Clustering Venue Categories in Seattle

Ashley Cacho

April 24, 2020

1 Introduction

For the longest time, Seattle, Washington has been known as the coffee capital of the United States. According to a recent study in September 2019 by WalletHub[1], Seattle came out on top as the Best Coffee City in America. The study measured several metrics across the top 100 most populous cities in the United States. With the cold and dreary days experienced for most of the year, it makes sense that Seattleites have turned to the delicious concoction to stay warm.

Suppose a person wanted to open a new shop in Seattle. Several considerations must be made to aid in its success. For example, the interior design is important in generating a welcoming environment to either stay and work or to bring customers in for the ever-growing popularity of aesthetic posts for social media. Another important factor, and arguably the most important, is location. Some neighborhoods may not have any coffee shops so opening one there may be a wise choice. However, in a populous city like Seattle, littered with emerging tech companies, one may wonder if there is a possibility of a district having no coffee shop.

Entrepreneurs looking to make an investment in the Seattle area would be interested in this information. Which area of Seattle should a new cafe open? The goal here is to leverage location data to understand the layout of the coffee shop industry in Seattle and to provide suggestion of which district to open a shop.

2 Data acquisition and cleaning

The data that might contribute to determining the best location to open a new coffee shop would involve the kinds of venues or establishments within a district. Two sources of data were used in this analysis. The first one is a list of districts within Seattle [2]. In order to obtain the location data for these districts, the *Nominatim* package from *geopy.geocoders* was used. The other source is from using the Foursquare API to obtain nearby establishments relative to the centers of districts.

Some steps are needed to prepare the district data set. Several columns were dropped as those pieces of information are not necessary to the question at hand. Annexed, Locator map, Street map, Image, and Notes were all dropped. Neighborhood name was also dropped because this analysis focuses on the larger district. Footnotes were parsed as “[number]” so some regex text processing was applied to remove those unnecessary characters. In addition, some districts have a “/” in their name since they represented neighborhoods that bordered between two districts. Since we only need the list of unique districts, these rows were dropped. The

Nominatim package did not return latitude and longitude values for the South End and Industrial Districts. These rows were also dropped giving a final district count of 21 (Table 1).

Table 1: Seattle districts with location information obtained from [1] and *Nominatim*.

District	Latitude	Longitude
Seattle	47.6028956	-122.33984
North Seattle	47.6607729	-122.2915
Northgate	47.7131534	-122.32123
Lake City	47.7191619	-122.29549
Ballard	47.6765073	-122.38622
Central Seattle	47.6126938	-122.3032
Magnolia	47.6468106	-122.39949
Queen Anne	47.6394805	-122.36075
Capitol Hill	47.6238307	-122.31837
Madison Park	47.6359301	-122.2802
Lake Union	47.6399187	-122.33556
South Lake Union	47.6231611	-122.33838
Downtown	47.6048723	-122.33346
Central Area	47.6038321	-122.33006
Minor	47.6265944	-122.33227
Atlantic	47.5904927	-122.32431
Madrona	47.6127915	-122.29123
Rainier Valley	47.552544	-122.29089
Seward Park	47.5512156	-122.2658
Beacon Hill	47.5792579	-122.3116

Using the *folium* library in Python, a map of these districts is provided (Figure 1). The interactive map can be viewed from the accompanied Jupyter notebook.

Foursquare was used to query establishments or venues within a radius of 500m to every latitude and longitude value of each district. For this data, only the ID, name of venue, category, latitude, and longitude data were kept, all other dropped. Results did not come back on venues from several districts for reasons that will not be explored here.

3 Methodology

3.1 Exploratory Analysis

To obtain a sense of the coffee shop industry layout, a specific query with “coffee” was used to obtain nearby venues and mapped out (Figure 2). The query was limited to the top 200 coffee shops for each district. South Lake Union, Central Area, and Downtown are the top 3 districts with the most coffee shops, colored in or dark orange, cyan, and orange, respectively. However, these districts would not be automatically disregarded for possible locations since it is also important to look at the distribution of venue types by each district.

In addition to looking at the total numbers of coffee shops in each district, the within district distribution for venue categories is also analyzed. Figure 3 shows the distribution of the

Figure 1: A map of Seattle with points for every district found in [2].

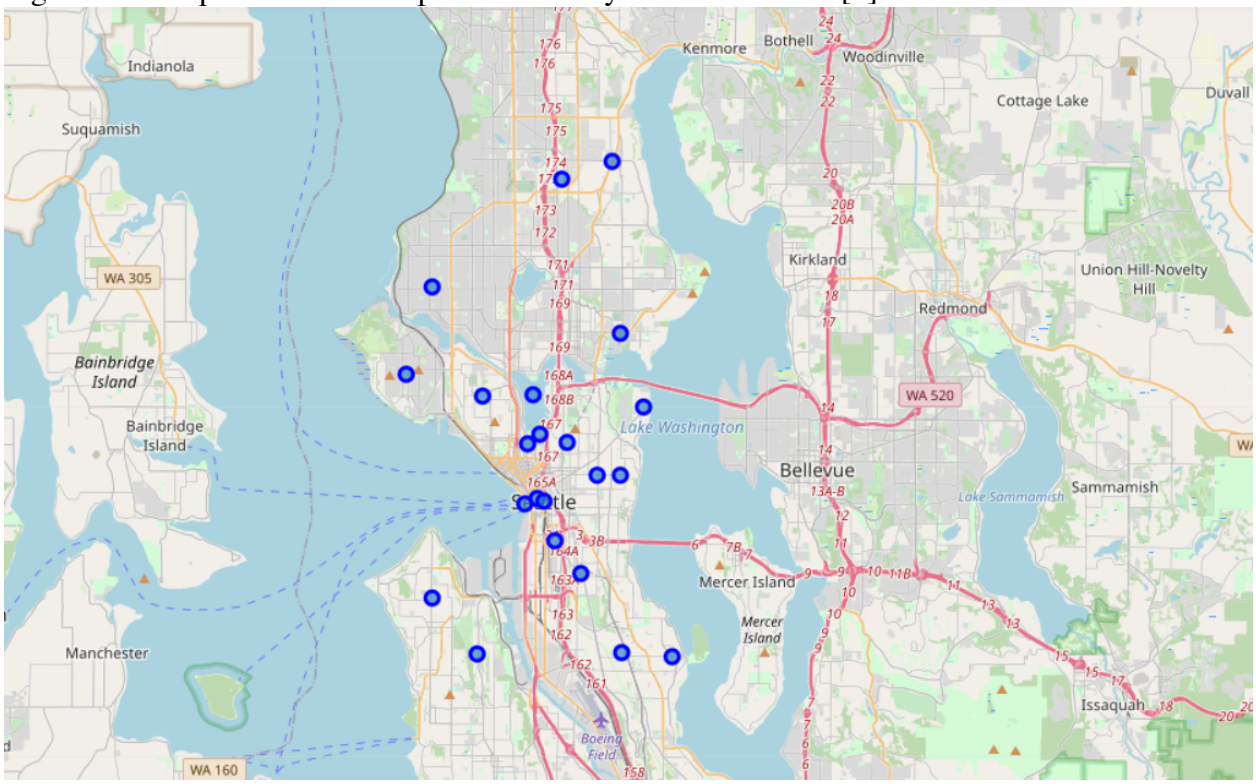
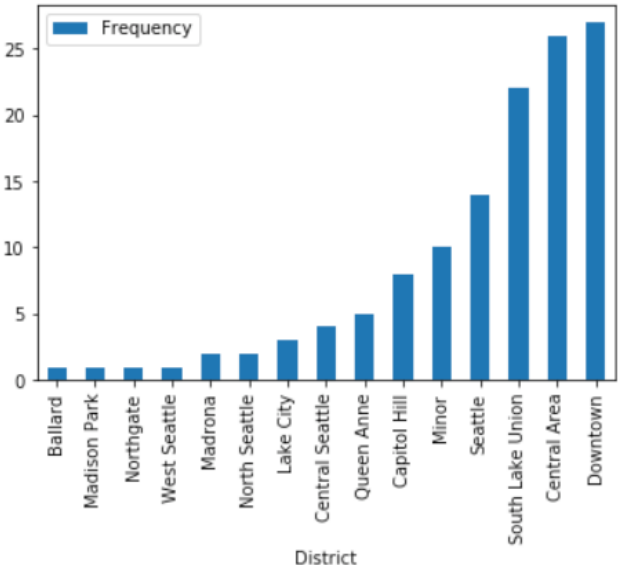
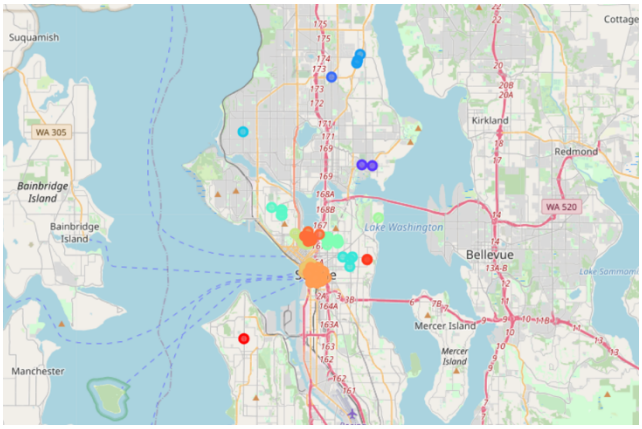


Figure 2: Frequency distribution of coffee shops in each district with accompanying map.





number of venues returned by Foursquare.

For each district, the top 5 most represented type of venue was obtained. Some example districts are shown in Table 2. Not including counts, a table summary of the top 5 most common venues are listed in Table 3. The table shows that only three districts appear to have a coffee shop or café in the top 5 most common establishments within the district and they are Central Seattle, Minor, and South Lake Union.

Table 2: Examples of the distribution of venue type by district. Remaining distributions available on Jupyter notebook.

District: Atlantic		District: Central Seattle	
Venue	Frequency	Venue	Frequency
Bus Line	9	Food Truck	9
Automotive Shop	8	Coffee Shop	5
Government Building	6	Building	4
Bus Stop	6	Bus Stop	4
Building	4	Salon / Barbershop	4

Figure 3: Frequency distribution of venues returned by Foursquare in each district.

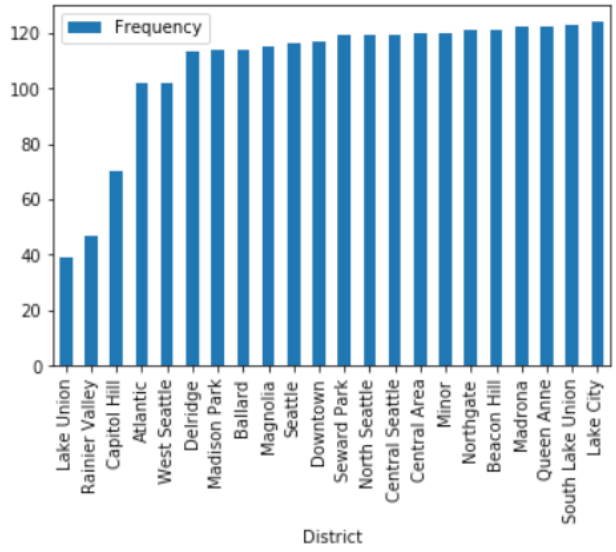


Table 3: Top 5 most common venues for each district, according to Foursquare.

District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Atlantic	Bus Line	Automotive Shop	Government Building	Bus Stop	Building

Ballard	Office	Bus Line	Church	Residential Building (Apartment / Condo)	Spiritual Center
Beacon Hill	Food Truck	Residential Building (Apartment / Condo)	Salon / Barbershop	Mexican Restaurant	Church
Capitol Hill	Residential Building (Apartment / Condo)	Dog Run	Bank	Gym / Fitness Center	Bed & Breakfast
Central Area	Office	Lawyer	Art Gallery	Courthouse	Bus Stop
Central Seattle	Food Truck	Coffee Shop	Salon / Barbershop	Building	Bus Stop
Delridge	Bus Line	Park	Garden	Bus Station	Playground
Downtown	Office	Bus Stop	Tech Startup	Building	Lawyer
Lake City	Bus Station	Salon / Barbershop	Residential Building (Apartment / Condo)	Bank	Asian Restaurant
Lake Union	Boat or Ferry	Harbor / Marina	Bus Stop	Lake	Deli / Bodega
Madison Park	Salon / Barbershop	Bus Stop	Office	Residential Building (Apartment / Condo)	Dentist's Office
Madrona	Bus Stop	Office	Gift Shop	Boutique	Furniture / Home Store
Magnolia	Bus Stop	Building	Bank	Church	Playground
Minor	Office	Building	Cafe	Boat or Ferry	Parking
North Seattle	Doctor's Office	Dentist's Office	Salon / Barbershop	Medical Center	School
Northgate	Dentist's Office	Office	Building	Doctor's Office	Assisted Living
Queen Anne	Bus Stop	Office	Dentist's Office	Yoga Studio	Nail Salon
Rainier Valley	Automotive Shop	Farm	Pawn Shop	Convenience Store	Light Rail Station
Seattle	Office	Boat or Ferry	Tech Startup	Pier	Gift Shop
Seward Park	Salon / Barbershop	Church	Building	Synagogue	Park
South Lake Union	Office	Food Truck	Coffee Shop	Hot Dog Joint	Rental Car Location

West Seattle	Doctor's Office	Church	Salon / Barbershop	Office	Cosmetics Shop

3.2 KMeans Clustering Analysis

Clustering of the venues is performed to see if there are any natural groupings of the venues. If districts are clustered because they appear to have more coffee shops than other clusters, then this is more information to provide a person who is interested in opening a new coffee shop.

In order to perform clustering on the data, each district is tabulated with information on the venues that were retrieved. Each unique category is one-hot encoded. For example, a categorical venue of Bus Line can be converted as a dummy variable such that it equals one if the venue is a Bus Line and 0 otherwise.

After converting each category of venue into a dummy variable, the data consists of 22 rows and 367 columns. For each column, the counts are standardized using the StandardScaler function of scikit-learn. This data is unsupervised so the number of clusters to use is not easily obtained. Several values of the n_clusters parameter of KMeans() was tested and k was chosen such that each cluster obtained a 'reasonable' amount of districts within each.

4 Results

The results of a particular instance of KMeans clustering analysis run with the number of clusters chosen as 4 are mapped in Figure 4. Cluster 1 (red) had a single district, Cluster 2 (purple) had 13 districts, Cluster 3 (cyan) had 4 districts, and Cluster 4 (yellow) had 4 districts.

Examining each cluster may lead to more insights. Table 4 shows the results of the clustering as well as the most common venues. Cluster 1 only had Central Seattle district in it. The first and second most common venue are Food Truck and Coffee Shop. Cluster 2 had 14 districts which have more offices and housing. Cluster 3 had 4 districts which have more Bus Stops and Salon / Barbershops. Finally, Cluster 4 had 4 districts which have more dentist's and doctor's offices as well as Bus Stops.

Table 4: Districts that appear in each cluster formed by KMeans clustering algorithm.

Cluster 1					
District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Central Seattle	Food Truck	Coffee Shop	Salon / Barbershop	Building	Bus Stop

Cluster 2					
District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Seattle	Office	Boat or Ferry	Tech Startup	Pier	Gift Shop
North Seattle	Doctor's Office	Dentist's Office	Salon / Barbershop	Medical Center	School

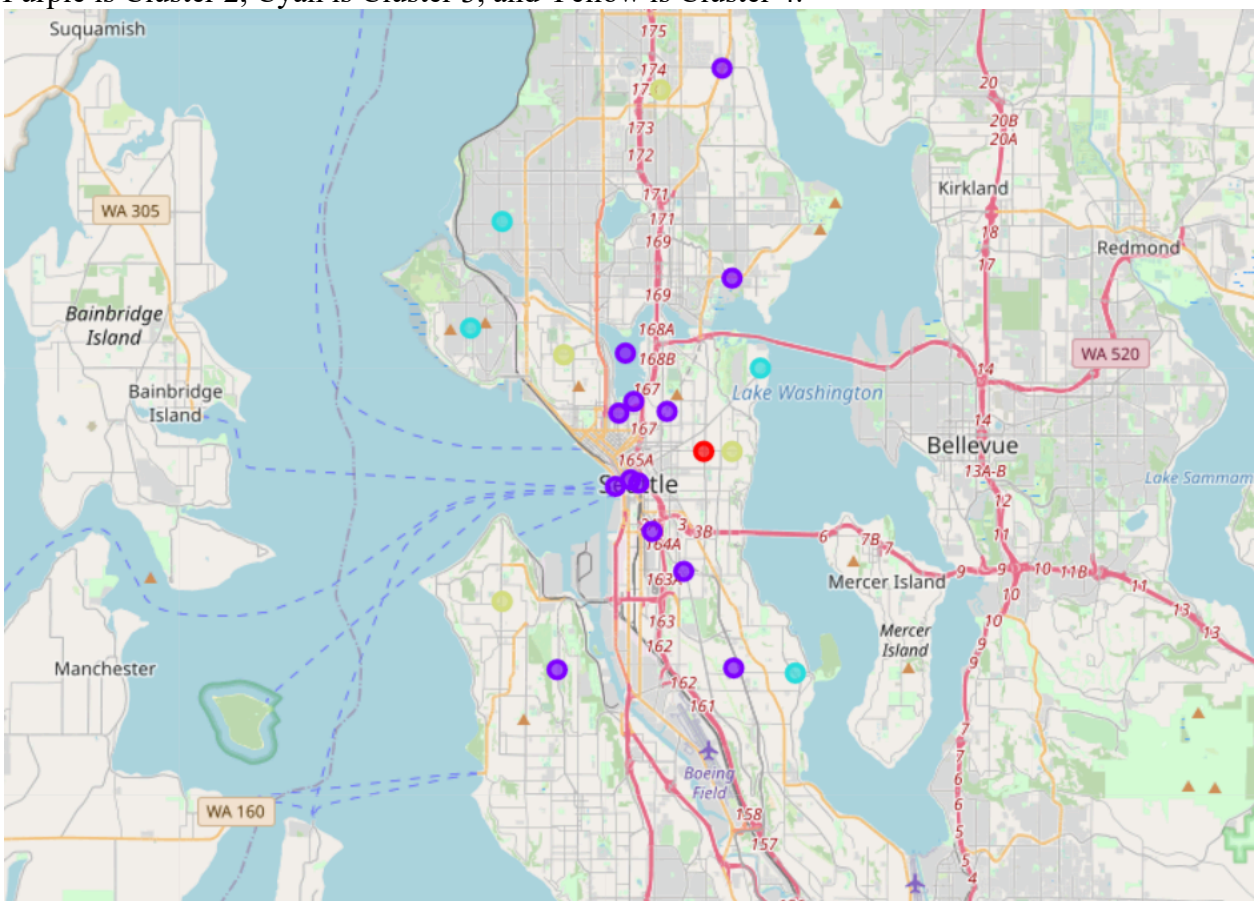
Lake City	Bus Station	Salon / Barbershop	Residential Building (Apartment / Condo)	Bank	Asian Restaurant
Capitol Hill	Residential Building (Apartment / Condo)	Dog Run	Bank	Gym / Fitness Center	Bed & Breakfast
Lake Union	Boat or Ferry	Harbor / Marina	Bus Stop	Lake	Deli / Bodega
South Lake Union	Office	Food Truck	Coffee Shop	Hot Dog Joint	Rental Car Location
Downtown	Office	Bus Stop	Tech Startup	Building	Lawyer
Central Area	Office	Lawyer	Art Gallery	Courthouse	Bus Stop
Minor	Office	Building	Cafv©	Boat or Ferry	Parking
Atlantic	Bus Line	Automotive Shop	Government Building	Bus Stop	Building
Rainier Valley	Automotive Shop	Farm	Pawn Shop	Convenience Store	Light Rail Station
Beacon Hill	Food Truck	Residential Building (Apartment / Condo)	Salon / Barbershop	Mexican Restaurant	Church
Delridge	Bus Line	Park	Garden	Bus Station	Playground

Cluster 3					
District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Ballard	Office	Bus Line	Church	Residential Building (Apartment / Condo)	Spiritual Center
Magnolia	Bus Stop	Building	Bank	Church	Playground
Madison Park	Salon / Barbershop	Bus Stop	Office	Residential Building (Apartment / Condo)	Dentist's Office
Seward Park	Salon / Barbershop	Church	Building	Synagogue	Park

Cluster 4

District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Northgate	Dentist's Office	Office	Building	Doctor's Office	Assisted Living
Queen Anne	Bus Stop	Office	Dentist's Office	Yoga Studio	Nail Salon
Madrona	Bus Stop	Office	Gift Shop	Boutique	Furniture / Home Store
West Seattle	Doctor's Office	Church	Salon / Barbershop	Office	Cosmetics Shop

Figure 4: A map of Seattle's districts colored by the KMeans clustering results. Red is Cluster 1, Purple is Cluster 2, Cyan is Cluster 3, and Yellow is Cluster 4.



5 Discussion

The Central Seattle district was clustered separately from the other districts based on food and coffee. Recall that in Figure 2 that only five coffee shops were returned by the Foursquare query for Central Seattle while Downtown had 5-fold more coffee shops. These results show that even

though absolute counts of the number of coffee shops show Central Seattle having much fewer than say, Downtown, the relative amounts within each district clearly shows that Central Seattle has more. This information would suggest that opening a new coffee shop in Central Seattle may not be the best choice.

Minor and South Lake Union also appear to have more coffee shops within their districts relative to other establishments with café being the third most and coffee shop being the third most, respectively. Those districts would also be eliminated from possible locations.

The districts in Cluster 2 have more offices and housing. Considering only the analysis using location data, a good choice for opening a new coffee shop would be in any of the districts found in Cluster 2, excluding Minor and South Lake Union.

6 Conclusions

In this study, I analyzed location data to provide a suggestion for the best district to open a coffee shop. Coffee shops are visited by many people but one can argue that the busiest ones are nearest offices. People working who need that extra caffeine to wake up in the morning can easily stop by a shop near work. Not to mention, the afternoon slump. Having a nice cup of coffee to wake from that would be nice and having a coffee shop nearby would be ideal. This analysis could help provide someone with the tool to look deeper into where to open a shop. This analysis should not be used alone as there are several factors that affect the success of a coffee shop. However, I believe this is a great starting place. For more details, please see the accompanying Jupyter notebook.

7 Resources

1. https://wallethub.com/edu/best-cities-for-coffee-lovers/23739/
2. https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle