Coursera Capstone Project

IBM Data Science Professional

New Movie Theaters in San Francisco, California

By: Kyle Tran

September 2019

Introduction/ Background

The Gold Rush built San Francisco, which brought a lot of immigrants and technology during its time. Earthquakes throughout history could not slow down the advancements. Before, there were groups of Native Americans who lived by hunting and gathering, but Spanish settlers discovered the bay and founded a mission. Now, San Francisco is known as a "City of Dreams" where the dream of a transcontinental railroad became a reality. Feats like these made San Francisco a diverse and bustling city full of adventure, work, and entertainment.

The idea of choosing the proper location for a business is time-consuming, stressful, and tedious. A company needs to know the amount of foot traffic and competitors to ensure the success and future of the business. The goal of opening a new movie theater is to avoid nearby competition and increase the amount of foot traffic. It is challenging to research which area in San Francisco would be best for a theater.

Movie theaters have become an integral part of the way people live since 1905. The first theater originated in Pittsburgh, Pennsylvania, and it revolutionized the way we spend our free time around the world. Short silent films eventually evolved into the motion pictures, which led to significant advancements in film production became the foundation in many technological types of equipment and techniques that are used today. As the popularity of movies increased the market, companies began to invest money into expanding more locations around the world.

Target Audience

The audience this project is pursuing are project developers in the San Francisco Bay Area. By providing insightful information about locations and competitors, the developers increase their opportunities for the success of movie theaters.

Business Problem

The purpose of my capstone project is to examine which locations in San Francisco, California, would be fit to open a movie theater.

Question:

Which area in San Francisco, California, would be recommended for project developers to open up a new movie theater?

Data

The data needed for this project would be the locations of each area(along with latitude and longitude), latitude and longitude of movie theaters nearby, and various information of movie theaters. The source is from a Wikipedia page(https://en.wikipedia.org/wiki/San_Francisco_Bay_Area) that contains tables of 10 areas. Web scraping techniques from before was used to extract the data we need using Python. Afterward, the location(latitude/ longitude) was provided for us through geocoder.

Foursquare API provides users social location services that provide information about the venues to 150, 000 developers. The project uses the service to obtain the data of movie theaters and their respective locations for plotting each area. After cleaning/ wrangling the data, the FourSquare API plotted locations of movie theaters and is clustered using KMeans. Perspectively, this process portrays the Data Science methodology and present an accurate analysis of the data.

Methodology

The data provided by Wikipedia did not need too much cleaning, but the table had no latitude or longitude, which was essential for plotting points using Folium. The dataset has 9 different counties containing the population, median income, latitude, and longitude.

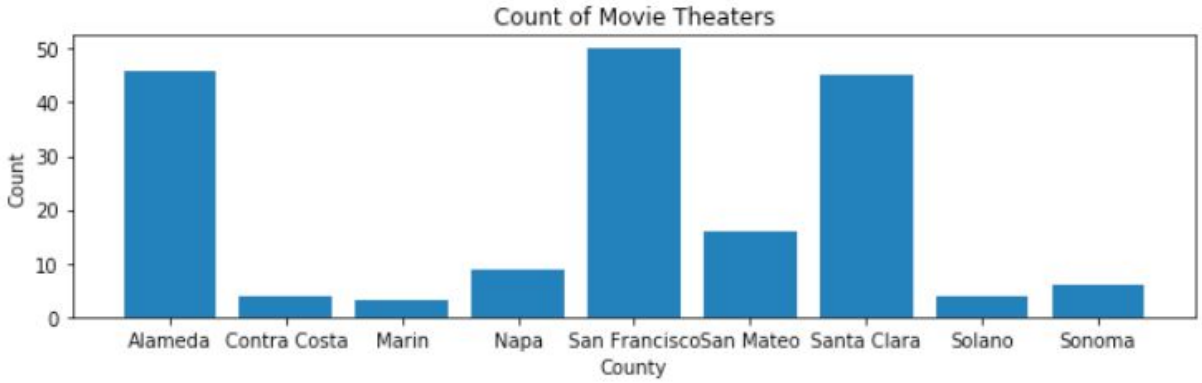
	County	Population	Median Income	Latitude	Longitude
0	Alameda	1,494,876	87,012	37.7799	-122.2282
1	Contra Costa	1,037,817	93,437	37.8584	-121.9018
2	Marin	250,666	113,826	38.0834	-122.7633
3	Napa	135,377	79,884	38.2975	-122.2869
4	San Francisco	870,887	87,329	37.7749	-122.4194
5	San Mateo	711,622	104,370	37.5630	-122.3255
6	Santa Clara	1,762,754	103,255	37.3541	-121.9552
7	Solano	411,620	79,316	38.3105	-121.9018
8	Sonoma	478,551	78,227	38.2919	-122.4580

In order to get the location, geocode was utilized to get a defined location for the San Francisco area. Geocoding is known as converting given addresses into coordinates. The coordinates were graphed on folium to see the distance and surrounding area for each county. It was crucial to take note which areas had a waterfront or were inland. Population and income gave a sense of which areas would be more willing to watch movies.

The Foursquare API retrieved venues for each county in a 10,000-meter radius. There were 183 venues in total.

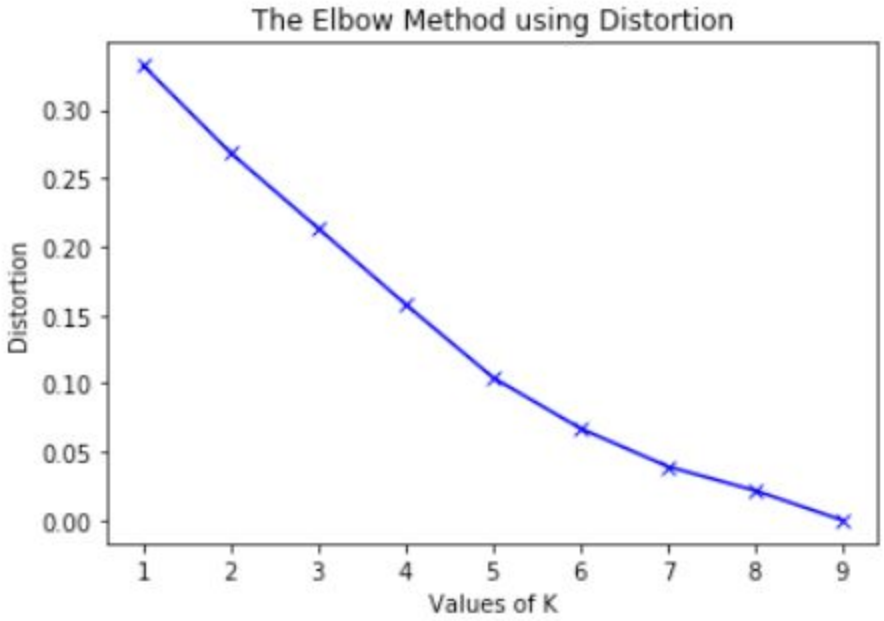
- **Alameda**- 46 venues
- **Contra Costa**- 4 venues
- **Marin**- 3 venues
- **Napa**- 9 venues
- **San Francisco**- 50 venues
- **San Mateo**- 16 venues
- **Santa Clara**- 45 venues
- **Solano**- 4 venues
- **Sonoma**- 6 venues

We can visualize the counts using the bar plot to see how each county compares to each other.



The algorithm used for cluster analysis of the venues was **k-means** from sklearn. K-means is an unsupervised clustering algorithm that tries to find the number of clusters in the data. Each centroid for the clusters is iterated to calculate the mean and find the new value. In order to change the venues into integer values, the data had to be **one-hot encoded** for the k-means.

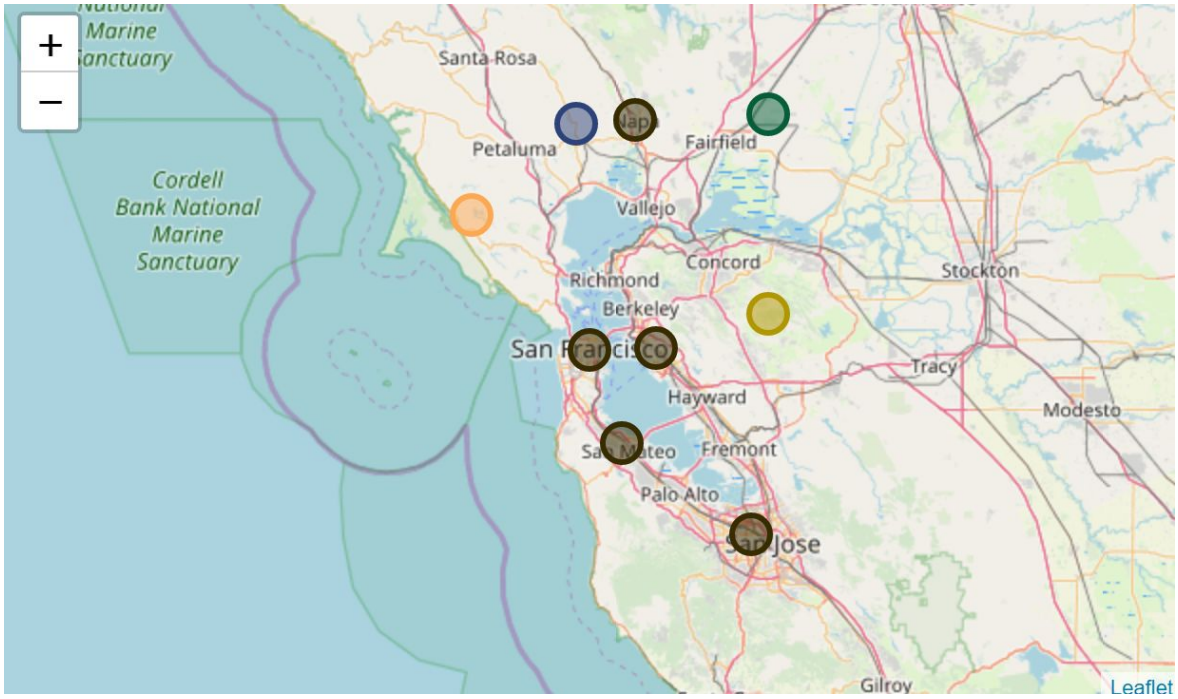
K-Means elbow method was used to determine the proper number of clusters in the dataset. The elbow point in the case could be 5 or 6, but since the score for k=5 was higher than k=6, 5 clusters were chosen to represent the data.



Results and Discussion

After cleaning the data, we can see the different counties in the San Francisco Bay Area and their concentration of movie theaters. Folium can display the 9 labels and clusters on the map. 5 numbers were used to determine which area would be best to open a movie theater, 0 is the worst, and 5 is the best.

- Cluster 0-Brown
- Cluster 1 -Blue
- Cluster 2-Orange
- Cluster 3-Yellow
- Cluster 4-Green



5 areas are Cluster 0: Alameda, San Francisco, San Mateo, Santa Clara, and Napa. These areas should be avoided at all costs for opening anything related to a movie theater because the concentration of entertainment would create unnecessary competition. The high demand for movie theaters is met with tens of theaters in the 6-mile radius.

Cluster 1 only has only Sonoma in its category. This county is large with little theaters, but it is not as populated as other areas. Sonoma is categorized as an avoid due to the lack of foot traffic.

Cluster 2 is another solo cluster due to its interesting characteristics. Marin has the highest median household income but almost the lowest population and count of movie theaters. Cluster 3 has contra costa in the Northeastern region of the Bay Area. Contra Costa is one of the prime candidates for a

new movie theater with its high population and low concentration of theaters.

Cluster 3 has contra costa in the Northeastern region of the Bay Area. Contra Costa is one of the prime candidates for a new movie theater with its high population and low concentration of theaters.

The best candidate is Solano in cluster 4. Cluster 4 is highly recommended as the ideal place to open a new movie theater. Solano has only 4 movie theaters for a decently populated area. The water is not accessible, meaning a movie theater would be a recommended form of entertainment for the population.

Conclusion

Throughout this project, we have identified a business problem, target audience, data, data science methodology, and results using the tools learned from the IBM Data Science course. Population and income were provided from the data, but the clustering was not taking the variables into account. The results from the project present a lot of insightful data for project developers by using the K-means clustering algorithm for opening a new movie theater in the San Francisco Bay Area. Hopefully, the future would provide more public data to analyze to provide concrete implementations of the results from today.

References

San Francisco.*Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/San_Francisco

Foursquare Developers Documentations. *Foursquare*. Retrieved from https://developer.foursquare.com/docs

Movie Theaters and Cinema Through the Decades. Retrieved from https://www.cheatsheet.com/entertainment/movie-theaters-and-cinema-through-the-decades.html/

Tim Lambert (2018 July 4). A Brief history of San Francisco, California. Retrieved from http://www.localhistories.org/sanfrancisco.html