***Best Areas to Open New Shopping malls and Shopping plazas in Texas***

By: Moushumi Zaman



1. **Introduction**

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. Shopping malls and shopping plazas are one – stop destinations for all types of shoppers. For retailers, the central location and the large crowd at the mall provide a great opportunity to market their products and services. Property developers are also taking the advantage of this trend to build more shopping malls to cater this demand. Opening shopping malls allows property developers to earn a consistent rental income. But as many other business decisions, opening a shopping mall also requires a serious consideration and is lot more complicated than it seems. Specially, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

* 1. **Business Problem**

The object of this capstone project is to analyze and select the best locations to open new shopping malls and shopping plazas in the state of Texas, USA. Using data science methodologies and machine learning techniques like clustering and with help of Foursquare API this project aims to answer the business question: In Texas, USA, if a developer is looking to open a new shopping mall, where would be your recommendation to open it?

* 1. **Target Audience**

This project is particularly targeted to the property developers and investors looking to open new shopping malls and plazas in the state of Texas, USA. Texas is a huge state with many cities. The state gained 187,545 people from migration between July 2017 and July 2018 — even after accounting for people leaving the state, according to U.S. Census data. In 2018, the majority of migrants to Texas — 104,976 people — came from other countries, with the rest arriving from other U.S. states. Though many of this cities already have shopping malls and plazas, with the increasing population which eventually leads to an increasing demand, opening shopping malls and plazas is a great investment idea.

1. **Data**
   1. **Data needed to solve the problem**

* List of cities in Texas, USA. This defines the scope of the project.
* Latitude, longitude coordinates of the cities. This is required in order to plot the map and to get the venue data.
* Venue data, specially data related to shopping malls and shopping plazas. I used this data to perform clustering on the cities.
  1. **Sources of data**

As the most populous cities has more demand for shopping malls and plazas, I used the list of Texas cities ranked by population from the Wikipedia page <https://en.wikipedia.org/wiki/List_of_cities_in_Texas_by_population>. This list consists of 68 cities. I used the top 40 cities as my dataset. I used pandas ‘read\_html’ method to extract the table from the Wikipedia page.

Then I used geographical coordinates of the cities by using Pyhton Geocoder package which gave me the latitude and longitude coordinates of the cities.

Finally, I used the Foursquare API to get the venue data for those cities. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers all around the world. Foursquare API provided me with many categories of venue data. But in this project I am particularly interested with the data of shopping malls and shopping plazas. Therefore, I built a new data frame with only the shopping malls and shopping plaza information.

1. **Methodology**
   1. **Getting the list of the cities**

For this project I needed the list of cities in Texas. As more population means more demand, I searched for the list ranked by the population size of the cities. Luckily I found such list in Wikipedia. Here is the link of the Wikipedia page (<https://en.wikipedia.org/wiki/List_of_cities_in_Texas_by_population>). I scraped the list from the web page by using the ‘read\_html’ method from pandas library. The list contains the census report from 2010 and 2019 and the change in these years. As the list is already ranked by the population size and I didn’t need the exact population size but just the list of most populous cities, I recreate the table only with the city names and only with the top 40 most populous cities.

* 1. **Getting the coordinates of the cities**

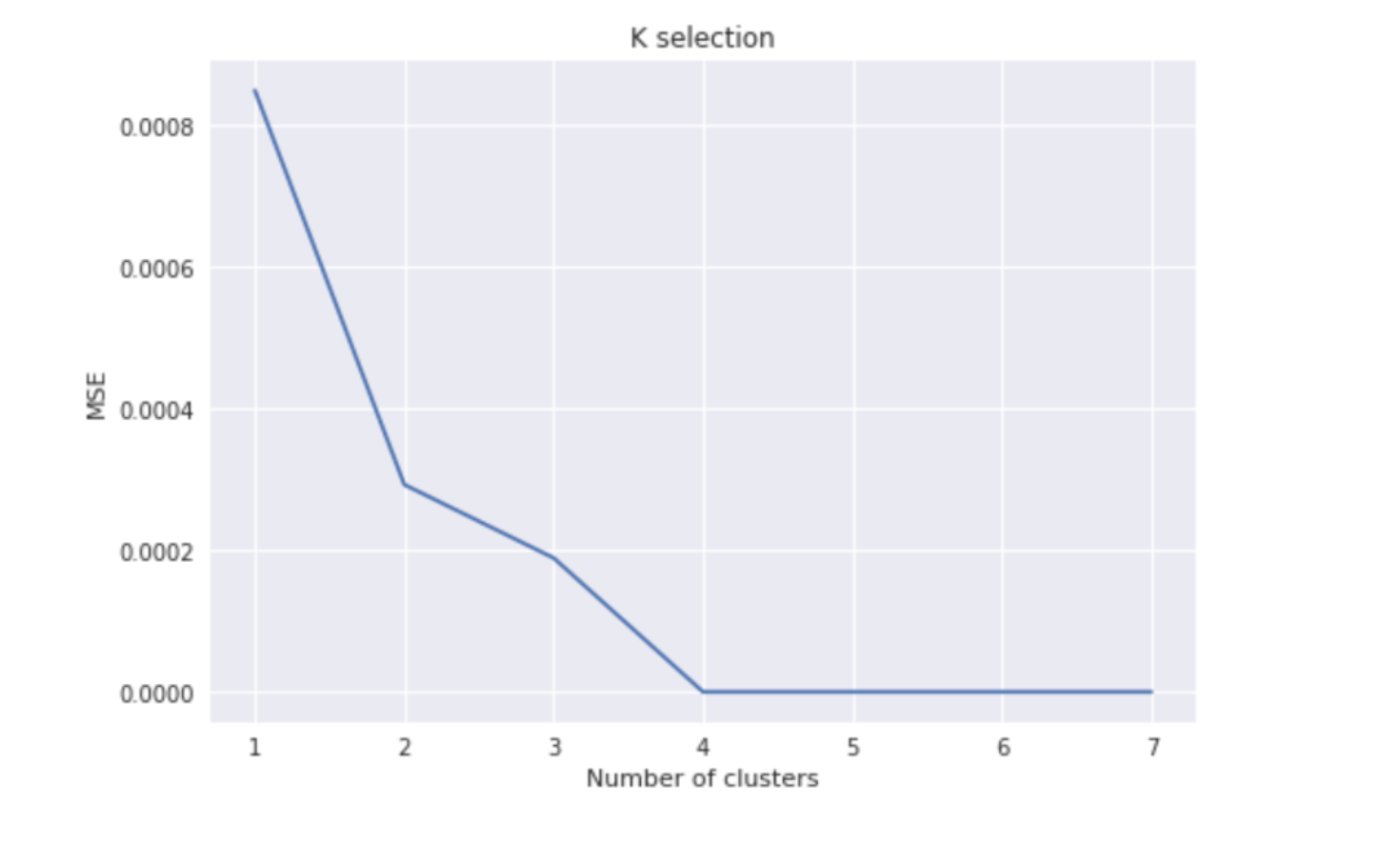
Then I needed the latitude and longitude of the cities. I needed the coordinates to be used by Foursquare API to get the venue data. So, I used the Python Geocoder package to get the coordinates for the listed cities. Then I populated a new pandas data frame with cities along with their coordinates. Next, I used Folium package to create a map to visualize the cities. This helps us to perform a sanity check to make sure that the coordinates provided by the Geocoder are correct.

* 1. **Using Foursquare API to get the venue data**

Next, I used Foursquare API to get top 100 venues that are within the radius of 2500 meters. I made API calls to Foursquare passing the coordinates of the cities. Foursquare, using these coordinates, returned the venue data in JSON file, and I extracted the venue names, venue category and venue coordinates from this file to a new data frame. I can see the number of different venues and the number of different categories of the venues in each city. Then I analyzed the data by grouping each city by rows and taking the mean of the frequency of occurrence of each venue category. As I am only interested in the shopping malls and shopping plazas, I filtered the table with only these categories.

* 1. **Finding the number of clusters**

Then I used ‘seaborn’ library and ‘matplotlib’ library to get the perfect number of clusters I need to perform clustering. I used elbow method for this step. According to the analysis the best number of clusters is 4.



* 1. **Clustering**

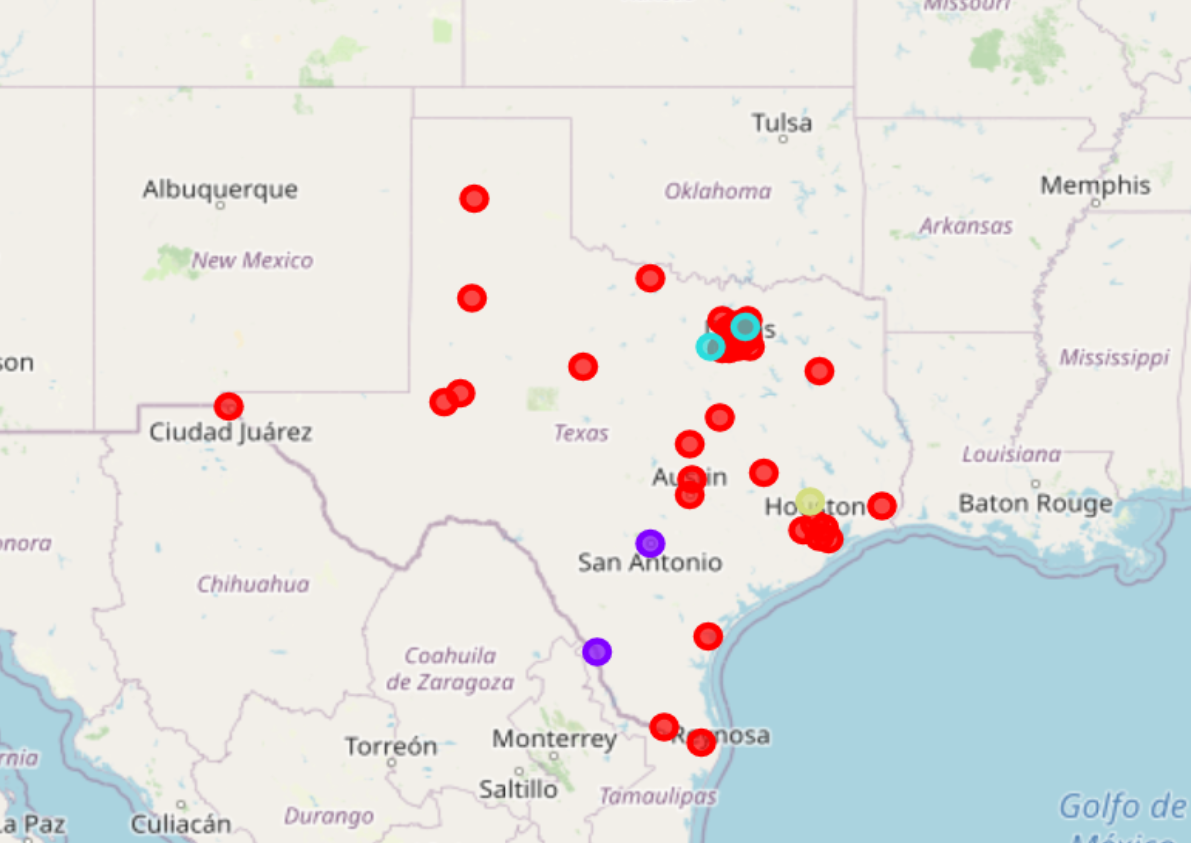
Finally, I performed k-means clustering on the data to get 4 clusters of the cities based on the frequency of the occurrence for Shopping malls and Shopping plazas. K-means clustering algorithm identifies k number of centroids (in my case it’s 4) and then allocates every data points to the nearest cluster, while keeping the centroids as small as possible. The result showed me which cities have a higher concentration of shopping malls and plaza and which have the lowest. Based on the occurrence of shopping malls and plazas in different cities, k-means clustering algorithm helped me to answer the question as to which cities will be most suitable to open new shopping malls.

1. **Result**

The result from k-means clustering showed me that I can categorized the cities into 5 clusters based the frequency of occurrence of shopping malls and shopping plazas:

1. Cluster 1: This cluster has the most number of shopping malls and plazas and the best one to invest
2. Cluster 2: This cluster has the moderate number of shopping malls.
3. Cluster 3: This cluster also has a moderate number of shopping plazas.
4. Cluster 4: This cluster has least number of shopping malls

The result of the clustering is visualized in the below map with cluster 1 in red color, cluster 2 in purple color, cluster 3 in blue color, cluster 4 in yellow color.



**5. Discussion**

As observations noted in the result section, most of the malls and plazas are located in cluster 1 whereas cluster 4 has the least number of shopping malls. That makes the area around cluster 4 a great place to open a new shopping malls and plazas and property developers are advised to avoid cluster 1 which already has a great number of malls and plazas.

**6. Limitations**

In this project I considered only the frequency of shopping malls and plazas, but for a business decision of opening a new shopping facility other factors such as size of the population and income of the population are also important. But in this project these factors were not taken into consideration. However, to my best knowledge as a researcher these data are not available to the city level as required by the project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm.

**7. Conclusion**

In this project I identified the business problem, extracted the required data, cleaned the data and used machine learning algorithm to answer the identified business problem of which location will be a good choice to open a new shopping facility. After all the processes I can conclude my research with the finding and recommend investors to invest in cluster 4 to open new shopping facility and to avoid cluster 1 as this is the cluster with too many already established shopping facility.