IBM – Coursera

Data Science Specialization

Capstone project - Final report

Open A new Shopping Mall in San Jose, California

Introduction

San Jose is a diverse city with many High Technology Firms and being heart of Word Famous Innovation built on Silicon fabricated System Applications not just Products such as AI application Robotics Internet of Things and many-others. On the other hands during weekend of holidays there are many shoppers spent their family life on leisure activities, relax, scroll around the shopping malls to dine and shop spent time and money in the malls.

Business problem

The objective of this capstone project is to analyze and select the best locations in the city of San Jose, California to open a new shopping mall. Use Data Science methodology and machine learning techniques such as clustering, this project intends to provide solutions to answer the business question: In the city of San Jose, if a property developer is looking to open a new

shopping mall, where would you recommend that they open it?

Target Audience of this project

This project is catering to property developers and investors looking to open or invest in new shopping malls in capital city of Silicon i.e. San Jose. This project is timely as the city and U.S.A. is currently suffering from oversupply of shopping malls due to unprecedent Coronavirus and malls are subjected to closing down due to the virus infections. However, when the Virus is gone there is always need for new business investment since San Jose is such a high business activities region deserved investment such as Malls or Business Offices.

Data

To solve the problem, we need the following information:

- Use of Neighborhood of San Jose California
- Latitude and longitude of those neighborhoods which is required to plot the map and use to get the venue data
- Venue Data particularly related to shopping malls and it is used to perform clustering of the neighborhood.

Source of the data and the method to extract them

The Wiki page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Jose,_California) contains the neighborhood of San Jose We will use web scraping with the help of Python Requests and BeautiSoap packages. Then we can use Python Geocoder to obtain the Latitudes and Longitudes of the Neighborhood. We will get the Venue Data using FourSquare API for the Neighborhood. FourSquare provide many categories of venue data we are particularly keen on the one related to Shopping Mall one to help us solve our problem putting forward.

This is a project using many data science using many data science skills from web scraping (Wiki) working with FourSquare API, data cleaning, data wrangling to machine learning (K-means Clustering)

and map visualization (Folium). In next section we will present Methodology where we have taken and data analysis, we have used with machine learning namely K-mean clustering.

Methodology

Firstly, we need to get the list of neighborhoods in the city San Jose, California. The list is available in the Wikipedia page: (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Jose,_California). We will do web scraping using Python requests and beautifulsoup to extract the list of the neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude to use Foursquare API. To do so, we will use Geocoder package will allow us to convert address into geographical coordinates. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using folium package. This allows us to perform a sanity check to make sure that the geographical coordinates returned by Geocoder are correctly plotted in the city of San Jose.

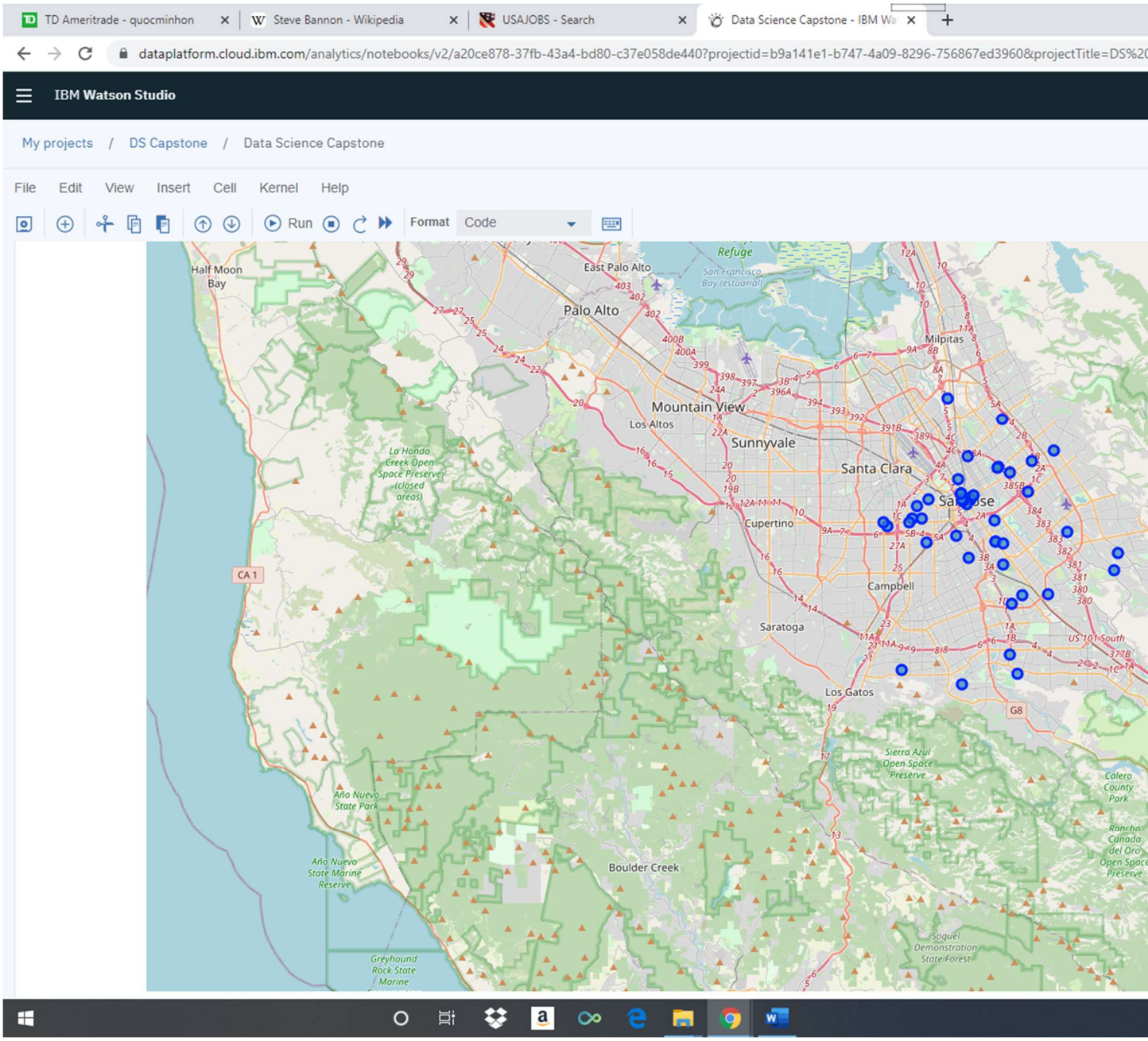

Fig 1 : 47 neighborhoods are plotted with Folium Package.

Next, we will use Foursquare API to get top 100 venues that are thin radius of 2000 meters.
We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and secret key. We then make API calls to Foursquare passing in geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will

extract the venue name, venue category, venue latitude and longitude. With data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data 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 it is particularly suited to solve the problems for this Project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence For the “Shopping Malls”. The results will allow us to identify which neighborhoods have higher Concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the Question as to which neighborhoods are most suitable to open new shopping malls.

Results

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

- Cluster 0: 30 Neighborhood with no existence of shopping malls
- Cluster 1: 7 Neighborhoods with moderate concentration of shopping malls
- Cluster 2: 10 Neighborhoods with high concentrated number of shopping malls

Discussion

The high concentration Cluster 1 would not be recommended to open new shopping malls at all while the big opportunity would be in Cluster 0 while there is no competition for foot traffics since there is no shopping malls opened yet.

Limitation and future research suggestions

There are other factors such as population density and income levels of the neighborhoods which is not available at the scope of this study which is based on mall density or frequency of occurrence of existence of the shopping mall. In future research those two factors should be considered in the future.

Conclusion

In this project we have identified the process identifying 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 proving recommendations to the relevant stakeholders, property developers and investors regarding the best locations to open a new shopping mall. To address the business question that was raised in the introduction section, the answered proposed by this project is: The neighborhoods in cluster 0 are the most preferable locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to Capitalize on opportunities on high potential locations while avoiding overcrowded areas In their decisions in opening a new shopping mall.