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 also look at the median price and rent of the apartments and homes in the neighborhoods. 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. It would make sense for most venues to be in or near the downtown so most of the people in the city and greater Bay Area can go to them. 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.

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 stored in a csv file and used for analysis through Python by storing it in a data frame through Pandas in Jupyter Notebooks. We then use NumPy, Matplotlib, Sklearn for machine learning and other Python libraries such as Seaborn 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 3rd 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.

Discussion:

Conclusion: Most of the restaurants are in or near the downtown or financial district and located near the ocean.

References:

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

https://www.zillow.com/san-francisco-ca/home-values/

https://www.businessinsider.com/san-francisco-housing-market-facts-rent-2019-5#and-the-median-rent-for-a-one-bedroom-apartment-in-san-francisco-hit-almost-3700-in-2019-3

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

https://www.rentcafe.com/average-rent-market-trends/us/ca/san-francisco/

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

Table of Figures: