**Optimizing the Location of a New Seattle-based Coffee Shop**

**Introduction:** This analysis is intended to aid an entrepreneur seeking to start a new café in the greater Seattle area. The question we are trying to answer is which zip code this person should choose for the location of their café. Importantly, Seattle is already known for being home to many great coffee shops. Therefore, we want to identify which zip codes can likely support another coffee shop profitably as evidenced by the density of coffeeshops in otherwise similar zip codes.

**Data:** To solve this problem, we will need three pieces of information: data on venues in the Seattle area, data on the number of people in each zip code, and data on the longitude and latitude of each zip code. The venue data will come from querying Foursquare, the population data will come from <https://www.zipdatamaps.com/zipcodes-seattle-wa>, and the location data will come from <https://gist.githubusercontent.com/erichurst/7882666/raw/5bdc46db47d9515269ab12ed6fb2850377fd869e/US%2520Zip%2520Codes%2520from%25202013%2520Government%2520Data>.

**Methodology:** Location data (longitude and latitude) were accessed using the ‘read\_csv’ method within the Pandas library. The population data was accessed via scraping the above webpage and using Beautiful Soup to parse the html data. This data was then cleaned by removing those zip codes with less than 5,000 people as reported by the webpage. The latitude and longitude for each zip code was then used to query Foursquare’s API and return those venues within 3,200 meters with a maximum number of results set to 1,000. A one hot encoding was applied to each venue according to its category, e.g., coffee shop, hotel, sushi restaurant, etc. The data were then grouped by zip code and normalized according to the total number of venues returned for that zip code. To test the accuracy of the parsed results, the 10 most common venues were displayed within several zip codes and it was verified that the venues belonged to standard categories. A k-means clustering algorithm was then applied to this data to identify zip codes with relatively similar shopping preferences. The clustering algorithm selected was the KMeans routine from the cluster package in scikit-learn and allowed for five clusters. Using Folium, each zip code was plotted on a map of the Seattle area with a circle at the center of the zip code given a color corresponding to the cluster to which it belongs. Next, the number of coffee shops in each zip code was calculated and normalized via the population in that zip code. Note that the term café was treated as synonymous to coffee shop in this exercise. The zip code with the minimum number of coffee shops per 10,000 people was then found within each cluster.

**Results:** Several pieces of data are included in this report. In Table 1, the population and longitude and latitude are shown for several Seattle zip codes. Next, Table 2 shows the most common venues in each of the zip codes from Table 1. Table 3 reports the five zip codes which had the fewest coffee shops per capita in each of the five clusters. Figure 1 shows the five clusters that were formed for the neighborhoods around Seattle. Figure 2 shows the density of coffee shops in each zip code according to population. In this figure, the zip codes are grouped according to their corresponding cluster. In Fig. 3, these five zip codes are plotted on a map of Seattle.

**Discussion:** In Table 1, a population of the same order of magnitude is observed in each of the zip codes. As zip codes are meant to divide an area into regions evenly, this makes sense. In Table 2, the most common venues in each category also seem sensible. A coffee shop is seen in the top four most common venues in all five zip codes shown. In Table 3, the least densely caffeinated zip code in each cluster is shown. From looking at the number of coffee shops per capita shown in the table, these values correspond to the minimum of each cluster in Fig. 2, suggesting that the data was parsed appropriately. In Fig. 1, the zip codes seem to cluster into similar neighborhoods reasonably well. The zip codes found in the very center of Seattle all fall into the same category (red). The more suburban areas are further out and roughly fall into the same cluster (green). West Seattle and the U-District are a bit more metropolitan and fall into the same cluster (purple). The other categories seem to make sense as well, for example Ballard and Magnolia were grouped together. In Fig. 2, each cluster has a fair amount of variability and some zip codes overlap. This is encouraging as it suggests we can in fact optimize our choice on a per cluster basis. In Fig. 3, we plot each of the most promising zip codes. Each zip code falls on the Seattle area, suggesting that the longitude and latitude values were carried throughout the analysis appropriately.

**Conclusion:** There are two issues confounding the selection of a ‘best’ zip code for a new coffee shop. The first is the difference in population density within each zip code and the other is that each neighborhood will likely have a different culture and needs. By using k-means clustering, each of 28 Seattle zip codes was sorted into a cluster according to similarity of venues in the area. The zip code with the least coffee shops per capita was identified within each of five clusters. These zip codes are the ones most likely to be able to support another coffee shop in the area.

Table 1: Five zip codes are shown with their corresponding latitutde, longitude, and population.

|  |  |  |  |
| --- | --- | --- | --- |
| **Zip Code** | **Latitude** | **Longitude** | **Population** |
| 98034 | 47.71577 | -122.214 | 40407 |
| 98101 | 47.6109 | -122.336 | 10238 |
| 98102 | 47.63714 | -122.322 | 20756 |
| 98103 | 47.67826 | -122.338 | 45911 |
| 98104 | 47.60188 | -122.329 | 13095 |

Table 2: The four most common venues are shown for five zip codes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ZIP** | **1st Most Common Venue** | **2nd Most Common Venue** | **3rd Most Common Venue** | **4th Most Common Venue** |
| 98034 | Coffee Shop | Sushi Restaurant | Sandwich Place | Pizza Place |
| 98101 | Hotel | Coffee Shop | Bakery | Sandwich Place |
| 98102 | Coffee Shop | Park | American Restaurant | Bakery |
| 98103 | Coffee Shop | Bar | Ice Cream Shop | Park |
| 98104 | Hotel | Vietnamese Restaurant | Cocktail Bar | Coffee Shop |

Table 3: The population and number of coffee shops per 10,000 people for each of the zip codes which had the fewest coffee shops in their cluster.

|  |  |  |  |
| --- | --- | --- | --- |
| **ZIP** | **Population** | **coffee per capita** | **Cluster Labels** |
| 98105 | 43924 | 1.593662 | 1 |
| 98118 | 42731 | 2.106199 | 4 |
| 98121 | 12628 | 4.751346 | 0 |
| 98133 | 44555 | 2.019975 | 3 |
| 98199 | 19686 | 2.031901 | 2 |

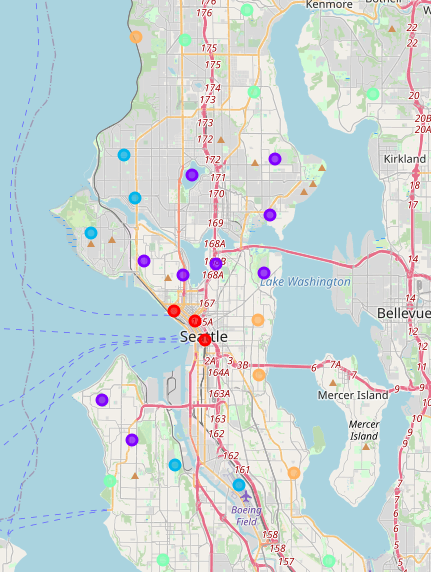


Figure 1: Each of the Seattle zip codes are shown with each marker’s color corresponding to the cluster to which that zip code belongs.

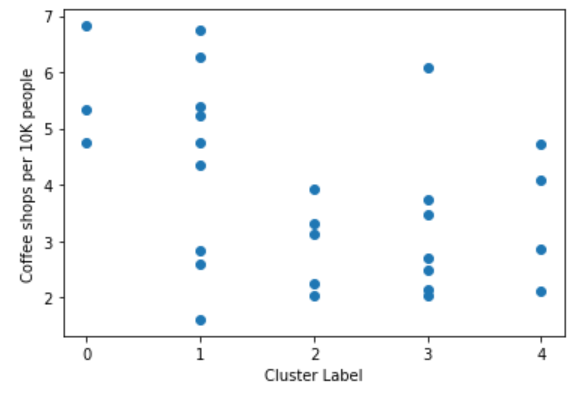


Figure 2: A scatter plot showing the number of coffee shops per 10,000 people in each zip code. Note that each zip code is organized according to the cluster to which it belongs.

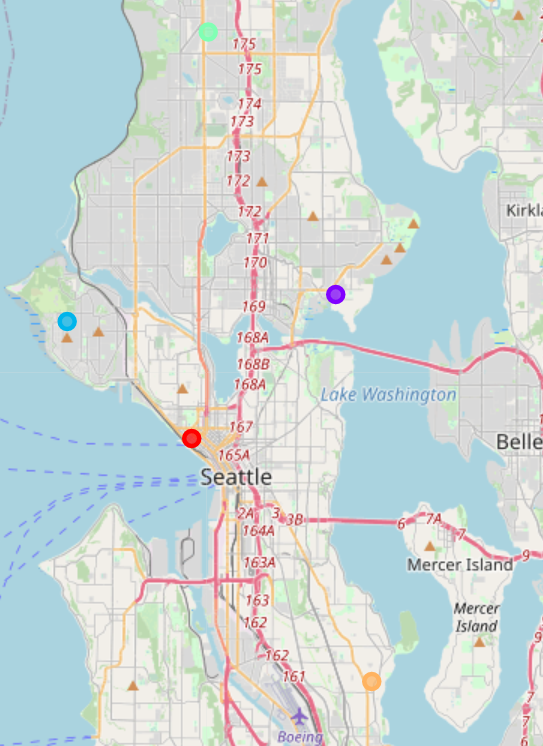


Figure 3: The location of the zip code in each category that would be most optimal for a new coffee shop to establish itself.