THE BATTLE OF NEIGHBORHOODS

-Kalyan Sundar

INTRODUCTION / BUSINESS PROBLEM

- One of the leading multinational Hotel and resort groups in Europe is interested in starting Luxury spa in Unites States of America and they would like to analyze the data pertaining to main cities and its localities of the country. The suitable locality can be identified based on the inputs such as Per Capita Income of the Main cities, Density of Population in the city, Population of each location and Various venues in the location

- We will consider various factors such as Travel & Transport, Residence, community, Work and offices, Events, Food and recreation, Arts & Entertainment, Shops & Service, College & University, Nightlife Spot and Outdoors & Recreation.

- Ultimate aim is to find the suitable locality which will attract more customers and it is highly dependent on the location and earning capacity of individuals visiting the location and residing at the vicinity.

DATA ACQUISITION AND PREPROCESSING

Datasets from Wikipedia and Foursquare will be used which will help us to identify the optimal locality to start the Luxury spas in the city of USA.

- population density and coordinates : https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population
- Per Capita Income : https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income
- Using Four Square API to get the following
- List of all venues in each city
- List of all venues in each locality in the selected city

Using the above data, we will first select best city to proceed with based on the values like Population density, per capita income of the state, number of venues (as we are giving weights to each venue based on its category).

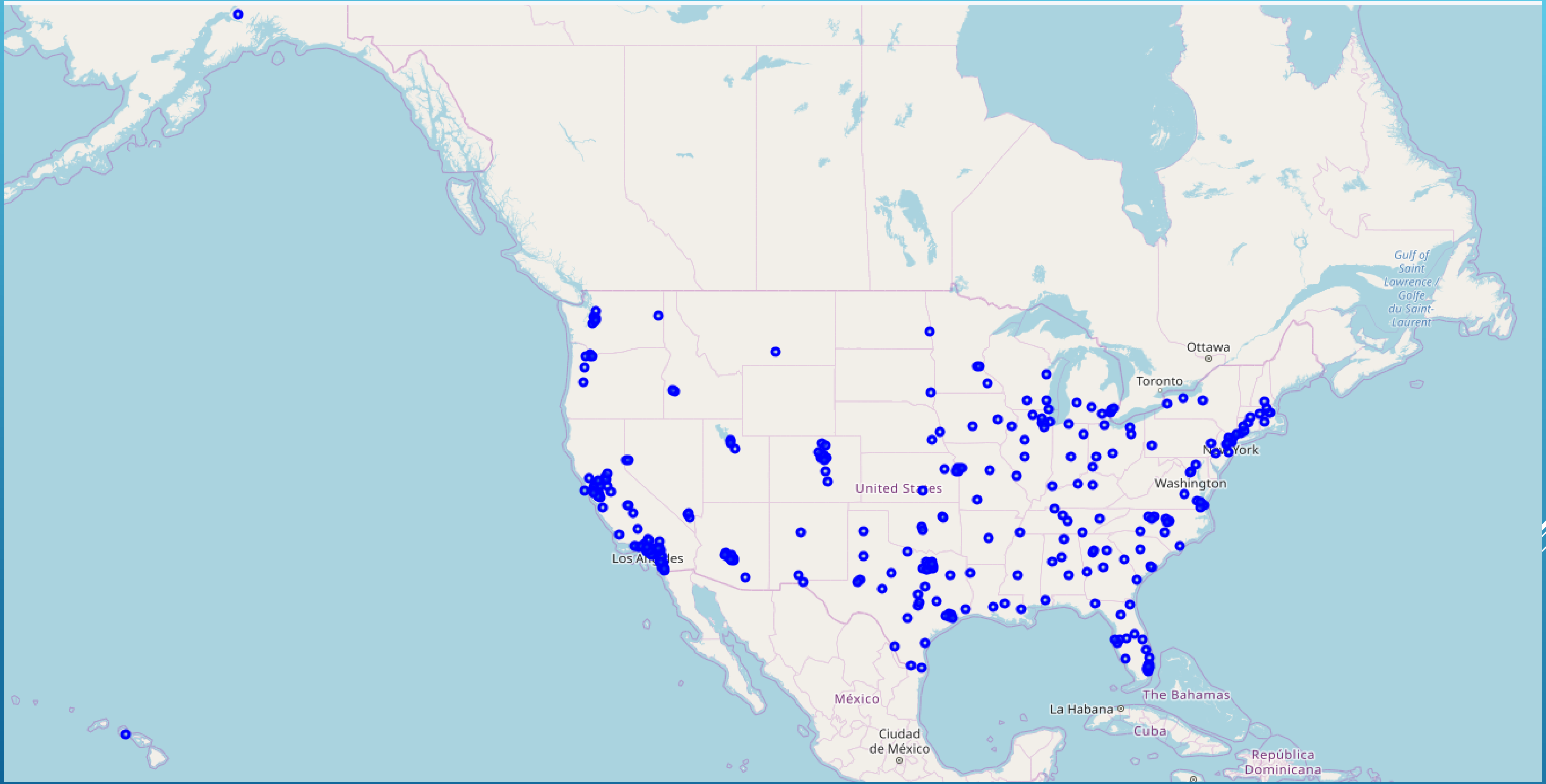
The first stage is to select the city in the USA based on available datasets, then to find the locality in the city which is influenced by the events and other above said factors.

METHODOLOGY

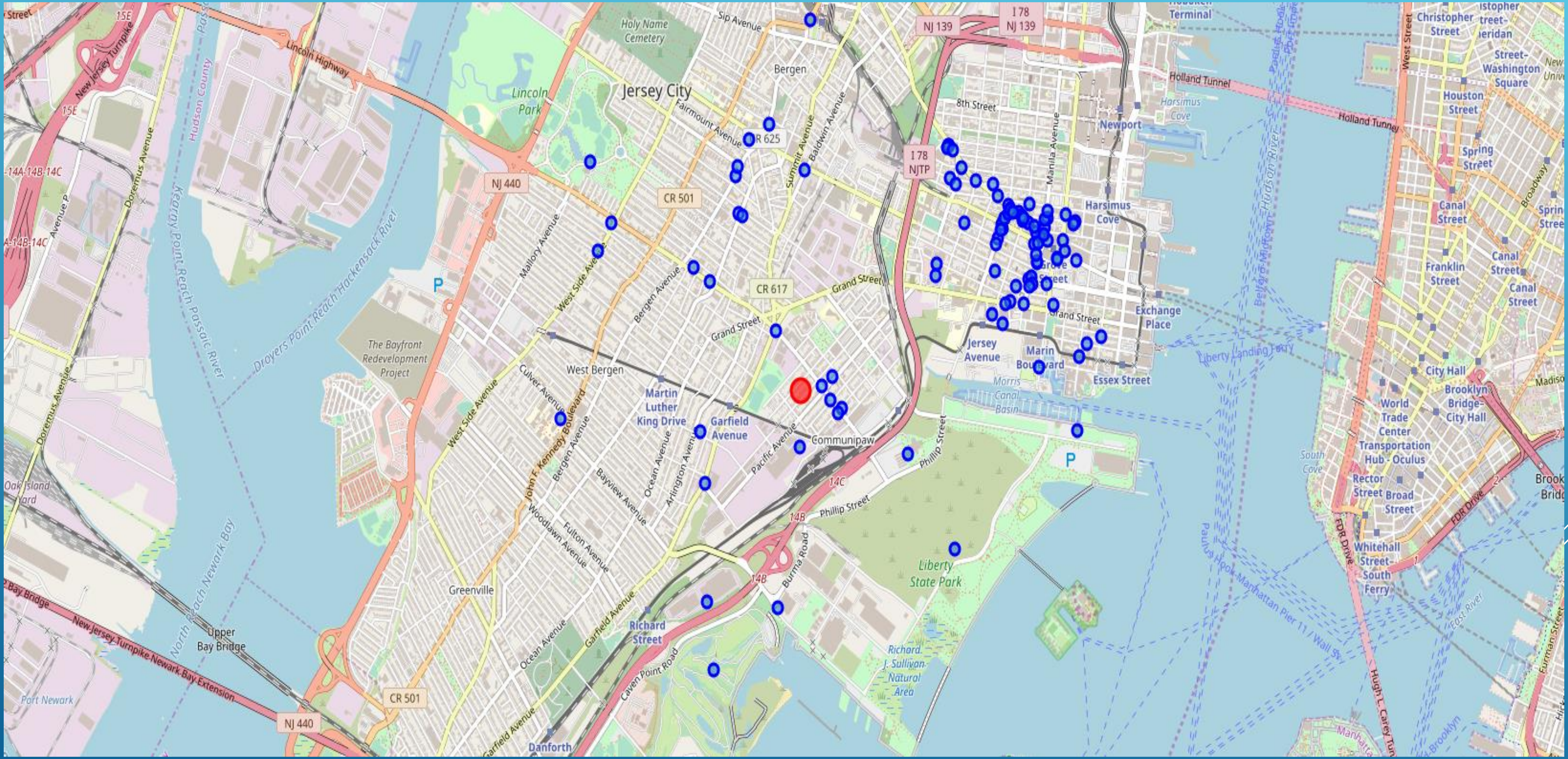
To identify optimal vicinity to start new Luxury SPA venture below methods are followed:

- The data pertaining to various factors are acquired from Wikipedia pages . The acquired data has been scraped using beautifulsoup into panda data frames which are categorized under cities, per capital income, states, co-ordinates, area and density of population
- Appropriate clean-up and processing of Data frame has been done.
- Venue details of various cities are extracted using Foursquare API . Each category has been weighted based on our preferences. City with maximum weightage is selected for our venture.
- Venues are clustered based on category using K-Means algorithm . Co-ordinates of the location having maximum weightage is found out.

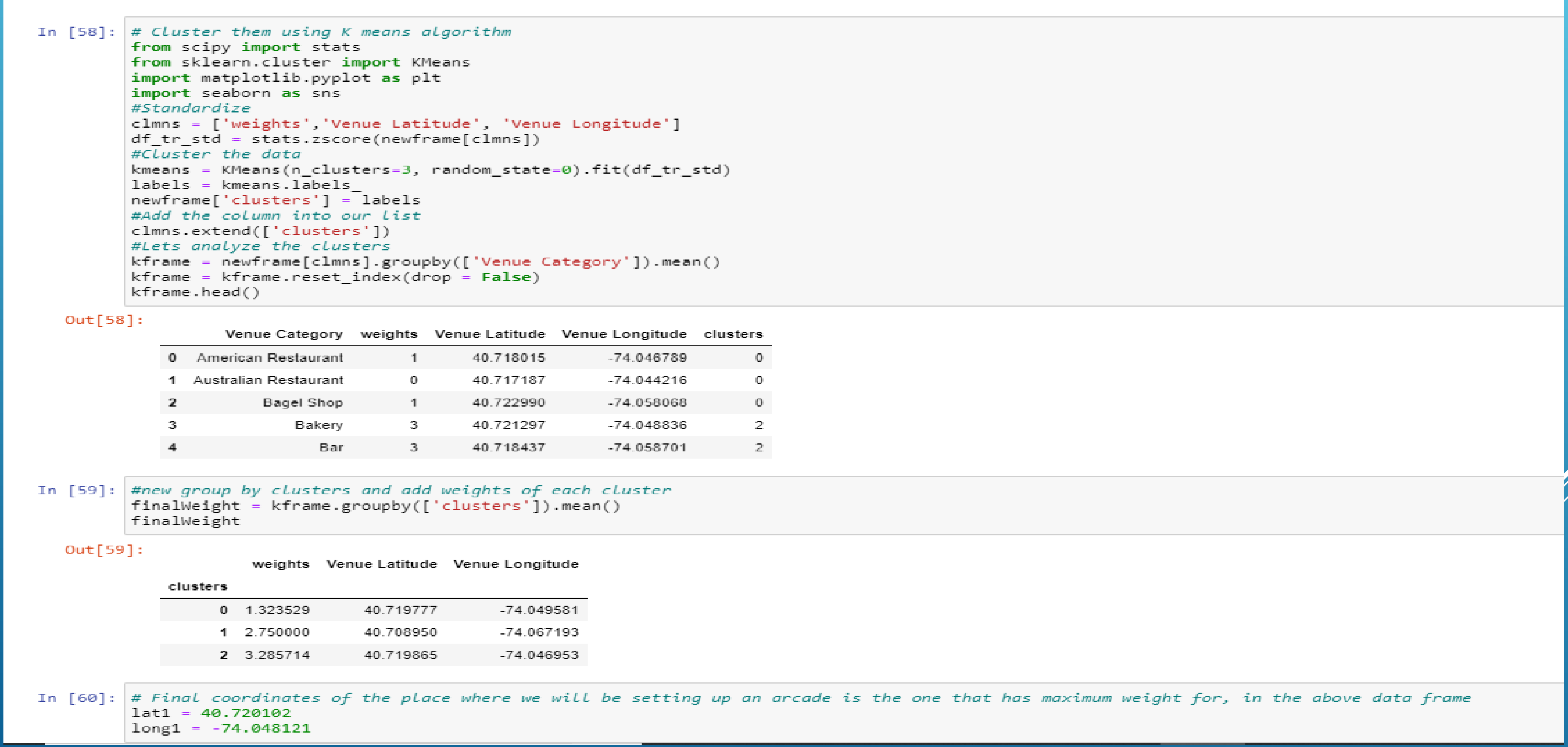
EXTRACTED DATA PLOTTED ON USA MAP



VENUES FROM FOURSQUARE API



K-MEANS CLUSTERING TO GET CO-ORDINATES



In [58]:

```
# Cluster them using K means algorithm
from scipy import stats
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns
#Standardize
clmns = ['weights','Venue Latitude', 'Venue Longitude']
df_tr_std = stats.zscore(newframe[clmns])
#Cluster the data
kmeans = KMeans(n_clusters=3, random_state=0).fit(df_tr_std)
labels = kmeans.labels_
newframe['clusters'] = labels
#Add the column into our list
clmns.extend(['clusters'])
#Lets analyze the clusters
kframe = newframe[clmns].groupby(['Venue Category']).mean()
kframe = kframe.reset_index(drop = False)
kframe.head()
```

Out[58]:

	Venue Category	weights	Venue Latitude	Venue Longitude	clusters
0	American Restaurant	1	40.718015	-74.046789	0
1	Australian Restaurant	0	40.717187	-74.044216	0
2	Bagel Shop	1	40.722990	-74.058068	0
3	Bakery	3	40.721297	-74.048836	2
4	Bar	3	40.718437	-74.058701	2

In [59]:

```
#new group by clusters and add weights of each cluster
finalWeight = kframe.groupby(['clusters']).mean()
finalWeight
```

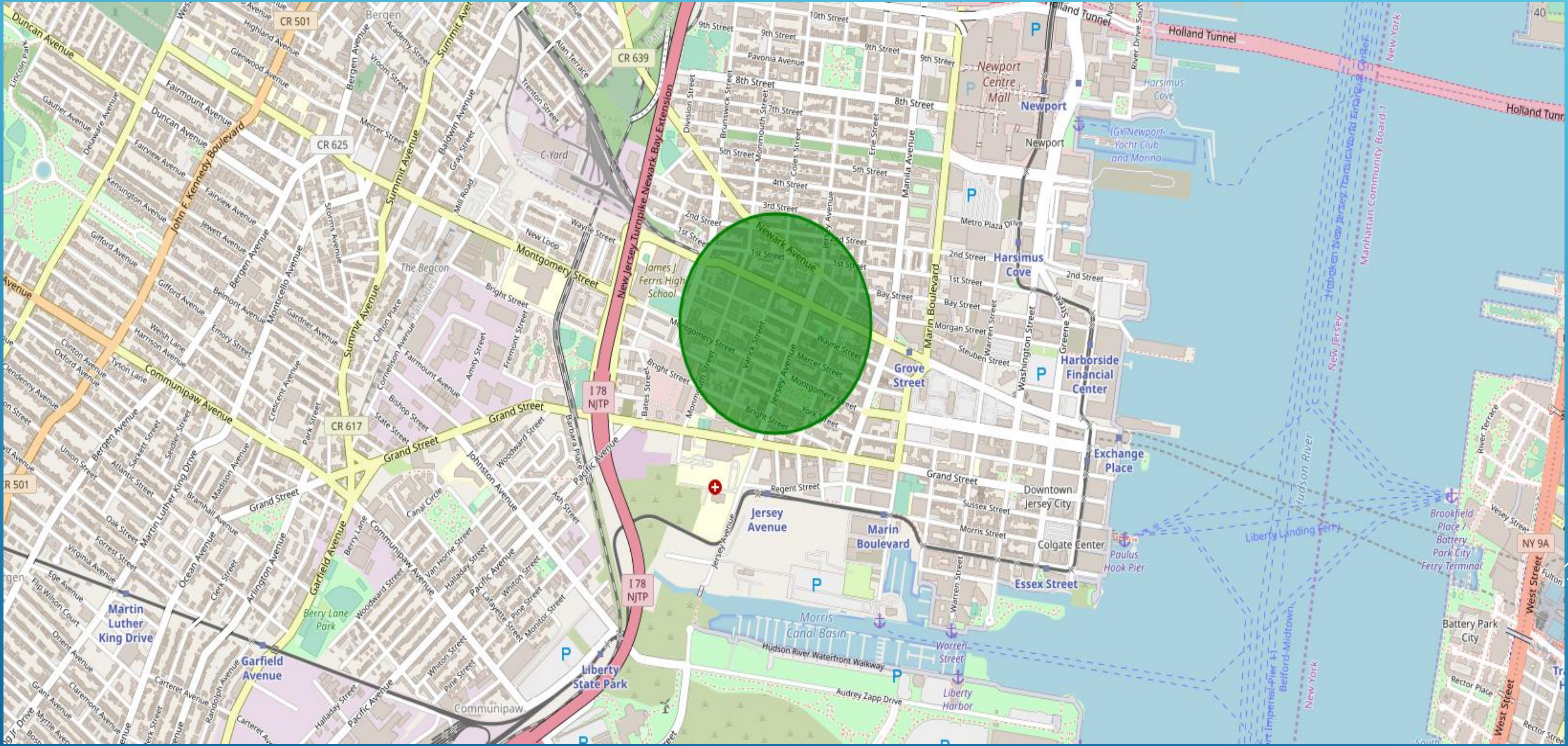
Out[59]:

	weights	Venue Latitude	Venue Longitude
clusters			
0	1.323529	40.719777	-74.049581
1	2.750000	40.708950	-74.067193
2	3.285714	40.719865	-74.046953

In [60]:

```
# Final coordinates of the place where we will be setting up an arcade is the one that has maximum weight for, in the above data frame
lat1 = 40.720102
long1 = -74.048121
```

OPTIMAL LOCATION FOR STARTING LUXURY SPA



RESULT / CONCLUSION

- We have arrived at the suggestion of better spot in the vicinity of Newark avenue and jersey Avenue.
- Based on the dependant factors and events that we considered in this project, we suggest to start the Luxury spa in the location where there will be more visitors be it tourists ,residents or professionals in that vicinity.
- This analysis can be further expanded deeply by considering other factors like crime rate, floating populations , rental expenses, customer data similar to other businesses such as Gyms, health and wellness centres, Saloons etc.