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

Introduction and Background: Business Problem

Capstone project

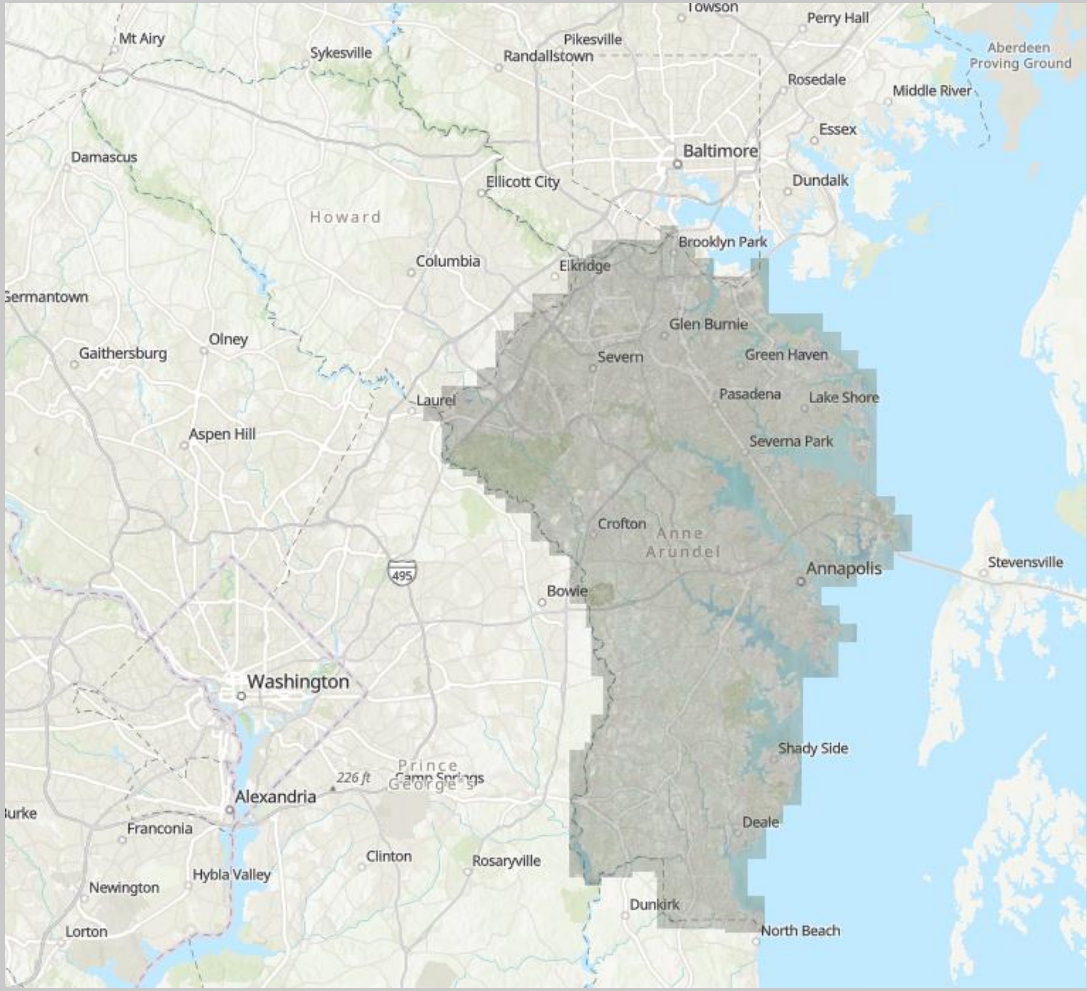
Scenario -- a large, national coffee chain is searching for new locations in which to open new coffee shops.

Region/Area: locations in Anne Arundel County Maryland. (suburbs between Baltimore Maryland and Washington D.C.)

Large metropolitan population with many coffee shops throughout the area to choose from.

Utilize 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.

Determine the best location to establish a new coffee shop branch for this coffee chain taking into account coffee shop density in the zip code areas of Anne Arundel County.



Data

United States Census Bureau

Search

BROWSE BY TOPIC | EXPLORE DATA | LIBRARY | SURVEYS/ PROGRAMS | INFORMATION FOR... | FIND A CO

// Census.gov > Reference Files > Gazetteer Files

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

Gazetteer Files

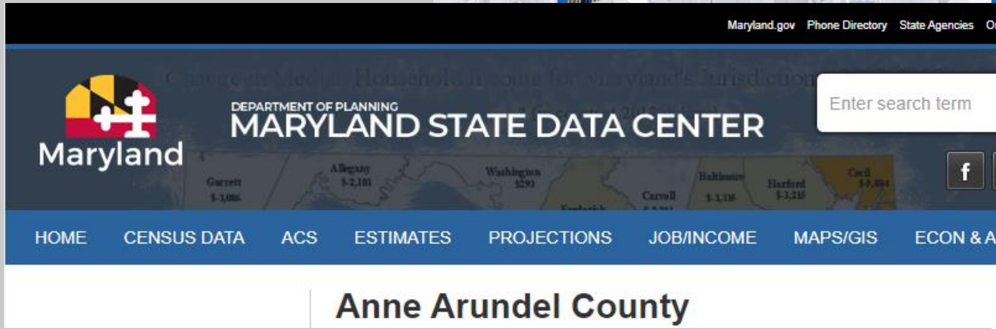
The U.S. Gazetteer Files provide a listing of all geographic areas for selected geographic area types. The files include geographic identifier codes, names, area measurements, and representative latitude and longitude coordinates.

2019 | 2018 | 2017 | 2016 | 2015 | 2014 | 2013 | 2012 | MORE

FOURSQUARE DEVELOPERS

Products | Docs | Log-in

Create Magical Real-World Moments for Your Users



Maryland.gov | Phone Directory | State Agencies | Online Services | Translate

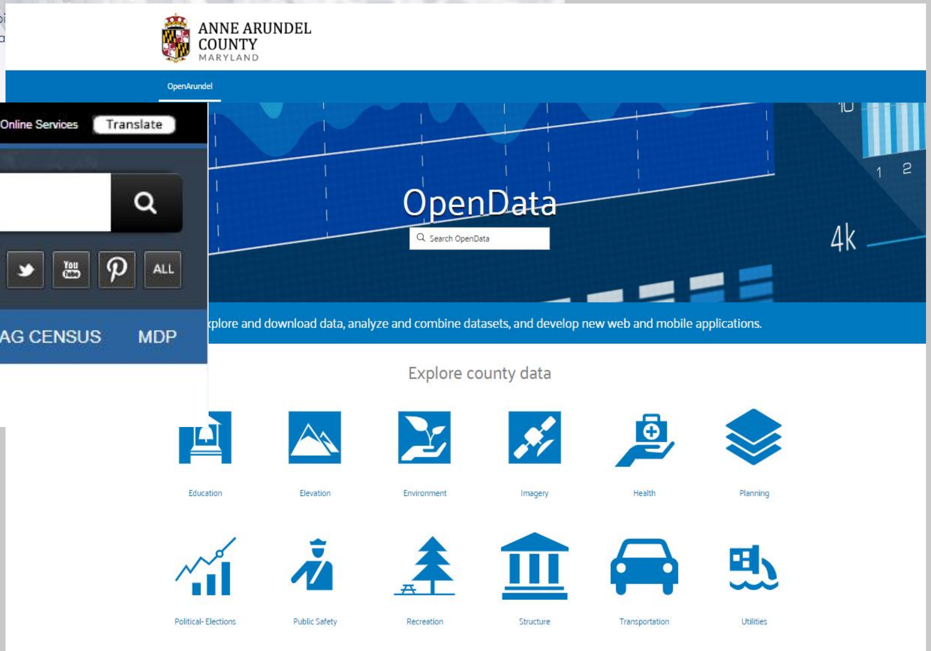
Maryland

DEPARTMENT OF PLANNING
MARYLAND STATE DATA CENTER

Enter search term

HOME | CENSUS DATA | ACS | ESTIMATES | PROJECTIONS | JOB/INCOME | MAPS/GIS | ECON & AG CENSUS | MDP

Anne Arundel County



ANNE ARUNDEL COUNTY MARYLAND

OpenArundel

OpenData

Search OpenData

xplore and download data, analyze and combine datasets, and develop new web and mobile applications.

Explore county data

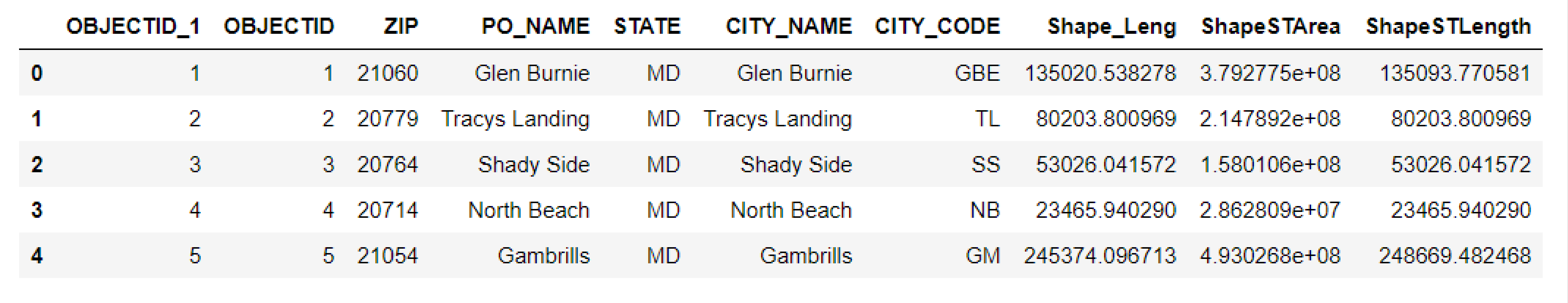
Data Sources:

- 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

Data

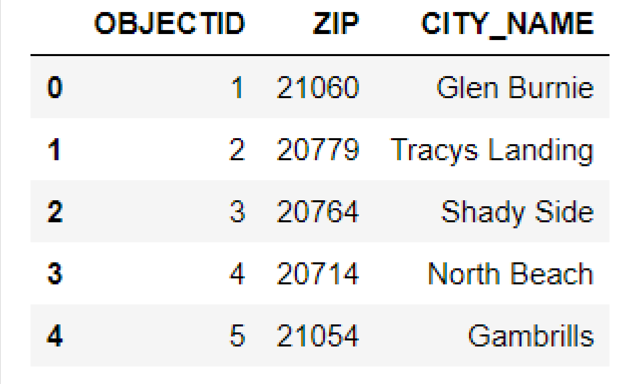
Obtaining target zip codes for Anne Arundel County Maryland
https://opendata.arcgis.com/datasets/899d24a210094af38a6ebe83c53c760f_7.csv

Initial (49,10) dataframe (head output below)



	OBJECTID_1	OBJECTID	ZIP	PO_NAME	STATE	CITY_NAME	CITY_CODE	Shape_Leng	ShapeSTArea	ShapeSTLength
0	1	1	21060	Glen Burnie	MD	Glen Burnie	GBE	135020.538278	3.792775e+08	135093.770581
1	2	2	20779	Tracys Landing	MD	Tracys Landing	TL	80203.800969	2.147892e+08	80203.800969
2	3	3	20764	Shady Side	MD	Shady Side	SS	53026.041572	1.580106e+08	53026.041572
3	4	4	20714	North Beach	MD	North Beach	NB	23465.940290	2.862809e+07	23465.940290
4	5	5	21054	Gambrills	MD	Gambrills	GM	245374.096713	4.930268e+08	248669.482468

Cleaned (49,3) dataframe



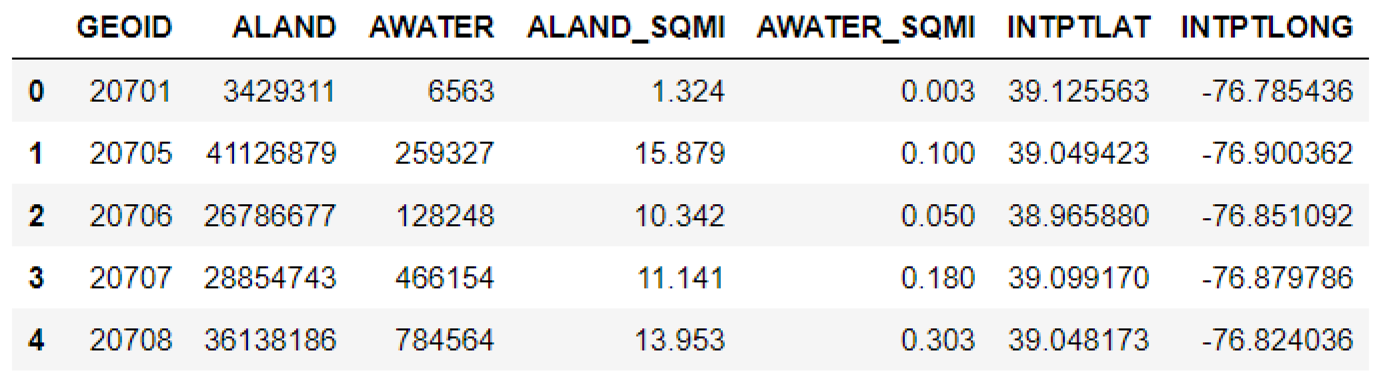
	OBJECTID	ZIP	CITY_NAME
0	1	21060	Glen Burnie
1	2	20779	Tracys Landing
2	3	20764	Shady Side
3	4	20714	North Beach
4	5	21054	Gambrills

Data

Obtaining latitude and longitude information for the target zip codes

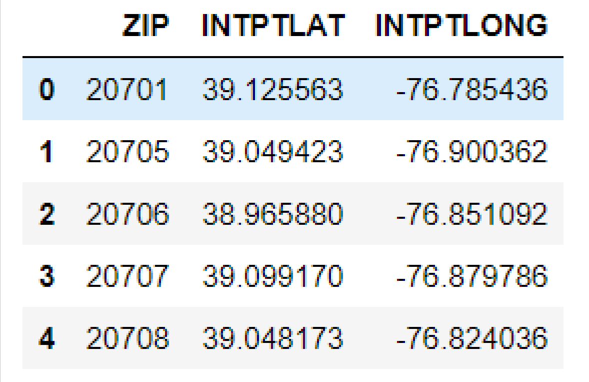
- https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html

Initial (219,7) dataframe (head output below)



	GEOID	ALAND	AWATER	ALAND_SQMI	AWATER_SQMI	INTPTLAT	INTPTLONG
0	20701	3429311	6563	1.324	0.003	39.125563	-76.785436
1	20705	41126879	259327	15.879	0.100	39.049423	-76.900362
2	20706	26786677	128248	10.342	0.050	38.965880	-76.851092
3	20707	28854743	466154	11.141	0.180	39.099170	-76.879786
4	20708	36138186	784564	13.953	0.303	39.048173	-76.824036

Cleaned (219,3) dataframe

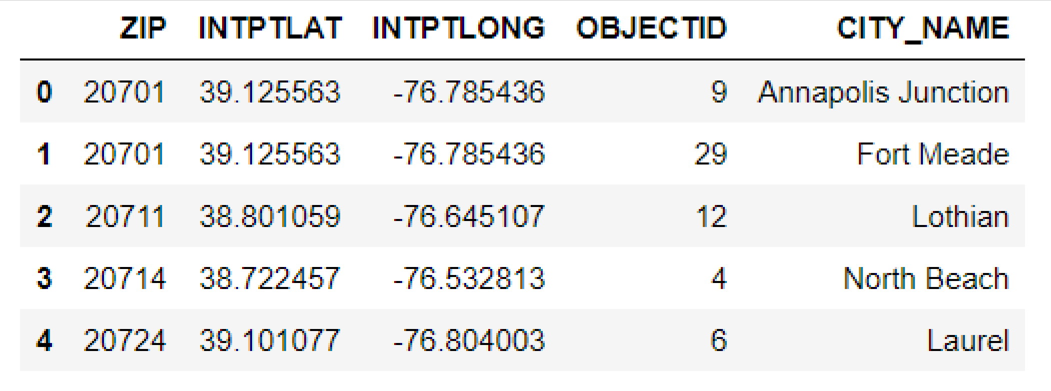


	ZIP	INTPTLAT	INTPTLONG
0	20701	39.125563	-76.785436
1	20705	39.049423	-76.900362
2	20706	38.965880	-76.851092
3	20707	39.099170	-76.879786
4	20708	39.048173	-76.824036

Data

Now merging the zip code area dataframe with the lat/long dataframe on the key value of zip code:

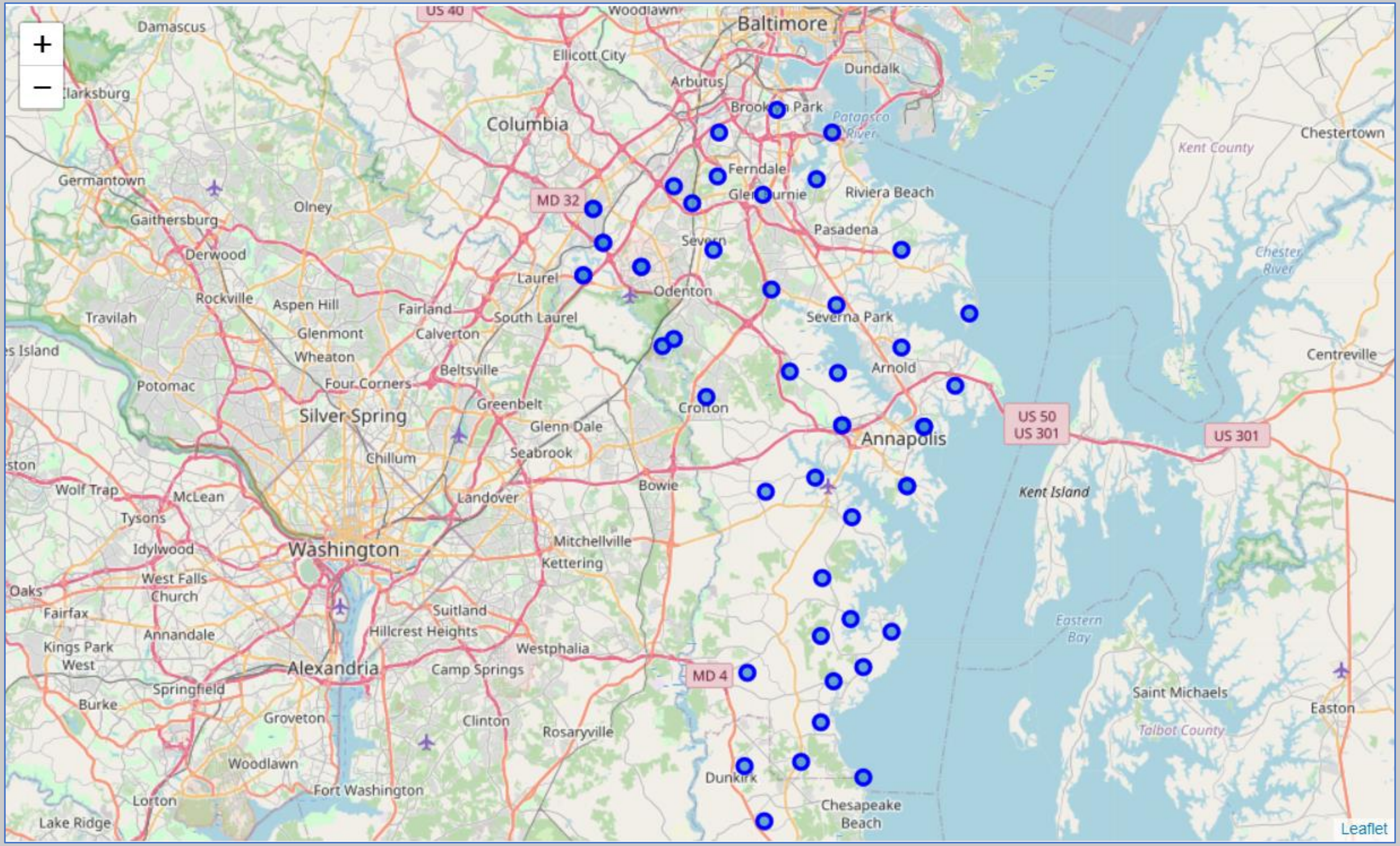
Produces the final (49,5) dataframe for all zip code areas in Anne Arundel County, dataset head output below:



	ZIP	INTPTLAT	INTPTLONG	OBJECTID	CITY_NAME
0	20701	39.125563	-76.785436	9	Annapolis Junction
1	20701	39.125563	-76.785436	29	Fort Meade
2	20711	38.801059	-76.645107	12	Lothian
3	20714	38.722457	-76.532813	4	North Beach
4	20724	39.101077	-76.804003	6	Laurel

Data

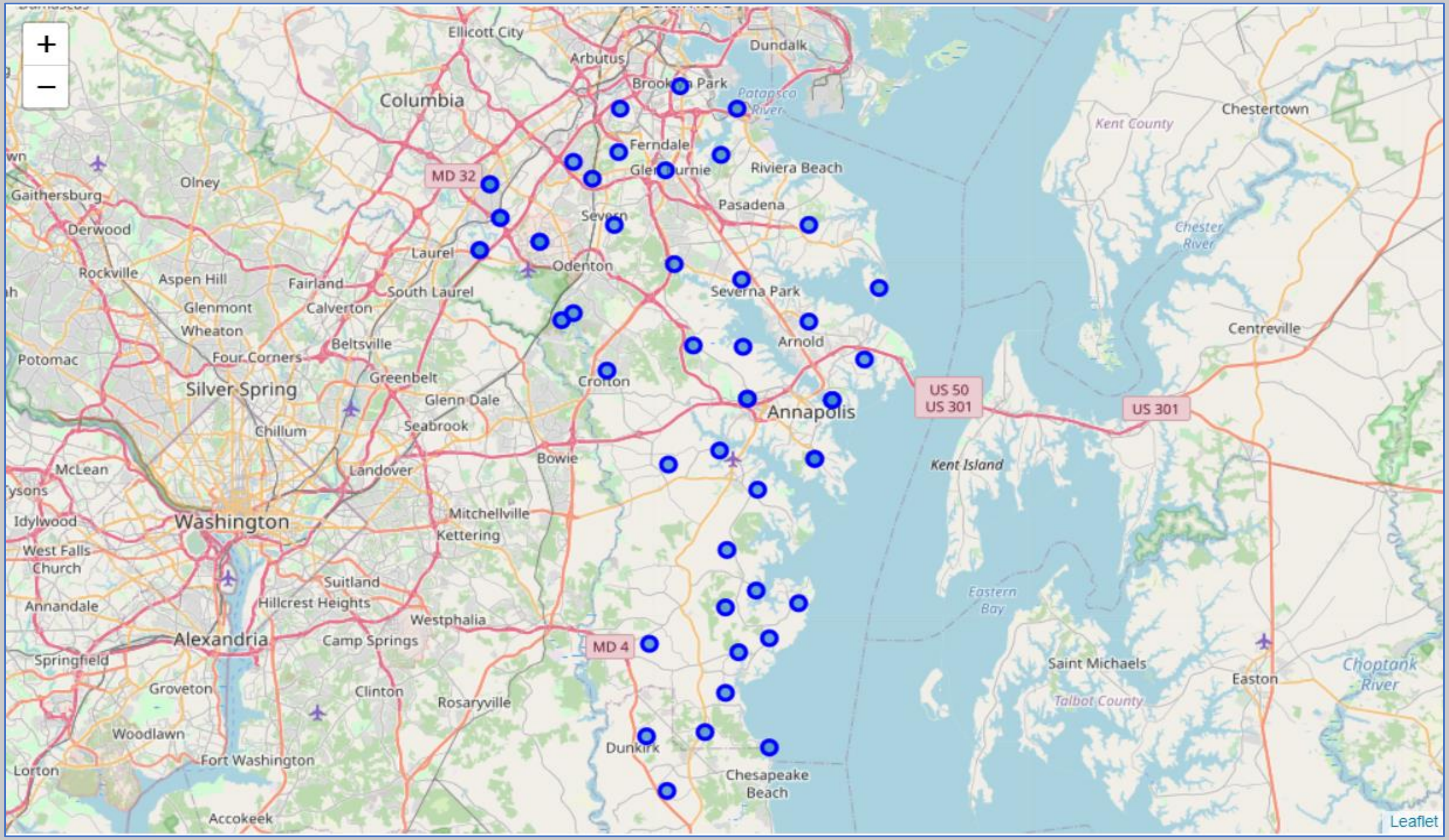
Plotting the zip code areas on the Anne Arundel County map produced:



Methodology

- 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.
- Perform 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.
- 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.
- Plot this dataframe with Folium to obtain a general zip code map of Anne Arundel County (next slide).

Methodology

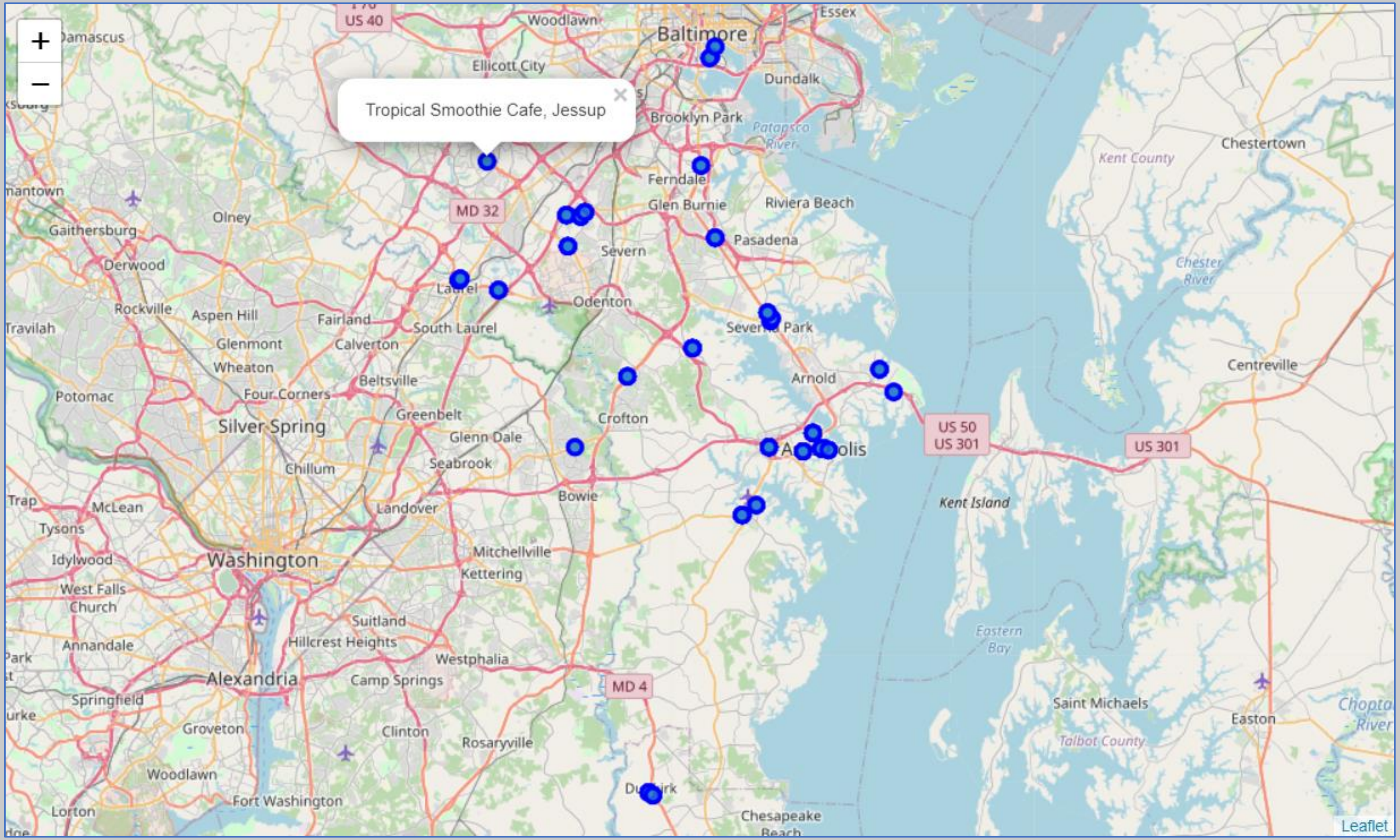


Map of Anne Arundel County zip codes

Methodology

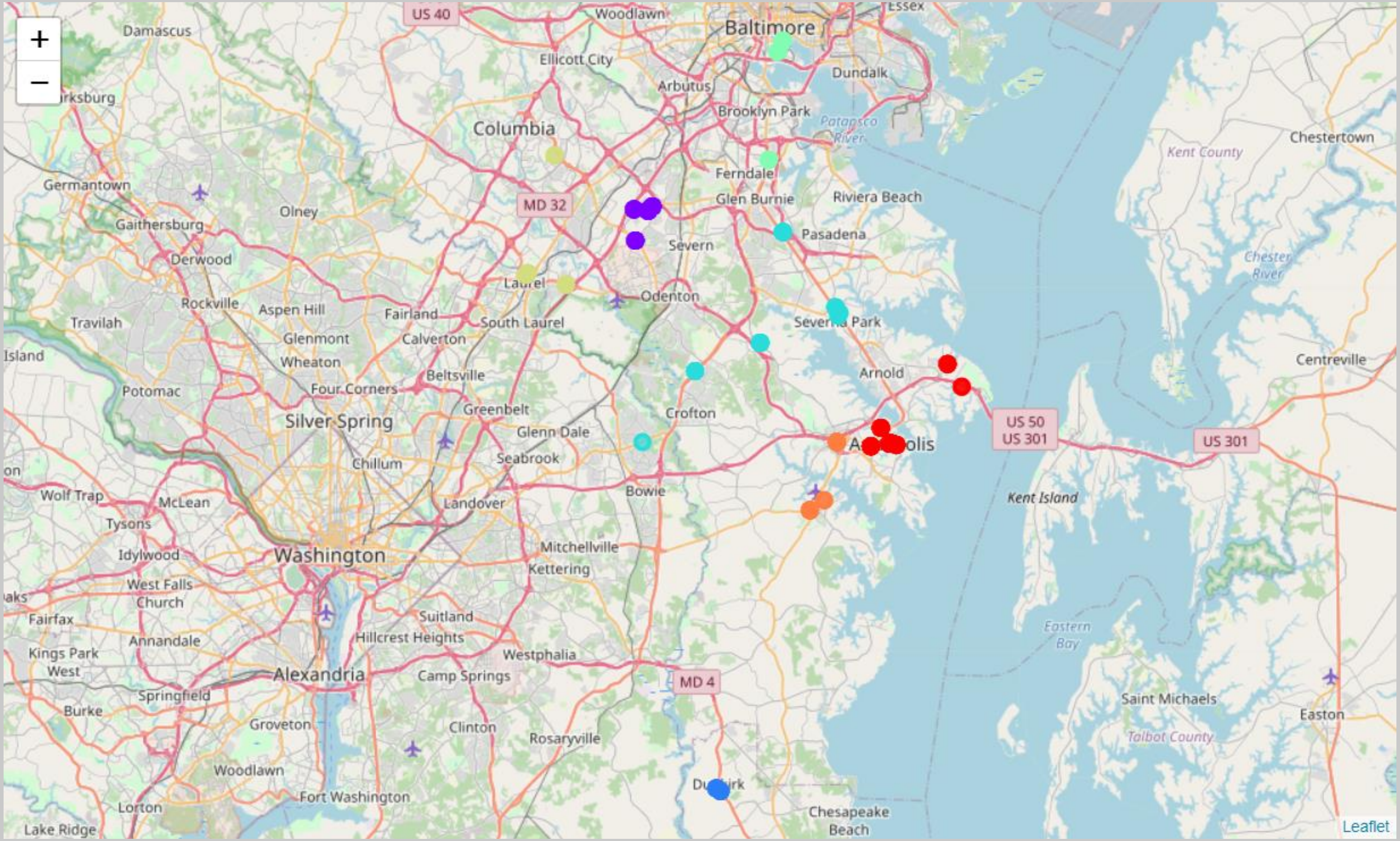
- Obtained a json dataset of all the coffee shops in Anne Arundel County from the Foursquare API.
- Modified the function from the peer assignment in week 3 to convert each json entry into a new Pandas dataframe containing each city name, city_lat, city_long, venue_name, venue_lat, venue_long and venue_category.
- 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.
- Used Folium again to plot the 184 coffee shops on a map of Anne Arundel County (below)

Methodology



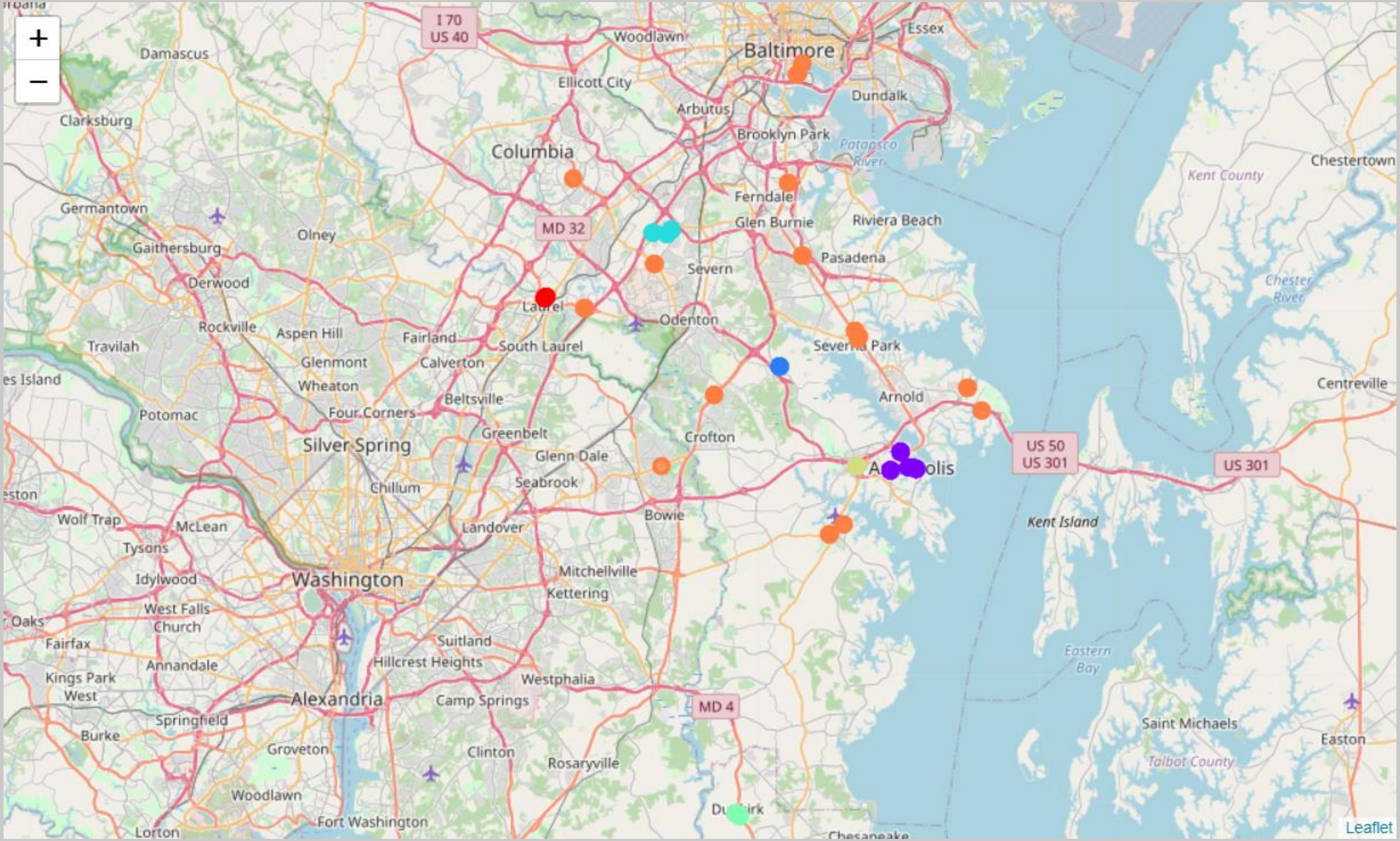
Map of Anne Arundel County Coffee Shops

Results



Coffee Shop Cluster Map from K-means Clustering Algorithm

Results



Coffee Shop Cluster Map from DBSCAN Clustering Algorithm

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.