Applied Data Science Capstone – Battle of the Neighborhoods: Which Neighborhood is for Me?

Introduction to the Problem

For this final capstone, I investigate a scenario in which someone is relocating to the county of Arlington, Virginia and they want to know information about cost of living and local venues in different neighborhoods.

Arlington, Virginia is a large, metropolitan city bordering Washington, DC. Due to its location, it has long been home to many United States Federal Employees. Recently, Amazon announced their second headquarters would be built in Arlington, and they intend to house 25,000 employees at the new location (https://en.wikipedia.org/wiki/Amazon_HQ2). Due to this, new residents are expected to move to the Northern Virginia area in the next few years, including data scientists who are being hired to work at Amazon.

Moving to a new area can be daunting, especially one as diverse and established as Arlington, Virginia. The goal of the project is to examine neighborhoods in Arlington, find venues in those neighborhoods such as restaurants, parks, or bars that new residents may want near where they live, and group the neighborhoods into clusters based on their unique venues. An unsupervised machine-learning algorithm creates the clusters of neighborhoods based on the local venue information. The analysis will be done by utilizing the Foursquare API and location data of each neighborhood. In addition to local venue data, the average rent cost will be examined by clusters, so that people researching moving to Arlington can not only make a decision on which neighborhood to move to based on venue data, but also average cost of living.

Description of Data

To complete this project, the following data is used:

1) List of Neighborhoods in Arlington, Virginia

2) Geo-coordinates of the neighborhoods in Arlington, Virginia

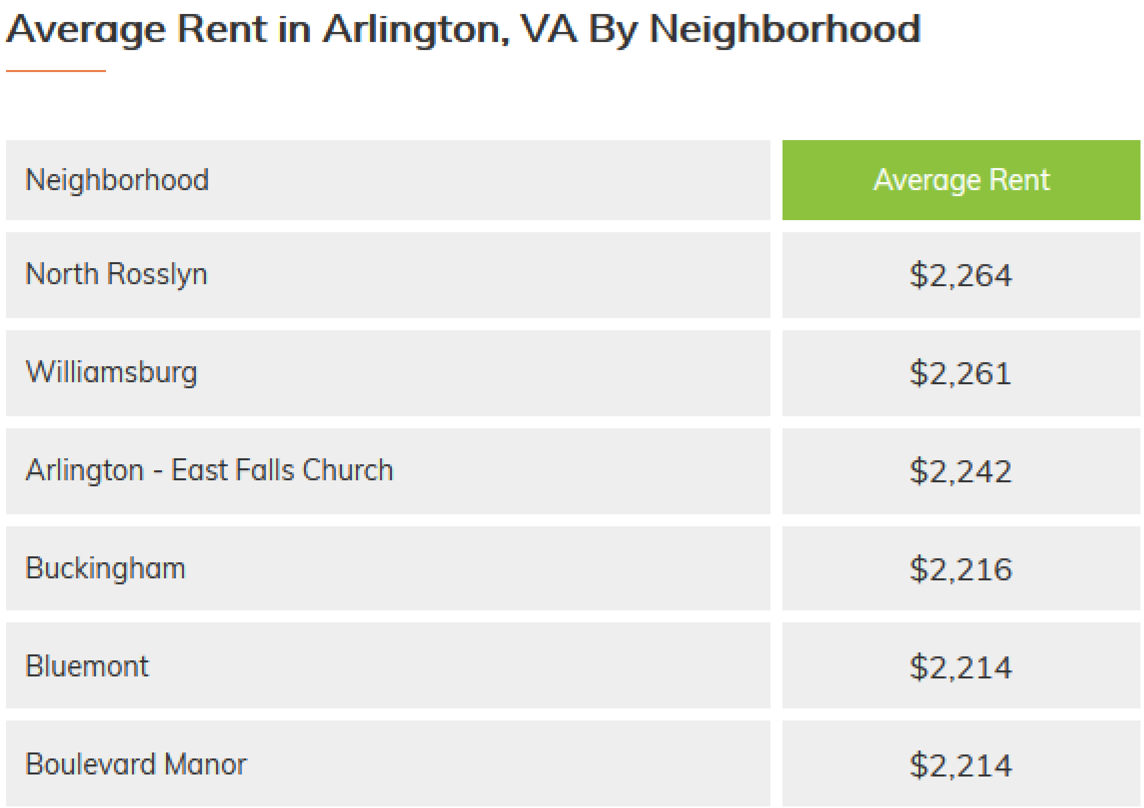
3) Average rent for each neighborhood in Arlington, Virginia

4) The top venues of each neighborhood

The list of neighborhoods and their average rent is obtained from RentCafe, a website where local rent market trends can be determined sorted by neighborhoods (https://www.rentcafe.com/average-rent-market-trends/us/va/arlington/).

Geolocation data of the neighborhoods is obtained from the geocoder tool.

The venue data will be obtained from Foursquare using their API, searching by geolocation.



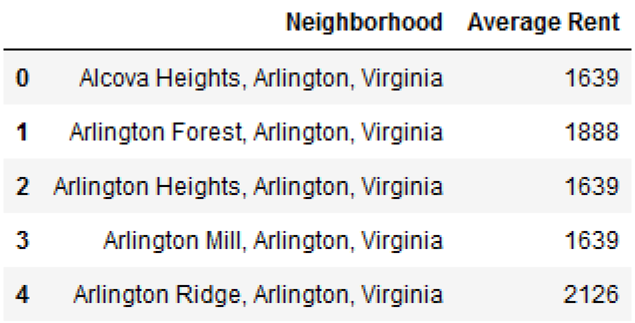
Average Rent in Arlington, VA By Neighborhood

Neighborhood	Average Rent
North Rosslyn	$2,264
Williamsburg	$2,261
Arlington - East Falls Church	$2,242
Buckingham	$2,216
Bluemont	$2,214
Boulevard Manor	$2,214

Figure 1: Example of the data that will be extracted from the RentCafe website

Methodology

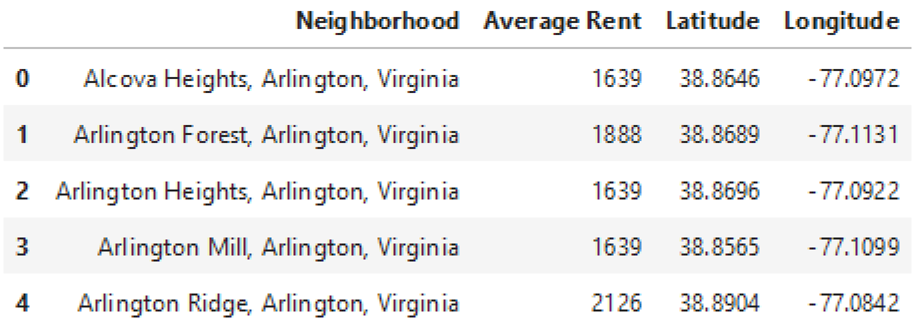
To start, the data from the RentCafe website was scraped and passed into a pandas dataframe using the BeautifulSoup Python package. The data had to be cleaned and formatted correctly into to later pass into the geopy package. For example, in Figure 1, the neighborhood “Arlington – East Falls Church” would be renamed to just “East Falls Church.” In addition, “, Arlington, Virginia” was appended to each of the neighborhoods to help the geopy package find the correct neighborhood latitudes and longitudes. The Average Rent columns were formatted as well, to drop the “$” and the comma so that the numbers could be made into floats for math operations later. Figure 2 represents an example of the formatted dataframe.



	Neighborhood	Average Rent
0	Alcova Heights, Arlington, Virginia	1639
1	Arlington Forest, Arlington, Virginia	1888
2	Arlington Heights, Arlington, Virginia	1639
3	Arlington Mill, Arlington, Virginia	1639
4	Arlington Ridge, Arlington, Virginia	2126

Figure 2: The cleaned and reformatted dataframe

In order to find the latitudes and longitudes of each of the neighborhoods, the Nominatim function from the conda-forge geopy package was used. The Neighborhood column was passed through the Nominatim function to find the latitudes and longitudes of the neighborhood data for Arlington, Virginia. The lat/longs were then appended to the existing dataframe, so that each row in the dataframe consists of the neighborhood, the average rent, the latitude, and longitude as shown in Figure 3.



	Neighborhood	Average Rent	Latitude	Longitude
0	Alcova Heights, Arlington, Virginia	1639	38.8646	-77.0972
1	Arlington Forest, Arlington, Virginia	1888	38.8689	-77.1131
2	Arlington Heights, Arlington, Virginia	1639	38.8696	-77.0922
3	Arlington Mill, Arlington, Virginia	1639	38.8565	-77.1099
4	Arlington Ridge, Arlington, Virginia	2126	38.8904	-77.0842

Figure 3: Neighborhood and Rent dataframe with the locations of each neighborhood appeneded.

The geopy package was unable to find certain neighborhoods. If this was the case, the rows with no location data were dropped from the dataframe. Then, the average rent data was made into a “float” type.

With the folium Python package, the 60 neighborhoods were plotted onto a map, as shown in Figure 4.

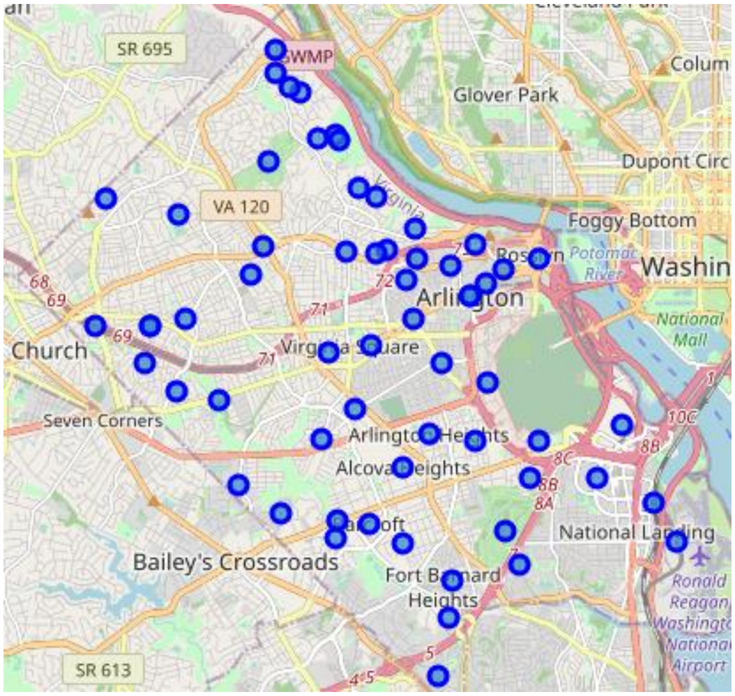
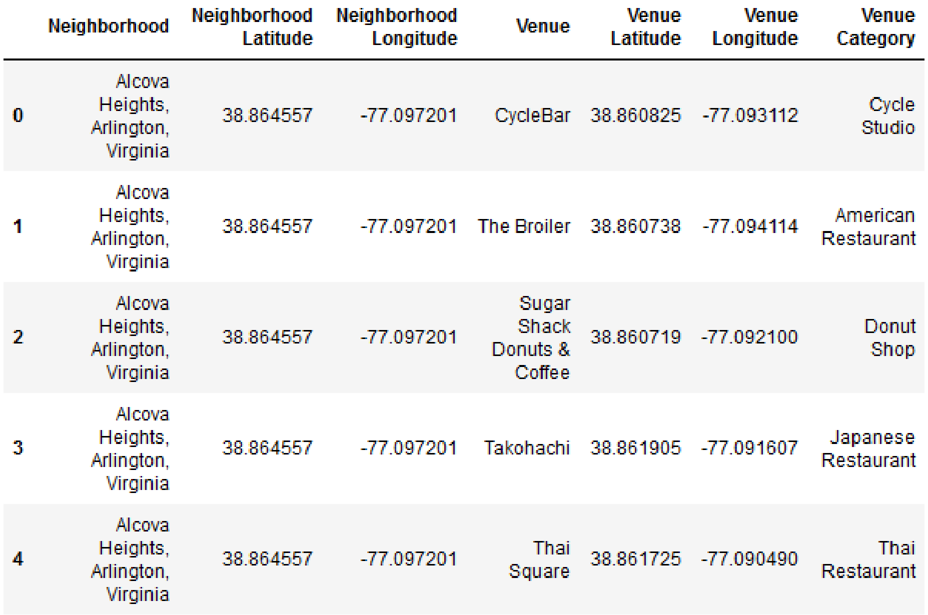


Figure 4: Map of neighborhoods in Arlington, Virginia

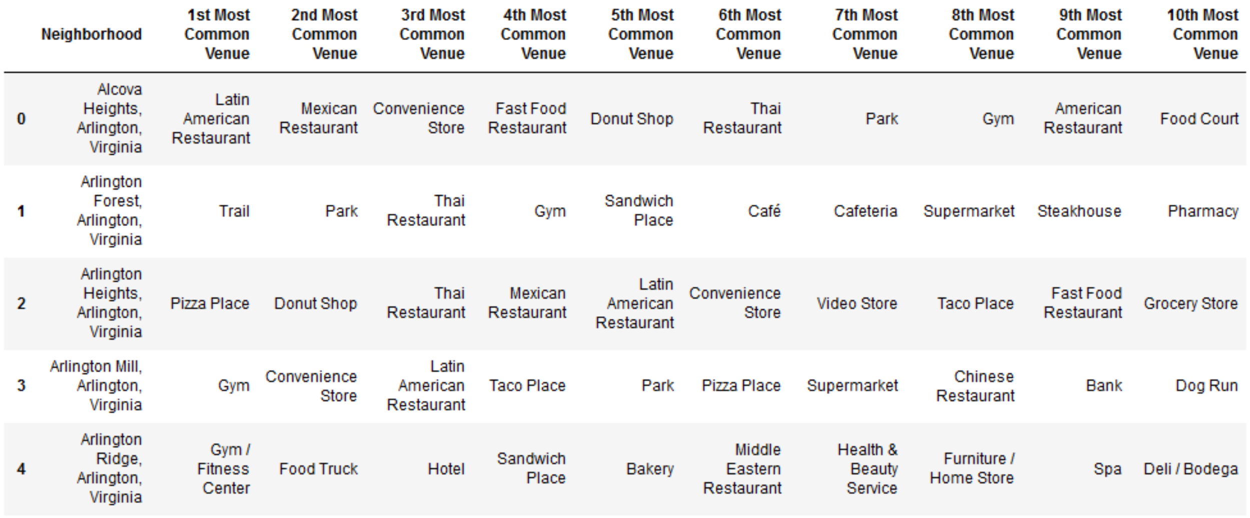
Next, the Foursquare API was used to find all the venues within a 1000 meter radius of each neighborhood point. 1000 meters was chosen because 1) some of the neighborhoods were close together and 2) the distance was small enough to be walking distance of each point, but large enough to get a lot of venues. After running the search, over 3,000 venues were discovered in Arlington that were within the set distance from the neighborhoods. The new data was passed into a dataframe (Figure 5) and then grouped by neighborhood names. This grouping resulted in 238 unique venue categories.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alcova Heights, Arlington, Virginia	38.864557	-77.097201	CycleBar	38.860825	-77.093112	Cycle Studio
1	Alcova Heights, Arlington, Virginia	38.864557	-77.097201	The Broiler	38.860738	-77.094114	American Restaurant
2	Alcova Heights, Arlington, Virginia	38.864557	-77.097201	Sugar Shack Donuts & Coffee	38.860719	-77.092100	Donut Shop
3	Alcova Heights, Arlington, Virginia	38.864557	-77.097201	Takohachi	38.861905	-77.091607	Japanese Restaurant
4	Alcova Heights, Arlington, Virginia	38.864557	-77.097201	Thai Square	38.861725	-77.090490	Thai Restaurant

Figure 5: Example of data found from the Foursquare API searching on the neighborhoods

In order to let a machine learning algorithm cluster the venues, the venue location was one-hot encoded and then grouped by means of occurrence for each neighborhood. From there, the top 10 most commonly occurring venue type was found for each neighborhood and put into a new dataframe (Figure 6).



	Neighborhood	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	Alcova Heights, Arlington, Virginia	Latin American Restaurant	Mexican Restaurant	Convenience Store	Fast Food Restaurant	Donut Shop	Thai Restaurant	Park	Gym	American Restaurant	Food Court
1	Arlington Forest, Arlington, Virginia	Trail	Park	Thai Restaurant	Gym	Sandwich Place	Café	Cafeteria	Supermarket	Steakhouse	Pharmacy
2	Arlington Heights, Arlington, Virginia	Pizza Place	Donut Shop	Thai Restaurant	Mexican Restaurant	Latin American Restaurant	Convenience Store	Video Store	Taco Place	Fast Food Restaurant	Grocery Store
3	Arlington Mill, Arlington, Virginia	Gym	Convenience Store	Latin American Restaurant	Taco Place	Park	Pizza Place	Supermarket	Chinese Restaurant	Bank	Dog Run
4	Arlington Ridge, Arlington, Virginia	Gym / Fitness Center	Food Truck	Hotel	Sandwich Place	Bakery	Middle Eastern Restaurant	Health & Beauty Service	Furniture / Home Store	Spa	Deli / Bodega

Figure 6: Example from the Dataframe of top 10 most commonly occuring venues in each neighborhood

With the data from the one-hot coding, a K-means clustering algorithm could be run to identify the groups based on frequency of venue types in the neighborhood. The average silhouette method was used in order to compute the optimal number of clusters. The silhouette method predicted that 5 was the optimal number of clusters, however, after analysis, 6 clusters was used because if only 5 were used, there was one large cluster, and four clusters of very few neighborhoods (fewer than 4 per cluster), resulting in an uninteresting conclusion. For this reason, 6 clusters were used.

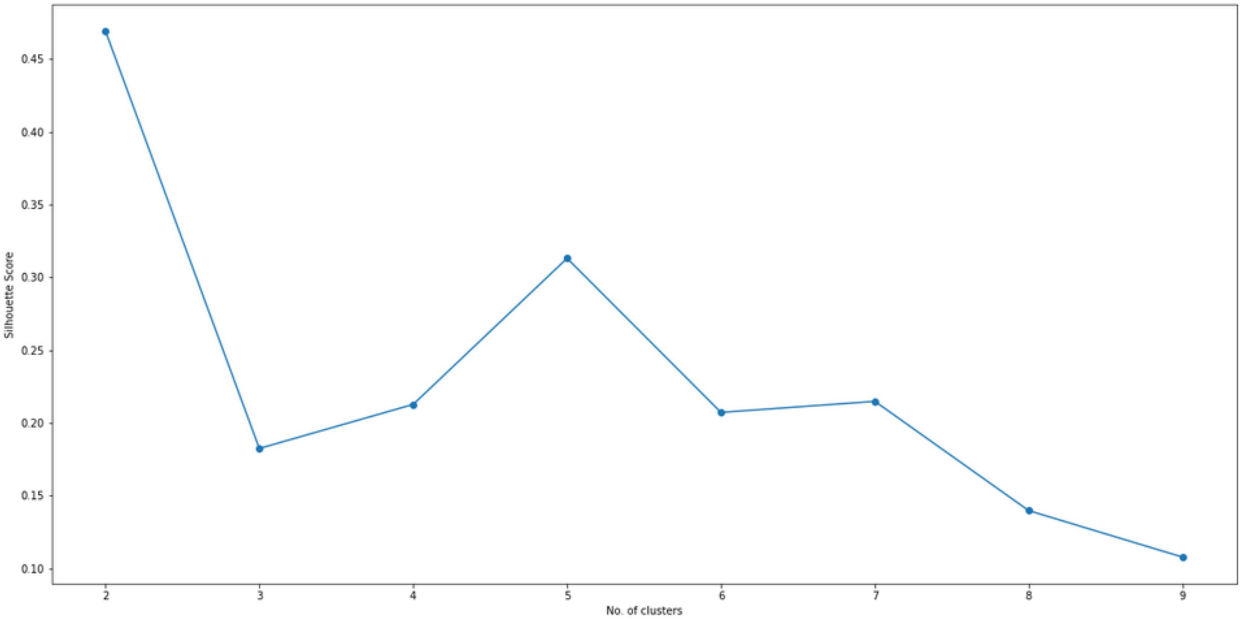
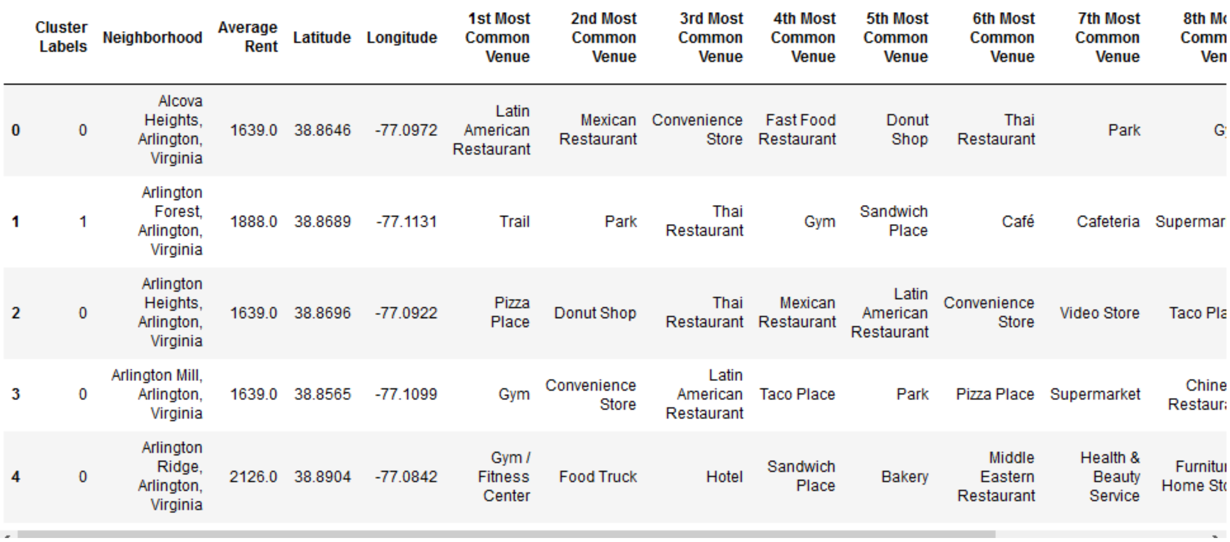


Figure 7: Results of the silhouette method to predict the optimal number of clusters. 6 was used for this analysis.

Results

Finally, the neighborhood cluster label, top 10 most common venue data, and neighborhood average rent/location dataframes were merged into one dataframe (Figure 8).



	Cluster Labels	Neighborhood	Average Rent	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mo Comm Ven
0	0	Alcova Heights, Arlington, Virginia	1639.0	38.8646	-77.0972	Latin American Restaurant	Mexican Restaurant	Convenience Store	Fast Food Restaurant	Donut Shop	Thai Restaurant	Park	G
1	1	Arlington Forest, Arlington, Virginia	1888.0	38.8689	-77.1131	Trail	Park	Thai Restaurant	Gym	Sandwich Place	Café	Cafeteria	Supermar
2	0	Arlington Heights, Arlington, Virginia	1639.0	38.8696	-77.0922	Pizza Place	Donut Shop	Thai Restaurant	Mexican Restaurant	Latin American Restaurant	Convenience Store	Video Store	Taco Pla
3	0	Arlington Mill, Arlington, Virginia	1639.0	38.8565	-77.1099	Gym	Convenience Store	Latin American Restaurant	Taco Place	Park	Pizza Place	Supermarket	Chine Restaur
4	0	Arlington Ridge, Arlington, Virginia	2126.0	38.8904	-77.0842	Gym / Fitness Center	Food Truck	Hotel	Sandwich Place	Bakery	Middle Eastern Restaurant	Health & Beauty Service	Furnitu Home Sto

Figure 8: Merged dataframes with cluster label, neighborhood, average rent, location, and top 10 most common venues

The neighborhoods were plotted again, but this time colored by neighborhood cluster label.

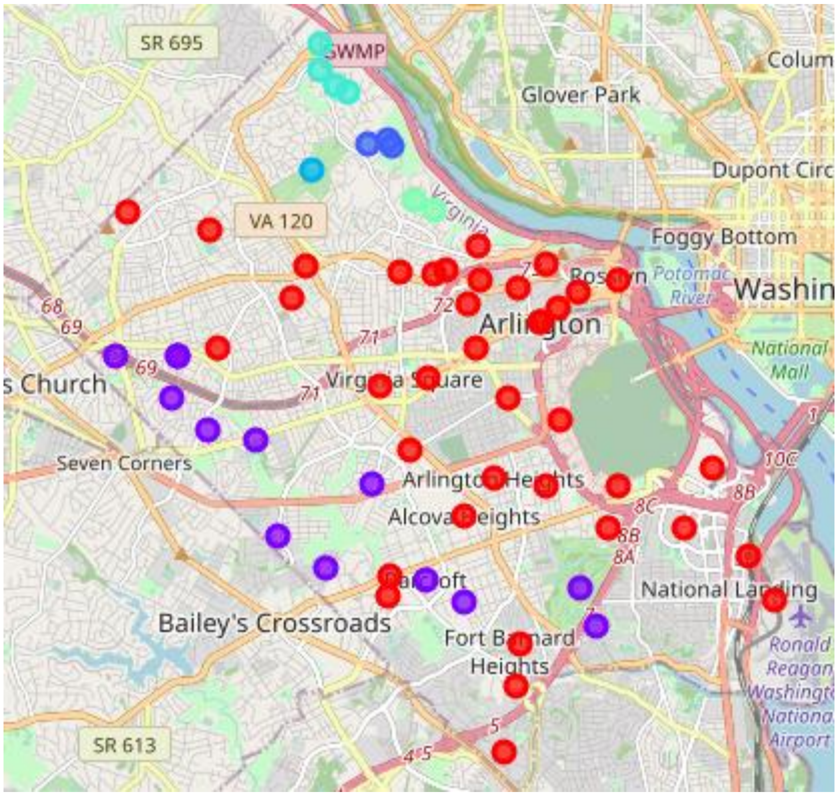
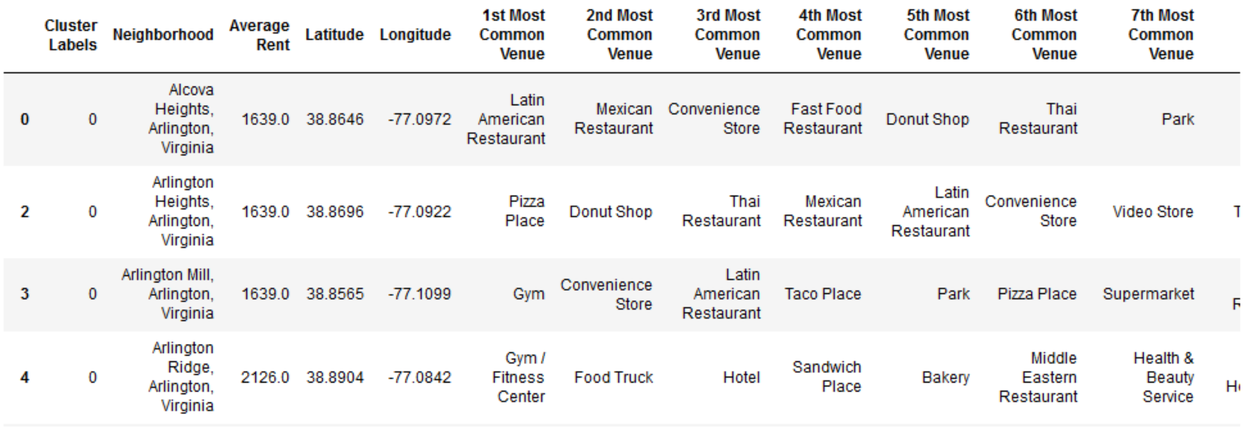


Figure 9: Neighorhoods, colored by cluster label

From the merged dataframe, the average rent was also calculated for each cluster label, to show how rent changed cluster to cluster.

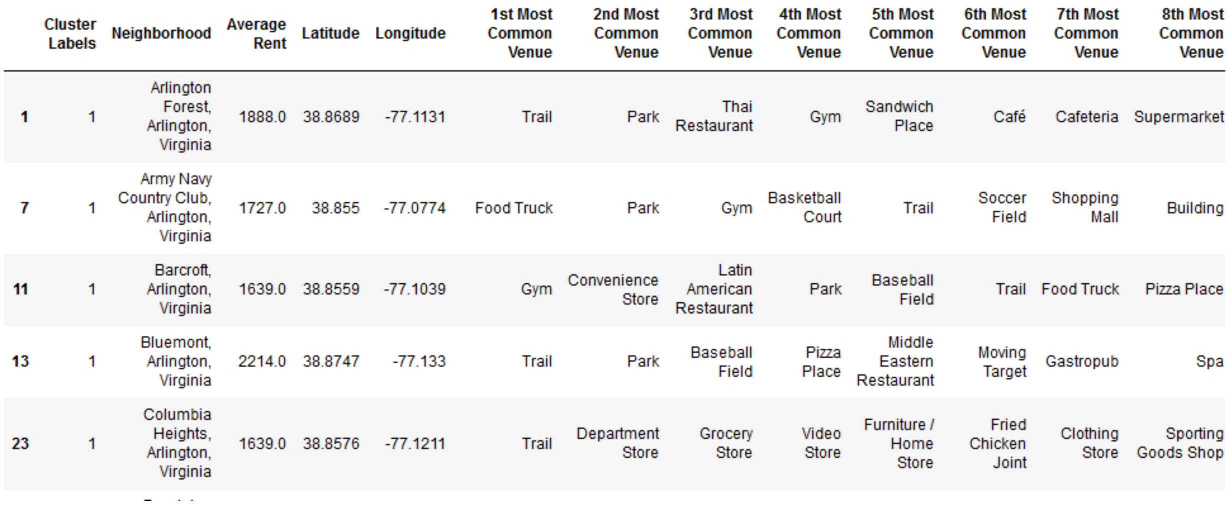
Cluster Zero is the largest of the clusters, and has an average rent of $2,012 per month. This cluster has a wide diversity of nearby venues including parks, theaters, gyms, and restaurants. If someone were moving to Arlington and wanted to make sure that the lived within walking distance of entertainment or amenities, they should start their search in neighborhoods from cluster zero.



	Cluster Labels	Neighborhood	Average Rent	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	
0	0	Alcova Heights, Arlington, Virginia	1639.0	38.8646	-77.0972	Latin American Restaurant	Mexican Restaurant	Convenience Store	Fast Food Restaurant	Donut Shop	Thai Restaurant	Park	
2	0	Arlington Heights, Arlington, Virginia	1639.0	38.8696	-77.0922	Pizza Place	Donut Shop	Thai Restaurant	Mexican Restaurant	Latin American Restaurant	Convenience Store	Video Store	T
3	0	Arlington Mill, Arlington, Virginia	1639.0	38.8565	-77.1099	Gym	Convenience Store	Latin American Restaurant	Taco Place	Park	Pizza Place	Supermarket	R
4	0	Arlington Ridge, Arlington, Virginia	2126.0	38.8904	-77.0842	Gym / Fitness Center	Food Truck	Hotel	Sandwich Place	Bakery	Middle Eastern Restaurant	Health & Beauty Service	H

Figure 10: Example of selections from Cluster Zero

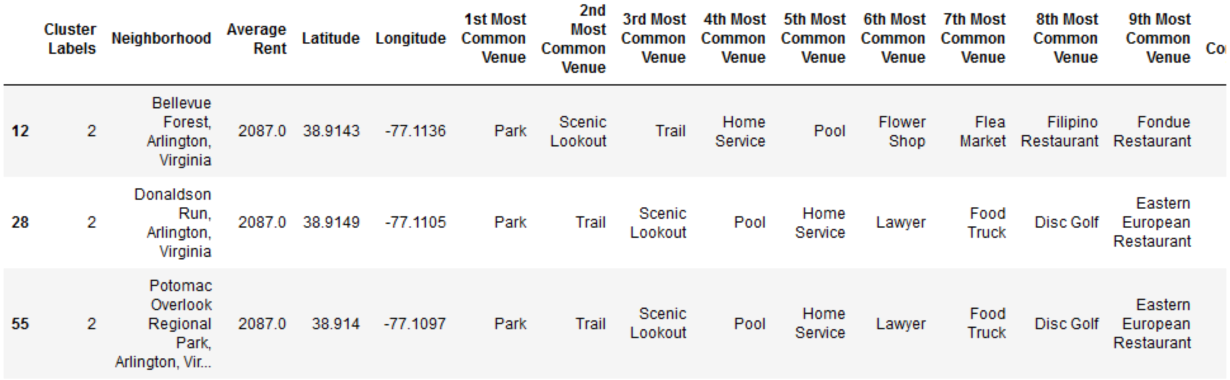
Cluster one is the second largest, and has an average rent of $1,848 per month. This cluster has some diversity, but a large number of the top nearby venues are parks and trails, but with some options of other amenities like stores and restaurants. If someone wanted to live in Arlington at a slightly cheaper rate and liked being near outdoor venues, cluster one neighborhoods would be the best option.



	Cluster Labels	Neighborhood	Average Rent	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
1	1	Arlington Forest, Arlington, Virginia	1888.0	38.8689	-77.1131	Trail	Park	Thai Restaurant	Gym	Sandwich Place	Café	Cafeteria	Supermarket
7	1	Army Navy Country Club, Arlington, Virginia	1727.0	38.855	-77.0774	Food Truck	Park	Gym	Basketball Court	Trail	Soccer Field	Shopping Mall	Building
11	1	Barcroft, Arlington, Virginia	1639.0	38.8559	-77.1039	Gym	Convenience Store	Latin American Restaurant	Park	Baseball Field	Trail	Food Truck	Pizza Place
13	1	Bluemont, Arlington, Virginia	2214.0	38.8747	-77.133	Trail	Park	Baseball Field	Pizza Place	Middle Eastern Restaurant	Moving Target	Gastropub	Spa
23	1	Columbia Heights, Arlington, Virginia	1639.0	38.8576	-77.1211	Trail	Department Store	Grocery Store	Video Store	Furniture / Home Store	Fried Chicken Joint	Clothing Store	Sporting Goods Shop

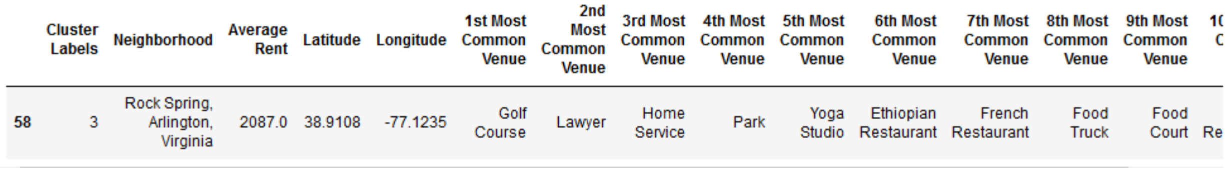
Figure 11: Example of selections from Cluster One

Cluster two, three, four, and five are all very small (fewer than four venues per cluster) and all actually have the same average rent of $2,087 per month. As small clusters, they don't have too much diversity, but the most common nearby venue for most neighborhoods in these clusters is parks. If someone didn't like cluster Zero or One neighborhoods, they could look at these neighborhoods.



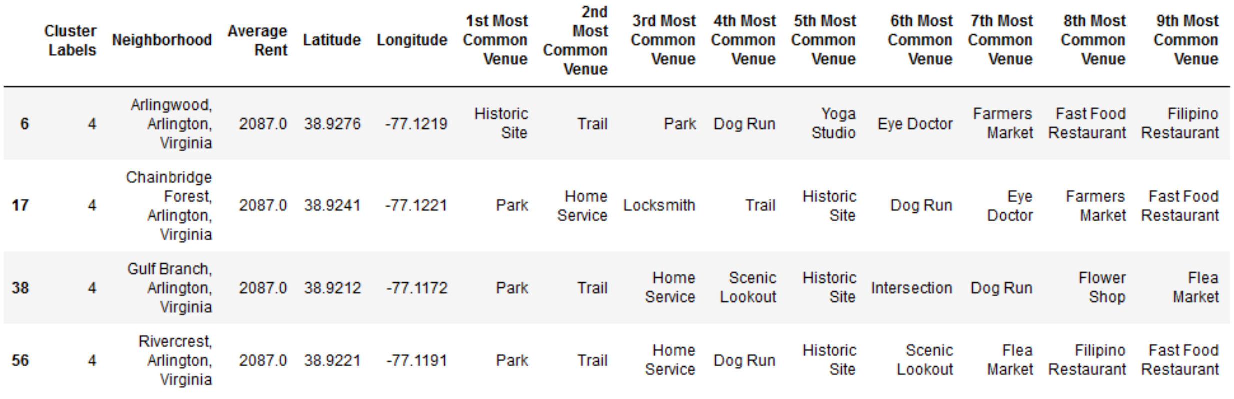
	Cluster Labels	Neighborhood	Average Rent	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	Co
12	2	Bellevue Forest, Arlington, Virginia	2087.0	38.9143	-77.1136	Park	Scenic Lookout	Trail	Home Service	Pool	Flower Shop	Flea Market	Filipino Restaurant	Fondue Restaurant	
28	2	Donaldson Run, Arlington, Virginia	2087.0	38.9149	-77.1105	Park	Trail	Scenic Lookout	Pool	Home Service	Lawyer	Food Truck	Disc Golf	Eastern European Restaurant	
55	2	Potomac Overlook Regional Park, Arlington, Vir...	2087.0	38.914	-77.1097	Park	Trail	Scenic Lookout	Pool	Home Service	Lawyer	Food Truck	Disc Golf	Eastern European Restaurant	

Figure 12: Neighborhoods in Cluster Two



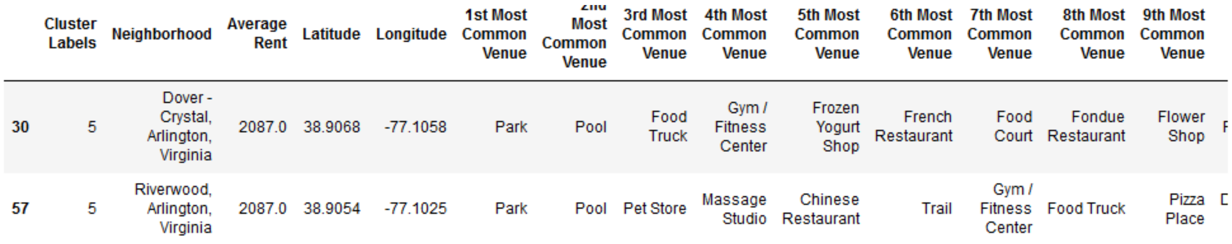
	Cluster Labels	Neighborhood	Average Rent	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	1(C
58	3	Rock Spring, Arlington, Virginia	2087.0	38.9108	-77.1235	Golf Course	Lawyer	Home Service	Park	Yoga Studio	Ethiopian Restaurant	French Restaurant	Food Truck	Food Court	Re

Figure 13: Neighborhoods in Cluster Three



	Cluster Labels	Neighborhood	Average Rent	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
6	4	Arlingwood, Arlington, Virginia	2087.0	38.9276	-77.1219	Historic Site	Trail	Park	Dog Run	Yoga Studio	Eye Doctor	Farmers Market	Fast Food Restaurant	Filipino Restaurant
17	4	Chainbridge Forest, Arlington, Virginia	2087.0	38.9241	-77.1221	Park	Home Service	Locksmith	Trail	Historic Site	Dog Run	Eye Doctor	Farmers Market	Fast Food Restaurant
38	4	Gulf Branch, Arlington, Virginia	2087.0	38.9212	-77.1172	Park	Trail	Home Service	Scenic Lookout	Historic Site	Intersection	Dog Run	Flower Shop	Flea Market
56	4	Rivercrest, Arlington, Virginia	2087.0	38.9221	-77.1191	Park	Trail	Home Service	Dog Run	Historic Site	Scenic Lookout	Flea Market	Filipino Restaurant	Fast Food Restaurant

Figure 14: Neighborhoods in Cluster Four



	Cluster Labels	Neighborhood	Average Rent	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	
30	5	Dover - Crystal, Arlington, Virginia	2087.0	38.9068	-77.1058	Park	Pool	Food Truck	Gym / Fitness Center	Frozen Yogurt Shop	French Restaurant	Food Court	Fondue Restaurant	Flower Shop	F
57	5	Riverwood, Arlington, Virginia	2087.0	38.9054	-77.1025	Park	Pool	Pet Store	Massage Studio	Chinese Restaurant	Trail	Gym / Fitness Center	Food Truck	Pizza Place	D

Figure 15: Neighborhoods in Cluster Five

Discussion

With just a few datasets, some time learning about coding, and basic information on machine learning algorithms, a city as complex as Arlington, Virginia can be categorized and better understood. From this analysis, it would be easy to look at other projects, maybe breaking down the types of restaurants in the each neighborhood, or how the venues in Arlington compare with another major metropolitan area. Based on the results, moving to Cluster Zero or Cluster One would give someone the most options in terms of local amenities and venues. Cluster Zero, also, is closest to the most public transportation and the city center of Arlington. Other analysis could look into mean distance from these transportation options, in case owning a car in this area is not viable. Future analysis could look further clustering neighborhoods with cost of living as an input, or look at the cost of buying a house vs renting, but these were outside the scope of this project.

Conclusion

Where someone decides to live is based on both personal preference and cost of living. Through this analysis, a potential future Arlington resident could get an understanding of neighborhoods without ever having visited the city. The code written for this project achieved the goal of grouping neighborhoods together through a machine learning algorithm and then presenting those results alongside rent data obtain from open source materials.