Food Trucks and Neighborhood Clustering: A Decision Tool

Capstone Project – Battle of the Neighborhoods (Part 2 – Week 5)
1.3.21

1. Introduction

Background
The restaurant industry represents 4% of the U.S. GDP and employees one in 10 Americans. The industry is fast-moving with ebbs and flows and low profit margins. A subset of this sector is food trucks. Food trucks have advantages to the traditional brick and mortar store front. First, they are mobile and allow operators to move to customers to increase their visibility and foot traffic. With the explosion of social media this allows repeat customers to easily locate the business and increase exposer to new customers. Next, their limited food production ability streamlines the meus, and lowers overhead costs. Additional, owning their “location”, truck, insulates owners against increasing rent and utilities combing to increase operator profit margins. Even with these advantages expanding into a new market brings high risk. Therefore, identifying locations that have similar characteristics to success cities with food truck businesses will reduce the risk of starting or expanding a business.

Data that can determine if a location is optimal for an enterprise includes nightlife, sporting events, outdoor recreation, university, and college venues. The purpose of this project is to predict an optimal location to begin a food truck business.

This project targets restauranters interested in starting a franchise, expanding their business, or start a business. Secondary targets are outside investors interested in reducing risk of capital investments in the restaurant sector.

2. Data

Data Sources
Two cities, Denver, CO (control) and Jacksonville, FL (target), will be used to identify similarities in the composition of the neighborhood venues. Denver, CO neighborhood names are sourced from denvergov.org (https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-statistical-neighborhoods) and Jacksonville, FL scraped from Wikipedia Neighborhoods of Jacksonville (https://en.wikipedia.org/wiki/Neighborhoods_of_Jacksonville). City geolocation data from Positionstack.com will include latitude and longitude of the city neighborhoods. From this geolocation data the venue composition data will be obtained from Foursquare. Venue categories will include Arts and Entertainment, Music and Sports, College and University, Nightlife, and Outdoors and Recreation. In addition, food trucks will be identified in Denver neighborhoods with Foursquare.

3. Methodology

3.1 Data Cleaning

Data was downloaded and scraped from two sources. For Jacksonville, the neighborhood names were scrape and required extensive cleaning to get a comprehensive list of names. Denver was directly downloaded. With the neighborhoods names geolocations where obtained from Positionstack. Positionstack was found to be unreliable so five calls were placed for each neighborhood name. Because of the multiple calls extra venues, blank returns, and unneeded columns were cleaned from the data frames.

3.2 Foursquare Venue Selection

The venue categories from Foursquare were Arts and Entertainment, Music and Sports, College and University, Nightlife, and Outdoors and Recreation. These were chosen because they are key sources of customers and supply additional foot traffic to increase business exposer. It was also decided to set a radius of two miles around each neighborhood to collect a comprehensive picture of each neighborhood. An additional dataset was obtained from Foursquare which was the number of food trucks associated with each Denver neighborhood. It was used to identify the clusters that had the highest density of food trucks. A heatmap was used to visualize the data.

3.3 K-means Clustering

K-means clustering was chosen for it's quick, robust, and reliable algorithm. The algorithm is also easy to code and results can be easily interrupted. Venue datasets were combined, and clusters determined. City cluster distribution was then calculated.

4. Results

4.1 Denver

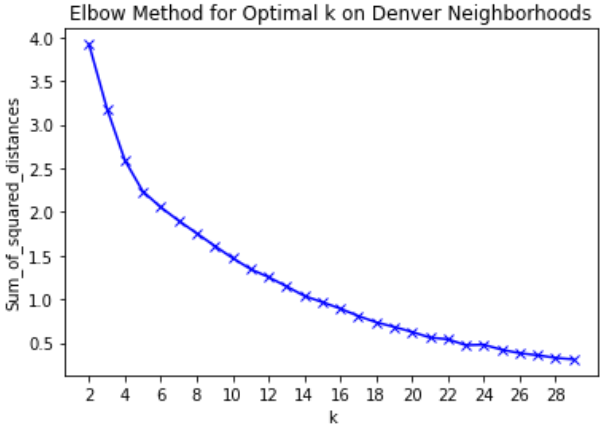
The Denver dataset was used as a psedo-test set. The frequency of venues was not part of the original dataset. The venue categories were one-hot coded and averaged to generate the data frame for clustering.

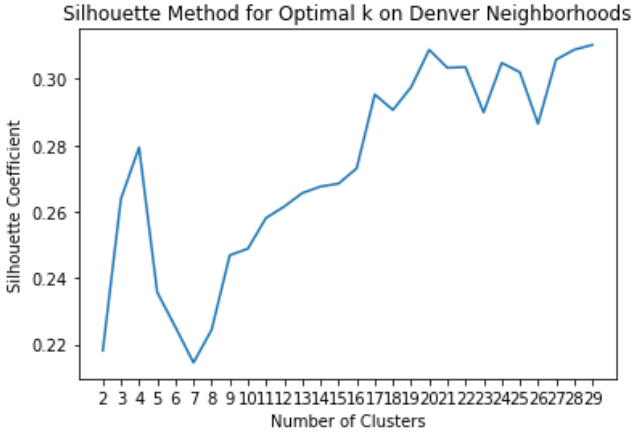
Out[34]:

	Neighborhood	Art Gallery	Art Museum	Arts & Entertainment	Circus	Comedy Club	Concert Hall	Country Dance Club	Dance Studio	Disc Golf	...	Piano Bar	Planetarium	Public Art	R C
0	Auraria	0.170000	0.050000	0.030000	0.0	0.040000	0.040000	0.000000	0.040000	0.000000	...	0.010000	0.0	0.020000	0.040
1	Baker	0.209302	0.046512	0.023256	0.0	0.011628	0.046512	0.000000	0.058140	0.000000	...	0.000000	0.0	0.034884	0.023
2	Barnum	0.451613	0.032258	0.032258	0.0	0.032258	0.000000	0.000000	0.161290	0.032258	...	0.032258	0.0	0.000000	0.000
3	Barnum West	0.235294	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.058824	0.000000	...	0.000000	0.0	0.000000	0.000
4	Bear Valley	0.111111	0.000000	0.000000	0.0	0.000000	0.000000	0.111111	0.333333	0.111111	...	0.000000	0.0	0.000000	0.000

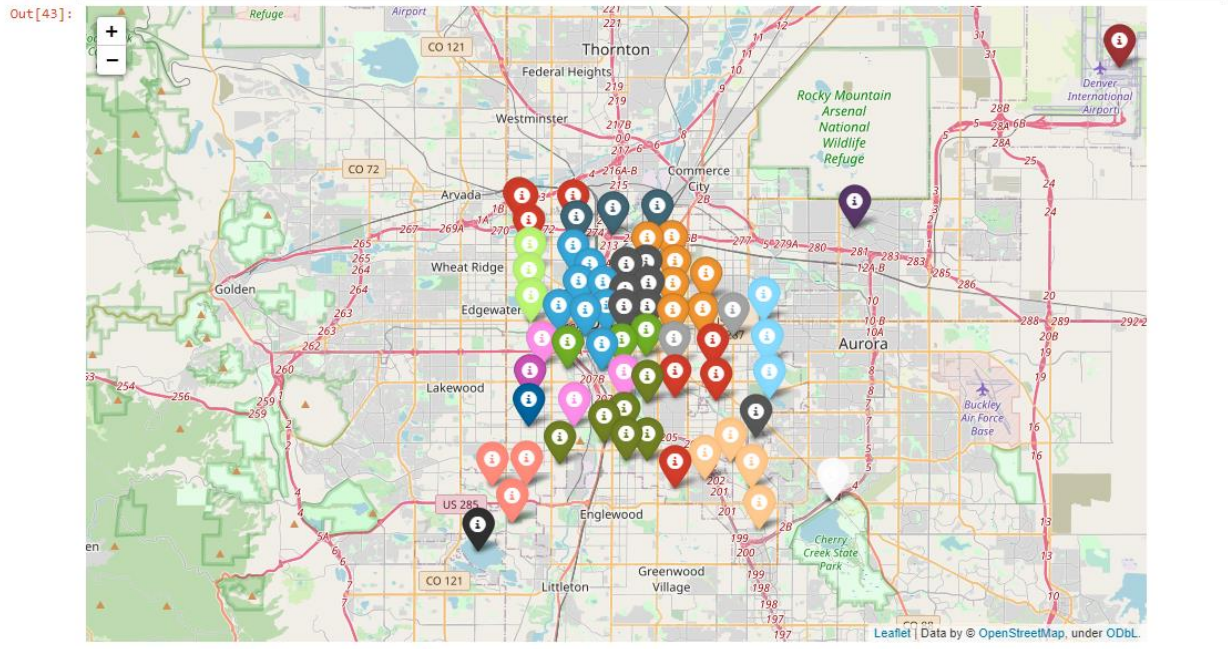
5 rows × 37 columns

To determine the optimal number of clusters the Elbow Method was used. There is no clear “elbow” and the Silhouette Method was also used.



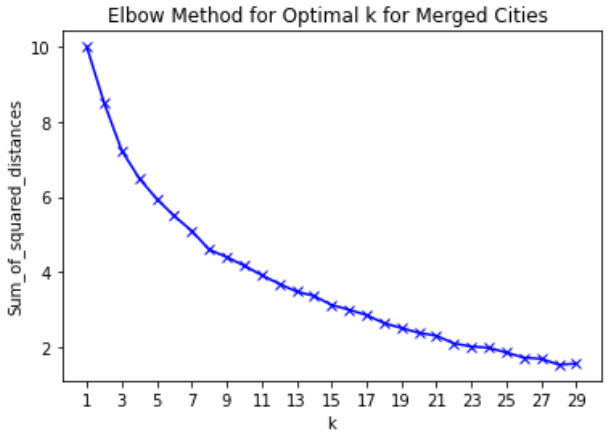


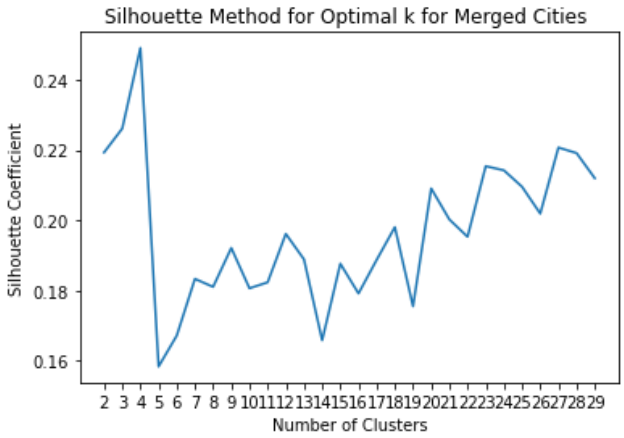
The Silhouette Coefficient is low however a k of 20 was chosen. The clusters were visualized on the Denver map.



4.2 Merged City

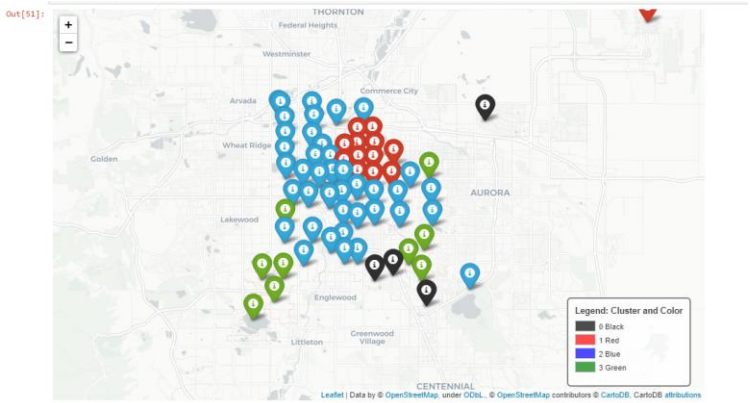
The procedure was repeated with the merged venue dataset. Again, both k optimization methods didn’t result in a clear choice.



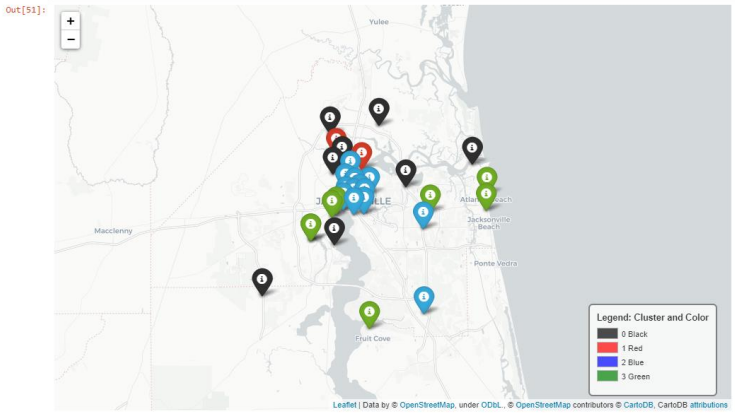


A k of 4 was chosen. The clusters were mapped to the cities.

Denver



Jacksonville

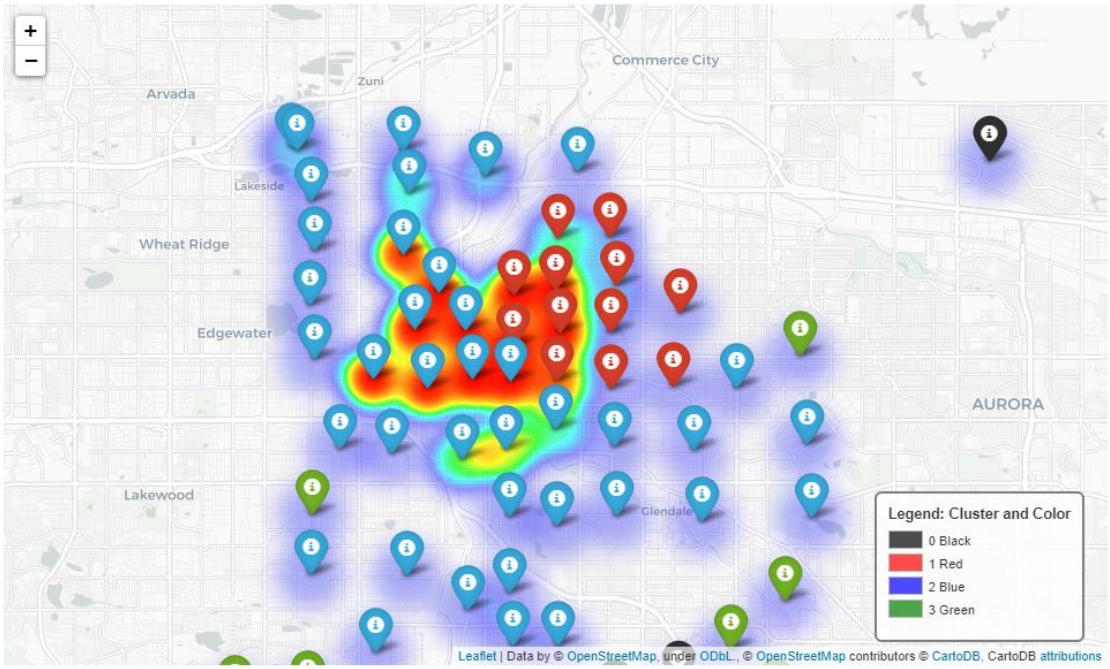


Examining the distribution of the clusters we see overlay of dominate clusters.

	Cluster Labels	Denver	Jacksonville	Percent_Denver	Percent_Jacksonville	Difference
0	0	4	8	6	26	-20
1	1	13	2	20	6	14
2	2	40	13	61	42	19
3	3	9	8	14	26	-12

Finally, the number of food trucks in Denver was mapped to the neighborhoods and it was found that clusters 1 and 2 were associated with them.

Out[56]:



5. Discussion

Through our analysis we were able to cluster the Denver and Jacksonville neighborhood venue datasets for use in determining similar characteristics within neighborhoods. Beginning with Denver as a psedo-test set, we used the K-means algorithm to cluster the neighborhoods by venue categories. The categories for this analysis were chosen for their ideal settings for food trucks. A k of 20 was determined optimal with the Elbow and Silhouette Methods, however, both methods were not definitive. The Elbow Method did not have a clear break point and the exponential decay curve had a steep linear decline. Additionally, the Silhouette Coefficient was low suggesting that little to no structure in the dataset. There was greater confidence in the K-means algorithm after the clusters were mapped. Based on geolocation and an understanding of the layout of Denver the clusters appeared to have some validity.

Merging the Denver and Jacksonville datasets and utilizing the K-means algorithm had similar problems as with the Denver only dataset. Both the Elbow and Silhouette Methods showed little

improvement in defining clusters. One reason for this may be the number of neighborhoods was too large with many different venue categories leading to high dimensionality, which K-means has problems with. Using principal component analysis to reduce the dimensions prior to clustering may improve results but was not performed here. A k of four was chosen as the optimal number of clusters. Visually the two cities appeared to be dominated by the same clusters. The majority of neighborhoods in both cities were in cluster 2. The greatest difference between the cities was observed in clusters 0 and 1. Denver lacking cluster 0 and Jacksonville lacking cluster 1. Finally, we overlay a density heatmap of numbers of food trucks in Denver. It shows that food trucks are associated with clusters 1 and 2. The strongest with cluster 2. We can see that cluster 2 dominates Jacksonville as well. Interestingly, it was found that every Denver neighborhood had at least a food truck associated with it.

Potential pitfalls in this analysis was that K-means clustering algorithm may not be the ideal choice. In both the merged and Denver datasets the cluster number optimizing methods demonstrated a lack of structure. Improving K-means clustering was proposed above, however, a density-based or t-SNE clustering may improve detection of irregular cluster shapes or address higher dimensional analysis.

5. Conclusion

The purpose of this project was to assist stakeholders in deciding to open or expand a food truck enterprise in new locations. Using Foursquare data on venue types this analysis was able to cluster neighborhoods in Denver, a thriving food truck culture, with neighborhoods in Jacksonville. Common clusters within both cities should be used as a starting point to determine if the target location is an ideal choice. Additional factors that may be considered in future analysis are population density, social-economics status, and local permit costs which were outside the scope of this project.