An Exploration of Cuisine in Denver, Colorado

Georges Brunet

February 10th, 2020

1. Introduction

1.1 Background

According to the U.S. Census, the city of Denver in Colorado is the nineteenth most populous city in the U.S. as of 2010 and the largest municipality in the state of Colorado with an estimated population of 716,492 in 2018. It has experienced a population growth of 19.38% since 2010 and makes it one of the fastest-growing major cities in the country.

The city of Denver officially consists of 78 neighborhoods. Among these neighborhoods, there are many places to eat and that number will only continue to grow as the city expands. In any growing city, the vitality of a neighborhood depends on the venues and attractions that it holds and so in order to optimize growth, predicting good business placement is key.

1.2 Problem

Like any major city, Denver has a good number of neighborhoods, each with varying levels of foot traffic and frequentation. This makes it so certain neighborhoods will inevitably have greater numbers of venues, notably restaurants. There are neighborhoods that are known for their quality and quantity of restaurants, but as the city continues to grow, this statistic is set to change, which begs interesting questions. What types of cuisine are more prevalent among which neighborhoods? Are there neighborhoods that hold considerably more food places that specialize in certain cuisines? And if so, where would be good places to open a restaurant of that type?

This project aims to predict which neighborhoods in Denver would be favorable spots to open restaurants of a specific type of cuisine based on the prevalence or lack of competition of that type of cuisine in that neighborhood. The type of restaurant will be decided on the prevalence of the type among all neighborhoods of Denver. The information on restaurants will be filtered repeatedly until it is confirmed that a type is dominant, which translates into popularity. The neighborhoods will be picked as favorable spots on three conditions. One, that the most popular type is not the most common type in that neighborhood. Two, that the neighborhoods have enough restaurants in them to be considered busy enough and so good candidates. Three, that the neighborhoods have enough diversity such that the certain type does not represent more than 12.5% of restaurant types in the area.

1.3 Interest

The project would provide actionable insights as to where that certain type of cuisine could thrive and in which neighborhood or neighborhoods. The insights can be leveraged for financial gain and potential business ventures. It can also provide insights for better quality of life, such as which neighborhood to go to for a certain type of cuisine. This project is mainly aimed towards aiding stakeholders or aspiring restaurant business owners that are interested in opening restaurants of a certain type in Denver. Others who live in Denver or those who may potentially want to live in the city could be interested in knowing what types of restaurants may or may not appear in what neighborhoods.

2. Data Acquisition and Cleaning

2.1 Data Sources

A csv file containing the official names of the neighborhoods of Denver as well as their identification number can be found on: https://www.denvergov.org/media/gis/DataCatalog/statistical_neighborhoods/csv/statistical_neighborhoods.csv.

This file, though, contains two additional categories, "Typology" and "Notes" that are empty for all entries, and does not contain latitude and longitudes. I used the Google API to find the latitudes and longitudes of the 78 neighborhoods, choosing the API's assigned coordinates of each respective neighborhood as the point of interest for further analysis using FourSquare. I visualize the span of the neighborhoods using folium.

I then used FourSquare's API through a jupyter notebook to determine the amount of venues in a neighborhood with a specific radius from the assigned coordinates collected earlier.

2.2 Data Cleaning

The csv file collected from the city of Denver's official website only had two main problems: it needed a bit of formatting as well as longitude and latitude coordinates for each neighborhood. The csv file was put into a dataframe in the jupyter notebook, and from there I removed the "Typology" and "Notes" categories. Next, the remaining problem was to get coordinates, so I inserted "Latitude" and "Longitude" categories to the frame and appended the coordinates to each neighborhood from the Google API.

Next, I had to collect information on what venues were in which neighborhood, so I used the FourSquare API to determine how many venues there were in each neighborhood. I placed a limit of 150 lookups per neighborhood with a radius of 500 meters, 1000 meters, and 2000 meters to get an idea of the span of venues in a neighborhood without having too much overlap into nearby neighborhoods and concluded the most accurate radius to consider a venue as having a notable presence in a neighborhood would be 1200 meters, around three quarters of a mile. I also figured this is reasonable as that makes things a walkable distance from the center of each neighborhood. This causes a few venues to appear in more than one neighborhood, but it is worth mentioning such venues if they are that close to another neighborhood's center.

I then needed to sort through the venues generated by the filter I applied. I generalized the search to all venues as the venue category that would appear with a venue's name was not always labelled as a restaurant – some said they were pizza places, burger joints, bars, breweries, bistros, etc. I applied a secondary filter to the dataframe and only kept the venues with the words "restaurant", "burger", "pizza", "tavern", "sushi", and "steakhouse" to the dataframe. I then removed the ones with "fast food" in the venue category as these types of restaurants are not of interest for this project. With the data cleaned, I had a little over 40 unique categories to use for exploratory data analysis.

2.3 – Data Features

From cleaning the data, I ended up with two dataframes. One dataframe is specifically focused on the neighborhoods and its distinguishable features (columns) are the neighborhood ID, neighborhood name, the neighborhood's center's latitude, longitude, the number of restaurants in that neighborhood, and the number of categories/types of restaurants in that neighborhood. It has 78 rows and 6 columns in total. This dataframe will be mainly used in the exploratory data analysis section to generate visualizations and insights on neighborhoods.

The second dataframe is the one generated from searching for venues with the FourSquare API. This dataframe is specifically for the restaurants. Its' columns are neighborhood, the neighborhood's center's latitude, longitude, the name of the venue, the venue's latitude, longitude, and the venue category. It has a total of 7 columns and over 1000 rows. Its purpose is to generate information to be used by the first dataframe.

3. Methodology

3.1 – Exploring the Data

The first step in exploring the data was to calculate different insights on different things. I chose to count the unique categories per neighborhood at first to get an idea of how diverse the restaurants were per neighborhood. I calculated this by using a groupby function using neighborhoods as the order and iterating over the "Venue Category" column with the nunique() function. I then calculated the number of restaurants of a certain type per neighborhood by using the same groupby method and iterating of the "Venue Category" column but then using a value_counts() function to count how many of each type of restaurant there are per neighborhood. This calculation provided me with the necessary values to determine what is the most common type of restaurant in a neighborhood for all the listed ones in our dataframe.

I calculated the number of restaurants per neighborhood by using the same groupby method and iterating over the neighborhoods this time, following with a count() function to get these values. This provided me with an idea of what the busiest neighborhoods are in terms of restaurants. I then calculated the number of restaurants per category to determine which type is the most popular in Denver. This calculation suggested that Mexican restaurants are the most numerous in Denver, and so, the most popular.

From these calculations, I added two columns to the main dataframe titled “Number of Restaurants” and “Number of Categories” holding the results. I then used this information to construct a stacked bar chart to visualize the difference between the two features. This provided insight on neighborhoods, since if the stacked bars where the same length or nearly the same, it suggests greater diversity in the neighborhood, but if the number of restaurants bar was much longer than the diversity one, it would show homogeneity in that neighborhood and that it may already be monopolized by a certain type.

3.2 – The Type of Restaurant to Pick

After having made a bar chart to visualize diversity vs dominance, I decided to elaborate on diversity of restaurant types to appropriately determine which type to pick for further analysis. This would mean that I would visualize the distribution of all types of restaurants in Denver, and the amount of types would be stacked against one another. This would mean that without some filter, the pie chart would be unreadable. I did not want to lose the data entirely, but I figured some features could be rewritten to simplify the analysis. I assigned a filter to the second dataframe to change all venue category names to “other” if there were less than 10 of such restaurants across all of neighborhoods of Denver. I then created a pie chart of this information, but quickly noticed that there were many venue categories that specified the same type of cuisine but were named differently.

To address this mislabeling issue, I renamed all “sushi” restaurants to Japanese restaurants as they are of the nations’ cuisine, all “New American” restaurants to American restaurants as they are redundant, and the same logic applied to a couple other types. The only two I did not apply this to were Pizza places to Italian restaurants since they can be very different from one another, so much so that they cannot be considered the same on a general basis. After relabeling the information in the dataframe, I displayed the number of restaurants per category as a table, which gave me the final results that Mexican restaurants dominated Denver in popularity by a large margin, while American restaurants and Pizza places were close to one another at second and third place, respectively. I visualized this difference in another pie chart with labeled percentages, which put Mexican restaurants as representing 19% of Denver restaurants, while American Restaurants made 15%, and Pizza places also making around 15%.

3.3 – What Makes a Good Spot to Place a Restaurant

From the information gathered through creating a pie chart, I decided to count how many Mexican restaurants there were per neighborhood. This would later be used to calculate the frequency of such restaurants among neighborhoods, but as integer values would give insight as to whether a neighborhood already has a significant number of Mexican restaurants. I created a double horizontal bar chart displaying the number of Mexican restaurants per neighborhood in comparison to other restaurants per neighborhood which demonstrated quite a few things to be considered. It showed that of all 78 neighborhoods, 12 did not contain any Mexican restaurants. Of those 12, only 3 had more than 10 restaurants. The earlier bar chart showed the same three had significant diversity in relation to their number of restaurants. This immediately led me to know they would be candidate neighborhoods at the end, but I would need to determine more from the other neighborhoods as 3 candidate neighborhoods are insufficient. There must be a generally low distribution of already existing Mexican restaurants in a neighborhood for the potential to place a new one to be viable.

The next step here was to use one hot encoding on the second dataframe to determine the frequency values of each type of restaurant per neighborhood. Using assistance from a previous lab in the course and the layout of the code as a template, I created a dataframe to display this and then displayed a table of the top 5 frequencies of restaurant types per neighborhood. This led to determining the most common type of restaurant per neighborhood in descriptive terms instead of numerical values. I displayed a value_counts table of this information. This table suggested that even though Mexican restaurants were the most numerous around Denver, American restaurants and Pizza places outnumbered them in commonality in neighborhoods. This suggested that distribution of American restaurants and pizza places were denser in some neighborhoods than others and that Mexican restaurants were more evenly spread. I appended the '1^{st} Most Common Restaurant" column to the main dataframe, and it supplied me with enough information to determine which neighborhoods would be the best candidates.

The main dataframe to be used to determine the candidate neighborhoods would have 9 descriptive columns: Neighborhood ID, neighborhood, latitude, longitude, number of Mexican restaurants, number of restaurants, number of categories, 1^{st} most common restaurant, and the frequency value of Mexican restaurants. Using these defining features, I would apply constraints as mentioned earlier to determine the proper candidate neighborhoods. I would base a candidate neighborhood if the number of restaurants per neighborhood is greater than 10, if the most common restaurant type is not Mexican, and if the Mexican restaurant frequency in the neighborhood is less than 0.125. I applied these measures and appended the remaining dataframe to a folium map.

4. Results

Applying the calculations above returned 17 candidate neighborhoods. Of the 17, 10 had 20 or more restaurants in the area, suggesting they are viable in competition. All neighborhoods displayed 8 or more restaurant categories, suggesting room for additional diversity and 12 of the 17 neighborhoods had 2 or less Mexican restaurants. Only 3 of the neighborhoods had 0 Mexican restaurants. The 17 neighborhoods are: Auraria, Cherry Creek, Congress Park, City Park, Highland, Rosedale, University Park, Mar Lee, Civic Center, CBD, Stapleton, Montbello, Lowry Field, Skyland, Platt Park, Hampden, and Bear Valley. It is important to note that these were calculated with the limitations of FourSquare's ability to detect venues in general and the limitations of my coding abilities. The reality may differ slightly from the values suggested.

Our analysis shows us that there are 17 neighborhoods in Denver that satisfy all statistical conditions to be optimal places to open up a new Mexican restaurant. These decisions were made by calculations based on a filter of 1200 meters from the center of each neighborhood that captured 4600+ venues at first, which we slimmed down to around 1000+ restaurants. Choosing the data features for the Venue Categories were vital to the analysis as it provided us with the range of possible restaurants to use in this project. As we can see from the map, the neighborhoods of focus are for the main part concentrated near the business center of Denver, Capitol Hill. This is great news as this is the busiest area of Denver, where socio-economic dynamics are optimal, and the wide range of diversity in the area makes it so a new restaurant would fit right in.

5. Discussion

I decided not to specify addresses and pinpointing potential locations within the neighborhoods as my capabilities are limited in those aspects and I wouldn't be able to decipher what is a realistic potential address from what could be in the middle of a building. I decided to leave that information for interpretation of the reader, who if interested would be keen to figure out a desirable spot within that neighborhood.

I also chose to remove neighborhoods with less than 10 restaurants as there are many reasons for a neighborhood to have little food venues. Those could be based on the structure of the neighborhood (if it is mainly residential, for example) or there could be important underlying reasons for the lack of food attractions. Anyways, I felt they would not be significant for analysis due to the many complications small numbers bring in making decisions on opening a new business.

6. Conclusions

The purpose of this project was to identify which neighborhoods could be viable candidates to open a new restaurant of a certain type. This type of restaurant would be based on popularity, which was determined by the total number of those restaurants in Denver. The neighborhoods were to be picked based on their lack of the most popular type of restaurant, the number of other restaurants in the area which would suggest if or if not the neighborhood is busy enough to be viable, and whether the neighborhood had already existing restaurants of that type or not. Fast Food restaurants were removed from analysis since they can pretty much be placed around a city like darts thrown randomly at a target. This information in all would aid stakeholders or aspiring restaurant business owners in Denver to open up a restaurant with a promised audience in a busy area, yet without a threatening amount of competition. The restaurants used for analysis had all the information about them gathered through FourSquare and some calculations made above.

The final component of such a decision, choosing the exact place in the neighborhood to open a restaurant, should be made entirely by the stakeholders or aspiring restaurant business owners.