complaint From Date	Report Date	Offense Classification Code	Offense Description	Crime successfully completed	Level of offense	Jurisdiction Description	Borough Name	Precinct	Location of Occurence	Premise Description	coordinat
0	12/31/2015	12/31/2015	113	FORGERY	COMPLETED	FELONY	N.Y. POLICE DEPT	BRONX	44.0	INSIDE	BAR/NIGHT CLUB	1007314.
1	12/31/2015	12/31/2015	101	MURDER & NON-NEGL. MANSLAUGHTER	COMPLETED	FELONY	N.Y. POLICE DEPT	QUEENS	103.0	OUTSIDE	NaN	1043991.
2	12/31/2015	12/31/2015	117	DANGEROUS DRUGS	COMPLETED	FELONY	N.Y. POLICE DEPT	MANHATTAN	28.0	NaN	OTHER	999463.
3	12/31/2015	12/31/2015	344	ASSAULT 3 & RELATED OFFENSES	COMPLETED	MISDEMEANOR	N.Y. POLICE DEPT	QUEENS	105.0	INSIDE	RESIDENCE-HOUSE	1060183.
4	12/31/2015	12/31/2015	344	ASSAULT 3 & RELATED OFFENSES	COMPLETED	MISDEMEANOR	N.Y. POLICE DEPT	MANHATTAN	13.0	FRONT OF	OTHER	987606.

Overview of the dataframe

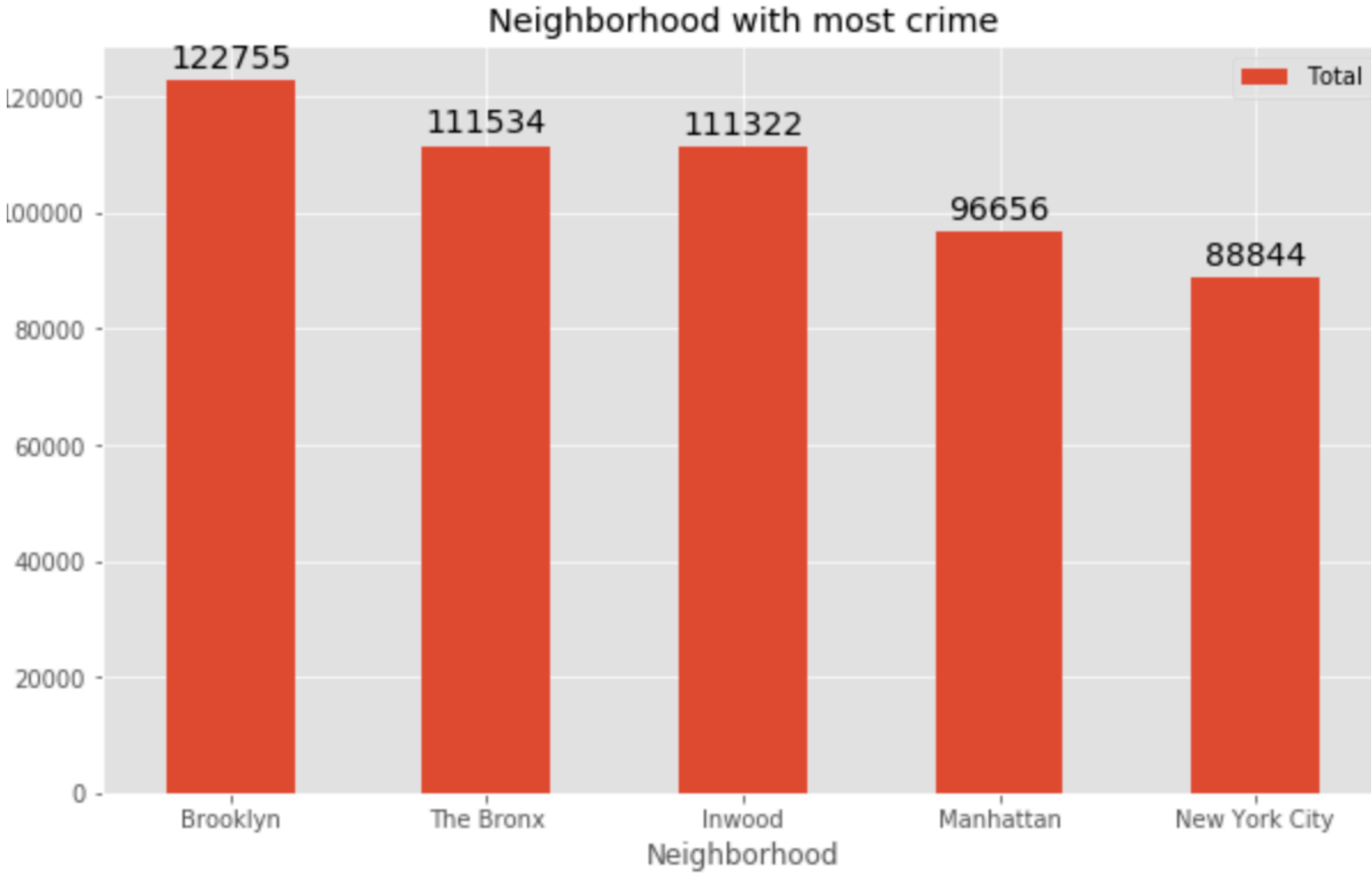
Data Visualization -1

When we look at this graph, we can see that Brooklyn has the most amount of crimes. Staten Island is much lower, almost 5 times less than Brooklyn. However, this could be misleading. Staten Island is the borough with the least amount of people, hence the decreased offense. Brooklyn is the borough with the biggest population. However, since this data is aggregate data of many years and the population number is per year, taking the total amount of offense divided by census data would not be a good indicator of crime per population.



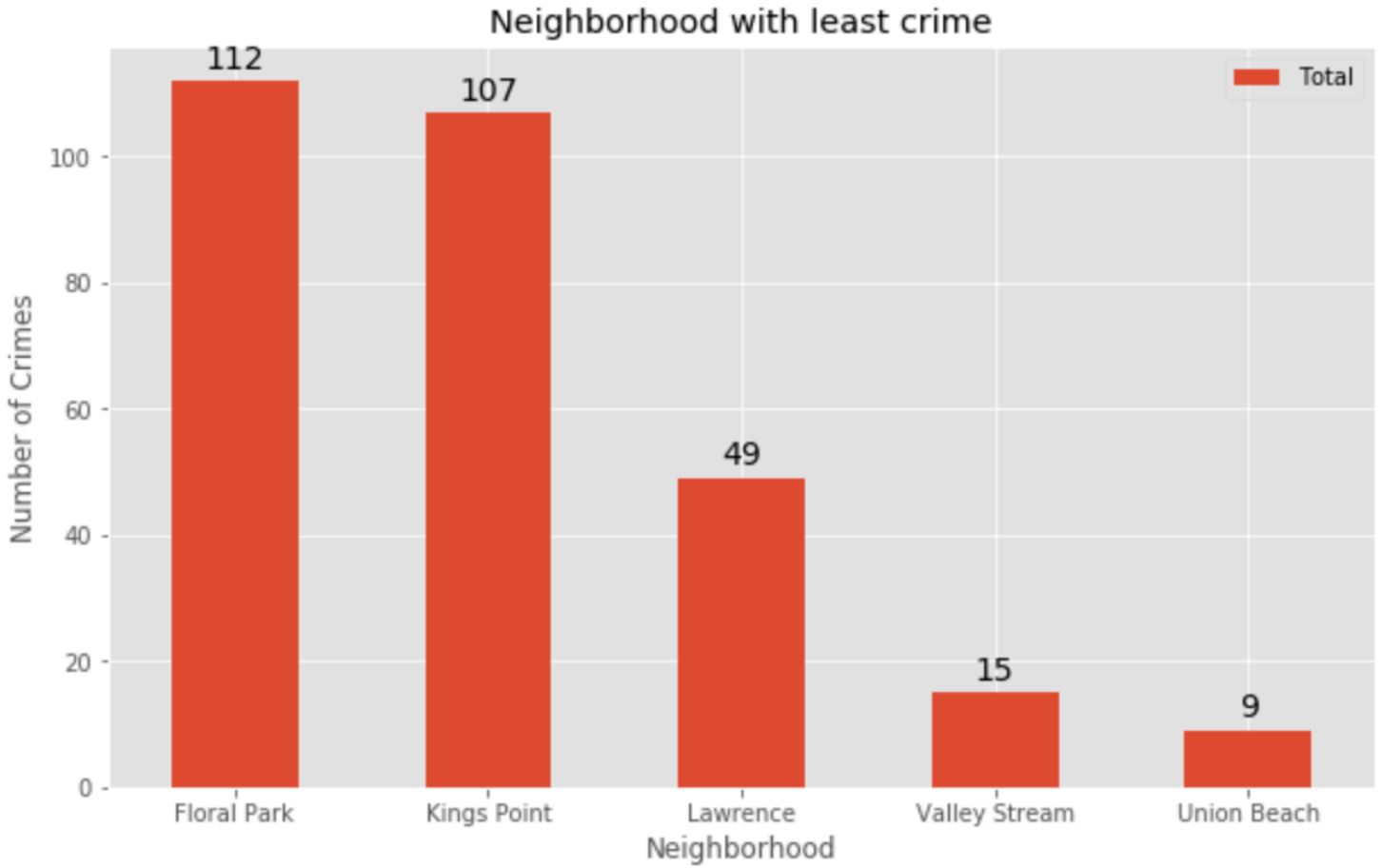
Data Visualization-2

As I said, the rg geocode is not perfect. The neighborhood division is not perfect, reflected by the two 'Manhattan' and 'New York City' category. It's confusing and misleading. Inwood is a specfic neighborhood in Upper Manhattan.



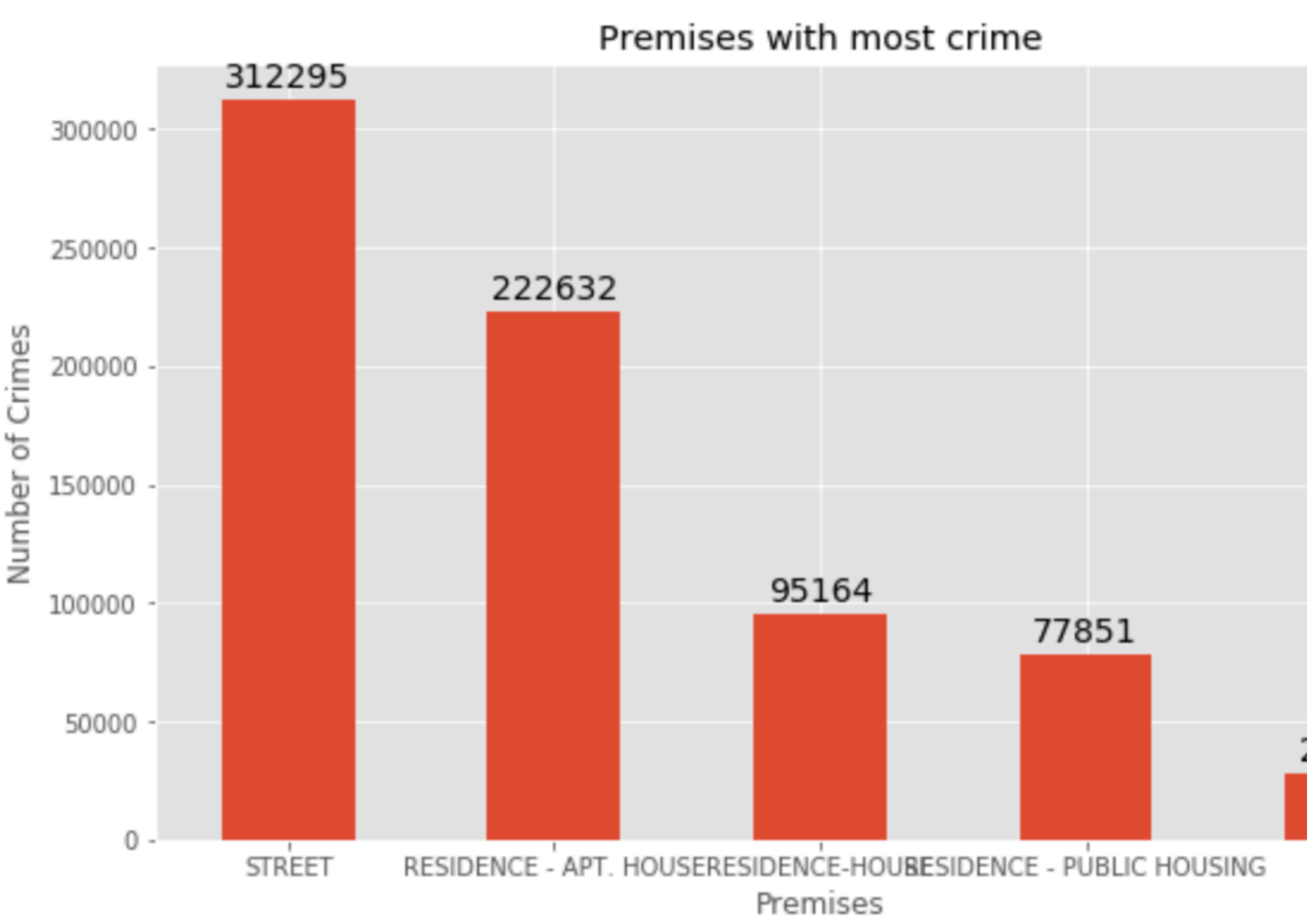
Data Visualization -3

I graphed the neighborhoods with least crime. These neighborhoods are all part of Long Island, NY. For those of you who don't live in New York the city is divided into five boroughs, and Long Island is not one of the five boroughs but it shares a border with Queens and part of NY state. Thus, many of the calls from Long Island would made it into NYC data cause of proximity. However, these neighborhoods would be considered 'bigger metro NYC area'.



Data Visualization-4

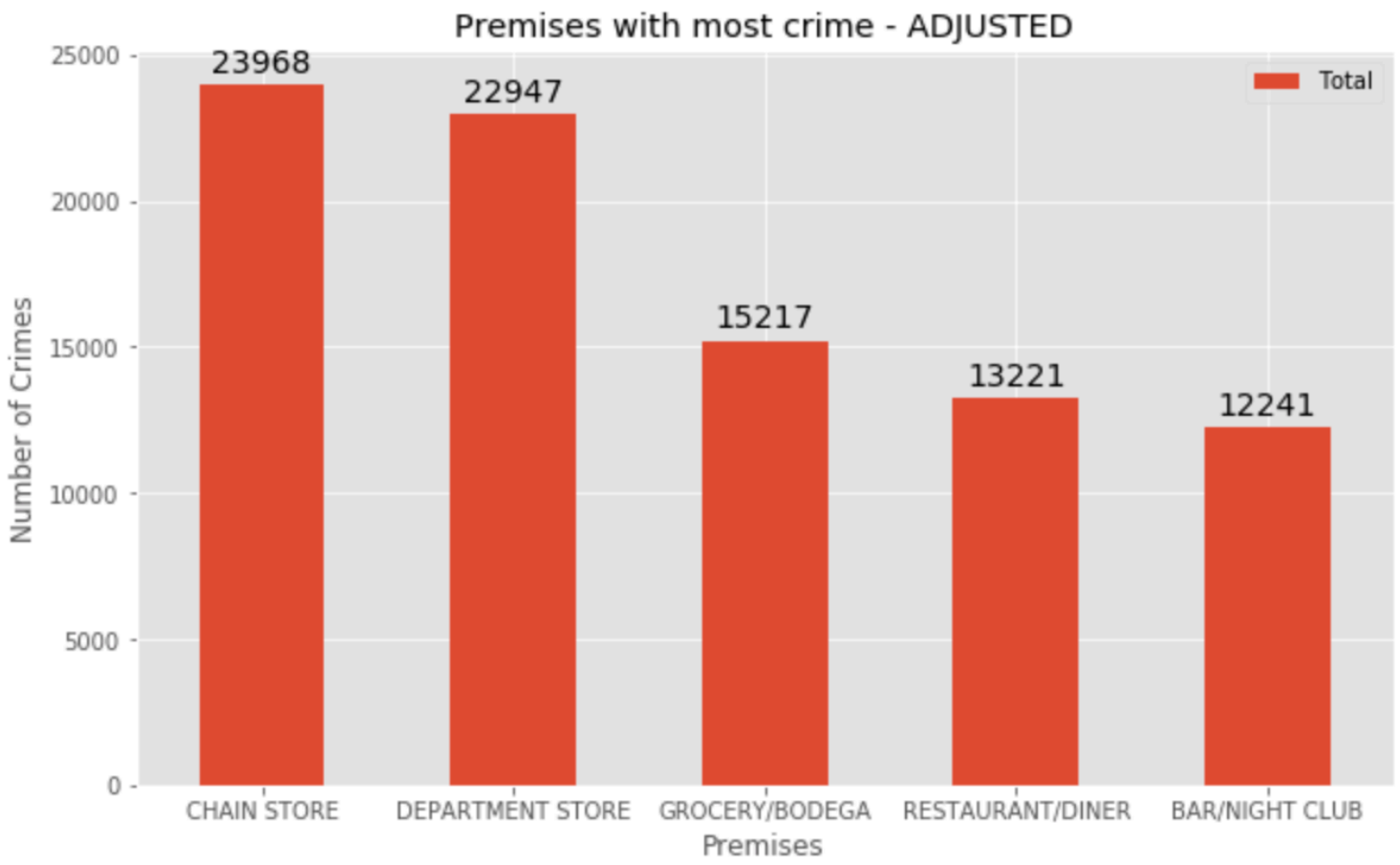
As we can see, the offenses often takes place in an apartment or a street. However, if we want to compare this data against popular venues in NYC generated by Foursquare API, we are more interested in places such as yoga, cafe, brunch spots... so we will remove some category of places of offenses from the dataframe.



Data Visualization -5

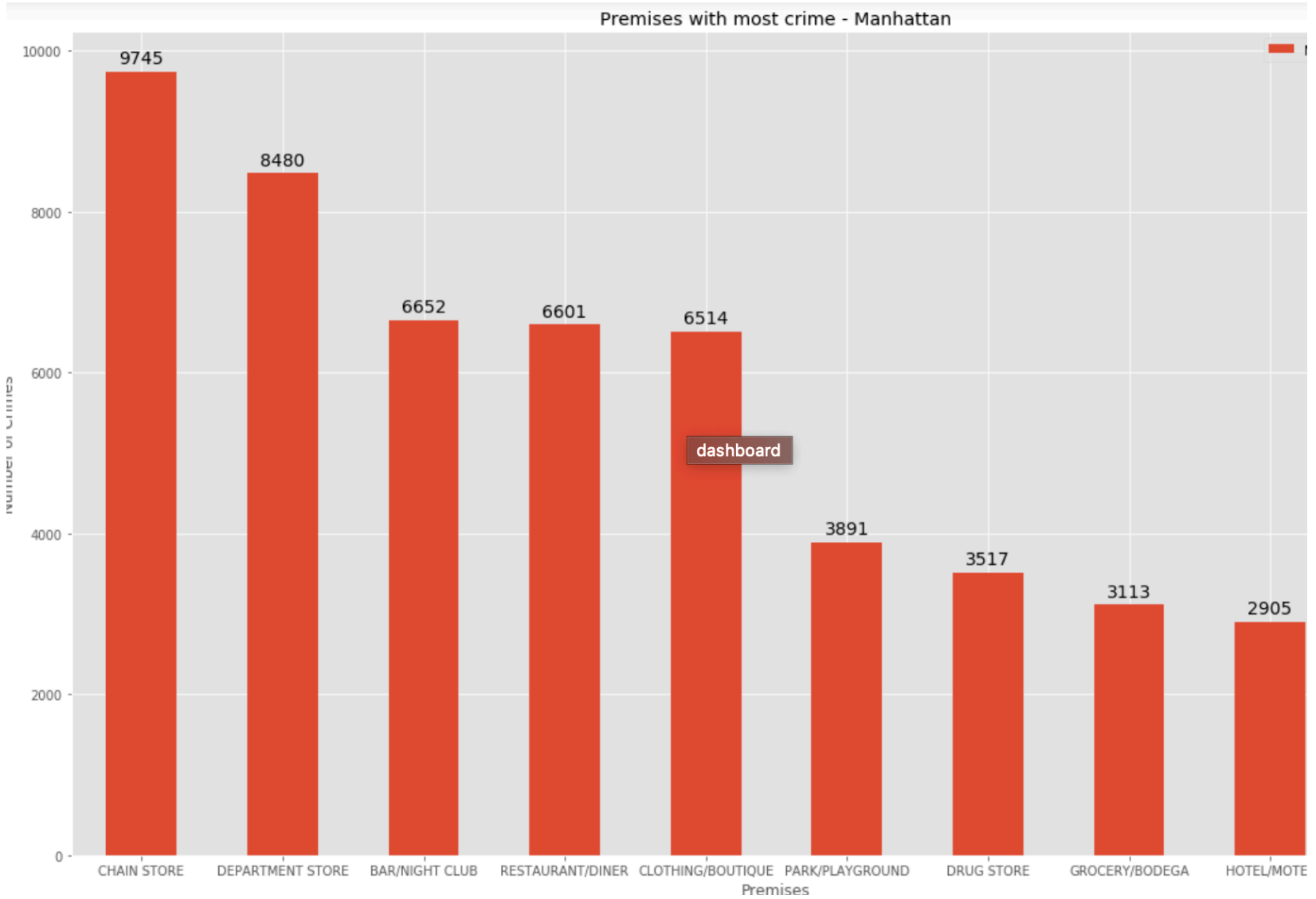
After removing residential building and ambiguous values, we got our top 5 places with most crime occurrences in NYC. We could cross-match this against our clusters and their most popular venues.

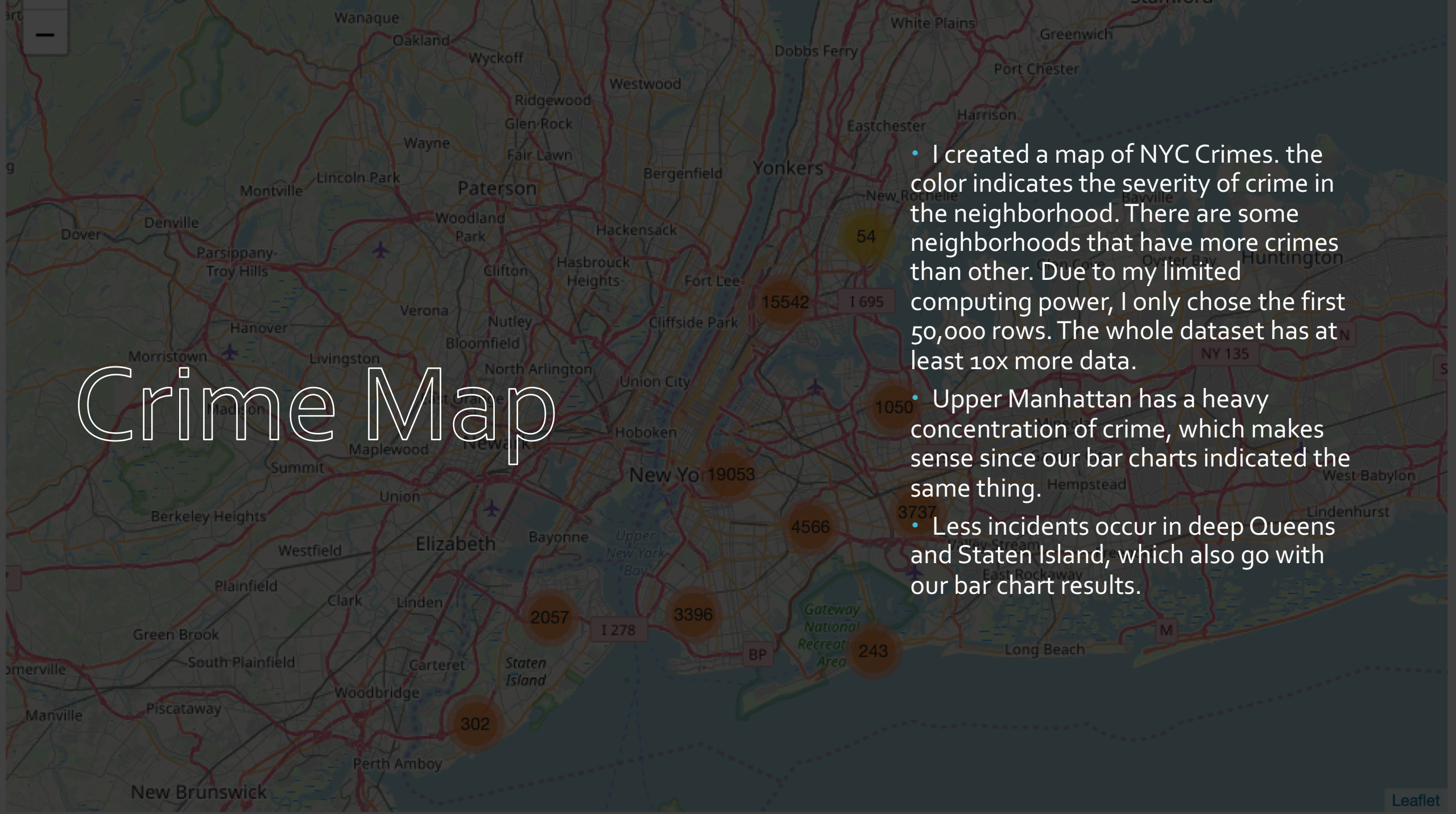
However, Our focus will be on Manhattan since the only part of NYC clustered is Manhattan.



Data Visualization – 6

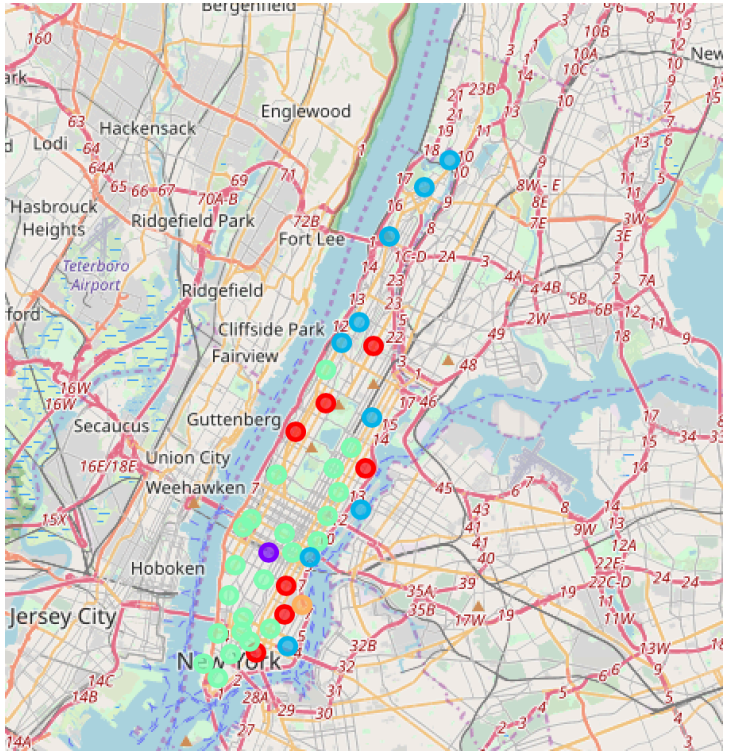
Since our cluster produces 10 most popular venues in the area, we are going to cross match the 10 premises with most crimes against the 10 most popular venues in the area.





- I created a map of NYC Crimes. the color indicates the severity of crime in the neighborhood. There are some neighborhoods that have more crimes than other. Due to my limited computing power, I only chose the first 50,000 rows. The whole dataset has at least 10x more data.
- Upper Manhattan has a heavy concentration of crime, which makes sense since our bar charts indicated the same thing.
- Less incidents occur in deep Queens and Staten Island, which also go with our bar chart results.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Coffee Shop	Park	Hotel	Gym	Wine Shop	Women's Store	Memorial Site	Italian Restaurant	Pizza Place	Plaza
1	Carnegie Hill	Coffee Shop	Pizza Place	Cosmetics Shop	Japanese Restaurant	Gym	French Restaurant	Yoga Studio	Café	Bookstore	Wine Shop
2	Central Harlem	Chinese Restaurant	African Restaurant	French Restaurant	Cosmetics Shop	Seafood Restaurant	Bar	American Restaurant	Cafeteria	Bookstore	Gym / Fitness Center
3	Chelsea	Coffee Shop	Italian Restaurant	Bakery	Ice Cream Shop	Theater	American Restaurant	Hotel	Nightclub	Cupcake Shop	French Restaurant
4	Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Vietnamese Restaurant	Optical Shop	Spa	Bakery	Hotpot Restaurant	Salon / Barbershop	Noodle House



5 clusters for Manhattan based on Foursquare API

```
.loc[manhattan_merged['Cluster Labels'] == 1, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
33	Midtown South	Korean Restaurant	Hotel Bar	Japanese Restaurant	Coffee Shop	Dessert Shop	Hotel	Cocktail Bar	American Restaurant	Cosmetics Shop	Gastropub

```
Cluster 1 has the following matches with our crime data: restaurant/diner, bar, hotel.
That's 3/10 matches
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Vietnamese Restaurant	Optical Shop	Spa	Bakery	Hotpot Restaurant	Salon / Barbershop	Noodle House
6	Central Harlem	Chinese Restaurant	African Restaurant	French Restaurant	Cosmetics Shop	Seafood Restaurant	Bar	American Restaurant	Cafeteria	Bookstore	Gym / Fitness Center
9	Yorkville	Italian Restaurant	Coffee Shop	Bar	Gym	Pizza Place	Deli / Bodega	Wine Shop	Sushi Restaurant	Japanese Restaurant	Diner
12	Upper West Side	Italian Restaurant	Wine Bar	Bar	Coffee Shop	Indian Restaurant	Bakery	Café	Mediterranean Restaurant	Restaurant	Pub
19	East Village	Bar	Ice Cream Shop	Wine Bar	Pizza Place	Mexican Restaurant	Chinese Restaurant	Italian Restaurant	Speakeasy	Coffee Shop	Cocktail Bar
25	Manhattan Valley	Bar	Indian Restaurant	Coffee Shop	Playground	Pizza Place	Mexican Restaurant	Thai Restaurant	Yoga Studio	Japanese Restaurant	Hawaiian Restaurant
27	Gramercy	Italian Restaurant	Bar	Pizza Place	Mexican Restaurant	Bagel Shop	Cocktail Bar	Thai Restaurant	Grocery Store	Thrift / Vintage Store	Comedy Club

Cluster 0 has the following matches with our crime data: restaurant/diner, bar, bodega, clothing/boutique, park/playground. That's 5/10 matches

Examine Clusters and Crime

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Marble Hill	Sandwich Place	Coffee Shop	Yoga Studio	Deli / Bodega	Supplement Shop	Steakhouse	Shopping Mall	Seafood Restaurant	Pizza Place	Department Store
2	Washington Heights	Café	Bakery	Grocery Store	Mobile Phone Shop	Coffee Shop	Deli / Bodega	Mexican Restaurant	Latin American Restaurant	New American Restaurant	Park
3	Inwood	Mexican Restaurant	Café	Restaurant	Lounge	Park	Chinese Restaurant	Deli / Bodega	Pharmacy	Wine Bar	Caribbean Restaurant
4	Hamilton Heights	Pizza Place	Coffee Shop	Café	Deli / Bodega	Mexican Restaurant	Sandwich Place	Indian Restaurant	Sushi Restaurant	Park	Yoga Studio
5	Manhattanville	Coffee Shop	Italian Restaurant	Mexican Restaurant	Bus Stop	Seafood Restaurant	Deli / Bodega	Park	Lounge	Gastropub	Supermarket
7	East Harlem	Mexican Restaurant	Thai Restaurant	Bakery	Latin American Restaurant	Pizza Place	Deli / Bodega	Taco Place	Cocktail Bar	Sandwich Place	Liquor Store
11	Roosevelt Island	Sandwich Place	Bridge	Greek Restaurant	Gym	Gym / Fitness Center	Coffee Shop	Liquor Store	Supermarket	School	Soccer Field
20	Lower East Side	Art Gallery	Coffee Shop	Café	Pizza Place	Shoe Store	Pharmacy	Japanese Restaurant	Cocktail Bar	Chinese Restaurant	Bakery
36	Tudor City	Café	Park	Mexican Restaurant	Greek Restaurant	Pizza Place	Diner	Coffee Shop	Deli / Bodega	Sushi Restaurant	Spanish Restaurant

Cluster 2 has the following matches with our crime data: restaurant/diner, bar, hotel, drugstore, department store, park/playground, bodega, fast food. That's 8/10 matches Inwood is actually one of the neighborhood with the most crimes, according to our bar plot

Examine Clusters and Crime-2

```
In [212]: erged.loc[manhattan_merged['Cluster Labels'] == 3, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape[1
```

Out[212]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	Upper East Side	Italian Restaurant	Exhibit	Art Gallery	Bakery	Juice Bar	Gym / Fitness Center	Coffee Shop	French Restaurant	Hotel	Pizza Place
10	Lenox Hill	Coffee Shop	Italian Restaurant	Pizza Place	Cocktail Bar	Sushi Restaurant	Gym / Fitness Center	Café	Burger Joint	Gym	Cycle Studio
13	Lincoln Square	Italian Restaurant	Café	Theater	Concert Hall	Plaza	Performing Arts Venue	Gym / Fitness Center	French Restaurant	Indie Movie Theater	Clothing Store
14	Clinton	Theater	Italian Restaurant	Gym / Fitness Center	Coffee Shop	American Restaurant	Hotel	Wine Shop	Gym	Spa	Sandwich Place
15	Midtown	Hotel	Sporting Goods Shop	Coffee Shop	Theater	Clothing Store	Bookstore	French Restaurant	Steakhouse	Café	Japanese Restaurant
16	Murray Hill	Sandwich Place	Coffee Shop	American Restaurant	Hotel	Japanese Restaurant	Italian Restaurant	Gym / Fitness Center	Sushi Restaurant	Bar	Restaurant
17	Chelsea	Coffee Shop	Italian Restaurant	Bakery	Ice Cream Shop	Theater	American Restaurant	Hotel	Nightclub	Cupcake Shop	French Restaurant
18	Greenwich Village	Italian Restaurant	Clothing Store	Sushi Restaurant	Café	Indian Restaurant	French Restaurant	Gym	Chinese Restaurant	Ice Cream Shop	Gourmet Shop

Examine Clusters and Crime-3

Cluster 3 has the following matches with our crime data: restaurant/diner, boutique, bar, hotel, drugstore, department store, park/playground, bodega, fast food.

That's 9/10 matches

However, this cluster is very wide spread and not every single neighborhood in the clusters have 9/10 matches.

```
loc[manhattan_merged['Cluster Labels'] == 4, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
37	Stuyvesant Town	Bar	Park	Fountain	Gas Station	Baseball Field	Harbor / Marina	Pet Service	Cocktail Bar	Coffee Shop	Heliport

Cluster 4 has the following matches with our crime data: restaurant/diner, bar, park/playground, bodega. That's 4/10 matches

Examine Clusters and Crime-4

Conclusions

- The dataset does have its setback. However, through this analysis I hope I can give you a guide into crime prevention in the Big Apple. Avoid neighborhoods with high crime matches, and review my crime stats visualization before you make a trip to NY
- There are limitations to the dataset and the process that I, however, do hope to improve:
- First of all, the reverse geocoding was not an exact match. Reverse geocoding is a costly process and to get precise neighborhood, we would need paid services such as Google API. The dataset is too big so it would be costly to run reverse geocoding. Thus, we cannot merge the crime and the cluster dataset based on neighborhood matches.
- NYC's venue definitions and terms vs. Foursquare's definition and terms. NYC data refers to any sit-down restaurant loosely as restaurant while Foursquare divided it into different cuisines. Thus, the matches are not exact match.
- **However**, there is potential for this project to scale further. I can cluster nyc neighborhoods. The only reason why I did not was because the Foursquare API and my code encountered some problems and it returned all neighborhoods as one big cluster. Thus, that would not provide any insight into the different clusters. However, we could scale this and do a borough-by-borough guide. we could also overlay the map of different neighborhoods (geojson data), and the clusters, and crime heatmap as one big choropleth map. However, this wasn't possible due to limited reverse geocoding