The Battle of Neighborhoods – Final Report

Introduction

Spas have become a pertinent cultural force, influencing not only how consumers manage their health, appearance, and stress, but also how consumers socialize, spiritualize, travel, and work. Rising levels of income, education, and sophistication among travelers and consumers worldwide have dramatically elevated the consciousness and desirability of spa treatments. The market potential of spa development is being captured by global and premium-brand spas that have expanded their service menus. As interest in physical well-being increases, spa therapy has become popular among consumers and has been recommended by many medical specialists

In 2017, there were over 149,000 spas, earning $93.6 billion in revenues and employing nearly 2.5 million workers. The spa sector has been growing by 9.9 percent annually from 2015–2017, and it is projected to reach $128 billion in 2022.

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

As a data scientist we have to take lot of venues into consideration which can directly or indirectly influence the customers at the locality to visit the Luxury spa.

In this project 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

In this project, I will be using datasets from Wikipedia and Foursquare 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).

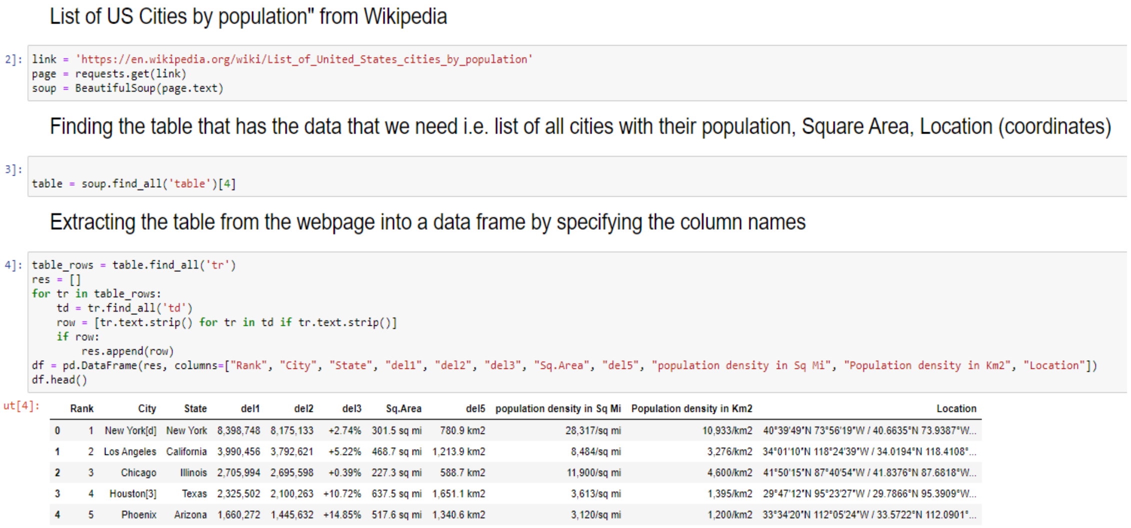
The first stage is to select the city in the USA based on available datasets, then to find the locality in the city which is influenced by the events and other above said factors.

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 beautiful soup 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.



List of US Cities by population" from Wikipedia

```
2]: link = 'https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population'
    page = requests.get(link)
    soup = BeautifulSoup(page.text)
```

Finding the table that has the data that we need i.e. list of all cities with their population, Square Area, Location (coordinates)

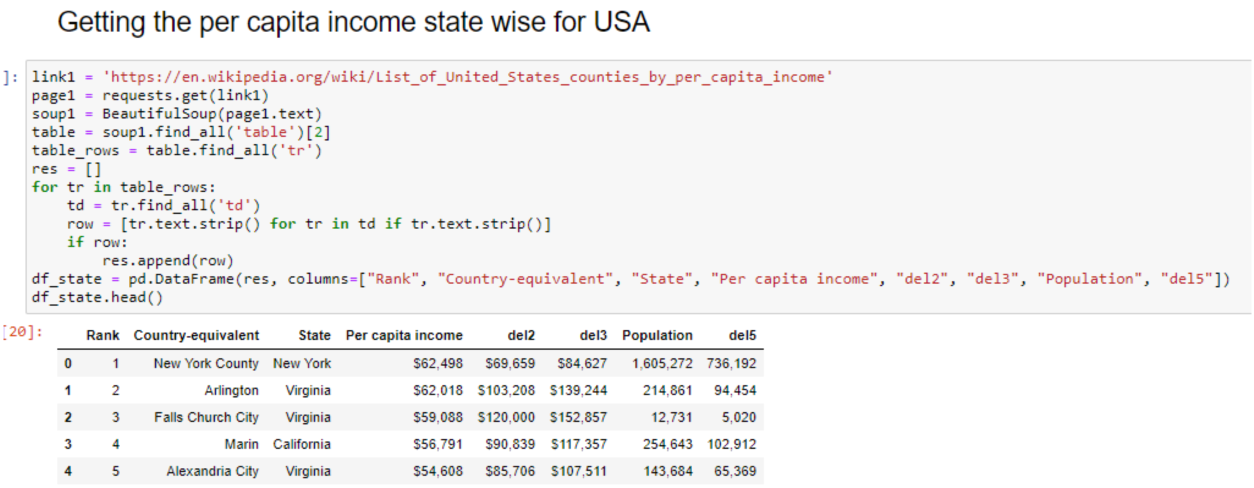
```
3]:
    table = soup.find_all('table')[4]
```

Extracting the table from the webpage into a data frame by specifying the column names

```
4]: table_rows = table.find_all('tr')
    res = []
    for tr in table_rows:
        td = tr.find_all('td')
        row = [tr.text.strip() for tr in td if tr.text.strip()]
        if row:
            res.append(row)
    df = pd.DataFrame(res, columns=["Rank", "City", "State", "del1", "del2", "del3", "Sq.Area", "del5", "population density in Sq Mi", "Population density in Km2", "Location"])
    df.head()
```

ut[4]:

	Rank	City	State	del1	del2	del3	Sq.Area	del5	population density in Sq Mi	Population density in Km2	Location
0	1	New York[d]	New York	8,398,748	8,175,133	+2.74%	301.5 sq mi	780.9 km2	28,317/sq mi	10,933/km2	40°39′49″N 73°56′19″W / 40.6635°N 73.9387°W...
1	2	Los Angeles	California	3,990,456	3,792,621	+5.22%	468.7 sq mi	1,213.9 km2	8,484/sq mi	3,276/km2	34°01′10″N 118°24′39″W / 34.0194°N 118.4108°...
2	3	Chicago	Illinois	2,705,994	2,695,598	+0.39%	227.3 sq mi	588.7 km2	11,900/sq mi	4,600/km2	41°50′15″N 87°40′54″W / 41.8376°N 87.6818°W...
3	4	Houston[3]	Texas	2,325,502	2,100,263	+10.72%	637.5 sq mi	1,651.1 km2	3,613/sq mi	1,395/km2	29°47′12″N 95°23′27″W / 29.7866°N 95.3909°W...
4	5	Phoenix	Arizona	1,660,272	1,445,632	+14.85%	517.6 sq mi	1,340.6 km2	3,120/sq mi	1,200/km2	33°34′20″N 112°05′24″W / 33.5722°N 112.0901°...



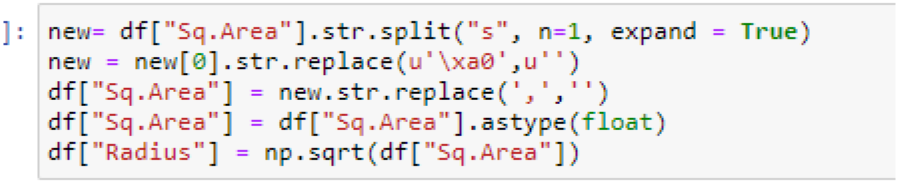
Getting the per capita income state wise for USA

```
]: link1 = 'https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income'
   page1 = requests.get(link1)
   soup1 = BeautifulSoup(page1.text)
   table = soup1.find_all('table')[2]
   table_rows = table.find_all('tr')
   res = []
   for tr in table_rows:
       td = tr.find_all('td')
       row = [tr.text.strip() for tr in td if tr.text.strip()]
       if row:
           res.append(row)
   df_state = pd.DataFrame(res, columns=["Rank", "Country-equivalent", "State", "Per capita income", "del2", "del3", "Population", "del5"])
   df_state.head()
```

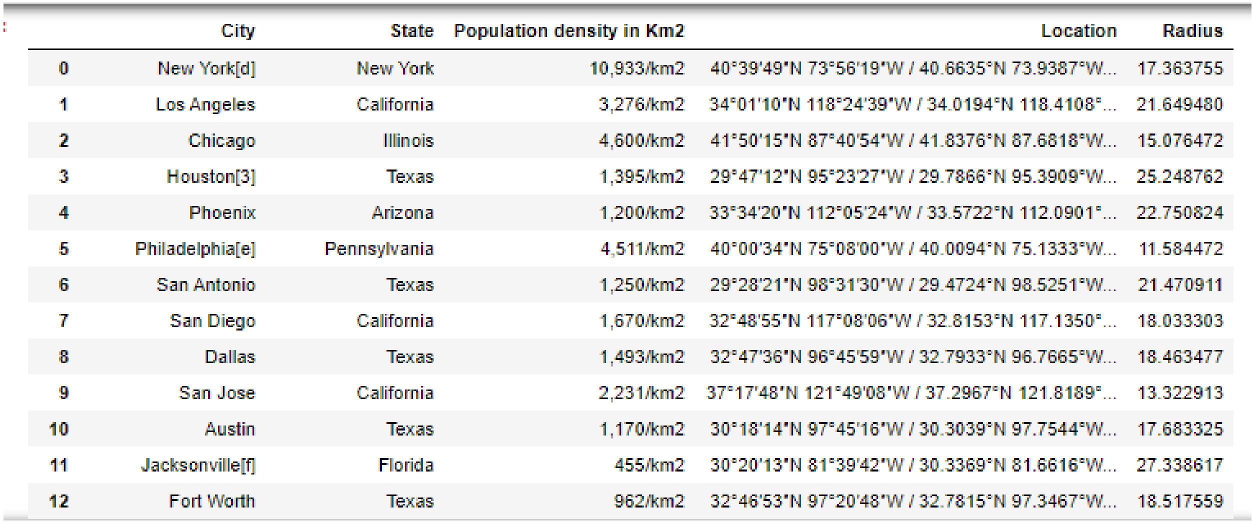
[20]:

	Rank	Country-equivalent	State	Per capita income	del2	del3	Population	del5
0	1	New York County	New York	$62,498	$69,659	$84,627	1,605,272	736,192
1	2	Arlington	Virginia	$62,018	$103,208	$139,244	214,861	94,454
2	3	Falls Church City	Virginia	$59,088	$120,000	$152,857	12,731	5,020
3	4	Marin	California	$56,791	$90,839	$117,357	254,643	102,912
4	5	Alexandria City	Virginia	$54,608	$85,706	$107,511	143,684	65,369

Finding the radius of each city with the help of Sq.Area, this step involves in preprocessing of the the column Sq.Area (changing its data type to float) then finding its square root



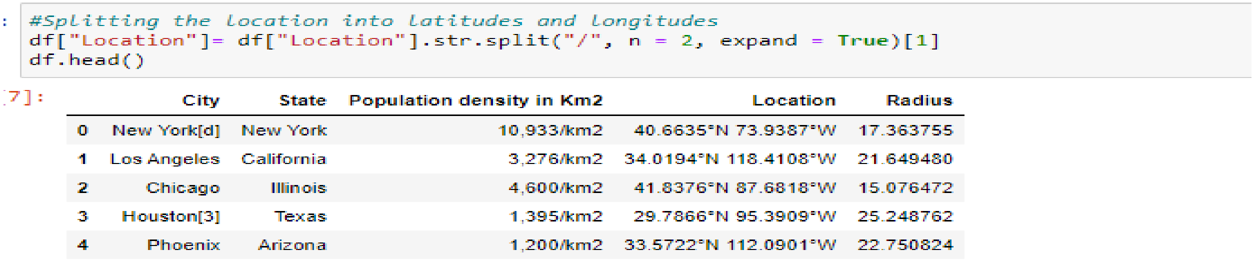
```
]: new= df["Sq.Area"].str.split("s", n=1, expand = True)
   new = new[0].str.replace(u'\xa0',u'')
   df["Sq.Area"] = new.str.replace(',','')
   df["Sq.Area"] = df["Sq.Area"].astype(float)
   df["Radius"] = np.sqrt(df["Sq.Area"])
```



:

	City	State	Population density in Km2	Location	Radius
0	New York[d]	New York	10,933/km2	40°39′49″N 73°56′19″W / 40.6635°N 73.9387°W...	17.363755
1	Los Angeles	California	3,276/km2	34°01′10″N 118°24′39″W / 34.0194°N 118.4108°...	21.649480
2	Chicago	Illinois	4,600/km2	41°50′15″N 87°40′54″W / 41.8376°N 87.6818°W...	15.076472
3	Houston[3]	Texas	1,395/km2	29°47′12″N 95°23′27″W / 29.7866°N 95.3909°W...	25.248762
4	Phoenix	Arizona	1,200/km2	33°34′20″N 112°05′24″W / 33.5722°N 112.0901°...	22.750824
5	Philadelphia[e]	Pennsylvania	4,511/km2	40°00′34″N 75°08′00″W / 40.0094°N 75.1333°W...	11.584472
6	San Antonio	Texas	1,250/km2	29°28′21″N 98°31′30″W / 29.4724°N 98.5251°W...	21.470911
7	San Diego	California	1,670/km2	32°48′55″N 117°08′06″W / 32.8153°N 117.1350°...	18.033303
8	Dallas	Texas	1,493/km2	32°47′36″N 96°45′59″W / 32.7933°N 96.7665°W...	18.463477
9	San Jose	California	2,231/km2	37°17′48″N 121°49′08″W / 37.2967°N 121.8189°...	13.322913
10	Austin	Texas	1,170/km2	30°18′14″N 97°45′16″W / 30.3039°N 97.7544°W...	17.683325
11	Jacksonville[f]	Florida	455/km2	30°20′13″N 81°39′42″W / 30.3369°N 81.6616°W...	27.338617
12	Fort Worth	Texas	962/km2	32°46′53″N 97°20′48″W / 32.7815°N 97.3467°W...	18.517559

Splitting the cooridnates to Latitudes and Longitudes for each city

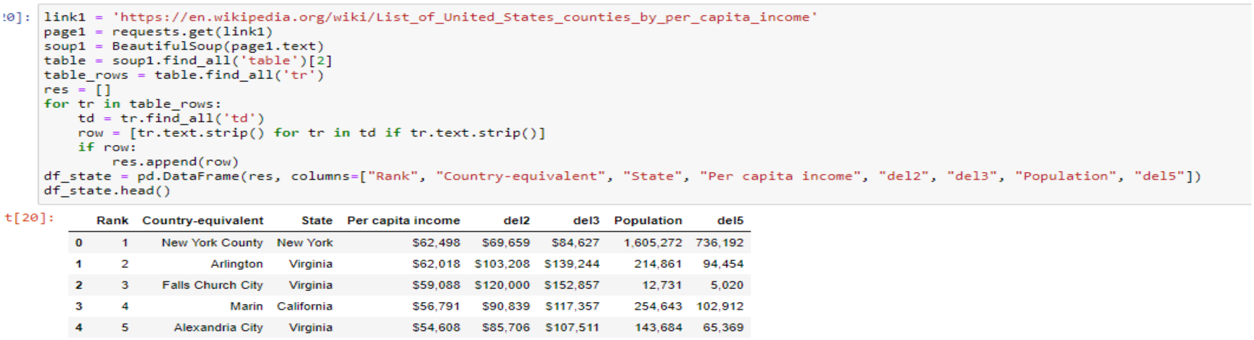


```
: #Splitting the location into latitudes and longitudes
  df["Location"]= df["Location"].str.split("/", n = 2, expand = True)[1]
  df.head()
```

7]:

	City	State	Population density in Km2	Location	Radius
0	New York[d]	New York	10,933/km2	40.6635°N 73.9387°W	17.363755
1	Los Angeles	California	3,276/km2	34.0194°N 118.4108°W	21.649480
2	Chicago	Illinois	4,600/km2	41.8376°N 87.6818°W	15.076472
3	Houston[3]	Texas	1,395/km2	29.7866°N 95.3909°W	25.248762
4	Phoenix	Arizona	1,200/km2	33.5722°N 112.0901°W	22.750824

Getting the per capita income state wise for USA

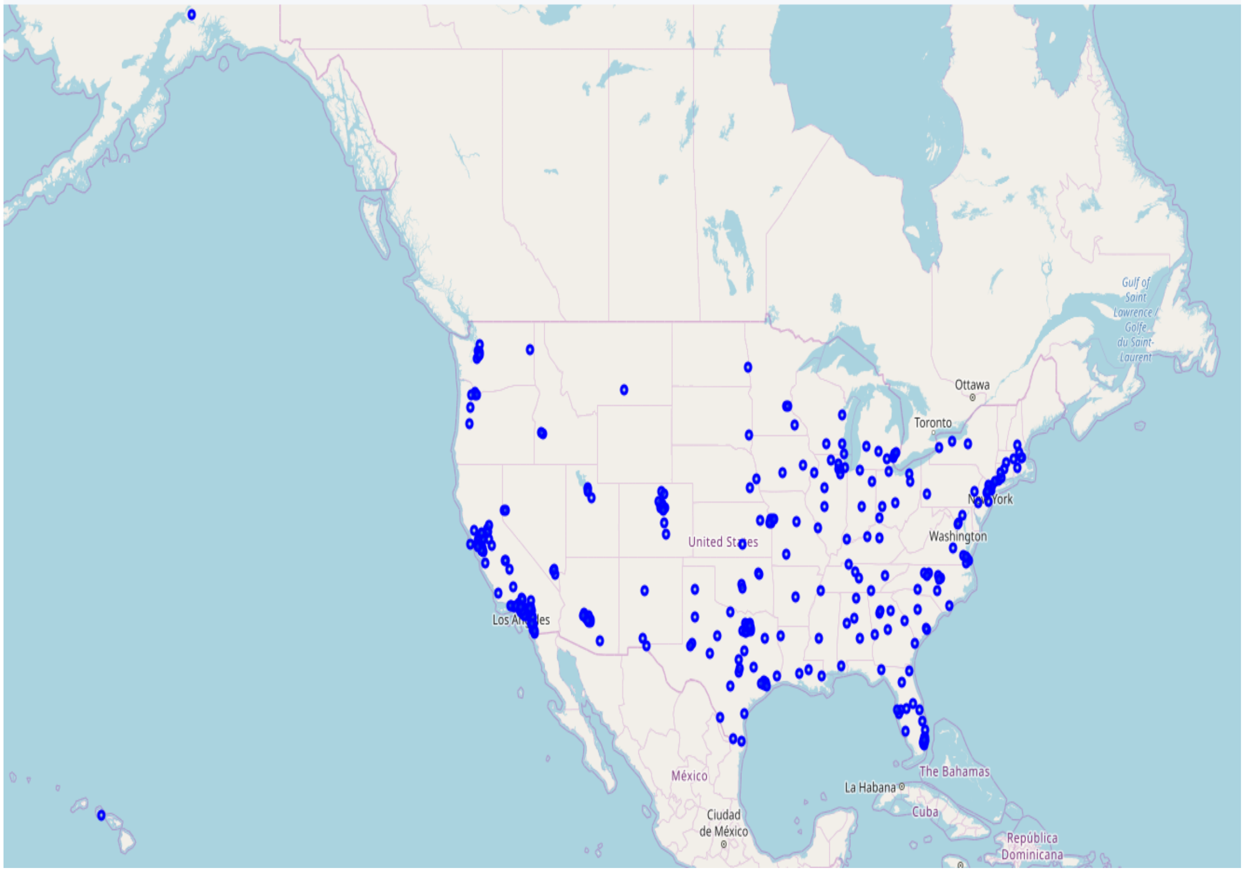


```
0]: link1 = 'https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income'
    page1 = requests.get(link1)
    soup1 = BeautifulSoup(page1.text)
    table = soup1.find_all('table')[2]
    table_rows = table.find_all('tr')
    res = []
    for tr in table_rows:
        td = tr.find_all('td')
        row = [tr.text.strip() for tr in td if tr.text.strip()]
        if row:
            res.append(row)
    df_state = pd.DataFrame(res, columns=["Rank", "Country-equivalent", "State", "Per capita income", "del2", "del3", "Population", "del5"])
    df_state.head()
```

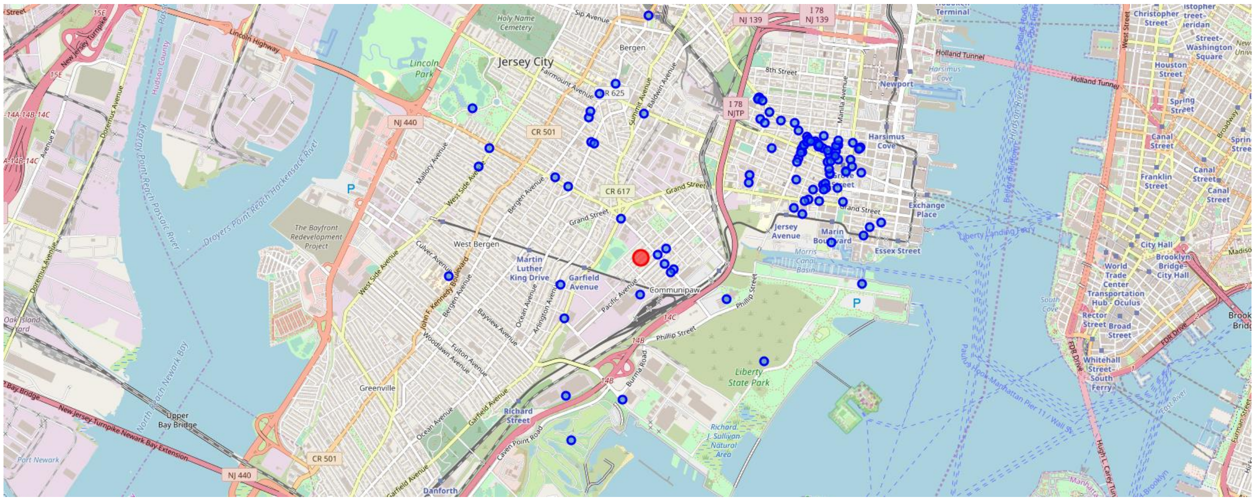
t[20]:

	Rank	Country-equivalent	State	Per capita income	del2	del3	Population	del5
0	1	New York County	New York	$62,498	$69,659	$84,627	1,605,272	736,192
1	2	Arlington	Virginia	$62,018	$103,208	$139,244	214,861	94,454
2	3	Falls Church City	Virginia	$59,088	$120,000	$152,857	12,731	5,020
3	4	Marin	California	$56,791	$90,839	$117,357	254,643	102,912
4	5	Alexandria City	Virginia	$54,608	$85,706	$107,511	143,684	65,369

Plotting all the cities of USA that we have extracted from Wiki page, using their coordinates



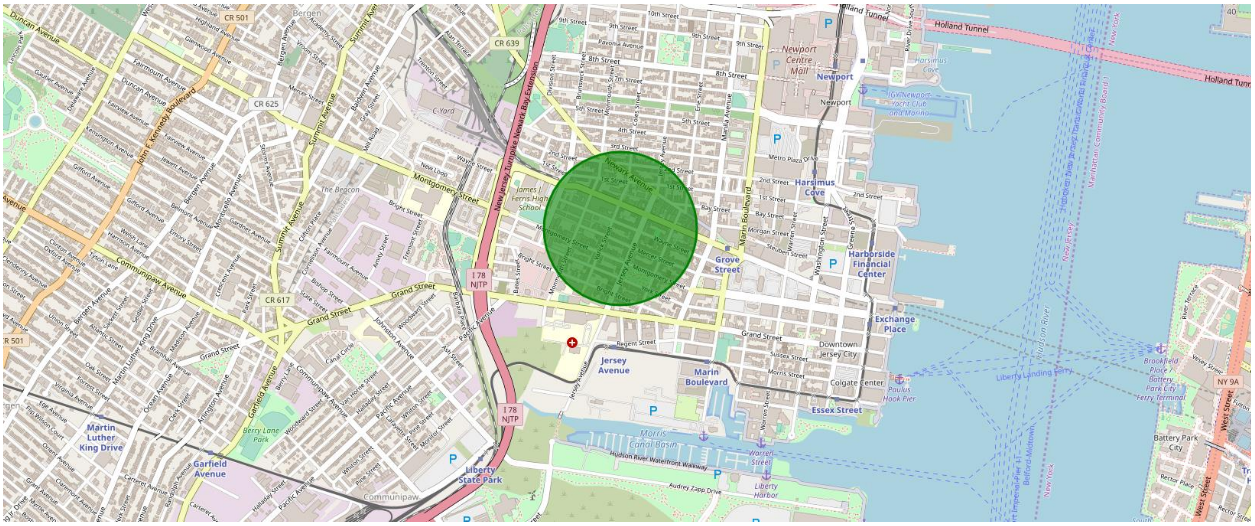
Plotting all the venues that we have got from the Four Square API



Result/Conclusion:

Based on the analysis we have done, the following is the result we have got:

Plot to show the location of final arcade that we are suggesting:



We have arrived at the suggestion of better spot in the best city of USA Based on the dependent factors and events 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, customer data similar to other businesses such as Gyms, health and wellness centers, Saloons etc.

Resources :

- List of all the cities in United States with population density and coordinates: https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population
- List of all the cities in United States with Per Capita Income : https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income
- Four Square API : https://foursquare.com