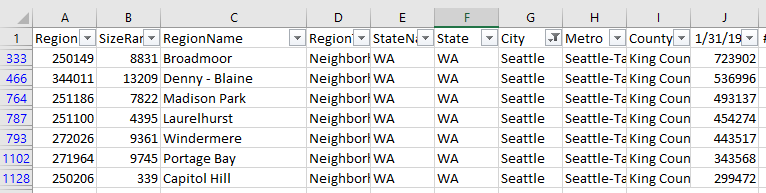
**Introduction**

The Seattle housing market has been extremely competitive for some time now, largely driven by high paying jobs from tech giants like Microsoft and Amazon. The prospect of buying a home in this environment can be quite daunting. Many people who move to Seattle initially live close to downtown in neighborhoods like Capitol Hill or Belltown and it is here that many grow to love the city. For many of us though, when it comes time to contemplate home ownership, we quickly realize that buying a home in those same trendy neighborhoods is out of reach for most, particularly when it comes to single family homes. The central problem of this capstone project is to find Seattle neighborhoods with lower average home costs that are the most similar to more centrally located and trendy neighborhoods such as Capitol Hill.

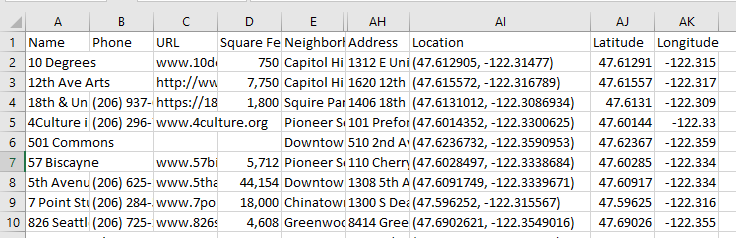
**Data**

The data used for this project will come from several different sources. Housing price data will come from Zillow (<https://www.zillow.com/research/data/>). Two different datasets will be pulled from Zillow, one that includes all home types and one which only includes single family homes. Seattle neighborhood latitude and longitude data will come from the city of Seattle’s open data program. The specific dataset used will be the Seattle Cultural Space Inventory (<https://data.seattle.gov/Community/Seattle-Cultural-Space-Inventory/vsxr-aydq>) as it conveniently contains both neighborhood and location data. However, this data will also require considerable cleaning as it contains multiple entries per neighborhood and additional unnecessary data. Multiple latitude and longitude entries exist for each neighborhood, the first entry of each will be taken. Folium will be used to generate maps of the neighborhoods and will further be used to ensure that the latitude and longitudes make sense. In instances in which the Seattle Cultural Space Inventory data is not accurate, entries will be manually updated with latitude and longitude from Google Maps and domain knowledge. This location data will be used in combination with the Foursquare API to pull relevant location data (venues, etc.…) for each of the neighborhoods.

Example of Zillow data:



Example of Seattle Cultural Space Inventory data:

Example of Foursquare API data:

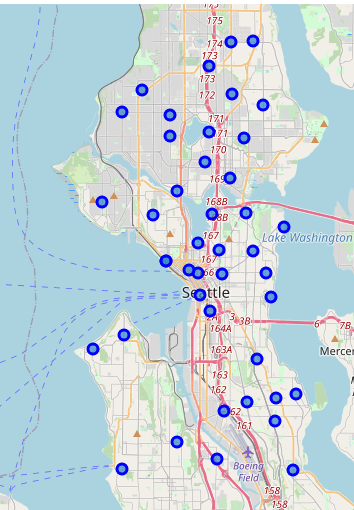


**Methodology**

Given the limitations of the Seattle Cultural Space dataset (1214 entries, extraneous columns, multiple entries, missing data, etc.…), extensive up-front cleaning was necessary. First, given that the neighborhood information was the most crucial piece of information, any rows that had missing values in the Neighborhood column were dropped. Next we removed duplicate neighborhood entries and finally we dropped any row which was missing latitude data (assuming any row missing latitude data would also be missing longitude). Finally, we dropped all columns except the relevant Neighborhood, Latitude, and Longitude columns.

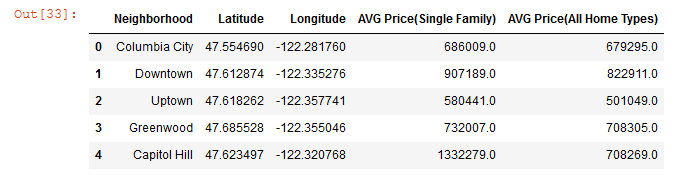
The Folium library was used to generate a map of Seattle and to visualize the newly cleaned data. Manual observation of this map revealed several issues. First there were two entries for Green Lake, one spelled correctly as “Green Lake” and another spelled incorrectly as “Greenlake.” In addition, it was observed that six neighborhoods (Capitol Hill, Madrona, Wedgwood, Eastlake, Phinney Ridge, and Beacon Hill) were not in the correct locations. The first issue was solved by deleting the misspelled Green Lake entry as it was also not in the correct position. The neighborhoods with incorrect locations were manually updated with the correct latitudes and longitudes (obtained using domain knowledge and Google Maps).

Visualization of included Neighborhoods:



Housing data from Zillow came in the form of two CSV’s, single family home and all home housing prices. Both datasets contain data for all cities in the United States in addition to extraneous columns such as State, City, County Name, etc.… Both data sets were winnowed to only contain entries for Seattle and extraneous columns were dropped. Housing prices are broken up into monthly chunks going back to 1996, furthermore these data ranges are used as columns, negatively impacting readability. The data was transposed so that each column represents a Neighborhood and each row represents monthly housing prices. Historical housing prices are not useful when trying to make comparisons in the current housing market therefore only the last two years of housing prices were considered, i.e. 2019 to present. Exploratory data analysis also revealed that the Zillow datasets break the Queen Anne neighborhood into 4 sub neighborhoods. For the purpose of this analysis only the Lower Queen Anne data was used as its average home prices for both single family and all homes was closest to the mean of all four sub-neighborhoods. Finally, the mean of home prices for the given data ranges were calculated on a per neighborhood basis. These dataframes were then merged to the Neighborhood dataframe using an inner join so that the resultant dataframe only has complete information. The most and least expensive neighborhoods for single family and all homes were calculated.

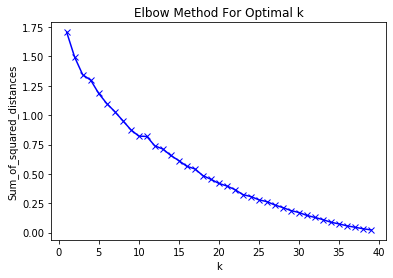
Final Merged Dataframe:



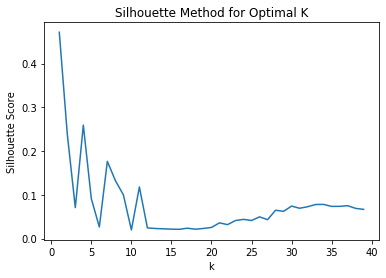
The Foursquare API was used to find venue information for each neighborhood. Up to 100 venues were pulled with a radius of 750 meters of the given latitude and longitude of each neighborhood. For some neighborhoods located close together this may have resulted in some unavoidable overlap. Venue information was transformed using one hot encoding and then the mean of each venue was found by neighborhood. The ten most common venues for each neighborhood was calculated.

K-means clustering was chosen to cluster the neighborhoods as it is the algorithm with which I have the most experience. To find the optimal number of clusters K with which to group the neighborhoods in to, a combination of the elbow and silhouette methods were used. With the elbow method, the sum of squared distances is graphed against K the point K where the slope of the line begins to level out is selected. The sum of squared distances will tend to zero as K increases towards the number of values in the dataset.

Elbow Method:

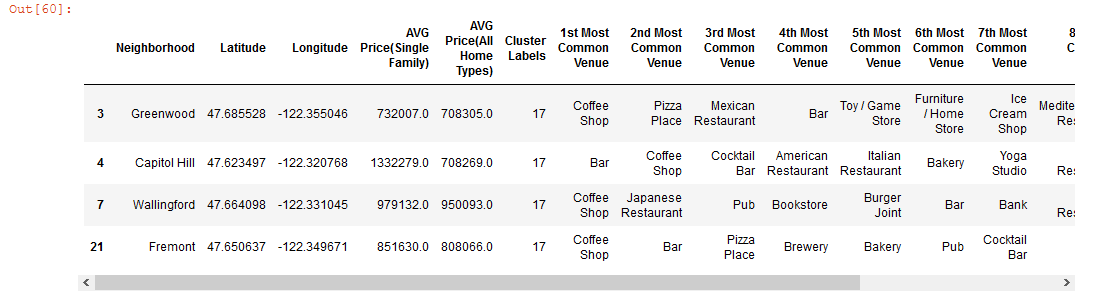


The choice of where the elbow is located is subjective, thus it can be combined with the silhouette method to verify choice of K. The silhouette method attempts to quantify entries similarity to each in a given number of clusters K. The higher the silhouette score, the more similar entries in a cluster are to each other. Ideally, the cluster number K which has both the highest silhouette score and corresponds to the elbow would be chosen. In this application, diversity of clusters must also be considered. The analysis means nothing if the target neighborhood is in a cluster by itself. Thus, a K of 34 was chosen as it corresponds to the elbow point, has a reasonably high silhouette score, and results in adequate cluster diversity.

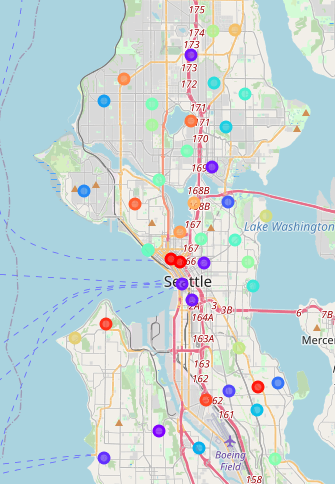
Silhouette Method: 

**Results**

Capitol Hill was assigned to cluster 17 which contains 3 other members: Fremont, Greenwood, and Wallingford. Observation of their most common venue information reveals that they all have a high prevalence of coffee shops, bars, and restaurants. According to this analysis these three neighborhoods are the most similar ones to Capitol Hill. In combination with the Zillow housing price data we see that these three neighborhoods all have lower average single-family home prices, while interestingly Capitol Hill has the lowest average home price for all homes. Greenwood has the lowest average single-family home price and only has a slightly higher average home price for all homes, as such it is the best candidate for someone looking to buy a home in a neighborhood that is similar to Capitol Hill.

Cluster containing Capitol Hill:

Visualization of clustered neighborhoods:



**Discussion**

The most surprising result of this analysis is that Capitol Hill has the lowest average housing prices when looking at all homes. This is likely due to the higher proportion of apartments, condos, and townhomes in the neighborhood though this would likely warrant a closer analysis. A shortcoming of this analysis is the dataset from which neighborhood location data was derived. The Seattle Cultural Space inventory has the advantage of having both neighborhood and latitude/longitude labels in a readily available format. However, as was seen in the data section, this data was not accurate for every entry. In addition, small things like the “Greenlake” and “Green Lake” entries probably occurred in other places but were not caught. This analysis would be stronger with more precise neighborhood location data.

The Zillow average housing prices for all homes is probably not the best metric (though most convenient) with which to compare between neighborhoods. Single family homes provided more delineation, however, picking the type and/or desired number of bedrooms of home (all of which is available from Zillow) which is desired would provide the best comparison between neighborhoods. The time series over which average prices are calculated could also probably be adjusted for more clarity. If a person is buying a house presently then the market value of a year ago is of little consequence.

Finally, K-means may not be the best clustering method available for this application. The fact that 34 clusters where needed when there were only 43 neighborhoods is slightly concerning. Seattle’s neighborhoods are quite unique, maybe just not *that* unique. This is reinforced by the fact that the silhouette score for 34 clusters is so low (~0.1) suggesting that inter-cluster similarity is not terribly strong. The fact that the only data used for clustering was Foursquare venue data may also have hampered analysis. Having additional information such as home type distribution (geographical and numerical) or demographic information could have helped to form more similar and distinct clusters.

**Conclusion**

Despite all of these challenges, the present analysis provides a strong starting point from which to build better datasets and algorithms. It also provides 3 neighborhoods which are more similar to Capitol Hill than they are to any other neighborhoods considered. When making such a long-term decision as buying a home more information is always preferred; this analysis provides 3 neighborhoods in which to start looking.

**References**

* Seattle Cultural Spaces Inventory: <https://data.seattle.gov/Community/Seattle-Cultural-Space-Inventory/vsxr-aydq>
* Zillow housing prices data: <https://data.seattle.gov/Community/Seattle-Cultural-Space-Inventory/vsxr-aydq>
* Google Maps: https://www.google.com/maps