Applied Data Science Capstone Project: Denver Neighborhood Ramen Restaurant

Author: Benjamin Melhorn

Date: 6/20/2020

Contents

1. Introduction/Business Problem

Denver is a young and expanding city with plenty of opportunities for new businesses. An entrepreneur in Denver is planning to start a chain of ramen bars and is interested in starting the chain in Denver. The entrepreneur understands the type of person mostly likely to frequent ramen restaurants and needs to identify the neighborhoods in Denver that contain this target demographic. In addition, the entrepreneur wants to limit direct competition and avoid neighborhoods that may already be saturated with ramen bars or other Japanese restaurants.

The goal of this analysis is to identify the top neighborhoods in Denver for the entrepreneur to launch their ramen bar chain. The neighborhoods must have a relatively high percentage of residents that fit the demographic the entrepreneur is targeting. In addition, desirable neighborhoods will have plenty of demand for a new ramen restaurant. The entrepreneur has identified the type of customer to target for their ramen bars. This customer is young or middle aged. They may or may not live alone, but they generally do not have kids or a family. This customer also lives and works in or near the same neighborhood and generally does not commute far from their neighborhood. The entrepreneur would also like to identify new markets where there are a low number of ramen bars or restaurants that will provide a similar experience.

2. Data

Solving this problem requires the following data:

A. Demographic and general data for Denver neighborhoods: Age, Living Situation, Marital Status, Commute Time, and Total Population

B. Location data for Denver neighborhoods

C. Venue location and category data for Denver

A. Demographic

The Denver neighborhood demographics data will be obtained from the City and County of Denver - American Community Survey Nbrhd (2014-2018), which can be found here: https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-american-community-survey-nbrhd-2014-2018

B. Location

The neighborhood location data will be acquired using Geopy.

C. Venue

The neighborhood venue information will be obtained using Foursquare.

3. Methodology

To solve this problem, I started by gathering the neighborhood demographic data and using it to create additional neighborhood features. These new features aligned with the features of the target demographic. I then used these features as the basis to group the different neighborhoods using K-means clustering. After running K-means with 2 to 8 clusters, I found that creating 6 clusters returned the most distinct results. From these results I selected the cluster that had the most favorable overall demographics. Next, I gathered the location and venue data for each neighborhood in the desired cluster. I limited venue data to only include Japanese restaurants. I selected this category, instead of just Ramen restaurant, to ensure that restaurants that may have ramen on their menus, but that are not solely Ramen restaurants, were considered. For the final two steps, I calculated the number of venues per capita for each neighborhood. Then created a map with each neighborhood and venue labeled.

Steps:

1. Gather Data
2. Calculate and Add Neighborhood Features
3. Run K-Means Clustering
4. Gather Neighborhood Location and Venue Data
5. Calculate and Add Venues Per Capita
6. Generate Map Showing Neighborhood and Venue Location

Gather Data

I first gathered the demographic data from the city of Denver government website. Then specified the desired columns and added them to a pandas dataframe.

	NBHD_NAME	TTL_POPULATION_ALL	AGE_20_TO_29	AGE_30_TO_39	AGE_40_TO_49	TOTAL_COMMUTERS	COMMUTE_LESS_15	NONFAMILY_HOUSEHOLD
0	Bear Valley	9247.0	1745.0	1289.0	1231.0	4710.0	647.0	1552
1	Harvey Park South	9410.0	1257.0	1313.0	956.0	4165.0	904.0	1368
2	Southmoor Park	5505.0	1729.0	1027.0	696.0	3611.0	884.0	1999
3	Hampden South	16259.0	2635.0	2547.0	2590.0	8738.0	2093.0	4034
4	Goldsmith	6045.0	1085.0	938.0	813.0	3121.0	477.0	1585

Calculate and Add Neighborhood Features

I used the neighborhood demographics data to calculate additional neighborhood features. These features are:

Percent of Population Ages 20 to 49 = Sum of Age Groups / Neighborhood Population

Percent of Population in a NonFamily Household = # in NonFamily / Neighborhood Population

Percent of Population with no Commute or a Commute that is 15 Minutes or Less = (No Commute + Commute Less 15) / Neighborhood Population

I defined two functions to calculate these features and add them to the existing dataframe. One that just required dividing by neighborhood population and another that required dividing the sum of multiple columns by neighborhood population.

I then used the functions to calculate Percent of Population in NonFamily Household and Percent of Population Ages 20 to 49.

For each neighborhood, I then calculated the number of individuals with no commute or a commute less than 15 minutes and added this to the existing dataframe.

	NBHD_NAME	TTL_POPULATION_ALL	AGE_20_TO_29	AGE_30_TO_39	AGE_40_TO_49	TOTAL_COMMUTERS	COMMUTE_LESS_15	NONFAMILY_HOUSEHOLD	PCT_NONFAMILY	PCT_20_TO_49	PCT_NO_OR_SHORT_COMMUTE
0	Bear Valley	9247.0	1745.0	1289.0	1231.0	4710.0	647.0	1552	0.167838	0.461231	0.560614
1	Harvey Park South	9410.0	1257.0	1313.0	956.0	4165.0	904.0	1368	0.145377	0.374708	0.653454
2	Southmoor Park	5505.0	1729.0	1027.0	696.0	3611.0	884.0	1999	0.363124	0.627066	0.504632
3	Hampden South	16259.0	2635.0	2547.0	2590.0	8738.0	2093.0	4034	0.248109	0.478012	0.591303
4	Goldsmith	6045.0	1085.0	938.0	813.0	3121.0	477.0	1585	0.262200	0.469148	0.562614

Run K-Means Clustering

I then used the new features to group the neighborhoods using K-means clustering. The first step is to create a dataframe with just the neighborhoods and their features. Then run K-means clustering and analyze the results.

Labels	PCT_20_TO_49	PCT_NONFAMILY	PCT_NO_OR_SHORT_COMMUTE
0	0.635898	0.329560	0.526592
1	0.485125	0.241640	0.638266
2	0.428585	0.109433	0.645660
3	0.677444	0.502774	0.515763
4	0.272889	0.069535	0.826480
5	0.512087	0.227569	0.546622

Of the clusters, it looks like cluster 3 is our desired cluster. Nearly 70% of its population is in the target demographic's age range and half of its population lives in a non-family household. The population of cluster 3 does not commute the least out of all the clusters, but its age and household demographic make up for this.

Gather Neighborhood Location and Venue Data

The next step is to gather location and venue data for just the neighborhoods in the desired cluster. The first step is to create a dataframe containing this cluster's data. I used geolocator to gather the neighborhood location data and then foursquare to gather the neighborhood venue data.

Desired Cluster Data:

	NBHD_NAME	TTL_POPULATION_ALL	PCT_20_TO_49	PCT_NONFAMILY	PCT_NO_OR_SHORT_COMMUTE	Labels
22	Jefferson Park	3165.0	0.774724	0.409795	0.441706	3
30	Cheesman Park	8998.0	0.633141	0.451656	0.483774	3
45	Union Station	6523.0	0.590986	0.506362	0.542695	3
69	Speer	11715.0	0.679300	0.501152	0.489799	3
74	Capitol Hill	16100.0	0.759627	0.604224	0.476957	3
75	North Capitol Hill	6360.0	0.725000	0.559591	0.483333	3
76	Civic Center	2202.0	0.569028	0.463669	0.603088	3
77	CBD	4253.0	0.687750	0.525747	0.604750	3

First 30 Rows of Venue Data:

	NBHD_NAME	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
0	Union Station	39.753630	-105.000748	Gyu-Kaku Japanese BBQ	39.755023	-105.001247
1	Union Station	39.753630	-105.000748	Blue Sushi Sake Grill	39.751519	-105.000317
2	Union Station	39.753630	-105.000748	Tokio	39.758531	-104.997483
3	Union Station	39.753630	-105.000748	Menya Noodle Bar	39.754866	-105.004671
4	Union Station	39.753630	-105.000748	Hapa Sushi	39.749680	-104.999860
5	Union Station	39.753630	-105.000748	Sakura House	39.751844	-104.993422
6	Union Station	39.753630	-105.000748	Noodles & Company	39.750139	-104.998999
7	Union Station	39.753630	-105.000748	Tom Tom Room	39.748241	-104.999992
8	Union Station	39.753630	-105.000748	Komotodo Sushi Burrito	39.748302	-104.998262
9	Speer	39.752540	-105.006965	Sushi Sasa	39.756701	-105.009367
10	Speer	39.752540	-105.006965	Gyu-Kaku Japanese BBQ	39.755023	-105.001247
11	Speer	39.752540	-105.006965	Menya Noodle Bar	39.754866	-105.004671
12	Speer	39.752540	-105.006965	Blue Sushi Sake Grill	39.751519	-105.000317
13	Speer	39.752540	-105.006965	Hapa Sushi	39.749680	-104.999860
14	Speer	39.752540	-105.006965	Noodles & Company	39.750139	-104.998999
15	Speer	39.752540	-105.006965	Sushi Han	39.748665	-105.007599
16	Speer	39.752540	-105.006965	Sushi King	39.748588	-105.007645
17	Speer	39.752540	-105.006965	Tai Tai Japanese Hawaiian	39.757061	-105.009392
18	Capitol Hill	39.735875	-104.979921	Tokyo Joe's	39.737657	-104.983505
19	Capitol Hill	39.735875	-104.979921	Tycoon Ramen & Sushi Bar	39.739958	-104.982459
20	Capitol Hill	39.735875	-104.979921	Imperial Village	39.736591	-104.984459
21	Capitol Hill	39.735875	-104.979921	Ikano Bowl	39.739673	-104.977948
22	Capitol Hill	39.735875	-104.979921	Katana Sushi	39.731372	-104.986137
23	Capitol Hill	39.735875	-104.979921	Kyu Ramen	39.739777	-104.979591
24	Capitol Hill	39.735875	-104.979921	Menya	39.739826	-104.979750
25	North Capitol Hill	39.745624	-104.981598	Tycoon Ramen & Sushi Bar	39.739958	-104.982459
26	North Capitol Hill	39.745624	-104.981598	Noodles & Company	39.742760	-104.988840
27	North Capitol Hill	39.745624	-104.981598	Tokyo Express	39.743500	-104.988645
28	North Capitol Hill	39.745624	-104.981598	Ikano Bowl	39.739673	-104.977948
29	North Capitol Hill	39.745624	-104.981598	Sabi Sushi - WTC Dropzone	39.742018	-104.987916
30	North Capitol Hill	39.745624	-104.981598	Menya	39.739826	-104.979750

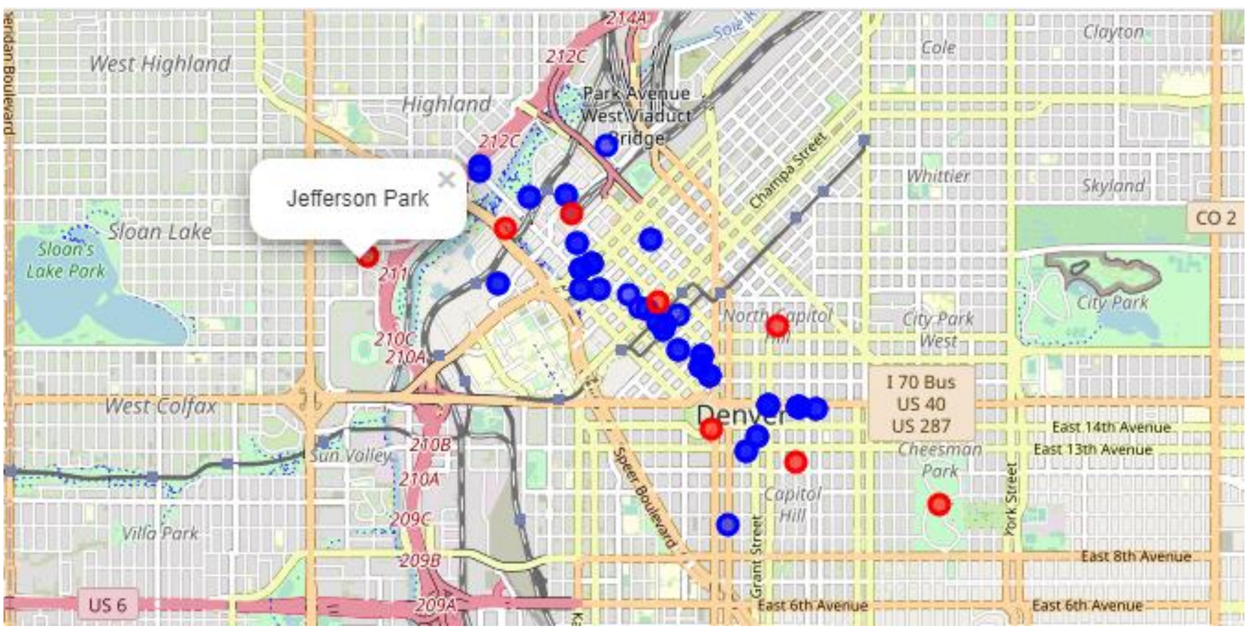
Calculate and Add Venues Per Capita

After gathering each neighborhood's venue information, I calculated the number of venues per capita for each neighborhood. Two of the neighborhoods in the cluster do not have any Japanese venues. This required changing NaN to zero in order to calculate venues per capita.

	NBHD_NAME	TTL_POPULATION_ALL	PCT_20_TO_49	PCT_NONFAMILY	PCT_NO_OR_SHORT_COMMUTE	Labels	Latitude	Longitude	Venue_Count	Venue_Per
0	Jefferson Park	3165.0	0.774724	0.409795	0.441706	3	39.750621	-105.019779	0.0	0.000000
1	Cheesman Park	8998.0	0.633141	0.451656	0.483774	3	39.732814	-104.966455	0.0	0.000000
2	Union Station	6523.0	0.590986	0.506362	0.542695	3	39.753630	-105.000748	9.0	0.001380
3	Speer	11715.0	0.679300	0.501152	0.489799	3	39.752540	-105.006965	9.0	0.000768
4	Capitol Hill	16100.0	0.759627	0.604224	0.476957	3	39.735875	-104.979921	7.0	0.000435
5	North Capitol Hill	6360.0	0.725000	0.559591	0.483333	3	39.745624	-104.981598	7.0	0.001101
6	Civic Center	2202.0	0.569028	0.463669	0.603088	3	39.738181	-104.987744	9.0	0.004087
7	Central Business District	4253.0	0.687750	0.525747	0.604750	3	39.747378	-104.992737	16.0	0.003762

Generate Map Showing Neighborhood and Venue Location

The final step I took was to generate a map of Denver with a marker and label for each neighborhood from the desired cluster and each of the venues within those neighborhoods. The red markers are the neighborhood centers and the blue markers are the venues.



4. Analysis

The analysis has two main parts. First, identify neighborhoods with a significant population of the target demographic. Second, assess the existing restaurant market in the identified neighborhoods to find the neighborhoods with the most favorable market for opening a new restaurant.

Neighborhood Demographics

The entrepreneur's target customers are young to middle aged (20-49), no family, and lives in, or near, the same neighborhood where they work. I created features for each to group the neighborhoods based on these characteristics:

> Percent of Population Ages 20 to 49 = Sum of Age Groups / Neighborhood Population
> Percent of Population in a NonFamily Household = # in NonFamily / Neighborhood Population
> Percent of Population with no Commute or a Commute that is 15 Minutes or Less = (No Commute + Commute Less 15) / Neighborhood Population

Using K-means clustering, I grouped the Denver neighborhoods into six clusters. Here are the results of that clustering: Run K-Means Clustering Results

From this we can see two of the clusters have greater than 60% of their population between the targeted ages. Of those two, one has a population that 50% non-family. For this reason, this is the cluster I selected to continue with the analysis. The commute feature turned out to be less distinct between the clusters with a range of 51% to 71% of the population having no or a short commute. In addition, the clusters with the largest population of no/short commuters, is also the cluster with the lowest percentage of nonfamily and desired age individuals. Suggesting that those who aren't commuting are likely not within the targeted group.

Neighborhood Restaurant Market

I next calculated Japanese restaurant per capita to assess the desirability of the restaurant market in each neighborhood in the selected cluster. Venues Per Capita

We can see that two neighborhoods don't have any Japanese restaurants, Jefferson Park and Cheesman Park. There are also two neighborhoods that have relatively low per capita rates, Speer and Capital Hill.

Neighborhood and Restaurant Map

Finally, I generated a map showing the location of each restaurant and the center of each desired neighborhood. From this we can see that the restaurants are primarily grouped in along the downtown area and slight south east of downtown. We see the Jefferson Park and Cheesman do not have any Japanese restaurants near their centers. In addition, we can see that North Capitol Hill, which has a high restaurant per capita, has all its restaurants to the far west of its center. It could be a good candidate for a new venue to far east of its center. A similar scenario is true for Civic Center, which does not have any Japanese restaurants south west of its center.

5. Results and Discussion

Two neighborhoods standout as ideal candidates to open a Ramen Bar: Jefferson Park and Cheesman Park. These two neighborhoods have a population that meets the target demographic and do not have any restaurants near their centers that would directly compete with a Ramen Bar. The analysis also identified three neighborhoods for additional exploration: Speer, Capitol Hill, and North Capitol Hill. These neighborhoods contain the target demographic. They already contain restaurants that may compete with a Ramen Bar, however, these are grouped to one side and the opposite side of the neighborhood may prove to be a desirable location to open.

6. Conclusion

The exploration of Denver neighborhood data and venue information identified two neighborhoods that are very strong candidates for a new Ramen Bar, Jefferson Park and Cheesman Park. In addition, it identified two three neighborhoods for additional exploration. These finds are based on the percentage of the population that meets the entrepreneur's target demographic and competitiveness of the restaurant marked.