Capstone Project - Battle of the Neighborhoods

May 2020

Content

What Problem Are We Trying to Solve?

Business Problem

- Amazon HQ2 will be deployed in Arlinton, in the DC area, and a large amount of new employees are expected to move in.
- Arlington is one of the best areas for living in the US.
- However, the area already has problems of traffic and cost of living.
- We want to analyze the Greater DC urban area and, using the commute time, housing costs, and amenities availability, identify other areas that may be good alternatives for prospective Amazon employees.

Data Acquisition and Cleaning

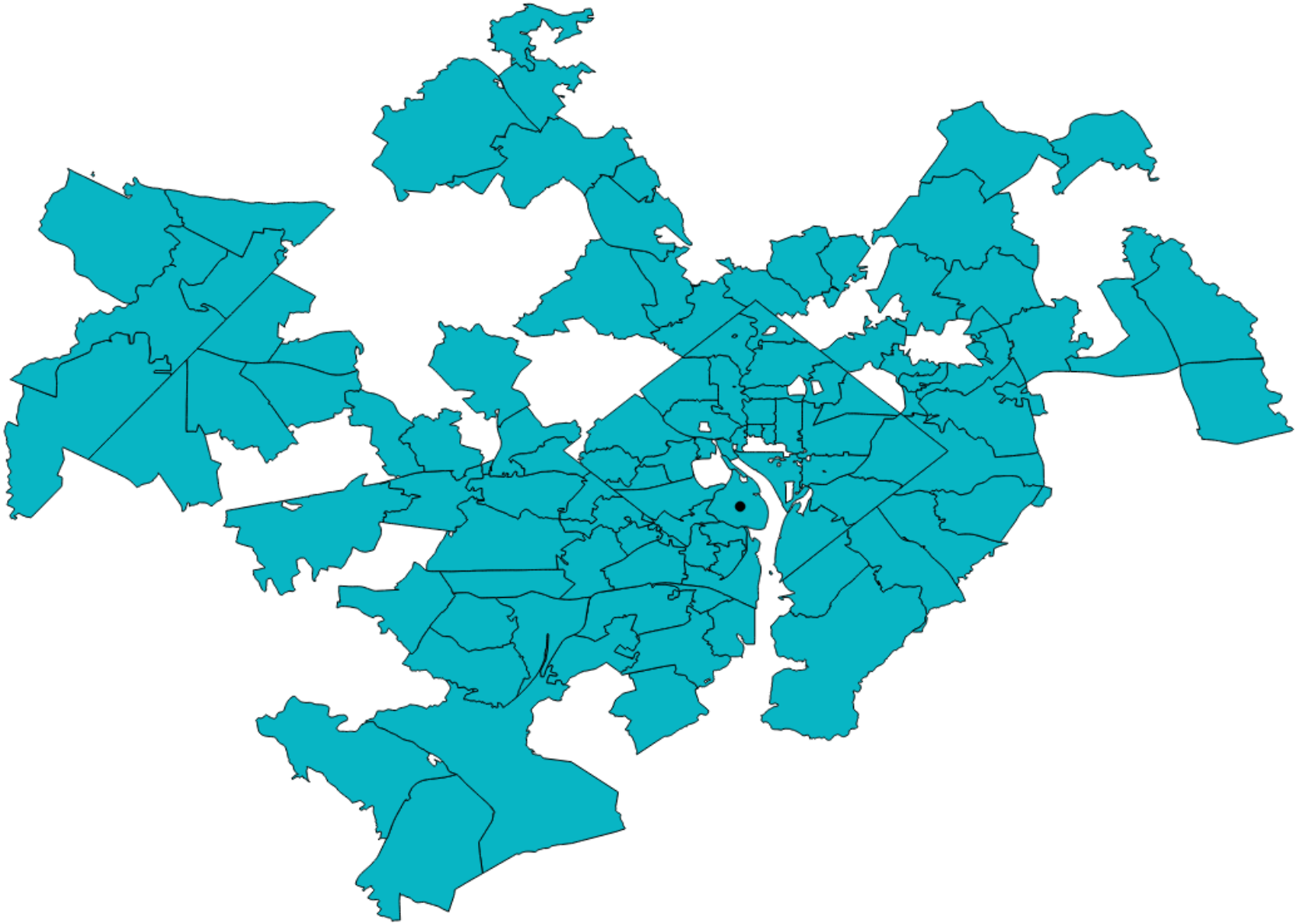
Data Acquisition

- There is not a definition of ‘neighborhood‘ that we can use reliably for this analysis.
- The Zip Code Tabulation Areas (ZCTA) defined by the US Census Bureau provide a good alternative, and we can acquire the dataset from the publicly available datasets in the US Census Bureau website.
- The costs of living can be obtained through Zillow, and are linked to Zip codes. There are multiple indexes for different types of homes.
- Commute times from each ZCTA to the National Landing site can be obtained by using Bing Maps API. We gathered the commute time at several times of the morning, to account for the morning rush hour.
- Finally, the venues and amenities availability was provided by FourSquare. Using their API we acquired a list of up to 100 venues per area.

Data Cleaning

- There were too many areas to consider: almost 1400 areas to cover Maryland, DC, and Virginia.
- We restricted the analysis to those areas that had a commute time by car under one hour.
- Additionally, the availability (or lack of) of housing and venue information for some areas also trimmed down the areas in the analysis.
- In the end, 102 areas were part of tha analysis.

Areas for Analysis

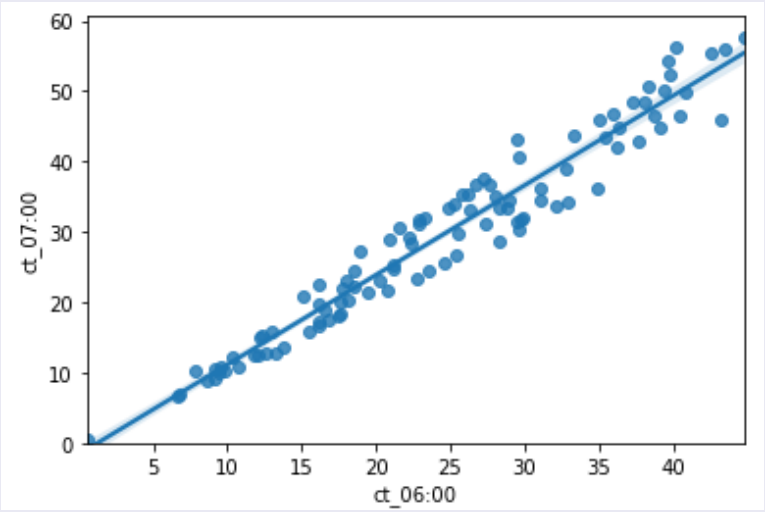


Exploratory Analysis

Correlation Between Commute Times

We found a strong positive correlation between all the commute times collected. This allowed us to replace all the features with a single one to ease the analysis. The correlation is stronger for short commutes (low values) and weaker for longer commutes.It is also stronger for commute times that are closer.

Correlation between 6:00 and 7:00 commute times



Correlation Between Housing Indexes

For housing prices we found that there were two groups. Within each group (identified by Single Family Homes and Multi-Family Homes) the correlation is significant to strong, but the correlation between groups is very low.

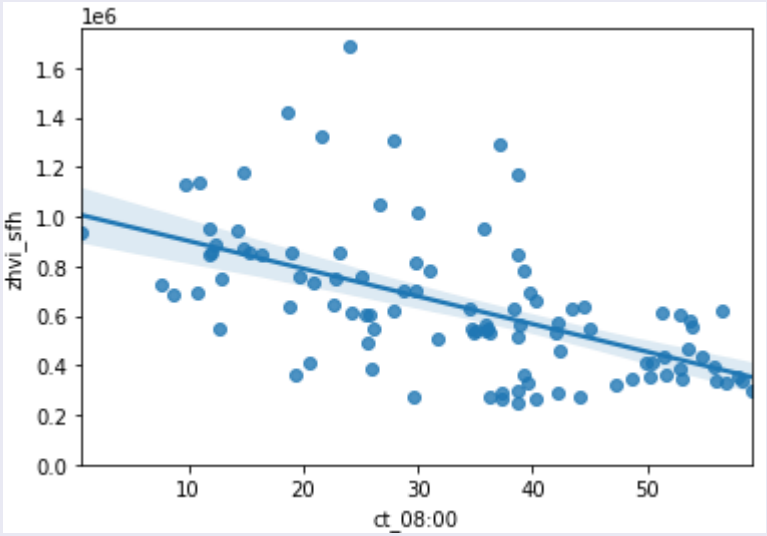
Coefficients and P-Values of single family homes, with the rest of types of homes

	Corr.Coeff	P-Value
MFH	0.6123	$7.98e^{-12}$
1Bed	0.4814	$3.03e^{-07}$
2Bed	0.7802	$4.24e^{-22}$
3Bed	0.8769	$1.41e^{-33}$
4Bed	0.9001	$5.77e^{-38}$
5Bed	0.9290	$5.94e^{-45}$
Rent	0.7994	$7.50e^{-24}$

Correlation Between Commute Times and Housing Indexes

The commute time did not have a strong correlation with the housing indexes. While it was negative(as expected), there are other factors than distance to the DC area that significantly affect the housing prices.

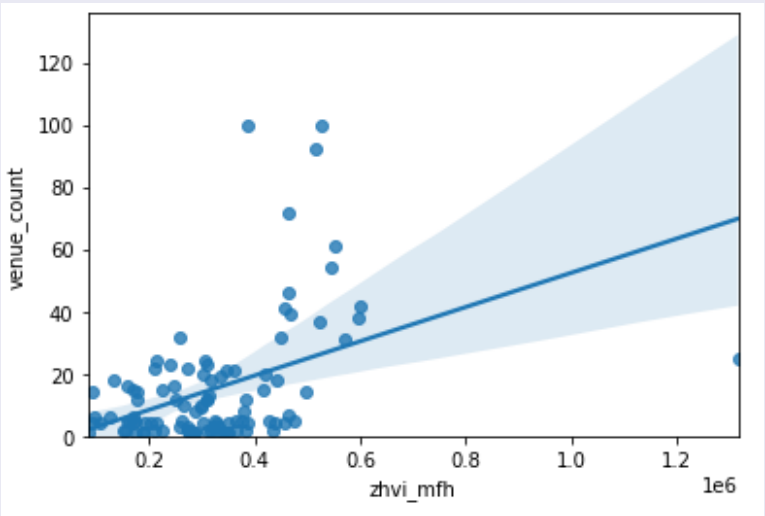
Correlation between commute times and housing indexes



Correlation Between Housing Indexes and Venue Availability

Similarly, the housing prices did not have a strong correlation with the availability of amenities and venues. In fact, these were some of the lowest values of correlation that we obtained throughout the whole study.

Correlation between housing indexes and amenities availability

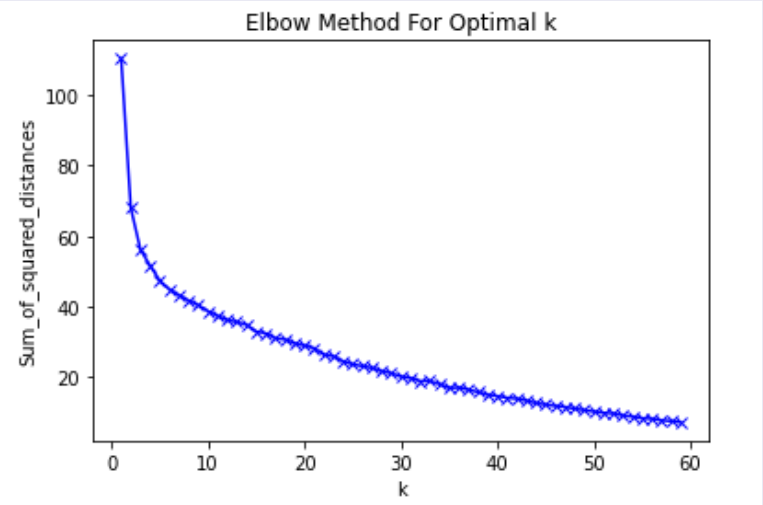


Clustering Analysis

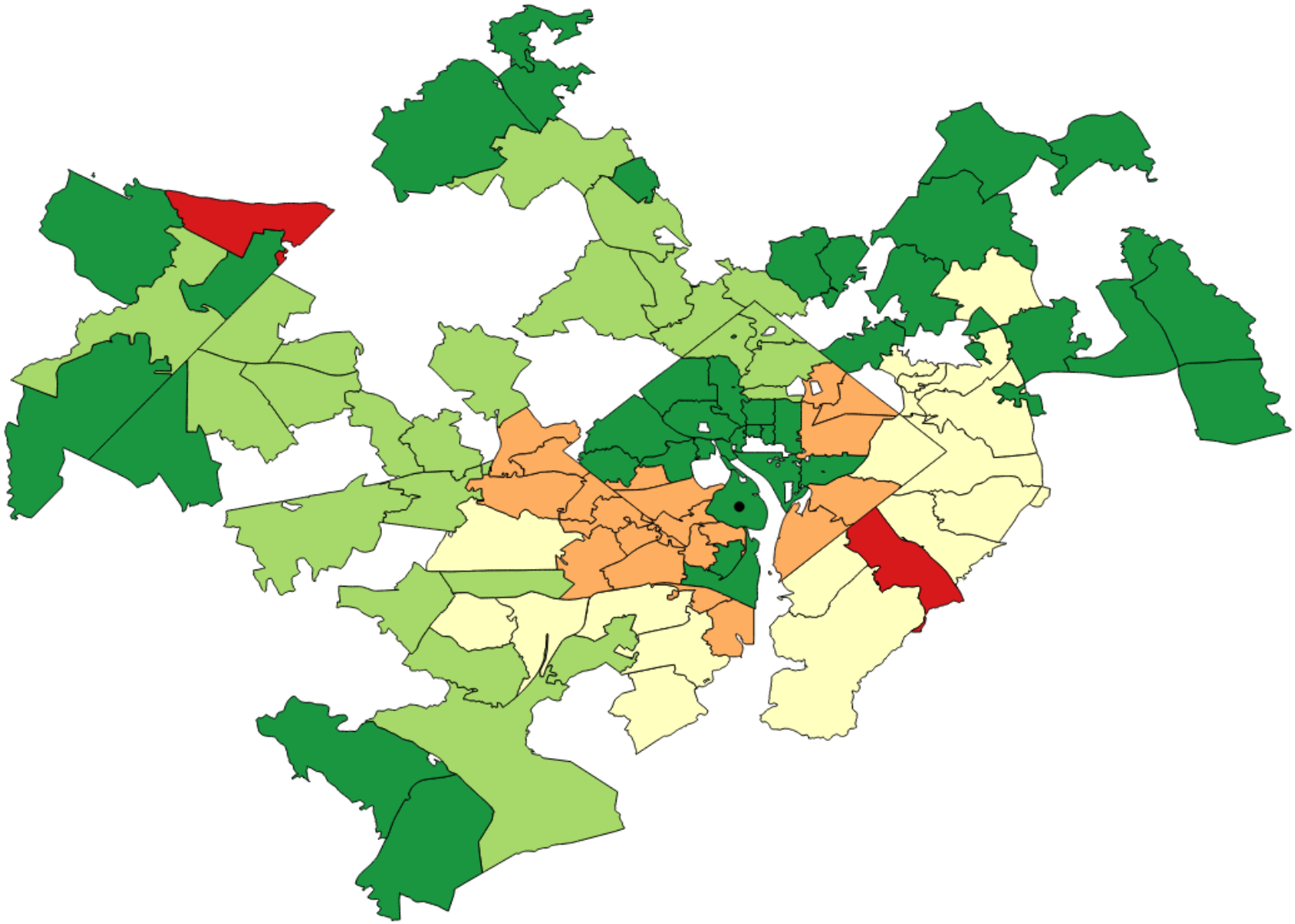
KMeans Clustering

KMenas had a problem in that when looking for the optimal number of clusters, there was not a clear candidate. The curve did not display a clear ‘elbow‘, so we performed the analysis with 6 and 30 clusters. 6 clusters was not enough, as some weird pairing of areas in the same cluster were clearly visible. 30 clusters provided much better results.

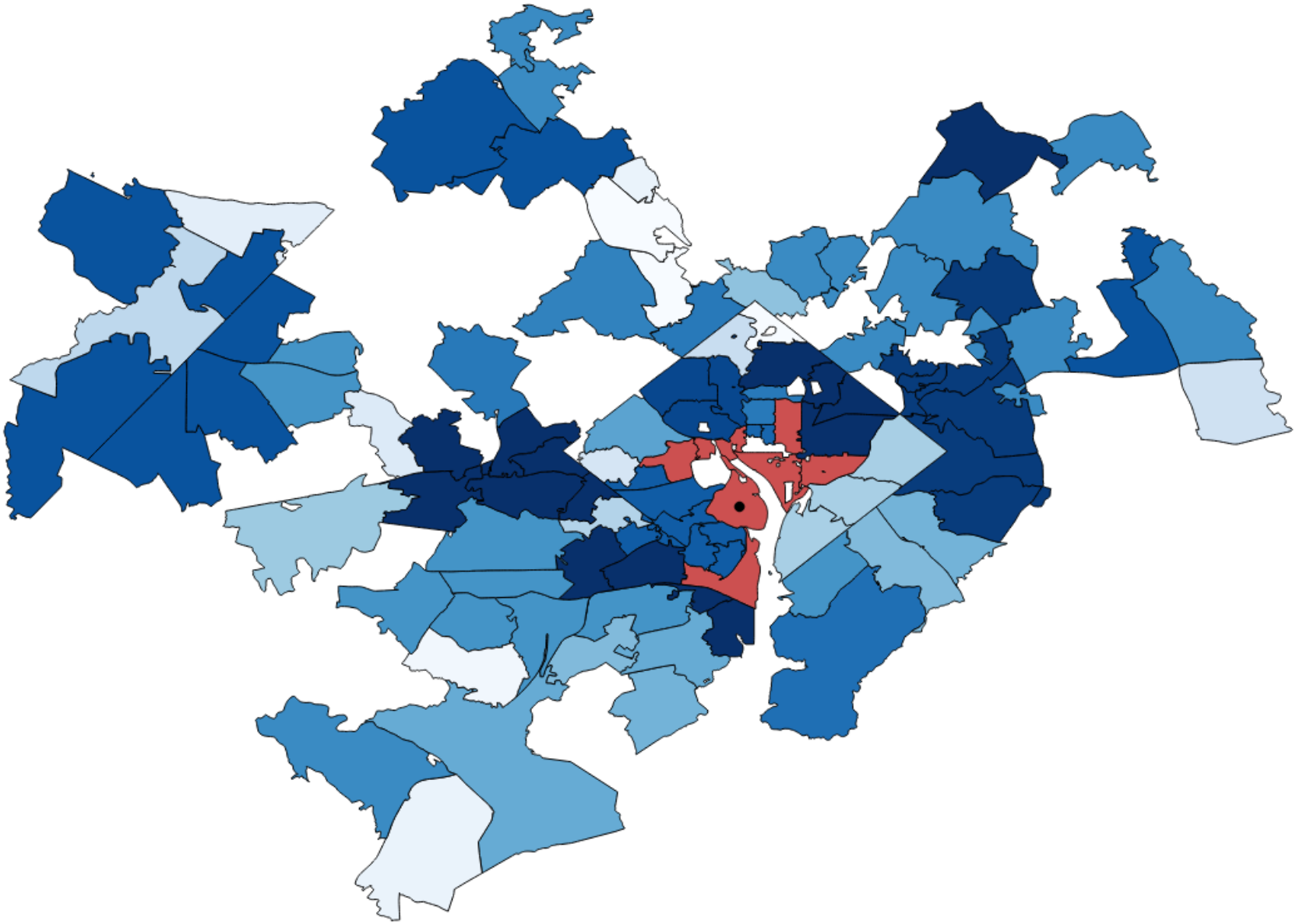
KMeans inertia with the number of clusters



KMeans Clustering with 6 Clusters

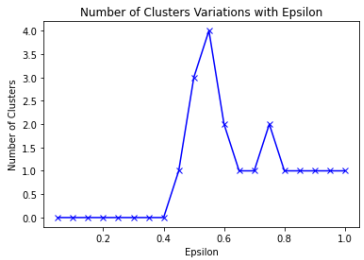


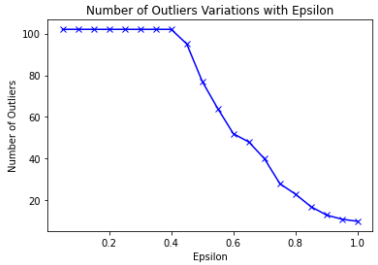
KMeans Clustering with 30 Clusters

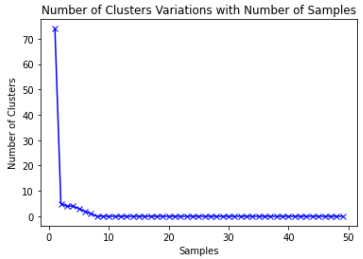


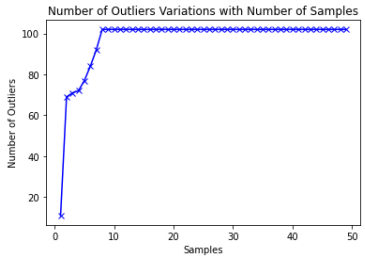
DBSCAN Clustering

- DBSCAN had trouble clustring these areas. We could not find combinations of epsilon and the minimum number of nodes in a cluster that would yield good results.
- In the best case we could achieve, most of the areas were considered outliers, including the area containing the National Landing.

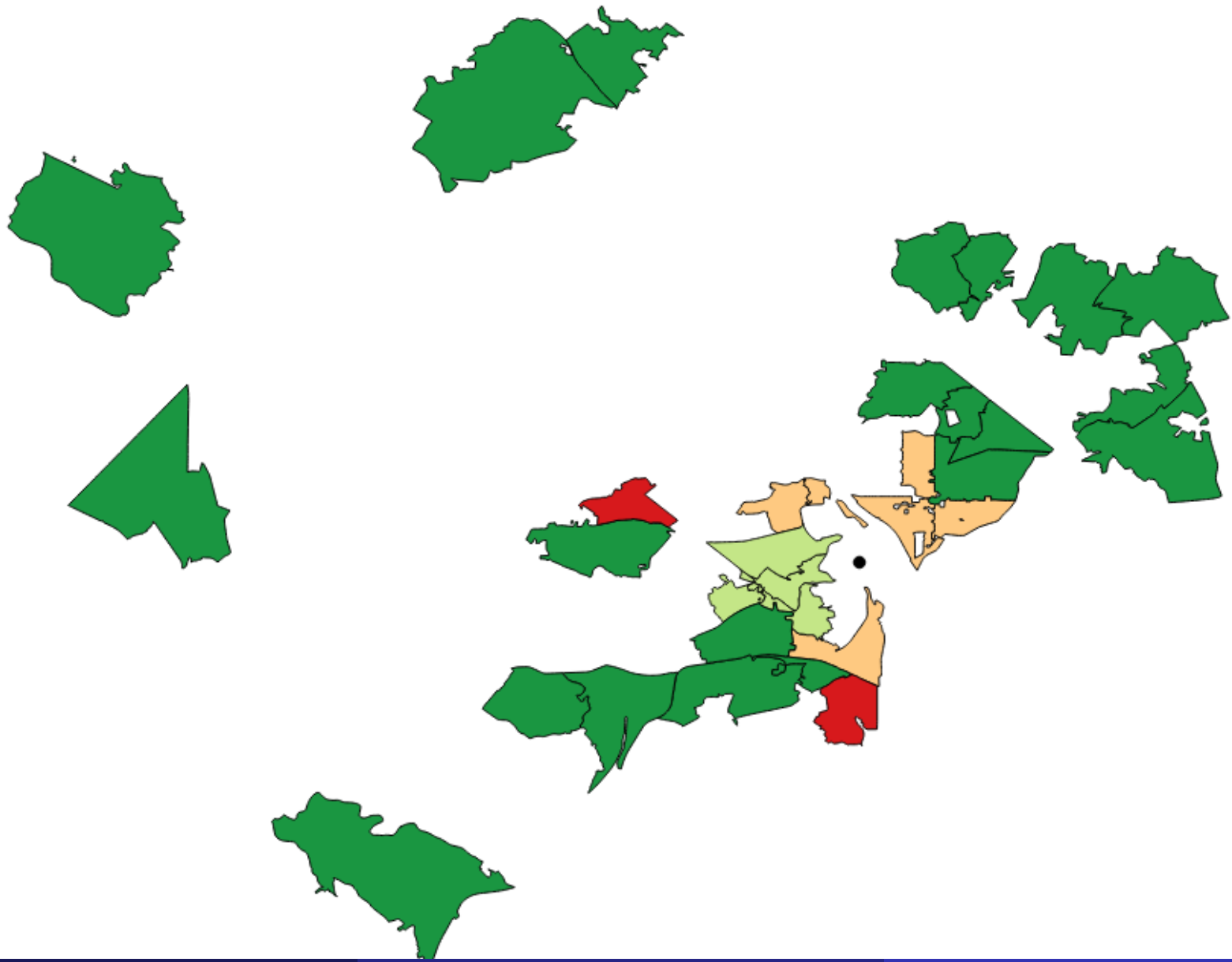








DBSCAN Clustering



Conclusion and Next Steps

Conclusion and Next Steps

- The clusting with KMeans has worked significantly better than DBSCAN, especially with a higher number of clusters.
- It is possible that one of the areas of interest (amenities and venue availability) has become dominant over the rest.
- Model can be extended with other metrics and features that are of intereset, like distance to emergency services.
- Additionally, other ‘distance‘ metrics can be used to compare with the current set of results.