# Applied Data Science Capstone by IBM through Coursera

## Capstone Project - The Battle of the Neighborhoods (Week 2)

**Brian Shelton**

## 1. Introduction

### 1.1 Background

The University of Washington is one of the top school choices for students to attend in the state of Washington. With a total student enrollment of 46,166 students and 15,022 of those students enrolling in the summer of 2019 at the Seattle campus alone shows how popular the school is. Attending college is a significant part in a lot of individuals' lives, whether they're attending for the first time or returning. Going to college can bring about a lot of stress due to tuition, courses, being in a new environment and much more.

### 1.2 Problem

One of the biggest concerns when going off to college is, where are you going to live? This is a significant concern because expenses can be too much for families or an individual to bear along with additional expenses such as tuition, parking, school supplies, etc. For those attending college, saving as much money as possible is a significant goal itself, outside of receiving a degree.

For my project, I will focus on identifying the most cost-efficient neighborhoods around the University of Washington within the city of Seattle for incoming students who are interested in living off campus. This analysis will provide an informative outlook on the many neighborhoods that are within a four-mile radius (North, South, East, West) of the UW Medical Center/UWSOM on campus.

### 1.3 Interests

This analysis should interest incoming students who are planning on attending the University of Washington at the Seattle campus and want to live near it. Others who may have an interest are students' families who may aid them or even be living with them during their time at the University.

## 2. Data acquisition

Using the definition of the business problem as a basis for this analysis and decision process, these factors will be taken under consideration:

- Average rental cost in each respective neighborhood.
- The total crimes committed/crime rate within each respective neighborhood.
- The type of venues that are provided within each neighborhood.

### 2.1 Data sources

The data used for this project will be the following:

- The neighborhoods I will be focusing on will be the neighborhoods suggested from the University of Washingtons' (https://blogs.uw.edu/esom/seattle-living/housing-neighborhoods/). Their respective coordinates I found using geopy library to convert neighborhood, city and state, which is the address into latitude and longitude values.
- I will be also collecting current average rental costs of the neighborhoods from a webpage that focuses on current rental market trends within the Seattle Area (https://www.rentcafe.com/average-rent-market-trends/us/wa/seattle/). This data will also be put into the data frame.
- To obtain the total crimes committed within each neighborhood, I will be using the crime dataset provided by the Seattle government website (https://data.seattle.gov/Public-Safety/Crime-Data/4fs7-3vj5/data). This data includes all of the crimes committed in Seattle.

Table 1. Neighborhood data frame with coordinates

| | Report Number | Occurred Date | Occurred Time | Reported Date | Reported Time | Crime Subcategory | Primary Offense Description | Precinct | Sector | Beat | Neighborhood |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 523586 | 2019000099944 | 03/20/2019 | 1713.0 | 03/20/2019 | 1713.0 | FAMILY OFFENSE-NONVIOLENT | CHILD-OTHER | SOUTH | O | O3 | MID BEACON HILL |
| 523587 | 2019000099946 | 03/20/2019 | 730.0 | 03/20/2019 | 1721.0 | BURGLARY-RESIDENTIAL | BURGLARY-FORCE-RES | EAST | C | C2 | MONTLAKE/PORTAGE BAY |
| 523588 | 2019000099949 | 03/20/2019 | 1724.0 | 03/20/2019 | 1724.0 | ROBBERY-COMMERCIAL | ROBBERY-BUSINESS-BODYFORCE | SOUTH | S | S2 | RAINIER BEACH |
| 523589 | 2019000099974 | 03/20/2019 | 1750.0 | 03/20/2019 | 1904.0 | THEFT-SHOPLIFT | THEFT-SHOPLIFT | NORTH | L | L2 | NORTHGATE |
| 523590 | 2019000099993 | 03/19/2019 | 1800.0 | 03/20/2019 | 2237.0 | THEFT-ALL OTHER | THEFT-OTH | NORTH | N | N1 | BITTERLAKE |

- The venues within each neighborhood will be found utilizing the search-and-discovery mobile app Foursquare to give almost a full view what each neighborhood has to offer providing information on top venues within each location.

The data collected will in turn will aid college students and their family's decision on where they should stay based on their interests and or needs.

## 3. Methodology

The basis of this project is to find the most cost-efficient neighborhood listed by the UW website where one could possibly live. In addition to finding the cheapest location, I also will provide a better look at each neighborhood obtaining the total crimes committed and calculate the crime rate within each neighborhood as well as the venues locally available.

### 3.1 Data Cleaning and Feature Selection

The initial phase of this project consisted of getting the collected data into data frames that only had the wanted features such as average rent totals, coordinates and crime totals. This process began by creating a data frame of the desired features. Then a list of all of the neighborhoods that were collected from the UW website was created and then added to the neighborhood column of the data frame. Since I am only working with neighborhoods in the city of Seattle, the city (Seattle) and state (Washington) were added to the data frame to help find the coordinates of each neighborhood using the geopy library. With the average rent totals already known before analysis, this data was added to the rent average column in data frame to complete the first data frame needed for analysis (see Table 1.).

Table 2. Neighborhood data frame with coordinates and average rent values

| | Neighborhood | City | State | Latitude | Longitude | Avg. rent |
|---|---|---|---|---|---|---|
| 1 | University District | Seattle | Washington | 47.661298 | -122.313152 | 1931 |
| 2 | Wallingford | Seattle | Washington | 47.659463 | -122.334342 | 1997 |
| 3 | Laurelhurst | Seattle | Washington | 47.663432 | -122.277070 | 1931 |
| 4 | Ravenna | Seattle | Washington | 47.675654 | -122.297626 | 1922 |
| 5 | Roosevelt | Seattle | Washington | 47.677305 | -122.313807 | 1909 |
| 6 | Eastlake | Seattle | Washington | 47.643145 | -122.326172 | 2357 |
| 7 | Wedgwood | Seattle | Washington | 47.690253 | -122.290811 | 1909 |
| 8 | Capitol Hill | Seattle | Washington | 47.623831 | -122.318369 | 2052 |
| 9 | Fremont | Seattle | Washington | 47.650453 | -122.349986 | 2072 |
| 10 | Green Lake | Seattle | Washington | 47.680155 | -122.324094 | 2003 |
| 11 | Mapleleaf | Seattle | Washington | 47.701368 | -122.294561 | 1727 |
| 12 | Lake City | Seattle | Washington | 47.719162 | -122.295494 | 1727 |
| 13 | Central District | Seattle | Washington | 47.603110 | -122.308270 | 2484 |
| 14 | Greenwood | Seattle | Washington | 47.690981 | -122.354877 | 1798 |
| 15 | Ballard | Seattle | Washington | 47.676507 | -122.386223 | 2036 |
| 16 | Montlake | Seattle | Washington | 47.641408 | -122.303044 | 1956 |
| 17 | Sand Point | Seattle | Washington | 47.682359 | -122.264312 | 1909 |

The crime data used in this analysis was retrieved from City of Seattle Open Data portal, downloading the csv file that contained all of the crimes committed within the city of Seattle up to date. For the sake of this analysis I decided to only keep the crimes from the years 2018 and 2019 because we are still in 2019, I would not be able to solve for the crime rate for the year and having last year's data could help give a better idea of how safe each neighborhood is due to it being the most recent complete year.

Upon downloading the crime dataset, I noticed a couple of problems with it. The first issue I ran into was that some of the crimes committed had unknown locations (neighborhoods) as well as unknown precincts to possibly match a crime to a neighborhood that way. Due to it being impossible to identify where these crimes took place, I decided to remove all of the cases where the neighborhood was unknown.

The second issue I noticed within the dataset was that the Seattle Police Department was using a condensed map to document crimes committed. Some neighborhoods were merged into another neighborhood, some neighborhoods were combined with another and some were sectioned with cardinal directions (Ballard North and Ballard South etc..). Seattle contains at least 30 neighborhoods within the city; with each neighborhood having their own boundary however, the crime map Seattle Police Department had created differed from the neighborhood map used on UWs' webpage of the city of Seattle in the labeling of the neighborhoods. This meant that for crimes committed in a neighborhood that was merged into a bigger neighborhood, it would count as the larger neighborhood. For crimes committed in neighborhoods that were combined together (Roosevelt/Ravenna) counted for that general area and not in particular neighborhood alone and for crimes committed in a neighborhood sectioned with cardinal directions, they counted for that particular section of the neighborhood. To fix this issue I compared the map provided on UWs' website and the Seattle Police Departments' map identify which neighborhoods were merged into larger neighborhoods, neighborhoods that were combined together and which neighborhoods were sectioned into cardinal direction. I then relabeled the data that had just the larger neighborhood name to all of the neighborhood (including itself) that it contained. I

relabeled the neighborhoods that were sectioned by coordinate directions to just its neighborhood name then used another line of code to combine those neighborhoods into a single neighborhood (Ballard North and Ballard South -> Ballard). For the neighborhoods that were combined with another I kept the same because identifying which particular crime took place in one of the combined neighborhoods would be impossible due to no coordinates or address being available in the dataset to match to the coordinates I found using the geopy library.

After making these adjustments, I was able to code the adjusted dataset into a proper data frame that contained the neighborhood(s) the crimes were committed, and the total amount of crimes committed the neighborhood(s). Features in the data frame that were dropped from the final data frame were the date the crimes occurred because it no longer mattered now that two separate data frames were created with just crimes committed in 2018 and 2019, the crime subcategory and primary offense were dropped because the overall number of crimes committed was the only data needed to provide an overview of the crimes committed in each neighborhood as well as providing me with the ability to calculate the crime rate with that value and lastly, the precinct, beat, and sector because they do not hold any significance for this particular analysis.

Table 3. Neighborhood data frame with Total crimes and crime rate

| | Neighborhood | City | State | Latitude | Longitude | Avg. rent | Total Crimes | crime_rate |
|---|---|---|---|---|---|---|---|---|
| 1 | BALLARD | Seattle | Washington | 47.676507 | -122.386223 | 2036.000000 | 959 | 126.859403 |
| 2 | CAPITOL HILL | Seattle | Washington | 47.623831 | -122.318369 | 2052.000000 | 884 | 116.938178 |
| 3 | CENTRAL DISTRICT | Seattle | Washington | 47.603110 | -122.308270 | 2484.000000 | 312 | 41.272298 |
| 4 | EASTLAKE | Seattle | Washington | 47.643145 | -122.326172 | 2357.000000 | 85 | 11.244056 |
| 5 | FREMONT | Seattle | Washington | 47.650453 | -122.349986 | 2072.000000 | 377 | 49.870693 |
| 6 | GREENWOOD | Seattle | Washington | 47.690981 | -122.354877 | 1798.000000 | 321 | 42.462845 |
| 7 | LAKE CITY | Seattle | Washington | 47.719162 | -122.295494 | 1727.000000 | 420 | 55.558863 |
| 8 | MONTLAKE | Seattle | Washington | 47.641408 | -122.303044 | 1956.000000 | 91 | 12.037754 |
| 9 | NORTHGATE | Seattle | Washington | 47.701368 | -122.294561 | 1727.000000 | 896 | 118.525574 |
| 10 | ROOSEVELT/RAVENNA/GREEN LAKE | Seattle | Washington | 47.677704 | -122.311842 | 1944.666667 | 686 | 90.746142 |
| 11 | SANDPOINT/LAURELHURST/WEDGWOOD | Seattle | Washington | 47.678681 | -122.277398 | 1916.333333 | 265 | 35.054997 |
| 12 | UNIVERSITY DISTRICT | Seattle | Washington | 47.661298 | -122.313152 | 1931.000000 | 664 | 87.835916 |
| 13 | WALLINGFORD | Seattle | Washington | 47.659463 | -122.334342 | 1997.000000 | 268 | 35.451846 |

### 3.2 Exploring the neighborhoods and data visualization

The final phase of this analysis consisted of exploring the neighborhoods and data visualization: The process in the phase required utilizing the Foursquare API to explore the neighborhoods of focus within this analysis. To explore the neighborhoods, what was needed was the converted address values (coordinates) to input into the Foursquare API to search for the top 100 venues within a 700-meter radius of the coordinates. Then the data was cleaned and structured into a pandas data frame to see how many venues were returned for each neighborhood. After further coding and data cleaning to narrow down the top venues in each neighborhood to 10, I performed the process of k-means to cluster the neighborhoods into 4 clusters that would be included in the data frame of each neighborhoods top 10 venues. After collecting all of the

results and placing them into a data frame, the average rent and crime data frames were both represented by bar graphs to compare each of the neighborhoods of focus.

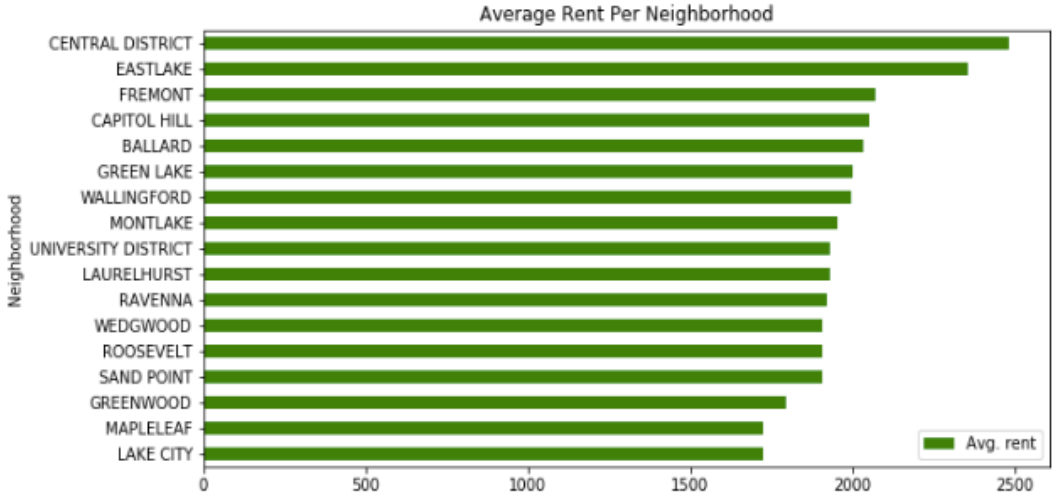


Figure 1. Neighborhood Average Rent

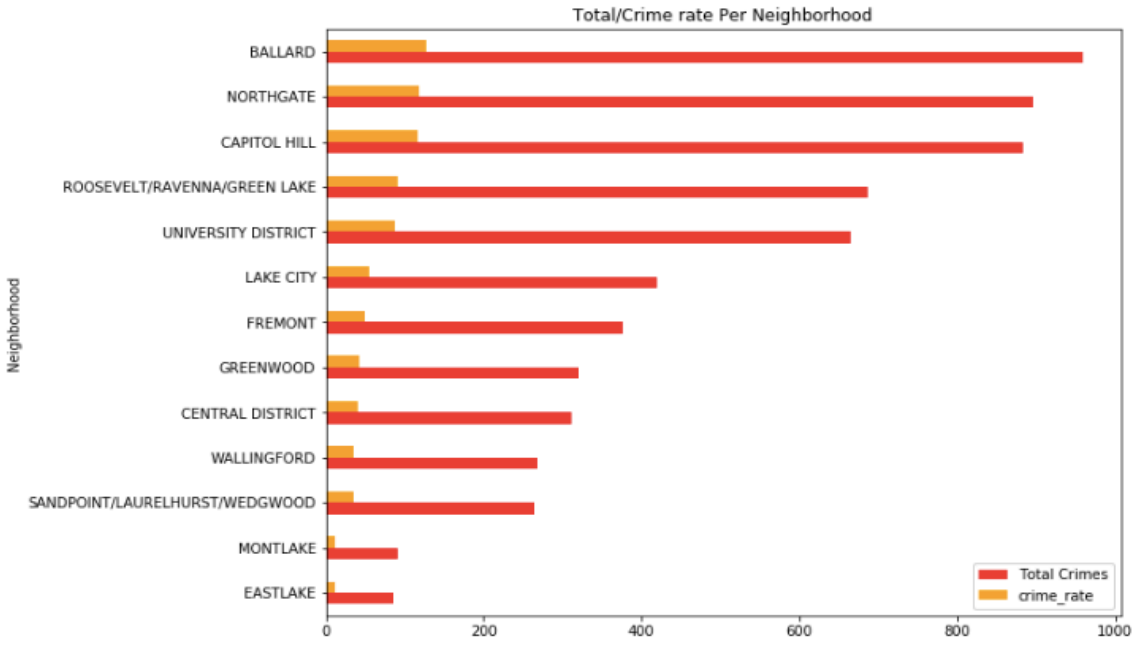


Figure 2. Neighborhood crime rate/total

To visualize the neighborhood clusters, the geopy and folium libraries were used to create a map of Seattle with the neighborhoods and their cluster color superimposed on top.

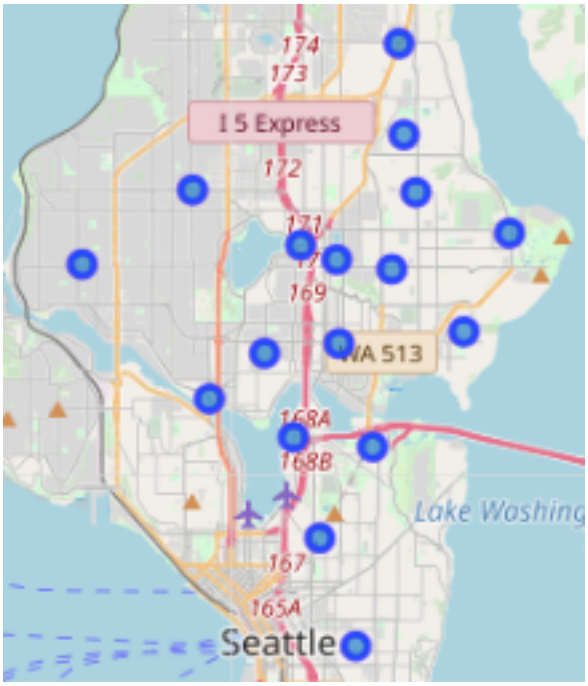


Figure 3. Cluster Map (Example)

### 4.1 Results and Discussion

This analysis showed that there are some great options as to where some students can live that is in close proximity to UWs' medical center that its' campus. The rental costs showed that the neighborhoods were fairly close in range with a few options that were cheaper as well as a few options that were more expensive. With the basis of this analysis focusing on identifying the most cost-efficient neighborhoods around the University of Washington within the city of Seattle, we narrowed our focus down to six cheapest options, Wedgwood, Roosevelt, Mapleleaf, Sand Point, Greenwood and Lake City. These neighborhoods make the most economical sense if trying to live near UWs' campus.

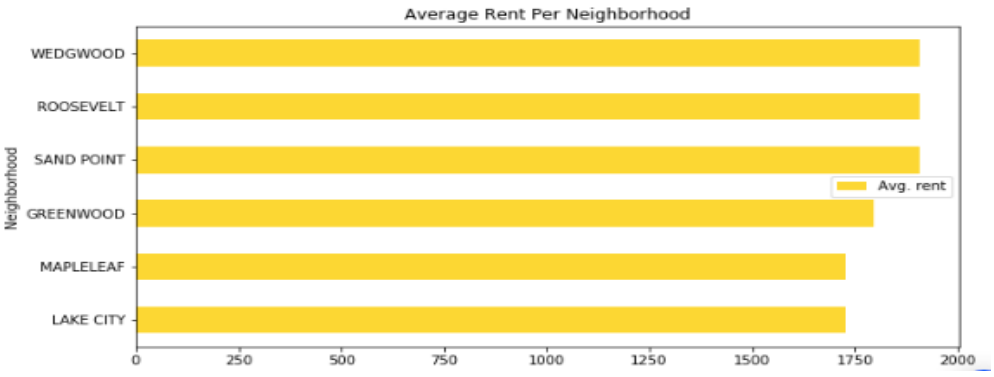


Figure 4. Neighborhood Average Rent (Top 6)

Though rental costs are the main focus for this analysis, this feature alone would not suffice when recommending a place to stay for incoming students. The safety of the new neighborhood they are considering living in was the next point of focus. Using the crime data frame, a bar graph was created to compare how much safer some of the cheaper neighborhoods of focus could be than the others.

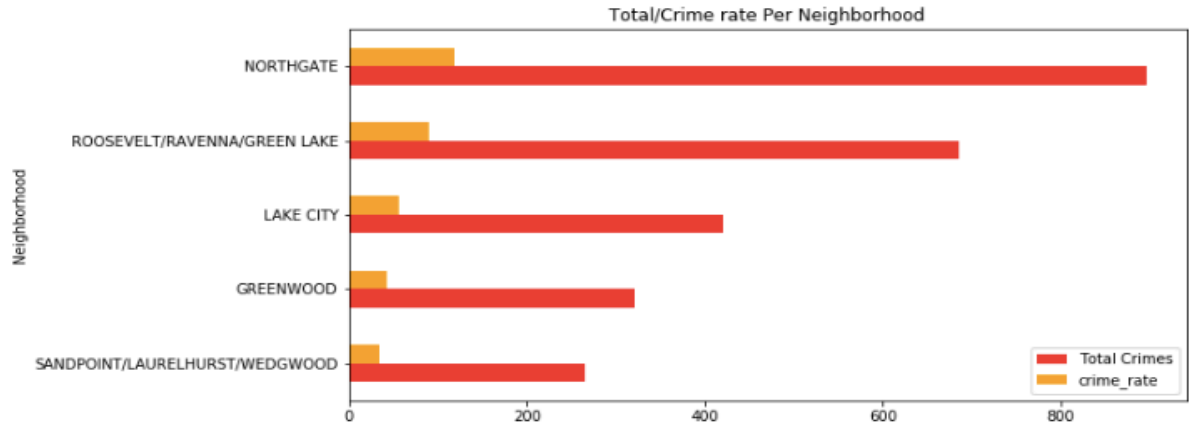


Figure 5. Neighborhood crime rate/total (Top 6)

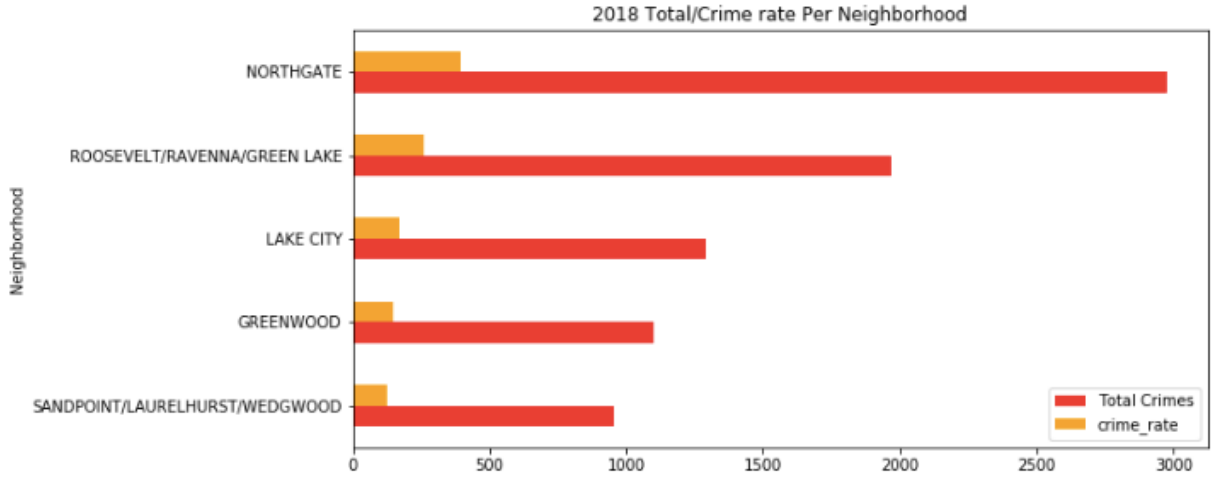


Figure 6. 2018 Neighborhood crime rate/total (Top 6)

From this graph I observed that two particular neighborhoods were not only top three in rental price, they are also in the top three in low crime totals and they were, Wedgwood and Lake City. Though these neighborhoods present an ideal environment for students being that they are the cheaper and safer neighborhoods, this does not imply that these locations in particular will be the best choice for them. This analysis only serves as a foundation or guide for students looking to live in the area.

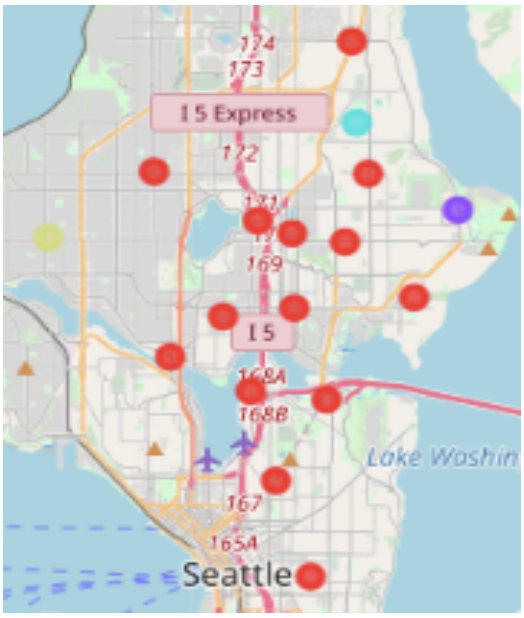


Figure 7. Cluster Map

From this map we can see that the cluster layout is one-sided. Identifying what each cluster represents using Foursquare to find the top venues cannot be determined due to similarities in the common venue citizens visit within their own neighborhood however, for this analysis the point of using Foursquare was to identify the top venues within each neighborhood so incoming college students could see what each neighborhood has to offer.

Table 4. Cluster 1 table

**Cluster 1** ¶

```
sea_merged.loc[sea_merged['Cluster Labels'] == 0, sea_merged.columns[[0] + list(range(5, sea_merged.shape[1]))]]
```

]:

| | Neighborhood | Avg. rent | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | LAKE CITY | 1727 | 0 | Pharmacy | Bank | Mexican Restaurant | Thai Restaurant | Pub | Pizza Place | Shipping Store | Sandwich Place | Breakfast Spot | Brewery |
| 14 | GREENWOOD | 1798 | 0 | Coffee Shop | Mexican Restaurant | Spa | Bookstore | Mediterranean Restaurant | Playground | Pizza Place | Bar | Sandwich Place | Lounge |
| 5 | ROOSEVELT | 1909 | 0 | Coffee Shop | Bar | Vegetarian / Vegan Restaurant | Burger Joint | Gym / Fitness Center | Grocery Store | Rental Car Location | Pub | Pizza Place | Pet Store |
| 7 | WEDGWOOD | 1909 | 0 | Park | Coffee Shop | Pub | Pharmacy | Supermarket | ATM | Grocery Store | Video Store | Gym | Italian Restaurant |
| 4 | RAVENNA | 1922 | 0 | Grocery Store | Mediterranean Restaurant | Pizza Place | Café | Southern / Soul Food Restaurant | Yoga Studio | Bagel Shop | Greek Restaurant | Donut Shop | Creperie |
| 3 | LAURELHURST | 1931 | 0 | Coffee Shop | American Restaurant | Café | Pharmacy | Gift Shop | Bank | Park | Bus Stop | Music Venue | Massage Studio |
| 1 | UNIVERSITY DISTRICT | 1931 | 0 | Coffee Shop | Korean Restaurant | Vietnamese Restaurant | Bubble Tea Shop | Thai Restaurant | Chinese Restaurant | Café | Indian Restaurant | Hotel | Mexican Restaurant |
| 16 | MONTLAKE | 1956 | 0 | Bus Stop | Park | Coffee Shop | Grocery Store | Trail | Tourist Information Center | Canal | Bus Line | Salon / Barbershop | Botanical Garden |
| 2 | WALLINGFORD | 1997 | 0 | Coffee Shop | Thai Restaurant | Japanese Restaurant | Bar | Asian Restaurant | Ice Cream Shop | Pizza Place | Pharmacy | Pub | Café |

In cluster one we can see a common theme between these neighborhoods in that their top venues are mainly places that sell food.

Table 5. Cluster 2 table

**Cluster 2**

```
sea_merged.loc[sea_merged['Cluster Labels'] == 1, sea_merged.columns[[0] + list(range(5, sea_merged.shape[1]))]]
```

| | Neighborhood | Avg. rent | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | SAND POINT | 1909 | 1 | Park | Soccer Field | Playground | Tennis Court | Trail | Indie Movie Theater | Rugby Pitch | Dog Run | Food Truck | Theater |

In cluster two we can see a common theme here, that most of the top venues are recreational places.

Table 6. Cluster 3 & 4 table

**Cluster 3**

```
sea_merged.loc[sea_merged['Cluster Labels'] == 2, sea_merged.columns[[0] + list(range(5, sea_merged.shape[1]))]]
```

| | Neighborhood | Avg. rent | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | MAPLELEAF | 1727 | 2 | Bus Line | Marijuana Dispensary | Liquor Store | Furniture / Home Store | Tennis Court | Automotive Shop | Auto Garage | Lake | Pub | Pool |

**Cluster 4**

```
sea_merged.loc[sea_merged['Cluster Labels'] == 3, sea_merged.columns[[0] + list(range(5, sea_merged.shape[1]))]]
```

| | Neighborhood | Avg. rent | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | BALLARD | 2036 | 3 | Park | Coffee Shop | Bakery | French Restaurant | Candy Store | Restaurant | Burger Joint | Korean Restaurant | Jewelry Store | Baseball Field |

Clusters three and four both show a diverse mix of top venues. For cluster three, we can see there is not a common venue that citizens go to in this neighborhood. As for cluster four, food is a common choice in this neighborhood.

## 5. Conclusions

As previously expressed, the purpose of this project was to identify the most cost-efficient neighborhoods within a four-mile radius (North, South, East, West) of the UW Medical Center/UWSOM on campus in the city of Seattle. This analysis' intent was to provide an informative outlook on the many neighborhoods which in turn will aid college students and their family's decision on where they should stay, based on their interests and or needs. This analysis began with obtaining the necessary data, average rental costs and crime totals, to analyze the neighborhoods. Next, the data was rearranged, modified and cleaned of any unwanted data that would not serve a purpose for this analysis. Coordinates of the neighborhoods were then obtained using the geopy library which in turn provided us with the capability of retrieving their top venues and clustering those locations using Foursquare.

Based off of the data that was collected for this project, This analysis concluded that safest and cheapest neighborhoods for an incoming college student to stay are, Lake City, which had the cheapest rental cost of 1727.00 with a total of 420 crimes committed currently in the year, Greenwood, which had the third lowest rental cost of 1798.00 with the second lowest crime total of 321 crimes committed currently in the year and with a tie, Sandpoint and Wedgwood both costing the most out of the 6 cheaper neighborhoods with a cost of 1909.00 with the lowest crime total, that is split between the neighborhoods along with Laurelhurst at, of 265 crimes currently committed in the year. All of these neighborhoods provide a plethora of venues to go to however, as far as necessities may go and safety, Wedgwood may be the most optimal neighborhood to stay providing all of the necessary venues within the neighborhood; Having venues such as Coffee Shop, Pharmacy, ATM, supermarket, gym and more. This does not serve as a final decision for the stakeholders but an aid for their own decision.