Coursera Capstone Project

Deciding Where to Move with Data Science

Braden Smolko

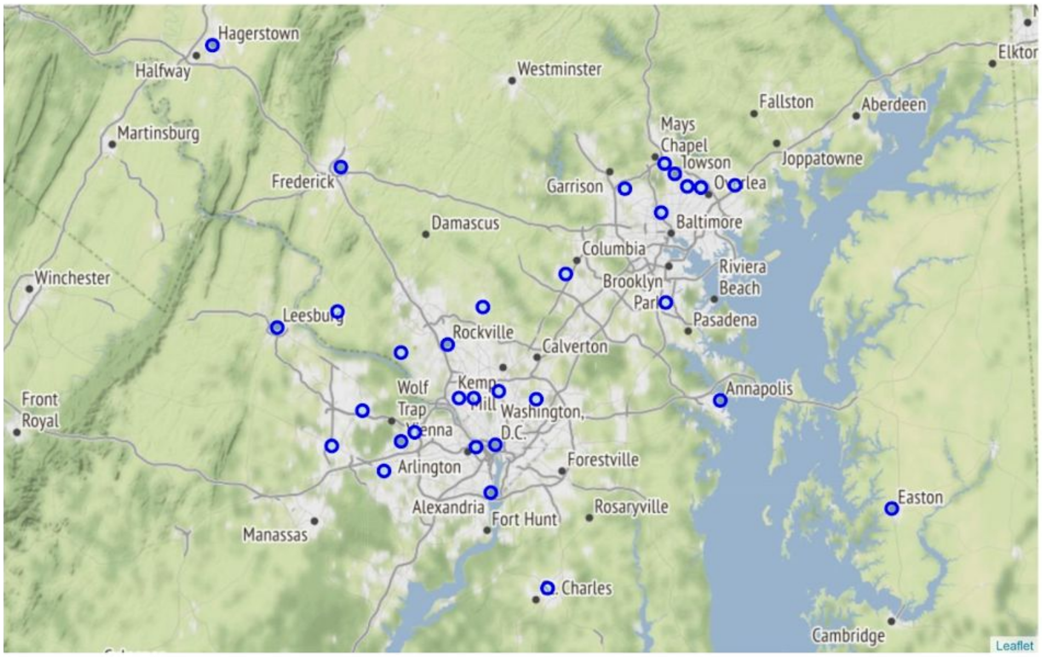
Introduction

My girlfriend and I are looking to move soon. She is applying to graduate school and when she gets in, we will have to move a bit closer. We really like where we are living currently, so I thought it would be cool to find other living areas that are similar to Frederick! I am also finishing up my IBM Data Science Professional Certificate, so it seemed like the perfect time to use my new skills to answer this question. Let’s dive into my process to find our potential new neighborhood.

Data Collection and Prep

Possible Cities and Geocoding

To start, I had my girlfriend create a list of possible cities for us to move to. To beef up this list, I added some extra towns from the Baltimore-Washington Metropolitan Area that were similar in size to Frederick. The final list included 31 cities in total. This list included just the names of the cities and the states that they were in, but in order to work with the cities as locations, I had to geocode them (get the latitude and longitude coordinates). I used the Geopy package to get these coordinates. Once the cities had coordinates connected to them, I was able to map them out using the Folium package!



Having these coordinates also allowed me to use the Foursquare API to search for venues around each city.

Foursquare API Data

The Foursquare API allows users to access location data for free! The API has multiple types of endpoints and searches, but the one we are interested in is the explore endpoint. The explore endpoint gives you a list of venues and information for a given location and radius.

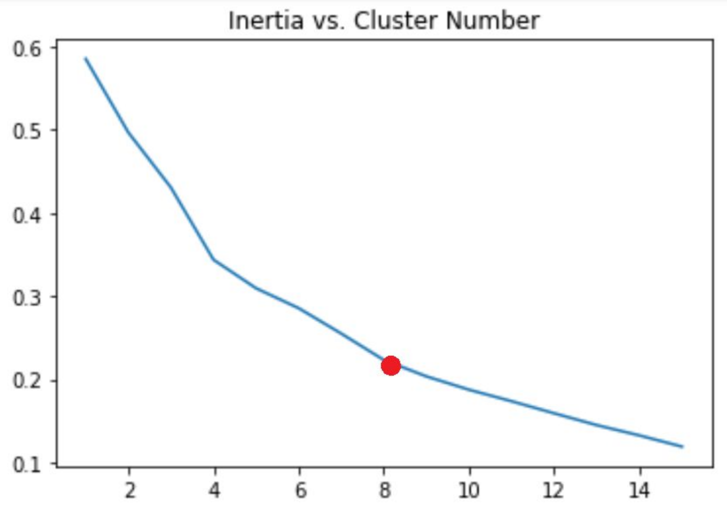
You can also specify how many venues that you want returned and how to sort them. I sorted by popularity to get the Top 100 most popular venues within 5km of each city. I then collected all of the categories from the returned venues (over 2400) and tagged each category with the venue's city. This allowed me to create frequency counts of all the different categories in each city. I converted these counts into percentages to be used by a clustering algorithm. Let's see how the cities match up!

Data Processing

Choosing Model Parameters

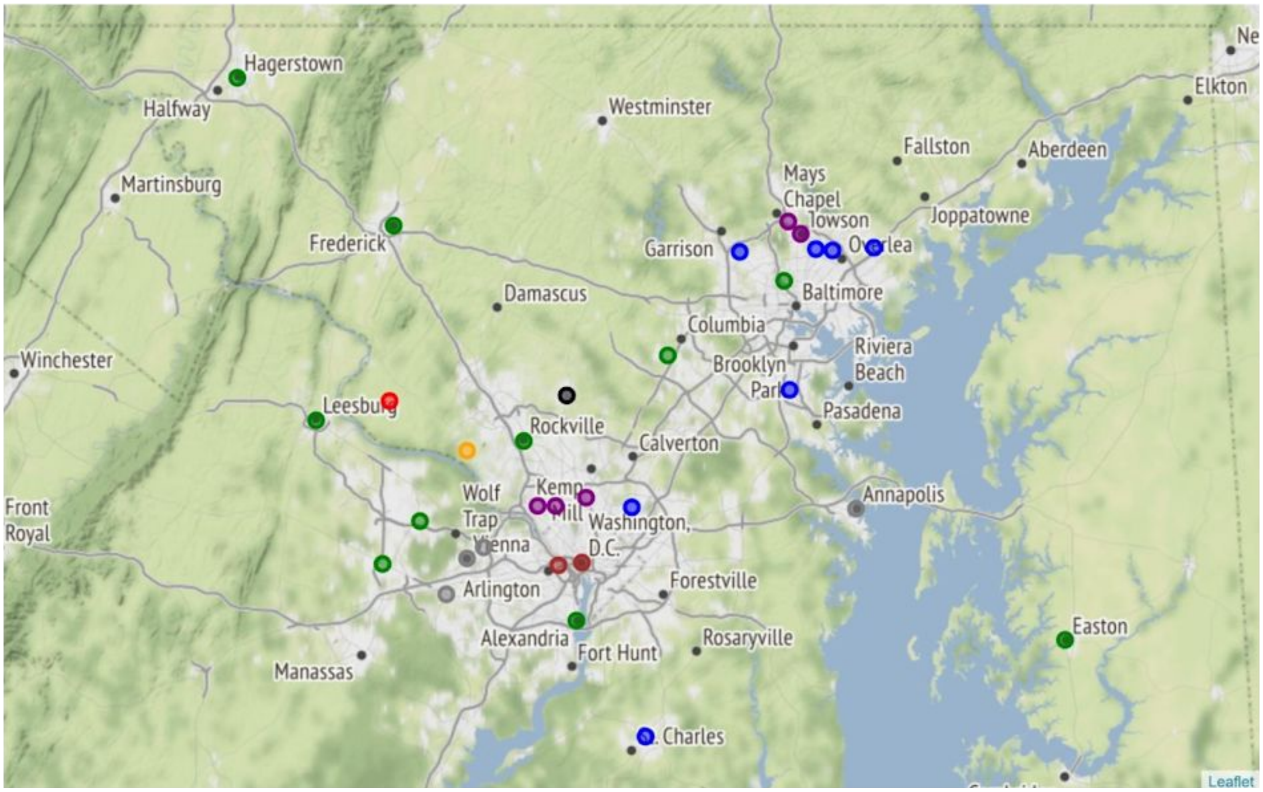
To cluster the cities together based on the similarity of their venues, I chose to use the K-means algorithm. This algorithm creates start points (initial centroids) that iteratively move to create clusters of data points. Eventually the centroids reach a local optimum such that the data points within the cluster are "closer" to points in the same cluster and "farther away" from points in other clusters. If you're confused, that's okay. The algorithm itself does most of the work, but we do have to choose the number of clusters to use and (preferably) the random state that the algorithm generates these initial centroids.

To choose these options I used something called the elbow method where I looked at graphs of within cluster variation for different numbers of clusters to determine the optimal number of clusters to use. I ended up using 8 clusters and a random state value of 4 because it produced the most clear elbow on the graph (seen below).



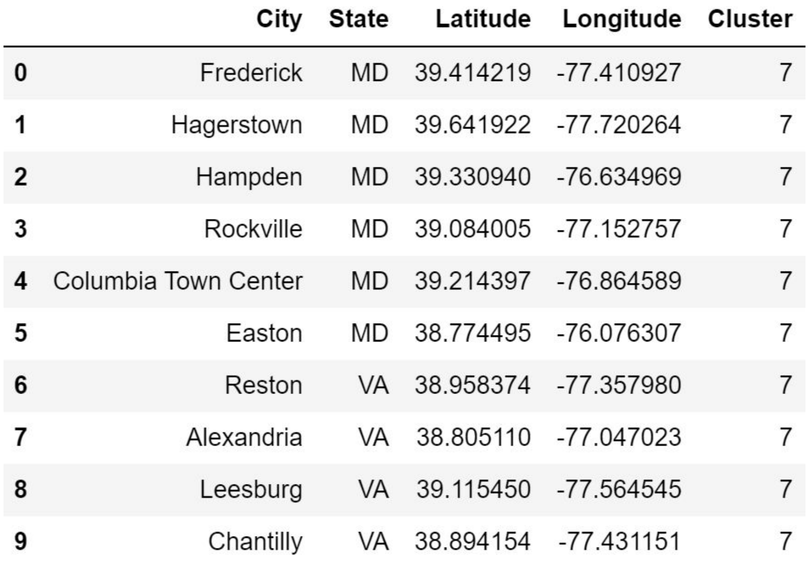
Clustering the Cities

Once I had the parameters chosen for the K-means algorithm, I was able to settle on my final model for analysis. I ran the algorithm one more time with 8 clusters and a random state value of 4 and the result had the cities clustered as seen below by the different colors!



Results and Analysis

I was interested in finding cities that are similar to Frederick, so I am interested specifically in the cities that were clustered with Frederick. The “Frederick Cluster” as I call it, was made up of Frederick plus 9 other cities. The cities can be seen in the table below.



	City	State	Latitude	Longitude	Cluster
0	Frederick	MD	39.414219	-77.410927	7
1	Hagerstown	MD	39.641922	-77.720264	7
2	Hampden	MD	39.330940	-76.634969	7
3	Rockville	MD	39.084005	-77.152757	7
4	Columbia Town Center	MD	39.214397	-76.864589	7
5	Easton	MD	38.774495	-76.076307	7
6	Reston	VA	38.958374	-77.357980	7
7	Alexandria	VA	38.805110	-77.047023	7
8	Leesburg	VA	39.115450	-77.564545	7
9	Chantilly	VA	38.894154	-77.431151	7

Examining the breakdown of venue categories for the different clusters, showed me that the Frederick Cluster cities were similar because they had a wide variety of venues and not just one prominent category like some of the other clusters. This variety of venues available is something that we enjoy about Frederick, so it is nice to see that the other cities in the cluster have this in common. Another part of Frederick that we enjoy, is the accessibility to parks in the area. We like to walk a lot and we want our new neighborhood to have areas for us to explore as well.

I used the Foursquare API again to search for parks and trails around the 9 cities in the Frederick cluster. The results showed me that most cities on the list have equal access or even more access to parks in comparison to Frederick.

Conclusion

The project as a whole helped me learn a lot about working with location data, using the K-means algorithm, and solving problems with data science skills. Using the clustering algorithm helped me pare down our list to 9 other similar cities. From these cities I was able to find specific ones that had the same number of parks around them as Frederick, or more! Once I had these final options, I was able to choose the option closest to my girlfriend’s graduate program and family. That city is…Hampden! I’m so glad to have been able to solve a problem in my life with these skills! We’re still a couple months out from our move, but we’re excited to be looking into Hampden.