**Coursera Capstone**

**IBM Applied Data Science Capstone**

***Locating good Residential Building having better Indian restaurant option in the neighborhood in Washington DC***

By: Sarika Rajeev

May 2019

# **Introduction/Business Problem**

The objective of this capstone project is to come up with a real problem and select a location for analysis. We need to use location data from Foursquare.com. We require to solve the problem using data science methodology and machine learning techniques.

In our problem statement, we have a group of software engineers from India who are planning to live in Washington DC for three months. Their hiring company is paying all expenditure. So, they would need to find flats/apartments having a good rating. They are going to work outside India for the first time. So, it's desirable that they are located nearby an Indian restaurant (having a great number of likes) preferred by others.

**The target audience for this project includes:**

1. Any professional/outsourcing employees who are going to relocate for short-term in Washington DC.
2. Real-estate agents who are interested in locating the attractive neighborhood in Washington DC.
3. Anyone who has an interest in location-based Data Science Project.

The code in python with data analysis is available on my GitHub page:

<https://github.com/sarikarajeev33/capstone-final-project/blob/master/capstone_final_project.ipynb>

**Data Section**

1. For the data collection part, we have used Wikipedia to find out Washington DC neighborhood names and link for the cities:

<https://en.wikipedia.org/wiki/Neighborhoods_in_Washington,_D.C.>

We have collected the data and after observation, we have neglected/dropped the unwanted data. With the help of link of the cities, we have found the Coordinates (longitude and latitude) of the cities in the neighborhood of Washington DC.



Fig. 1. Using Wikipedia for neighborhoods of Washington DC

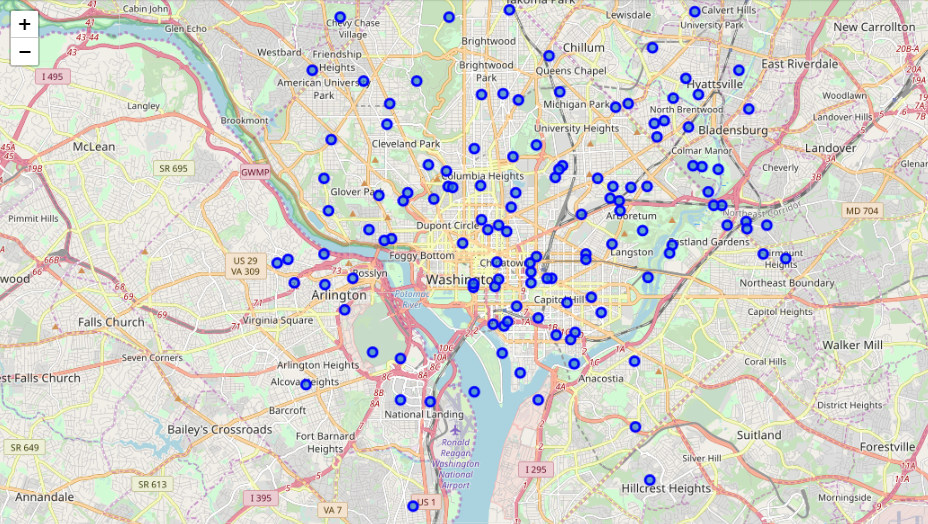


Fig. 2. Map of adding marker (neighborhoods) on the map of Washington DC

1. To find the **Residential Building (Apartment / Condo)** with good rating we have utilized FourSquare API, namely the `explore the venues` request with a corresponding category-Id of ` 4d954b06a243a5684965b473.



Fig. 3. Screenshot of utilizing the FourSquare API for finding homes

1. To find the **Indian Restaurant** with more number of likes (preferred by others), we have used FourSquare API, namely the `explore the venues` request with a corresponding category-Id of ` 4bf58dd8d48988d10f941735.



Fig. 4. Screenshot of utilizing the FourSquare API for finding nearby resturents

# **Methodology**

After the data is collected we need to understand data using approaches such as univariate statistics, and histogram. We have evaluated the mean and standard deviation of the collected data to find that whether data collected is useful or there is redundancy of data for example: same neighborhoods could have multiple option of homes and restaurants. We have used “groupby” command to get rid of redundancy of data. So, after we understand the data we need to prepare the data using the models such as feature engineering. We have used google map for data visualization of neighborhoods of Washington dc.

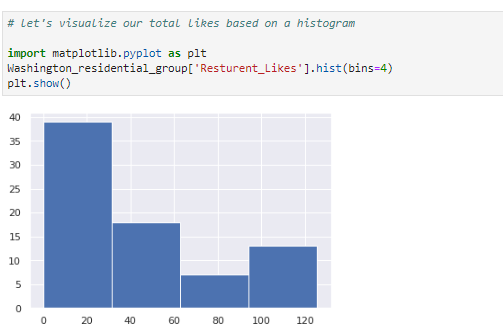


Fig. 5. Data visualization using histogram for no. of likes of restaurants in the neighborhoods of Washington DC

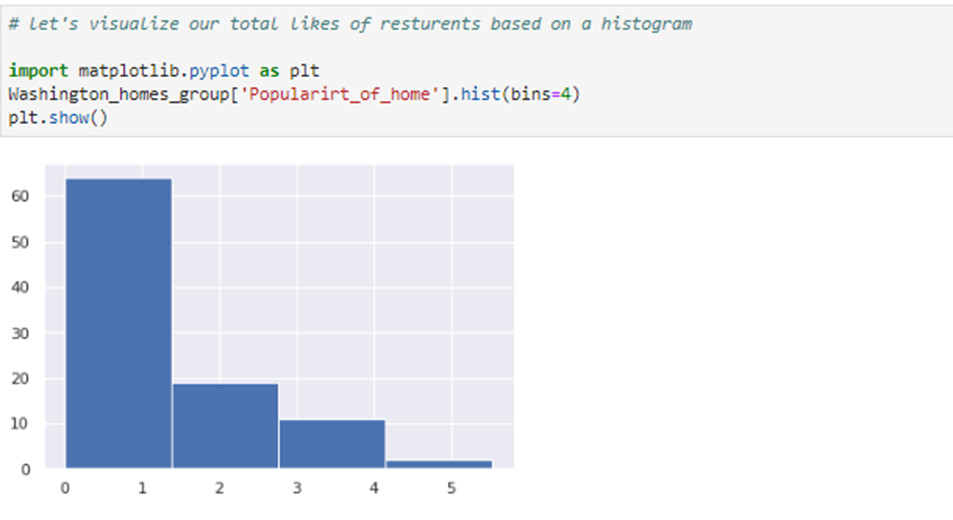


Fig. 6. Data visualization using histogram for popularity of homes in the neighborhoods of Washington DC

We have utilized the API of foursquare.com to obtain the number of likes of restaurant and homes. After data is prepared we can model the data using machine learning models. We have chosen the K-means clustering approach. We have divided the data set into four clusters using k-means clustering model. We have also created a cluster map to show the four cluster of neighborhoods.

# **Result**

I have used K-means clustering algorithm, to solve the problem of favorable accommodation of software professionals. I could generate four clusters of neighborhoods of Washington DC. These are as follows:

**Cluster 1:**

* Somewhat Popular homes
* Mostly poor or average restaurant

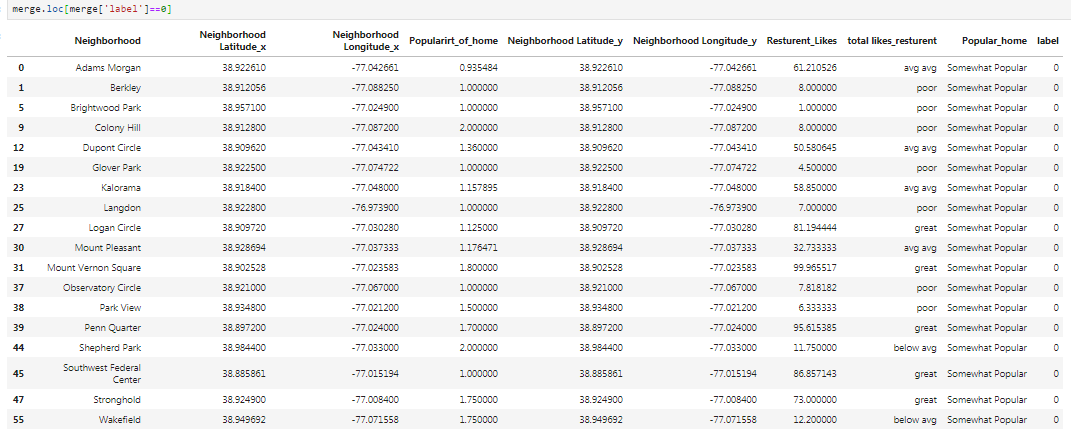


Fig. 7. Screenshot of Cluster 1.

**Cluster 2:**

* . Most Popular Home
* . Great and above avg. restaurants

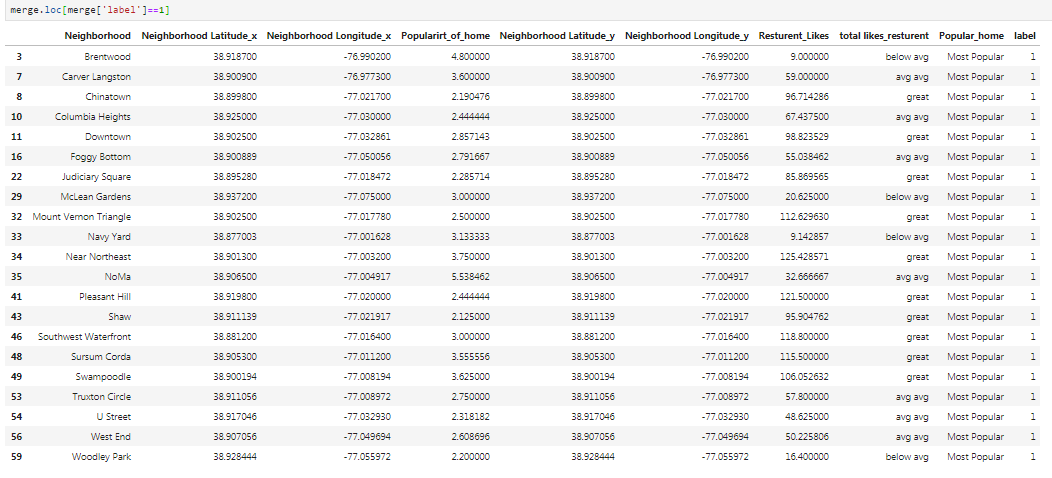


Fig. 8. Screenshot of Cluster 2

.

**Cluster 3:**

* . Least Popular Home
* . Poor and below avg. restaurants

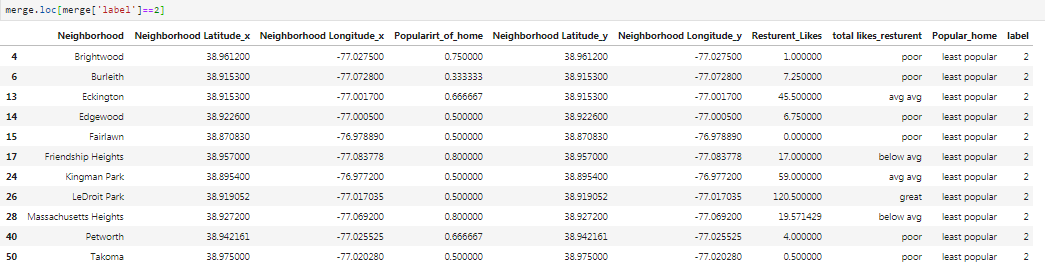


Fig. 9. Screenshot of Cluster 3

**Cluster 4:**

* . unpopular Home
* . Average restaurants

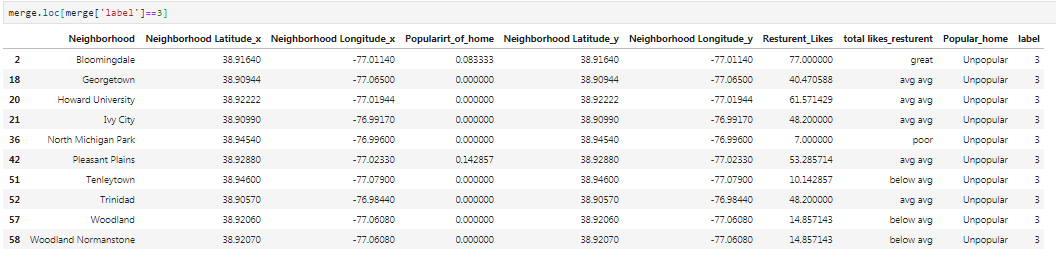


Fig. 10. Screenshot of Cluster 4

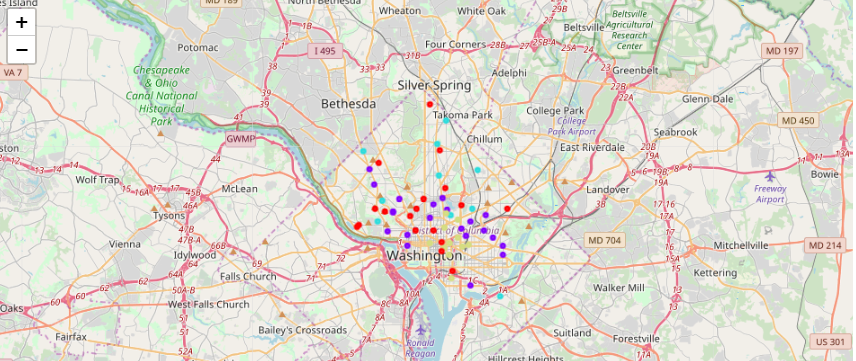


Fig. Cluster Map for neighborhoods of Washington DC

# **Discussion**

For the problem of exploring favorable residence for a group of software professional who are coming to Washington Dc for 3 months, I have utilized API of foursquare.com to obtain location data. I have also used Wikipedia of neighborhoods of Washington DC to find the neighborhoods of Washington DC. I have used likes as a proxy for quality of restaurant. The more likes imply, the better the restaurant is. I have used “likes” for the popularity of homes. Initially I have tried to get the rating for good home using API, but from free subscription of foursquare.com, I was only able to get rating of few homes. Then I have switched to the “likes” criteria for evaluating the popularity of homes.

I have used K-means approach to divide the data set into four cluster. I have got four cluster with different categories (options) of homes and restaurants. I have also used cluster-maps for data visualization. So, the clients will have a clear image of all four options of clusters and the tradeoff is evidently visible.

# **Conclusion**

Although it is hard to find balance between good housing and fine restaurant. But we have provided our consumers/end users four clusters of neighborhoods of Washington DC to meet their desired criteria. This will make it easier to understand them the trade-off for each offer.