IBM CAPSTONE PROJECT

DETERMINING WHICH NEIGHBORHOOD IN SEATTLE TO PLACE A BILLBOARD FOR A NEW ONLINE RESTAURANT COMPANY

Sandra Bamfo | IBM Data Science | July 2021

Introduction

IBM Data Science Professional Certificate course on Coursera is an opportunity for new learners of data science to learn the prerequisite skills for forging ahead in a career as a business analyst . In the last module of this course, students are required to complete a capstone project. This project is about using data science tools on a real-life problem and demonstrating the creation of value by applying the learned skills. In this post is a summary of my hypothetical project and my inferences drawn. All analysis was performed in Python. Below this post is a more detailed report and the Jupyter notebook at the end of the post.

A.2. Description & Discussion of the Background

To begin the project, I chose a hypothetical business problem.

As quoted in an article by small businesses - Advertising's Effects on Demand

A food business owner needs to decide on where to advertise his new business strategy of an online restaurant in Seattle, Washington(USA).food business owner of multiple mid-high-end restaurants decide on where to advertise his new business strategy of an online restaurant in Seattle, Washington(USA).

B. Assumptions and business thinking

The assumption behind this analysis is that we can use unsupervised machine learning to crate clusters of districts that will provide an idea of a concentration of restaurants in Seattle.

C. Audience

The restaurant business owner.

D. DATA

To consider the problem we can list the datas as below:

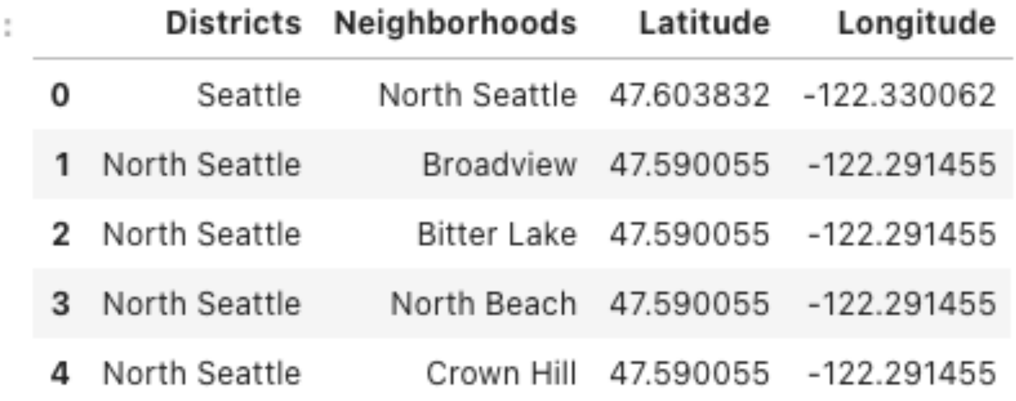
- I found a list of Districts in Seattle through Wikipedia. I cleaned the data because the HTML file had information I required but other information I didn't require in my analysis. (https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle)

- I used **Foursquare API** to get the most common venues of given town of Washington State.

- Geo-coordinates of districts will be obtained with the help of the geocoder tool in the notebook.

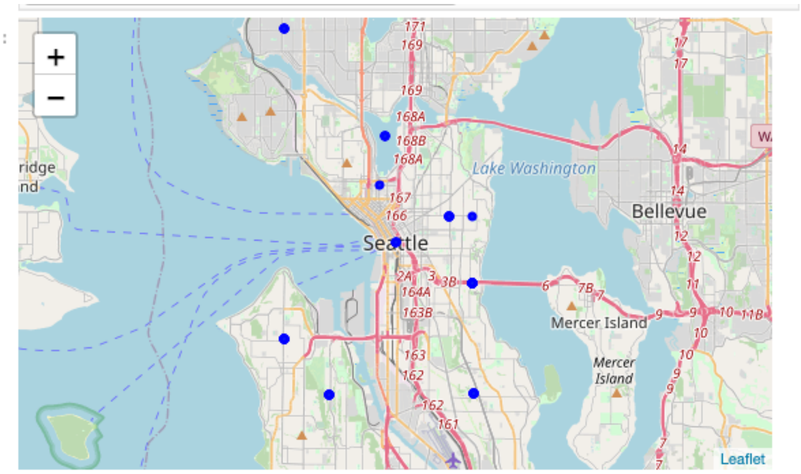
E. Methodology

After downloading all the required libraries in executing this project, I generated a dataframe by scrapping Wikipedia for a list of districts in Seattle incorporating beautiful soup and pandas. I then retrieved geo coordinates of the collected district names and merged it with my cleaned data from Wikipedia on Seattle. For this exercise, the geocode Python library was used.

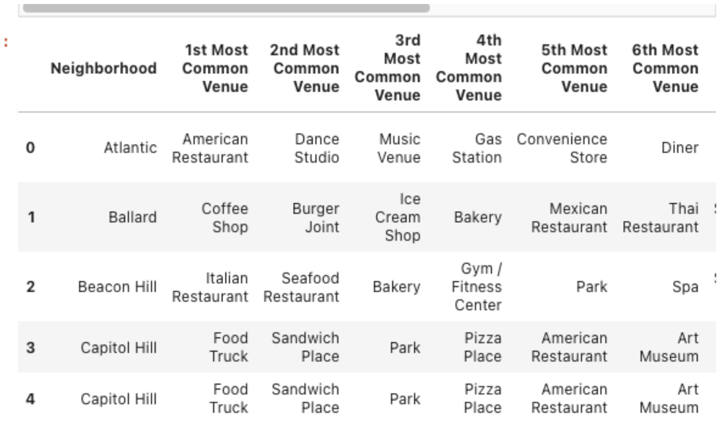


	Districts	Neighborhoods	Latitude	Longitude
0	Seattle	North Seattle	47.603832	-122.330062
1	North Seattle	Broadview	47.590055	-122.291455
2	North Seattle	Bitter Lake	47.590055	-122.291455
3	North Seattle	North Beach	47.590055	-122.291455
4	North Seattle	Crown Hill	47.590055	-122.291455

I used python **folium** library to visualize geographic details of Seattle. I created a map of Seattle with districts superimposed on top. I used latitude and longitude values to get the visual as below:

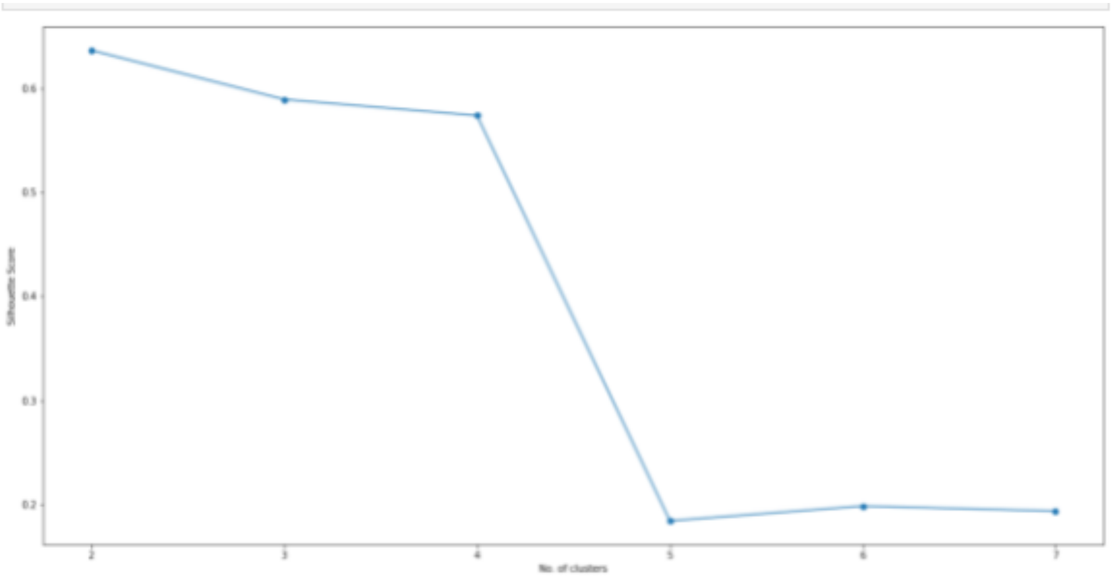


I utilized the Foursquare API to explore the districts and segment them. To explore venues in these districts, I designed the limit as 100 venue and the radius 1000 meter for each district from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

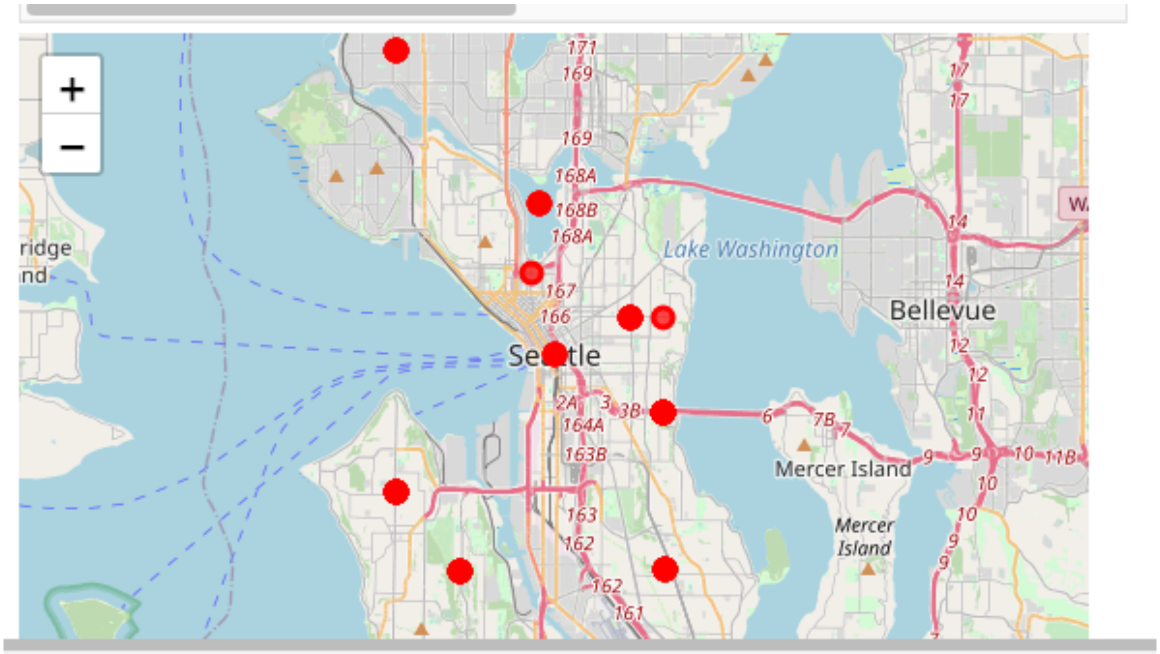


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Atlantic	American Restaurant	Dance Studio	Music Venue	Gas Station	Convenience Store	Diner
1	Ballard	Coffee Shop	Burger Joint	Ice Cream Shop	Bakery	Mexican Restaurant	Thai Restaurant
2	Beacon Hill	Italian Restaurant	Seafood Restaurant	Bakery	Gym / Fitness Center	Park	Spa
3	Capitol Hill	Food Truck	Sandwich Place	Park	Pizza Place	American Restaurant	Art Museum
4	Capitol Hill	Food Truck	Sandwich Place	Park	Pizza Place	American Restaurant	Art Museum

I used the one-hot encoding technique which converts the categorical values into dummies so they can be used for machine learning.
For the clustering process, the K-means approach was used, which is an unsupervised machine learning algorithm. This process also requires setting the parameter for the number of clusters. To be able to identify the optimal number for this parameter, the silhouette score was used. This provided us with the value 4 as the best number to be used for clusters.



Then, the K-means process with 4 clusters were performed, which provided me with the following visualized clustering below:



F. RESULTS

By looking at the cluster data, we can see that cluster 1 is the one that we are the most interested in.
Cluster 1 is the biggest cluster, food consumption is highly rated, but behind that parks, bars are also present. These are mainly areas with family houses where people live and a lively part of the city.
Cluster 3 is districts where hospitality is paramount.
Cluster 4 contain only one district. Here we see shopping is at the top.
Cluster 2 is characterized by minimum consumerism traits. It's more for events and sporting activity.

H. RECOMMMENDATION

Based on this, I can advise the restaurant owner to consider restaurant advertising in districts in cluster 1. There is potential for increased food demand by the populace in that area.

H. CONCLUSION

Providing a solution to the business owner in this example is an instance of a real life business problem the businesses encounter on a day to day basis. Data science is a brilliant and factual way that businesses can make decisions that will translate into profit maximization. The output of the analysis provided a thorough base for the recommendation for the business problem in question.