COURSERA CAPSTONE PROJECT

IBM APPLIED DATA SCIENCE CAPSTONE

Opening a New Korean Restaurant in Northern Virginia, USA

By: Espen Matel

January 2021

INTRODUCTION

- Northern Virginia, the suburbs just outside of Washington D.C., contain an estimated 3,159,639 residents.
- 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 this population distribution, there is a high frequency of Asian restaurants in the area and an increasingly growing customer base

The Northern Virginia Region
Leesburg
Loudoun County
Herndon
Dulles Airport
Vienna
Falls Church
Fairfax City
Arlington County
District of Coumbia
Alexandria
Manassas Park
Manassas
Fairfax County
Prince William County
Dumfries

BUSINESS PROBLEM

- 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, with some specific areas having a disproportionate amount of Korean restaurants, some areas with a disproportionate amount of Thai restaurants, etc.

DATA

- To answer the question, we need to retrieve the following data:
  - Postal Codes of Northern Virginia
    - Northern Virginia 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.
    - Virginia postal codes that are not within this area will be removed
  - Latitude and Longitude of each postal code
    - Required to plot the maps and retrieve nearby venue data from Foursquare
  - Venue Data
    - To be retrieved from Foursquare API
    - 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
  - Using Python requests and beautifulsoup packages, we can scrape the website and place the data into a pandas dataframe
- Virginia geojson file was acquired from the following url:
  - https://github.com/jalbertbowden/open-virginia-gis/raw/master/zip-codes/json/zt51_d00.geojson
- Venue data was retrieved from a call to Foursquare API

METHODOLOGY

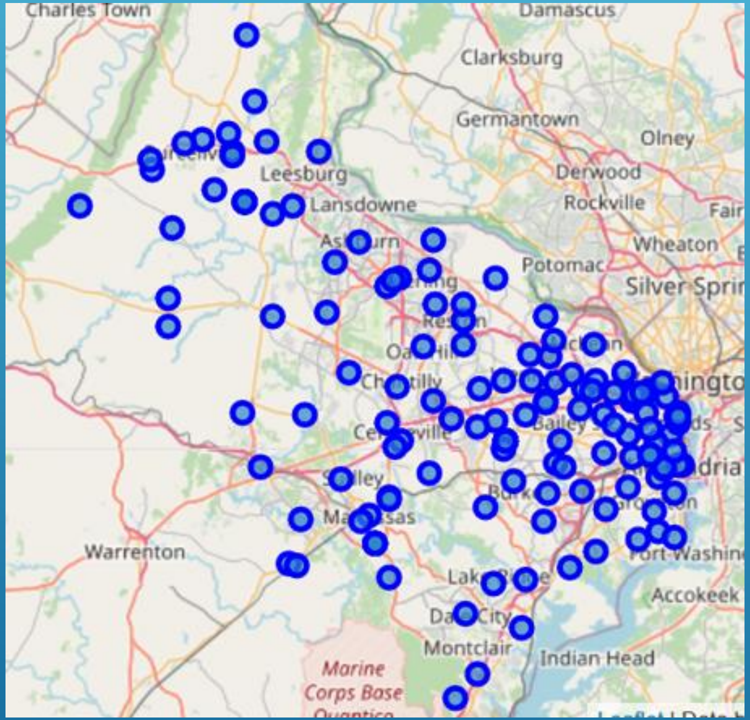
- Scrape postal code data from https://public.opendatasoft.com/
- Refine data, including only postal codes within Northern Virginia
    - 178 unique postal codes
- Use postal code data in a call to Foursquare API, retrieving venue data within proximity of each postal code
    - A total of 16,698 venues were retrieved.
    - Venues that were not relevant for analysis, such as hotels, retail stores, bus stations, etc., were removed from the data
    - A total of 1,557 venues were selected for analysis

METHODOLOGY CONT.

- Calculated frequency counts for each venue by postal code
- Data were visualized using a bar chart and choropleth maps
- Data were then used in a k-means clustering algorithm, which clustered 178 postal codes into six clusters

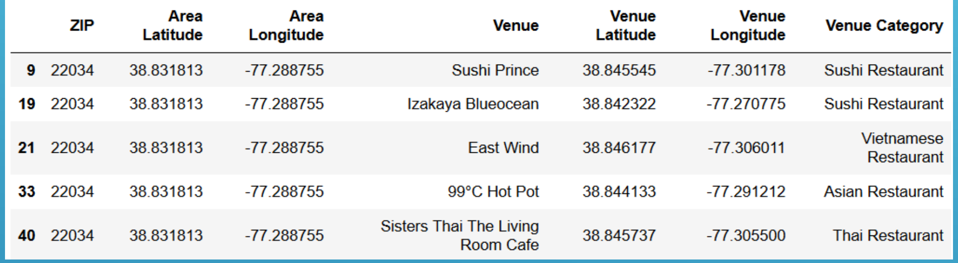
RESULTS

- 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, and the areas to the east, particularly Alexandria, are denser

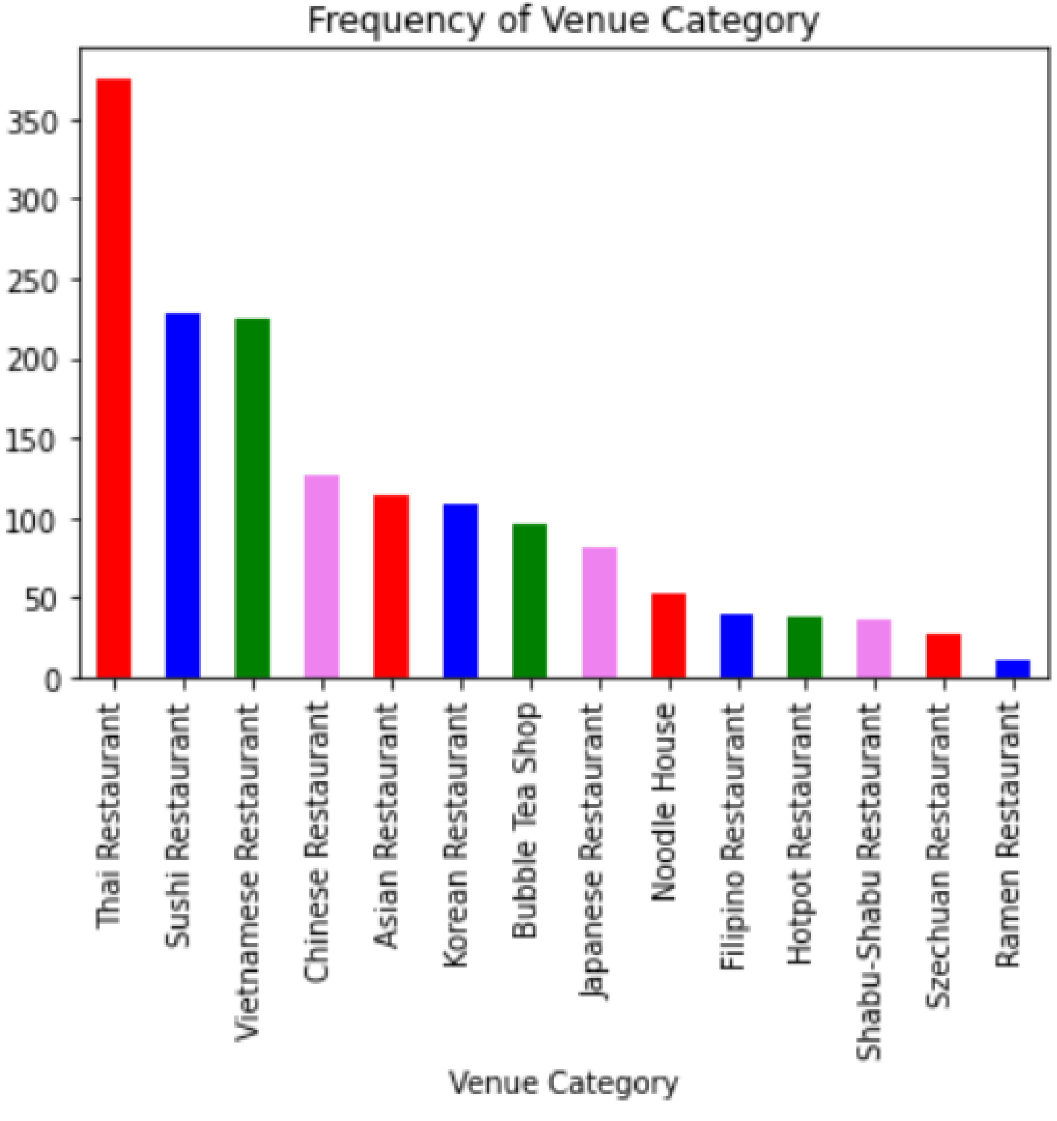
- After calling Foursquare API for nearby venues, the results were put into a dataframe with each venue represented in a row of data:



	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

RESULTS CONT

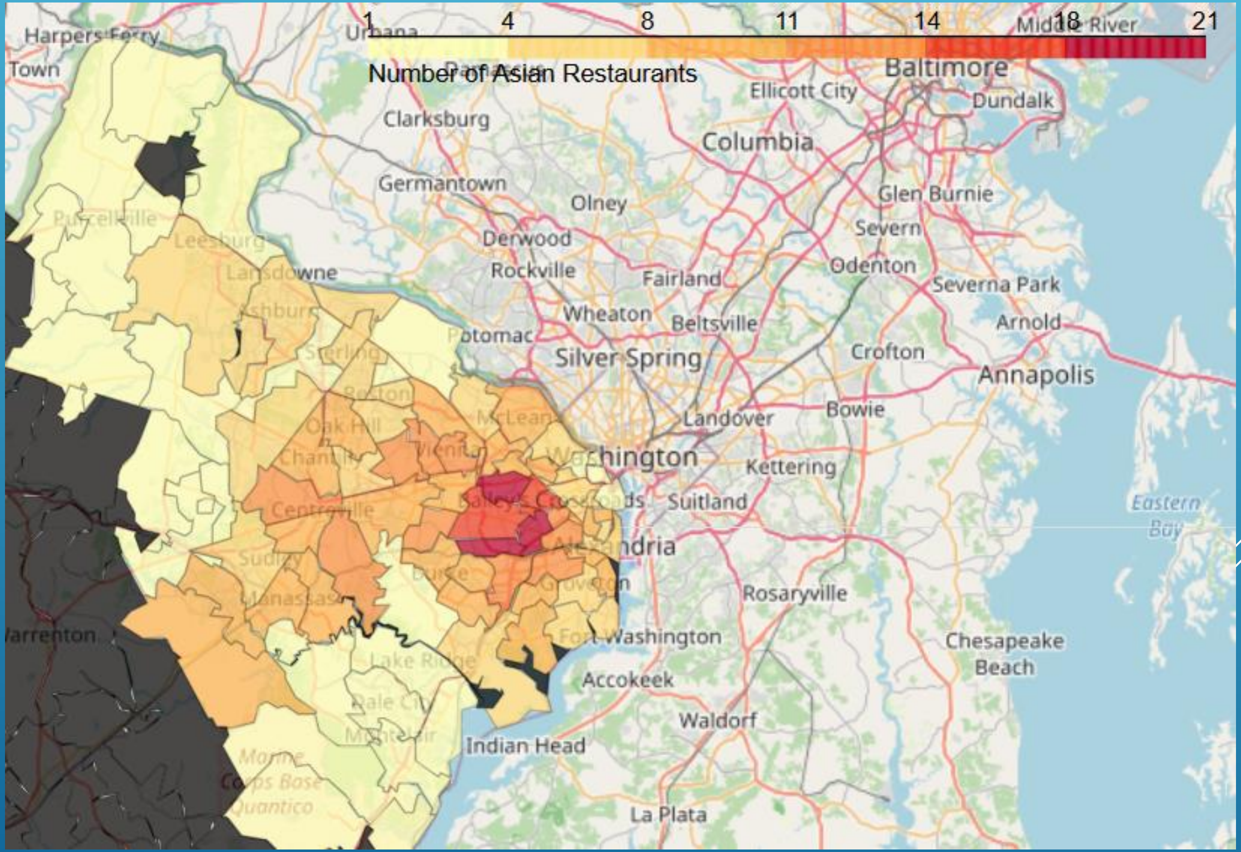
- The column "Venue Category" contains 14 unique categories.
- To the right is a bar chart of the frequency of each category
- Thai restaurants were most common with 376 results
- Ramen restaurants were the least common with 11 results
- Korean restaurants are somewhere in the middle, with 108 results
- Thai, Sushi, and Vietnamese restaurants are the 3 most common Asian restaurants in Northern Virginia



RESULTS CONT

- 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

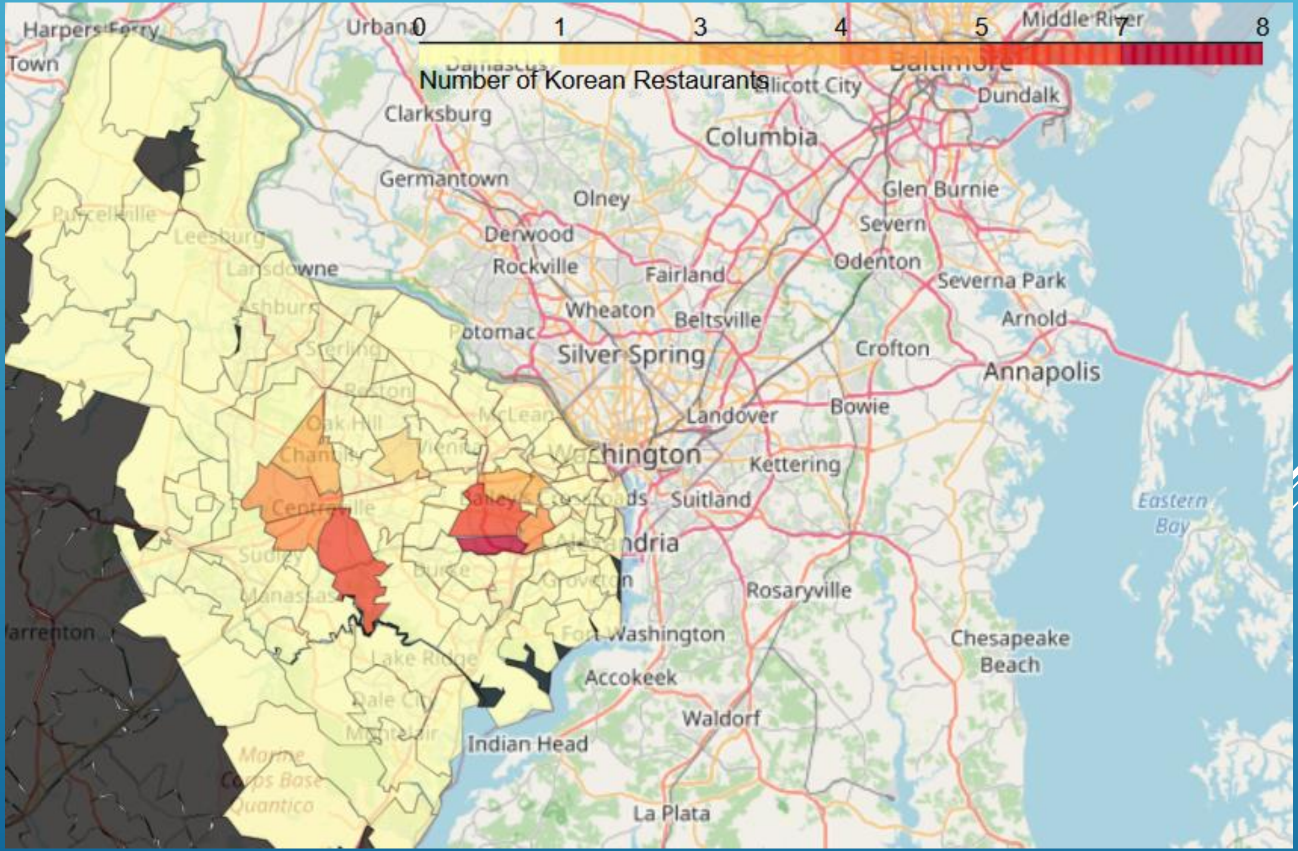
- To the right is a choropleth map of the frequency of **all Asian restaurants** in Northern Virginia by postal code
- While there are numerous Asian restaurants spread out over Northern Virginia, a concentration of Asian restaurants can be seen in the Centreville and Annandale areas



RESULTS CONT

- The data were further refined to include only venues that were labeled with the category "Korean Restaurant."

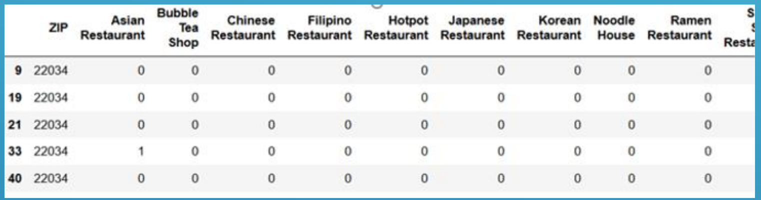
- To the right is a choropleth map of the frequency of **all Korean restaurants** in Northern Virginia by postal code
- After refining the data to include only Korean restaurants, we also see a concentration of Korean restaurants in Centreville and Annandale



RESULTS CONT

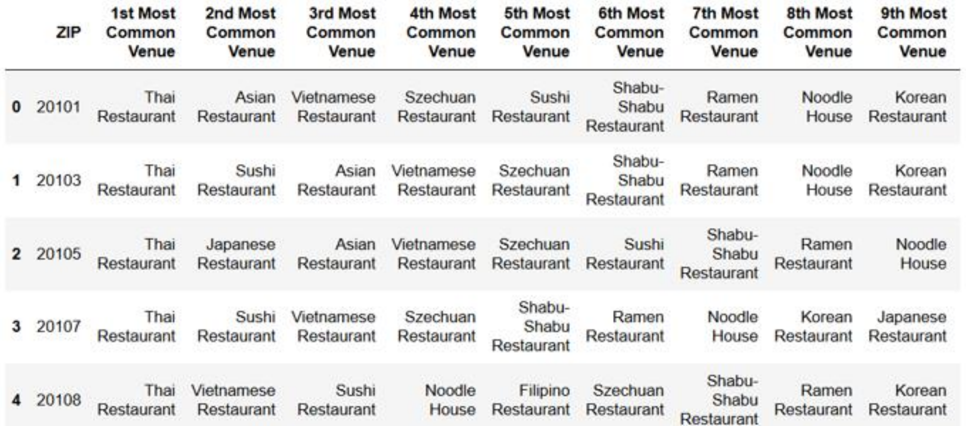
- Next, we process the data in order to employ a k-means clustering machine learning algorithm

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



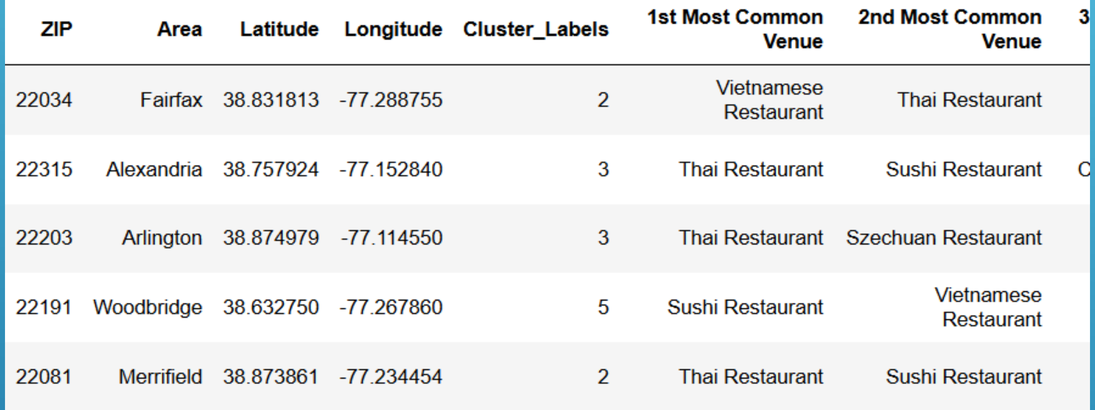
	ZIP	Asian Restaurant	Bubble Tea Shop	Chinese Restaurant	Filipino Restaurant	Hotpot Restaurant	Japanese Restaurant	Korean Restaurant	Noodle House	Ramen Restaurant	S S Resta
9	22034	0	0	0	0	0	0	0	0	0	
19	22034	0	0	0	0	0	0	0	0	0	
21	22034	0	0	0	0	0	0	0	0	0	
33	22034	1	0	0	0	0	0	0	0	0	
40	22034	0	0	0	0	0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

- Create a new dataframe that displays the postal codes along with "1st Most Common Venue" up to "10th Most Common Venue"



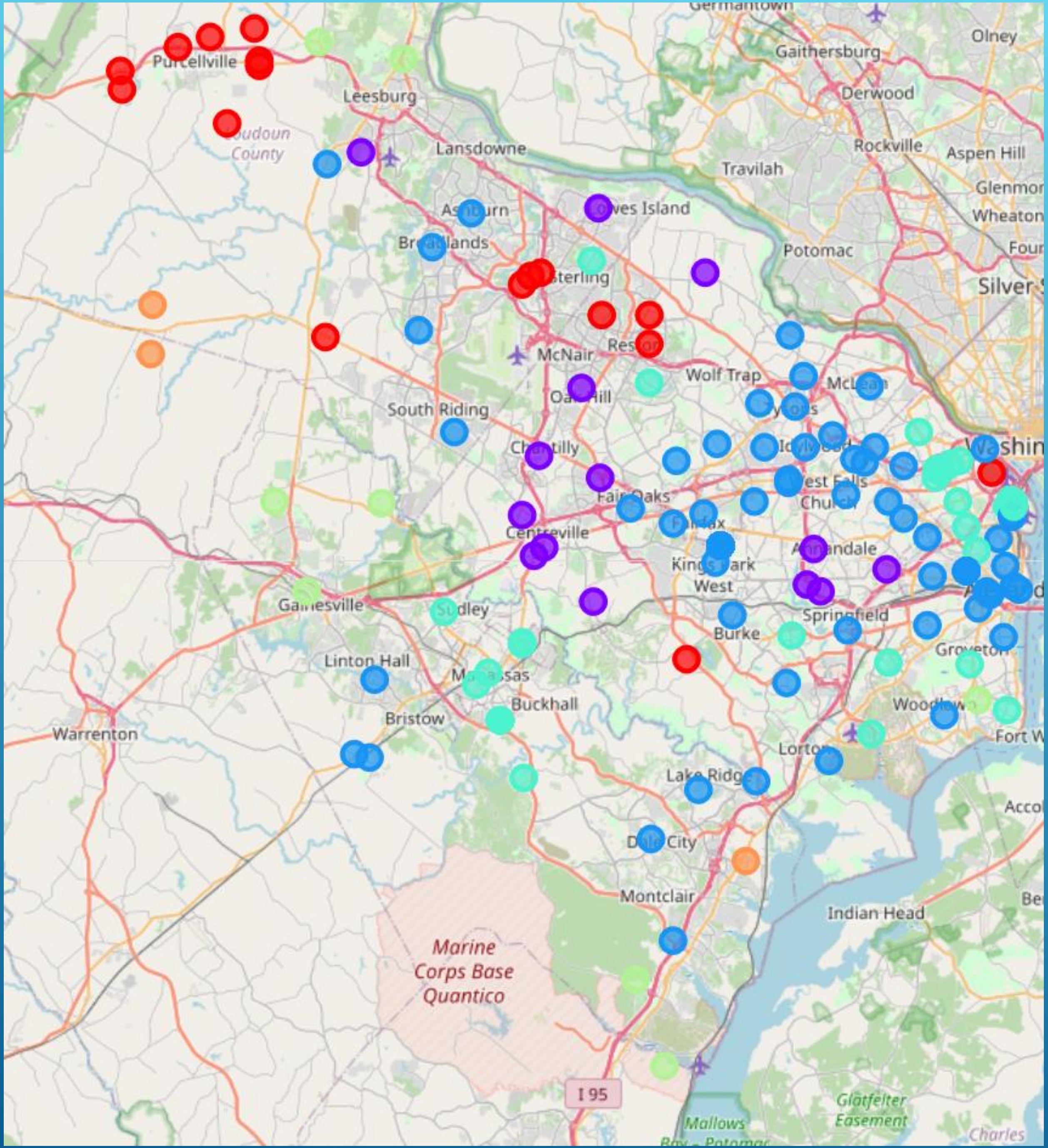
	ZIP	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	20101	Thai Restaurant	Asian Restaurant	Vietnamese Restaurant	Szechuan Restaurant	Sushi Restaurant	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

- K-means clustering algorithm was run, and the resulting clusters were added to the dataframe:



ZIP	Area	Latitude	Longitude	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3
22034	Fairfax	38.831813	-77.288755	2	Vietnamese Restaurant	Thai Restaurant	
22315	Alexandria	38.757924	-77.152840	3	Thai Restaurant	Sushi Restaurant	C
22203	Arlington	38.874979	-77.114550	3	Thai Restaurant	Szechuan Restaurant	
22191	Woodbridge	38.632750	-77.267860	5	Sushi Restaurant	Vietnamese Restaurant	
22081	Merrifield	38.873861	-77.234454	2	Thai Restaurant	Sushi Restaurant	

- The clusters were plotted on to a map of Northern Virginia, resulting in the visual on the next slide



DISCUSSION

- **Cluster 0 (Red)** is most associated with the slightly more rural and less dense areas in the north-west
- Annandale and Centreville, which contain a high amount of Korean restaurants, were grouped into **Cluster 1 (Purple)**
  - While most of Cluster 1 can be found around the Annandale and Centreville area, Leesburg, which is a somewhat distant north-west, was also grouped into Cluster 1
- 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 Washington D.C.
- Arlington, an area with a high density of postal codes, was mostly grouped into **Cluster 3 (Green)**
- **Clusters 4 (Yellowish/Mint Green)** and **Cluster 5 (Orange)** are mostly on the far edges of Northern Virginia

DISCUSSION CONT.

- 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.
- Additional recommendations:
  - Fairfax, 22033, Cluster 1
  - Arlington, Cluster 3

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
- 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
- Reliance on “Venue Category” from Foursquare. 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.
- 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