Coursera Capstone Project

The Battle of Neighborhoods

Guy Ravet

COURSERA CAPSTONE

– INTRODUCTION SECTION :

Discussion of the business problem and interested audience.

– DATA SECTION:

Description of the data that will be used to solve the problem and the sources.

– METHODOLOGY SECTION

Discussion and description of exploratory data analysis carried out.

– RESULTS SECTION

Discussion of the results.

– DISCUSSION SECTION

Elaboration and discussion on any observations noted and any recommendations suggested based on the results.

– CONCLUSION SECTION

Report Conclusion.

1. INTRODUCTION SECTION

DESCRIPTION OF THE PROBLEM AND BACKGROUND

SCENARIO:

I'm a data scientist in Switzerland. I recently got promoted but will need to move to New York City. I don't really know the city so I need a way to find out which neighborhood would be best for me. I have a comfortable life in Switzerland and would like to find the same, or better, in this gorgeous city.

Here are my requests :

- 2 or 3 bedrooms apartment
- Near a metro station in the Manhattan area and within 1.0 mile (1.6 km) radius
- Rent below $7,000 per month
- Ammenities in the selected neighborhood shall be similar to my current situation
- Would like to have venues such as coffee shops, asian food restaurants, wine stores, gym club and food shops

BUSINESS PROBLEM:

The challenge is to find a suitable apartment for rent in Manhattan that complies with the demands on location, price and venues. The data required to resolve this challenge is described in the following section 2.

INTERESTED AUDIENCE

I believe this is a relevant challenge with valid questions for anyone moving to any other large city. The same methodology can be applied in accordance to demands as applicable. It definetely a good exercise to develop Data Science skills.

2. DATA SECTION

DATA REQUIRED TO RESOLVE THE PROBLEM

In order to make a good choice of AN apartment in Manhattan NY, the following data is required: List/Information on neighborhoods form Manhattan with their Geodata (latitude and longitude) List/Information about the subway metro stations in Manhattan with geodata. Listed apartments for rent in Manhattan area with descriptions (how many beds, price, location, address) Venues and ammenities in the Manhattan neighborhoods

The list of Manhattan neighborhoods has been done in this course before. A csv file was created which will be read in order to create a dataframe and its mapping.

A list of Manhattan subway metro stops was complied in Numbers and it was complemeted with wikipedia data (https://en.wikipedia.org/wiki/List_of_New_York_City_Subway_stations_in_Manhattan) and information from NY Transit authority and Google maps (https://www.google.com/maps/search/manhattan+subway+metro+stations/@40.7837297,-74.1033043,11z/data=!3m1!4b1) for a final consolidated list of subway stops names and their address

A list of places for rent was collected by web-browsing real estate companies in Manhattan : http://www.rentmanhattan.com/index.cfm?page=search&state=results https://www.nestpick.com/search?city=new-york&page=1&order=relevance&district=manhattan&gclid=CjwKCAiAjNjgBRAgEiwAGLlf2hkP3A-cPxjZYkURqQEswQK2jKQEpv_MvKcrIhRWRzNkc_r-fGi0lxoCA7cQAvD_BwE&type=apartment&display=list https://www.realtor.com/apartments/Manhattan_NY A csv file was compiled with the rental place that indicated: areas of Manhattan, address, number of beds, area and monthly rental price. The csv file « nnnn.csv » had the following below structure. An algorythm was used to create all the geodata using Nominatim, as shown in section 3.0. The actual algorythm coding may be shown in 'markdown' mode becasues it takes time to run. With the use of geolocator = Nominatim() , it was possible to determine the latitude and longiude for the subway metro locations as well as for the geodata for each rental place listed. The loop algorythms used are shown in the execution of data in section 3.0 « Great_circle » function from geolocator was used to calculate distances between two points , as in the case to calculate average rent price for units around each subway station and at 1.6 km radius. Foursquare is used to find the avenues at Manhattan neighborhoods in general and a cluster is created to later be able to search for the venues depending of the location shown.

HOW THE DATA WILL BE USED TO SOLVE THE PROBLEM

The data will be used as follows: Use Foursquare and geopy data to map top 10 venues for all Manhattan neighborhoods and clustered in groups
Use foursquare and geopy data to map the location of subway metro stations , separately and on top of the above clustered map in order to be able to identify the venues and ammenities near each metro station, or explore each subway location separately
Use Foursquare and geopy data to map the location of rental places, in some form, linked to the subway locations.
Create a map that indicates, for instance, the average rental price per square ft, around a radious of 1.0 mile (1.6 km) around each subway station – or a similar metrics. I will be able to quickly point to the popups to know the relative price per subway area.
Addresses from rental locations will be converted to geodata(lat, long) using Geopy-distance and Nominatim.
Data will be searched in open data sources if available, from real estate sites if open to reading, libraries or other government agencies such as Metro New York MTA, etc.

MAPPING OF DATA

The following maps were created to facilitate the analysis and the choice of the palace to live. Manhattan map of Neighborhoods manhattan subway metro locations Manhattan map of places for rent Manhattan map of clustered venues and neighborhoods Combined maps of Manhattan rent places with subway locations Combined maps of Manhattan rent places with subway locations and venues clusters

3. METHODOLOGY SECTION

This section represents the main component of the report where the data is gathered, prepared for analysis. The tools described are used here and the Notebook cells indicates the execution of steps.

THE ANALYSIS AND THE STRAGEGY:

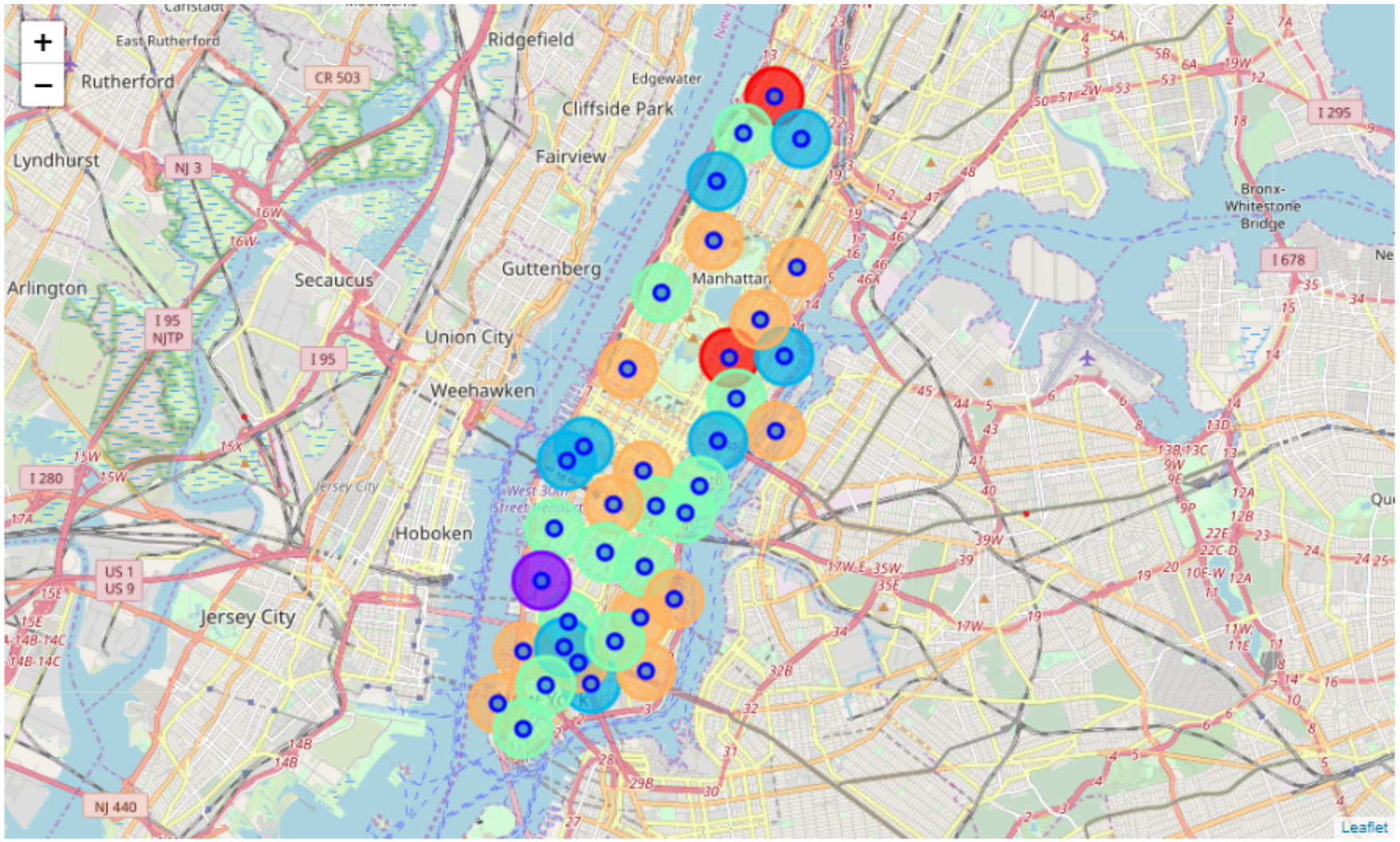
The strategy is based on mapping the above described data in section 2.0, in order to facilitate the choice of at least two candidate places for rent. The choice is made based on the demands imposed : location near a subway, rental price and similar venues to Singapore. This visual approach and maps with popups labels allow quick identification of location, price and feature, thus making the selection very easy.

The procesing of these DATA and its mapping will allow to answer the key questions to make a decision:

- what is the cost of available rental places that meet the demands?
- what is the cost of rent around a mile radius from each subway metro station?
- what is the area of Manhattan with best rental pricing that meets criteria established?
- What is the distance from work place (Park Ave and 53 rd St) and the tentative future rental home?
- What are the venues of the two best places to live? How the prices compare?
- How venues distribute among Manhattan neighborhoods and around metro stations?
- Are there tradeoffs between size and price and location?
- Any other interesting statistical data findings of the real estate and overall data.

4. EXECUTION AND RESULTS

MANHATTAN MAPS – NEIGHBORHOODS AND CLUSTER OF VENUES



MANHATTAN APARTMENTS FOR RENT – GEODATA

In [76]:
```
# csv files with rental places with basic data but still wihtout geodata ( latitude and longitude)
# pd.read_csv(' le.csv', header=None, nrows=5)
mh_rent= df_data_2
mh_rent.head()
```

Out[76]:

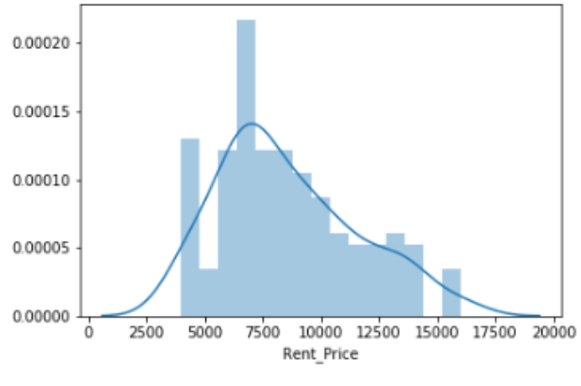
	Address	Area	Price_per_ft2	Rooms	Area-ft2	Rent_Price	Lat	Long
0	West 105th Street	Upper West Side	2.94	5.0	3400	10000	NaN	NaN
1	East 97th Street	Upper East Side	3.57	3.0	2100	7500	NaN	NaN
2	West 105th Street	Upper West Side	1.89	4.0	2800	5300	NaN	NaN
3	CARMINE ST.	West Village	3.03	2.0	1650	5000	NaN	NaN
4	171 W 23RD ST.	Chelsea	3.45	2.0	1450	5000	NaN	NaN

In [77]:
```
mh_rent.tail()
```

Out[77]:

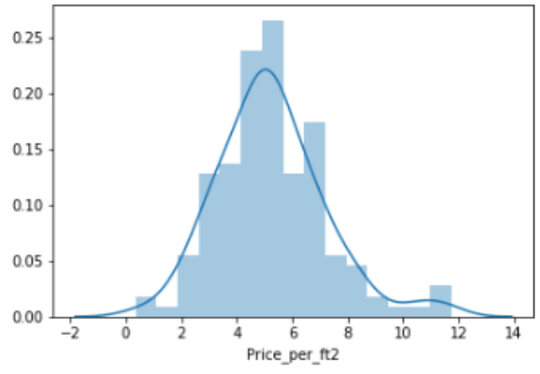
	Address	Area	Price_per_ft2	Rooms	Area-ft2	Rent_Price	Lat	Long
139	200 East 72nd Street	Rental in Lenox Hill	5.15	3.0	1700	8750	NaN	NaN
140	50 Murray Street	No fee rental in Tribeca	7.11	2.0	1223	8700	NaN	NaN
141	300 East 56th Street	No fee rental in Midtown East	3.87	3.0	2100	8118	NaN	NaN
142	1930 Broadway	No fee rental in Central Park West	5.06	2.0	1600	8095	NaN	NaN
143	33 West 9th Street	Rental in Greenwich Village	6.67	2.0	1500	10000	NaN	NaN

MANHATTAN RENTAL PRICES STATISTICS



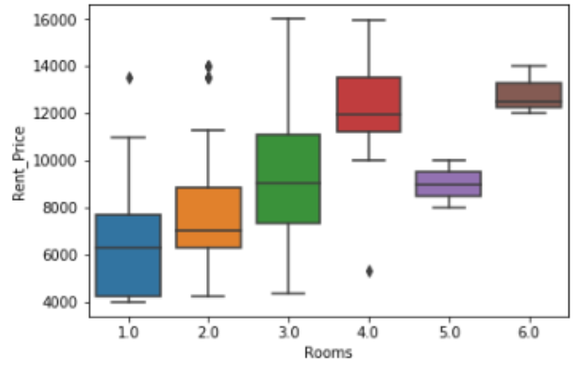
In [82]:
```
import seaborn as sns
sns.distplot(mh_rent['Price_per_ft2'],bins=15)
```

Out[82]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff9c6678ba8>

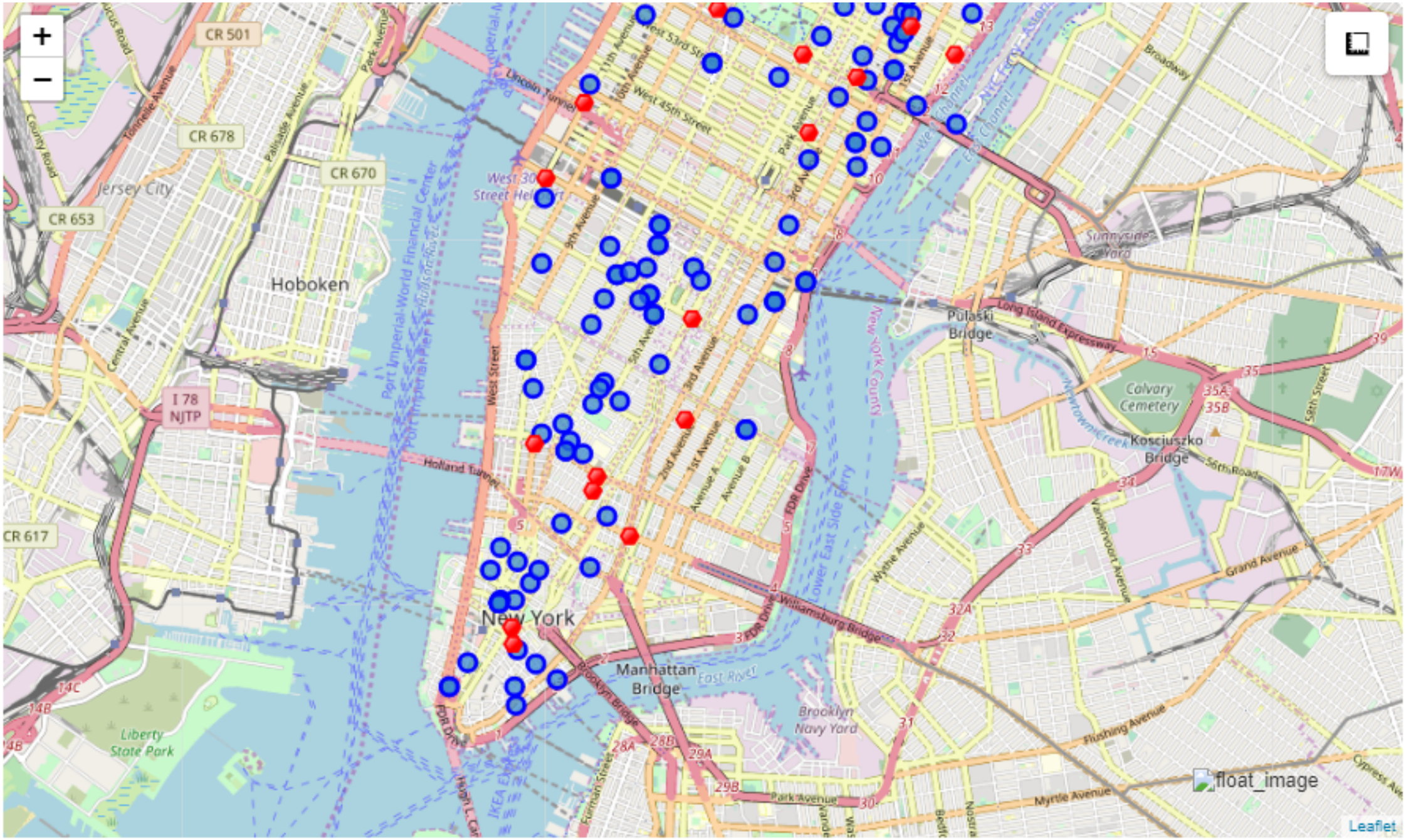


In [83]:
```
sns.boxplot(x='Rooms', y= 'Rent_Price', data=mh_rent)
```

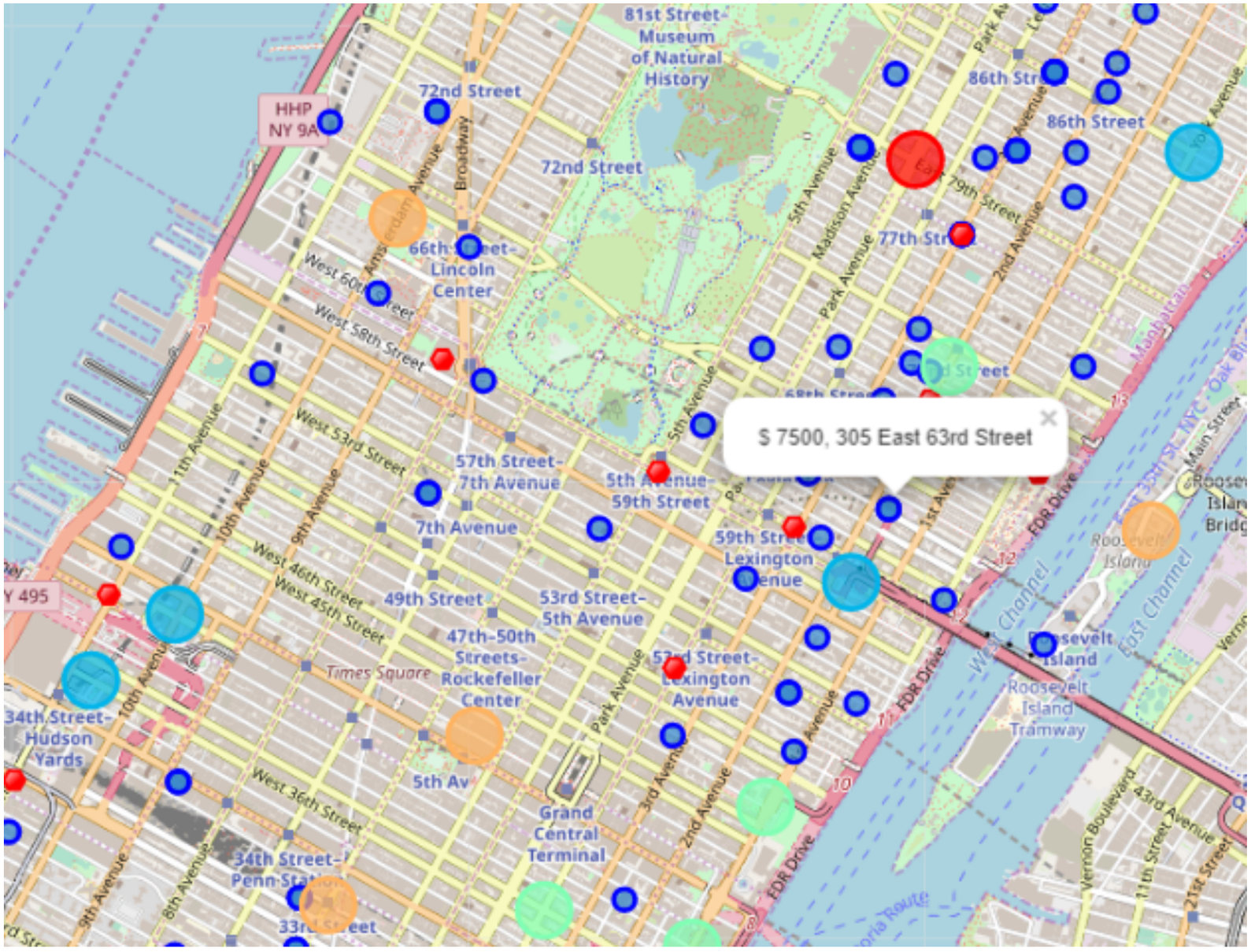
Out[83]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff9c65fe3c8>



MANHATTAN APARTMENT FOR RENT (BLUE) AND SUBWAY STATIONS (RED)



SELECTED APARTMENT



APARTMENT SELECTION

Using the « one map » above, I was able to explore all possibilities since the popups provide the information needed for a good decision.

Apartment 1 rent cost is US7500 slightly above the budget. Apt 1 is located 400 meters from subway station at 59th Street and work place (Park Ave and 53rd) is another 600 meters way. I can walk to work place and use subway for other places aroung. Venues for this apt are as of Cluster 2 and it is located in a fine district in the East side of Manhattan.

Apartment 2 rent cost is US6935, just under budget. Apt 2 is located 60 meters from subway station at Fulton Street, but I will have to ride the subway daily to work , possibly 40-60 min ride. Venues for this apt are as of Cluster 3.¶

I feel that Cluster 2 type of venues is a better option. That means that APARTMENT 1 is a better choice since the extra monthly rent is worth the conveniences it provides.

5. DISCUSSION

This project gave me a good opportunity to apply and exercice all of the skills I learned in the past two months while taking this course.

6. CONCLUSIONS

I’m actually impressed of all the things you can do when you know how to manipulate data. What is amazing is the amount of data available open source. It’s a very powerful ressource that will most likely become more and more valuable in the future.