USING MACHINE LEARNING TO FIND LOCATIONS TO OPEN A MEXICAN RESTAURANT

In these Corona times we spend time in different ways. I thought of learning something new. I searched for online courses to understand the fundamentals of data science and how to practically use various tools such as IBM Watson Studio, Python and the libraries, such as pandas, folium, numpy, geocoder, Scikit learn etc. I plan to work on data analysis and program machine learning applications. I would like to absolutely recommend this online course, the IBM Data Science Professional Certificate.

The course has a final assignment in which many of the tools and methods learned throughout the recent months are applied in a self-chosen challenge around the general idea. I am analyzing the probability of opening a new Mexican restaurant in Phoenix. As part of this, I will be creating a Jupyter notebook with Python as the programming language, with comments and this presentation as a final report. I am assuming a scenario of opening a reliable Mexican restaurant in Phoenix. The most important and crucial decision behind this project is to find an optimal location in Phoenix for a great opportunity to open the Mexican restaurant.

PHOENIX

Phoenix is the anchor of the Phoenix metropolitan area, also known as the Valley of the Sun. Phoenix has long, extremely hot summers and short, mild winters. The metropolitan area is the 11th largest by population in the United States

Business Problem

The objective of this project is to find the most suitable location to open a New Mexican restaurant in Phoenix. By using data analysis, and machine learning algorithms like clustering, this project aims to provide solutions to answer to the business problem.

Data

To solve the problem, we need the following data:

- List of Neighborhoods in the city of Phoenix
- Latitude and Longitude of this Neighborhood
- Venue data related to Mexican restaurants

Extracting the data

- Scrapping of Phoenix Neighborhood data via Wikipedia
- Getting the Latitude and Longitude of this neighborhood using the Geocoder package
- Using Foursquare API to get the venues related to this neighborhood

Methodology

Importing the important python packages and libraries used in this project.

Pandas: For creating and manipulating data frames.
Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
Geocoder: To retrieve Location Data.
Scikit Learn: For importing k-means clustering.
silhouette_score: For finding the optimal value for K
JSON: Library to handle JSON files.

Beautiful Soup: Web scraping
Requests: library to handle http requests.
Matplotlib: Python Plotting Module.

Web Scraping

The first step I performed was to scrape data from the Wikipedia page

https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Phoenix,_Arizona that consisted of all the neighborhoods of Phoenix. I used Beautiful Soup to scrape the data. I converted it into a data frame since they are the best data structure to work with when it comes to analysis using visualization techniques. Next step I did was adding latitudes and longitudes of the neighborhood in the dataframe using geocoder.

Out[4]:

	Neighborhood	Latitude	Longitude
0	Arcadia (Phoenix)	33.502560	-111.973986
1	Biltmore Area	33.518471	-112.025525
2	Brentwood Historic District	38.919790	-76.986420
3	Central Avenue Corridor	33.557690	-112.074490
4	Chinatown, Phoenix	33.448250	-112.075800
5	Desert Ridge	33.634140	-111.930610
6	Downtown Phoenix	33.447992	-112.073579
7	Golden Gate Barrio	33.566061	-112.122063
8	Maryvale, Phoenix	33.492882	-112.174764
9	Moon Valley, Phoenix	33.619229	-112.084337
10	North/Northwest Phoenix	33.581205	-112.007520
11	Sacred Heart Church (Phoenix, Arizona)	33.434282	-112.056118
12	South Phoenix	33.406720	-112.071180
13	F. Q. Story Neighborhood Historic District	33.448250	-112.075800
14	Woodlea Historic District	33.448250	-112.075800

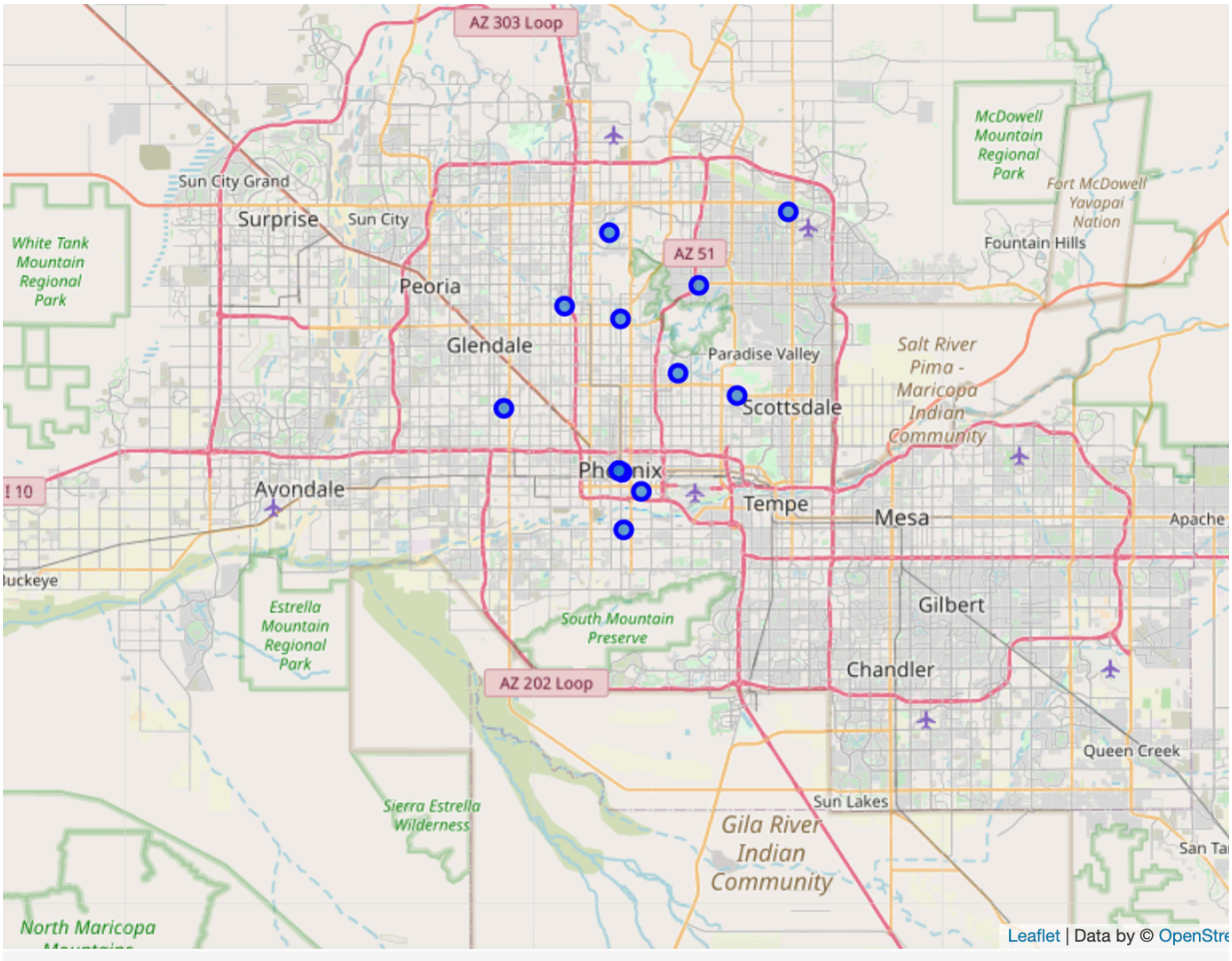
Caption

Data Cleansing

I analyzed the Data Frame and understood that while obtaining the latitudes and longitudes, one of the neighborhood has pulled a wrong latitude and longitude. Since it affects my data I dropped that neighborhood from my dataframe.

Data Visualization

I used folium to visualize the data. I created the map of Phoenix using latitude and longitude values with its neighborhoods marked.



Foursquare API

I have used foursquare Api to explore the neighborhood of Phoenix. “Foursquare” locational information is used to gain the information. Foursquare is a location data provider with information about all types of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. The foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

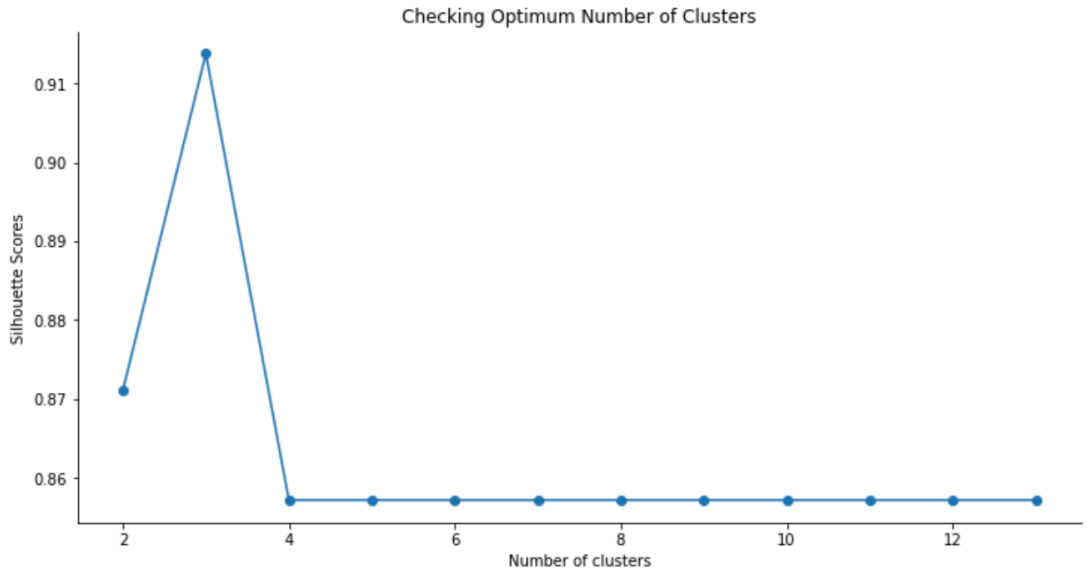
After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, I have chosen the radius to be 500 meter. To find clusters of restaurant types in the different neighborhood, I first transformed the data frame with the restaurant venues, associated to the neighborhood, by one-hot encoding (0/1). Next I grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arcadia (Phoenix)	33.502560	-111.973986	Shemer Art Center & Museum	33.505025	-111.971996	Art Gallery
1	Arcadia (Phoenix)	33.502560	-111.973986	Monsoon Driving	33.505769	-111.973916	Lake
2	Arcadia (Phoenix)	33.502560	-111.973986	Reed's Candle Light Observatory	33.499291	-111.977236	Planetarium
3	Biltmore Area	33.518471	-112.025525	Arizona Biltmore Golf Course	33.521478	-112.023028	Golf Course
4	Biltmore Area	33.518471	-112.025525	Wrigley Mansion	33.522777	-112.026997	American Restaurant

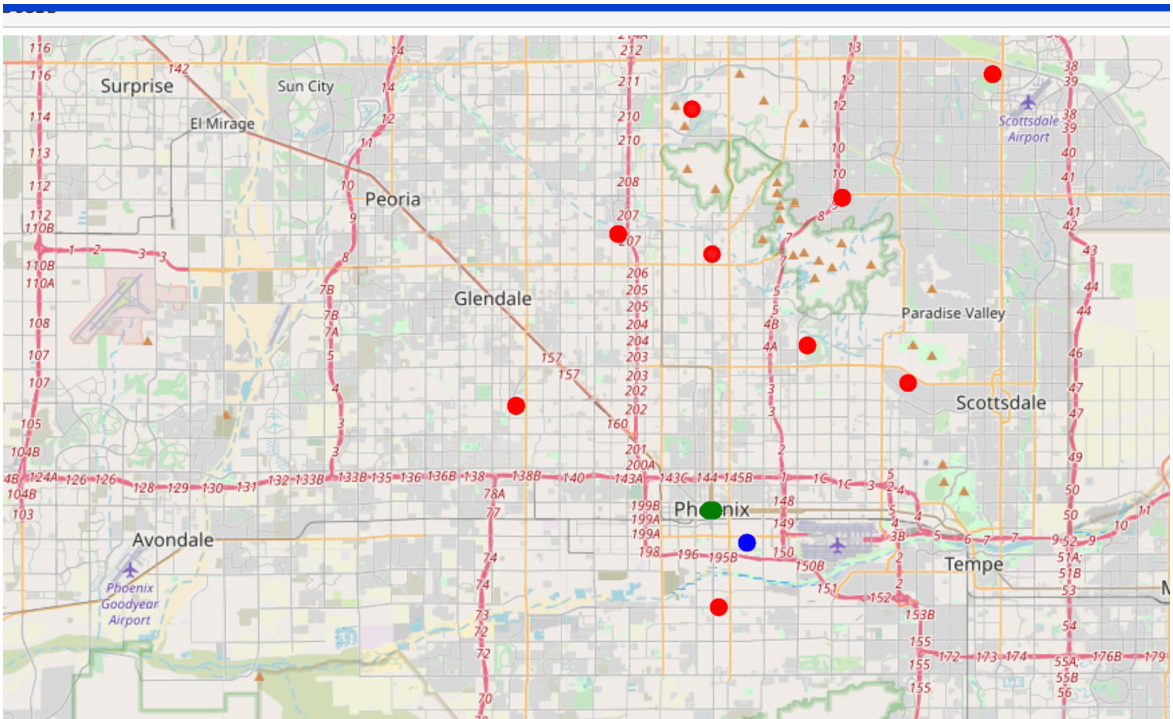
Machine Learning

Kmeans Clustering

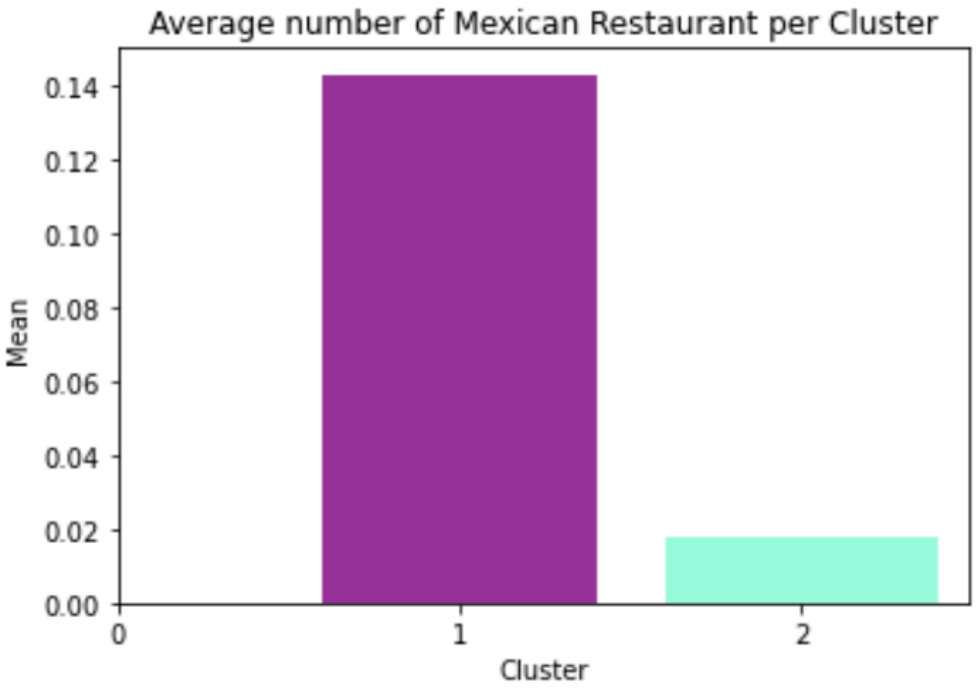
For finding the Optimum number of Clusters I used Silhouette Scores method.



I got the optimal value of K as 3. Then I ran unsupervised k-means clustering machine learning algorithm from the scikit-learn package. We can now use the cluster labels to show the neighborhoods marked with a cluster-specific color on a map, again using folium.



Next I filtered out the Mexican Restaurants from each cluster and plotted the graph for all the Neighborhood Cluster to see the Mexican Restaurant occurrence.



Cluster 1 has most of the Mexican Restaurant followed by Cluster 2.

Result

In the map given above, Cluster 0 (Red color) has no Mexican restaurants. Therefore, this project recommends the entrepreneur to open Mexican Restaurant in these neighborhoods with little to no competition.

Limitations

In this project, I took only one factor into consideration: the occurrence of Mexican restaurants in each neighborhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a new restaurant.

Conclusion

Finally to conclude this capstone project, I have got a glimpse of what a real life data-science project is. In this project, I have imported different types of python libraries and packages such as panda, numpy, matplotlib, Scikit learn, geocoder, Folium etc. I have used BeautifulSoup package to web scrape data. I have used Machine learning technique K-Means clustering to cluster the neighborhoods and predicted the location that is optimal for opening a Mexican Restaurant in Phoenix city, where the demand is high and profit will be maximum. This kind of data analysis provides me initial guidance and interest to analyze more real-life challenges using data-science.