The Denver Brewery Project

Coursera / IBM Data Science Capstone

1. Introduction

1.1

I am looking to enter the craft beer sector by opening a new brewery in Denver, Colorado. As with any business endeavor, there are a number of unknowns. The purpose of this project is to help identify which neighborhoods in Denver would be the best potential locations for a new brewery business. The target audience for this project will be myself. It will help me make an efficient decision on selecting a location.

The ideal neighborhood for this new business would have a large number of potential customers and few direct competitors. I will look for neighborhoods that already support a number of drinkeries, like pubs, but have no or few existing breweries. I want to avoid opening the new brewery in a location with existing breweries who might already have loyal clients that can't be swayed away. However, neighborhoods that have few drinkeries of any type may not support any similar businesses.

2. Data

2.1 Data Sources

To solve the problem, we used the following data: a list of neighborhoods in Denver, Colorado; latitude and longitude coordinates of those neighorhoods; and venue data to perform clustering on the neighborhoods.

The list of Denver neighborhoods can be extracted from Wikipedia by web scraping and using the beautifulsoup package. The geographical coordinates of the neighborhoods can be determined using Python Geocoder package. Data on neighborhood businesses can be found on FourSquare. FourSquare is a location technology company that gives developers access to their location data. Location data includes details of "venues" - businesses, restaurants, parks, etc in relation to a specified geolocation. This was done using the FourSquare API, and sending a list of neighborhood coordinates and requesting a list of the top 100 venues within 1 km of that location. The json data retrieved from FourSquare are read into a dataframe.

2.2 Data Cleaning

After creating dataframes to store the neighborhood names retrieved using beautifulsoup and the geolocation coordinates using GeoCoder, the two datasets were merged using the neighborhood name to match. One neighborhood (Whittier) was not located using GeoCoder and was dropped using the "dropna" function.

3. Methodology

The data collected as described above was used to build neighborhood profiles of breweries. For the 51 neighborhoods, we retrieved data for a total of 2204 venues. One hot encoding was used to turn the list of Venue Categories into columns with a binary value indicating whether the venue was of that category (0 = it is not of the category, 1 = it is of that category). The venue categories were summed and plotted as a rainbow bar chart to analyze the frequency of venue types.

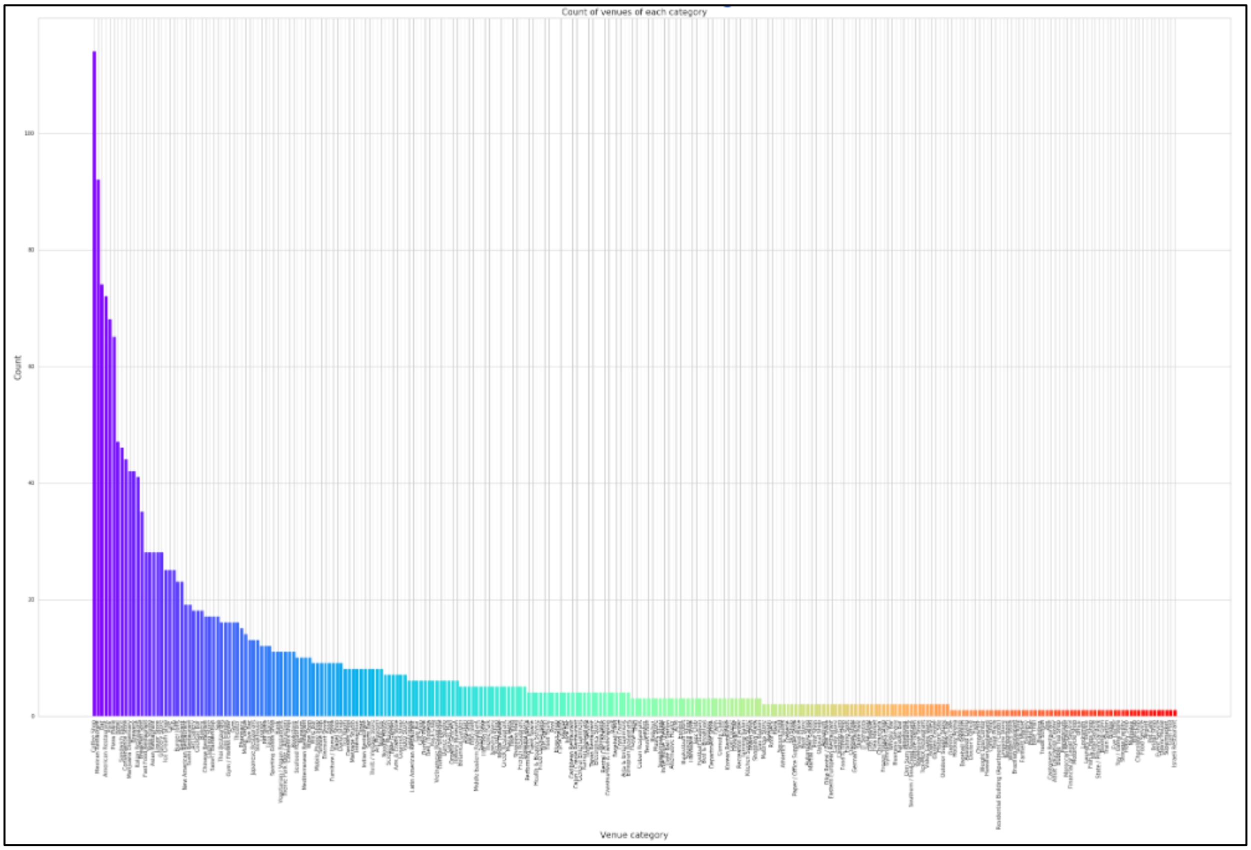

Figure 1. Rainbow Bar Chart of All Venues

The rows of venue data were then grouped by neighborhood, so that each column would show the total number of venues in the neighborhood of that venue category. The numerical data was normalized using scikit-learn's preprocessing module.

Clustering on the data was performed using K-means clustering. K-means clustering algorithm identifies K number of centroids, and then allocated 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 is particularly suited to solve the problem for this project.

The first step in K-means clustering is determining the proper K value to use, as this can have a big impact on the output of the model. The Elbow Method was used to determine the best value for K, specifically the KElbowVisualizer from the yellowbrick package. The elbow value was determined to be 5.

Neighborhoods were first clustered into 5 clusters based only on brewery data. A dataframe was then created containing normalized data for venues in the following categories: brewery, bar, pub, cocktail bar, and wine bar. The additional four venue categories were the most common competing types of venues, as shown in the rainbow bar chart. Neighborhoods were also clustered into 5 clusters based on the five types of venues.

4. Results

4.1 Breweries Only Dataset

Using the coordinates within the data, the Folium package was used to plot the neighborhoods on a map, using color to differentiate between the five clusters.

The clusters showed fewer numbers of neighborhoods in the clusters with the higher number of breweries. Cluster 1, which had no breweries, contained 22 neighborhoods. Clusters 2 and 3, with the highest number of breweries, contained 4 and 1 neighborhoods, respectively. This indicates that existing breweries are concentrated in a relatively small number of neighborhoods.

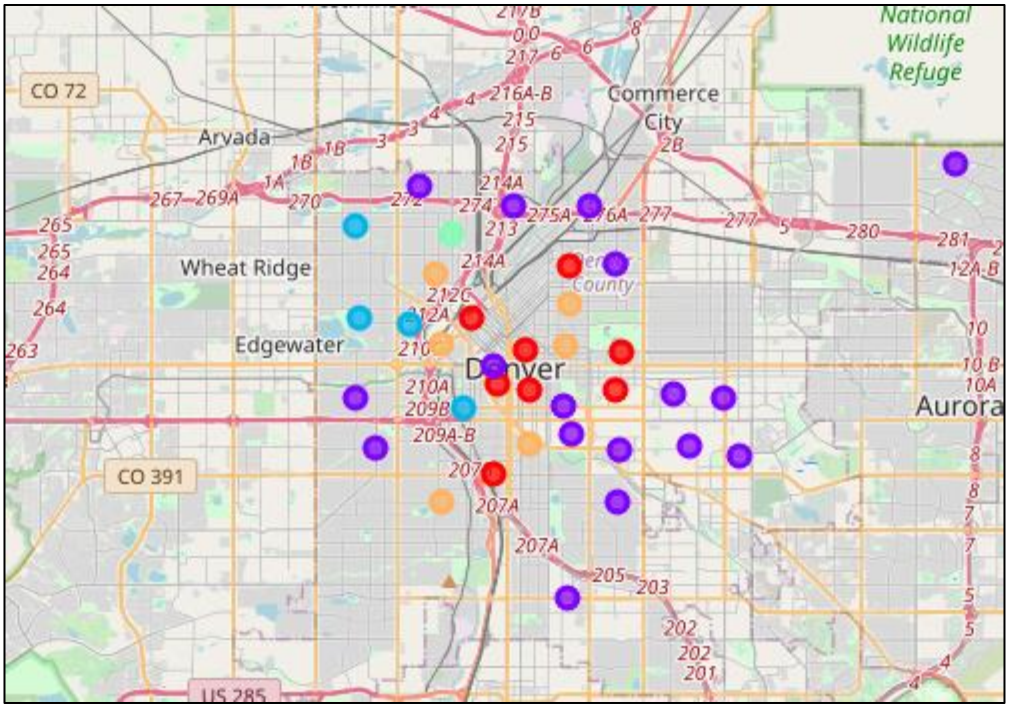

Figure 2. Neighborhood Clusters Based on Brewery Venues

4.2 All Drinkeries Dataset

The clusters for the five venue categories was similarly plotted using Folium

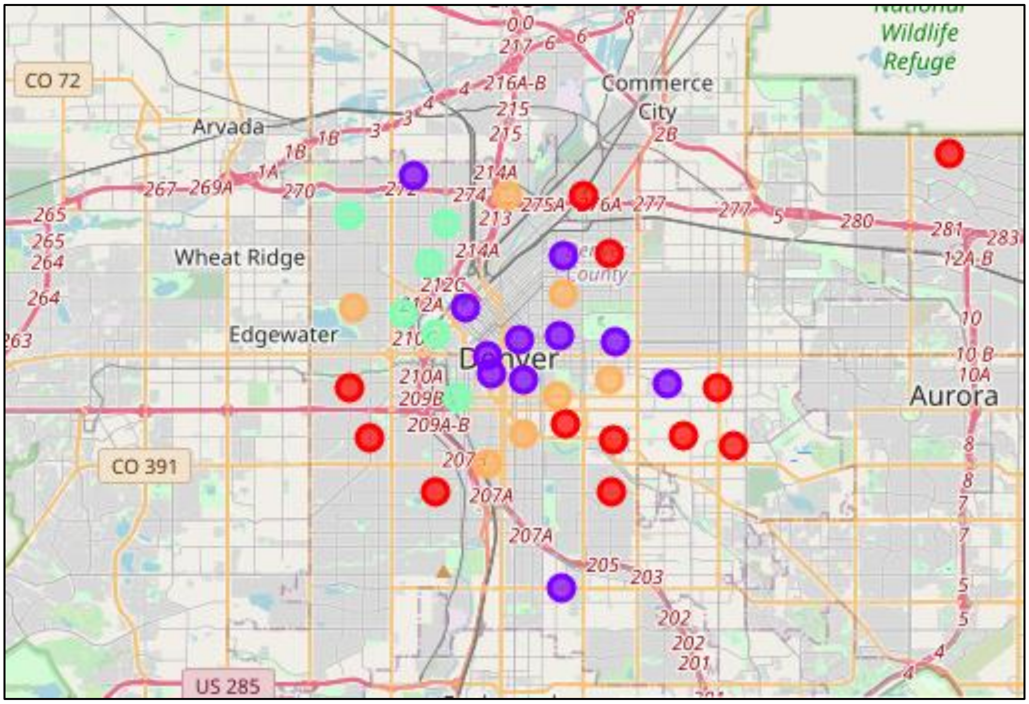

Figure 3. Neighborhood Clusters Based on Multiple Venues

Neighborhoods in Cluster 0 (red) and Cluster 1 (purple) have a very low or low number, respectively, of any drinking venues. Cluster 2 only contained one neighborhood (Capital Hill), which only contained Pub venues from the five categories considered. Cluster 3 (green) consists of neighborhoods with a high number of existing breweries and cocktail bars, and a moderate amount of bars. Cluster 4 (orange) has a low to medium number of existing breweries, and a medium to high number of other drinking venues, such as bars.

5. Discussion

Cluster 4 from the multiple-venue dataset identified the types of neighborhoods we were most interested in: low number of direct competitors (existing breweries) and a medium to high number of related businesses. This cluster consists of seven neighborhoods: Auraria, Ruby Hill, Sloan Lake, Elyria-Swansea, Chaffee Park, Cole, and West Colfax. Of these seven neighborhoods, two (Elryia-Swansea and Chaffee Park) have no breweries identified in our FourSquare venue data. The neighborhoods identified in this cluster will be prioritized for further research in identifying our ideal brewery location.

One interesting observation from our clusters, is the correlation between breweries and cocktail bars. All neighborhoods in our target neighborhood cluster (Cluster 4) have no identified cocktail bars. While Cluster 3, with a high number of breweries, has several neighborhoods with a moderate to high amount of cocktail bars.

Incorporating additional data into our modeling could help determine market opportunities. A population dataset including data such as age, gender, and income could be combined with venue data to predict groups of customers that prefer the venue type we want to open and help identify neighborhoods that contain a large group of under-served customers.

6. Conclusions

From the clustering analysis done in this report, it is likely that a neighborhood within Cluster 4 would be the most suitable location for a new brewery. A next step would be to take the neighborhoods of Cluster 4 and find more data on the area, possibly economic data, and perform another cluster analysis to break down the cluster even further. The findings of this project will help the relevant stakeholders to identify and capitalize on the opportunities of locations with high potential.