Portland Neighborhood Selection

NOVEMBER 14

COMPANY NAME
Authored by: Sean Rumberger

Contents

1. Introduction

1.1 Background

Portland is the largest city in Oregon both by total area and population (Wikipedia, 2019). It was founded in 1845 and was named after Portland, Maine by Francis W. Pettygrove who won the right to name the city in a coin toss (Wikipedia, 2019). The city has seven boroughs known as neighborhood coalitions which are essentially groups of neighborhoods (ARCGIS, n.d.). In terms of population, it was estimated in 2018 that Portland was home to 653,115 residents (Wikipedia, 2019). Both the Willamette and Columbia rivers run through the city and Portland is about two hours from the Oregon Coast and the Cascade mountains. This allows residents access to many outdoor activities, including hiking, mountain biking, wind surfing, skiing, and surfing. Additionally, Portland is the third most affordable city on the West Coast of the United States, with a median home price of $375,425, and was ranked number 8 on the "Overall Best Places to Live List" (Thorsby, 2019). Therefore, it is an attractive place to both live and/or invest in a home.

1.2 Problem

The author currently lives in Southern California where Real Estate is simply out of his price range. He wishes to own a home so he can A) begin putting his money towards equity instead of rent and B) secure a home to live for himself and his family. However, in order to solve this problem, the following questions must be addressed: What neighborhoods would be within his price range? And of those neighborhoods within the author's price range, which neighborhoods would be suitable to live based on crime statistics and nearby venues?

1.3 Interest

This problem and analysis presented below is potentially of interest to those seeking to relocate to the city of Portland. It is also of interest to potential real estate investors looking to purchase property in Portland. Furthermore, the analysis can be duplicated for any city for those looking to invest or move to a particular location.

2. Data

2.1 Data Sources

The following data and accompanying sources were required to determine which neighborhoods were acceptable for the author's subjective needs:

- **Neighborhood Data:** This includes a JSON file to render choropleth maps of the neighborhoods and two csv files which contained the Coalition Names, Neighborhood Names and Area. All of which was obtained from ARCGIS.
- **Geocoordinates:** Longitude and Latitude values were obtained using ARCGIS as well.
- **Crime Data:** Specifically the Offense Counts, Offense Category, and Neighborhood of Occurrence were obtained via a csv file from portlandoregon.gov/police/71978.
- **Real Estate Pricing Data:** Which included Median home price, Price per Square Foot sorted by the Neighborhood Name were scraped from Portland Monthly.
- **Venue Data:** Venue data was obtained via Foursquare API.

2.2 Data Cleaning

Three data sets in total were obtained via csv files and one was obtained using the lxml scraping tool. The Coalition Code and Name were kept from the Coalition csv file, while the Coalition Code, Neighborhood name and neighborhood area were kept from the neighborhood csv file. The two tables were then joined on the Coalition Code and unclaimed areas were dropped from the dataframe because they increased the length of the dataframe from 98 entries to 118. The resulting dataframe contained 94 neighborhoods.

Concurrently, the housing price data from the lxml scraping was processed and the neighborhood name along with the corresponding median prices, average prices and price per square foot for each of those neighborhoods were kept. This dataframe was then joined with the Neighborhood/Coalition dataframe using the neighborhood name as an index.

Finally, the crime data was processed and added to the dataframe (see description of processing in Methodology Section). Upon inspection of the data it was discovered that a number of neighborhoods had neither housing price data nor crime data. This suggested an absence of human activity in these neighborhoods, so they were dropped from the data set. The resulting dataset contained 70 neighborhoods.

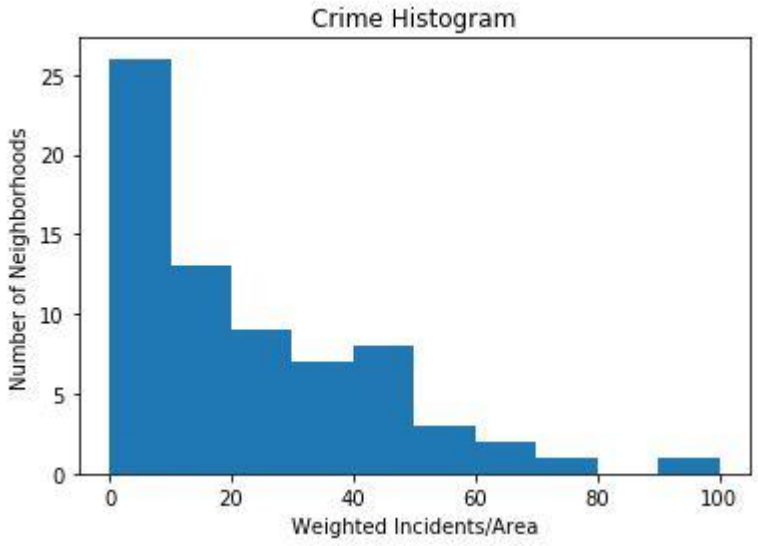
3. Methodology

3.1 Data Processing and Weighting Methodologies

An ARCGIS loop was created to get the longitude and latitude values for each neighborhood. This data was then appended to the dataframe. ARCGIS was also used to get the coordinates of the city of Portland which was used to visualize the neighborhoods in a choropleth map.
The Median home price data was used to assess which neighborhoods were within my price range and which were not. Median price was chosen as the metric for home pricing data over the average home price to avoid any potential skewing of the data by outliers. The acceptable price threshold was set to $500,000 and below; however, the author's target price is in the low $400,000 range. He set the median price threshold to $500,000 because there are likely homes within his target price range even if the median price is $500,000 and because housing prices are negotiable. After setting the threshold, a function was created to convert the median home price data into binary data. Meaning, those houses above the median home price threshold received affordability score of zero and those within the threshold received a one. The results of this analysis are displayed in figure 2 of the results section.

The Offense Category from the crime data for the current year was weighted according to the offense category. This weighting was a subjective measure of how the author viewed the seriousness of the offenses. For example, Homicides received a weighted score of five, while petty theft received a weighted score of one. After the Offense Categories were weighted, the weighted value was then multiplied by the number of counts associated with the particular incident. The weighted counts were then grouped by neighborhood, divided by neighborhood area in order to normalize the data for neighborhood size. Finally, the data was set to a maximum score of 100 by dividing each weighted incident/area score by the maximum score and multiplying the result by 100, thus creating a crime score from 0 to 100 where 100 is the highest weighted incident count by area. The resulting crime score data was then visualized in a histogram plot (Figure 1) to determine an appropriate threshold crime score.

Figure 1: Histogram plot showing the number of neighborhoods on the y axis and the crime score on the x axis.



The crime data was separated into bins of ten, the lowest bin contained neighborhoods with a crime score of 0 to 10 while the highest bin contained neighborhoods with a score from 90 to 100. From to histogram plot it is evident that the data is skewed to the left, meaning that there are many neighborhoods with low crime and as the crime score increases the number of neighborhoods in those subsequent bins tends to decrease. This allowed the author to confidently set the crime score threshold at 20. A function was created to assign a one value to any neighborhood with a crime score of less than 20 and a zero to any neighborhood with a crime score greater than or equal to 20. See the results section figure 3 for further details.

The binary data for both crime and median price were then be multiplied together in order to generate a list of neighborhoods that met both thresholds. Neighborhoods that did not meet both thresholds were dropped from the dataframe. See the results section of section 4 for further details.

The geocoordinates for each neighborhood were used to collect the venue data by calling the Foursquare API. The radius limit was set to 1000 meters to 1) ensure that enough venues were collected for each neighborhood and 2) because 1000 meters can be walked within 10 to 15 minutes, so it is a reasonable distance for walking access to venues. The crime and housing price data was dropped to convert the venue categories per neighborhood to dummy variables. The dummy variables were then converted into ratios to discover which five venues had the highest ratios. The median home price, cost per square foot, and crime data was added back into the dataframe. The median home price was divided by 10,000 and the Cost per square foot was divided by 100 in order to weight them for K-Means. The author chose to divide the median home price by 10,000 because he wanted home price to be weighted the most in the decision, followed by the crime data, the cost per square

foot, and finally the venue data, since home prices and crime are more important than the type of venues in the neighborhoods. Theoretically the highest value for median home price would be 50 if the home price were $500,000. As mentioned earlier, the highest crime score is 20. Dividing the cost per square foot by 100 meant that the values would be somewhere between 3 and 1, while the venue data would be less than one.

3.2 Machine Learning Methodology

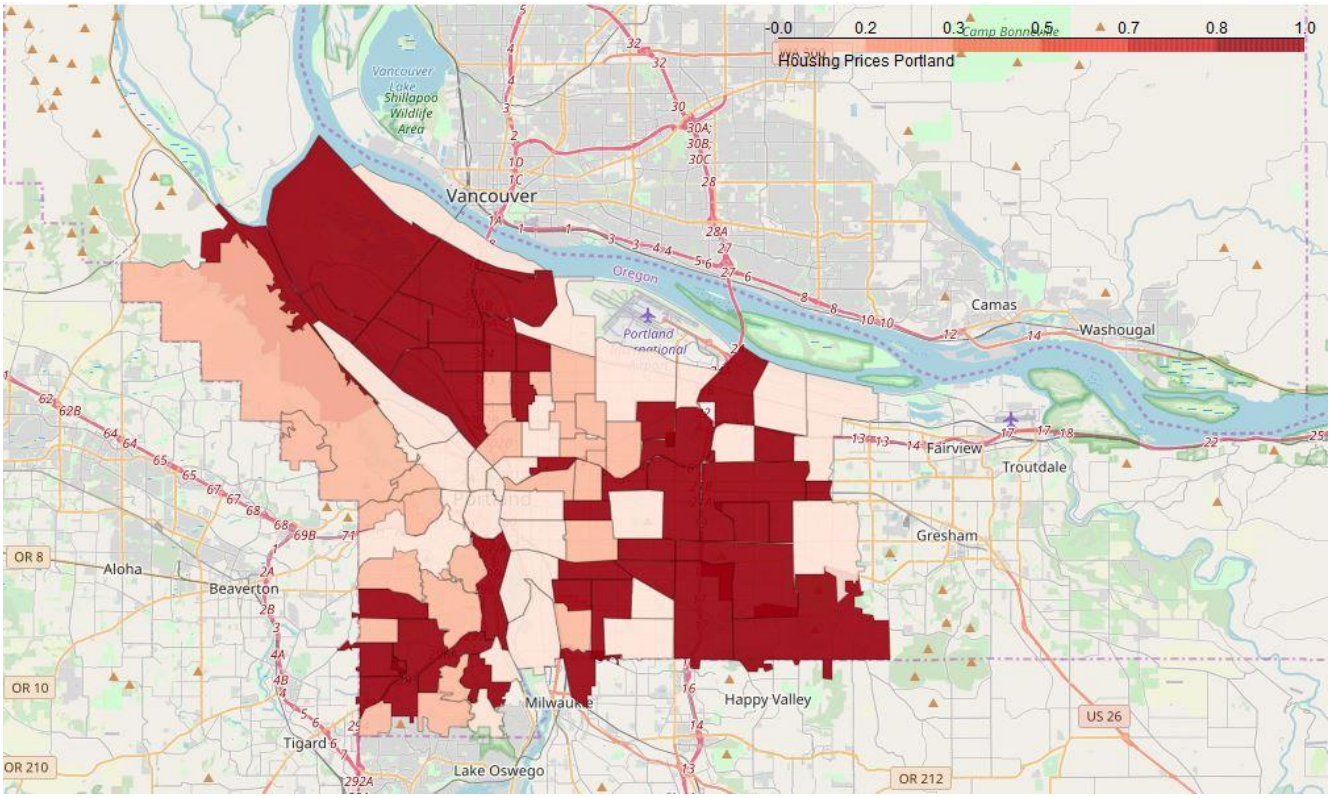
K-Means is a form of unsupervised machine learning that can only be applied to numerical data. With the exception of the neighborhood names, which were dropped to apply K-Means to the data, all of the collected data was numerical. Furthermore, the goal of the project was to group the data into clusters in order to find a suitable set of neighborhoods to begin house hunting. Therefore, K-Means was deemed to be the most suitable machine learning algorithm to apply to the problem. K-Means was be applied to 20 neighborhoods, all of which had been selected based on criteria mentioned above. Three clusters were chosen because each cluster should yield somewhere between five and eight neighborhoods per cluster, which for the author's purposes seemed to be reasonable as it allowed for options while preventing the paradox of choice from coming into play. K-Means was then applied using neighborhood venue data, median price data, cost per square foot, and crime score to generate clusters. These clusters were then analyzed based on their suitability and the cluster deemed most suitable was used to narrow the home search to those particular neighborhoods.

4. Results

4.1 Neighborhoods with a Median Home Price of less than $500,000

Figure 1 shows the results of the binary median home price analysis. Neighborhoods in dark red have a median home price of less than $500,000, while neighborhoods in the lighter red color have a median home price of $500,000 or greater. There is no median home price data associated with the neighborhoods in the lightest shade of red.

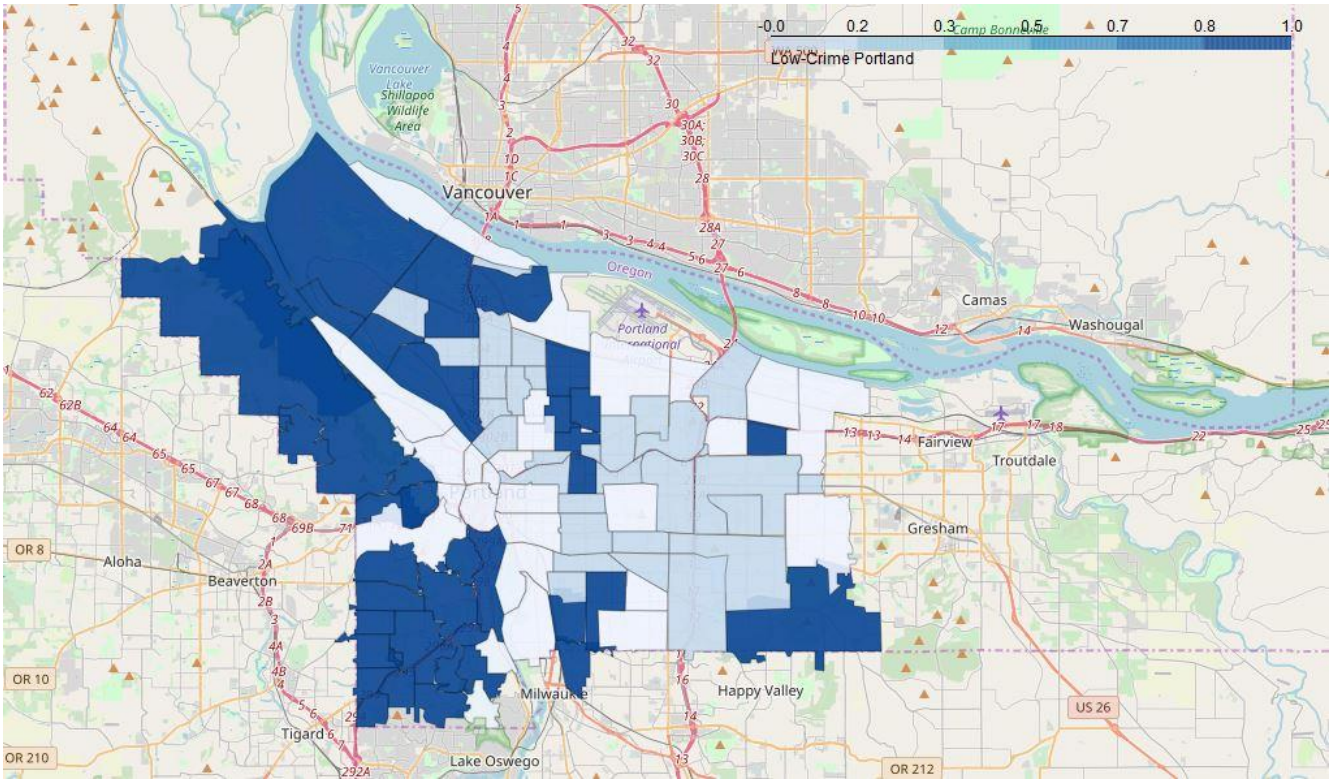
Figure 2: Results of the binary median home price threshold analysis. Neighborhoods with a median home price of less than $500,000 are displayed in dark red and are considered acceptable.



4.2 Neighborhoods with a Crime Score of less than 20

Figure 3 shows the results of the binary crime score analysis. Those neighborhoods with a crime score of less than 20 are displayed in dark blue. Those neighborhoods with a crime score of greater than or equal to 20 are displayed in the lighter blue color. No data is available for neighborhoods which are displayed in the lightest shade of blue.

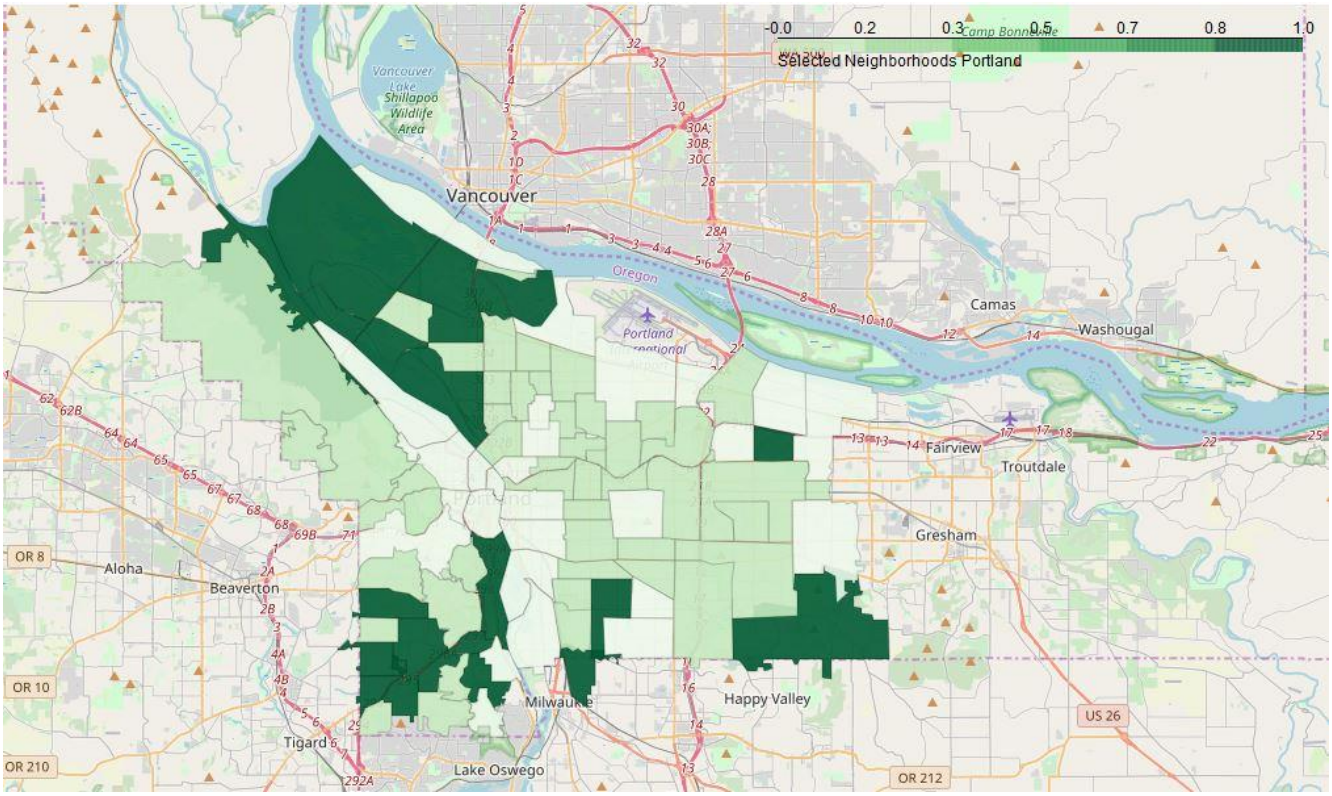
Figure 3: Binary Crime Score data plotted on a choropleth map. Neighborhoods in dark blue have a crime score of less than 20 and are considered acceptable.



4.3 Suitable Neighborhoods Based on Crime and Housing Data

Those neighborhoods with both a crime score of less than 20 and a median home price of less than $500,000 are displayed as dark green in figure 3. Those neighborhoods in the lighter shade of green either have a median home price of more than $500,000, a crime score greater than 20 or both a crime score greater than 20 and a median home price of $500,000 or greater. No data is available for the neighborhoods in the lightest shade of green.

Figure 4: Result of multiplying the median price and crime threshold data. Acceptable neighborhoods are displayed in dark green.

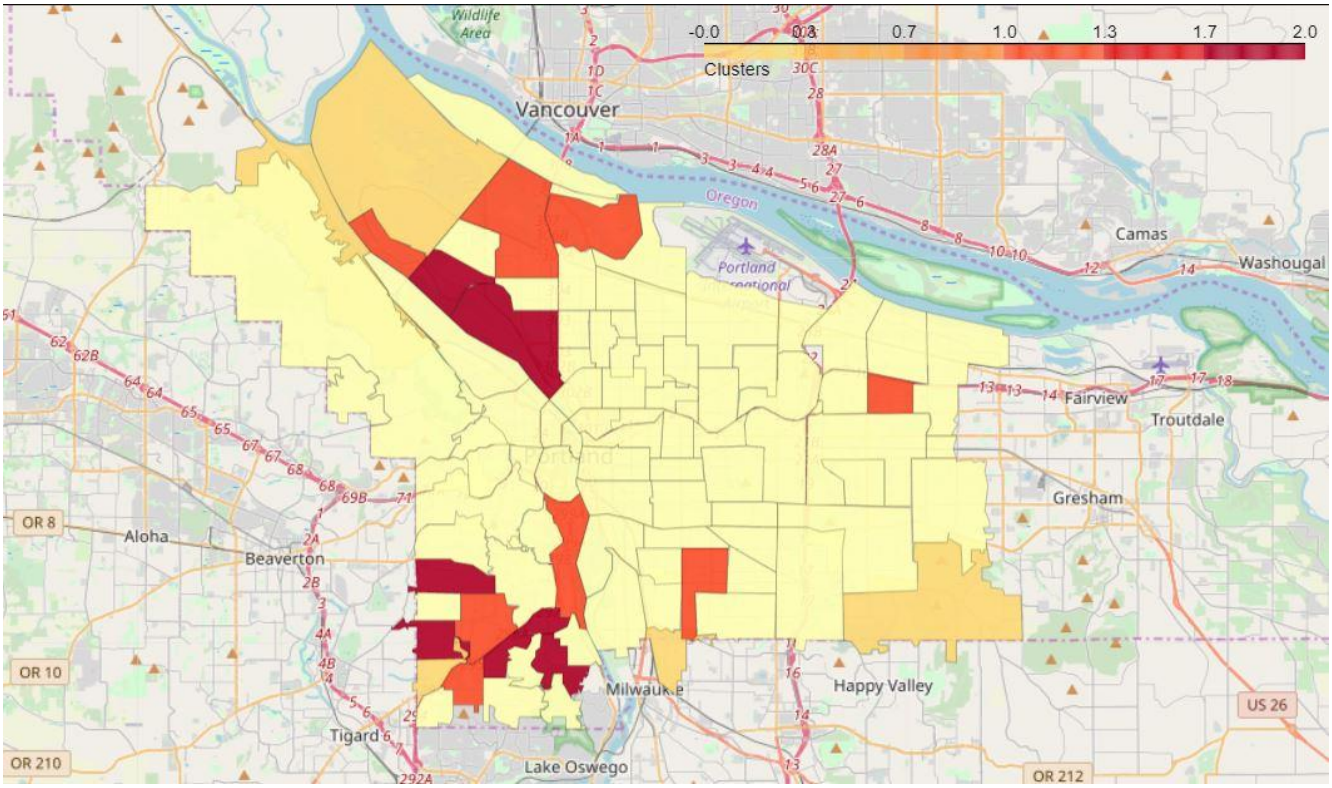


4.4 Clusters

Figure 5 displays the neighborhoods after K-Means was applied to the Median Home price, Crime Score, Cost per square foot, and venue data. The three clusters are displayed below. Cluster one is yellow, cluster two is orange, cluster 3 is red. Those neighborhoods in light yellow were excluded from the analysis because they either did not meet the criteria mentioned above or there was no data for these neighborhoods. Upon analyzing the clusters, a pattern emerged. The neighborhoods within cluster one had a relatively low in median price and low crime score. Neighborhoods in cluster two had a relatively higher crime score and a range of home prices. Finally, the neighborhoods within cluster three had higher prices and a lower crime score.

Figure 5: Results of K-Means clustering. Cluster one is Yellow, cluster two is orange, cluster three is red.

1



5 Discussion and Limitations

5.1 Discussion

Based on the results (see Table 1 below) Cluster 1 looked to be the most attractive because its median home price and crime score was low. Furthermore, these neighborhoods tend to be near natural settings, such as parks, forests, and nature preserves. Therefore, the author would recommend that a home buyer or real estate investor interested in purchasing a house in the city of Portland begin looking within neighborhoods contained in cluster one. These neighborhoods include Pleasant Valley, St. Johns, Linnton, Ardenwald-Johson Creek, and Crestwood.

Table 1: Cluster 1 Neighborhoods and their corresponding house, crime, and venue data.

Neighborhood	MEDIAN PRICE	COST/SQ_FT	CRIME SCORE	Common Venue 1	Common Venue 2	Common Venue 3	Common Venue 4	Common Venue 5
PLEASANT VALLEY	420000	182	3.1	Snack Place	Park	Deli / Bodega	Nature Preserve	Stables
ST. JOHNS	360000	242	0	Dog Run	Park	Disc Golf	Yoga Studio	Fish Market
LINNTON	377450	183	0.48	Food	Convenience Store	Bar	Trail	Forest
ARDENWALD-JOHNSON CREEK	434500	185	0	Garden	Stables	Hotel	Nature Preserve	Dive Bar
CRESTWOOD	388500	230	2.2	Gym / Fitness Center	Park	Bank	Smoke Shop	Playground

5.2 Limitations

Unfortunately, the author was unable to find population data related to each of the neighborhoods. Had the author encountered such data he would have used it to obtain a weighted crime per capita number instead of normalizing crime by area. This value would be likely more reflective of actual crime activity than normalizing using area. Furthermore, the area used was not in standard units, instead it was for the purposes of describing the area in the JSON file. This shouldn't matter however as the important detail is the area value relative to the other neighborhoods.

6. Conclusion

Price, crime, and venue data was collected. The crime data was weighted and normalized to reflect a crime score that took the size of the neighborhood and offense category into account. Thresholds were set for both median price and crime score. All neighborhoods not meeting this threshold were dropped. K-Means was then used to cluster neighborhoods based on the crime, price and venue data, resulting in a group of three clusters. Of the three clusters the author determined that the most attractive cluster, for the purposes of finding a home was cluster one.

The application of this analysis can be used with any threshold for crime and median home price. It could be potentially beneficial for Portland home buyers and real estate investors.

References

ARCGIS. (n.d.). *City of Portland, Oregon*. Retrieved from GIS-PDX: http://gis-pdx.opendata.arcgis.com/

DeNies, R. (2019, March 26). Portland Neighborhoods by the Numbers 2019: The City. *Portland Monthly*. Retrieved from https://www.pdxmonthly.com/articles/2019/3/26/portland-neighborhoods-by-the-numbers-2019-the-city

Portland Police Bureau. (2019, October 29). *Crime Statistics*. Retrieved from The City of Portland: https://www.portlandoregon.gov/police/71978

Thorsby, D. (2019, October 2). The Best Affordable Places to Live on the West Coast. *US News*. Retrieved from US News: https://realestate.usnews.com/real-estate/slideshows/best-affordable-places-to-live-on-the-west-coast?slide=9

Wikipedia. (2019, November 5). *Portland, Oregon*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Portland%2C_Oregon