Project Title:

Exploring the Business Prospect Opportunity in Washington DC

Washington District of Columbia Neighborhood Exploration, Segmentation and Clustering using unsupervised machine learning K-Means!

Table of Contents

Introduction/Business Problem

Washington, DC, the U.S. capital, is a compact city on the Potomac River, bordering the states of Maryland and Virginia. It's defined by imposing neoclassical monuments and buildings – including the iconic ones that house the federal government's 3 branches: The Capitol, White House and Supreme Court. It's also home to iconic museums and performing-arts venues such as the Kennedy Center. It also has US Border Patrol and U.S. Customs & Border Protection.

In this project, I am working on if someone is looking to open a cafe/coffee shop, what should they open and where should they open it? Since its the capital of United States and a spot for government staff for meetings and tourists whether they come to see museums, arts center, white house or for business meetings. Due to extreme busy and fast pace atmosphere in District of Columbia, there is always a need for cafe/coffee shops to meet the need for thousands of visitors along with its locals every day.

Audience

Any business person who wants to open a cafe/coffee shop/restaurant.

Data Section

Data Sources

I downloaded the data from

'https://opendata.dc.gov/datasets/neighborhood-labels/data'

I imported all the libraries and packages (numpy, panda, types, geocoder, matplotlib, folium, K-means, lxml, html5lib, bs4) required for python coding in the Jupyter Notebook.

Data Cleaning

After downloading data, I received the dataframe with an index and 8 columns; X, Y, OBJECTID, GIS_ID, NAME, WEB_URL, LABEL_NAME, DATELASTMODIFIED. Data wrangling was performed to drop unnecessary columns. Since we are exploring, segmenting and clustering neighborhood so we only needed neighborhood name, their latitude and longitude value and I used drop function to remove all the other columns mentioned above. Then I renamed the columns 'X' and 'Y' as 'Latitude' and 'Longitude' and 'name' as 'Neighborhood'. I proceeded with rearrange the columns in their order as "Neighborhood", "Latitude" and "Longitude" to perform analysis. I checked the dataframe data, its types using dtype function and its shape using shape(). Followed upon that I looked for statistical values of the dataframe using 'info' function with the original dataframe.

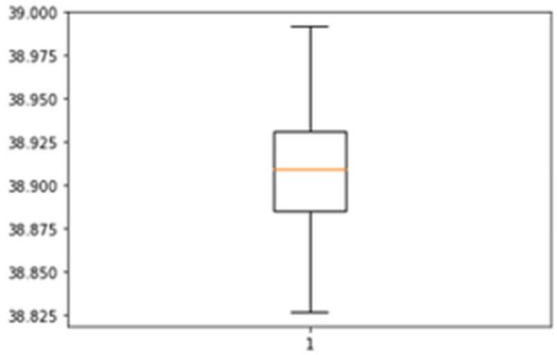
Total of 132 neighborhoods showed up with their Latitude and Longitude after all the pre-processing. Geographical coordinates were collected for the city to visualize map using Folium. Foursquare API is being used to further processing with segmentation. In the Methodology section I would share how I worked on segmentation (Explore Neighborhoods, Analyze Neighborhoods, Examine Neighborhoods) and clustering (Cluster Neighborhood).

Methodology Section

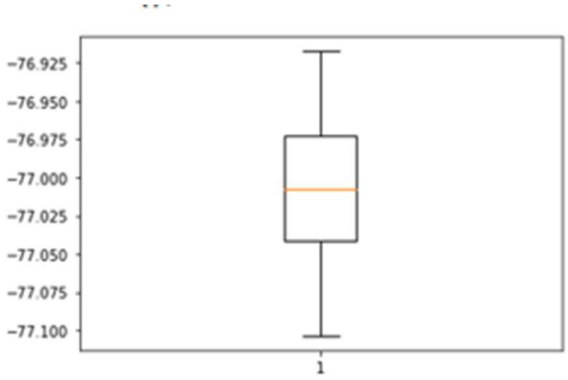
Exploratory Data Analysis

I explored the data using box plot, histogram, and scatter plot. Box Plot was being used to concentrate distribution of the Latitude and Longitude of all the Neighborhood. These data visualizations indicate that both the north-south distribution and east-west distribution of the neighborhood is approximately normal. The scatter plot indicates that the neighborhoods are evenly spaced across the diamond shape Washington DC area.

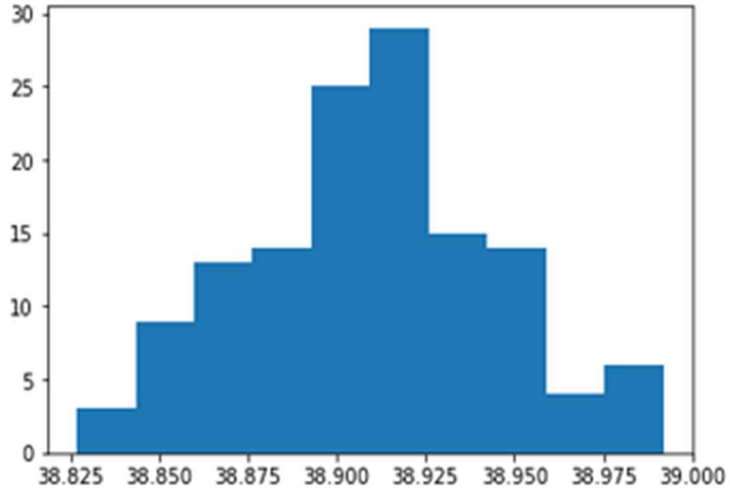
Box plot of the Latitude



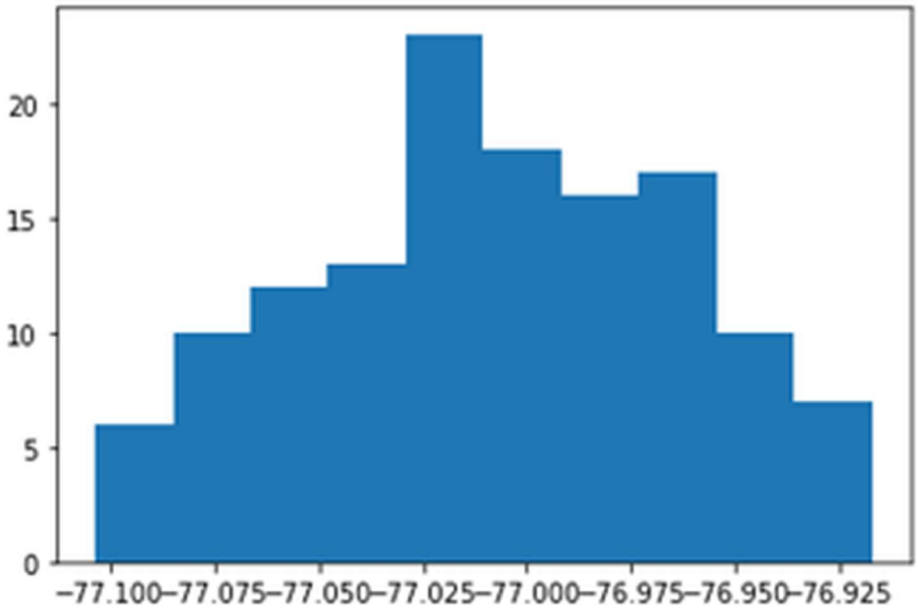
Box plot of the Longitude



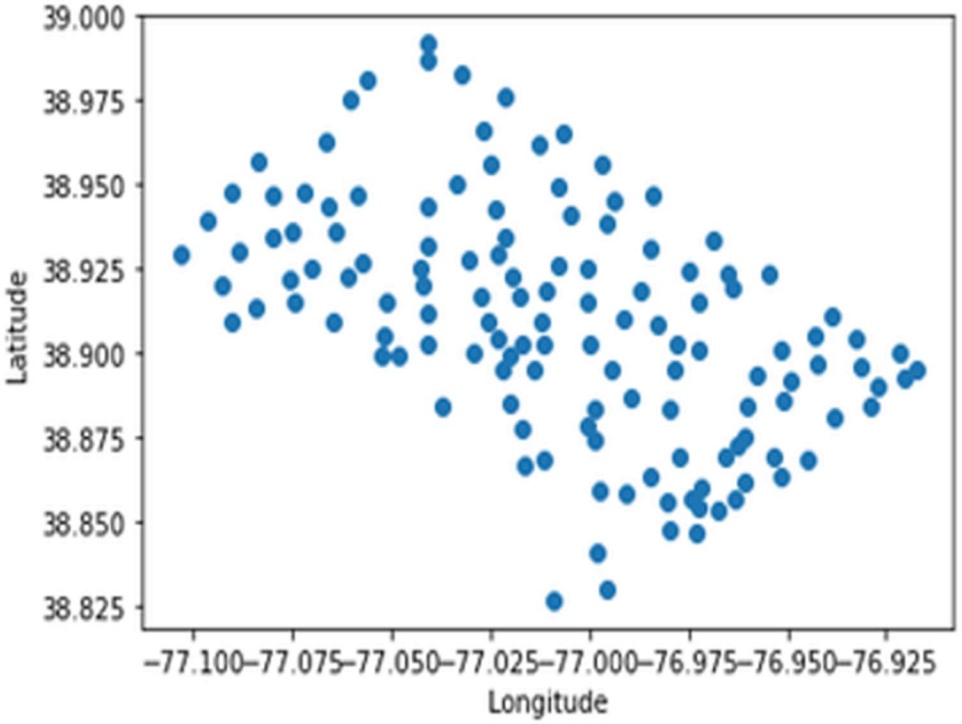
Histogram of Latitude



Histogram of Longitude

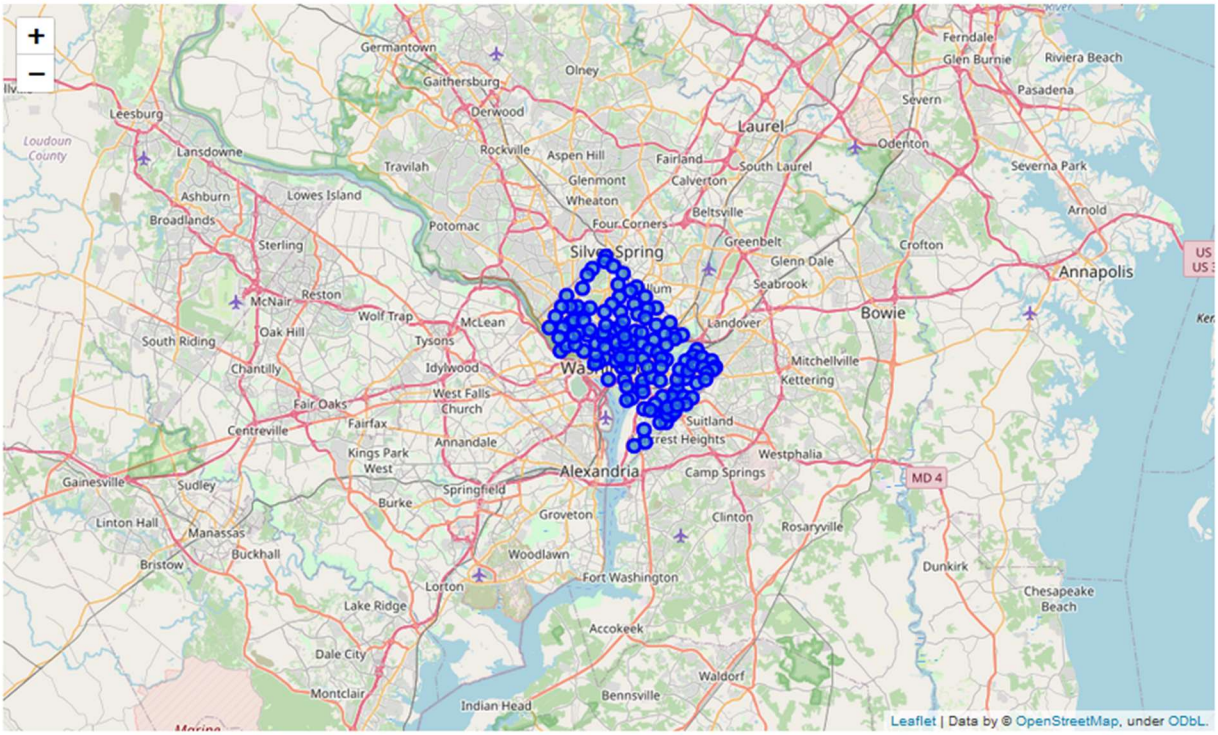


Scatter Plot of Latitude and Longitude



Map of Washington DC using latitude and longitude values

```
map_Washington_DC
```



Predictive Modeling

Regression, classification and clustering are among the possible models that could be used to analyze neighborhoods. Since we are trying to find a pattern in the data I chose to use unsupervised machine learning method of K-Means Clustering to determine the best place to open a cafe.

Unsupervised Machine Learning Method K-Means was used for Segmentation and Clustering

I used Geocoder to get the geographical coordinate of the center of Washington DC to establish the center of folium map. I used Folium to generate the map of Washington DC. I added blue markers to the map based upon the neighborhood coordinates in my original dataset. I looked for missing values and NAN values in the dataframe to proceed further with segmentation.

I created an account with Foursquare API so that I can used credential (CLIENT_ID and CLIENT_SECRET) to proceed with segmenting the Neighborhoods. Since the very first Neighborhood came as 'Fort Stanton', I looked for 100 venues that are in Fort Stanton within a radius of 500 meters. I used GET API of Four Square to obtain results. We created new dataframe with get_category_type and structure it into a pandas dataframe. We created a new dataframe nearby venues, GetNearbyVenue using Foursquare Credentials along with VERSION, latitude, longitude, radius, LIMIT. We further append Venue Categories to the existing dataframe. Further I created an array washingtonDC_venues and printed number of venues returned by Foursquare. Checked the size of it. Further I grouped the venues as per individual neighborhood.

I transformed the neighborhood data using one hot encoding. Further, I grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category by One Hot Encoding. With the help of One Hot Coding and its grouping for I printed each neighborhood along with the top 5 most common venues

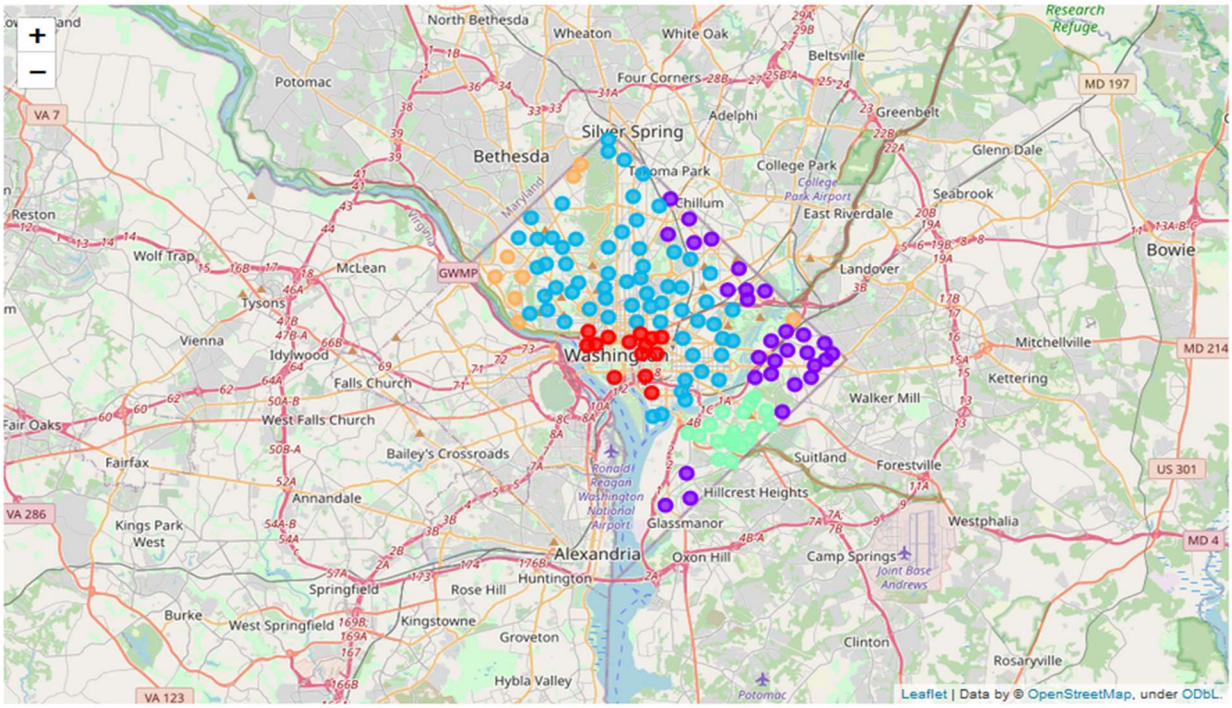
Finally, I used K-Means to cluster the Neighborhoods. I created a new dataframe Washington DC sorted that includes the cluster as well as the top 10 venues for each name. Then I merged the Original dataframe with new sorted dataframe under a new dataframe called Washington Merged. My output from the above merging finally displayed all neighborhoods with their long and latitude value along with cluster labels and 10 most common venues in the table. I used Folium to generate the folium map using cluster results. I used rainbow colors along with an opacity of 0.7.

Results

Finally, I examined Clusters. Through Elbow method I used K equals 5. I found out neighborhoods in Cluster 4 which is the 5th cluster (as the number began from 0 to 4) is best to open a new cafe coffee shop since there are less number of cafe coffee shops in those areas despite a spot for visitors and students which are showing through top 10 common venues trails, college stadium, bus stops, picnic areas, scenic lookouts showing in that cluster.

Cluster Map

```
map_clusters
```



Cluster 0 has 14 neighborhoods and there are multiple cuisines eatery American Restaurant, Italian Restaurant, French Restaurant, Cafe/Coffee Shops, Mediterranean Restaurant, Sandwich Places, Pizza Place Vegetarian Restaurant. So this is not at all good place for any food business.

Cluster 1 has 29 neighborhoods and this is showing output for Convenience Store, Park, Gas Station, Baseball Field along with lots of Cafe/Coffee shop/Breakfast Spots, Fast Food Centers and Sandwich Place. This cluster neighborhood does not seem good idea for opening a business.

Cluster 2 is the biggest since it has 69 neighborhoods. The top most common venue here are Bars. It seems like Rich Neighborhood Clusters. It has all the Venues which one could

imagine. All kind of eatery such as Ethiopian Restaurant, fast food corners, Bars, Liquor stores, Brewery, Park, Trail, Gym, Deli, Bakery, Mediterranean Restaurant, Mexican Restaurant, Pizza Place, Thai Restaurant, French Restaurant, Greek Restaurant, Burger Joint, Sandwich Place, Breakfast Spot, Sushi Restaurant, Vegetarian / Vegan Restaurant. Definitely a very bad choice to open any eatery or food business here.

Cluster 3 has 18 neighborhoods. It is more common for Banks. Fair amount of eatery such as Sandwich Shops, Cafe Coffee shops, Breakfast Spot, available to meet locals needs. Seems like residential area. Enough businesses are present in this neighborhood.

Cluster 4 has only 8 neighborhoods. As per my analysis it has very less number of Coffee Shops and lots of trails and parks. It would be ideal place to open cafe/coffee shop around parks and trails.

Discussion

While examining clusters I found out cluster 2 would be the worst choice to open any cafe coffee shop/restaurants or any kind of eatery. In Cluster 1 all neighborhoods have more than 2 food business apart from a neighborhood named "Congress Heights". So if we proceed to further segment each neighborhood in individual cluster someone can open a good eatery business here. It required thorough study and analysis further. Cluster 0, 1 and 2 have the busiest neighborhoods with less personal transport convenience there's good prospect for any contractor who wants to run rickshaw/auto-rickshaw business there. Since my project focused on opening a Cafe/Coffee shop so my choice of cluster is different in above Results Section.

Conclusion

This data analysis is important for any business-man who wants to open a food business in Washington DC without being failure about choosing the wrong place and losing millions of dollars. Money matters! Still there's a lot to learn about python libraries/packages which can be make many data scientist/data analyst work easy.

References

All the python codes and libraries that were used in this Project are as per Applied Data Science Capstone Course from IBM Data Science Professional Course. I would like to thank Alex Aklson for explaining unsupervised K-Means clustering in an easy way. I would also like to thank other faculties Romeo Keinzler, Joseph Santarcangelo, Nikolay Manchev, Lakshmi Holla and Saishruthi Swaminathan who helped learners to get clear understanding of Data Science Concepts. This report may subject to change if needed at any point of time. Any mistake in this project belongs to me.