Peer-graded Assignment: Capstone Project - The Battle of Neighbourhoods (Week 1)

1.Introduction/Business Problem

The Business problem to be discussed here is that to setup a business of any kind whether it can be related to clothing, restaurant, Grocery store or a pharmacy store and the list can go on. To establish something which is related to business you just can't establish in whichever place you want if profit and success matters to you. There are some particular/important areas and statistics based on the region or otherwise that are needed to be taken care of for the success of the business. For Example: Population of the particular area can really effect the business. There are many important areas which are to be concentrated such as quality, stock, price affordability. Every customer just keeps price affordability and quality because the customer compares price with another store. The price should be neither high nor too low. To establish a business there should be uniqueness in the product or products depending on the type of product. The problem to be solved here is Where to establish a business and what criteria should be used to solve this problem and how to increase the number of stakeholders to acquire profits in the business?

2. Data section/Solution

Data scientist and Data science plays a very crucial role in this aspect. To solve the above problem there are many techniques in data science which give clear and clean statistics and show you the right solution for the problem mentioned above. In this problem we'll also use foursquare location application to get the data about the searched location. We can directly put in the address and search for the nearby stores their details their ratings etc. But as we are going to map the data we use jupyter notebook using python. We query for the foursquare location with respective client id, client secret, version to obtain the results. The results we get will be in the form of json file and we have to convert it into pandas data frame for easier access of the values and manipulate the data for good results.

Packages that are to be imported are:

1.geopy package for convert an address into latitude and longitude values of the particular location.

2.requests package to handle requests.

3.folium to map rendering library.

4.json to handle JSON files.

5.KMeans clustering form sklearn to cluster similar kind of data.

6.matplotlib to plot modules.

7.pandas to convert data into a data frame and numpy to handle data in a vectorized manner.

8.Json\_normalize to transform JSON file into a pandas data frame

After gathering all the information/data about the particular location we perform queries and clustering techniques and we choose which location is appropriate for us to establish business in preferred location.

In this aspect we use New York data set https://cocl.us/new\_york\_dataset from the link and transfer it into a data frame to obtain nearby stores in the location and cluster them to obtain the results and choose the appropriate location where we can establish the business.

We are going to establish a Italian restaurant. For this we are going to explore which location suits the best in the notebook.

To establish a business many things should be considered and then take a forward step because if there is huge success then more branches can be established otherwise there will be huge loss.

Methodology Section

In this section we will download the data sets and explore the data sets and manipulate the data accordingly. First downloading the data set and the data is a json file. Json package is imported and used to retrieve the data in the json file and json normalize package to normalize the data in the json file. pandas package is used convert into a data frame for easier access of data and manipulation of data. A data frame is created with 4 columns named as Borough, Neighbourhood, latitude and longitude.

A screenshot of a cell phone

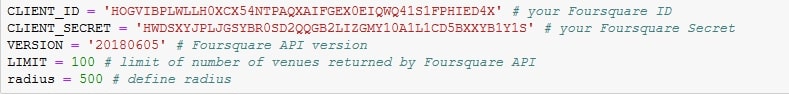
Description automatically generated

Geopy package is imported and used to get latitude and longitude values of the given address. Here new York NY is used as the address. The latitude and longitude values for the particular location is appended in the data frame with their respective locations. Here folium package to plot a map with their respective labels of neighbourhood.

A screenshot of a social media post

Description automatically generated

The next step is we use four square location app to get the nearby venues of the given address. For that client id, client secret and version are needed.



The main motive here is to open a Italian restaurant so we group the data based on the Italian Cuisine and assign a new variable to that data frame.

A screenshot of a cell phone

Description automatically generated

After grouping we create clusters using sklearn package and K Means module is imported of clustering similar kind of data with K Means clustering. There are 4 clusters.

Results Section

These are the results for the 4 clusters

Cluster 0:In this cluster there are no Italian restaurants within the neighbourhood

A close up of a piece of paper

Description automatically generated

Cluster 1:Even in this cluster there are no Italian restaurants within their respective neighbourhoods

A close up of a piece of paper

Description automatically generated

Cluster 2:In this cluster we find most number of Italian restaurants compared to other clusters

A close up of a piece of paper

Description automatically generated

Cluster 3:Less number of Italian restaurants

A close up of a piece of paper

Description automatically generated

Here K Means clustering is used because it is one of the easiest and best way of machine learnings to cluster the data. That is the reason it is used to cluster the similar neighbourhoods in this project.

Recommendations

Here by observing the clusters cluster 3 is has more Italian restaurants than any other clusters when compared and it seems that Cluster 3 will be the area that gathers the Italian Restaurant most intensively.

Benson hart seems to be the neighbourhood that incorporates the most Italian restaurants, which could also be verified by venues names. There are less number of Italian restaurants so this project recommends the entrepreneur to open an authentic Italian restaurant with wide variety of Italian cuisines in the this location with no competition.

Limitations and Suggestions for Future Research

In this project, I only take into consideration of one factor: the occurrence / existence of Italian Restaurants in each neighbourhood. 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. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors.

Conclusion section

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilising k means clustering and providing recommendation to the stakeholder. An entrepreneur can establish in either of the locations in cluster 3 as mentioned above.

References

Foursquare Developer Documentation: https://developer.foursquare.com/docs

New York Venues: https://cocl.us/new\_york\_dataset