Effects of nearby venues on housing prices in Seattle

By: Yash Singh

The problem

When looking to move to a new place, there are so many things that we consider. So many variables we juggle, to try to find a sort of right fit. However, the one thing that becomes the major constraint is the housing price. After which we rank things as per our own preference, one way to look at a place would be to consider neighborhoods, the availability of housing and the places or venues nearby. The last one is really intriguing to me, as the places can be varied to a great extent. There can be places that everyone wants to live close to (e.g. public transport, grocery stores, etc.), other times we may avoid some places completely (e.g. factories, airports), while there are many that depend on personal preferences (e.g. different restaurants).

In this project, I wanted to look at what affect do nearby venues have on housing market prices. I have decided to perform this study on the city of Seattle as it is where I currently reside. The city of Seattle has a population of around 745,000 people, with over 2 million people living around the area in King county [1]. Seattle is surrounded by water with a total of 200 miles of shoreline [1], and sits close to various hiking trails, national parks and much more [1]. It is also the city with highest number of bookstores and libraries per capita. This showcases the variety of places and therefore we can see their impact on housing prices.

Data sources and description

To consider the problem, we are going to be utilizing

- Zillow housing data: Data provided by Zillow for housing prices and rent prices in the various neighborhoods of Seattle. In this dataset we would be looking at the Zillow Home Value Index (ZHVI) and Zillow Rent Index (ZRI). We shall only be looking at the current values of both ZHVI and ZRI, and later would look into various venues, to see if they affect rent prices or home values differently. Data taken from Zillow in July 2020.

- American Community Survey (ACS): The ACS 5 year-survey (2013-2017) focusing on the neighborhoods of Seattle is taken from Seattle.gov. This dataset contains the basic demographics of each neighborhood and provides various important metrics like household income and number of people in neighborhoods. It is according to this datasets demarcation that we define the various neighborhoods in the city. According to which there are 53 neighborhood districts. We will first look at this dataset to build our first model, which will form the control subject of our test. This will explain the effects of demographics and then we can build upon the differences caused by nearby venues.

- Foursquare API: Lastly, we will look at nearby venues using the Foursquare API. This would provide us with venues in a neighborhood and their associated details.

- In the ACS dataset, there were no coordinates for a neighborhood and since I wanted to map the results, I utilized Google Maps to find centralized positions for each neighborhood, and then use these as the positions for the coordinates of the neighborhoods. This data was attached to the ACS dataset.

Methodology

Data Cleaning and Feature Selection

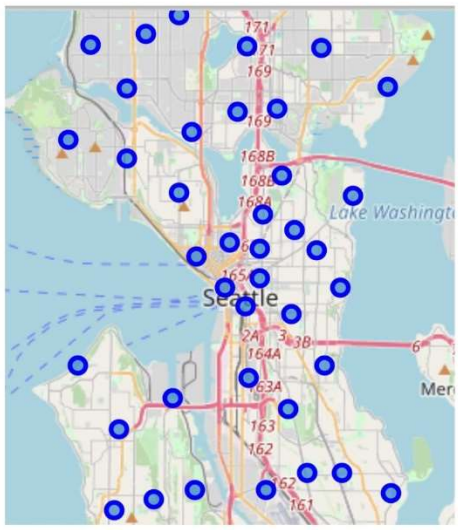
After finding the data sources, I cleaned each dataset, this involved understanding the basic features of each dataset determining which ones can help point me to the right direction. The largest dataset was the ACS survey, with a wide variety of topics from economic features (household income, labor force, etc.) to demographics (race make-up, age make-up of each area). I chose multiple features from ACS like population density and housing density to help understand the differences in the neighborhoods and how the housing may compare. This dataset also provided us with the basic outlines for the neighborhoods, with a total of 53 neighborhoods. From the Zillow dataset, I mainly focused on the Zillow Home Value Index (ZHVI) and the Zillow Rent Index (ZRI) and in those, I only focused on the current value of housing as per July 2020. Zillow's neighborhoods were divided further than the ACS neighborhoods, so I combed through each neighborhood and grouped them based on the outline of ACS neighborhoods using google maps and the Seattle GIS. For each group the ZHVI and ZRI were calculated by taking the mean value of all the small neighborhoods that formed the larger neighborhood. One neighborhood did not have sufficient data (ZHVI and ZRI), which was Licton Springs. Given that the study was focused on the housing prices, I decided to exclude the dataset. There were some other neighborhoods missing ZRI data, I chose to keep the neighborhoods on account of ZHVI.

	Community Reporting Area Name	Area in Acres	Total Population	Median Age	Total Housing Units	Occupied Housing Units (%)	Population per Gross Acre	Housing Units per Gross Acre	Latitude	Longitude	Home Value Index	Rent Index
0	Ballard	492.9	8649	34.3	5580	95.5	17.55	11.32	47.665804	-122.379226	765650.0	2724.0
1	North Beach/Blue Ridge	1284.4	12701	42.6	5532	96.0	9.89	4.31	47.699993	-122.374384	1082000.0	3200.0
2	Montlake/Portage Bay	951.1	9732	37.3	4831	95.9	10.23	5.08	47.639969	-122.312162	1240050.0	3569.0
3	Interbay	1214.6	11024	34.4	6114	91.9	9.08	5.03	47.645169	-122.379354	742400.0	NaN
4	North Capitol Hill	283.7	4807	36.1	2599	95.5	16.94	9.16	47.628489	-122.320303	710600.0	2472.0

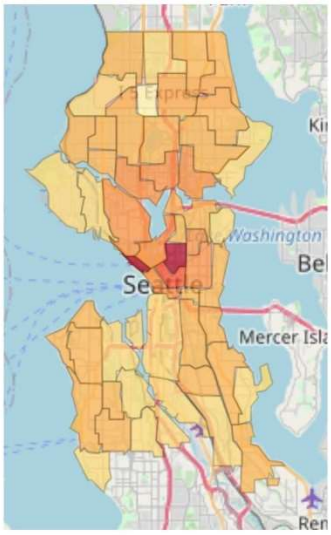
Features kept: Neighborhood (Community reporting area name), Area in acres, Total population, Median Age, Total Housing Units, Occupancy Rate (Occupied Housing Units (%)), Population Density (Population per gross acre), Housing density (Housing units per gross acre), Latitude, Longitude, ZHVI (Home Value Index) and ZRI (Rent Index).

Exploratory Data Analysis

To start out, I visualized the neighborhood center as a way to showcase the location of each neighborhood across the city. Then, I decided to see if there are any differences in population density and housing density.



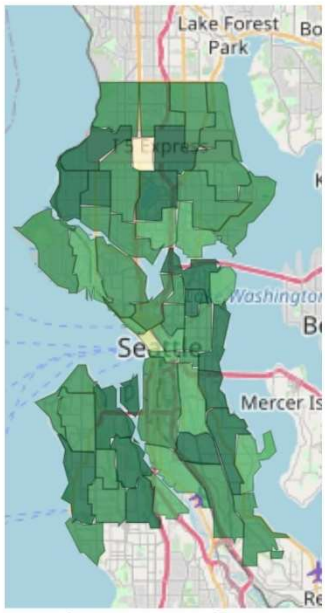
Neighbourhood Centers



Population Density

Housing Density

At first glance the population and housing density seem to be similar, however, we see that the population density is higher in some areas than others, which is not represented by the housing density. This can be seen in the outskirts of the city with the top and bottom most areas having little difference in housing density but differences in population density. To investigate this, I thought that I should check the occupancy rates in each neighborhood. I also decided to see visualize the ZHVI and ZRI across the neighborhoods to get a better idea of what could cause this, and what other questions I can raise.



Occupancy Rate


House Value Index

Rent Index

The occupancy rate map showcased that there are two neighborhoods with low occupancy, which were Licton Springs and Downtown (Commercial Core). Since we do not have the data for Licton Springs, I looked into the bottom 5 neighborhoods in terms of occupancy.

	Community Reporting Area Name	Area in Acres	Total Population	Median Age	Total Housing Units	Occupied Housing Units (%)	Population per Gross Acre	Housing Units per Gross Acre	Latitude	Longitude	Home Value Index	Rent Index
42	Downtown Commercial Core	206.5	4872	43.3	3480	79.0	23.59	16.85	47.606689	-122.336548	815000.0	2656.0
17	Georgetown	1183.9	1045	41.0	697	89.0	0.88	0.59	47.546259	-122.319237	607300.0	NaN
31	First Hill	440.9	16895	33.4	9952	89.3	38.32	22.57	47.609422	-122.321996	632200.0	2113.0
38	Pioneer Square/International District	180.6	5289	46.6	3795	90.1	29.29	21.01	47.601071	-122.327994	548050.0	NaN
29	Belltown	176.5	10282	35.6	7719	90.5	58.25	43.73	47.615817	-122.349439	572000.0	2302.0

This is where we see that other than Georgetown, the other neighborhoods surround Downtown. There is a noticeable difference in Downtown's occupancy rate of 79% compared to the next lowest of 89%. However, this can be explained by the home value index, as even the surrounding areas to downtown (First hill, Pioneer square, Belltown) have much lower home values. Looking at the housing density map, we can also see that these areas have a particularly high housing density. One reason for the low prices and low occupancy rates in these neighborhoods (excluding Downtown) could be the high number of houses in them. The difference between them and Downtown could be one of nearby venues.

Nearby Venues

To get nearby venues, we utilized Foursquare API and the centralized locations of each neighborhood. We limited each neighborhood to 100 venues and chose a radius of 1000 meters. This radius was the approximate distance between the two closest neighborhood centers. From the retrieval, we got 2868 venues in total. We then utilized one-hot encoding to get the frequency of each venue category for each neighborhood, then we grouped all venues by neighborhood, taking the distribution of each venue category throughout the city. That is, each venue category is normalized to represent a total contribution of 1, and each neighborhood represents the contribution to the whole venue category. This tells us which neighborhoods have the most coffee shops, but also that coffee shops do not just by there numbers overrepresent other categories like theaters or parks.

We then picked the top 10 most common venues for each neighborhood. Here is what that looks like.

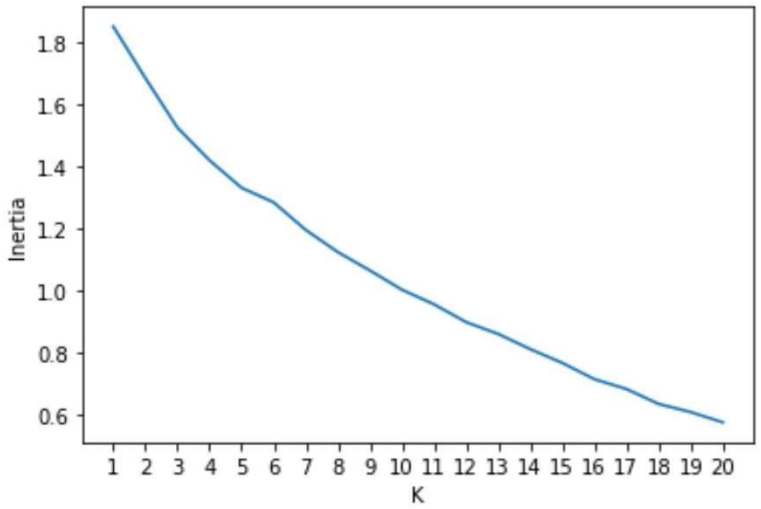
	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alki/Admiral	Ice Cream Shop	Coffee Shop	Trail	Park	Beach	Food Truck	Thai Restaurant	Mexican Restaurant	Clothing Store	Seafood Restaurant
1	Arbor Heights	Home Service	Pool	Bowling Alley	Flower Shop	Trail	Bus Station	Other Repair Shop	Ethiopian Restaurant	Donut Shop	Dry Cleaner
2	Ballard	Brewery	Coffee Shop	Mexican Restaurant	Bar	New American Restaurant	Sandwich Place	Cocktail Bar	Pizza Place	Ice Cream Shop	Clothing Store
3	Beacon Hill	Intersection	Pizza Place	Grocery Store	Sporting Goods Shop	Supermarket	Gas Station	Garden Center	Bar	Café	Marijuana Dispensary
4	Belltown	Coffee Shop	Sushi Restaurant	Sculpture Garden	Pizza Place	Bar	Breakfast Spot	Bakery	Hotel	Movie Theater	Hotel Bar

Next, I decided to cluster similar neighborhoods to see trends across the city. To do so, I utilized the K-Means algorithm, which is an unsupervised machine learning technique. The first thing that I had to look for was determining the value of K.

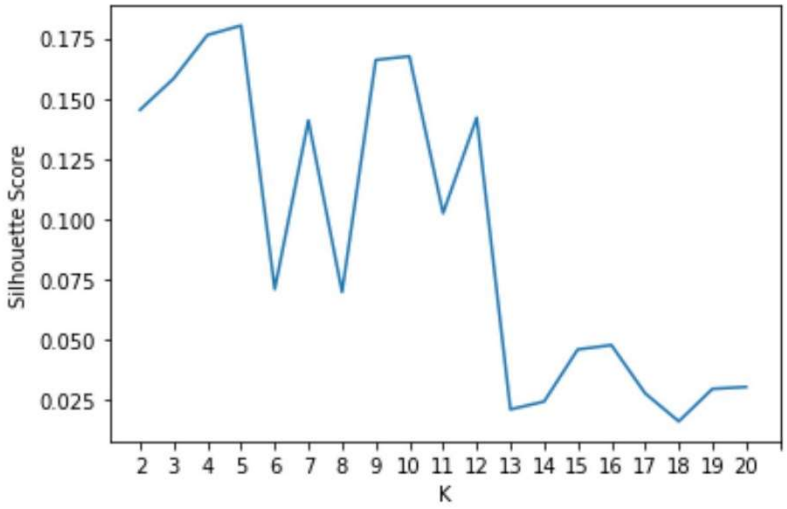
K-Means Classifier

Finding K

To get the optimum value of K, I ran the elbow point test. In which, I compared the inertia over k ranging from k = 1 to k = 20. The inertia is the sum of squared distances of all points from closest cluster center, which represents the error in our model.



Since there is no clear elbow point in the chart, I looked at the silhouette scores, which score the distance between two cluster centers. Here we look to maximize the distance between two cluster centers.



The silhouette scores tell us that the optimum value for K is 5. Now, we run the K-Means algorithm with K=5.

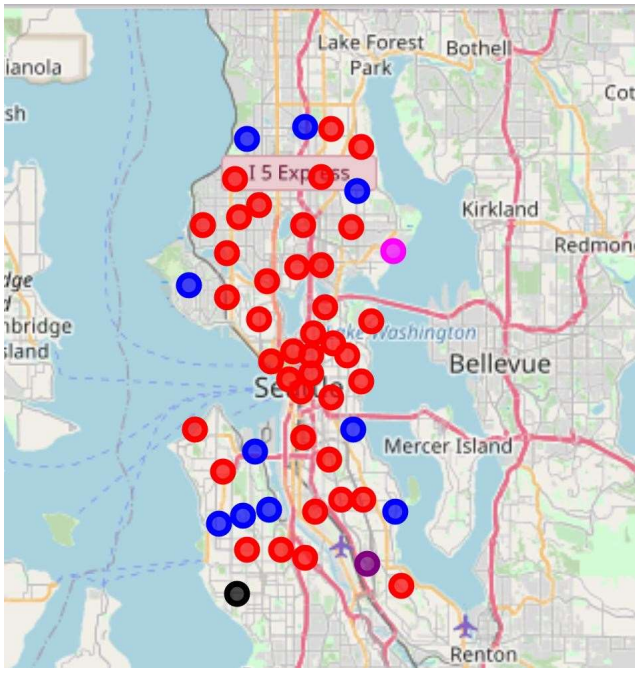
Clustering

We ran the K-Means algorithm to classify each neighborhood into 5 clusters based solely on the 10 most common venue categories in each neighborhood. We then visualized this by putting the different clusters on a map.

Results and Discussion

After running K-Means, we got the following cluster. I numbered the clusters as follows:

Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5



Now we examine each cluster and see what makes them different and how this relates to hosing prices.

Cluster 1

Latitude	Longitude	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
47.665804	-122.379226	...	Brewery	Coffee Shop	Mexican Restaurant	Bar	New American Restaurant	Sandwich Place	Cocktail Bar	Pizza Place	Ice Cream Shop	Clothing Store
47.699993	-122.374384	...	Pizza Place	Coffee Shop	Pet Store	Asian Restaurant	Mexican Restaurant	Chinese Restaurant	Thai Restaurant	Massage Studio	Baby Store	Taco Place
47.639969	-122.312162	...	Garden	Playground	Coffee Shop	Scenic Lookout	Park	Bus Stop	Cemetery	Bike Shop	Steakhouse	Dog Run
47.645169	-122.379354	...	Bus Stop	Golf Course	Park	Asian Restaurant	Gym	Grocery Store	Coffee Shop	Burger Joint	Boat or Ferry	Bar
47.628489	-122.320303	...	Coffee Shop	Cocktail Bar	Bar	American Restaurant	Yoga Studio	Park	Thai Restaurant	Garden	Seafood Restaurant	Italian Restaurant
47.618192	-122.321713	...	Coffee Shop	Pizza Place	Bar	Mexican Restaurant	Thai Restaurant	Café	Ice Cream Shop	Cocktail Bar	Italian Restaurant	Bakery
47.681815	-122.371196	...	Park	Deli / Bodega	Pizza Place	Food Truck	Coffee Shop	Thai Restaurant	Café	Restaurant	Pub	Soccer Field

Cluster 2

ed ng %)	Population per Gross Acre	Housing Units per Gross Acre	Latitude	Longitude	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
.0	7.92	3.2	47.666513	-122.265822	...	Massage Studio	Hardware Store	Bank	Automotive Shop	Café	Greek Restaurant	Dry Cleaner	Pizza Place	American Restaurant	Event Space

Cluster 3

pied sing (%)	Population per Gross Acre	Housing Units per Gross Acre	Latitude	Longitude	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
94.3	7.99	3.64	47.508091	-122.372651	...	Home Service	Pool	Bowling Alley	Flower Shop	Trail	Bus Station	Other Repair Shop	Ethiopian Restaurant	Donut Shop	Dry Cleaner

Cluster 4

upied using ts (%)	Population per Gross Acre	Housing Units per Gross Acre	Latitude	Longitude	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
95.6	9.9	3.22	47.52185	-122.283226	...	Light Rail Station	Mexican Restaurant	Pool	Tennis Court	Plane	Rental Car Location	Deli / Bodega	Business Service	Moving Target	Plaza

Cluster 5

ion oss cre	Housing Units per Gross Acre	Latitude	Longitude	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
.77	4.60	47.693939	-122.289757	...	Park	Bus Line	Pub	Coffee Shop	ATM	Grocery Store	Supermarket	Chinese Restaurant	Italian Restaurant	Video Store
.53	1.85	47.573781	-122.359622	...	Park	Coffee Shop	Food Truck	Bus Station	Beer Bar	Bar	Dive Shop	Sandwich Place	Diner	Restaurant
.19	5.35	47.717887	-122.366402	...	Trail	Pizza Place	Thai Restaurant	Food Truck	Beer Bar	Video Store	Bus Station	Antique Shop	Waterfront	Sushi Restaurant
.07	5.83	47.543675	-122.367893	...	Convenience Store	Food Truck	Park	Playground	Rental Car Location	Field	Storage Facility	Bus Line	Dive Bar	Deli / Bodega
.96	3.75	47.545644	-122.264726	...	Park	Trail	Playground	Video Store	Nature Preserve	Seafood Restaurant	Arts & Crafts Store	Intersection	Greek Restaurant	Salon / Barbershop
.93	5.11	47.540527	-122.385239	...	Park	Coffee Shop	Sandwich Place	Food Truck	Pub	Pizza Place	Pet Store	Bookstore	Beach	Deli / Bodega
.96	2.95	47.650742	-122.405051	...	Park	Bus Stop	Trail	Playground	Video Store	Tennis Court	Baseball Field	Chinese Restaurant	Athletics & Sports	Grocery Store

We take the mean values and standard deviations to compare neighborhoods. (Clusters 2, 3 and 4 do not have standard deviations as they have only one neighborhood). We also look at other interesting comparisons.

Observations:

From a quick glance at the clusters, we see that there is a distinction between places that are closer to the city, against those that are not. We see that the city/urban areas have grouped together in cluster 1, whereas the places that are closer to nature or outdoor venues have grouped together in cluster 5. Cluster 1's top venues are mostly restaurants, bars, coffee shops, whereas cluster 5 has parks, gardens, scenic spots as the top venues.

We also notice that clusters 2, 3 and 4 have only one neighborhood. However, we can tell that cluster 3 and 4 are outside the city and the most common areas here are more small shops or recreational activities. Cluster 2 is in the city but has a larger variety of entertainment and leisure venues. Now, let's consider their effects on housing price.

Home Value Means and Standard Deviations for each Cluster:

```
Mean for cluster 1:  817366.8966666667
Mean for cluster 2:  1433000.0
Mean for cluster 3:  624000.0
Mean for cluster 4:  530400.0
Mean for cluster 5:  748575.0
Standard Deviation for cluster 1:  331694.8717674027
Standard Deviation for cluster 5:  158417.24487006527
```

Now, if we compare cluster 1 and 5 on home value, we see that there is a larger variation in pricing for cluster 1 when compared to cluster 5, more than 2 times the variation in cluster 1 than cluster 5. We see that there is a difference in that cluster 1 is more expensive than cluster 5, but not noticeably so.

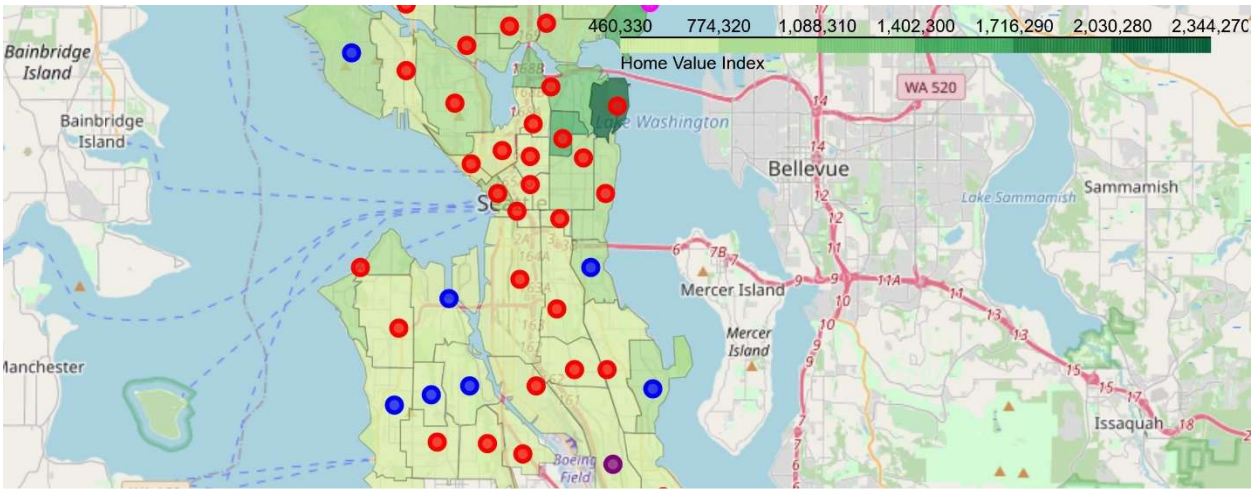
Cluster 3 and 4 which are on the outskirts of the city, and do not have as many eating venues, tend to have much lower housing prices. Cluster 2 on the other hand is in the city but has a variety of different leisure and entertainment venues that may explain the increased housing prices.

```
Population Density for cluster 1:  17.759487179487184
Population Density for cluster 2:  7.92
Population Density for cluster 3:  7.99
Population Density for cluster 4:  9.9
Population Density for cluster 5:  9.056999999999999
```

Another noticeable difference is in the population density of cluster 1 compared to others as cluster 1 has nearly 2 times as many people per acre than any other cluster. This is interesting as we see that densely populated neighborhoods have similar general makeup then those that are not as densely populated. One can assume that limitations of space lead to only a few of each category surviving, but this may show that there is a threshold for a venue category in each densely populated neighborhood.

To end and summarize the project, I created a choropleth map of Seattle, distinguishing the neighborhood on housing price. We then add the central markers to each neighborhood representing

the cluster. Finally, we add popups to the map, that tell us the top 3 most common venues in that neighborhood.



Conclusion

The above data analysis showcases the different impact of venues on housing. We see that there is some clear differences that can be formed by the type of venues, and these can form the basis of explanation for the price differences. I think the best representation of the effects of venues can be seen between cluster 1 (red) and cluster 2 (pink), which are both focused on the city, and yet have a difference in terms of venues, which leads to a housing price differentiation.

References

[1] Visit Seattle website for facts about Seattle: https://visitseattle.org/press/press-kit/seattle-facts/

[2] Geo-data from seattle.gov to form the neighborhood shapes.