**Seattle Neighborhood Clustering**

Joe Battafarano

August 25, 2020

**1. Introduction**

* 1. **Background**

Seattle has been and continues to be one of top grossing technological cites in the world. It is the headquarters of some of the most value companies such as Amazon, Google, and Microsoft. All whom attract the best talent in the world. As the population grows so does the landscape evolve for better and worse. The technological advancements bring wealth in the form of tax dollars and innovations from bright minds. On the flipside, Seattle’s housing prices increase nearly at the same rate as San Francisco. Many businesses struggle and homelessness is on the rise. There is no stopping technological advancement. It would be trivial to assume someone can predict their net outcomes. Therefore, a company or an individual can only do what’s best with the information they are given. Information can be turned into insight and each party can choose what neighborhood best fits their situation.

* 1. **Problem**

Find similar neighborhoods based on proximity, popularity, and pricing for companies and individuals. Companies want to expand, but don’t want to be the tenth coffee shop on the block. An individual has their own preferences and budget. There may be a similar neighborhood with lower costs. The right answer for a company or individual is not as clear cut as classification. Therefore, clustering will be utilized and insight for both parties will be gained.

* 1. **Interest**

Seattleites, new hires, and growing businesses alike can extrapolate insight to make more economical decisions increasing their quality of life and lowering cost of living. A business can find their target consumers and keep track of competition. An individual can stay on budget while enjoying the highest-ranking venues around.

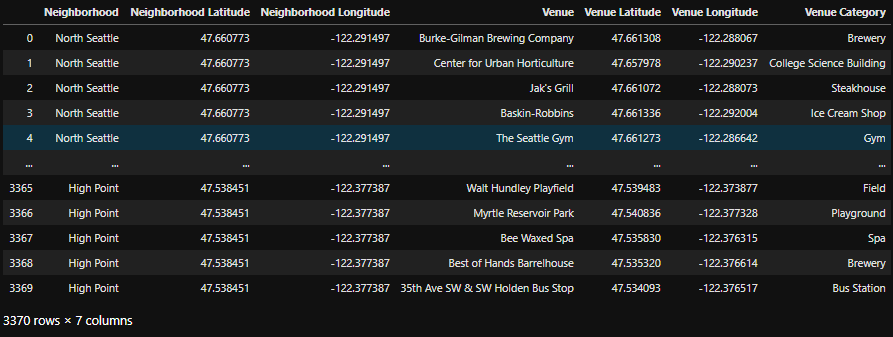
**2. Data**

**2.1 Data Acquisition**

The Seattle neighborhoods and their addresses were scraped via Wikipedia using Python’s Beautiful Soup. The addresses were then processed into GeoPy to obtain their longitude and latitude. The addresses and their coordinates enabled a series of Foursquare API calls via the free endpoint explore. Each unique venue id had a one to optional many relationship with the neighborhoods. The venue ids were used as a parameter to obtain pricing and rating information via Foursquare’s premium endpoint details. Since details is a premium endpoint only 500 calls can be executed per day. Therefore, the unique venues were parsed out over a series of days. The data was concatenated together and formed two major datasets.

**2.2 Data Understanding**

All of the data sources had clean data. The only exception was the Seattle neighborhoods from the web. Several special characters needed to be removed and the data shaped accordingly. Nonethless, the seattle neighborhoods were merged with the respective Foursquare data to form two datasets: *Seattle\_Venues(see Table 1) and Seattle\_Venues\_PnR(see Table 2).* *Seattle\_venues* contains all 118 neighborhoods. There are a total of 3370 venues, 2285 of which are unique venues across 311 different categories. *Seattle\_Venues\_PnR* is a smaller dataset only containing the venues with pricing and rating information. Thus, the spread is far different. There are a total of 1113 venues across 90 different categories in 82 different neighborhoods. Many of the categories consist of various restaurant types. The venue price is tiered one through four. At price tier one an item is ten dollars or less. Whereas at price tier four sells many items around the thirty to forty-dollars. Even with a smaller dataset the pricing information can tell us a lot about the neighborhoods’ cost of living. The quality of life or rather the perceived quality of life can be examined via the venue rating. The range of values is from one to ten. Both *Seattle\_Venues* and *Seattle\_Venues\_PnR* are reformatted later to accommodate k-means clustering. *Seattle\_Venues* uses the most frequent venues to establish similarity and dissimilarity between neighborhoods. *Seattle\_Venues\_PnR­* will also include pricing and rating averages on a neighborhood basis. This will give a more in-depth view on the cost of life as well as the quality of life.



***Table 1:*** Seattle\_venues



***Table 2:*** Seattle\_venues\_PnR(price and rating)

**2.3 Data Relevance**

A business must act strategically prior to opening its doors or expanding. Thus, knowing which neighborhoods have venue compliments or competing venues will be crucial. Additionally, a business, especially restaurants want to price their items according to the target audience. Thus, knowing the price range in the neighborhood will be crucial from the start of business to years later. Restaurants are risky and a misstep in strategy can hinder small businesses with little cashflow.

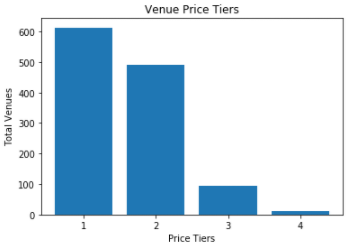
An individual fresh out of college or a family looking to settle down can use the same data differently. Many people walk or bike everywhere rain or shine. The convenience of your favorite food can be found in a cheaper or more aesthetically pleasing place with a park. Regardless of your budget the neighborhood clusters should be able to group up similar alternatives to your current living situation.

**3. Exploratory Data Analysis (Methodology)**

**3.1 Pricing**

The pricing data set contains all 1211 venues that had a price tier. It is a subset of all the venues such that pricing tier exists. The variety of restaurant pricing can be shown in *Figure 1*. Most of the venues fall at price tiers one and two with 613 and 492 venues, respectively. That means about 90% of all priced venues charge twenty dollars or less for their items. The near identical proportions of tier one and tier two could hint at differences in price levels. Later, neighborhoods with similar venues, but different price levels could be captured in the same cluster. A clear indication that one should move to the lower cost neighborhood that has the same venue availability.

It is expected price tiers three and four are far rarer. Although cost of living differs in each neighborhood, it would be irrational to assume all expensive venues exist in one place. Yet, the proximity of such expensive options should show differences in neighborhoods. The clustering will identify neighborhoods with the similar price tiers. It would not be surprising to see neighborhoods slightly differ in venue availability but have similar cost of living. Such finds would be less impactful than difference in tier one and tier two. However, it may indicate to a restaurant to move to a less competitive area. For an individual or family, it may not save them money. Rather, moving would only increase their availability to venues they enjoy.

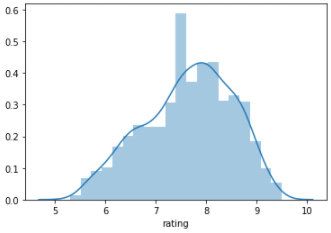


***Figure 1:*** *Venue Price Tiers*

**3.2 Rating**

The rating dataset contains far more venues than the pricing dataset at a total of 1890. It is a subset of all venues such that a rating exists. Ratings are represented by discrete values from one to ten. They can be non-whole number values such as 6.5. Although keep in mind there is only one decimal place. Therefore, the values are not continuous and remain between one and ten. As you can see in *Figure 2* no ratings exist below 5.3. Foursquare might have a feature that allows venues to not show their ratings below this threshold. It is unclear why 5.3 is the cut off. Additionally, it is unclear why the maximum is 9.5 and not say 9.7 or even ten. Again, this might be something Foursquare does behind the scenes.

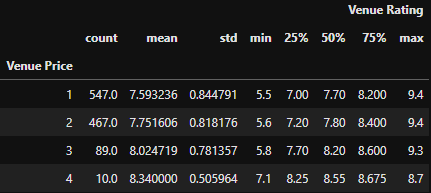
Something fascinating about the distribution in *Figure 2* is its almost normally distributed. The mean is 7.6 and the median 7.7. The standard deviation is close to one at approximately, .87. Yet, the distribution still is skewed. For whatever reason there is a massive portion of venues around the 7.5 rating mark. It might be a threshold venue strive for or where consumers are satisfied. The ratings on their own to not tell us enough information for our business problem. The next section looks at the relationship of pricing and rating. A combination which should indicate a balance of cost of living and quality of living.



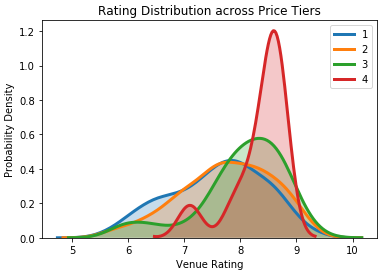
***Figure 2:*** *Rating Distribution*

**3.3 Pricing and Rating**

The pricing and rating dataset *Seattle\_venues\_PnR* will also be used in cluster. It contains all venues such that a pricing tier and rating exist. There is a total of 1113 venues. Price and rating are not correlated with only a slight positive value of 0.16. As you can see in *Table 3* each price tier differs immensely. The only clear relationship between all price tiers and ratings is the left skewed distribution. You can see this illustrated in *Figure 3*. Keep in mind it is a probability density function, not a cumulative density function. Therefore, values can exceed one since it is not representing a probability.



***Table 3:*** *Venue Rating Descriptive Stats Across Price Levels*



***Figure 4:*** *Rating Distribution Across Price Tiers*

As the price tier increases so does the average rating at restaurants. However, the difference in averages is minimal. All price tiers appear to be located around a similar range of values. The only exception is price tier four. The minimum rating value is far greater than the other three tiers. Each price tier follows an upward trend in 25th percentile, 50th percentile, 75th percentile. The exception being the maximum. Although smaller sample sizes price tier three and four especially have lower maximums. Such a finding may hint at consumer behavior patterns. Simply put, consumers expect more when spending more. Thus, the overall ratings hit a lower ceiling. The gap between perception and expectation can only go so high. The most important take away is ratings may be slightly higher on average as price increases. However, the difference is marginal. Most venues at price tier one and two are only about a 0.5 difference in average rating. Thus, individuals may care less about the rating and companies may be satisfied with a 7.5 rating regardless of price level.

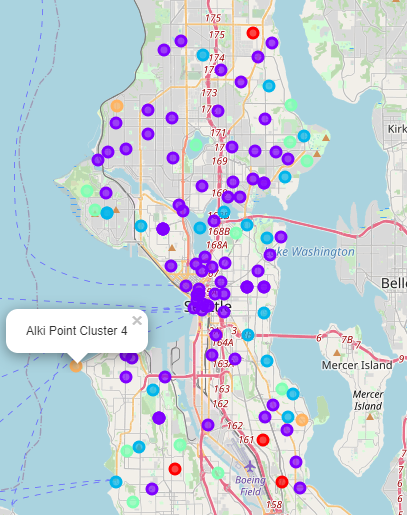
**4. Modeling (Results and Discussion)**

**4.1 K-Means Clustering: 10 Most Common Venues Nearby**

Companies need to know if surrounding venues are competitors or compliments. Opening the fifth coffee shop on the block would be ill advised. However, placing a café near a park in a different neighborhood may yield additional traffic. An individual has specific tastes and limited time on their hands. Thankfully, they can take the bike trail to work, stop by the only Ethiopian restaurant around for lunch, and make their eye doctor appoint before dinner. The scenarios posed are not easy to meet. It is not easy to draw the line between similarities and dissimilarities. Therefore, clustering comes into play.

Prior to clustering the data, it must be reshaped. There is a total of 3370 venues across 311 categories in 118 different neighborhoods. The categories are one hot encoded and their frequencies measured. Many categories do not exist in other neighborhoods. For example, not every neighborhood has a beach, a park, or a gym. Due to their lower frequency neighborhoods with these specific venues may group closely. In contrast, many neighborhoods have coffee shops, gas stations, or pubs. Their importance in differentiating one neighborhood from another is marginal. All venue frequencies are rank ordered and narrowed down to the most common ten. The most common ten venues will be the only input into this clustering portion. The next section will explore the most common ten venues along with their pricing and rating.

Below you can see *Figure 5*. It is a map of the Seattle neighborhoods after clustering. Each neighborhood can be clicked on for more information. Their color represents which cluster they belong to. *Table 4* is a key identifying each cluster and its respective color. What immediately pops out is the amount of purple representing cluster one. The density of purple in the middle is not surprising. Since they are tightly packed many venue frequencies are similar. Some of the most common venues include trails, parks, and coffee shops. They differ slightly in frequency but are close enough in value to cluster together. Another notable cluster is cluster four. It only compromises of three neighborhoods. Each neighborhood in a completely different area of Seattle. However, due to their costal locations share rare venues in common. These include beaches, fabric shops, falafel restaurants, and fairs. An odd combination no doubt. One a company or person would not intuitively anticipate such similarities. The dissimilarities are as important. A family may love the beach but dreads the long commute to work. Luckily, Seward Park has multiple bus stations. The other two beach neighborhoods don’t have any at all. A practical solution would be to move, not compromising your interests while saving time in traffic.



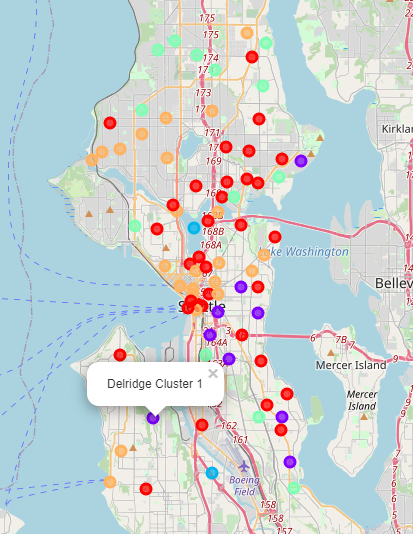
***Figure 5 and Table 4:*** *K-Means Clustering Interactive Map*

**4.2 K-Means Clustering: Pricing and Rating**

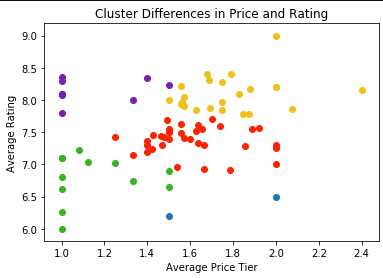
The top ten most common venues are helpful, but not enough for a company or individual to plan. Thus, cost of living and quality of living must be considered. Companies have different target audiences and markets. Even though a company might be the only bakery in town, people might not be able to afford it. Alternatively, venue rankings in an area could be higher than average. An additional risk analysis may be necessary. It is the company’s choice to decide their risk versus reward. A new hire now moving to Seattle knows housing prices are on the rise. To not get off on the wrong foot, picking a lower price level with similar venues is a great compromise. Alternatively, a slightly more expensive neighborhood near work may yield a more economical outcome. The scenarios represent common decisions by any company or newcomer to Seattle. Whoever utilizes information the best is commonly the victor in today’s age.

There are only 1113 venues across 90 categories in 82 different neighborhoods. Most venues did not have both pricing and rating information. However, over one thousand venues are more than enough to establish price levels and quality. The preparation of data is identical to the past clustering. The only additions are venue pricing average and venue ratings average per neighborhood. It should extenuate the similarities and differences between neighborhoods seen prior.

Below you can see *Figure 6* and *Table 4.* The colored key is identical. However, keep in mind this does not mean the neighborhoods will fall in the same clusters. Nor do the clusters represent any similarities with their past cluster classification. Unlike *Figure 5*, *Figure 6* has two major clusters in central Seattle, clusters zero and four. Cluster zero neighborhoods contain venues with lower ratings, but around the same price as cluster four. Cluster four has many of the highest-ranking venues. There are a couple neighborhoods averaging above a price tier of two. However, most of the neighborhoods range from 1.5 to 1.9 price tier. Therefore, it may be advantageous to move from cluster zero to four. On average the pricing is about the same but has higher rankings. Perhaps the most notable comparison is clusters one and three. Cluster one has the highest-ranking venues at the lowest cost. Many of the neighborhoods have an average rating of eight and above with price tier one. The wide spread of the cluster shown on the map allows businesses and individuals flexibility. Each neighborhood has its own set of venues. Cluster three neighborhoods may yield opportunity for business, but not be advantageous for individuals. Similar to cluster one the prices are quite low. However, the range of ratings is around six to seven. Since the bar is low a business may be able to integrate into the ecosystem smoothly. *Figure 7* illustrates the clear divisions of each cluster. The colors are identical to *Figure 5* and *Figure 6*.



***Figure 6 and Table 4:*** *K-Means Clustering Interactive Map*



***Figure 7 and Table 4:*** *Clustering Differences in Price and Rating*

**5. Conclusion**

Identifying which neighborhood is best for business expansion is complex. A company must understand the expectations of the consumer, their propensity to spend, and know the competition. Many commercial leases can last multiple years, even decades. Thus, pricing information and ratings is essential. *Figure 6* and *Figure 7* enable a company to make the informed decision. Cluster three has a wide spread of neighborhoods across Seattle. The lower average ratings and price can help small to medium businesses get a footing. In contrast, Central Seattle sees the most customer traffic. Therefore, medium to large businesses can look to expand in cluster zero. Cluster zero has the same pricing range as cluster three, expect the ratings are lower. It is still a risk, but with the right placement a company can exceed their limitations.

An individual or family can now make more informed decisions on where they live. Everyone has their own needs and wants. Therefore, a tailored experience is necessary. Luckily, anyone can find similar venues they enjoy. They can look to move there given it is within their price range. A great example is from *Figure 7* cluster one. Every neighborhood has above average rating and below average pricing. The selection of common venues includes pizza places, bakeries, breweries, and diners. Compared to almost any other cluster a move here would yield cost savings and a high probability of increasing happiness.