Coursera
Applied Data Science Capstone
December 14th, 2020

To: Arlington Restaurant Entrepreneurs
From: Kent Sullivan
CC: Coursera Classmates
Subject: **Best Neighborhood to Open a Restaurant – Arlington, VA**

Introduction

Several entrepreneurs and restaurant connoisseurs are looking to open different types of restaurants in Arlington, VA and they want to know which neighborhoods are going to be the best for each type of restaurant. There are a variety of factors that can affect this decision, such as:

- A City's Population (its size, density, and ethnic/cultural/socioeconomic backgrounds)
- Surrounding Competition (the type of restaurants and total number of restaurants.
- Zoning Restrictions
- Property Values
- Local Taxes & Fees

So, using the available location data for Arlington, VA and focusing on the existing competition that is already there, we can help narrow down the scope of neighborhoods that are best suited for our client's restaurants. This initial analysis is an important step in finding a neighborhood, but more importantly will help reduce the risk associated with opening a new restaurant and may even help improve the return on its investment.

Data

To solve this problem, we will need to know what neighborhoods are in Arlington, VA and specifically where they are located (latitude and longitude) as well as information on the venues in/surrounding each neighborhood.

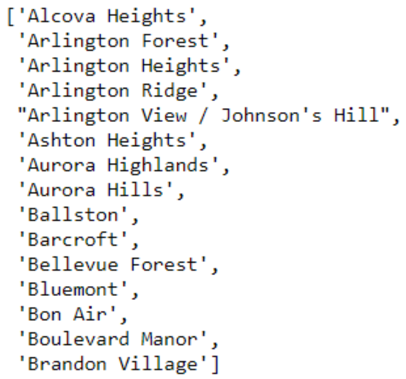
A list of neighborhoods can be found at:
'https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Arlington_County,_Virginia'

To determine each neighborhood's latitude and longitude, Nominatim from the Geopy library, can be used will be used – simply by providing an address for each neighborhood – "Neighborhood Name, Arlington, VA" – Nominatim can return its coordinates.

From here, to determine what restaurants are in each neighborhood, by calling to the Foursquare API, we can request venue information for each neighborhood. This information can include everything from the types of venues at a given location to individual user's reviews for a specific venue.

Methodology

Before any data can be analyzed, it first must be pulled and cleaned. The initial data needed for this project is a list of neighborhood names in Arlington, VA. Using the Beautiful Soup library, this information was scrapped of the web – results can be seen in Figure 1.

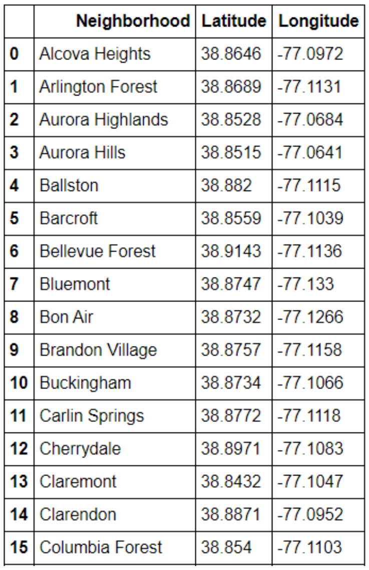


```
['Alcova Heights',
 'Arlington Forest',
 'Arlington Heights',
 'Arlington Ridge',
 "Arlington View / Johnson's Hill",
 'Ashton Heights',
 'Aurora Highlands',
 'Aurora Hills',
 'Ballston',
 'Barcroft',
 'Bellevue Forest',
 'Bluemont',
 'Bon Air',
 'Boulevard Manor',
 'Brandon Village']
```

Figure 1. Arlington, VA neighborhood web scrapped names.

To help improve the quality of the search results from Nominatim, the neighborhoods in Figure 1, are processed and cleaned to remove second names (separated with a "/") or neighborhoods that also have additional information (names followed "()"). Nominatim can be simultaneous called as the neighborhoods are processed to create lists for each of their latitudes and longitudes.

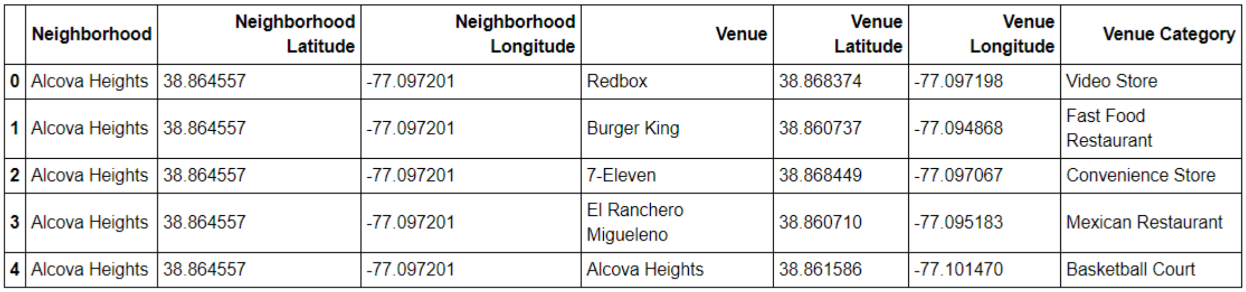
Once the neighborhood names and their latitudes/longitudes are gathered, they can be combined into a pandas dataframe. Unfortunately, Nominatim is not perfect and multiple neighborhoods do not return location coordinates or the coordinates do not make geological sense relative to Arlington. For simplicity these are filtered out. Figure 2 shows a section of the resulting dataframe. In total, there are 73 neighborhoods in Arlington, and after filtering 50 remained.



	Neighborhood	Latitude	Longitude
0	Alcova Heights	38.8646	-77.0972
1	Arlington Forest	38.8689	-77.1131
2	Aurora Highlands	38.8528	-77.0684
3	Aurora Hills	38.8515	-77.0641
4	Ballston	38.882	-77.1115
5	Barcroft	38.8559	-77.1039
6	Bellevue Forest	38.9143	-77.1136
7	Bluemont	38.8747	-77.133
8	Bon Air	38.8732	-77.1266
9	Brandon Village	38.8757	-77.1158
10	Buckingham	38.8734	-77.1066
11	Carlin Springs	38.8772	-77.1118
12	Cherrydale	38.8971	-77.1083
13	Claremont	38.8432	-77.1047
14	Clarendon	38.8871	-77.0952
15	Columbia Forest	38.854	-77.1103

Figure 2. Dataframe of neighborhoods and their accompanying latitudes/longitudes.

Now that we have the necessary neighborhood information, we can call to the Foursquare API to get information on the different venues in and around each neighborhood. Foursquare returns 924 venues that make up these neighborhoods, shown in Figure 3.

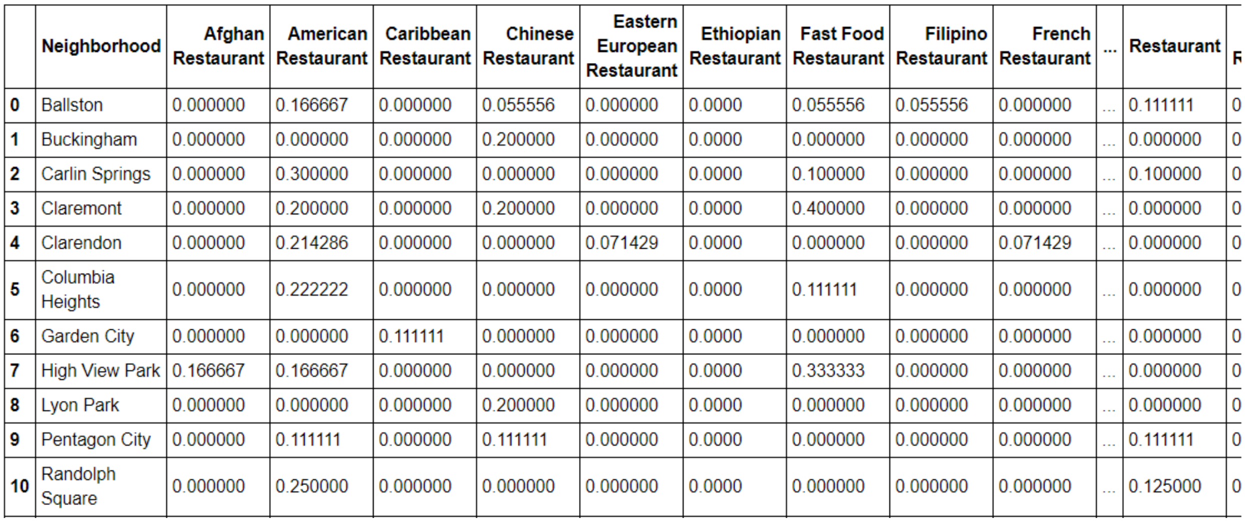


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alcova Heights	38.864557	-77.097201	Redbox	38.868374	-77.097198	Video Store
1	Alcova Heights	38.864557	-77.097201	Burger King	38.860737	-77.094868	Fast Food Restaurant
2	Alcova Heights	38.864557	-77.097201	7-Eleven	38.868449	-77.097067	Convenience Store
3	Alcova Heights	38.864557	-77.097201	El Ranchero Migueleno	38.860710	-77.095183	Mexican Restaurant
4	Alcova Heights	38.864557	-77.097201	Alcova Heights	38.861586	-77.101470	Basketball Court

Figure 3. List of venues in Arlington, VA selected Neighborhoods.

Since we are primarily interested in restaurants, the list of venues is filtered down further, by focusing on “Venue Categories” that contain the word “Restaurant.” Furthermore, the data is filtered down to focus on neighborhoods that have at least 5 restaurants. After each filtering process, we’ve gone from 924 venues to 181/144 restaurants with 34 different restaurant categories that are focused on.

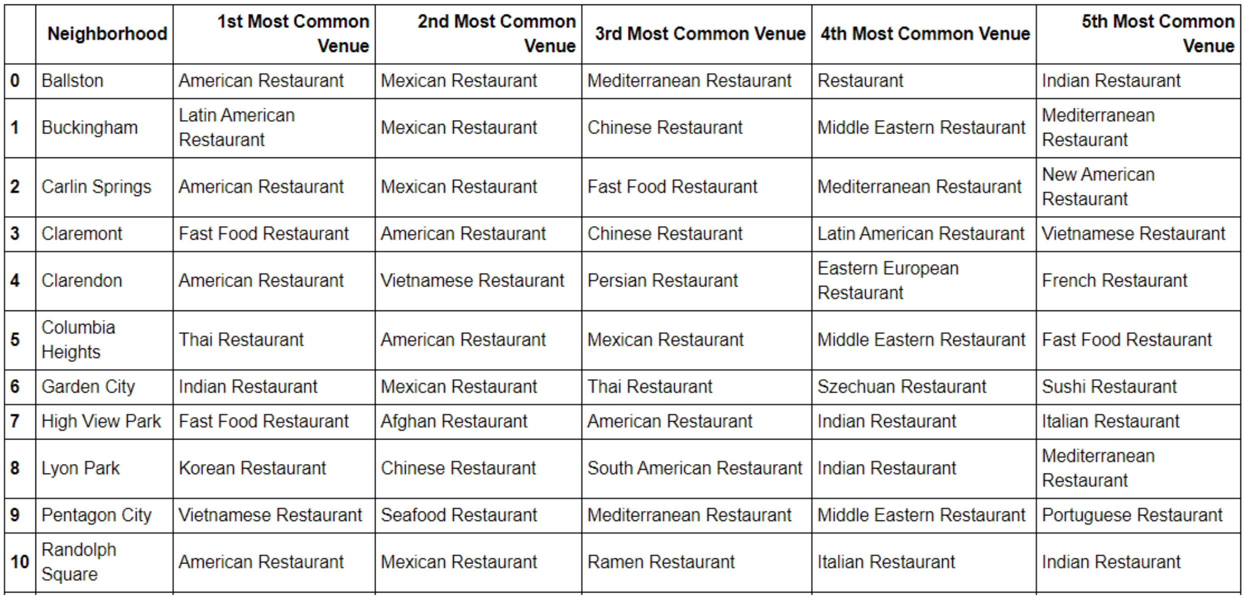
With the restaurant data collected, it is then restructured in preparation for clustering. First using one-hot encoding, we transcribe our categorical venue types into numerical values for easier analysis and then take the mean of our data to see the percentage of restaurant types that make up each neighborhood (Figure 4.)



	Neighborhood	Afghan Restaurant	American Restaurant	Caribbean Restaurant	Chinese Restaurant	Eastern European Restaurant	Ethiopian Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	...	Restaurant	F
0	Ballston	0.000000	0.166667	0.000000	0.055556	0.000000	0.0000	0.055556	0.055556	0.000000	...	0.111111	0
1	Buckingham	0.000000	0.000000	0.000000	0.200000	0.000000	0.0000	0.000000	0.000000	0.000000	...	0.000000	0
2	Carlin Springs	0.000000	0.300000	0.000000	0.000000	0.000000	0.0000	0.100000	0.000000	0.000000	...	0.100000	0
3	Claremont	0.000000	0.200000	0.000000	0.200000	0.000000	0.0000	0.400000	0.000000	0.000000	...	0.000000	0
4	Clarendon	0.000000	0.214286	0.000000	0.000000	0.071429	0.0000	0.000000	0.000000	0.071429	...	0.000000	0
5	Columbia Heights	0.000000	0.222222	0.000000	0.000000	0.000000	0.0000	0.111111	0.000000	0.000000	...	0.000000	0
6	Garden City	0.000000	0.000000	0.111111	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	...	0.000000	0
7	High View Park	0.166667	0.166667	0.000000	0.000000	0.000000	0.0000	0.333333	0.000000	0.000000	...	0.000000	0
8	Lyon Park	0.000000	0.000000	0.000000	0.200000	0.000000	0.0000	0.000000	0.000000	0.000000	...	0.000000	0
9	Pentagon City	0.000000	0.111111	0.000000	0.111111	0.000000	0.0000	0.000000	0.000000	0.000000	...	0.111111	0
10	Randolph Square	0.000000	0.250000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	...	0.125000	0

Figure 4. Percentage of restaurant categories by neighborhood.

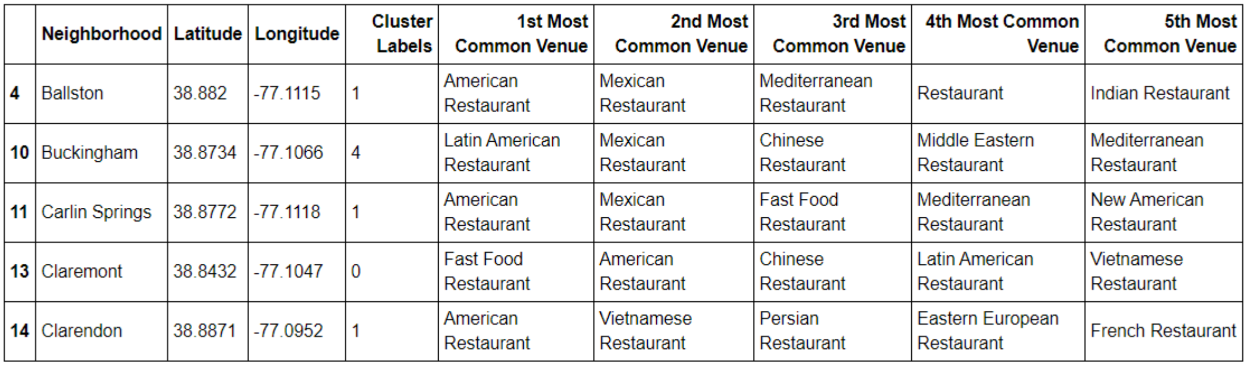
From here, we can pull out the top 5 most common types of restaurants for each neighborhood. The results are shown in Figure 5.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Ballston	American Restaurant	Mexican Restaurant	Mediterranean Restaurant	Restaurant	Indian Restaurant
1	Buckingham	Latin American Restaurant	Mexican Restaurant	Chinese Restaurant	Middle Eastern Restaurant	Mediterranean Restaurant
2	Carlin Springs	American Restaurant	Mexican Restaurant	Fast Food Restaurant	Mediterranean Restaurant	New American Restaurant
3	Claremont	Fast Food Restaurant	American Restaurant	Chinese Restaurant	Latin American Restaurant	Vietnamese Restaurant
4	Clarendon	American Restaurant	Vietnamese Restaurant	Persian Restaurant	Eastern European Restaurant	French Restaurant
5	Columbia Heights	Thai Restaurant	American Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Fast Food Restaurant
6	Garden City	Indian Restaurant	Mexican Restaurant	Thai Restaurant	Szechuan Restaurant	Sushi Restaurant
7	High View Park	Fast Food Restaurant	Afghan Restaurant	American Restaurant	Indian Restaurant	Italian Restaurant
8	Lyon Park	Korean Restaurant	Chinese Restaurant	South American Restaurant	Indian Restaurant	Mediterranean Restaurant
9	Pentagon City	Vietnamese Restaurant	Seafood Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Portuguese Restaurant
10	Randolph Square	American Restaurant	Mexican Restaurant	Ramen Restaurant	Italian Restaurant	Indian Restaurant

Figure 5. Top 5 most common restaurant types by neighborhood.

With all of our data collected, cleaned, and formatted - we are finally ready to utilize machine learning. In this case, we will cluster the data using K-Means clustering. Since there are no predefined categories describing each neighborhood's collection of restaurants, we need to cluster them together to understand which neighborhoods have similar/different dining scenes. Additionally, we are trying to find groups that have not been explicitly labeled, the k-means clustering algorithm is a good choice for our application. After clustering into 5 groups, each cluster label is added to our final dataframe shown in Figure 6.



	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Ballston	38.882	-77.1115	1	American Restaurant	Mexican Restaurant	Mediterranean Restaurant	Restaurant	Indian Restaurant
10	Buckingham	38.8734	-77.1066	4	Latin American Restaurant	Mexican Restaurant	Chinese Restaurant	Middle Eastern Restaurant	Mediterranean Restaurant
11	Carlin Springs	38.8772	-77.1118	1	American Restaurant	Mexican Restaurant	Fast Food Restaurant	Mediterranean Restaurant	New American Restaurant
13	Claremont	38.8432	-77.1047	0	Fast Food Restaurant	American Restaurant	Chinese Restaurant	Latin American Restaurant	Vietnamese Restaurant
14	Clarendon	38.8871	-77.0952	1	American Restaurant	Vietnamese Restaurant	Persian Restaurant	Eastern European Restaurant	French Restaurant

Figure 6. Arlington neighborhood characteristics and cluster labels.

Results & Discussion

Figure 7 shows the geological location of each neighborhood and is color coded by its assigned cluster.

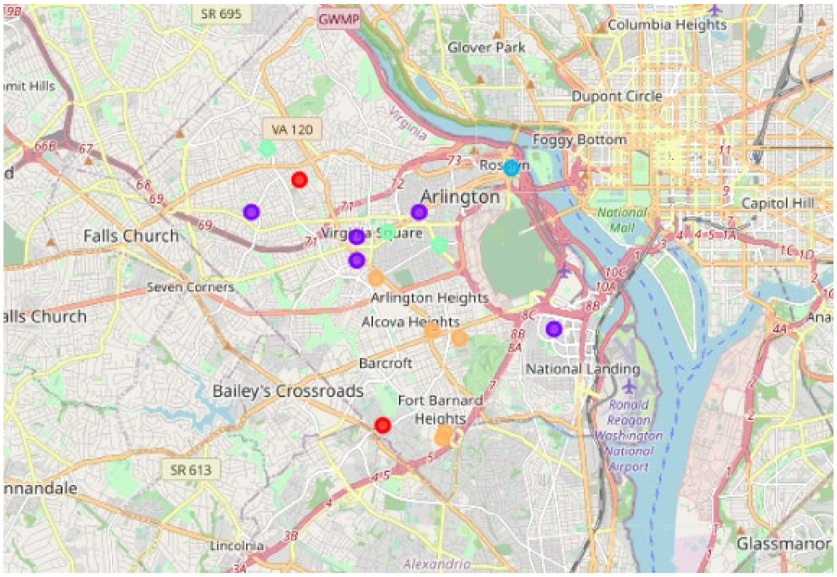
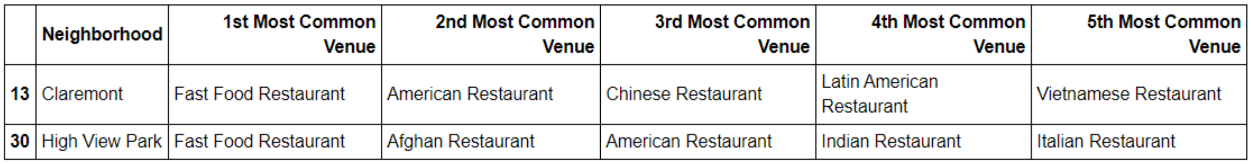


Figure 7. Map of Arlington, VA with Neighborhoods clustered by the type of its most common restaurants.

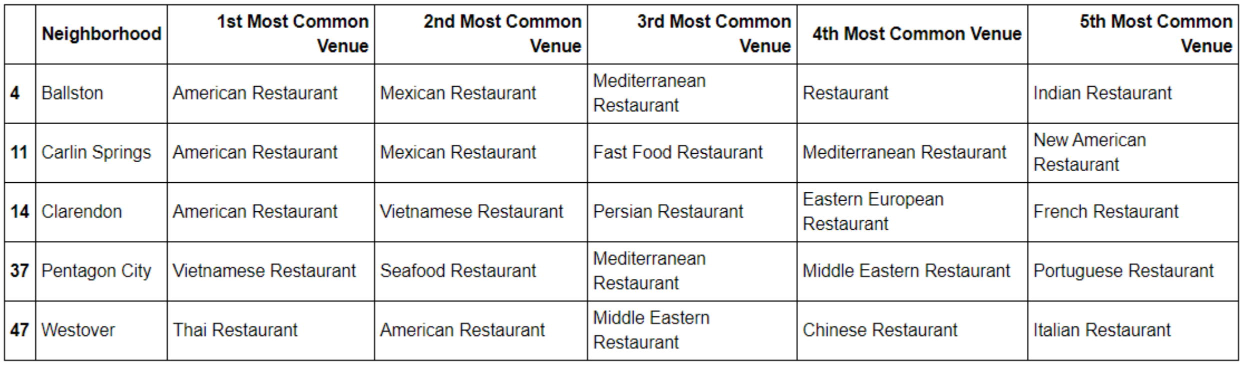
Figures 8 through 12 look at the most common venues in each cluster.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
13	Claremont	Fast Food Restaurant	American Restaurant	Chinese Restaurant	Latin American Restaurant	Vietnamese Restaurant
30	High View Park	Fast Food Restaurant	Afghan Restaurant	American Restaurant	Indian Restaurant	Italian Restaurant

Figure 8. Cluster 0.

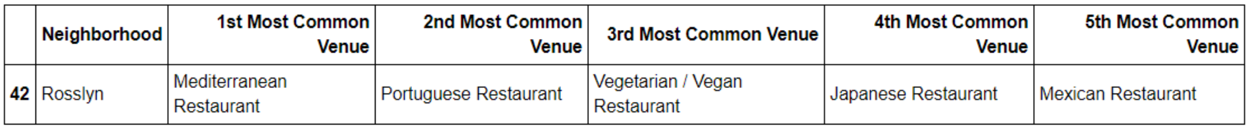
Cluster 0 seems to be based off fast food & American restaurants.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Ballston	American Restaurant	Mexican Restaurant	Mediterranean Restaurant	Restaurant	Indian Restaurant
11	Carlin Springs	American Restaurant	Mexican Restaurant	Fast Food Restaurant	Mediterranean Restaurant	New American Restaurant
14	Clarendon	American Restaurant	Vietnamese Restaurant	Persian Restaurant	Eastern European Restaurant	French Restaurant
37	Pentagon City	Vietnamese Restaurant	Seafood Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Portuguese Restaurant
47	Westover	Thai Restaurant	American Restaurant	Middle Eastern Restaurant	Chinese Restaurant	Italian Restaurant

Figure 9. Cluster 1.

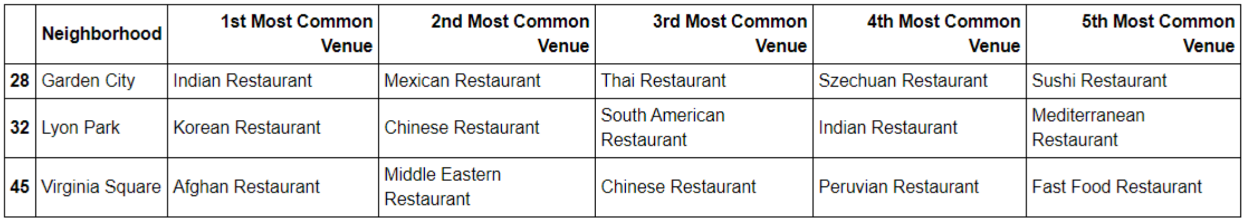
Cluster 1 looks to contain predominately American restaurants, as well as some Mexican restaurants and Mediterranean restaurants.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
42	Rosslyn	Mediterranean Restaurant	Portuguese Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Mexican Restaurant

Figure 10. Cluster 2.

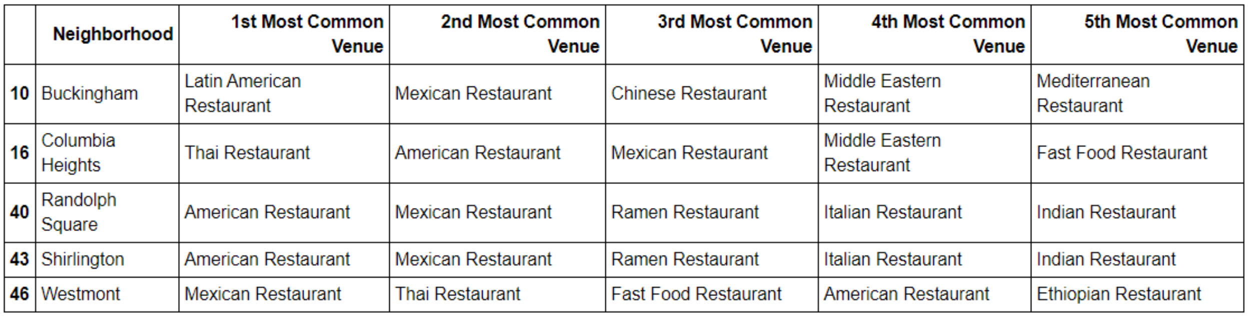
Cluster 2 only contains one neighborhood, Rosslyn, so it must have a unique list of most common venues. This could mean it is a niche neighborhood and hard to get a foot hold in.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
28	Garden City	Indian Restaurant	Mexican Restaurant	Thai Restaurant	Szechuan Restaurant	Sushi Restaurant
32	Lyon Park	Korean Restaurant	Chinese Restaurant	South American Restaurant	Indian Restaurant	Mediterranean Restaurant
45	Virginia Square	Afghan Restaurant	Middle Eastern Restaurant	Chinese Restaurant	Peruvian Restaurant	Fast Food Restaurant

Figure 11. Cluster 3.

Cluster 3 contains some middle eastern and southeast Asian restaurants.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
10	Buckingham	Latin American Restaurant	Mexican Restaurant	Chinese Restaurant	Middle Eastern Restaurant	Mediterranean Restaurant
16	Columbia Heights	Thai Restaurant	American Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Fast Food Restaurant
40	Randolph Square	American Restaurant	Mexican Restaurant	Ramen Restaurant	Italian Restaurant	Indian Restaurant
43	Shirlington	American Restaurant	Mexican Restaurant	Ramen Restaurant	Italian Restaurant	Indian Restaurant
46	Westmont	Mexican Restaurant	Thai Restaurant	Fast Food Restaurant	American Restaurant	Ethiopian Restaurant

Figure 12. Cluster 4.

Cluster 4 seems to focus on American & Mexican restaurants (like cluster 0, but with more emphasis on Mexican restaurants). Additionally, several neighborhoods share Ramen, Middle Eastern, and Indian restaurants.

Conclusion

This clustering analysis gives our stakeholders an idea of feature restaurant types prominent in each neighborhood in Arlington, VA. Ideally, the results will aid in the neighborhood selection process for opening a new restaurant. It will help in finding the right balance between a market oversaturated in one type of restaurant and a competitive one.

Moving forward to improve this analysis, it would be a good idea to include the remaining neighborhoods in Arlington that were left out due to bad or no location data. It would also be interesting to take financial, demographic, and popularity data into consideration as it would offer a more pinpointed and final neighborhood selection.