Coursera Capstone Project

Exploring Venues in Northern Virginia using Foursquare API

Espen Matel

January 2019

Table of Contents

Introduction

Northern Virginia, the suburbs just outside of Washington D.C., contain an estimated 3,159,639 residents. US Census data indicates that the Asian population is roughly 5-6% of the total US population. The Asian population is roughly 10-12% of the total population in Northern Virginia, with some cities (e.g., Annandale) having over 20% of the population identifying as Asian. Given the high percentage of the Asian population, there are many Chinese, Thai, Vietnamese, Korean, etc., restaurants to patronize in the area. Anecdotally, as a resident of the area, I have noticed more and more Korean restaurants opening. While many of these restaurants are concentrated in particular cities, such as Annandale, there seems to be an increase in openings all across Northern Virginia. Using postal code data with coordinates, we can use Foursquare API to create a dataset listing all the Asian restaurants in close proximity to each postal code. Using that data, we can observe the frequency of Asian restaurants in each postal code area, visualize the density of Asian restaurants in each area, and use k-means clustering to partition all postal codes into clusters with similar characteristics.

Business Problem

Given this trend and the popularity of Asian restaurants in the area, the goal of this analysis is to find the best location to open a Korean restaurant in Northern Virginia. Other Asian restaurants, such as Chinese, Thai, Japanese, etc., are also popular in the area and may be competitors. However, these restaurants are not equally distributed across the area geographically, with some specific areas having a disproportionate amount of Korean restaurants. Using data science methodology and the machine learning technique k-means clustering, this analysis aims to answer the question: In Northern Virginia, if an entrepreneur is looking to open a new Korean restaurant, what would be the best location to look into?

Target Audience

This analysis is useful to entrepreneurs wanting to open a new restaurant or to investors who want to be able to make better decisions about where to put their money in the restaurant industry.

Data

To answer the question, we need to acquire the following data:

- Postal codes of Northern Virginia.
    - Virginia postal codes that do not fall into the area considered to be Northern Virginia need to be removed.
- Latitude and Longitude of each postal code.
    - Required to plot the map and get nearby venue data.
- Venue data
    - To be retrieved from Foursquare.
    - Data to be used in k-means clustering

Data Sources

Virginia postal codes along with associated geographical coordinates will be acquired from the following url:

- https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=VA

This website contains postal code, area associated with postal code, and geographical coordinates. Using Python requests and beautifulsoup packages, we can scrape the website for this data and place it into a pandas Dataframe.

Once this data is acquired, we will use Foursquare API to retrieve nearby venue data for each postal code. Foursquare has one of the largest databases for venue data, as such, we will retrieve many results that are not relevant for analysis (i.e., gyms, parks, retail stores). The dataframe will need to be refined to exclude these results and only include results that fall under the umbrella of "Asian Restaurant."

Methodology

In this section, I will describe the data analysis and how I used the data to yield the results.

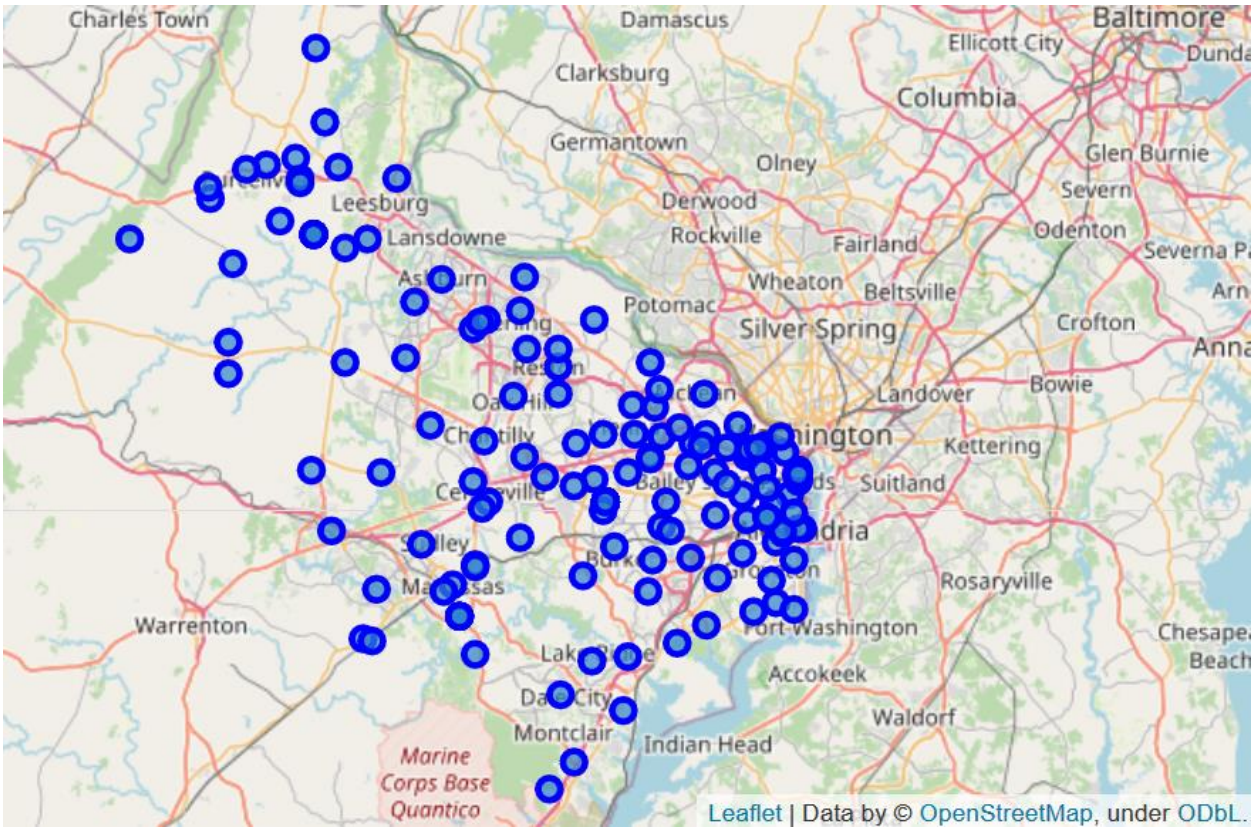
To start, I needed postal codes and geographical data for Northern Virginia. Northern Virginia is not an official area and is generally used in the colloquial sense. However, it is narrowly defined as the counties of Arlington, Fairfax, Loudoun, and Prince William as well as the independent cities of Alexandria, Fairfax, Falls Church, Manassas, and Manassas Park. I retrieved a dataset available on https://public.opendatasoft.com that contained postal codes for Virginia and the corresponding geographical coordinates. After retrieving postal codes, I refined the data, keeping only postal codes that belong to Northern Virginia. I plotted the postal code data on a folium map to confirm that the correct data were used. Once I had the correct postal codes, I passed in the geographical coordinates for each postal code using a for loop to call Foursquare API to retrieve all venues that were in proximity (radius = 7000 meters) of each postal code. A total for 16,968 venues were retrieve from the Foursquare API call. Venues that were not relevant for analysis, such as hotels, retail stores, bus stations, etc., were removed from the data. After cleaning the data, there were a total of 1,557 venues in the dataset.

Dummy variables were assigned to each venue category; for example, if a postal code has a Korean restaurant in proximity, the value for "Korean Restaurant" in the corresponding postal

code is equal to 1. This was done across 178 unique postal codes. Using this data, a new dataframe with the 10 most common venues for each of the 178 postal codes was created. This data were used in the k-means clustering algorithm, which clustered the 178 postal codes into six clusters based on their most common venues. The results will allow us to identify which neighborhoods which neighborhoods already have a higher concentration of Asian restaurants and which have a low concentration. This information, along with the results of the k-means clustering algorithm, will help us answer the question as to which areas are most suitable to opening a new Korean Restaurant.

Results

After retrieving and cleaning postal code data, the data were plotted on to a map of Northern Virginia:

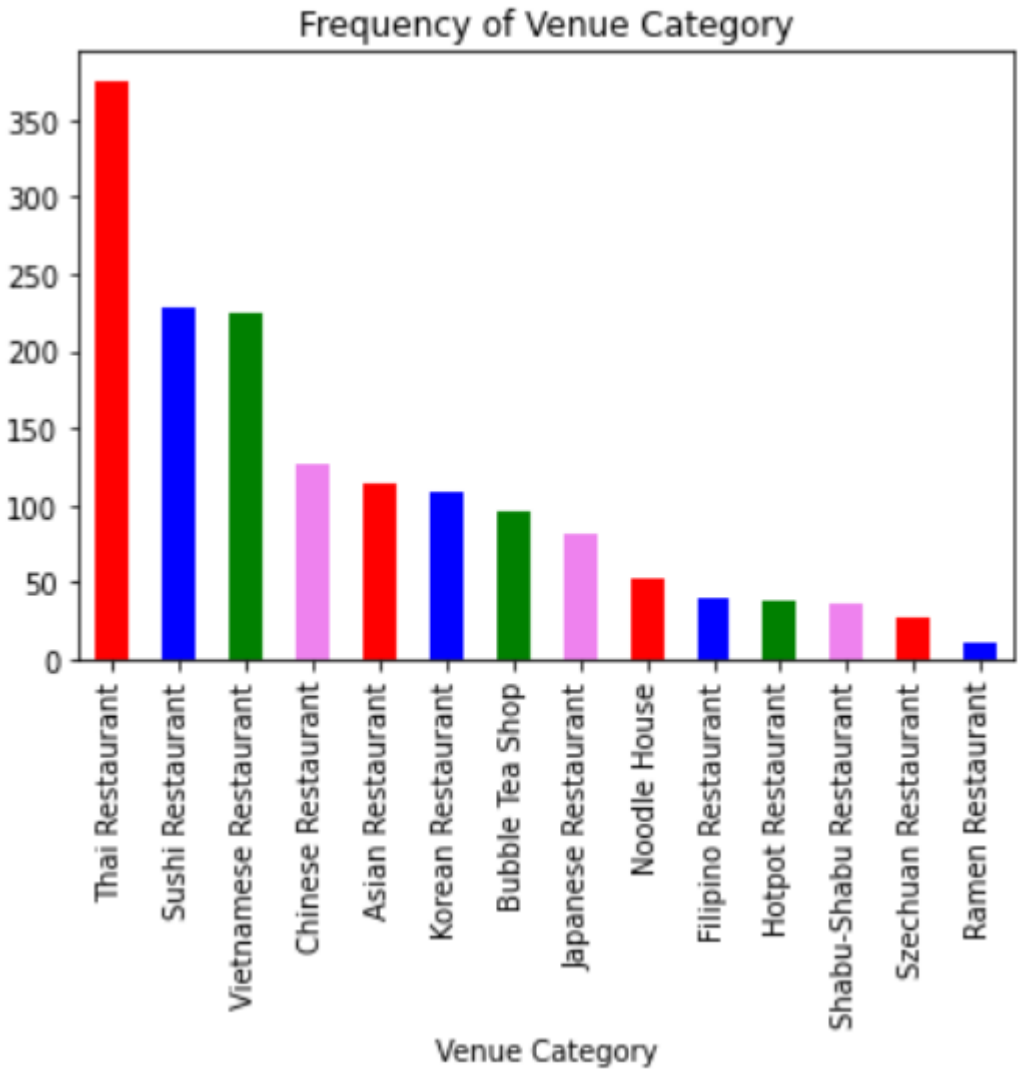


Notice that the areas to the west and north-west are more rural, while the areas to the east, particularly Alexandria, are denser. Some postal code areas in Alexandria only cover a couple blocks, so the venues in proximity may be similar.

After calling the Foursquare API for nearby venues, the results were put into a dataframe with each venue represented in a row of data:

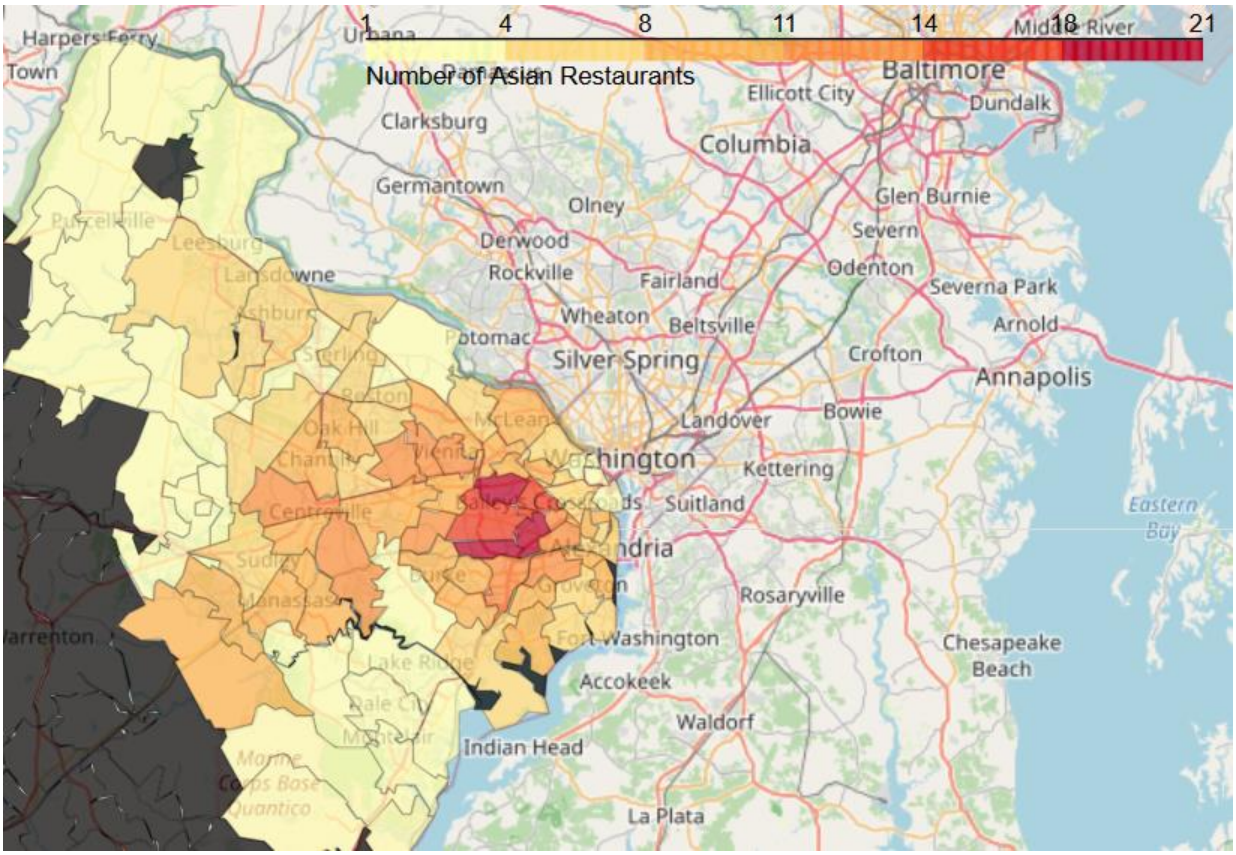
	ZIP	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
9	22034	38.831813	-77.288755	Sushi Prince	38.845545	-77.301178	Sushi Restaurant
19	22034	38.831813	-77.288755	Izakaya Blueocean	38.842322	-77.270775	Sushi Restaurant
21	22034	38.831813	-77.288755	East Wind	38.846177	-77.306011	Vietnamese Restaurant
33	22034	38.831813	-77.288755	99°C Hot Pot	38.844133	-77.291212	Asian Restaurant
40	22034	38.831813	-77.288755	Sisters Thai The Living Room Cafe	38.845737	-77.305500	Thai Restaurant

The column “Venue Category” contains 14 unique categories. Below is a bar chart of the frequency of each category:



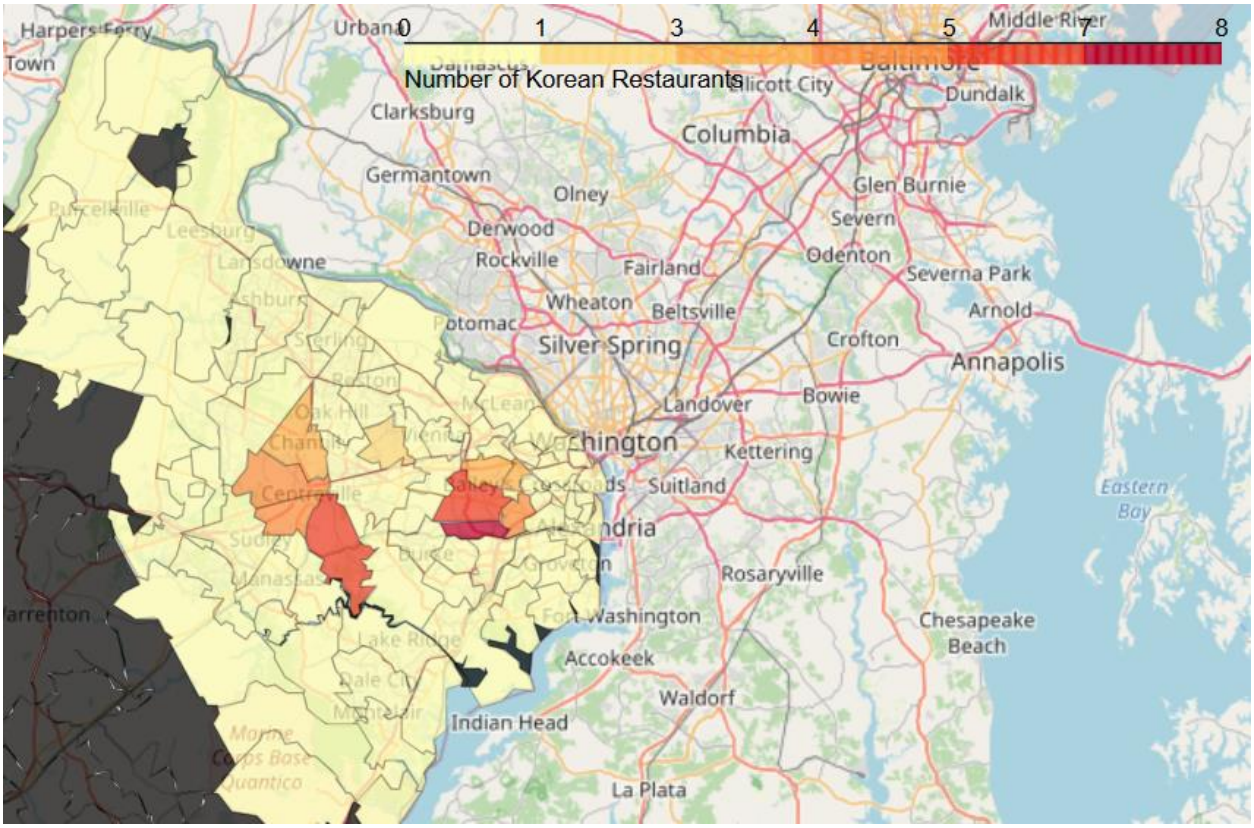
Thai restaurants were the most common with 376 results and Ramen restaurants were the least common with 11 results.

In order to visualize the frequency of Asian restaurants and Korean restaurants by postal code, the data were processed to be plotted as a choropleth map. Below is a choropleth map of the frequency of all Asian restaurants in the Northern Virginia area by postal code.



While there are numerous Asian restaurants spread out over Northern Virginia, a concentration of Asian restaurants in the Centreville and Annandale areas is evident.

The data were further refined to only include venues that were labeled with the category, “Korean Restaurant.” This data were then plotted to another choropleth map:



After refining the data to include only Korean restaurants, we also see a concentration of Korean restaurants in Centreville and Annandale.

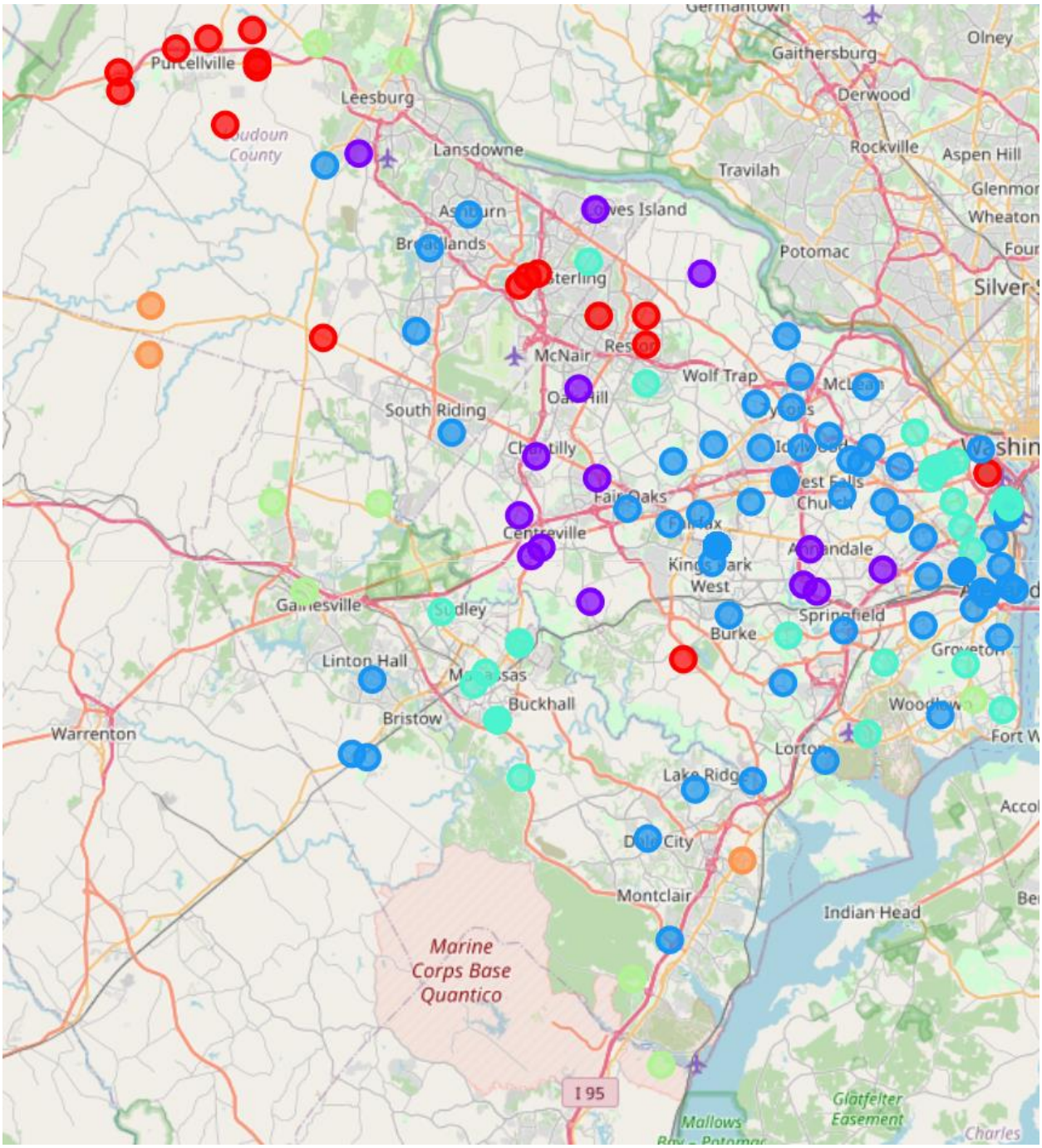
Dummy variables were retrieved for each venue, and put into a new dataframe:

	ZIP	Asian Restaurant	Bubble Tea Shop	Chinese Restaurant	Filipino Restaurant	Hotpot Restaurant	Japanese Restaurant	Korean Restaurant	Noodle House	Ramen Restaurant	S Resta
9	22034	0	0	0	0	0	0	0	0	0	
19	22034	0	0	0	0	0	0	0	0	0	
21	22034	0	0	0	0	0	0	0	0	0	
33	22034	1	0	0	0	0	0	0	0	0	
40	22034	0	0	0	0	0	0	0	0	0	

The data were processed in order to create a dataframe that displays postal codes along with columns "1st Most Common Venue" up to "10th Most Common Venue."

	ZIP	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
0	20101	Thai Restaurant	Asian Restaurant	Vietnamese Restaurant	Szechuan Restaurant	Sushi Restaurant	Shabu-Shabu Restaurant	Ramen Restaurant	Noodle House	Korean Restaurant
1	20103	Thai Restaurant	Sushi Restaurant	Asian Restaurant	Vietnamese Restaurant	Szechuan Restaurant	Shabu-Shabu Restaurant	Ramen Restaurant	Noodle House	Korean Restaurant
2	20105	Thai Restaurant	Japanese Restaurant	Asian Restaurant	Vietnamese Restaurant	Szechuan Restaurant	Sushi Restaurant	Shabu-Shabu Restaurant	Ramen Restaurant	Noodle House
3	20107	Thai Restaurant	Sushi Restaurant	Vietnamese Restaurant	Szechuan Restaurant	Shabu-Shabu Restaurant	Ramen Restaurant	Noodle House	Korean Restaurant	Japanese Restaurant
4	20108	Thai Restaurant	Vietnamese Restaurant	Sushi Restaurant	Noodle House	Filipino Restaurant	Szechuan Restaurant	Shabu-Shabu Restaurant	Ramen Restaurant	Korean Restaurant

Then, a k-means clustering machine learning algorithm was employed to cluster each postal code based on the percentage frequency of each venue category. This frequency was the percentage of total venues within proximity of each postal code. K-means is a type of partitioning clustering that divides the data into non-overlapping subsets (clusters) without any cluster-internal structure. The goal is to have examples within a cluster be very similar, and examples across different clusters very different. Using the k-means clustering algorithm imported from sklearn, the data were clustered into six clusters. These clusters were color-coded and plotted to a map:



For each cluster, I retrieved the value counts of the 1st most common venue for each postal code in each cluster:

Cluster 0 (Red):

Thai Restaurant	14
Vietnamese Restaurant	2
Asian Restaurant	1

Cluster 1 (Purple):

Korean Restaurant	8
Vietnamese Restaurant	3
Japanese Restaurant	1
Thai Restaurant	1
Asian Restaurant	1

Cluster 2 (Blue):

Vietnamese Restaurant	50
Thai Restaurant	21
Sushi Restaurant	15
Chinese Restaurant	3

Cluster 3 (Green):

Thai Restaurant	45
Noodle House	1
Chinese Restaurant	1

Cluster 4 (Yellow):

Japanese Restaurant	6
Chinese Restaurant	2

Cluster 5 (Orange):

Sushi Restaurant	3

Discussion

As noted in the Results section, Annandale and Centreville contain a high amount of Korean restaurants and as such, were both grouped into Cluster 1 (Purple). Cluster 1 can be considered the Korean restaurant cluster, as the most popular venue in this cluster overall are Korean restaurants. While most of the clusters for Cluster 1 can be found around the Annandale and Centreville area, Leesburg, which is in the somewhat distant north-west, was also grouped into Cluster 1. Arlington, an area with a high density of postal codes, was mostly grouped into Cluster 3 (Green). Alexandria, another area with a high density of postal codes, was mostly grouped into Cluster 2 (Blue), along with most of the suburban areas just outside of Washington D.C.. Cluster 0 (Red) is most associated with the slightly more rural and less dense areas in the north-west. Clusters 4 (Yellowish/mint green) and Cluster 5 (Orange) are mostly on the far edges of Northern Virginia. These results could help an entrepreneur or investor select an area to look into for opening a Korean restaurant. One area to look into is Leesburg, 20175, as this postal code was grouped into Cluster 1. Additionally, we saw on the choropleth map that this area does not have a high density of Korean restaurants as opposed to Annandale and Centreville, whom are also in Cluster 1. This suggests that there would not be as much competition from other Korean restaurants. There is also a customer base to serve in the area, as Korean restaurants are the 3rd most popular venue in Leesburg, 20175.

Another area an entrepreneur could look into is Fairfax. This may not appear as promising, as geographically, Fairfax is sandwiched in between Centreville and Annandale, two areas with a high density of Korean restaurants. However, Korean restaurants are in the top five most popular for most Fairfax postal codes. While most of the postal codes within Fairfax were grouped into Cluster 2, one postal code within Fairfax, 22033, was grouped into Cluster 1. The

1st most popular venue for postal code 22033 are Japanese restaurants, with Korean restaurants being the 2nd most popular.

If a risk-seeking entrepreneur wanted to try an area with little competition from other Korean restaurants, I would recommend looking into postal codes within Cluster 3. Cluster 3 has a disproportionate amount of Thai and Chinese restaurants and few Korean restaurants and is also more associated with dense urban areas.

Conclusion

In this project, we have identified a business problem, determined what data were needed, extracted, preprocessed, and cleaned the data, visualized the data, and clustered the data using a k-means machine learning algorithm. Lastly, recommendations were provided to address the initial question in the business problem. These recommendations included Leesburg, Fairfax, and Arlington (Cluster 3).

Limitations

This analysis relies on results from a call to Foursquare API. This carries an assumption that the data from Foursquare are complete, and accurately represent the number of venues in proximity to each postal code. While working on this analysis, I noticed that the results of each call to Foursquare were not entirely consistent, even when doing the same exact call. Another limitation is the variance in geography of Northern Virginia: some postal codes encapsulate a couple blocks in Arlington, while some postal codes encapsulate larger rural areas. As a result, venues retrieved for smaller and dense postal code areas may be similar. Another limitation of Foursquare API is the reliance on "Venue Category." Of the 1,557 venues, 114 are classified generally as "Asian restaurant." These restaurants could either be "fusion" style restaurants that

combine multiple cuisines, or that Foursquare does not have the proper venue category assigned. One would need to look at each venue individually to classify these venues appropriately.

The goal of this analysis is find the best location to open a Korean restaurant, however, other things an entrepreneur would need to consider, such as tax codes by county, regulations by county, rent costs, etc., were not included in this analysis. This analysis should be used as a starting point for an entrepreneur who is researching which areas in Northern Virginia are most optimal for opening a new Korean restaurant.