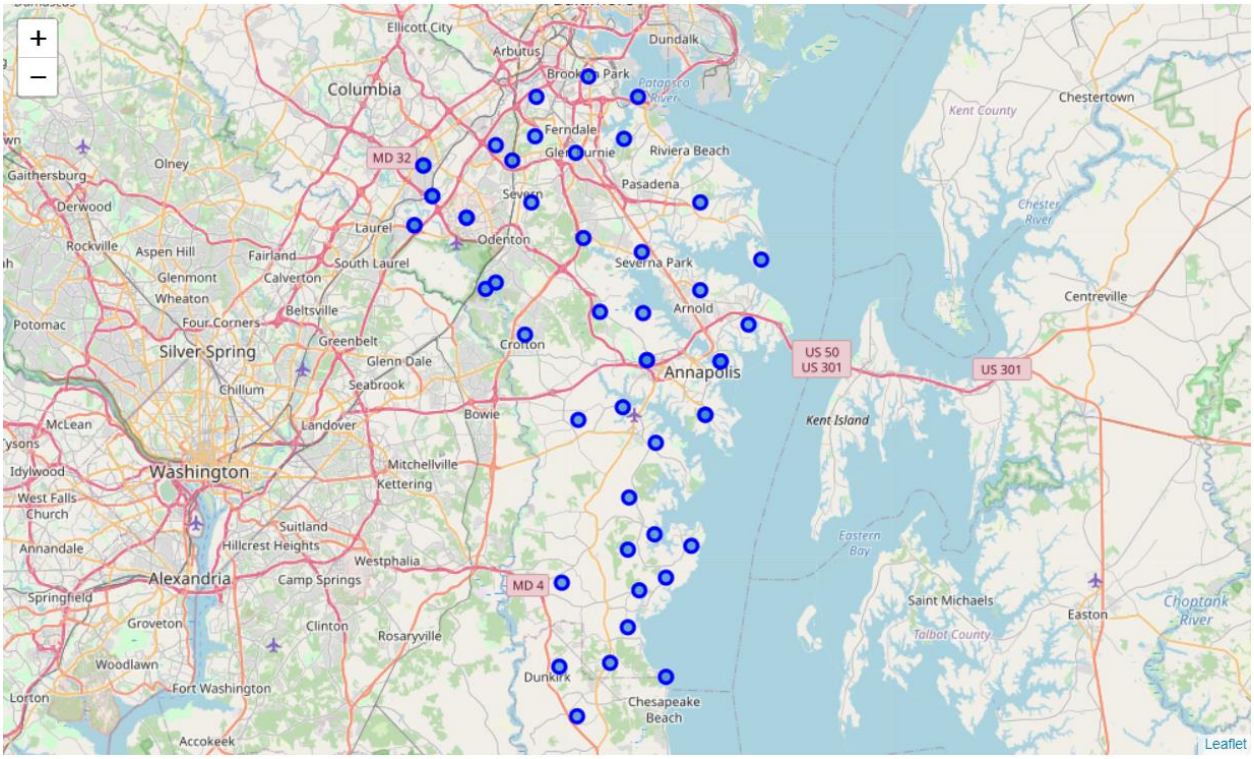
Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Christopher Lawrence

Data Analysis for New Coffee Shop Locations in Anne Arundel County Maryland

Contents



Overview

For the Capstone project, I chose a scenario in which a large, national coffee chain is searching for new locations in which to open new coffee shops. For the area, I chose to explore the region in which I currently live, Anne Arundel County Maryland. This region is located in the suburbs between Baltimore Maryland and Washington D.C. This area has a large metropolitan population with many coffee shops throughout the area.

I will use the Foursquare data API to establish the current coffee shop environment in Anne Arundel County in conjunction with other data sets from the Maryland and Anne Arundel Maryland County open data set web portals.

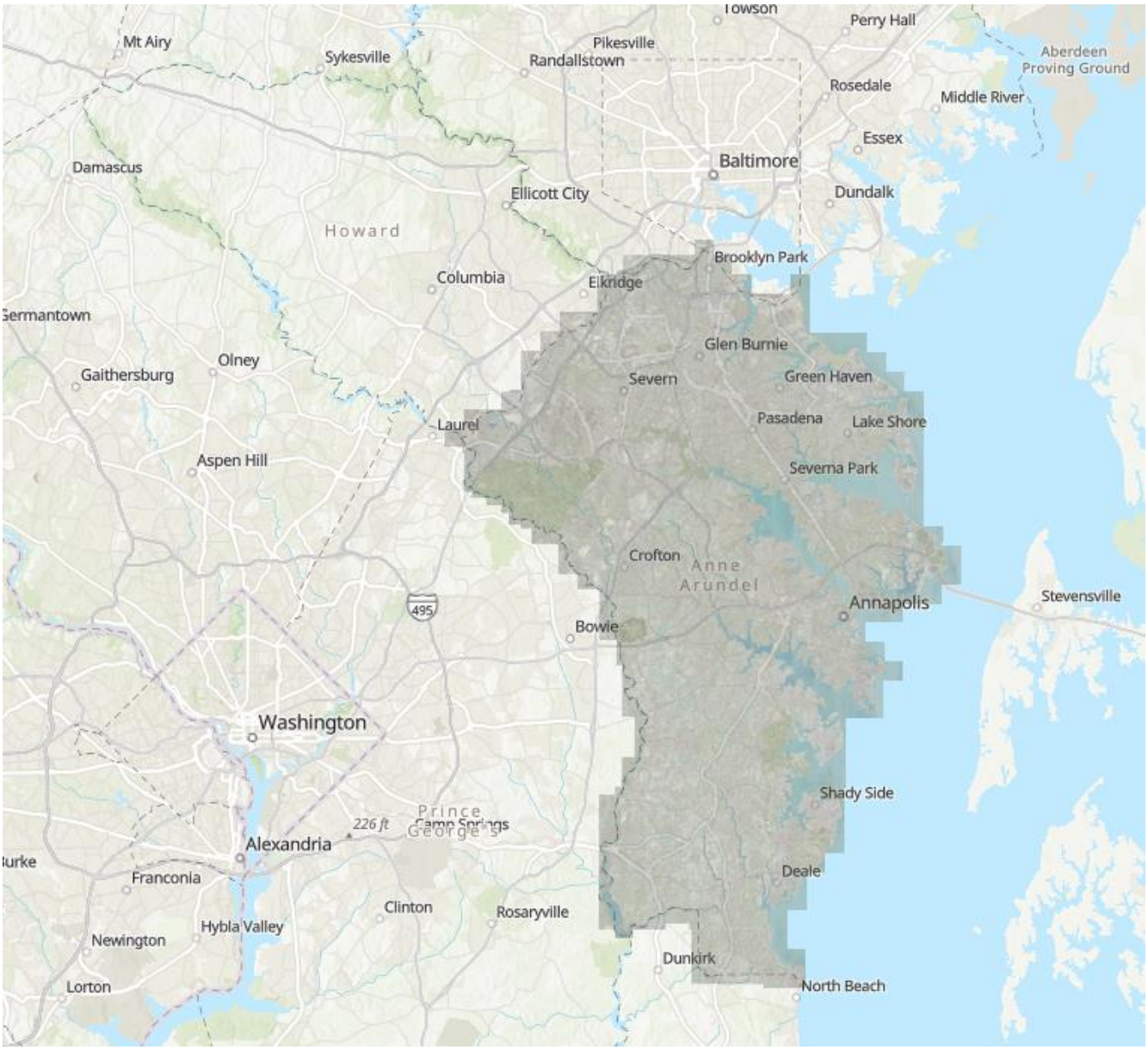
Introduction and Background: Business Problem

This section will present a brief description of the problem/question to be explored and a discussion of the background of the scenario presented in this Capstone.

As stated in the overview section, this Capstone will present the problem of searching for an 'ideal' location to expand a new coffee shop store somewhere in the county of Anne Arundel Maryland. Anne Arundel Maryland is a large, suburban county in the state of Maryland located between the cities of Baltimore Maryland and Washington D.C. The problem/scenario presented in this Capstone will center around defining the properties of an 'ideal' location given the Foursquare API dataset information for existing coffee shops throughout the Anne Arundel county area.

After exploring the various attributes of the existing coffee shop data in the Foursquare database (in conjunction with a systematic partitioning of Anne Arundel county via county open data sets (discussed in the data section), the business problem of where to locate a new coffee shop will become further defined. In this scenario, the company wishes to expand its current number of coffee shops in the Anne Arundel county area, but also wants to maximize its business opportunities by avoiding areas of the county that are saturated in existing coffee shops and other establishments that cut into the coffee shop business.

Map of Anne Arundel County, Maryland

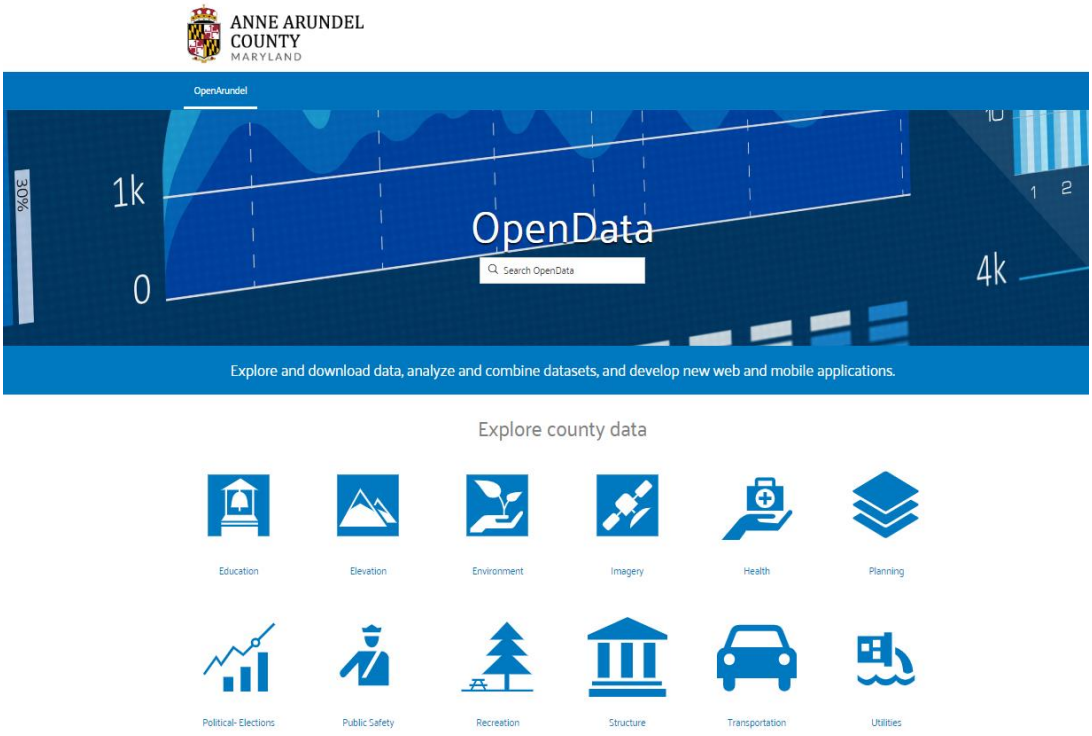


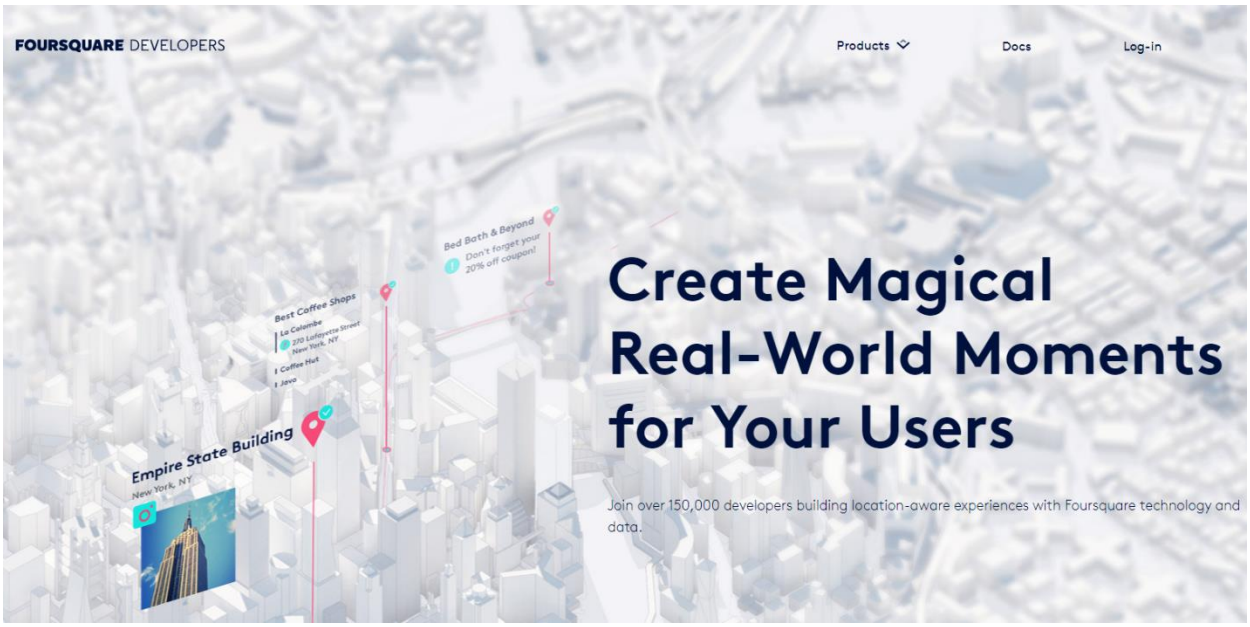
Data

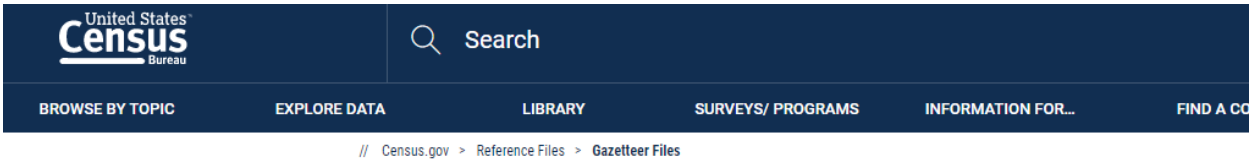
This section will provide a description of the data, the data sources, and how it will be used to solve the problem presented in this Capstone project.

This project will utilize the Foursquare API database to obtain data on all the existing coffee shops in Anne Arundel County Maryland. To further explore this data, the coffee shop location information will be partitioned in the Anne Arundel County area by using locational data from the Anne Arundel County and Maryland State open data web portals.

This will create a picture of the current coffee shop 'landscape' which will help to avoid areas that are congested with existing competitive shops. An examination of the Foursquare data on existing coffee shops may illuminate further attributes or areas of investigation that would enhance the decision making process of where to locate a new coffee shop in the county.







// Census.gov > Reference Files > **Gazetteer Files**

GEOGRAPHIES
- Mapping Files
- Mapping Tools
- Reference Files
- Reference Maps
- Back to Reference Files

Gazetteer Files

The U.S. Gazetteer Files provide a listing of all geographic areas for selected geographic area types. The files include geographic identifier codes, names, area measurements, and representative latitude and longitude coordinates.

2019 2018 2017 2016 2015 2014 2013 2012 MORE

Anne Arundel County

Data Sources

- https://opendata.aacounty.org/
- https://planning.maryland.gov/MSDC/Pages/Cnty_Menu/Anne.aspx
- https://www.aacounty.org/
- https://developer.foursquare.com/
- https://public.opendatasoft.com/
- https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html

Methodology

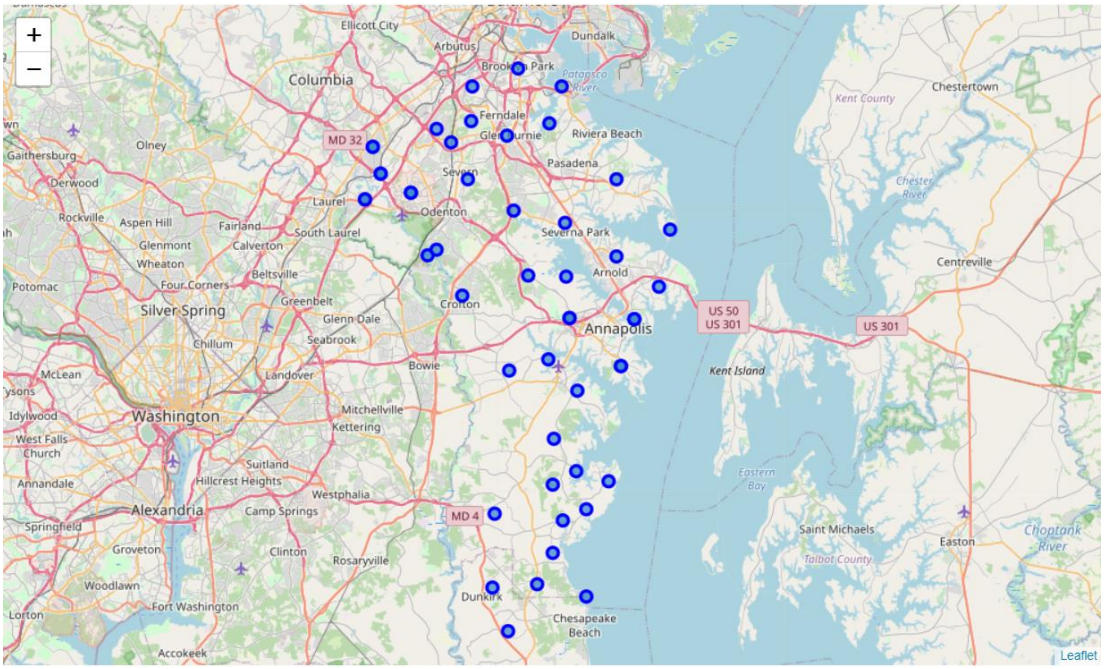
The main idea is to partition-up Anne Arundel County into its zip code area constituents. So, I obtained the zip code dataset for Anne Arundel County from the open data web portal (https://opendata.aacounty.org/datasets/zip-codes) and imported the csv file into a Pandas dataframe.

I performed the necessary data wrangling and cleaning on the dataframe until it resulted in a simple dataframe with a city/neighborhood name, zip code, and object_id. This dataframe showed that there are a total of 49 zip code cities in Anne Arundel County.

Next, I needed to obtain the latitude and longitude data for each zip code area in order to plot the dataset. For this, I imported a Maryland State zip code dataset from Census.gov which also contained the lat/long for each zip code. After wrangling and cleaning this new dataframe, I merged the two dataframes on the common zip code entries to obtain a new dataframe with 49 zip code areas with city names and lat/long information.

I then plotted this dataframe with Folium to obtain a general zip code map of Anne Arundel County (below).

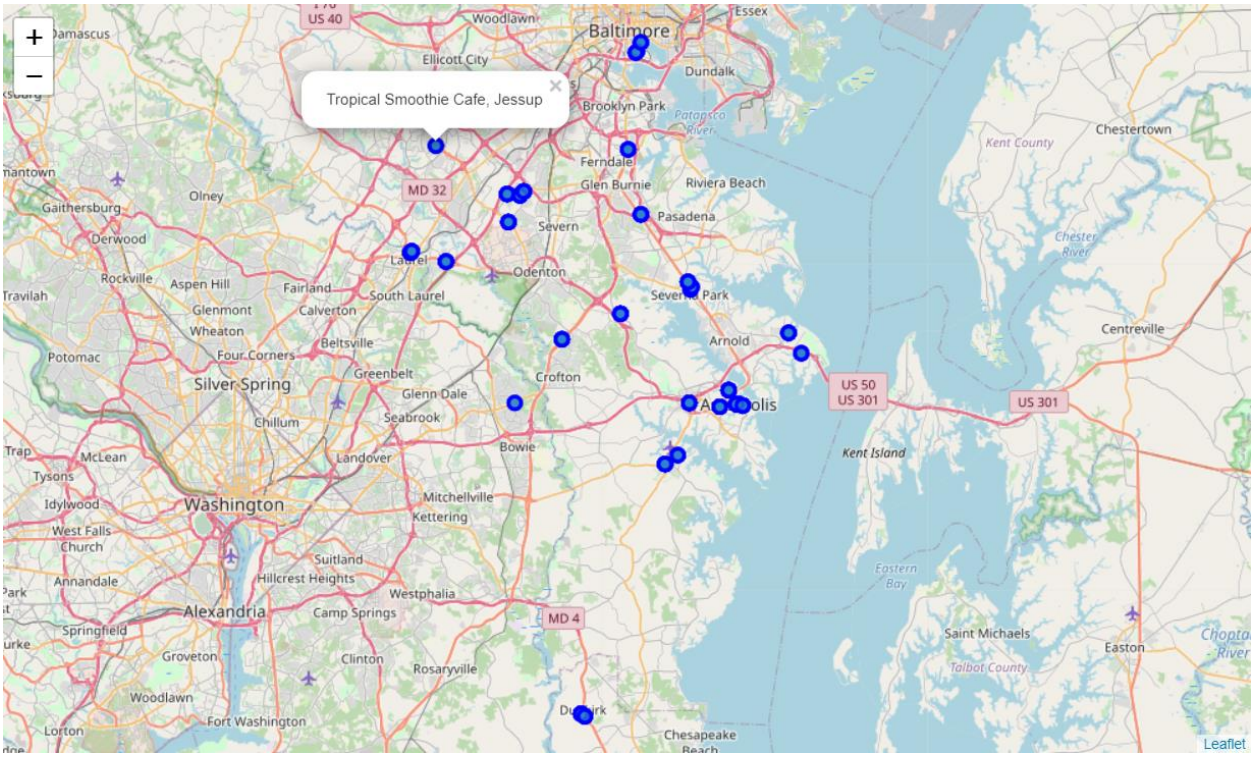
Map of Anne Arundel County zip codes



I then proceeded to obtain a json dataset of all the coffee shops in Anne Arundel County from the Foursquare API. I modified the function from the peer assignment in week 3 to convert each json entry into a new Pandas dataframe containing each city name, city_lat, city_long, venue_name, venue_lat, venue_long and venue_category. After a considerable amount of data wrangling and cleaning with this set, I finally obtained a usable dataframe of only coffee shops in Anne Arundel County along with their city and latitude and longitude.

I then used Folium again to plot the 184 coffee shops on a map of Anne Arundel County (below)

Map of Anne Arundel County Coffee Shops



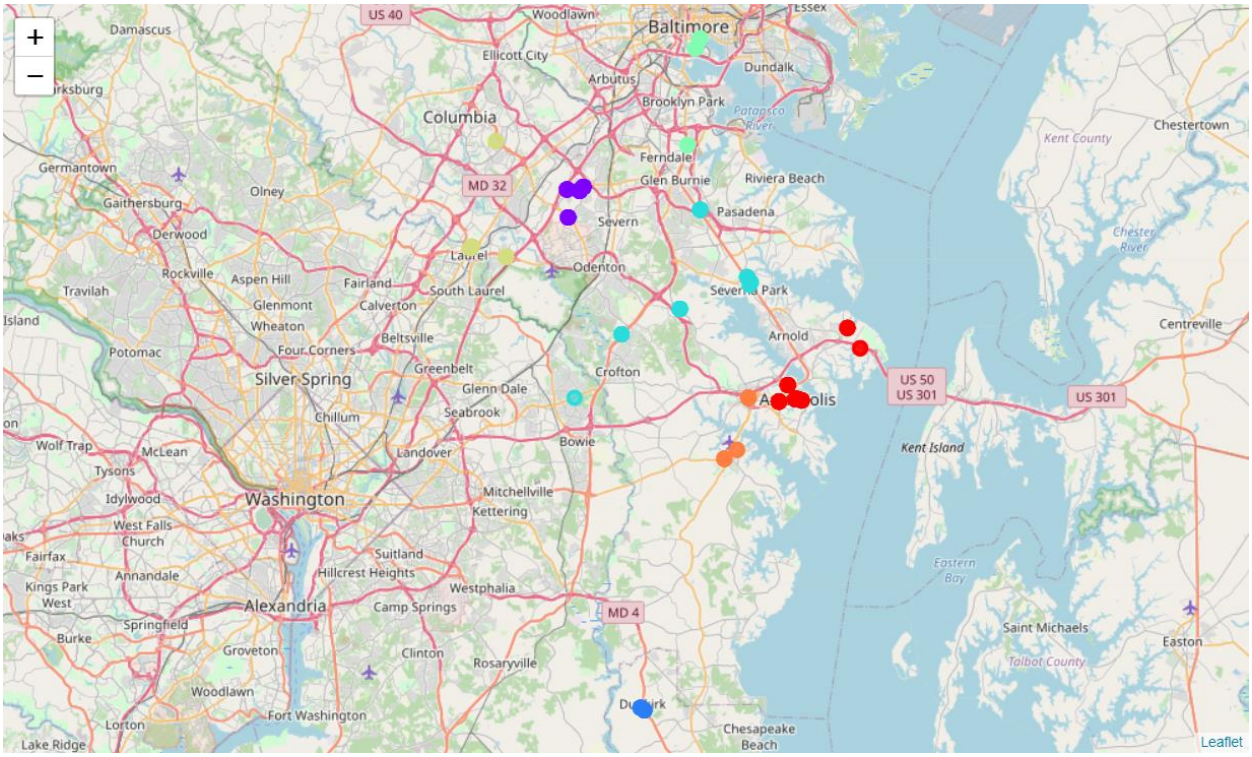
Results

The process now turned to how to analyze this data to obtain the best reasoning for my criteria of where to place a new coffee shop in Anne Arundel County. The simple criteria that I've chosen for this assignment is to use geolocational methods to observe the density of coffee shops in different zip code areas of Anne Arundel County as the primary factor in considering where to locate a new coffee shop.

I proceeded to use and compare two types of clustering algorithms to see what results they might produce when clustering pure locational data. The first algorithm that I tested was K-means clustering.

This resulted in the following cluster map of coffee shops using k=7 clusters:

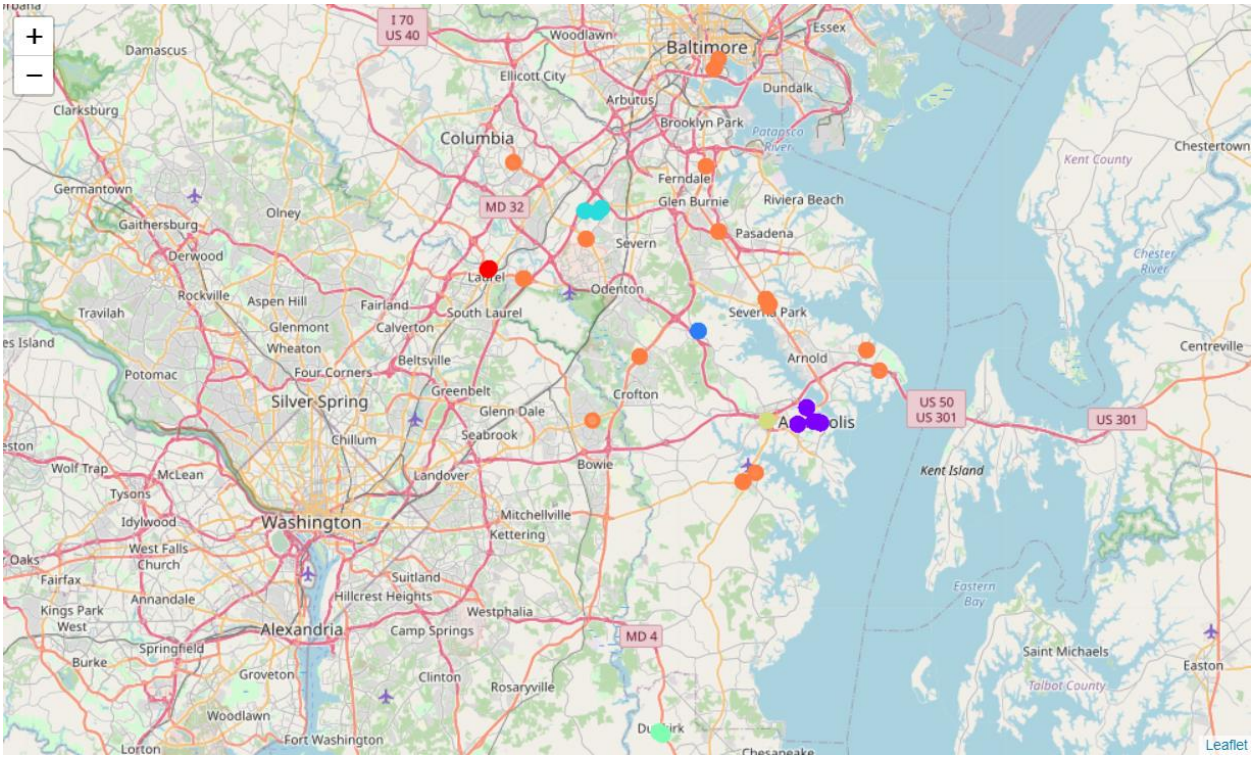
Map of K-means clustering of coffee shops



The second algorithm that I tested was DBSCAN clustering.

This resulted in 7 clusters of vary different categories than that of the K-means 7 clusters.

Map of DBSCAN clustering of coffee shops



Discussion

This project had quite a number of challenges. For instance, given the coffee shop data from Foursquare, the dataset contains two different venue categories for a ‘coffee shop.’ One was simply, ‘Coffee Shop’ while the other one was ‘Café’ in which the character encoding did create a few problems with Pandas and Boolean logic (since it was also rendered as CafÃ©).

Given more time, I might have wanted to combine other relevant datasets that may have given additional insight into where new coffee shop locations might work better than others. For instance, a dataset with the demographics of each of the zip code areas of Anne Arundel County (from either the Maryland State open data portal or the Anne Arundel County open data portal) to add population data for each zip code area as an added feature, or perhaps some type of business metrics from another dataset that would provide another feature for consideration.

I only used the locational data from the Foursquare API in this project. However, I might be able to include other features about each coffee shop from the Foursquare data and add it to the decision process.

Conclusion

For this project, mastering the two clustering algorithms and getting them to work correctly was a major challenge with my combined datasets. This was the major factor in only using locational data as my primary source or feature when answering my primary question of where to locate a new coffee shop in Anne Arundel County.

However, I was able to obtain a dataset of coffee shops by zip code area for Anne Arundel County, Maryland. I was then able to plot the raw dataset of coffee shops on a map of Anne Arundel County. I then utilized two clustering algorithms to obtain two different partitions of the current coffee shop clustering in the county.

With these two clustering maps, I was able to locate a few potential areas in Anne Arundel County in which to locate a new coffee shop for this national coffee chain.