IBM Data Science Capstone

Introduction: This is the final report for the IBM Data Science Capstone Project on Coursera which is part of a 9-course specialization of the IBM Data Science specialization. The main objective is to use the Foursquare API to analyze the most popular venue of neighborhoods in San Francisco and see where the best neighborhood or part of San Francisco is to open a restaurant. We use the internet to gather data about the various neighborhoods of San Francisco by using websites such as Wikipedia for the whole list of neighborhoods and look at which neighborhoods have the most population. The Foursquare API contains a vast amount of information for any venue that you want to analyze from restaurants to clubs to coffee shops and the prices of them as well and can be accessed by creating a Foursquare developer account. San Francisco is a densely populated city with a decent public transportation such as BART and MUNI. San Francisco is a city with around 800,000 people living in it. San Francisco also has a very large technology presence so a very large amount of tech workers and a large amount of educated folks with college degrees make up most people living in San Francisco with around 57% of residents having a bachelor's degree or higher and where the median household income is around $100k. There will always be a demand for new restaurants to open as there are a huge amount of various cuisines around the world that people are willing to try.

Data Description: The list of neighborhoods from San Francisco was gathered from Wikipedia. The venues in each neighborhood of San Francisco was gathered from Foursquare. There are around 119 neighborhoods in San Francisco in total. The most closer you get to the downtown district which is the Financial District and the Embarcadero where most of the office buildings and skyscrapers are located, the real estate gets more expensive The list from Wikipedia is gathered through html using BeautifulSoup and used for analysis through Python by storing it in a data frame through Pandas in Jupyter Notebooks. Geocoder is also used to get the latitude and longitude coordinates of the different neighborhoods. We then use NumPy, Matplotlib, Sklearn for machine learning and other Python libraries such as Folium to make statistical analysis about the data such as creating maps and charts of the data.

Methodology: KMeans clustering using Sklearn is the best way to find the best place to open a restaurant in San Francisco because it clusters the neighborhoods and checks which neighborhoods have the most restaurants that are popular.

Results: According to the analysis, most of the restaurants are in the green cluster, which is in the downtown or near the ocean. The places with the least amount of restaurant in highly populated neighborhoods would be the best way to open them so there is no competition for that restaurant among the other restaurants which is in the red cluster.

Map of SF with the coordinates as neighborhoods

Map of SF with the neighborhood coordinates divided in clusters
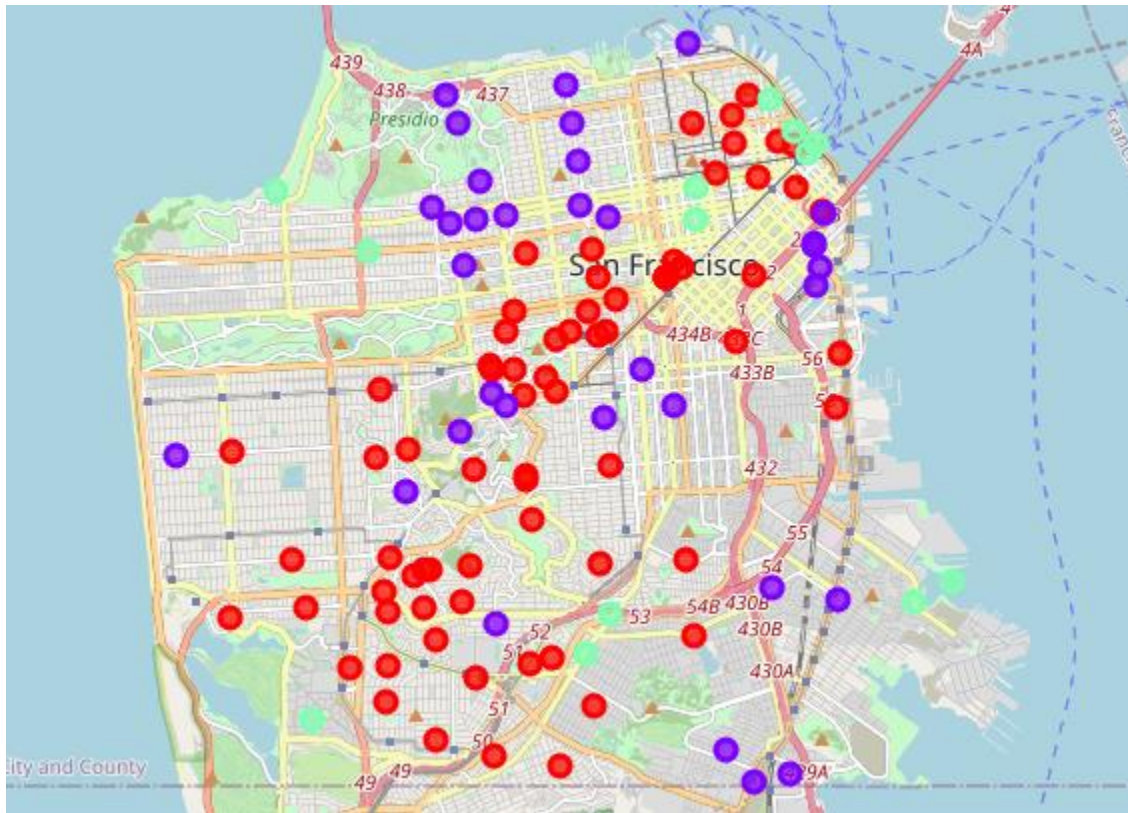
Discussion: The least amount of restaurants appear to be in the red cluster which seems to be the best place to open up a new restaurant because there are still a huge population in the red cluster and less restaurants so there is a higher chance of people coming to this restaurant.

Conclusion: Most of the restaurants are in or near the downtown or financial district and located near the ocean. The best place to open a restaurant would be in the neighborhoods where most of the people live which are in the red cluster and towards the middle of San Francisco. There is always a demand for restaurants to open and there is a large business in the restaurant industry because there are many cuisines from all around the world that people are willing to always try.

References:

https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco

https://en.wikipedia.org/wiki/San_Francisco

https://www.niche.com/places-to-live/san-francisco-san-francisco-ca/

https://developer.foursquare.com/