INDENTIFYING NEIGHBORHOODS OF VALUE IN A CROWDED SEATTLE MARKET

A Clustering Analysis and Machine Learning Exploration

By C.P.

JUNE 1, 2019

APPLIED DATA SCIENCE CAPSTONE
IBM via Coursera

This document is intended to serve as a template for a future Data Science Consultancy Firm. It will follow a standard Machine Learning Project Checklist.

Table on Contents

Section 1: Summary

A. Introduction: Frame the Problem and Look at the Big Picture

Background
I am a management consultant expanding into the data science realm. I friend of mine is a product manager at one of the FANG companies. He has built an extensive network and amassed a great deal of wealth. He is ready to start investing aggressively on real assets located in the major technology hubs of the U.S. He's interested in buying property, opening restaurants, and potentially other retail establishments.

He has asked me to conduct exploratory analysis in the city of Seattle as a test run. If I do a good job, he is willing to hiring me on as a partner in his future investment firm. The goal is to find any arbitrage opportunities in the already overpriced market. Examples of questions that he would like answered are:

- Can the density of a certain venues serve as an indicator of average real estate price in the area? *(ie. Can a high concentration of "Poke" restaurants indicate the average home price in a given neighborhood?)*

- Are certain types of neighborhoods undervalued compared to similar neighborhoods? *(ie. Do all of the expensive neighborhoods have a high concentration of bars? If so, is there a cheap neighborhood that clusters with them?)*

Defining the objective in business terms:

Our purpose is to find the neighborhood(s) and or retail establishments that have the most potential for future value growth.

How will our solution be used?

When a working method is established, our solution will be used to further explore Seattle, as well as other tech hubs in the United States. Austin, San Francisco, and New York City are other potential targets of the investment firm.

What are the current solutions/workarounds (if any)?

Currently, the only solutions for evaluation and purchasing real estate in Seattle is our knowledge of the local neighborhoods.

How should we frame this problem (supervised/unsupervised, online/offline, etc.)?

We will use both supervised and unsupervised learning since we have labeled data that must be transformed into unlabeled clusters. We are using historical data points, so our model only needs to be offline.

How should performance be measured?

Performance will be measured by finding statistical outliers in our datasets.

Is the performance measure aligned with the business objective?

Yes, Seattle has experienced tremendous growth since the 2010 Census. We need to maximize value opportunities by identifying underdeveloped areas. At the same time, we need to preserve capital and manage risk by avoiding overdeveloped "trendy" neighborhoods.

What would be the minimum performance needed to reach the business objective?

A single neighborhood outlier or under-represented venue would warrant further research for serious investing.

What are comparable problems? Can you reuse experience or tools?

- Does this area have high crime?
- Does the geographic location pose any environmental hazards (flooding, landslide)?
- Does the existing structure have modern earthquake standards?
- Does the homeless population effect future sale value?

Is human expertise available?

Yes, I spent many years living in Seattle, and many hours property hunting on Zillow. I've also spent way too much money dining out, exploring the city, and experiencing the culture as if I were a local hipster.

How would you solve the problem manually?

In order to solve the problem manually, we would have to search many properties that are on the market in each neighborhood as well as previous sale prices in the past decade. We would also have to compile a list of retail establishments for each neighborhood (*ie. There are 5 pizza shops in Ballard but only one in Magnolia*).

List the assumptions you (or others) have made so far.

1) Seattle has been overpriced by high paying, entry level tech jobs

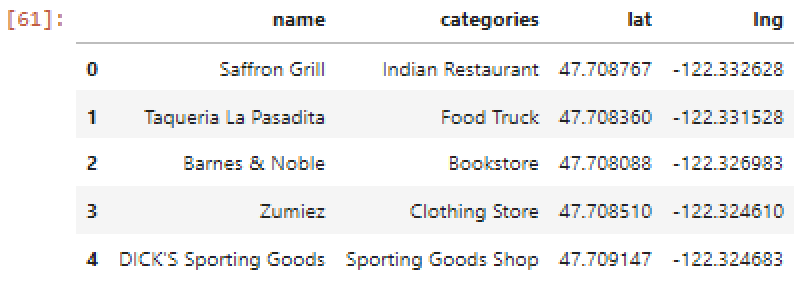
2) Young locals can no longer afford to move out of their parent's house and buy property
3) There are too many nail salons
4) Neighborhoods with Single family homes will have a higher average Zillow Value.
5) Retail establishments in neighborhoods with extremely high Zillow Value will suffer from poor foot traffic

Note: Verification of assumptions will be completed when Phase 2 of this Project is funded.

B. Data Description: Source and Prepare the Data

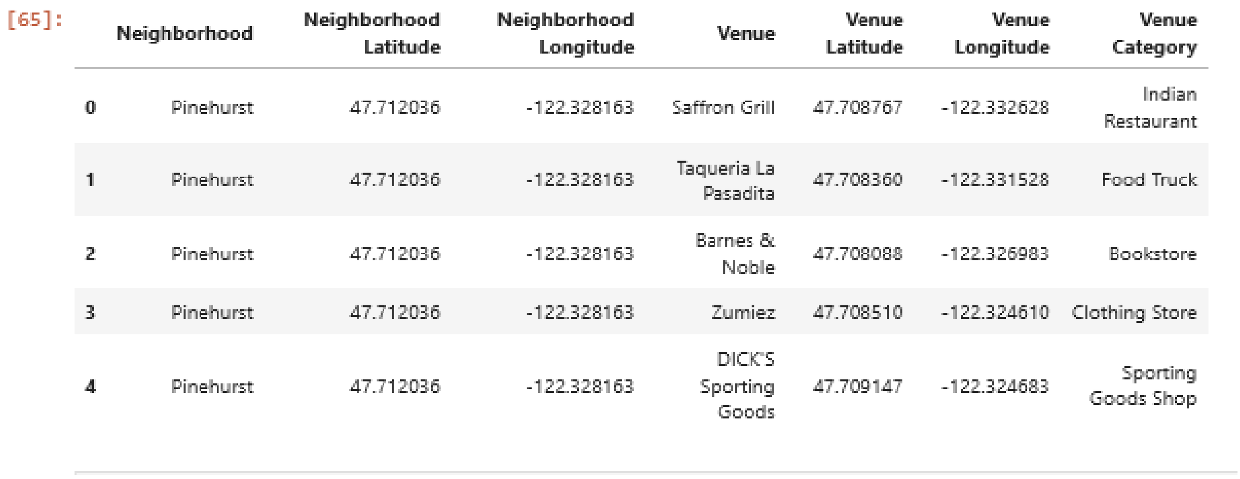
In order to complete the initial analysis, we will use 3 data sets.

1. Foursquare Location Data (meta data via Foursquare API)
   Capture Date: Active
   - To query, map, and correlate venue types, density, location, etc.



[61]:

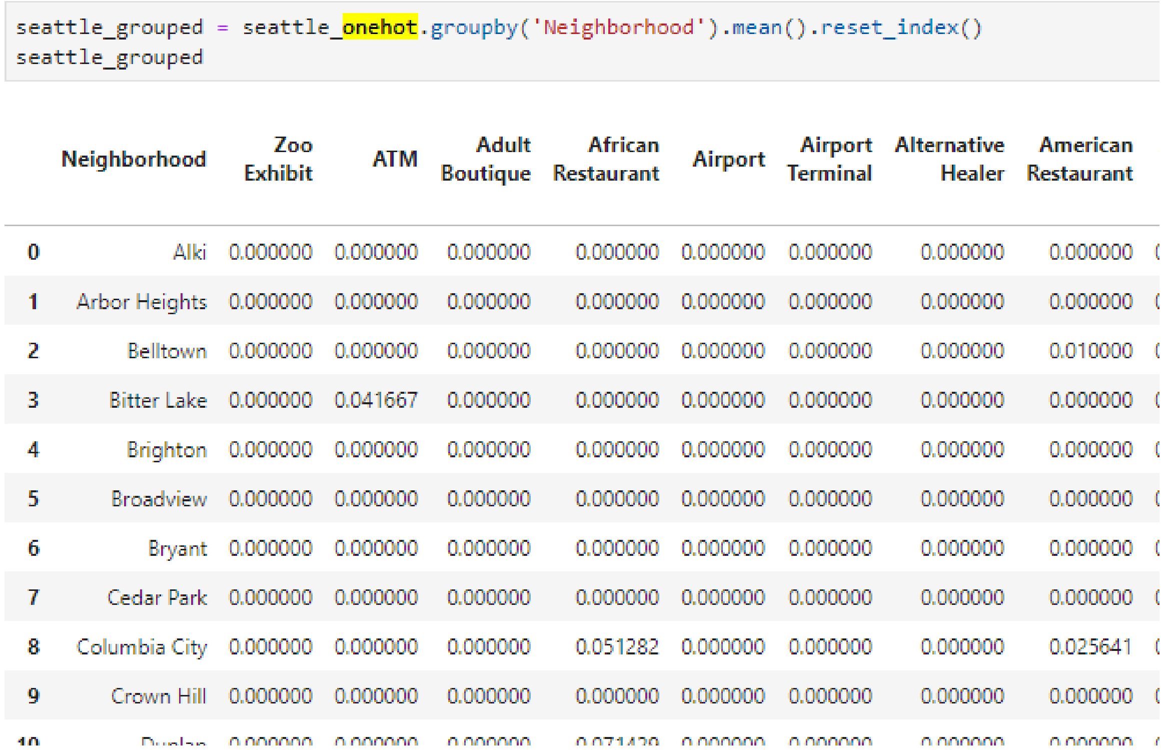
	name	categories	lat	lng
0	Saffron Grill	Indian Restaurant	47.708767	-122.332628
1	Taqueria La Pasadita	Food Truck	47.708360	-122.331528
2	Barnes & Noble	Bookstore	47.708088	-122.326983
3	Zumiez	Clothing Store	47.708510	-122.324610
4	DICK'S Sporting Goods	Sporting Goods Shop	47.709147	-122.324683



[65]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Pinehurst	47.712036	-122.328163	Saffron Grill	47.708767	-122.332628	Indian Restaurant
1	Pinehurst	47.712036	-122.328163	Taqueria La Pasadita	47.708360	-122.331528	Food Truck
2	Pinehurst	47.712036	-122.328163	Barnes & Noble	47.708088	-122.326983	Bookstore
3	Pinehurst	47.712036	-122.328163	Zumiez	47.708510	-122.324610	Clothing Store
4	Pinehurst	47.712036	-122.328163	DICK'S Sporting Goods	47.709147	-122.324683	Sporting Goods Shop

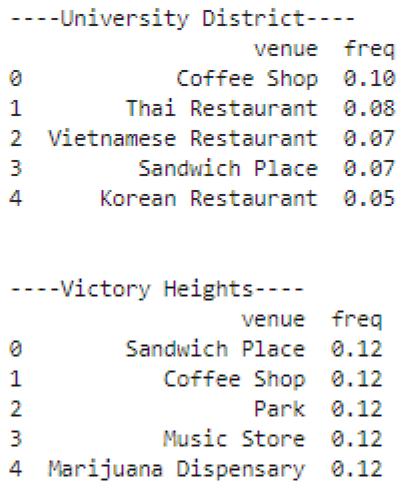
The one hot encoding method was used to weigh the 220 unique venue categories.



```
seattle_grouped = seattle_onehot.groupby('Neighborhood').mean().reset_index()
seattle_grouped
```

	Neighborhood	Zoo Exhibit	ATM	Adult Boutique	African Restaurant	Airport	Airport Terminal	Alternative Healer	American Restaurant
0	Alki	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Arbor Heights	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Belltown	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.010000
3	Bitter Lake	0.000000	0.041667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Brighton	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Broadview	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Bryant	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Cedar Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	Columbia City	0.000000	0.000000	0.000000	0.051282	0.000000	0.000000	0.000000	0.025641
9	Crown Hill	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10	Dunlap	0.000000	0.000000	0.000000	0.071429	0.000000	0.000000	0.000000	0.000000

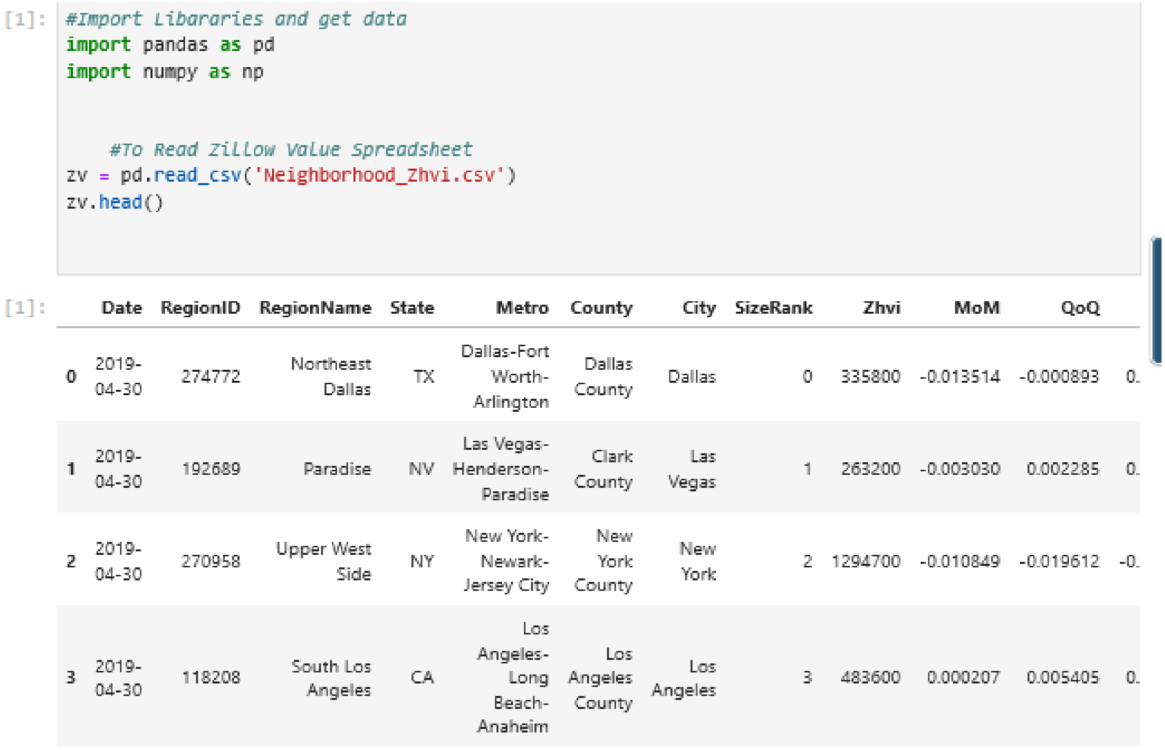
Snapshot of the top 5 most common venues and their frequency



```
----University District----
                   venue  freq
0            Coffee Shop  0.10
1        Thai Restaurant  0.08
2  Vietnamese Restaurant  0.07
3         Sandwich Place  0.07
4      Korean Restaurant  0.05


----Victory Heights----
                  venue  freq
0        Sandwich Place  0.12
1           Coffee Shop  0.12
2                  Park  0.12
3           Music Store  0.12
4  Marijuana Dispensary  0.12
```

2. **Zillow Value Real Estate Data (CSV format from the Zillow Research website)**
   Capture Date: April 2019
   - To examine Zillow Value and 5 Year Growth rate among Seattle neighborhoods

The spreadsheet contained neighborhoods from all over the country

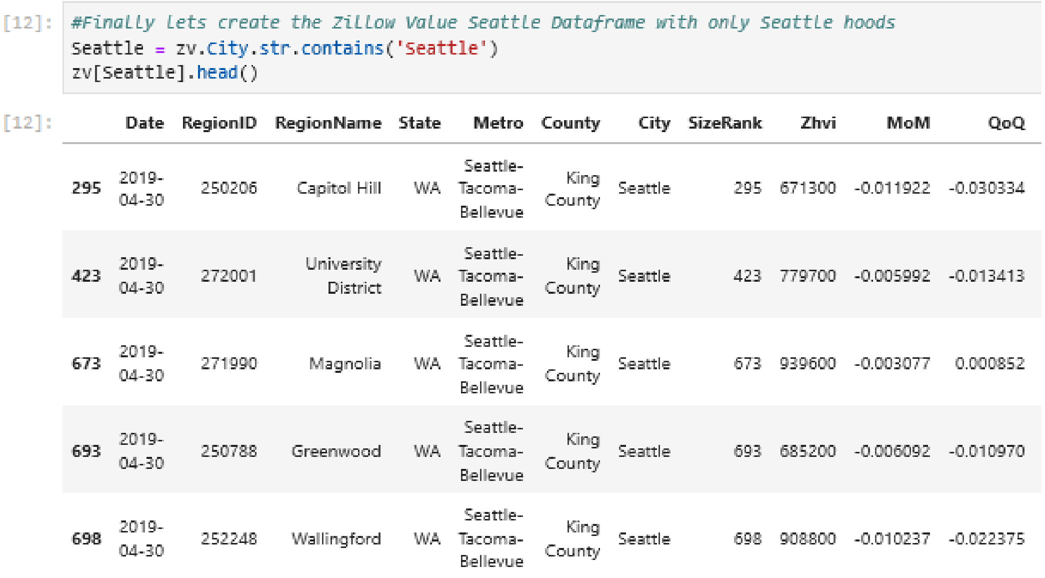


```
[1]: #Import Libararies and get data
     import pandas as pd
     import numpy as np


         #To Read Zillow Value Spreadsheet
     zv = pd.read_csv('Neighborhood_Zhvi.csv')
     zv.head()
```

[1]:

	Date	RegionID	RegionName	State	Metro	County	City	SizeRank	Zhvi	MoM	QoQ	
0	2019-04-30	274772	Northeast Dallas	TX	Dallas-Fort Worth-Arlington	Dallas County	Dallas	0	335800	-0.013514	-0.000893	0.
1	2019-04-30	192689	Paradise	NV	Las Vegas-Henderson-Paradise	Clark County	Las Vegas	1	263200	-0.003030	0.002285	0.
2	2019-04-30	270958	Upper West Side	NY	New York-Newark-Jersey City	New York County	New York	2	1294700	-0.010849	-0.019612	-0.
3	2019-04-30	118208	South Los Angeles	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	Los Angeles	3	483600	0.000207	0.005405	0.

Seattle neighborhoods were separated from the rest of the country, then wrote it to a new data frame in case we want to compare to the national average during further study



```
[12]: #Finally lets create the Zillow Value Seattle Dataframe with only Seattle hoods
      Seattle = zv.City.str.contains('Seattle')
      zv[Seattle].head()
```

[12]:

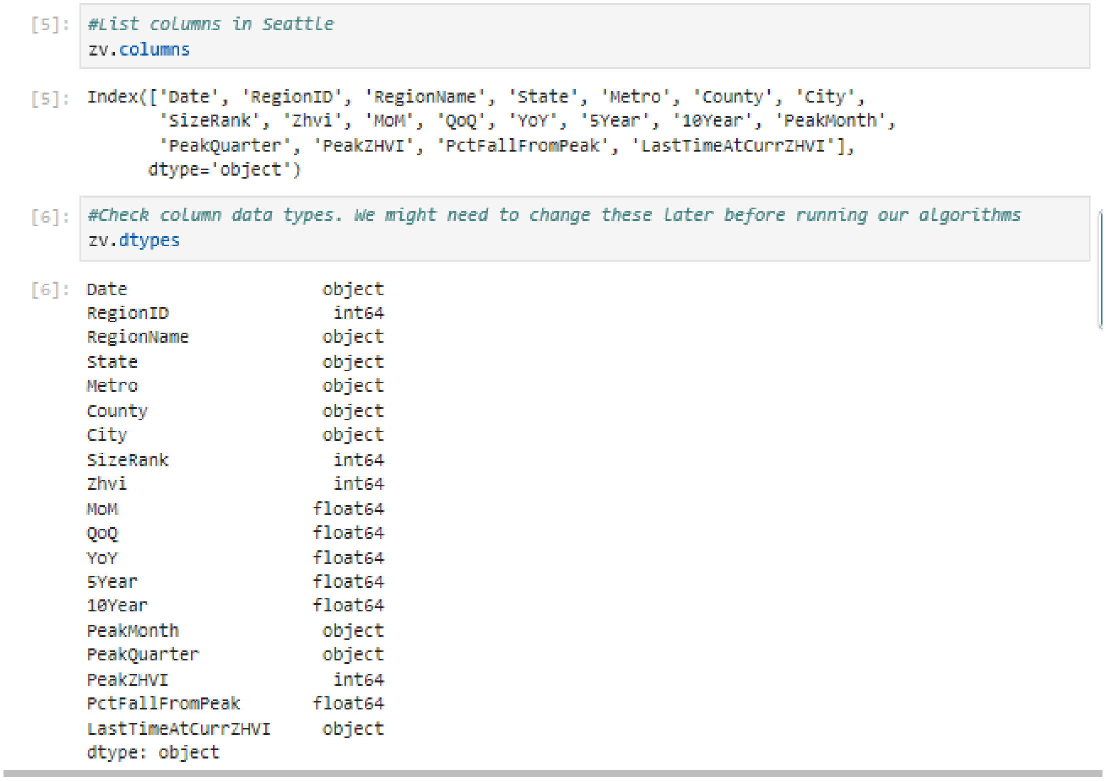
	Date	RegionID	RegionName	State	Metro	County	City	SizeRank	Zhvi	MoM	QoQ
295	2019-04-30	250206	Capitol Hill	WA	Seattle-Tacoma-Bellevue	King County	Seattle	295	671300	-0.011922	-0.030334
423	2019-04-30	272001	University District	WA	Seattle-Tacoma-Bellevue	King County	Seattle	423	779700	-0.005992	-0.013413
673	2019-04-30	271990	Magnolia	WA	Seattle-Tacoma-Bellevue	King County	Seattle	673	939600	-0.003077	0.000852
693	2019-04-30	250788	Greenwood	WA	Seattle-Tacoma-Bellevue	King County	Seattle	693	685200	-0.006092	-0.010970
698	2019-04-30	252248	Wallingford	WA	Seattle-Tacoma-Bellevue	King County	Seattle	698	908800	-0.010237	-0.022375

Problems with the Data: The 80/20 Myth is Real

3. Getting Longitude and Latitude Data for Mapping was a Major Problem

Wrangling “accurate” Latitude and Longitude data ended up taking up the majority of the project time. I’d like to outline the issues below because many people believe that Data Scientists spend there day building and fine-tuning models. However, it is often *stressed* that Data Scientists spend the majority of there time cleaning the data.

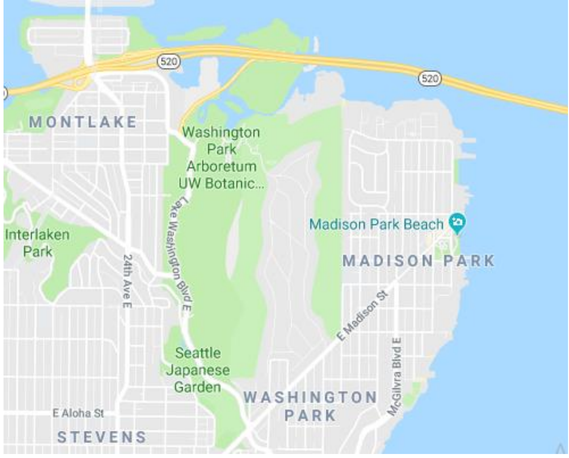
Problem #1: Zillow no longer provided downloadable neighborhood coordinates or shapefiles without request. No latitude and longitude boundaries in the CSV.

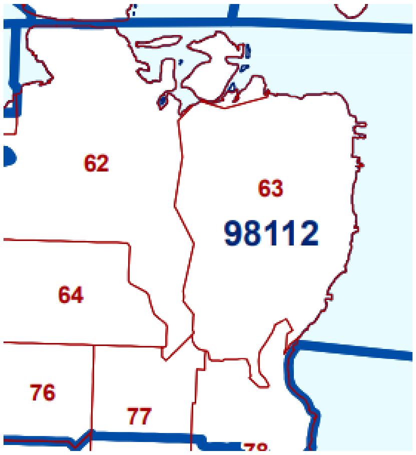


```
[5]: #List columns in Seattle
     zv.columns

[5]: Index(['Date', 'RegionID', 'RegionName', 'State', 'Metro', 'County', 'City',
            'SizeRank', 'Zhvi', 'MoM', 'QoQ', 'YoY', '5Year', '10Year', 'PeakMonth',
            'PeakQuarter', 'PeakZHVI', 'PctFallFromPeak', 'LastTimeAtCurrZHVI'],
           dtype='object')

[6]: #Check column data types. We might need to change these later before running our algorithms
     zv.dtypes

[6]: Date                   object
     RegionID                int64
     RegionName             object
     State                  object
     Metro                  object
     County                 object
     City                   object
     SizeRank                int64
     Zhvi                    int64
     MoM                   float64
     QoQ                   float64
     YoY                   float64
     5Year                 float64
     10Year                float64
     PeakMonth              object
     PeakQuarter            object
     PeakZHVI                int64
     PctFallFromPeak       float64
     LastTimeAtCurrZHVI     object
     dtype: object
```

Problem #2: Zip codes do not accurately reflect neighborhood boundaries

A Google Maps query outlines the Washington Park Arboretum separating the Montlake and Madison Park neighborhoods. These distinct areas are among two of the most expensive places to buy property in the city.



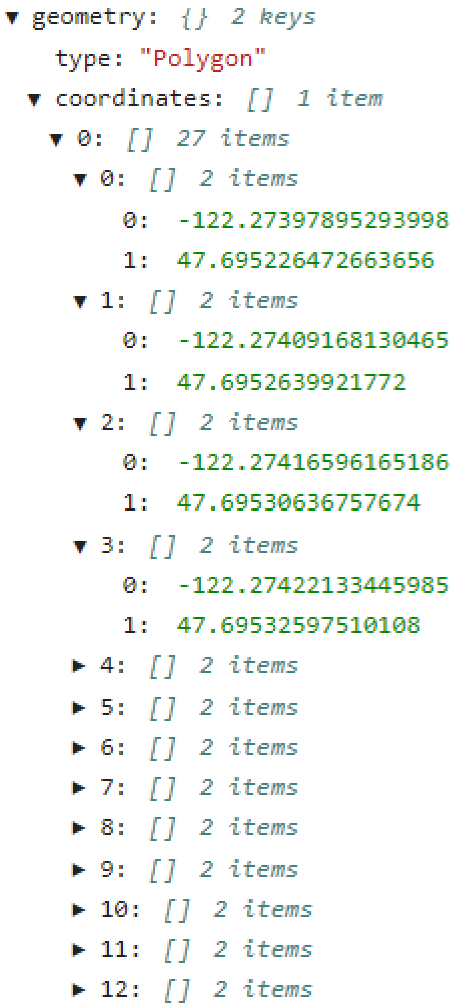


The map on the right from the 2010 U.S. Census shows Montlake and Madison Park sharing the same zip code.

Problem #3 Polygon and Multi-Polygon Datasets are Difficult to Put into a Dataframe

I scoured the internet and located a study titled **U.S. Neighborhoods Greenness Measures and Social Variables** from the **NYU Spatial Data Repository**. I was able to parce the GeoJson from the file. However, I had trouble mapping the Multi-Polygon neighborhood coordinates type using Folium, GeoPandas, and other Python Geospatial libraries.

Polygon and ***Multi-Polygon*** geometry types use nested coordinates to draw neighborhoods. Populating a dataframe isn't as simple as ***Point*** type geometry. I noticed these are more popular with GIS and JavaScript. Folium maps did not render in Jupyter notebook because the data is too complex.



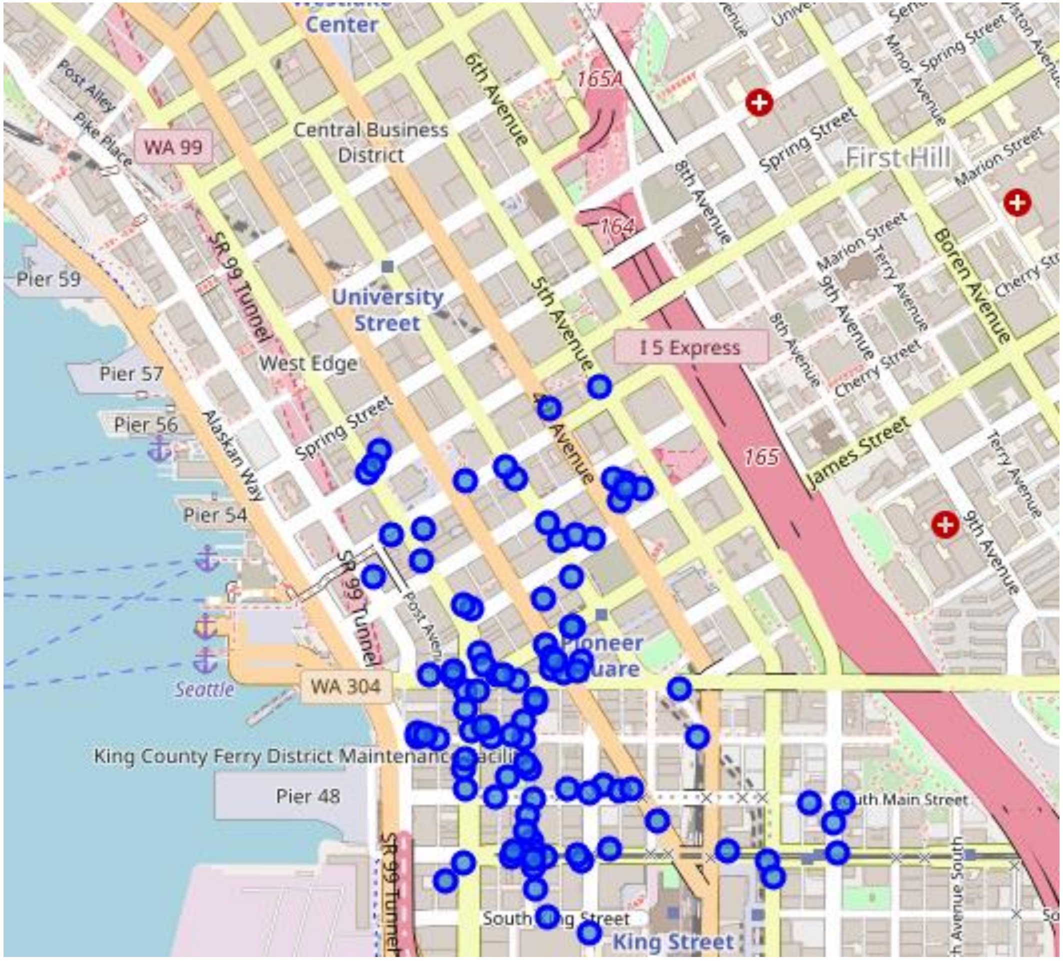
```
▼ geometry: {} 2 keys
    type: "Polygon"
  ▼ coordinates: [] 1 item
    ▼ 0: [] 27 items
      ▼ 0: [] 2 items
          0: -122.27397895293998
          1: 47.695226472663656
      ▼ 1: [] 2 items
          0: -122.27409168130465
          1: 47.6952639921772
      ▼ 2: [] 2 items
          0: -122.27416596165186
          1: 47.69530636757674
      ▼ 3: [] 2 items
          0: -122.27422133445985
          1: 47.69532597510108
      ▶ 4: [] 2 items
      ▶ 5: [] 2 items
      ▶ 6: [] 2 items
      ▶ 7: [] 2 items
      ▶ 8: [] 2 items
      ▶ 9: [] 2 items
      ▶ 10: [] 2 items
      ▶ 11: [] 2 items
      ▶ 12: [] 2 items
```

Problem #4

As an alternative, the BBox Longitude and Latitude data was used as Point geometry. However, this resulted in the bubbled neighborhoods being slight shifted.

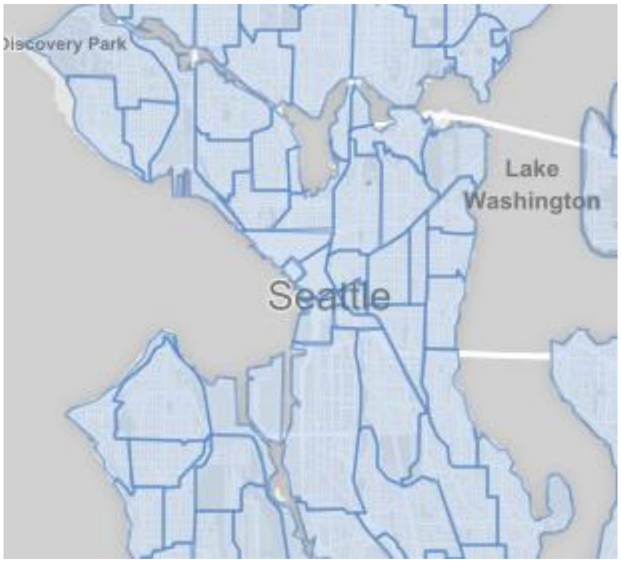
When combining the GeoJson BBox coordinates with the Foursquare data, Venues did not cluster correctly. This was evident when one “land-locked” neighborhood showed a beach as its most common venue.

In the example below, mapped First Hill Venues were overlaying on the Pioneer Square neighborhood.



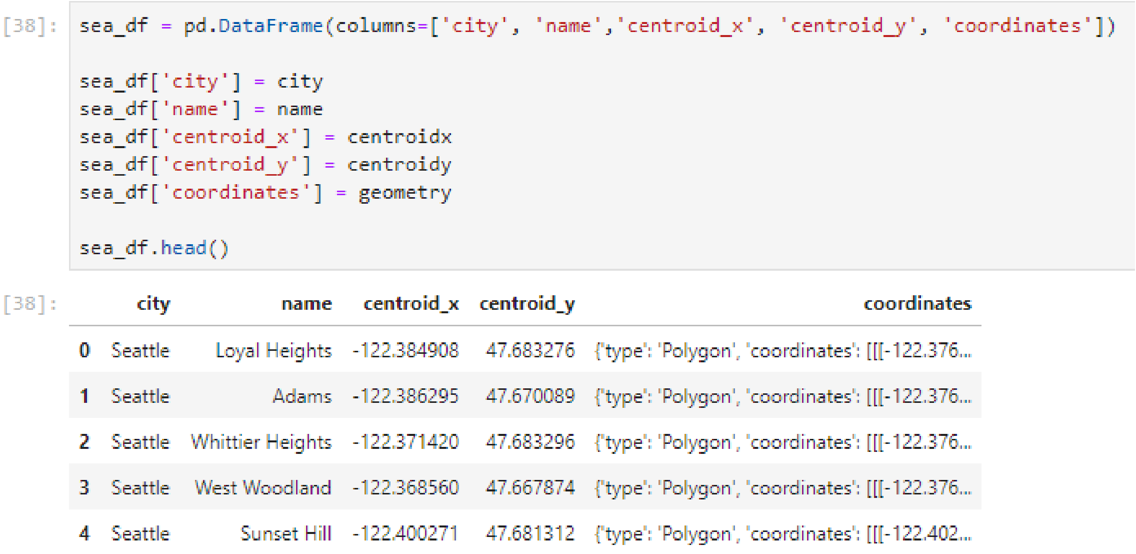
Solution

Finally, a GeoJson file was located on Github that had accurate Polygon coordinates. SeattleIO has different boundary maps located on their page provided from various government sources.



I attempted to use the Shapely library and import its GeometryCollection and Polygon features to draw out the neighborhoods over the Folium map. Drawing individual neighborhoods did work, but the entire map would not render in Jupyter Notebook.

Alternatively, I pulled the Centroid coordinates from the GeoJson and created a dataframe similar to what I did with the BBox coordinates in the other file. This time the neighborhoods did bubble in the accurate locations.



```
[38]: sea_df = pd.DataFrame(columns=['city', 'name','centroid_x', 'centroid_y', 'coordinates'])

      sea_df['city'] = city
      sea_df['name'] = name
      sea_df['centroid_x'] = centroidx
      sea_df['centroid_y'] = centroidy
      sea_df['coordinates'] = geometry

      sea_df.head()
```

[38]:

	city	name	centroid_x	centroid_y	coordinates
0	Seattle	Loyal Heights	-122.384908	47.683276	{'type': 'Polygon', 'coordinates': [[[-122.376...
1	Seattle	Adams	-122.386295	47.670089	{'type': 'Polygon', 'coordinates': [[[-122.376...
2	Seattle	Whittier Heights	-122.371420	47.683296	{'type': 'Polygon', 'coordinates': [[[-122.376...
3	Seattle	West Woodland	-122.368560	47.667874	{'type': 'Polygon', 'coordinates': [[[-122.376...
4	Seattle	Sunset Hill	-122.400271	47.681312	{'type': 'Polygon', 'coordinates': [[[-122.402...

Section 2: Exploration, Methodology, and Analysis

B. Finding Value in the Zillow Data

Valuing Seattle Neighborhoods

Initially, I intended to look at historical sales for each neighborhood and compare them to the density of venues. However after much research, I realized that the Zillow Value (ZHVI) could sufficiently replace historical sales, or any value I could give a Seattle neighborhood.

Let me start off by talking about Zillow, and explaining Zillow Value. If you have not heard of Zillow at this point, you obviously haven't looked for a place to live in the past decade.

Zillow is an internet based Real Estate company that was founded in 2006 by former Microsoft executives. Naturally, it has become the pinnacle of Real Estate and Technology. If you are looking for an apartment to rent, or a home to purchase in the United States, Zillow is often the best place to begin research. Consumers are able to conduct and filter a search across many different categories free of charge.

What is Zillow Value (ZHVI)?

Zillow Value is essentially a sophisticated algorithm that Zillow has created to value Real Estate across the country. It is a complex machine learning model that uses techniques such as:

- Market Clustering (segmentation)
- Historical Sales
- Seasonality
- Moving Average Filters

Detailed methodology can be found here

A PDF summary can be found here: Zillow Home Value Index

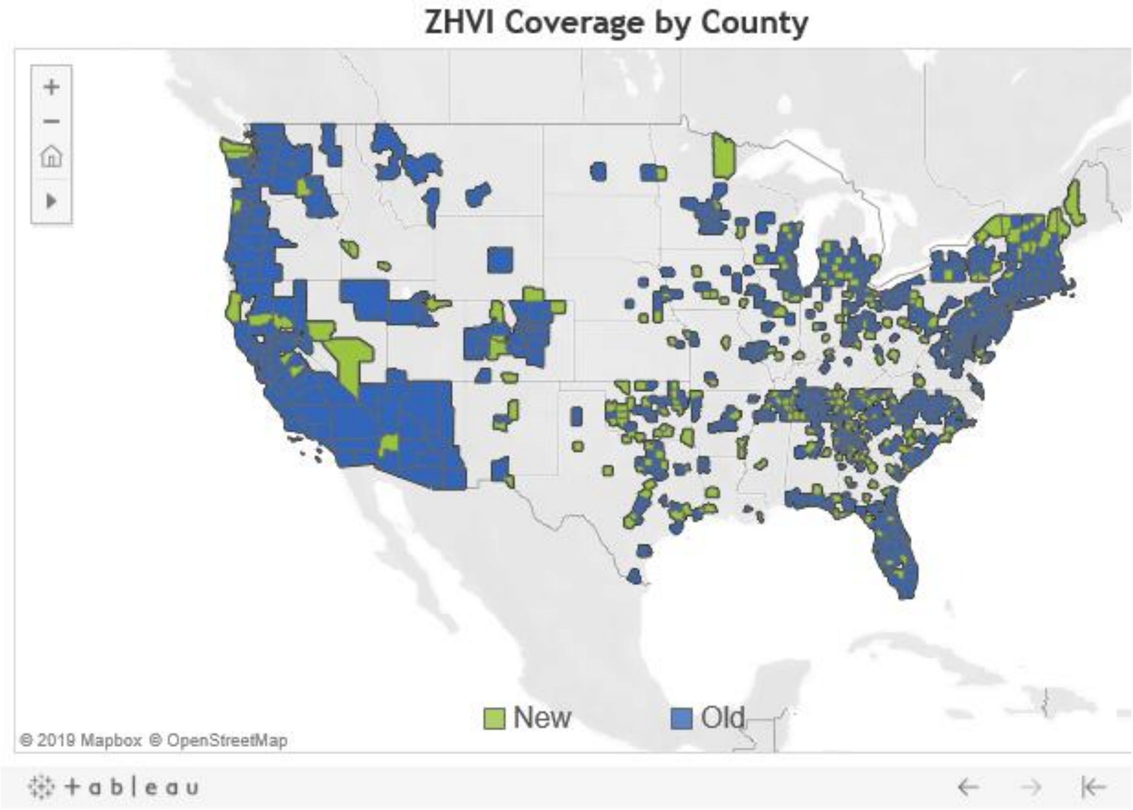
A Note of Caution

Zillow Value or "Zestimate" has been often criticized for its inaccuracies among homeowners and real estate investors. However, as with Machine Learning algorithms, the more data it receives the better it performs. A major overhaul of the system was performed in 2013, and research and technology has only improved since then.

I believe the major flaw in ZHVI is the inaccuracy it presents in rural and suburban markets. As you can imagine, the less sales in a local market, the worse the prediction of value would be. In hot, urban markets, ZHVI excels.

Seeing that Zillow is a Seattle based company, we must trust in the fact that they have studied their local market well. I personally cannot value neighborhoods better than a team of experienced data scientists at the world's leading Real Estate tech company.

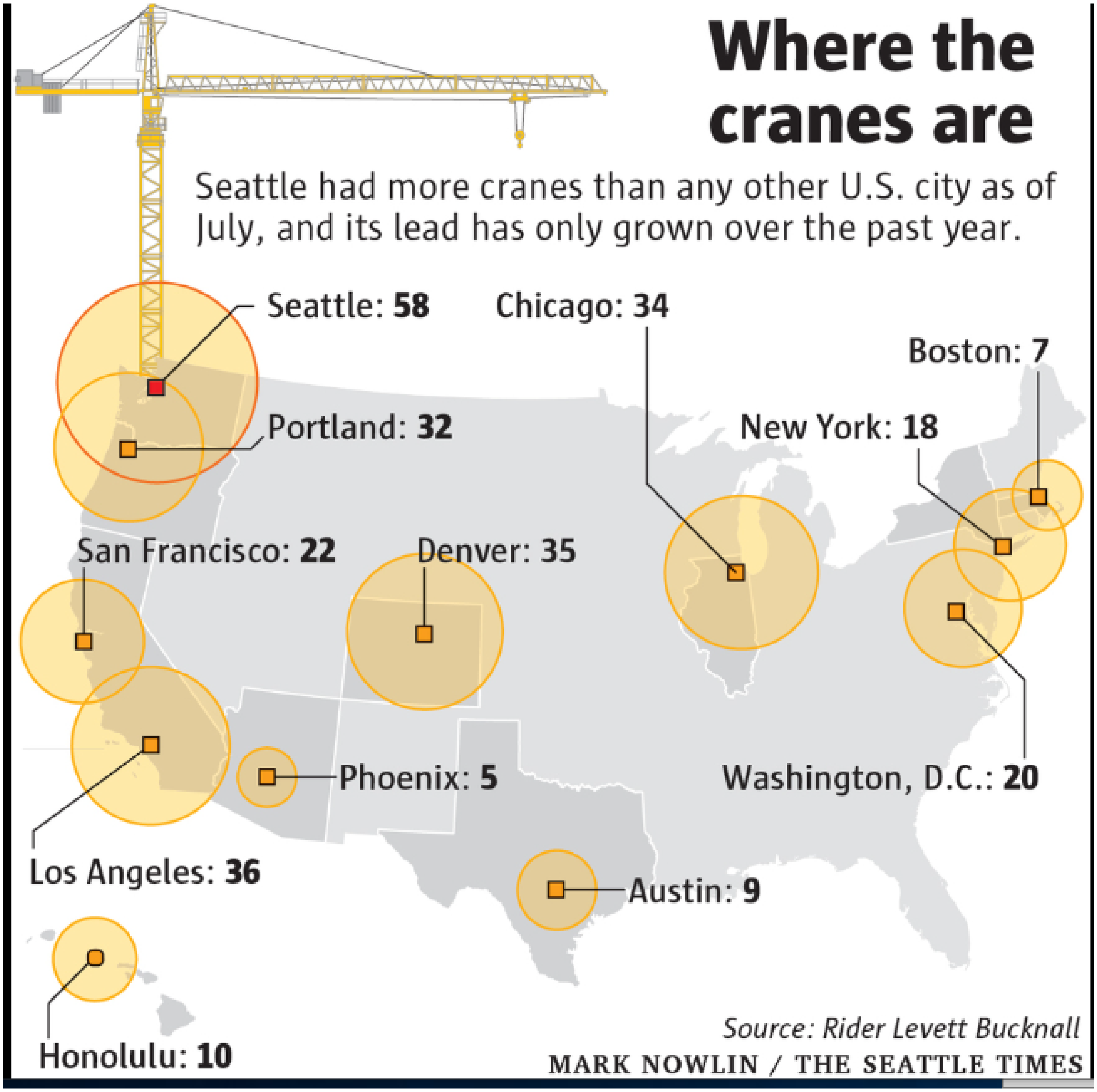
As you can see in the map below ZHVI does not cover every county in the United States. It would appear their new algorithm has been adopted to serve more rural counties better.



5 Year Growth of ZHVI

The other data point I used from Zillow was the 5 Year Growth of ZHVI (labeled: 5Year). I figured this would accurately reflect Value Growth in the hot tech market. With the explosion of Amazon, and the rapid construction around the city, the past 5 years seemed the most appropriate to examine.

A 2017 article by the Seattle Times states that Seattle has the most cranes in the country for the second year in row. It is estimated that Seattle has seen a 22% increase in population since the 2010 census. It is currently estimated as the 18th most populated city in the country, and 15th by metro area population density. We won't know for sure until the 2020 census, but it is likely Seattle has been among the fastest growing cities in the country.



Where the cranes are

Seattle had more cranes than any other U.S. city as of July, and its lead has only grown over the past year.

Source: Rider Levett Bucknall

MARK NOWLIN / THE SEATTLE TIMES

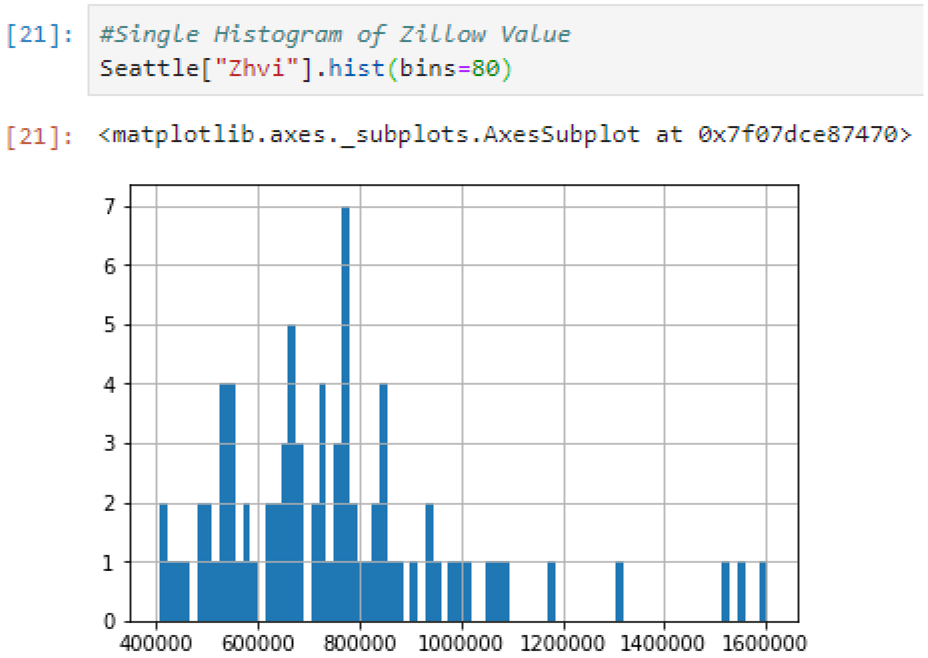
Why Not 10 Year Growth?

The reason I chose to not explore the 10 Year Growth of ZHVI are as follows:

- The 10Year column in the Zillow dataset did not seem accurate to me
- The market crash and economic recession of 2009 would poorly skew the data
- The impact of high paying tech jobs from Amazon would not be fully reflected in the local economy

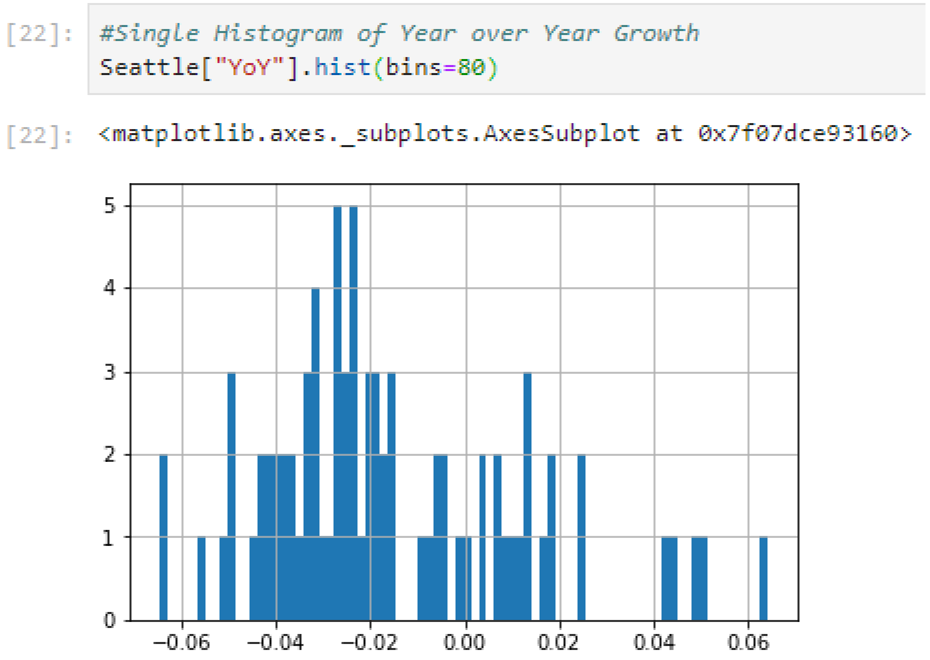
Zillow Neighborhood Analysis

The histogram below demonstrates the distribution average Zillow Value of the 80 Seattle neighborhoods in the data set.



```
[21]: #Single Histogram of Zillow Value
      Seattle["Zhvi"].hist(bins=80)

[21]: <matplotlib.axes._subplots.AxesSubplot at 0x7f07dce87470>
```

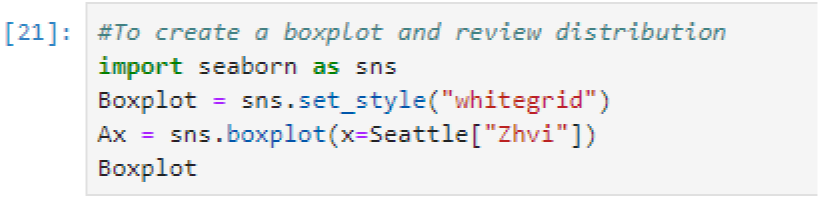
The histogram below shows the distribution of the Year over Year growth of Zillow Value. Besides being nearly identical, the distribution shows how Seattle Real Estate prices have stabilized in the past year. Becoming a buyer at this point has downside risk.



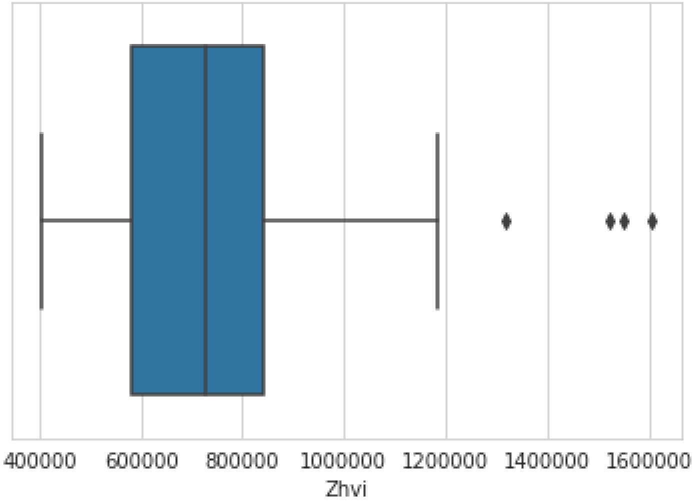
```
[22]: #Single Histogram of Year over Year Growth
      Seattle["YoY"].hist(bins=80)

[22]: <matplotlib.axes._subplots.AxesSubplot at 0x7f07dce93160>
```

Boxplot

- I used the Seaborn library to create a boxplot to further examine the distribution of neighborhood ZHVI.
- Note the 4 outliers that could potentially skew our data.
- The majority of neighborhoods are concentrated between 600K and 800k.



```
[21]: #To create a boxplot and review distribution
      import seaborn as sns
      Boxplot = sns.set_style("whitegrid")
      Ax = sns.boxplot(x=Seattle["Zhvi"])
      Boxplot
```



Correlations

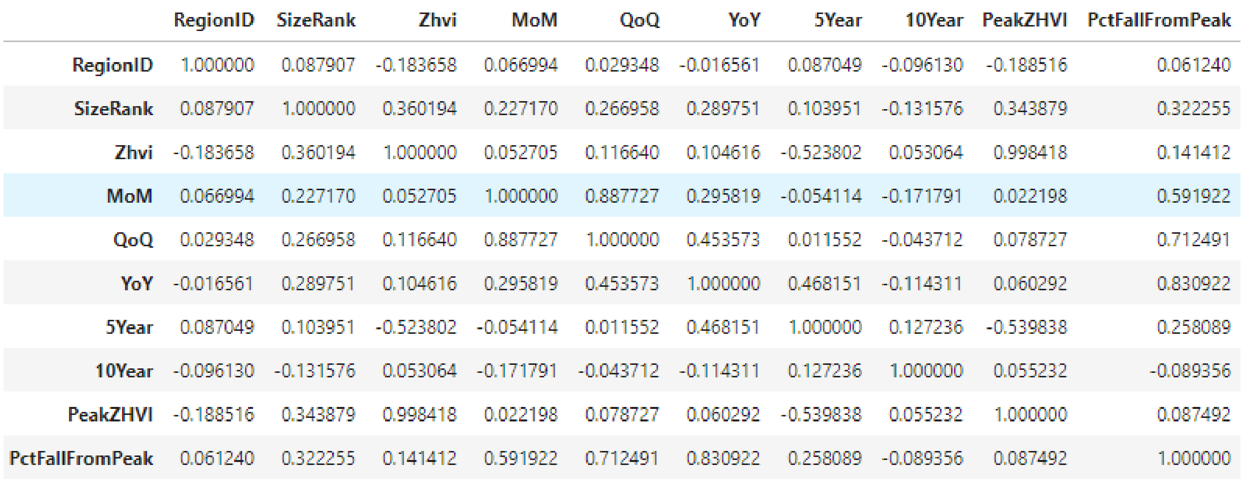
Next, a correlation table was created to examine which variables could be related. Although we are working with limited data points, this table could warrant further investigation.

Obvious correlations between variables such as QoQ (quarter over quarter) and YoY (year over year) were ignored. The main takeaway was **a moderate negative correlation (-0.52)** between **ZHVI** and **5Year**(growth)**.**

The data shows cheaper neighborhoods (if you call 400k cheap) have seen the highest increase in Zillow Value over the past 5 years. Thus indicating high competition among tech workers and other professionals competing for a home. This falls in line with the notion that "6 figure" Amazon workers have been making it difficult for locals to settle down in the area.

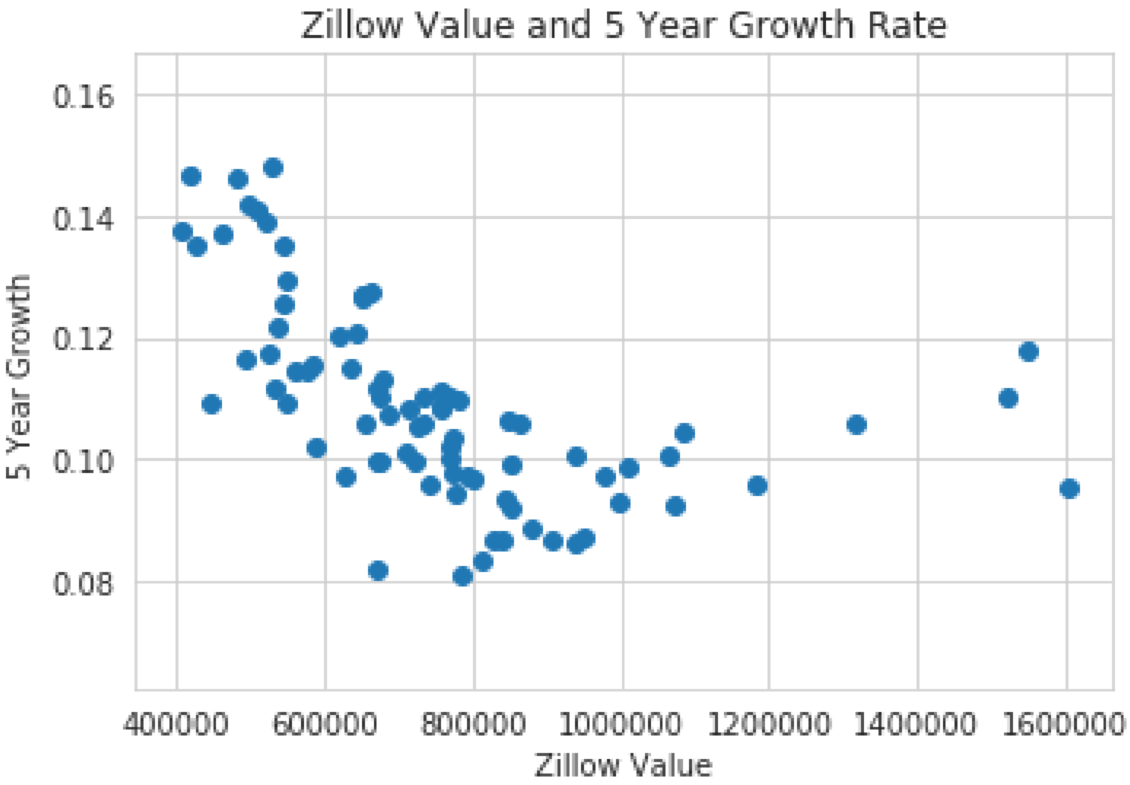
However, the trend breaks in neighborhoods with the average Zillow Value over $1.2 million. This could indicate a prestige factor among the richest neighborhoods. It also might indicate "tech executives" have swelled the demand for high end property.

Zillow Data Correlation Table



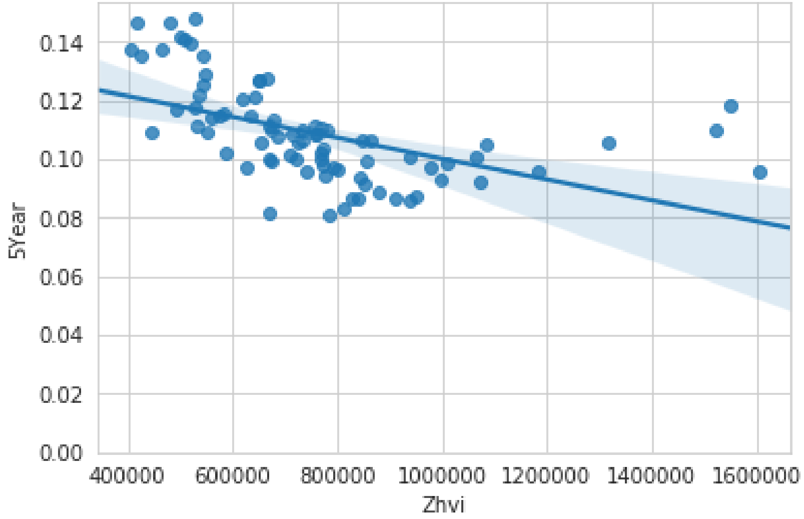
	RegionID	SizeRank	Zhvi	MoM	QoQ	YoY	5Year	10Year	PeakZHVI	PctFallFromPeak
RegionID	1.000000	0.087907	-0.183658	0.066994	0.029348	-0.016561	0.087049	-0.096130	-0.188516	0.061240
SizeRank	0.087907	1.000000	0.360194	0.227170	0.266958	0.289751	0.103951	-0.131576	0.343879	0.322255
Zhvi	-0.183658	0.360194	1.000000	0.052705	0.116640	0.104616	-0.523802	0.053064	0.998418	0.141412
MoM	0.066994	0.227170	0.052705	1.000000	0.887727	0.295819	-0.054114	-0.171791	0.022198	0.591922
QoQ	0.029348	0.266958	0.116640	0.887727	1.000000	0.453573	0.011552	-0.043712	0.078727	0.712491
YoY	-0.016561	0.289751	0.104616	0.295819	0.453573	1.000000	0.468151	-0.114311	0.060292	0.830922
5Year	0.087049	0.103951	-0.523802	-0.054114	0.011552	0.468151	1.000000	0.127236	-0.539838	0.258089
10Year	-0.096130	-0.131576	0.053064	-0.171791	-0.043712	-0.114311	0.127236	1.000000	0.055232	-0.089356
PeakZHVI	-0.188516	0.343879	0.998418	0.022198	0.078727	0.060292	-0.539838	0.055232	1.000000	0.087492
PctFallFromPeak	0.061240	0.322255	0.141412	0.591922	0.712491	0.830922	0.258089	-0.089356	0.087492	1.000000

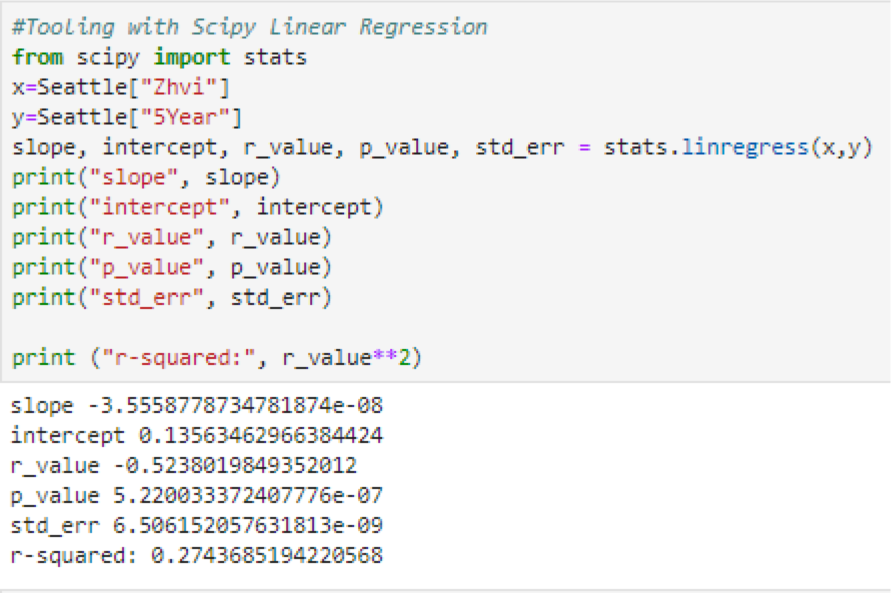
A scatterplot further demonstrated the negative correlation between ZHVI (x-axis) and 5Year(y-axis).



Running a Linear Regression Model

```
[23]:  (0, 0.15325769654018181)
```



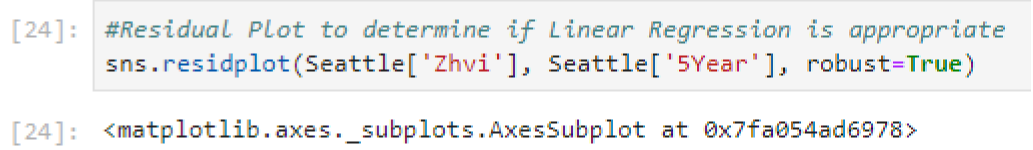


```
#Tooling with Scipy Linear Regression
from scipy import stats
x=Seattle["Zhvi"]
y=Seattle["5Year"]
slope, intercept, r_value, p_value, std_err = stats.linregress(x,y)
print("slope", slope)
print("intercept", intercept)
print("r_value", r_value)
print("p_value", p_value)
print("std_err", std_err)

print ("r-squared:", r_value**2)
```

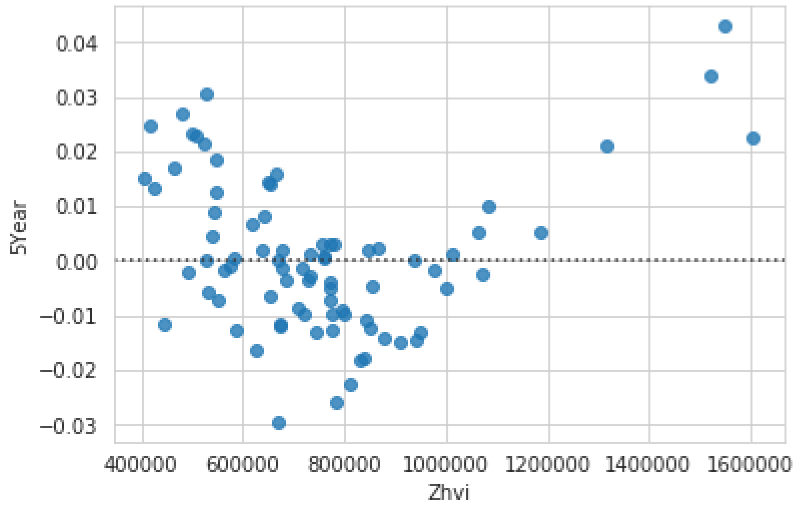
```
slope -3.5558778734781874e-08
intercept 0.13563462966384424
r_value -0.5238019849352012
p_value 5.220033372407776e-07
std_err 6.506152057631813e-09
r-squared: 0.2743685194220568
```

Note the residual plot below is "u shaped", indicating, linear regression may not be the most appropriate model. However, we don't have a ton of data points, and we may end up removing those expensive outlier neighborhoods later. We’ll keep things simple at this point and skip polynomial regression.

Running a Residual Plot

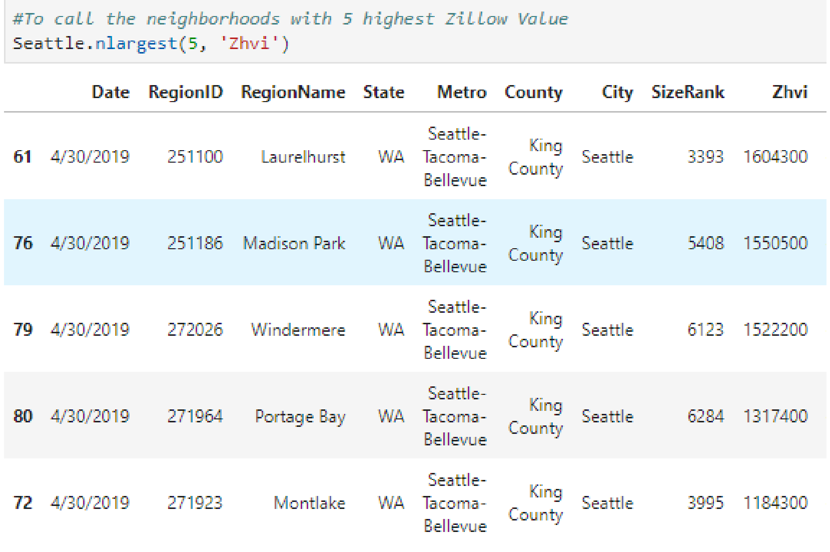


```
[24]: #Residual Plot to determine if Linear Regression is appropriate
      sns.residplot(Seattle['Zhvi'], Seattle['5Year'], robust=True)

[24]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa054ad6978>
```



Examining Outliers

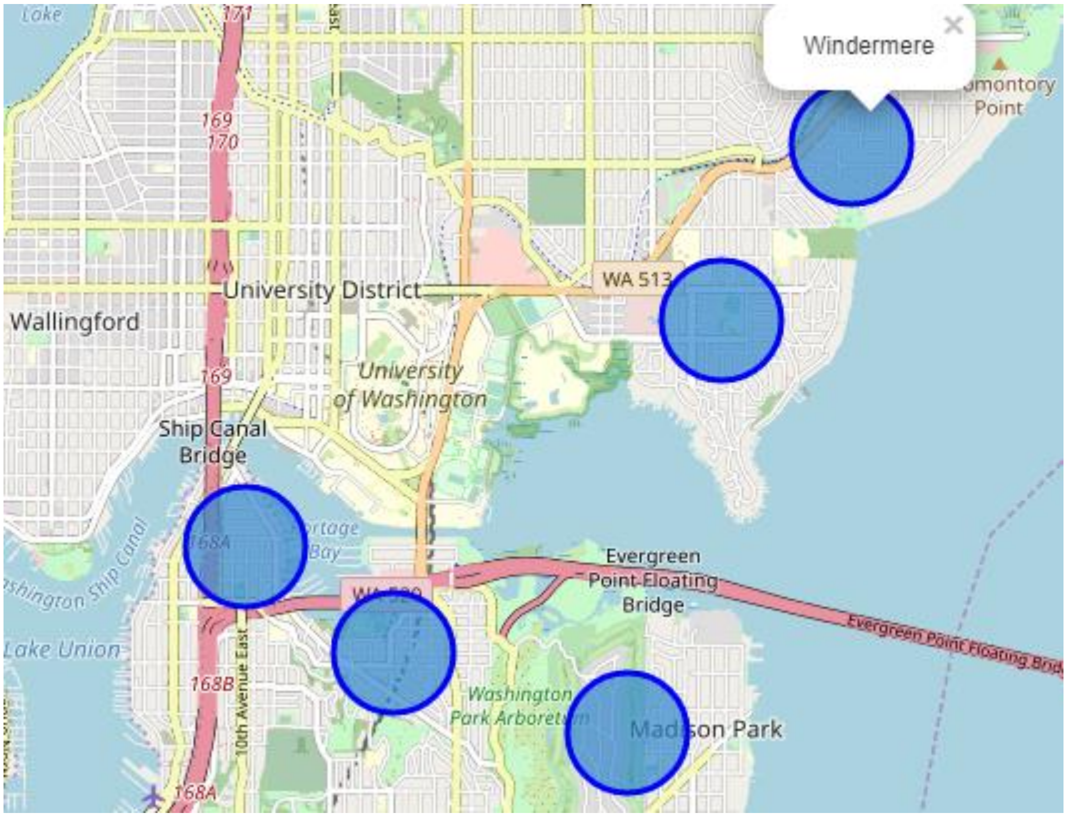
Taking a closer look at the expensive outliers on the scatterplots, we can make a few noteworthy observations. The 5 highest neighborhoods are northeast of downtown and clustered around the University of Washington. They lack an "urban" vibe and contain a many single-family homes.



```
#To call the neighborhoods with 5 highest Zillow Value
Seattle.nlargest(5, 'Zhvi')
```

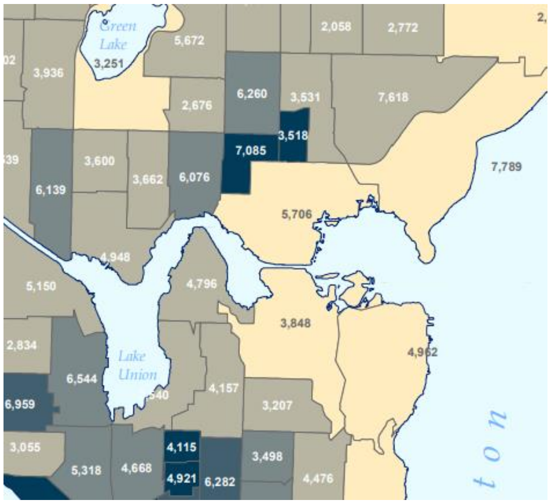
	Date	RegionID	RegionName	State	Metro	County	City	SizeRank	Zhvi
61	4/30/2019	251100	Laurelhurst	WA	Seattle-Tacoma-Bellevue	King County	Seattle	3393	1604300
76	4/30/2019	251186	Madison Park	WA	Seattle-Tacoma-Bellevue	King County	Seattle	5408	1550500
79	4/30/2019	272026	Windermere	WA	Seattle-Tacoma-Bellevue	King County	Seattle	6123	1522200
80	4/30/2019	271964	Portage Bay	WA	Seattle-Tacoma-Bellevue	King County	Seattle	6284	1317400
72	4/30/2019	271923	Montlake	WA	Seattle-Tacoma-Bellevue	King County	Seattle	3995	1184300

Neighborhoods with the Highest Average Zillow Value in Seattle

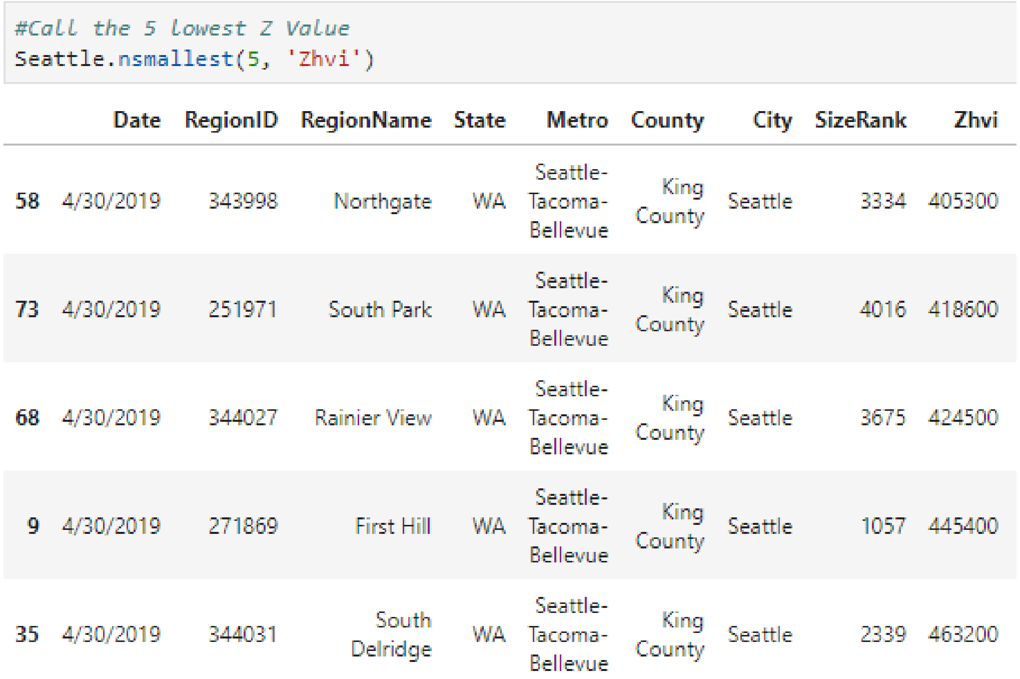


Note the 5 most valuable neighborhoods in Seattle are clustered around the University of Washington. They all include waterfront property and are northeast of Downtown. However, before jumping to conclusions, a note of caution must be exercised. According to Seattle.gov and the 2010 Census, this area is one of the least densely populated areas of the city.

<u>2010 Census Population Density via Seattle.gov</u>



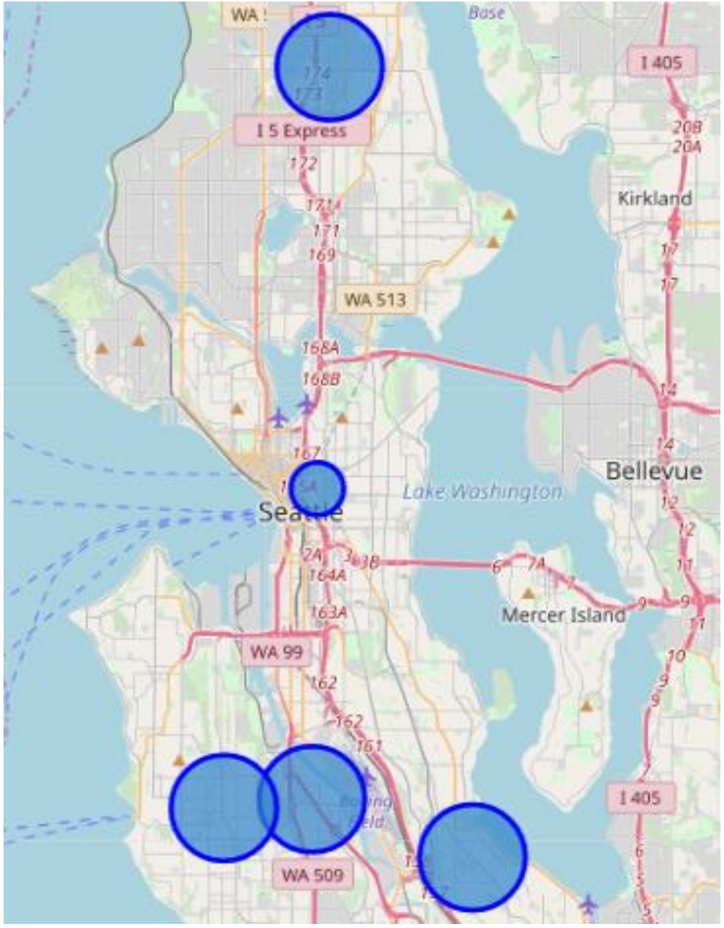
5 Cheapest Neighborhoods (if you call $400k cheap)



```
#Call the 5 lowest Z Value
Seattle.nsmallest(5, 'Zhvi')
```

	Date	RegionID	RegionName	State	Metro	County	City	SizeRank	Zhvi
58	4/30/2019	343998	Northgate	WA	Seattle-Tacoma-Bellevue	King County	Seattle	3334	405300
73	4/30/2019	251971	South Park	WA	Seattle-Tacoma-Bellevue	King County	Seattle	4016	418600
68	4/30/2019	344027	Rainier View	WA	Seattle-Tacoma-Bellevue	King County	Seattle	3675	424500
9	4/30/2019	271869	First Hill	WA	Seattle-Tacoma-Bellevue	King County	Seattle	1057	445400
35	4/30/2019	344031	South Delridge	WA	Seattle-Tacoma-Bellevue	King County	Seattle	2339	463200

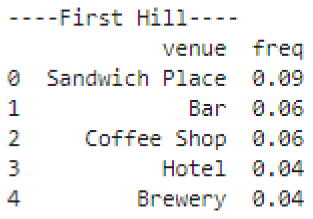
Among the 5 lowest Z Value's, 4 out of 5 are far out from Downtown, and even lack an "urban" vibe.



Hidden Value in First Hill

Among the 5 lowest Z Value's, **First Hill** is the only one with an "urban vibe", and the only one located near **Downtown** (5-10 min walk). It has excellent transportation due to the intersection of a major bus route, street car, and a 5 minute walk to light rail.

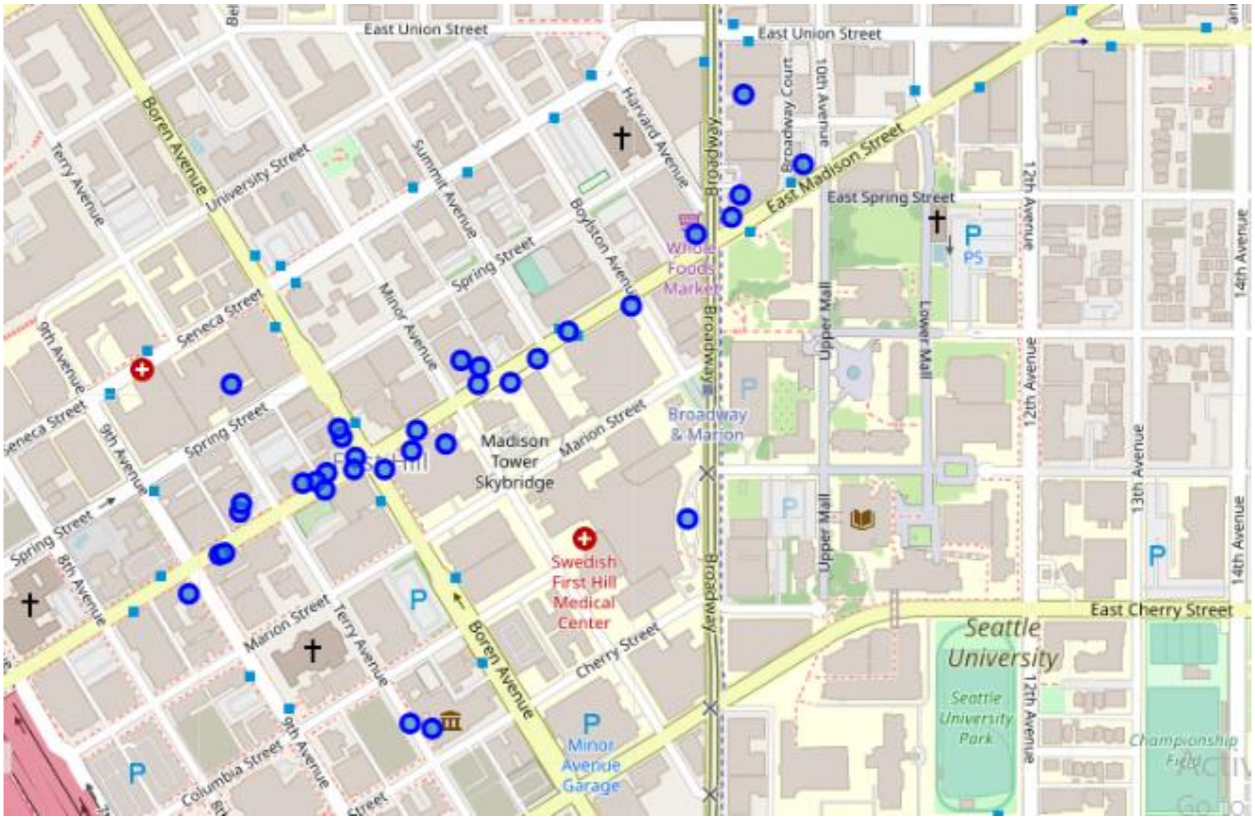
In addition, First Hill borders **Capitol Hill**, Seattle's trendiest neighborhood and liveliest nightlife. It is located near two higher education institutions (Seattle University and Seattle Central Community College) which will help to stabilize the real estate market in case of an economic downturn.



```
----First Hill----
           venue  freq
0  Sandwich Place  0.09
1            Bar  0.06
2     Coffee Shop  0.06
3           Hotel  0.04
4         Brewery  0.04
```

First Hill is often referred to as "Pill Hill" among locals due to the high concentration of hospitals, which again, helps to hedge against an economic downturn. It also might contribute to the density of lunch spots.

It also recently received infamous, "Whole Foods indicator", with one being constructed in the center of the neighborhood *(see below on the intersection of East Madison and Broadway).*



Recommendations:

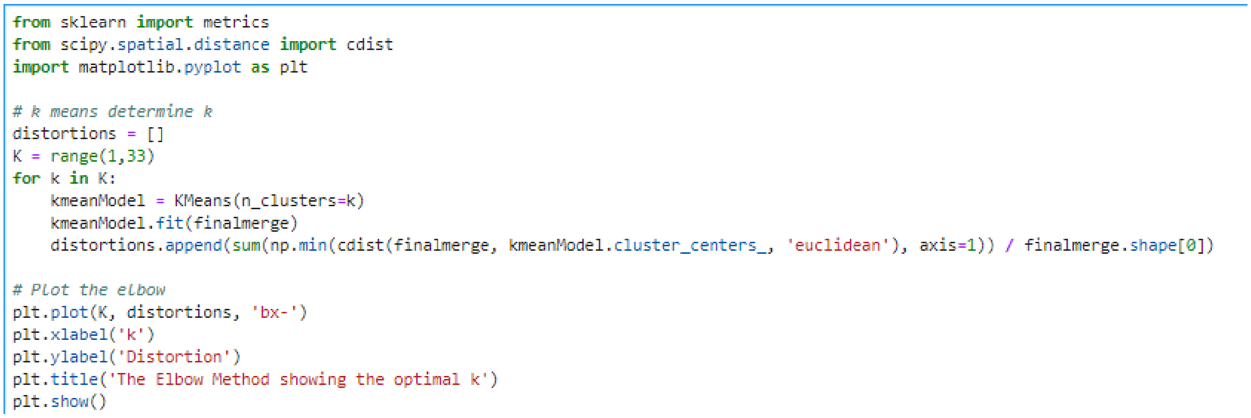
One caveat about First Hill, is it lacks single family homes, and many of the high-rise apartments are of older construction. Thus, leading to a lower ZHVI relative to the core downtown areas.

However, the right condo or retail venue could turn out to be a good investment. First Hill warrants further investigation as a potential area in the crowded Seattle market. I would recommend purchasing the FourSquare premium data to study foot traffic and consumer trends in the neighborhood.

C. Clustering the Neighborhoods by Venue Distribution

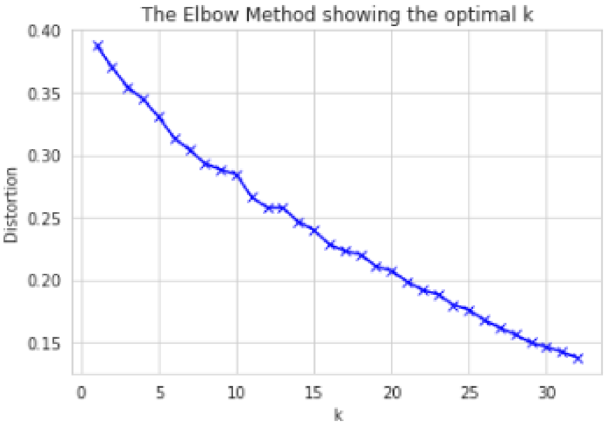
The K-Means Clustering algorithm from Scikit-Learn was used to group together similar Seattle neighborhoods. With a total of 66 different neighborhoods, it was difficult to discern the optimal number of “K” clusters. After much trial and error in plugging in various “K” values, it was decided to use a more scientific approach.

The Elbow Method plots the numbers of clusters against the distortion in the data. When marginal gain drops off, an “elbow” joint appears in the plot, thus determining the optimal “K”.

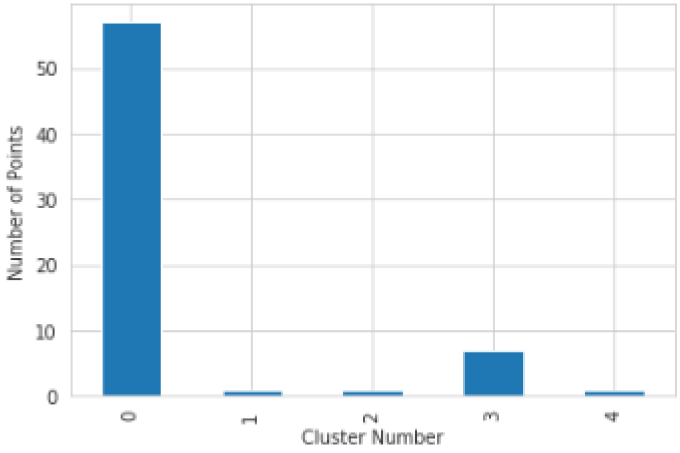


```
from sklearn import metrics
from scipy.spatial.distance import cdist
import matplotlib.pyplot as plt

# k means determine k
distortions = []
K = range(1,33)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(finalmerge)
    distortions.append(sum(np.min(cdist(finalmerge, kmeanModel.cluster_centers_, 'euclidean'), axis=1)) / finalmerge.shape[0])

# Plot the elbow
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```

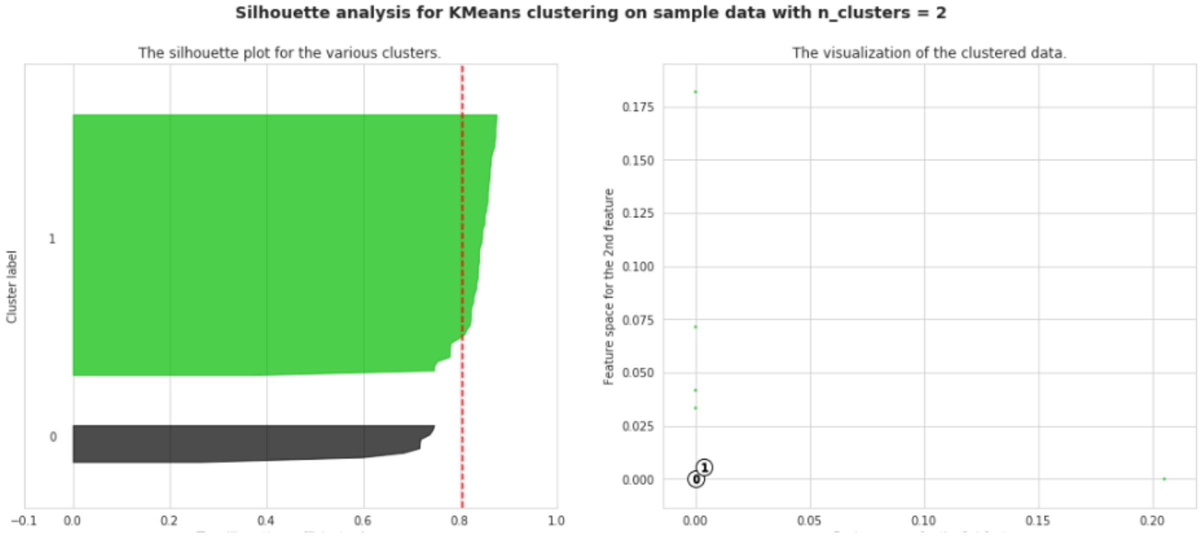
By adjusting the “range” in the code above, one can tune the models “Euclidean distance” across various amounts of clusters.



The Elbow Method proved unfruitful however, with optimal K's often resulting in clusters that contained 0 or 1 neighborhoods. Most parameters decided that 5 was the optimal number of K Clusters.

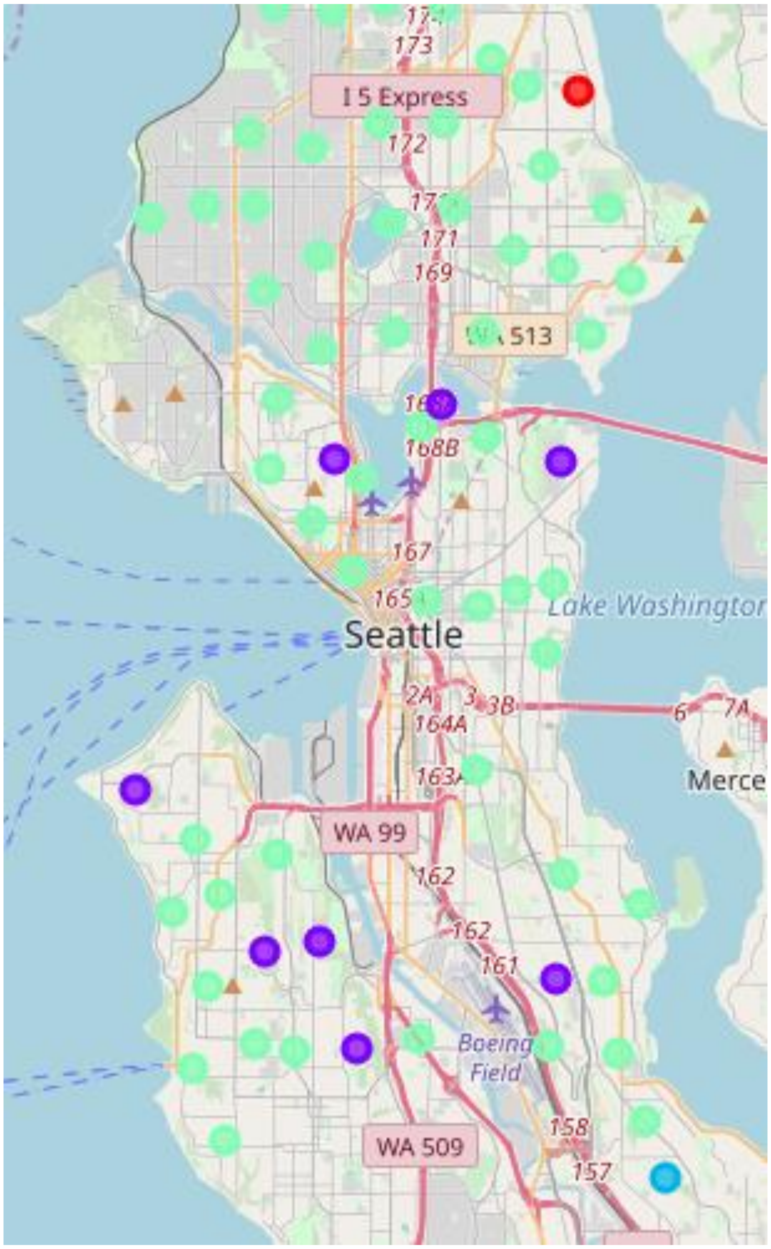


As an alternative to the Elbow Method the Silhouette Analysis confirmed 2 and 3 Clusters yielded the most optimal results



Coffee Shop Cluster

In either case, the K Means algorithm often created two main Clusters; Coffee Shops and Green Space. Not surprisingly, the green dots in the map below were grouped together by the high frequency of coffee shops. The purple dots represent neighborhoods with a high density of Gardens, Parks, and Playgrounds.

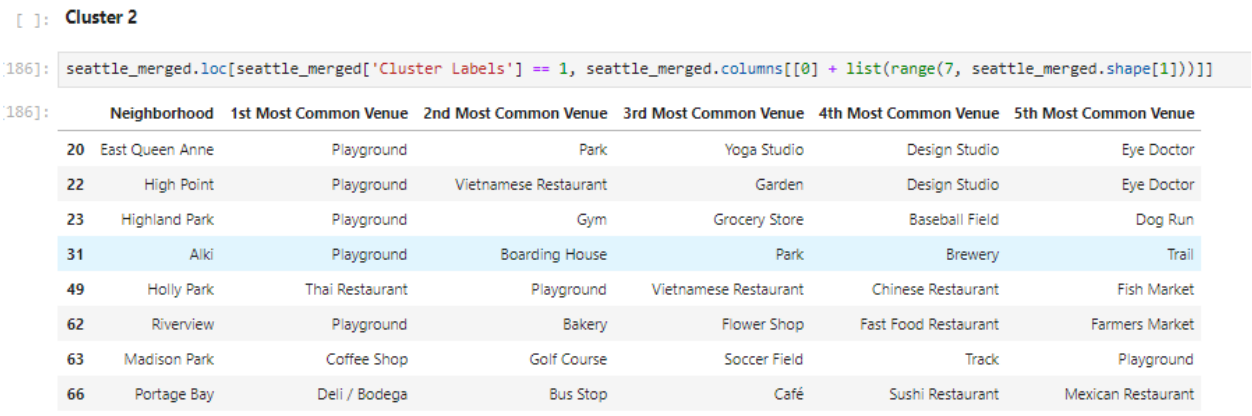


Green Space Cluster

While it was obvious that Coffee Shops were inherently common in Seattle, it was interesting to see the K Means group together Green Space neighborhoods. The **Madison Park** neighborhood in the Cluster below stands out in particular.

The 1st Most Common Venue in **Madison Park** is Coffee Shop. However, it was not grouped with the main Coffee Shop cluster. The next most common venues were Golf Course, Soccer Field, Track and Playground, which resulted of being placed in the Green Space cluster.

If you recall **Madison Park** was one of the neighborhoods with the 2nd highest Zillow Value. Curiously enough, **Portage Bay**, the 4th highest Zillow Value also shows up in this cluster. The other neighborhoods (save for East Queen Anne) are in the less desirable neighborhoods located South of Downtown.



[ ]: **Cluster 2**

```
186]: seattle_merged.loc[seattle_merged['Cluster Labels'] == 1, seattle_merged.columns[[0] + list(range(7, seattle_merged.shape[1]))]]
```

186]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
20	East Queen Anne	Playground	Park	Yoga Studio	Design Studio	Eye Doctor
22	High Point	Playground	Vietnamese Restaurant	Garden	Design Studio	Eye Doctor
23	Highland Park	Playground	Gym	Grocery Store	Baseball Field	Dog Run
31	Alki	Playground	Boarding House	Park	Brewery	Trail
49	Holly Park	Thai Restaurant	Playground	Vietnamese Restaurant	Chinese Restaurant	Fish Market
62	Riverview	Playground	Bakery	Flower Shop	Fast Food Restaurant	Farmers Market
63	Madison Park	Coffee Shop	Golf Course	Soccer Field	Track	Playground
66	Portage Bay	Deli / Bodega	Bus Stop	Café	Sushi Restaurant	Mexican Restaurant

Section 3: Scratching the Surface and Next Steps

A. The Curse of Dimensionality: Predicting Zillow Value from 220 Unique Venues

Does a Relationship Exist Between Venue Type and Zillow Value?

Finally, we wanted to examine the relationship between Seattle Venues and Zillow Value. Multivariate Regression was used to determine the relationship between ZHVI and **all 220** unique Venue categories.

Of course, data becomes sparse when adding so many dimensions to the model. As more dimensions are added to the model, an exponential amount of data points are required to keep the same data density.

The Curse of Dimensionality shows us that the data density of 20 data points in a 1-dimensional space is equivalent to 8000 datapoints in a 3-dimensional space. With over 220 variables, its impossible to predict Zillow Value with such a simple setup.

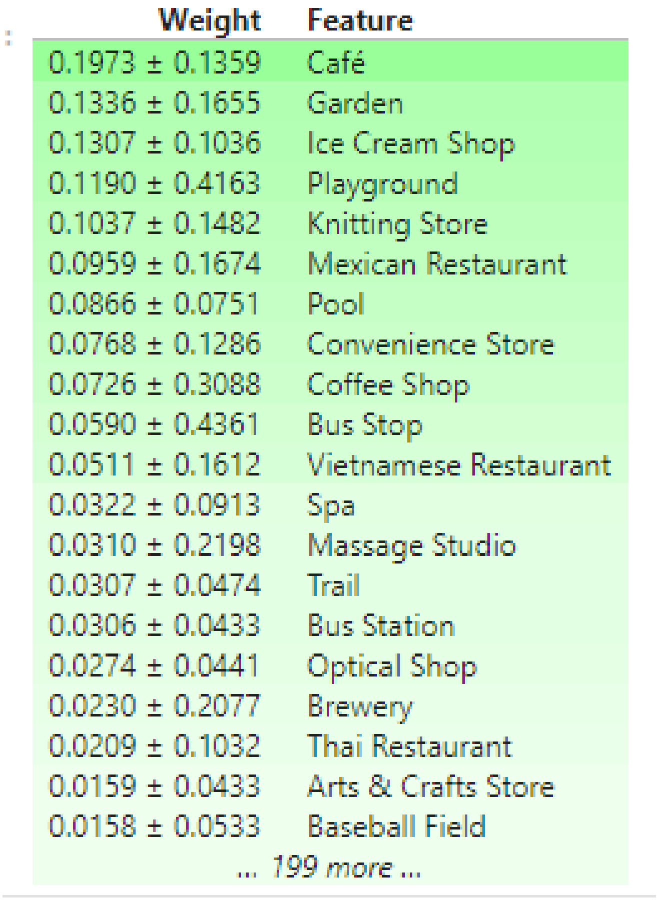
To illustrate this issue, our Foursquare query is yielding only 79 Coffee Shops!

Feature Extraction

The only way to accurately to find any predictors of Zillow Value in the Venue data is to reduce the dimensionality of the data. Gardens, Parks, Playgrounds, and other green spaces need to be combined into one feature called Green Spaces. Café's need to be further binned into the Restaurant or Coffee Shop category.

A few techniques were also used to try to find out Feature Importance

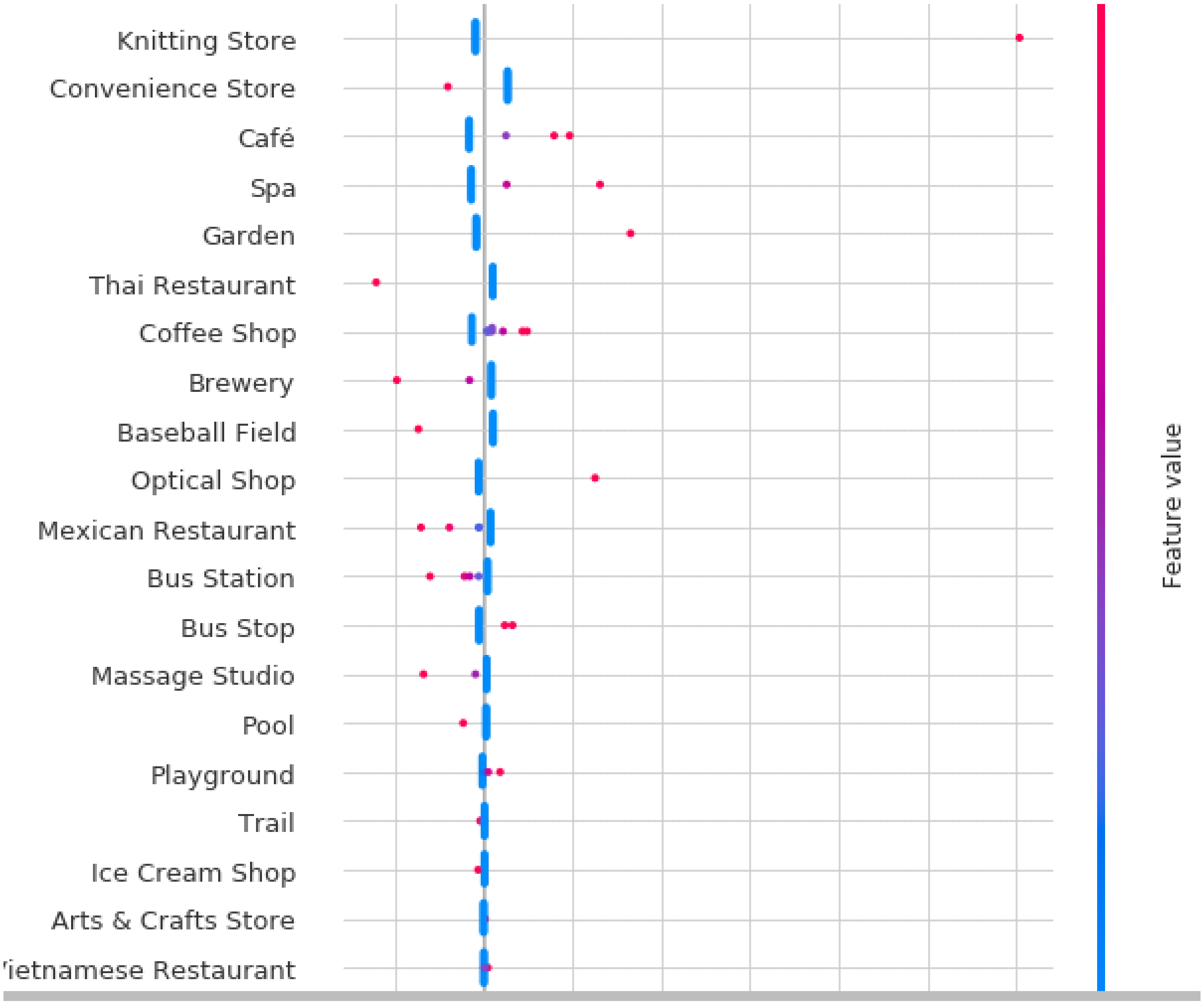
The ELI5 library was used to examine the Permutation Importance of the Regression Model (below). While no predictions can be made, it helps to highlight the weights of certain venues for feature extract.



Weight	Feature
0.1973 ± 0.1359	Café
0.1336 ± 0.1655	Garden
0.1307 ± 0.1036	Ice Cream Shop
0.1190 ± 0.4163	Playground
0.1037 ± 0.1482	Knitting Store
0.0959 ± 0.1674	Mexican Restaurant
0.0866 ± 0.0751	Pool
0.0768 ± 0.1286	Convenience Store
0.0726 ± 0.3088	Coffee Shop
0.0590 ± 0.4361	Bus Stop
0.0511 ± 0.1612	Vietnamese Restaurant
0.0322 ± 0.0913	Spa
0.0310 ± 0.2198	Massage Studio
0.0307 ± 0.0474	Trail
0.0306 ± 0.0433	Bus Station
0.0274 ± 0.0441	Optical Shop
0.0230 ± 0.2077	Brewery
0.0209 ± 0.1032	Thai Restaurant
0.0159 ± 0.0433	Arts & Crafts Store
0.0158 ± 0.0533	Baseball Field

... 199 more ...

The SHAP package was also imported to further examine the weights and feature importance of the model. The SHAP package is often used to help explain the predictions of complex models. The algorithm computes the Shapley Value, which utilizes Game Theory. It is named after the Nobel Prize winning Economist Lloyd Shapley.



Based on the Permutation and SHAP Value I decided to look for correlations between Zillow Value and a few of the important features. None of the Correlation Coeffients yielded anything significant. For example, Garden and ZHVI came in a 0.38.

Out of a thirst for curiosity and validation, I did explore the relationship between a few of the Venues that were One Hot Encoded. One strange discovery was that Antique Shops had many correlations with other venues. Unsurprisingly, Airports had a 99% Correlation Coeffient with both Airport Terminals and Heliports.

B. Final Recommendations: We Have Alot to Learn

In summary, using data science to identify hidden value in a crowded Seattle market has proven to be much more complex than initially hypothesized. I’ve learned many techniques from this study, and even applied some concepts that I started learning in more advanced courses.

I realize that even after committing several dozens of hours to this study, I’m only just scratching the surface of discovering informed statistical investment opportunities.

While exploring the FourSquare geolocation data was interesting, I did find several flaws and holes that would take much more due diligence to accurately correct.

Moving forward, I would recommend purchasing the FourSquare Premium account in order ensure full venue querying and access to consumer trends. I would also recommend sourcing and purchasing other alternative forms of data that could explore:

- Foot Traffic
- Parking Lot Traffic
- Cell Phone Geo Location Data
- Consumer Trends

In the meantime, I will continue exploring this data and crafting it into a professional project. I plan to further reduce the dimensionality of the data, and categorize all “Green Space” or “Breakfast+Lunch” venues one feature. Examples of desired outcomes I would like to explore are:

- **X** Green Space Venues = **Y**% Increase in ZHVI
- Does the density of bars predict ZHVI?
- Do any neighborhoods need more Coffee Shops?
- Do “Breakfast and Lunch Only” venues indicate lower ZHVI?
- Create a heatmap of foot traffic in the First Hill neighborhood
- Cell Phone geo location density by Day/Night cycle in certain neighborhoods

Section 4: Bibliography

FourSquare. (2019, May). *FourSquare API.* Retrieved from 'https://api.foursquare.com/v2/venues/explore

Kienzler, R., & Manchev, N. (2019, June). *Advanced Machine Learning and Signal Processing by IBM*. Retrieved from Coursera: https://www.coursera.org/learn/advanced-machine-learning-signal-processing

Rosenberg, M. (2019, January 15). *The Seattle Times.* Retrieved from Seattle Still Has the Most Cranes in America, and Construction isn't Losing Much Steam: https://www.seattletimes.com/business/real-estate/seattle-still-has-the-most-cranes-in-america-and-construction-isnt-losing-much-steam/

SeattleIO. (2019, May). *seattle-boundaries-data.* Retrieved from Github: https://github.com/seattleio/seattle-boundaries-data

Zillow. (2019, April). *Zillow Research.* Retrieved from Economic Data: https://www.zillow.com/research/data/