Capstone Project - Portland, Oregon's Bicycle Traffic

Lisa Myers
November 14, 2020

1. Introduction

1.1 Background

Over the last decade, Portland, Oregon's traffic congestion has significantly increased. The average Portland driver is spending around 50 hours in rush-hour; making Portland traffic the 12th worst in the nation.

During this time, the Portland Bicycle Plan for 2030 was adopted unanimously by Portland's City Council on February 11, 2010; which calls for more than a quarter of all trips to be made by bicycling by 2030. The plan's core elements are to make bicycling more convenient, comfortable, and accessible to more people throughout Portland.

The physical proximity of the bicycle routes, shopping venues, parking areas and affordable housing are important factors in the successful implementation of this plan. In order for commuters to choose cycling over driving, the shopping venues need to be grouped together, routes to affordable housing needs to be short and parking needs to be close to venues.

Therefore, it is advantageous for Portland to understand how currently placed bicycle routes, parking, stores, and affordable housing affect cyclist ridership. Conversely, this information can be used to target future locations of new routes, appropriate new venues, affordable housing, and new bicycle parking strategies.

1.2 Problem

In this project I will try to determine if Portland, Oregon's neighborhoods can increase bicycle ridership.

This can be done by looking at the current:

- number of existing bicycle routes,
- number of existing bicycle parking areas,
- number of and distance to existing venues (which are appropriate for bicycle commuters), if any
- median house sale price for each Portland neighborhood

1.3 Interest

Not only would the Portland's City Council be interested in understanding cyclist commuting needs, the Department of Transportation and other bicycle services (ex. Nike's Biketown) could better understand their impact in Portland. Accurate location of bicycle infrastructure would provide better budgeting, help resolve traffic congestion problems, reduce vehicle emissions, and ultimately increase the livability of Portland Oregon.

2. Data

2.1 Data sources

These sources will be used to extract/generate the required information:

- *Coordinates* of **Portland** center will be obtained using **GeoPy API geocoding** for Portland, Or location.

- *Bicycle parking* will be extracted from: PortlandMaps - Open Data, updated on 11/4/2020 ... Downloaded *Bicycle_Parking.kml* file and used the pykml parser to read the data with 7618 records.

- *Bicycle routes* will be extracted from: PortlandMaps - Open Data (Bicycle Routes), updated in 11/5/2020 ... Downloaded *Recommended_Bicycle_Route_Points.kml* file and used the pykml parser to read the data with 554 records.

- *Median House Sale Price* will be scraped from a web table: Portland Neighborhoods by the Numbers 2019: The City... web table consists of 96 records of Portland neighborhood data.

- Number of *venues* and *location* in every *neighborhood* will be obtained using **Foursquare API.**

2.2 Data cleaning

Data downloaded or scraped from multiple sources will be combined into one table using outer joins. There is the potential for missing values, so a neighborhood will be eliminated if the datasets don't provide coordinates or Median House Sale Price. I will only use the most current bicycle route and parking data, and will use housing data from 2019, as to get the median from a completed year.

There are several problems with the housing dataset. First, the neighborhood names are all in upper case, so I have decided to make all datasets contain Neighborhoods with uppercase letters. Then there are several neighborhood names that used abbreviations, so I will have to update the

neighborhood names to match the other datasets. Also, the Median House Sales Price is stored as a string with commas. The column will be modified by removing the commas and converting the sale price to type float.

Both the Bicycle Parking and Route datasets will be extracted with coordinates, and will have to be split into the Latitude and Longitude values. Also, these datasets didn't contain address information. I used the reverse API call from **Google Maps API geocoding** to retrieve the Neighborhood from the dataset coordinates.

2.3 Feature selection

After all data transformations and cleansing have occurred, the combined dataset will be reduced to having seven features and 74 rows. Features that don't relate to housing prices or bicycle ridership will be candidates for elimination. An example, all neighborhood divorce information was removed from the housing dataset.

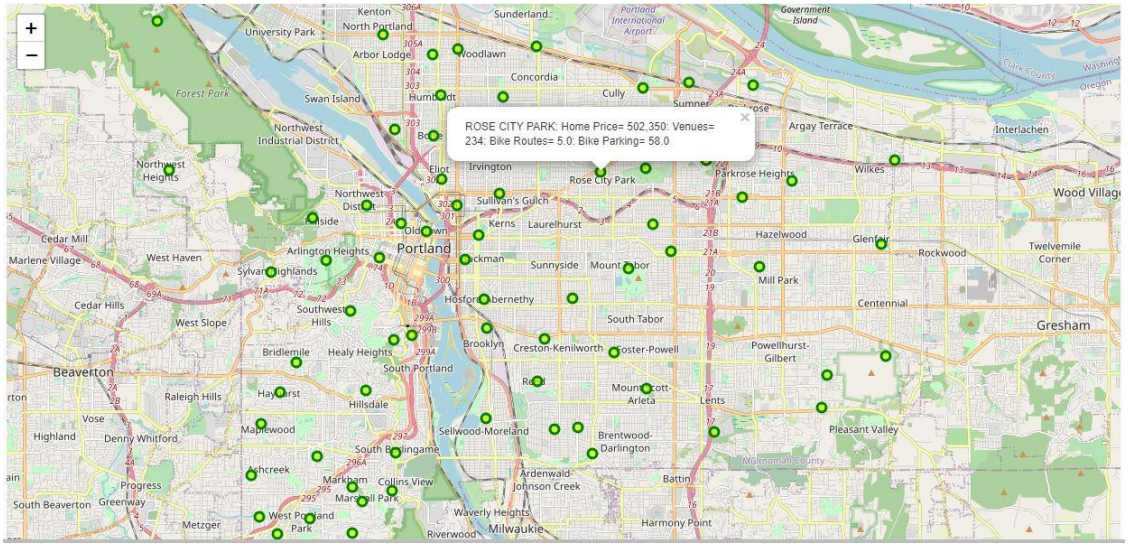
	Neighborhood	Latitude	Longitude	Route Count	Parking Count	MedianHomePrice	Venue Count
0	ALAMEDA	45.549715	-122.637829	1.0	14.0	738000	231
1	ARBOR LODGE	45.576960	-122.689652	1.0	44.0	459450	147
2	ARLINGTON HEIGHTS	45.517047	-122.711445	15.0	8.0	862500	233
3	ARNOLD CREEK	45.444224	-122.701538	6.0	0.0	680000	60
4	ASHCREEK	45.459642	-122.740055	4.0	0.0	455000	146
...	...	...	...	...	...	...	...
69	VERNON	45.560344	-122.644026	3.0	172.0	539250	219
70	WEST PORTLAND PARK	45.448154	-122.717725	4.0	0.0	442000	63
71	WILKES	45.543739	-122.495289	2.0	0.0	330000	52
72	WOODLAWN	45.573263	-122.661376	1.0	56.0	439000	227
73	WOODSTOCK	45.472327	-122.615555	3.0	83.0	467500	175

74 rows × 7 columns

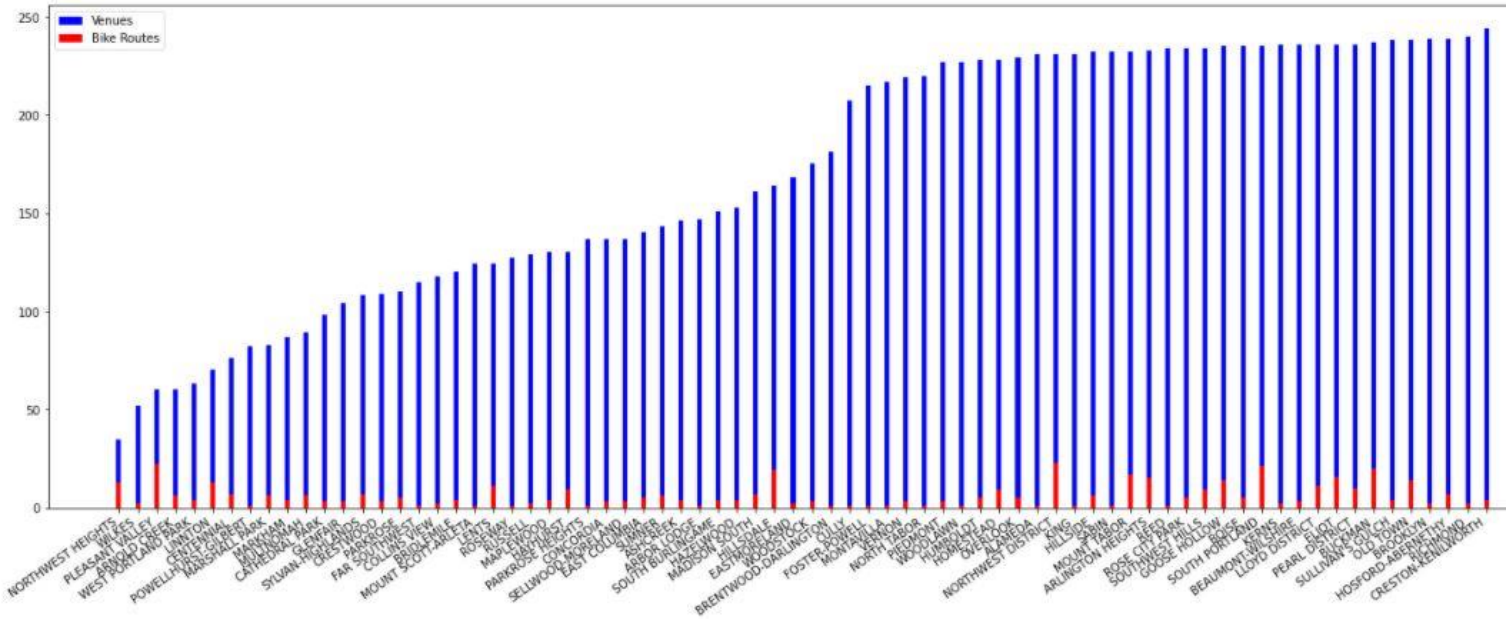
3. Methodology

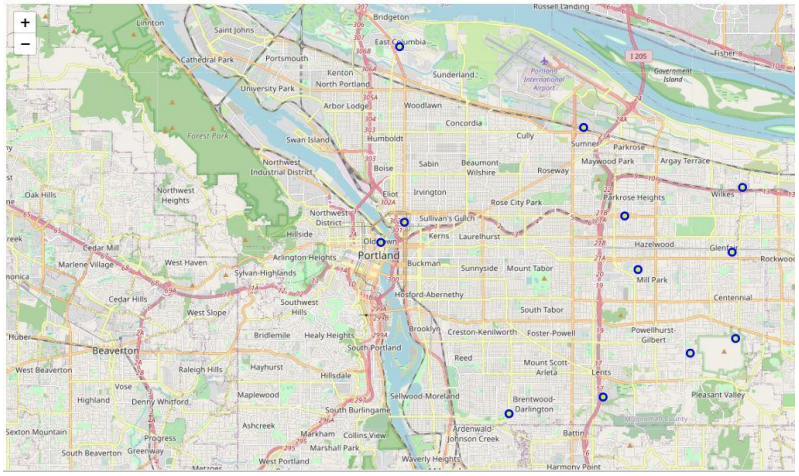
3.1 Exploratory Data Analysis

I used the python visualization library, Folium to map the Neighborhoods on a Leaflet map. Embedded into each neighborhood locator, are the associated feature counts. By clicking on any of the neighborhood locators, I can quickly view the Median Home Price for that neighborhood. Thus allowing the ability to spatially see where the housing is expensive or affordable.

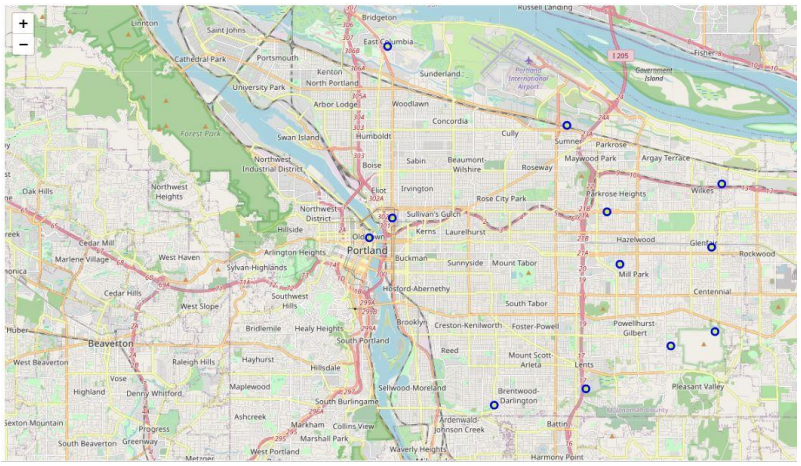


Next, I compared *venue count* to *bicycle count* with a bar chart. Here, I can see the red bars are the bike route counts and the blue bars are the venue counts for each neighborhood.

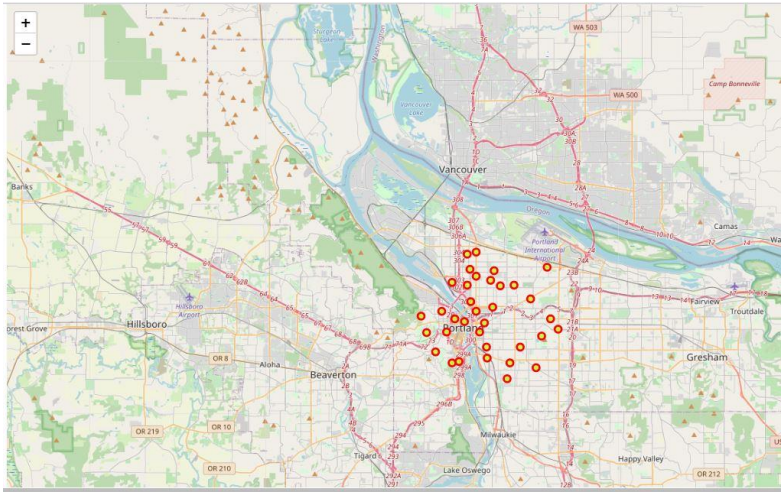




Above is a map showing the neighborhoods that have a median house price less than or equal to $330,000

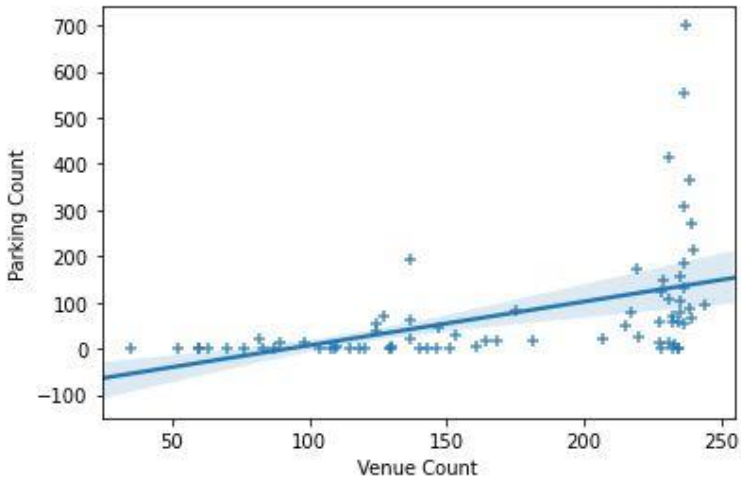


Above is a map of the neighborhoods where the bicycle routes were greater than 15.

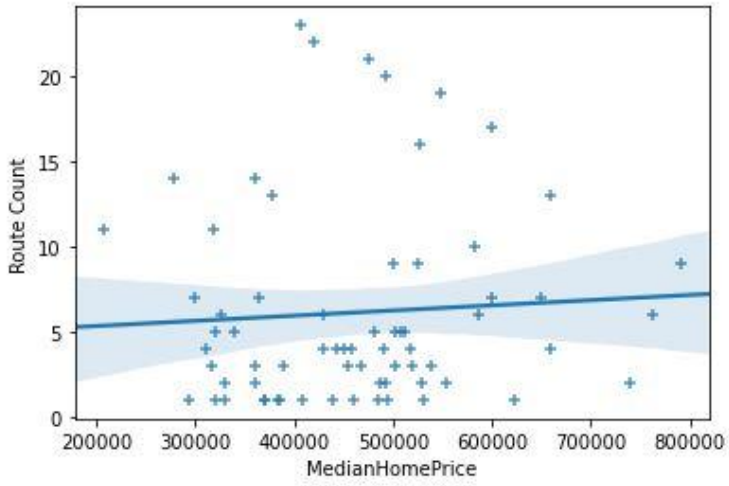


Above is a map of the neighborhoods where the venues were greater than 200.

Several linear regression models were used to determine if one feature has an impact on another.



With the regression line graph of the independent variable (Route Count) and independent variable (Median Home Price), it would appear that there is a moderate relationship between the parking counts on the venue count, which would indicate that planning bicycle parking is considered when planning venue locations.



With the regression line graph of the independent variable (Route Count) and independent variable (Median Home Price), there seems to be no impact.

3.2 Machine Learning

3.2.1 Hot Encoding

I created a secondary dataset that contains the default 100 venues returned from **FourSquare** for each neighborhood; with a radius equaling two miles. I chose the hot encoding process because the venues are categorical variables that can be converted into a form that could be provided to machine learning algorithms.

In this instance, the hot encode process was used to find the frequency of the top 5 venues for each neighborhood:

```
----ALAMEDA----
           venue  freq
0    Coffee Shop  0.07
1            Bar  0.06
2  Grocery Store  0.05
3    Pizza Place  0.04
4           Café  0.04
```

```
----ARBOR LODGE----
         venue  freq
0  Coffee Shop  0.09
1          Bar  0.08
2         Park  0.07
3   Food Truck  0.05
4  Pizza Place  0.04
```

```
----ARLINGTON HEIGHTS----
           venue  freq
0           Park  0.09
1    Coffee Shop  0.07
2        Brewery  0.04
3      Bookstore  0.04
4  Grocery Store  0.04
```

```
----WILKES----
             venue  freq
0      Coffee Shop  0.06
1  Harbor / Marina  0.06
2   Sandwich Place  0.06
3     Intersection  0.04
4  Thai Restaurant  0.04
```

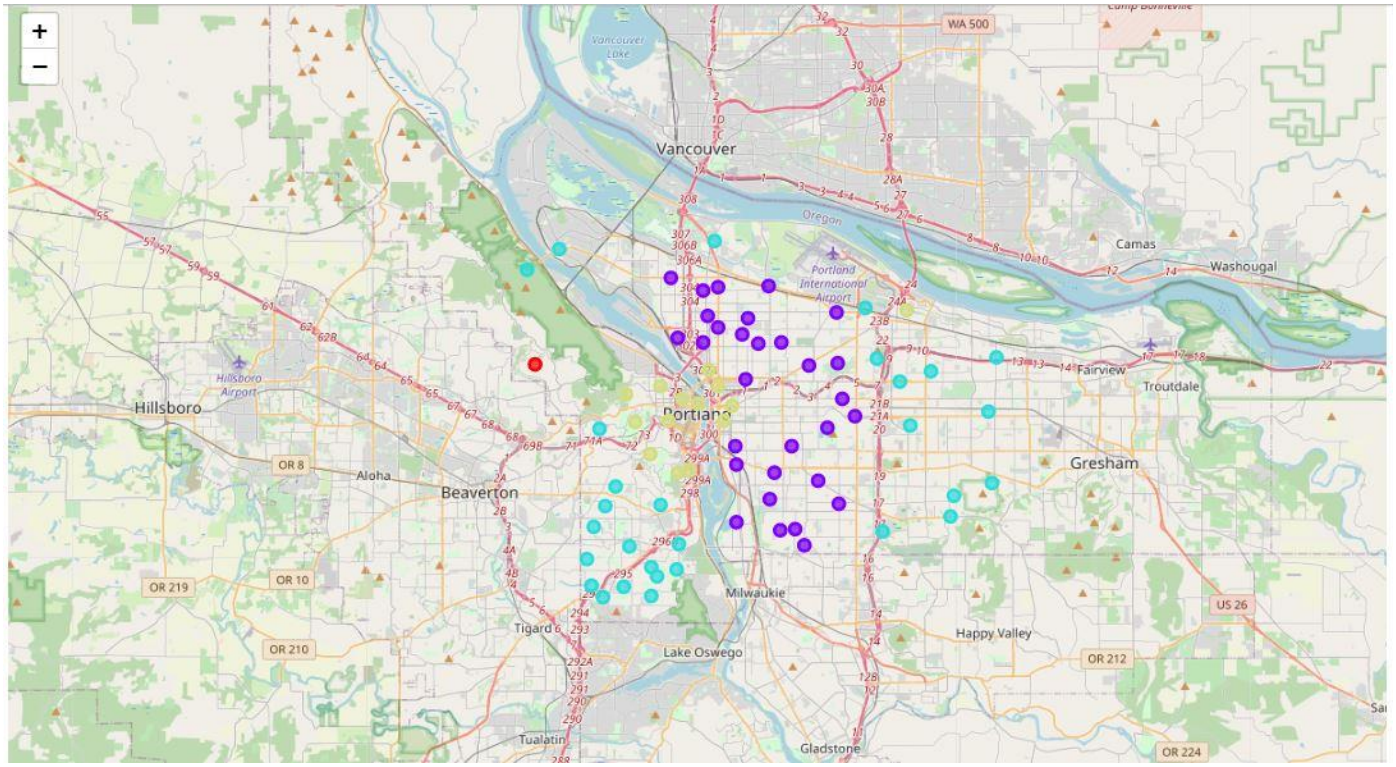
```
----WOODLAWN----
         venue  freq
0  Coffee Shop  0.09
1   Food Truck  0.06
2  Pizza Place  0.05
3          Bar  0.04
4         Park  0.04
```

```
----WOODSTOCK----
                  venue  freq
0           Food Truck  0.06
1              Brewery  0.05
2  Vietnamese Restaurant  0.04
3                 Park  0.04
4    Mexican Restaurant  0.04
```

Next, I ran the *k-mean* to cluster the neighborhood into 4 clusters, with the top 10 venues for each neighborhood.

	Neighborhood	Latitude	Longitude	Route Count	Parking Count	MedianHomePrice	Venue Count	Cluster Labels	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	ALAMEDA	45.549715	-122.637829	1.0	14.0	738000	231	2.0	Coffee Shop	Bar	Grocery Store	Pizza Place	Café	Park	Bakery	Pet Store	Farmers Market	French Restaurant
1	ARBOR LODGE	45.576960	-122.689652	1.0	44.0	459450	147	2.0	Coffee Shop	Bar	Park	Food Truck	Pizza Place	Cocktail Bar	Mexican Restaurant	Breakfast Spot	Beer Store	Beer Garden
2	ARLINGTON HEIGHTS	45.517047	-122.711445	15.0	8.0	862500	233	0.0	Park	Coffee Shop	Bookstore	Grocery Store	Brewery	Pizza Place	Italian Restaurant	Café	Garden	Zoo
3	ARNOLD CREEK	45.444224	-122.701538	6.0	0.0	680000	60	1.0	Grocery Store	Sports Bar	Gym / Fitness Center	Bar	Park	Convenience Store	Pizza Place	Coffee Shop	Café	Burger Joint
4	ASHCREEK	45.459642	-122.740055	4.0	0.0	455000	146	1.0	Coffee Shop	Bar	Sports Bar	Pizza Place	Brewery	Park	Breakfast Spot	Ice Cream Shop	Thai Restaurant	Grocery Store

Here, I rendered the Folium map with the resulting cluster dataset so that I could see the visual distribution of the clusters:



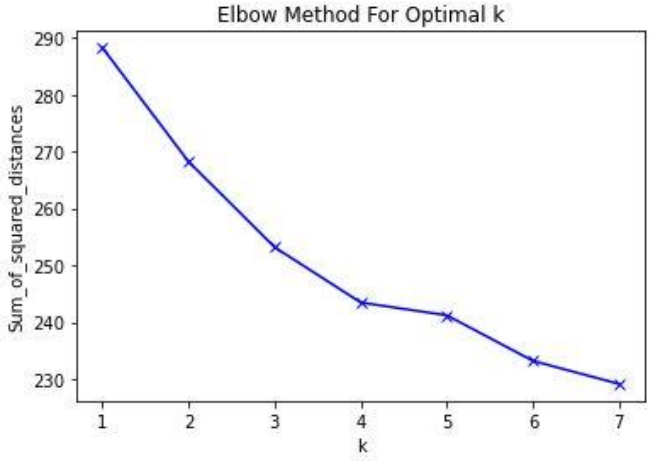
3.2.2 K-Means Clustering

The goal of using the *K-Means Clustering*, unsupervised learning, is to find groups within the unlabeled data (no categories or groups); and this fits my columns of count and price data.

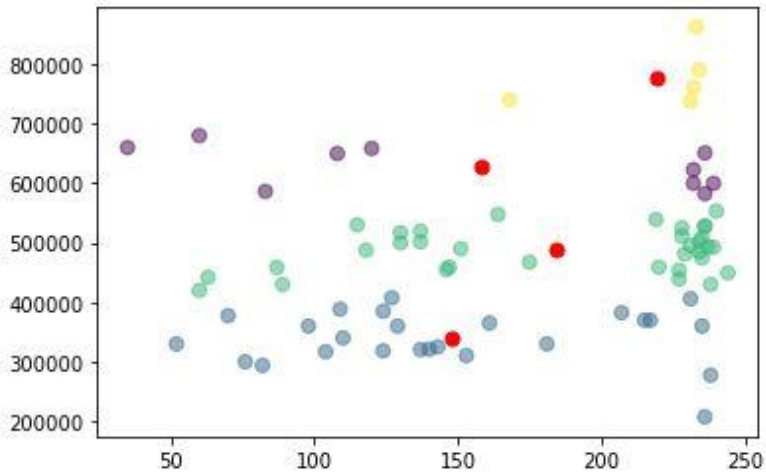
	Venue Count	Route Count	Parking Count	MedianHomePrice
count	74.000000	74.000000	74.000000	74.000000
mean	170.864865	6.040541	74.229730	476762.378378
std	63.721209	5.727701	127.586291	131667.628946
min	35.000000	1.000000	0.000000	207450.000000
25%	121.000000	2.000000	0.000000	371862.500000
50%	171.500000	4.000000	20.500000	470990.000000
75%	232.750000	7.000000	82.000000	529500.000000
max	244.000000	23.000000	700.000000	862500.000000

I ran the *best fit for K* process on the dataset with venue, route, parking, and median home price data.

First I generated the features in a dummy dataset, normalized the data, and processed the results in the *KMeans* model. Next, I plotted the graph with K and sum of squared distances.



Visually I can see that the elbow is starting at 4, so this model is suggesting that k=4 is the best fit.



The scatter plot is showing the four centroids in red, with x=venue count and y=median home price; it appears the data has some similarities.

4. Results

I found that the neighborhoods with the most affordable housing (Median Home Price <= $330,000) were on the outskirts of Portland. This tread also appeared with the neighborhoods with the most bicycle routes. Also, all neighborhoods with higher densities on venues were located near the center of Portland.

The models predict that the venues that will be frequented the most will be Café, bars, and other non-essential shopping, and the cluster map indicates this similarity by being able to cluster most of the neighborhoods into groups.

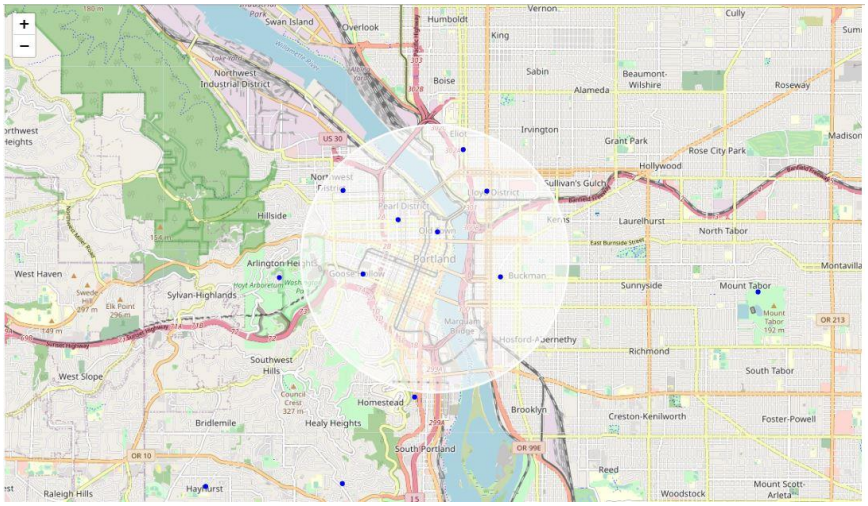
Further analysis indicated that there is a moderate relationship between bicycle parking and neighborhoods that had a higher number of venues, but there was no correlation between Median Home Prices and any of the other features.

5. Discussion

Consider that the median income for a family in Portland is around $50,000, which financial advisers will tell you mean they should not spend more than $315,000 on a house. Also consider that the national average commuting time is 25 minutes each way. So can you find an affordable house in a place that's about a one-hour round-trip commute to downtown Portland by bike?

There are other studies that indicate only 9% of Portland cyclist commuter to work; 53% of Portland cyclist are engaging in recreational activities. Also, the majority of cyclists are of younger age and possibility living near universities and schools.

I believe the biggest opportunity for improvement would be around the center of Portland.



It would be my recommendation to have more targeted bicycle routes in the areas where they connect younger cyclist with essential shopping, work/school and affordable housing. Maybe develop bicycle routes closer to park and ride locations.

4. Conclusion

The purpose of this project was to analyze existing Portland, OR bicycle infrastructure, in order to aid stakeholders in determining how to encourage commuter to increase bicycle ridership. By calculating venue density distribution from FourSquare data, median house sale price data, bicycle route and parking data, we have predicted venue frequencies.

We have concluded that if the stakeholders want to increase bicycle ridership, they should focus efforts on having high value venues near highly traveled bicycle routes, where affordable house in close by and obtainable. At the same time, this should cultivate an environment bringing the cyclist a great sense of pride and community. There are many opportunities to improve the downtown area, and this would have the biggest impact of achieving the goal of increasing bicycle ridership.