The Battle of neighborhood (Week 2) - Report

Table of contents

Introduction: Business Problem

In this project I will try to find an optimal location for a restaurant. Here I will try finding if someone wants to open a new restaurant in the city, since there are alot of restaurants in SF, I will try to detect locations that are not already crowded with restaurants. I would also prefer locations as close to city center as possible, assuming that the first two conditions are met.
I will use our data science polrs to generate a few most enticing neighborhoods based on this criteria.

Data

a) I scrape the following page, http://www.healthysf.org/bdi/outcomes/zipmap.htm , in order to obtain the data that is in the table of postal codes and to transform the data into a pandas dataframe.

b) number of restaurants and their type and location in every neighborhood will be obtained using Foursquare API

Methodology

I use “K-Means Clustering Algorithm ". K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:
The centroids of the K clusters, which can be used to label new data Labels for the training data (each data point is assigned to a single cluster) Also, I will be utilizing different maps in-order to give a clear vision to the target audience.
Steps I took for the analysis:
a) Collected required **data: location and type (category) of every restaurant within our lat and lng. I have also the type of restaurants in particular locality.

b) Explored the 'restaurant density' across different areas of SF using K- mean to identify a few promising areas close to center with low number of restaurants and their type.

c) Explored the most promising areas and within those create clusters of locations that meet some basic requirements I will take into consideration locations with less restaurants in a radius of 500 meters, I will present map of all such locations but also create clusters (using k-means clustering) of those locations to explore neighborhood.

Analysis

Data identification, capturing and cleaning.
Search & Identify the relevant data source and capture it, here I are using wikipedia to get data about SF, California. Then I remove all the redundant value(data cleaning). Now the data is clean and ready to use.

Combining different data source and sorting neighborhood based on Longitude and latitude
Now, I will combine neighborhood dataset with postal address and dataset with Latitude & Longitude and save them in a separate data frame.The resultant data frame with contain details about Neighborhood, Latitude & Longitude. Then visualize it using a folium map.

Explore the Toronto's neighborhoods
Firstly, I explored all the neighborhoods in the city of SF, using the Latitude & Longitude data, using Foursquare API to get the Restaurant venues available in SF. Explore the unique categories in the neighborhood. Filter the Venues details for all possible ‘Restaurants’. Find each neighborhood along with the top most common venues.Identify the top 10 venues for each neighborhood.

Clustering
With an assumption of 5 clusters, use K-Cluster algorithm to come up with 5 different clusters in SF with a similar set of Venues. Explore each cluster and determine the discriminating venue categories that distinguish each cluster. Identify the clusters & neighborhoods with minimum number restaurants and their types.

Results and Discussion

Our analysis shows that although there is a great number of restaurants in SF, there are pockets of low restaurant density fairly close to the city center.
Based on our initial assumption of the cluster with maximum number of restaurants will have the best possibility to have a new restaurant due to the need in the area. Based on the resultant clusters it looks like Cluster 3 has the least number of restaurants than the rest of the clusters.
It is entirely possible that there is a very good reason for a small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

Conclusion

The purpose of this project was to identify areas in SF with a low number of restaurants in order to narrow down the search for the optimal location for a new restaurant. By calculating restaurant density distribution from Foursquare data I have first identified general neighborhoods that justify

further analysis, Clustering of those locations was then performed in order to create major zones of interest.

Based on the fact there are only a few neighborhoods in San Francisco, I can't say with complete confidence that our clusters are insightful. I had to use only these clusters. The best place to open a restaurant would be Hunters point since it has the least amount of restaurants that are most visited.