IBM Data Science Certificate Capstone Project

California Exodus: Neighborhood & Venue Analysis of San Francisco and Austin

Jewon Ju

March 14, 2021

1 Introduction

1.1 Background

The Bay Area has been experiencing a mass exodus of Big Tech companies in recent years. In 2020, California suffered the largest outflow of residents among all 50 states and ranked last for migration growth. This phenomenon has been exacerbated by the growth of remote working opportunities due to the COVID-19 pandemic, high rents, skyrocketing cost of living, and long commutes. Tech giants, such as Apple, Google, Salesforce, and Twitter have sent their employees home to work remotely, allowing workers to look for cheaper alternative cities. Additionally, businesses in California are also looking for alternative locations for their company headquarters due to high corporate and income taxes. Companies like Oracle and Hewlett Packard Enterprise have moved their entire headquarters to Texas.

Austin, Texas has emerged as one of the most popular alternative destinations for workers and businesses moving out of the Bay Area. Austin offers significantly lower housing costs, property taxes, cost of living, and state income taxes.

1.2 Problem

People moving to new locations face many difficulties. Especially during a pandemic, it may be impossible to visit the location prior to deciding where to live. Moreover, if people are having to move abruptly due to their employment situation, people may not have enough time to consider all the variables to take into account when choosing their new home. Many

people could also face hardships adjusting to their new environment if the changes are too drastic. Therefore, the neighborhood to which one migrates can be a huge factor in whether people have a positive experience adjusting to their new lifestyle.

1.3 Target Audiences

1. San Francisco residents looking to move to Austin.
2. Austin residents looking to move to San Francisco.
3. Anyone deciding between Austin and San Francisco.

2 Data

This analysis aims to solve the issue of finding the best alternative neighborhood for Bay Area residents looking to move to Austin, Texas. The different types of venues available in the vicinity of the neighborhood will be considered to group similar neighborhoods between San Francisco, California, and Austin, Texas.

2.1 Data Sources

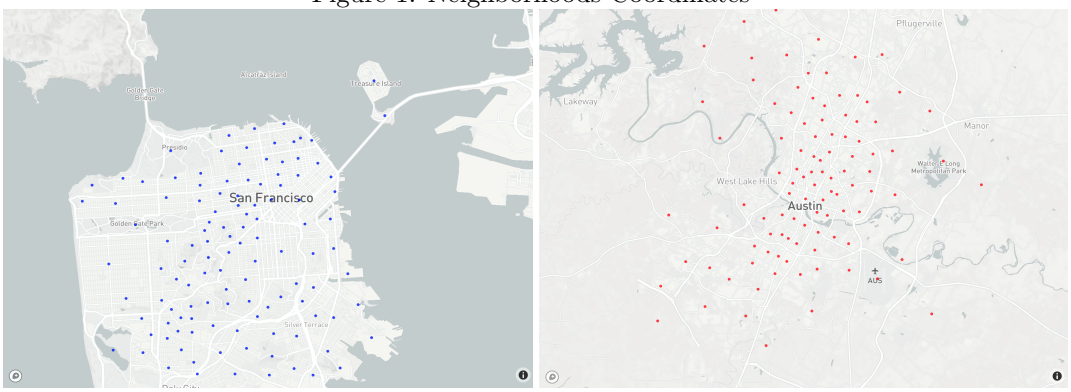
I will be using geospatial data from SF Open Data Portal and Austin Open Data Portal to obtain the coordinates of the centroid for each neighborhood. I will obtain venue data for each neighborhood from the Foursquare API. I will download the top 100 venues located in 1 mile radius around each neighborhood using the API.

2.2 Data Collection & Cleaning

The downloaded data from SF Open Data Portal and Austin Open Data Portal contained MULTIPOLYGON geometries of each neighborhoods. I used the Shapely Python package to calculate the centroid of each geometry, which allowed me to extract the coordinates. Then I used the Mapbox Maps visualization library and Plotly graphing library to display and confirm that the coordinates for each neighborhood were extracted correctly. The resulting maps are shown in Figure 1.

Using the coordinates extracted from the downloaded data, I scraped venues data from the Foursquare API. I built a function that makes API calls and extracts all the venues located in a given radius around the coordinates provided. I decided to use different radii for Austin and San Francisco since there is a significant difference in area between the two cities. The area of San Francisco is 46.87 mi^2 whereas the area of Austin is 271.8 mi^2. As a result, calculated the average radius of neighborhoods by dividing the total area of the city by the number of neighborhoods and using the formula for the area of a circle. However, this is based on the assumption that the neighborhoods are all shaped like a circle and that the

Figure 1: Neighborhoods Coordinates



distribution of neighborhood areas is rather uniform. The assumption was mainly made due to the difficulty of obtaining accurate areas for each neighborhoods in both cities.

Foursquare API provides information about venues and geolocation. The API call returns a list of venues near the specified coordinates and radius. I am most interested in the categories of the venues for our analysis. However, the venue category returned only contains the subcategories. This could be problematic for cluster analysis because there may be too many unique venue subcategories that make it difficult for the algorithm to cluster the neighborhoods into meaningful clusters. Therefore, I converted the subcategories into top-level categories. There were total of 10 top-level categories in the data set as shown in Figure 2. I used this table to assign top-level categories to the corresponding subcategories.

Figure 2: Top-Level Venue Categories

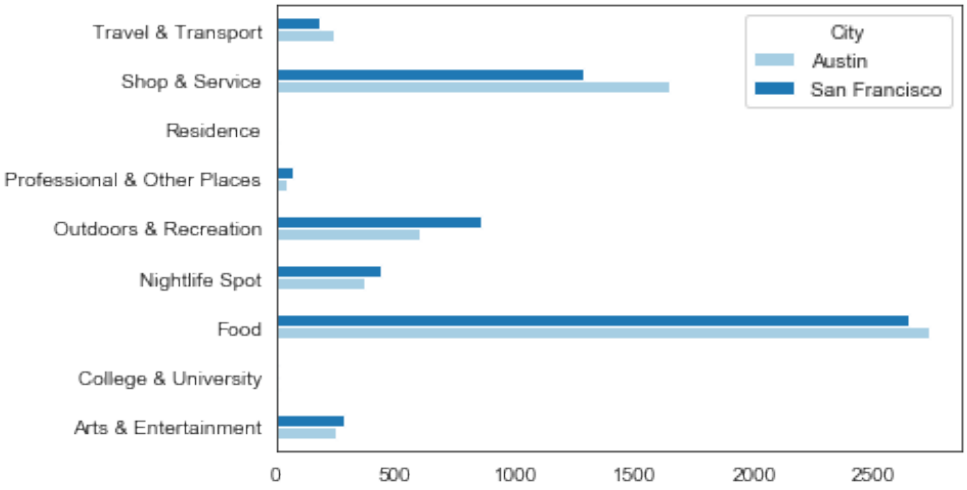
Top-Level Categories
Arts & Entertainment
College & University
Event
Food
Nightlife Spot
Outdoors & Recreation
Professional & Other Places
Residence
Shop & Service
Travel & Transport

3 Exploratory Data Analysis

3.1 City Comparison

There seems to be some similarities between Austin and San Francisco in terms of the different types of venues available in the city. Most notably, the number of food venues are quite similar. It is interesting to see that in both cities food venues are the most prevalent category, followed by shops and services venues. Some differences between the two cities includes the fact that San Francisco has more outdoors and recreation venues, whereas Austin has more shops and services venues.

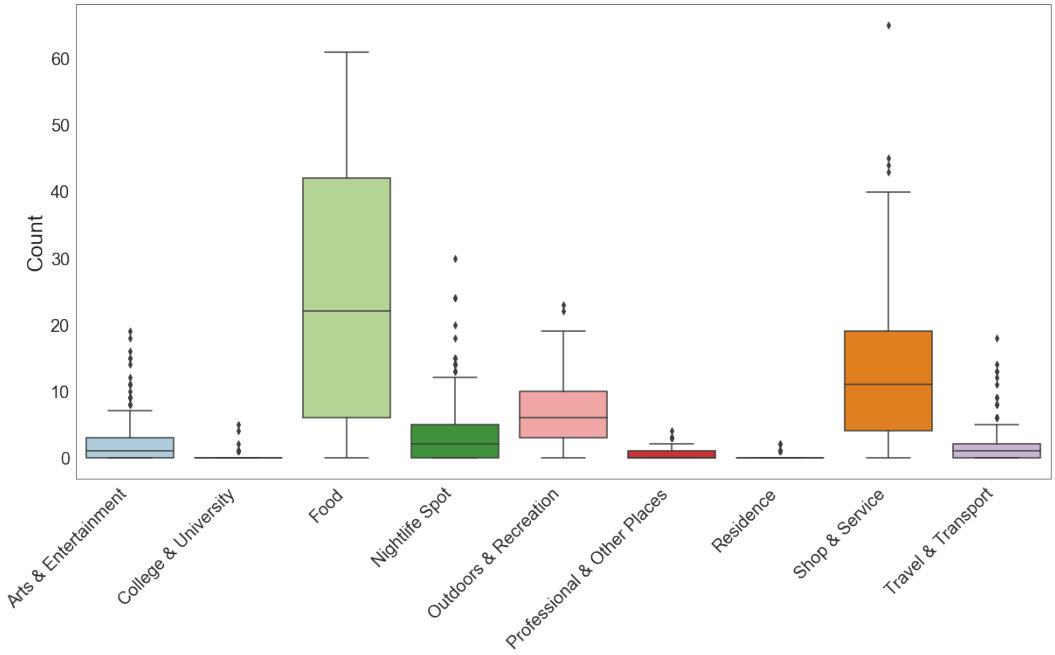
Figure 3: Bar chart of venue categories



3.2 Venue Categories Overview

Looking at the boxplot in Figure 4, there are only 9 different top-level venue categories. This is because there was no venue classified as Event venue in either of the cities. It is apparent that Food venues have the highest median count as well as the largest variability. This is followed by Shop & Service venues. Art & Entertainment, Nightlife Spot, and Travel & Transport venues have the most significant outliers. College & University and Residence venues are omitted from the analysis since the number of neighborhoods with those venues are comparatively low.

Figure 4: Boxplot of Venues Categories



4 Methodology

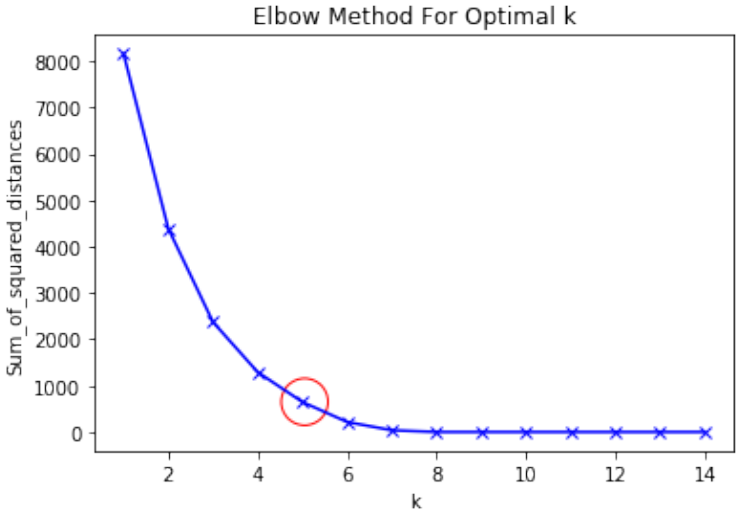
4.1 K-Means Clustering Analysis

In this investigation, the k-means clustering algorithm will be performed to categorize similar neighborhoods into clusters based on the types of venues present in the vicinity of the neighborhoods.

A frequent problem in data clustering is determining the number of clusters (k) in a data set. The correct choice of k is often ambiguous. Increasing the number of clusters improves the ability of the algorithm to explain more of the variation in the data. However, this could result in over-fitting of the data and increase the difficulty of interpreting the result of the clustering.

As a result, the Elbow method is employed to determine the optimal k for k-means clustering algorithm. The optimal k will be chosen where the diminishing returns are no longer worth the additional cost of additional cluster. A plot of sum of squared distances for k in the specified range of ks will be generated. If the plot ends up looking like an arm, then the optimal k is the elbow of the arm. Looking that Figure 5, the elbow occurs at k=5. Therefore, the k-means clustering analysis will be performed using 5 clusters.

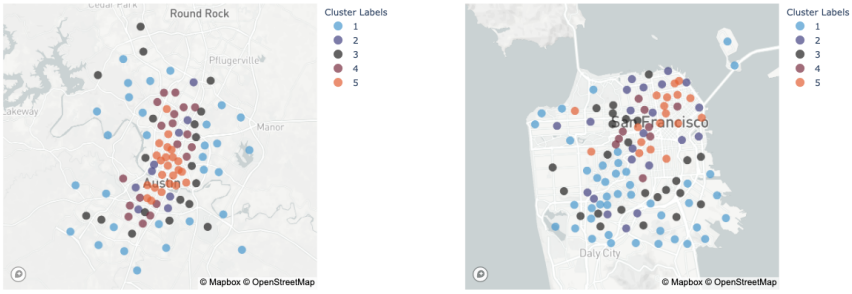
Figure 5: The Elbow Method for determining the optimal k



5 Results

The results of the k-means clustering analysis are shown in Figure 6 and 7. The full interactive version of the maps in Figure 6 are available here: Austin and San Francisco.

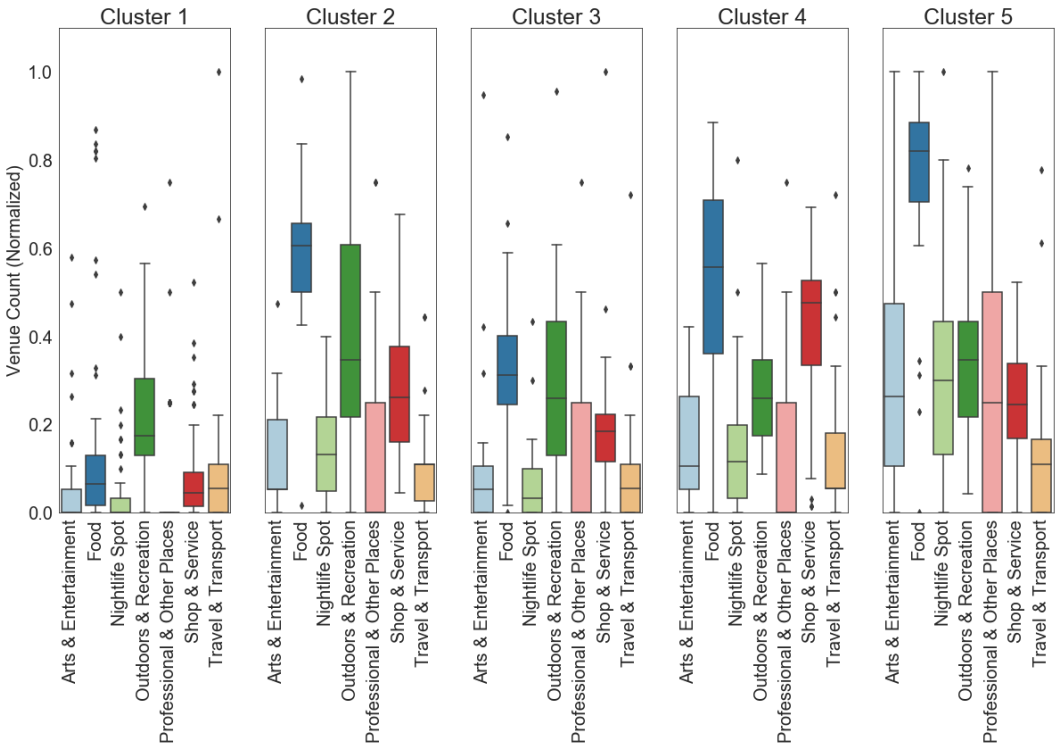
Figure 6: Maps of cluster groups



6 Discussion

Looking at the boxplots and maps above, the cluster groups can be characterized according to which type of venues are most prevalent near the neighborhoods in the two cities.

Figure 7: Boxplot of CLuster Groups



- **Cluster 1** has the lowest number of venues across the board. The neighborhoods in this cluster seem to be areas that are located further away from the city centers; therefore, don't have as many venues as neighborhoods closer to the city centers. It is also worth noting that the most popular venue category in Cluster 1 is Outdoors & Recreation. This could imply that in both cities, living further away from the city center could be best suited for residents that enjoy outdoor activities and don't wish to be disturbed by heavy human traffic in their neighborhoods.

- **Cluster 2** has a high number of Food venues as well as one of the highest number of Outdoors & Recreation venues. This cluster seems to group neighborhoods that have a great balance of amenities and infrastructure. Their geographic locations also seem to suggest they are situated just outside of the main city center. Therefore, this cluster could represent residential neighborhoods with many residents that work in the city. Perhaps, the neighborhoods in this cluster could be a perfect destination for city commuters that like to enjoy a wide variety of venues closer to their homes.

- **Cluster 3** seems to exhibit similar characteristics with a slightly fewer number of venues than Cluster 2 when looking at the boxplots above. However, when looking at the maps, we can see that Cluster 3 neighborhoods tend to be located much further away from the city centers. The fact that Cluster 3 has lower numbers of Food,

Nightlife Spot, and Arts & Entertainment could imply that Cluster 3 neighborhoods are more concentrated with residential housing compared to Cluster 2 neighborhoods.

- **Cluster 4** can be characterized by the high number of Food and Shop & Service venues in the neighborhoods. Looking at the maps, we can see that the neighborhoods are located right outside of the city centers. The neighborhoods in the cluster could be well suited to people that work in the retail business or want to open shops and services after moving to their new neighborhood.

- **Cluster 5** definitely can be characterized as the neighborhoods in the city center, indicated by the high number of venues across almost all the categories. The cluster has the highest count of Food venues as well as Professional & Other Services venues. The map of Austin paints a clear picture of Cluster 5 is the center of the city. In San Francisco's case, many neighborhoods near Union Square have been assigned to Cluster 5. The neighborhoods in this cluster are ideal for workers who don't want to commute longer distances to work and want to enjoy the city lifestyle.

7 Conclusion

The results from the maps of the clusters and the boxplots show us that the k-means clustering algorithm was successful in grouped similar neighborhoods in Austin and San Francisco. The resulting clustering groups can help people make decisions on which neighborhood to move to when moving from either one of the two cities.

7.1 Potential Improvements

Although this investigation has provided us with some useful information, there are many ways in which we can improve our methodology to provide more detailed and accurate analysis for our target audiences.

- The investigation can be advanced by adding more variables that characterize certain neighborhoods, such as demographics, socioeconomic data, crime rate, housing types, and political orientations.

- The extraction method for the venues data from Foursquare API can be improved by adjusting the radius based on the area of each neighborhood.

- The investigation could be extended to include other popular cities like New York City, Seattle, and Boston for more comprehensive results for the target audiences.

- Perhaps using recommendation algorithms to make neighborhood suggestions based on certain neighborhood qualities can be a useful tool for our target audiences.