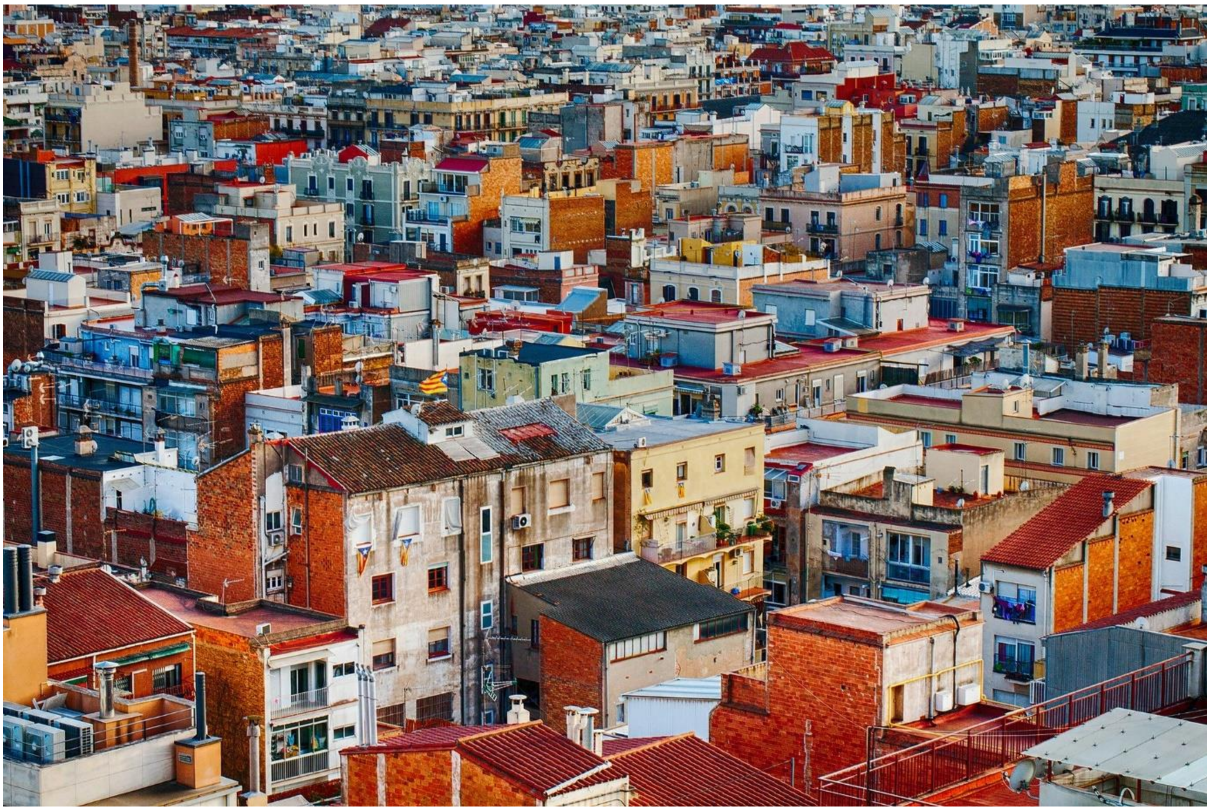
Coursera Capstone project

IBM Applied Data Science Capstone

Know the locality of your new house

By: Sumit Kumar

March, 2020



1.Introduction:

Many of the time we have to move from one city to another city, so we look for house to get on rent in the new city on some housing websites. The price of the house is not the only factor people consider before taking house on rent. We want to know the locality more and more. We want to live in an area where we get all the essential things close to our house we want stores places like grocery shop, gym or Pharmacy store around us, and some people also wants to have some restaurant or playground or park around the house they looking to take on rent.

Now problem with most of the housing website is that it shows how good or bad the house is but there is a very less information about the locality of the house, how is the locality, if there all the essentials shops and stores are present in the area or not. To make a better decision to buy a new house we have to look on many other factors like how expensive the house is in compare to the other house in the area or are the houses in the other area is cheaper with similar facilities and similar locality.

1.1 Business Problem:

In this project we are going to analyse a city on basis of price of houses in different area. We will also analyse price of houses on the basis of type of houses. We will analyse and visualise if there is all the needed stores, shops or playground in the area so that a person can make a better decision to rent a house.

2. Data:

- The data used in this is New York City Airbnb Open Data housing dataset. The data is cleaned and filled the missing values in the dataset.
  The source of the Data is https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
- This data is used to analysis the price of the house in different area of the New York City and know the price of different type of houses like single room of multiple room.
- I have also used this data to select a sample vacant apartment to do our analysis.
- To know the details of locality of a house, I have used Foursquare API. The data comes in .json format which I have cleaned and reduced it to get only the details of venue.

3. Methodology:

3.1 Data Cleaning:

There were lots of missing data due to some lack to records keeping. I have cleaned the data by filling the missing values. I am targeting economical class people, so I only took the data of house whose price is below 200 dollars per month.

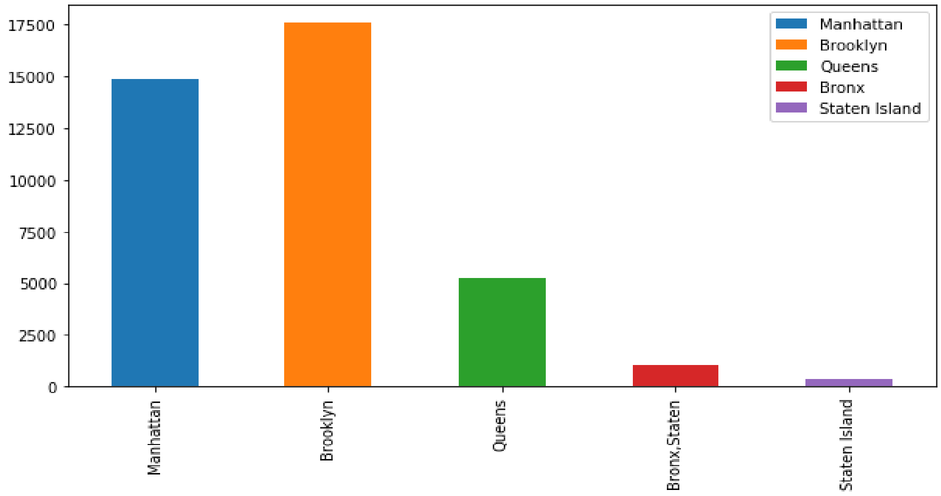
3.2 Feature Selection:

For analysing the data, the features we have used from the data are price of the rooms, latitude, longitude, neighbourhood group and room type. I have used latitude, longitude and price to get to know the price distributions on different areas of the New York State. Used neighbourhood group and price to analysis the price of houses in different borough. Also used room type and price to analysis the price of different types of room in the State. Also I have analysed the number of available houses of different types and availability of houses in different borough of the State. I have not dropped any feature as we have to use it as sample available house.

3.3 Exploratory Data Analysis:

3.3.1Number of vacant apartments in different districts of New York:

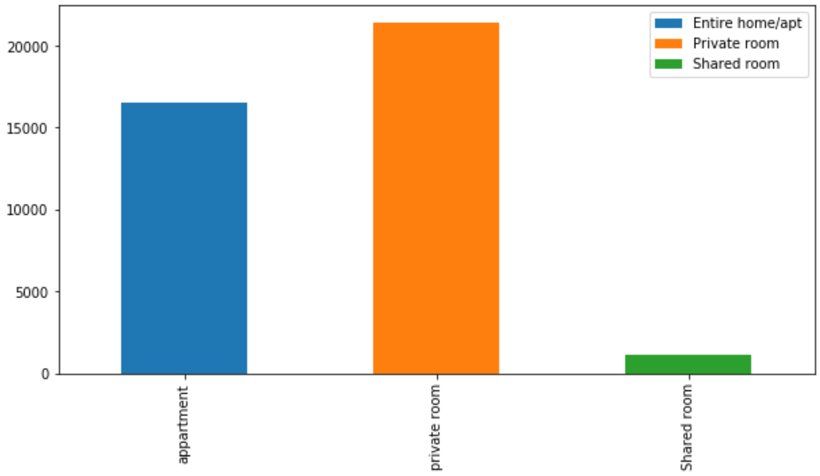
There are five borough of New York, Manhattan, Brooklyn, Queens, Staten Island and The Bronx. I have visualised the number of empty apartments on bar graph.



As the graph says most numbers of empty houses are in Brooklyn followed by Manhattan and others. Brooklyn and Manhattan combined has most empty houses. But as we know these two places Manhattan and Brooklyn are known for its expensive living, so we will analyse the cost of apartment in different area of the state.

3.3.2Type of vacant houses and their numbers:

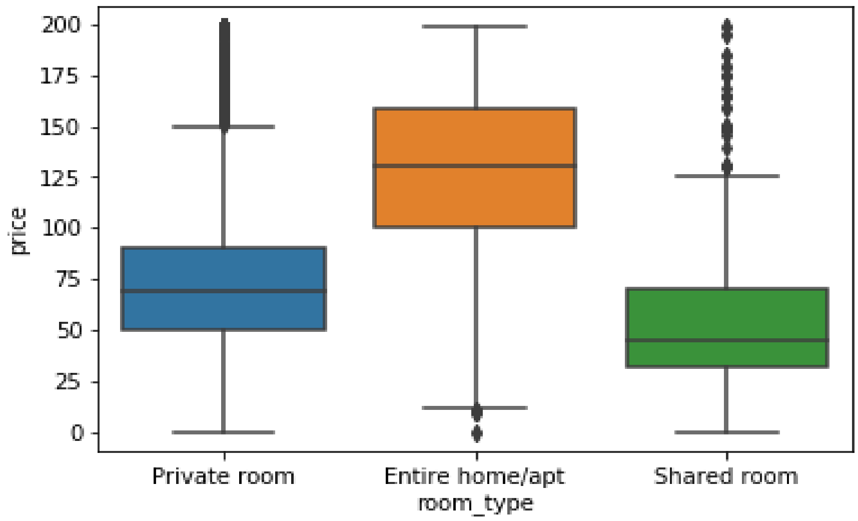
There could be different type of houses available on rent like whole apartment, private room or shared room. I have visualised the number of different types of room in the state on a bar graph



As we can see in the graph there is very few numbers of shared rooms which is considers to be least expensive if someone is living in some expensive city. Private rooms are available the most.

3.3.3 Relation between the price and different types of rooms:

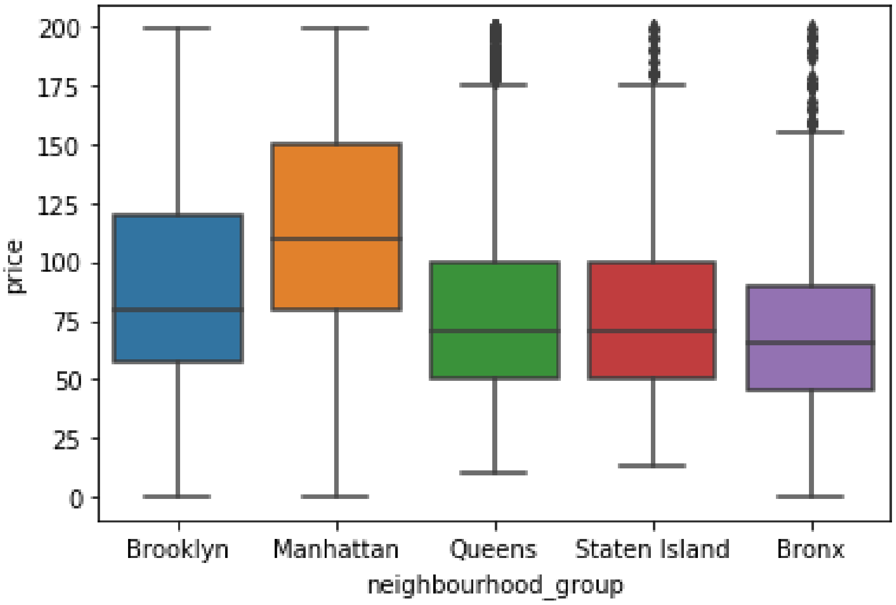
To make a better decision on what kind of house one should rent on the basics of price I have plotted a box plot between price and different types of apartment. The different types of houses are



Well as per the graph Entire home is most expensive with average cost of 130 dollars, the private room cost 70 dollars on average, the shared room cost average of 40 dollars but many of the shared rooms also cost 70 dollars which is average cost of private rooms so one can consider taking private room instead of shared room.

3.3.4 Relation between price of houses in different boroughs:

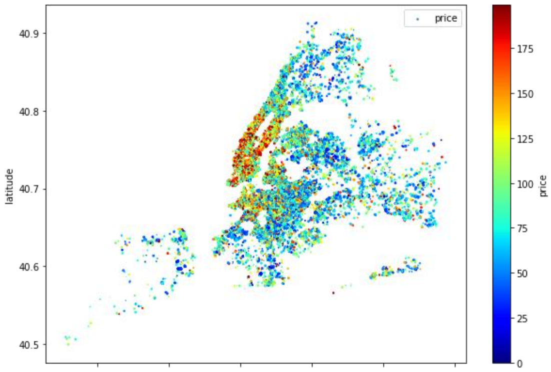
The price may vary in different borough of New York, to find out I have done analysis by plotting a box graph between price and different borough.



Manhattan has highest average cost of rent which is around 110 dollars per month, the average cost in other cities doesn't vary much so we don't have to look much into it.

3.3.4 The cost of house in different area:

To know the cost of houses for rent in different part of the city I have plotted price of houses with respect to longitude and latitude of the house.



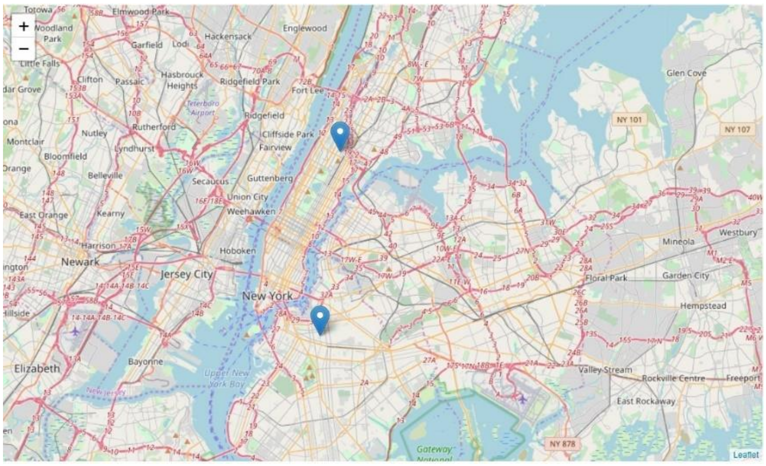
New York City
N
Bronx
Manhattan
LaGuardia Airport
Queens
Newark Liberty International Airport
Brooklyn
John F. Kennedy International Airport
Staten Island

The figure in the left hand side is a graph which shows the distribution of price of houses according to the longitude and latitude and figure in the right and side shows the New York city we can see that the most expensive houses are in Manhattan and the area close to Manhattan in different borough, also it shows that the price around the coast is more expensive than places away from the coast.

3.3.5 Sample Available House data:

Now we need to know the vacant house available in the city. Now as we already have data houses of New York city, we can take data of houses from it, so the sample data of house to that we are considering in this project is from the New York Airbnb dataset only. To make analysis easy I have only chosen data of two houses which we I will compare and select which house is better to choose to take on rent.

I have used folium library in python to visualize details of the locality of house.



As given in the figure I have taken two houses as sample from which a person who is looking to take house on rent will choose from. To make a person choose from the vacant house we will analyse the places around each of the two available houses and let the user know its locality apart from its price and location so that the user can make a better decision.

3.3.6 Getting data of locality:

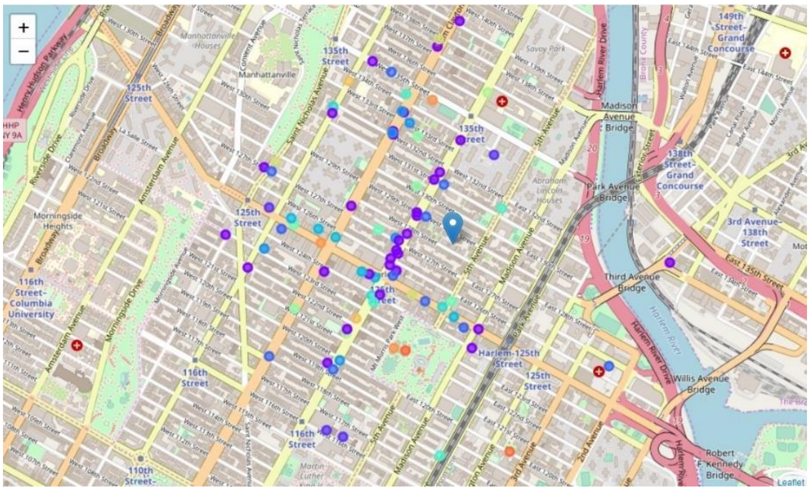
To know the data of its locality I used FourSquare and got the details of the venues around the house. To fetch the data I have used Foursquare API and request Foursquare to get the details of a place by constructing URL.

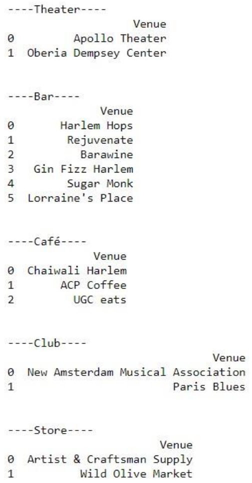
3.3.7 Machine Learning model used:

As I have to classify this data into different types of venues like restaurant, grocery shops, café etc we need a model which can classify data into different groups. We first need to get the data of places around a location for which I have used FourSquare API . Then I need to classify those places into different groups, to classify the places I have chosen KNN model to process this. After processing the data through KNN model it got the venue classified into 15 different groups. This classification will be visualised on map.

3.3.7(a) Details of the first house:

I have classified the extracted places in the area into 15 different categories like café shops, parks etc. Then this data is visualised on the map. The venues around the first empty data is classified and shown on the map and its name on text.

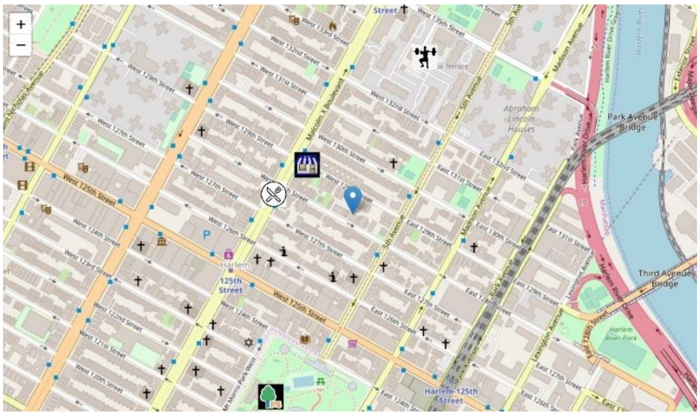


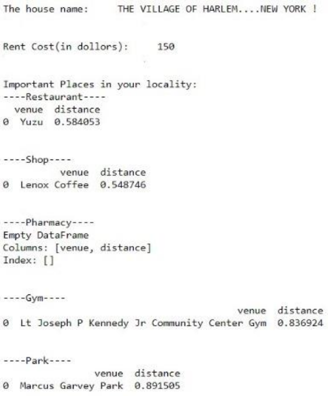


```
----Theater----
                   Venue
0         Apollo Theater
1  Oberia Dempsey Center


----Bar----
              Venue
0       Harlem Hops
1        Rejuvenate
2          Barawine
3   Gin Fizz Harlem
4        Sugar Monk
5  Lorraine's Place


----Café----
             Venue
0  Chaiwali Harlem
1       ACP Coffee
2         UGC eats


----Club----
                                  Venue
0  New Amsterdam Musical Association
1                        Paris Blues


----Store----
                        Venue
0  Artist & Craftsman Supply
1           Wild Olive Market
```

All the places around our first location is listed as shown in the right side of the map.

I have looked for the important places in the locality that are most important to have around like gym, playground, shops etc and also calculated their distance from the house and visualised the closest one of each category on the map.



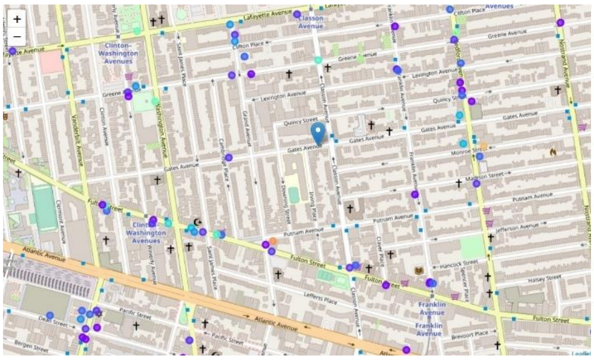


```
The house name:     THE VILLAGE OF HARLEM....NEW YORK !


Rent Cost(in dollors):     150


Important Places in your locality:
----Restaurant----
  venue  distance
0  Yuzu  0.584053


----Shop----
         venue  distance
0  Lenox Coffee  0.548746


----Pharmacy----
Empty DataFrame
Columns: [venue, distance]
Index: []


----Gym----
                                           venue  distance
0  Lt Joseph P Kennedy Jr Community Center Gym  0.836924


----Park----
                venue  distance
0  Marcus Garvey Park  0.891505
```

So the important and closest places are visualised on the map and the final details are given in the list which has details of the important places in the area listed in category with its distance from the house, it also shows the name and cost of the house.

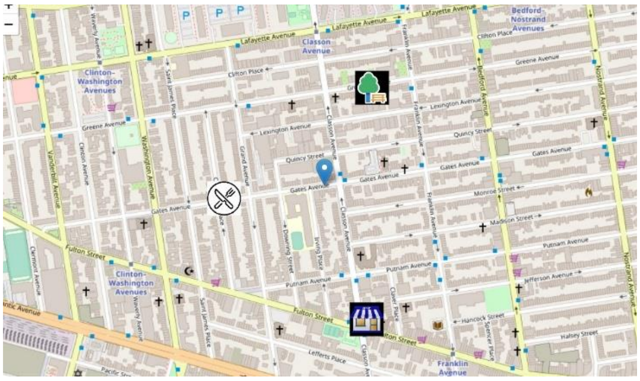
3.3.7(b) Details second house:

Similar to the first house I have used FourSquare to extract data of the places around the second house and then classified them using KNN model and visualised it on a map and on a list.



```
----Restaurant----
                     Venue
0  Peaches Shrimp & Crab
1                  The Fly
2                The Finch
3                    Golda
4                   Hart's
5       Pilar Cuban Eatery
6      Joloff Restaurant 2


----Bakery----
           Venue
0  The Good Batch
1   Beny's Delice


----Pub----
Empty DataFrame
Columns: [Venue]
Index: []


----Shop----
                          Venue
0            Always and Forever
1      Leon & Son Wine & Spirits
2              The Mixtape Shop
3            Daily Press Coffee
4  Stonefruit Espresso + Kitchen
5          Playground Coffee Shop
6            Brooklyn Kolache Co.
```

Then I have looked for the important places around the house find distance of each one of them and visualised the closest one of them on the map.



```
The house name:     Cozy Entire Floor of Brownstone


Rent Cost(in dollors):     89


Important Places in your locality:
----Restaurant----
                 venue  distance
0  Locanda Vini & Olii  0.646749


----Shop----
                venue  distance
0  Always and Forever  0.739786


----Gym----
Empty DataFrame
Columns: [venue, distance]
Index: []


----Garden----
                          venue  distance
0  Jane Bailey Memorial Garden  0.685217
```

The important places are visualized on the map and the details are listed along with its distance form house and price of the house. One thing which can be noted here is that there is no gym around this house.

4. Discussion:

In this project I have visualised the important venues on the map and distance between them. The major goal of this project was to categories and visualised extracted data of places around the houses on map to reduce effort of user to do more research about the locality of the houses on internet also the thoughts that comes into mind of a person before renting a house is how would be the society and will I get to settle there well.

One thing which I found surprising was there were no shop of Pharmacy in one kilometre of either of the house we choose as sample house. Pharmacy shop is one of the important shops one wanted to have around their house. There could be possibility that there is very low number of pharmacy stores in New York City. Another thing is that there are lots of restaurant in New York sometimes there are more than fifteen restaurants within one kilometre of area.

4.1 Limitations:

There are plenty of limitations of this research

- There are many features that is required to make a decision on which house to select to take on rent but this project focus on only few of such features
- We have used KNN model to categorise the venues but we have dropped those data which has no category which means if there is avenue which is very unique in the area it would be used in any process.

- I have only analysed data of houses whose price is less than 200 dollars per month but some people may consider living in a high standard house with high expense.

5. Conclusion:

The major goal of this project was to extract information of locality of a house which is supposed be listed on a housing website and let the customers or users know the locality of the house they are considering to rent on. Then the customer take decision on what do they really want in their locality based on information that has been shown. For ease of customer I have visualised the important venues on the map, the venues are displayed as icon of its category.

The result of this is totally depends on the person who is suppose to rent the house. But I can have my opinion on the result, the first house we have more number essential places in the locality like restaurant, gym etc but I the second house there is no gym in the area. If we compare the cost of rent of first house it is 150 dollars bit second house just cost 79 dollars which is a pretty good deal. I am not a gym freak so I would prefer to take the second house as it cost me less and it has almost all venues I need to have around a house other than a gym which is not required for me much.

5.1 Future Directions:

I am using very limited data extracted from FourSquare. There is possibility that the closest venue we are choosing to display on the map has very poor review or it is not visited frequently, there could be another place of same category which is visited more and has a better review, so we can extract more information like photos of venues, customer review and rating all the most frequently visited places and then we can also implement it using an algorithm which consider all other factors. Then the best results came out of this can be displayed on the map.