Capstone Project

Applied Data Science Capstone

Best Location to Build a New Shopping Center in San Pedro Garza Garcia, MX

David Resendez

Introduction

There is no doubt about the great interest of real estate investors in shopping malls, due to their constant income. For the vast majority of people, they choose to have their convenience stores in one place, in addition to having a pleasant time to relax or enjoy good entertainment. Serving these centers as a good sales channel for large and small sellers.

Business Problem

The project would focus to analyze locations using machine learning and clustering and select the best one to build a new shopping mall in San Pedro Garza Garcia, one of the richest cities in Latin America.

Target Audience of this project

The main objective of these projects are investors, who with the help of developers, can make better decisions to increase their profits.

Data

The following data is need to carry out the project:

- List of neighborhoods of San Pedro Garza Garcia
- Coordinates of the city and neighborhoods
- Venue data

Sources of Data

To extract the neighborhoods and postal codes table, we will use the webpage:

https://micodigopostal.org/nuevo-leon/san-pedro-garza-garcia/

and we'll use the library BeautifulSoup from bs4 to scrap all the information required from the webpage.

For the geographical coordinates we will use Python Geocoder package.

And for the venues data information, we will use the Foursquare API. Foursquare is a technology company that built a massive dataset of location data.

Methodology

The way the project was carried out was that we first extracted the table with the required information using BeuatifulSuop.

Following this, we cleaned up the DataFrame a bit to be left with only the necessary information.

To get the coordinates in the proper format to use the Foursquare API, we use the Geocoder package to convert the address into coordinates.

With the previous steps you can obtain a new DataFrame with more complete information.

Then we use the Foursquare API to get the venues information within a radius of 1000 meters.

To do this, a previous registration was made for a Foursquare Developer Account to obtain the necessary data.

The data was prepared for use in clustering.

In the end, the data was clustered using K-means Clustering. A cluster of neighborhoods was made into 3 clusters based on their frequency of occurrence for "Shopping Mall" and "Shopping Plaza". The results allow us to identify which neighborhood have the fewer number of shopping malls.

Results

The results from the K-Means Clustering show that we categorize the neighborhoods into 3 clusters:

- Cluster 0 (red): More concentration of shopping malls
- Cluster 1 (purple): Moderate concentration of shopping malls
- Cluster 2 (mint): Low concentration of shopping malls
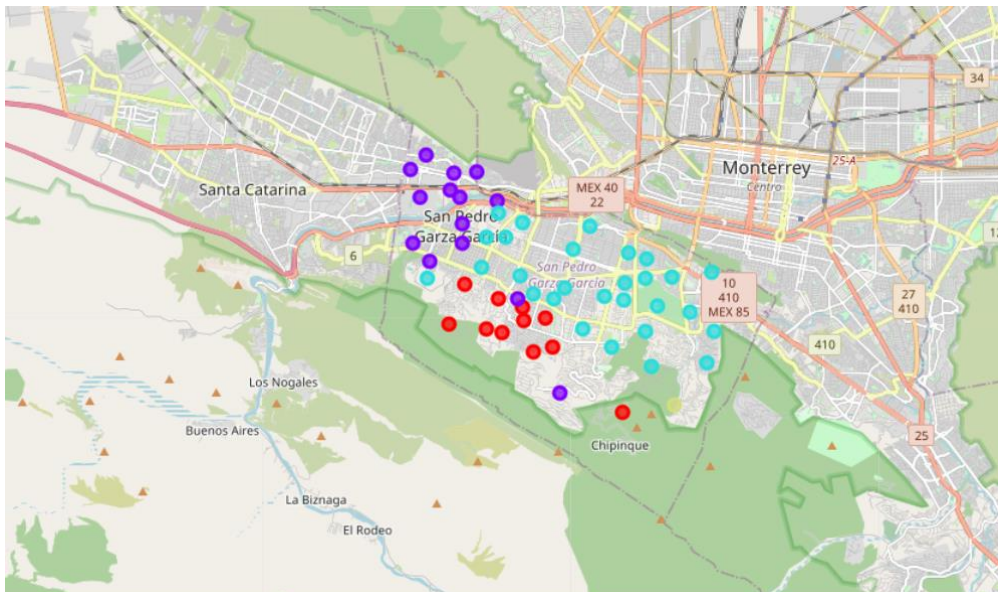
Discussion

As we can see from the results, in the areas with the highest concentration of shopping centers, they are areas closer to other municipalities in the metropolitan area.

The areas with moderate and low concentration of shopping malls are a good option for the investor, it should only be taken into account that there are areas that are close to natural protected areas.

Conclusion

With the project, it was possible to achieve the problem posed in the introduction, helping investors to interpret the information easier and in a more agile way than the conventional one.
Proving that with K-Mean Clustering I can find the most suitable area to start a new shopping real estate.