Opening a Coffee Shop in Seattle

Alexander Garcia

November 18, 2019

1. **Introduction**

   1. **Background**
      Seattle is a Mecca for coffee. There are many options for customers and fierce competition among providers.

      This notebook will be looking to determine which neighborhood would be the optimal place to open a new cafe. It will leverage the data provided by the Foursquare API among other data sources to compare various possible neighborhoods.

   2. **Problem**
      Simply put, the question to be considered here is, in a place with as much competition in the coffee industry as Seattle, is there room to open another if you desired? Where might be the best place for one to do this?

   3. **Interest**
      This project will be of interest to anyone who wishes to open a coffee shop in Seattle, from a hipster who believes they have a secret sauce in their cappuccinos, to Starbucks looking to continue their expansion.

2. **Data Acquisition**

1. **Data Sources**
   Three sources of data were used in this project. Data on the zip codes for Seattle and their longitudes and latitudes were found [here](). Data on existing coffee shops and other venues to help guide location selection came from the Foursquare database. Lastly, I used statistics on house sales prices in Seattle found [here]() to further narrow down the remaining options.

2. **Data Cleaning**
   Cleaning the zip code data frame was fairly straightforward. It involved dropping the first approximately 25 rows that had been scraped from the webpage and mostly involved the program grabbing data from the pages' sidebar and leaving most of the data in the first section as NaN. Several columns were also dropped for the same reason.

   The next data set retrieved was concerning existing coffee shops. Cleaning this required first extracting the category from the JSON and then dropping unnecessary or redundant data such as the latitude, longitude, and address all being repeated. The city and state data was also unnecessary as we are only concerned with one city. To try and ensure that the data set was as comprehensive as possible, the radius of each search was set to be almost a mile for each zip code. The function drop_duplicates was used on the id column to eliminate identical responses. Finally, I did a series of queries on the categories column to ensure that as many responses not in the scope were eliminated.

   Another set of queries was done on the Foursquare database to investigate other venues in each zip code. Knowing what other attractions are nearby is an important consideration for a prospective coffee shop as foot traffic is central to its viability. The first step to cleaning this data set was dropping the coffee related columns from the response as we do not want those responses to be a part of this analysis. I performed a one hot encoding on the result so that different neighborhoods could be more easily compared. The responses were grouped by zip code and the mean of the responses taken. This provided a data set from which I could extract the most common venue in each zip code.

   The previous data sets were the primary vehicles by which the choice of area was narrowed down. A data set on house sales prices was used to help make the final determination. While not a perfect representation of the potential business available, a direct comparison of house sales prices serves as a reasonable approximation for affluence and expendable cash for the given areas. I extracted the data based zip codes still under consideration and compared their prices.

3. Exploratory Data Analysis

1. Existing Coffee Shops

To exploring the data set, I started by mapping all the coffee shops in Seattle according to Foursquare as seen below. The density of coffee shops was particularly remarkable for the downtown area, with some areas averaging more than one per block.
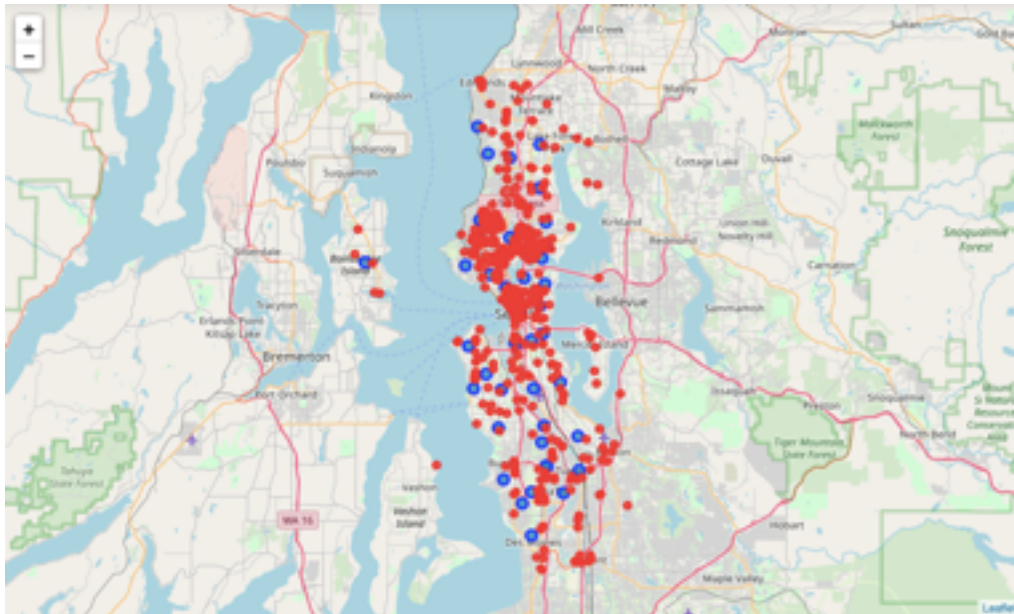

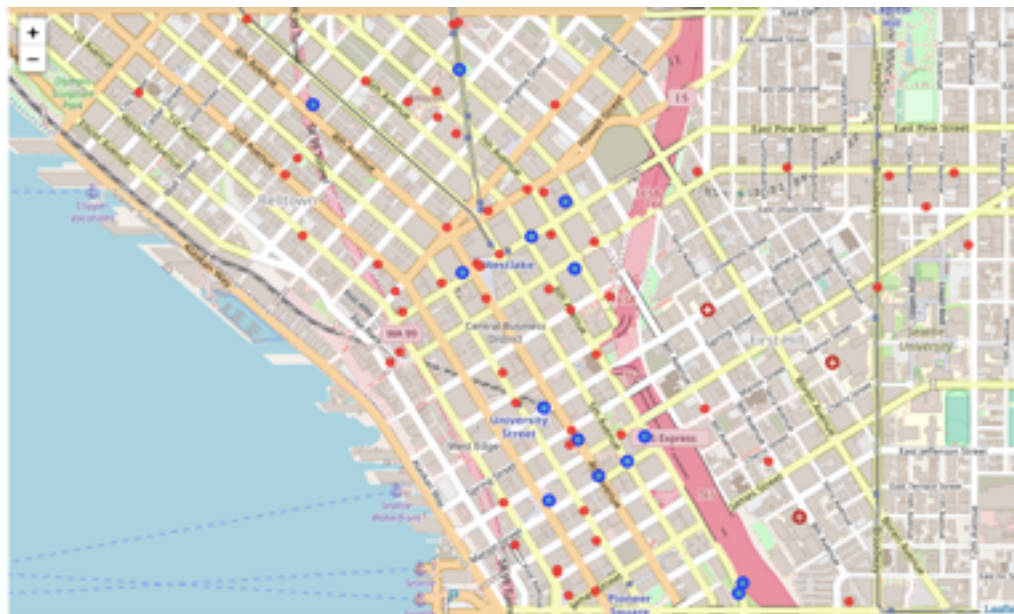*Figure 1: Map of coffee shops in Seattle. Shops marked red and zip codes marked blue.*

Figure 2: Map of coffee shops in downtown Seattle.

I then created a histogram to compare the amount of coffee shops in each zip code as seen below.
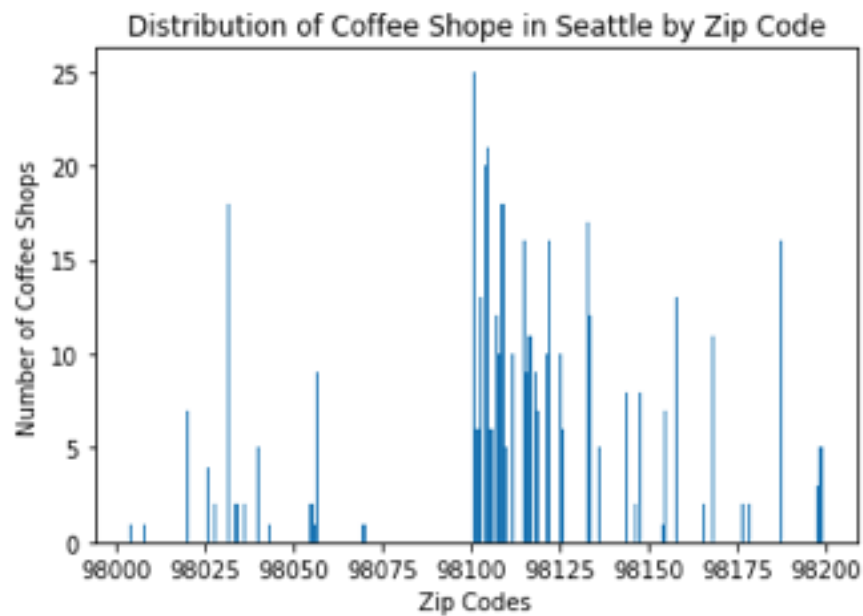

Figure 3: Distribution of coffee shops in Seattle by Zip Code.

This led to the first round of eliminations for areas that could be considered. I determined that 15 coffee shops in a single zip code was the threshold above which an area was too dense to reasonably be considered if the prospective shop wished to be able to compete.

4. Modeling

In order to compare the remaining neighborhoods, I performed a k-means clustering on the data pertaining to other venues in each area. I then mapped each cluster onto a map of Seattle as seen below.
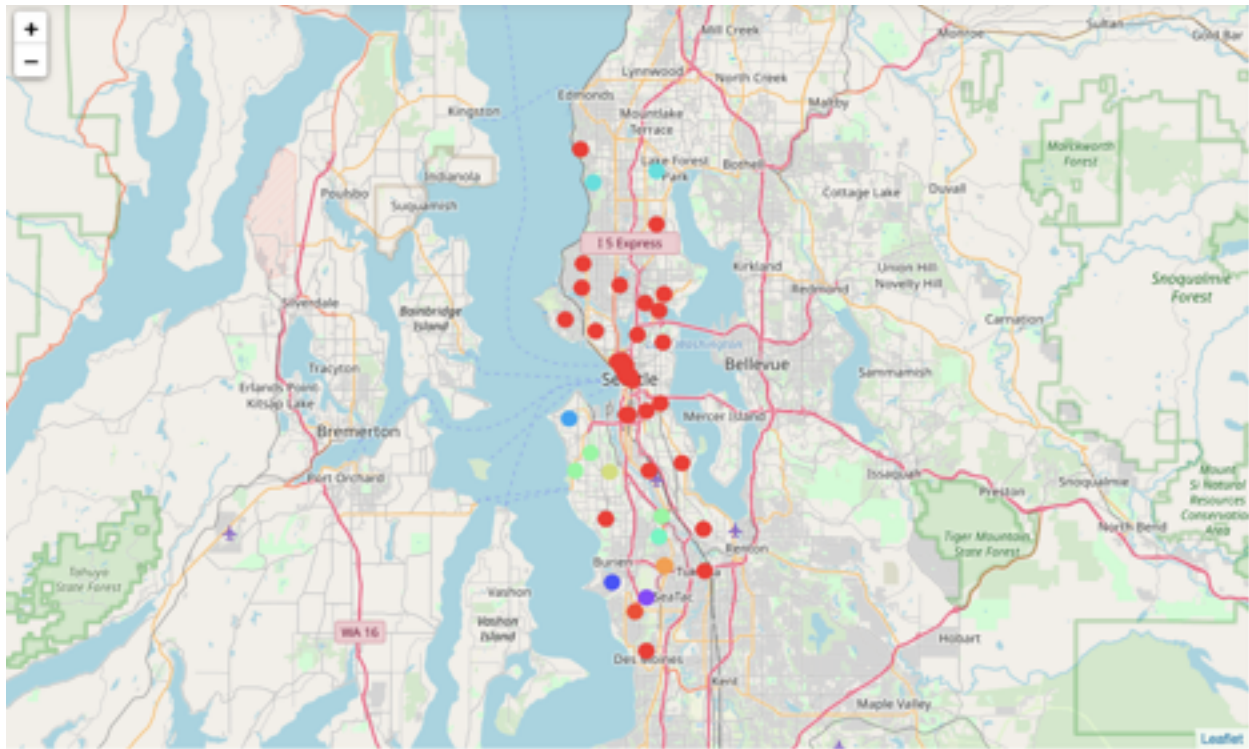

Figure 4: Zip codes of Seattle in clusters.

I ran the clustering several times with various numbers of clusters. Often most zip codes fell into a single cluster (seen in red above). I eventually settled on 10 clusters as the amount that showed consistent patterns within each cluster.

The cluster in red seemed to mostly be connected by the presence of hotels and various types of restaurants. This, combined with the high density of coffee shops in these areas led me to determine that they would not be the best choices for our new coffee shop. The full breakdowns of each cluster are available in the notebook.

The cluster that grabbed my attention was cluster 6. The most common venues in these areas were primarily parks, playgrounds, pools, food trucks, gyms, etc. These are idea neighbors for a new coffee shop as they all attract significant foot traffic. They are marked in green on the map with two located in west Seattle and the third south and east of there.

With these as my targets I used the house sales prices to make a final determination. The data set was lacking information for one of the three zip codes so at this point I had to accept that my choice would be between just two. I created a box plot for prices in these two areas as seen below.
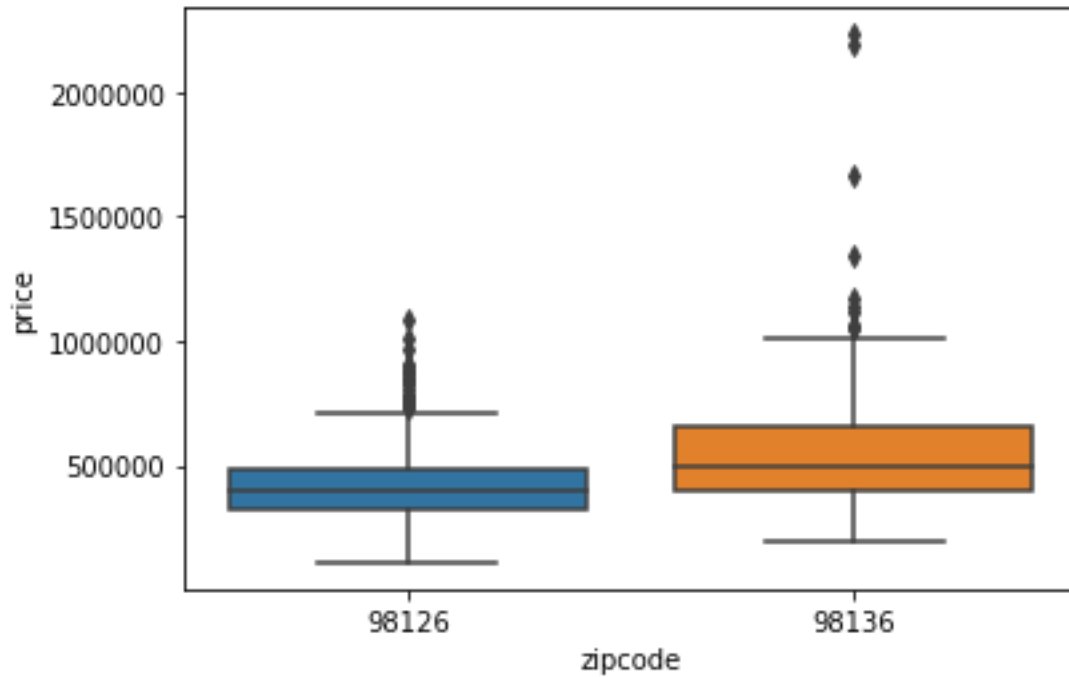

Figure 5: Housing sales prices for two similar neighborhoods in Seattle.

Based on all this information, I determined that the optimal place to open a new coffee shop in Seattle is the zip code 98136.

5. Conclusion

Coffee is undeniably popular in Seattle. It is able to support a large and vibrant array of shops at an impressive level of density. It is clear from this analysis, however, that opportunities remain for those interested and areas that are yet to be fully developed. When considering location, the presence of other stores is of the upmost important. Stay away from the densest areas if you wish to stand out. From there, think about your customer base and what areas will be most conducive to bringing them to you. This approach can be useful to anyone looking to open a business, particularly in the coffee industry.