IBM Applied Data Science Capstone Project

Starting a Grocery Store in Dallas, Texas

By Jianing He

May 2020

Introduction

Dallas is the largest city in North Texas. Its estimated population was 1,323,573 in 2019 and is the ninth most-populous city in the U.S. and third in Texas. Dallas is also the main core of the largest metropolitan area in the Southern United States with diverse economy including defense, financial services, information technology, telecommunications and transportation. With this large amount of population and economy, the requirement of grocery stores or supermarkets is huge. In addition, the Dallas City Council encourages grocery-anchored developments with given million-dollar plus subsidies. According to a March 2017 City Council briefing from the Economic Development Department, the Dallas City has given more than $8.4 million incentive to four grocery-anchored developments and will fund more. Thus, it is a good investment to start a grocery store in Dallas with a development grand offered by the city. There is a lot more to be concerned to start a grocery store for a certainty. Especially, the location of the grocery store is one of the most essential decisions that will determine the target market and the profit.

Business Problem

The objective of this project is to analyze the neighborhoods of Dallas, Texas and to select the best locations to open a grocery store. With data science methodology and K-means clustering algorithm, we will give grocery-anchored developers the recommended locations where to open their new grocery stores.

Data we need to solve this problem

1. List of neighborhoods in Dallas, Texas.
   https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Dallas
2. Latitude and longitude coordinates of the neighborhoods in Dallas for map visualization and for Foursquare locations.
3. Venue data from Foursquare API to show how many grocery stores in each neighborhood and then segment and cluster these neighborhoods in Dallas.

Methodology

Firstly, we need to scrap the list of neighborhoods in Dallas from Wikipedia Page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Dallas) with Python requests and beautiful soup packages to extract the neighborhoods data from website. Then, we install and use Geocoder package to get the latitudes and longitudes of the neighborhoods for further map visualization and venues data from Foursquare API. After merging the names and geographical coordinates of the neighborhoods, we save all this information in a pandas DataFrame and then save it as a csv file. Using Folium package, we can visualize the neighborhoods of Dallas, Texas in a map.

Secondly, we request the top 1000 venues within a radius of 20k meters in each neighborhood in Dallas from Foursquare API with the Foursquare ID and secret key. The venue data includes venue names, venue category, venue latitude and longitude. Then, we check how many venues returned for each neighborhood and how many unique categories in all venues. The neighborhoods are grouped by the mean of the occurrence frequency of each venue category. Since we are focus on the “Grocery Store” data, a new target data frame is created just with “Grocery Store” data for each neighborhood.

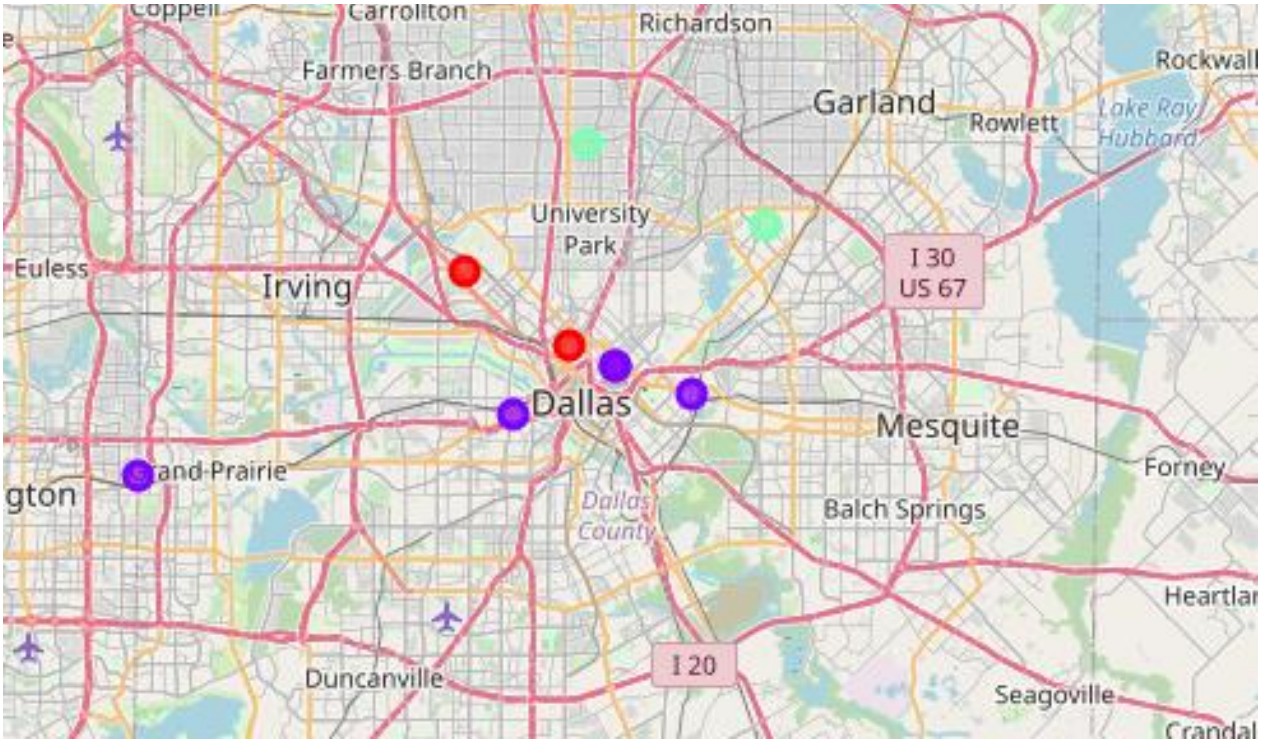
Thirdly, we will perform clustering on the target data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Grocery Store”. The results will allow us to identify which neighborhoods have higher concentration of Grocery Stores while which neighborhoods have fewer number of Grocery Stores. Based on the occurrence of Grocery Stores in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Grocery Stores.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Grocery Stores”:

- Cluster 0: Neighborhoods with moderate number of s Grocery Stores
- **Cluster 1: Neighborhoods with low number to no existence of Grocery Stores**
- Cluster 2: Neighborhoods with high concentration of Grocery Stores

The results of the clustering are visualized in the map below with cluster 0 in red color, **cluster 1 in purple color**, and cluster 2 in minty green color.



Discussion

As observations noted from the map in the Results section, this project recommends grocery-anchored developers to capitalize on these findings to open new Grocery Stores in neighborhoods in cluster 1 with little to no competition. grocery-anchored developers with unique selling propositions to stand out from the competition can also open new Grocery Stores in neighborhoods in cluster 0 with moderate competition. Lastly, grocery-anchored developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of Grocery Stores and suffering from intense competition.

In this project, we only consider frequency of occurrence of Grocery Stores, there are other factors such as population and income of residents that could influence the location decision of a new Grocery Store. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

References

[1] Category: Neighborhoods in Dallas. *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Dallas

[2] Foursquare Developers Documentation. *Foursquare.* Retrieved from https://developer.foursquare.com/docs

[3] 4 times the City of Dallas shelled out millions for grocery stores. *Advocate.* Retrieved from https://oakcliff.advocatemag.com/2019/11/4-times-dallas-shelled-out-millions-for-grocery-stores-in-southern-dallas/

[4] Dallas. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Dallas

[5] How to start a grocery store. *Truic.* Retrieved from https://howtostartanllc.com/business-ideas/grocery-store