Problem and Background

If someone wanted to move from one neighborhood in Washington DC to a similar neighborhood, what are their options? Washington DC is a very diverse city and the capital of the United States. The city is broken up into 8 wards numbered 1 -8, and is comprised of 131 neighborhoods. Washington DC has a competitive job market, with opportunities all over the city. Business professionals often like to live within walking/biking distance of their job. If a person accepts a job offer in a different ward, and would like to move closer to their new office building, but likes the characteristics of their current neighborhood, it would be desirable to identify similar neighborhoods in the new area.

Data

Wikipedia has a page that lists the 8 wards of DC and the neighborhoods belonging to each ward (https://en.wikipedia.org/wiki/Neighborhoods_in_Washington,_D.C.). I will scrape this data from the webpage and use the Nominatim library to enrich that data with the latitude and longitude of each neighborhood. This data will be used to query Foursquare to gather data about popular venues in each neighborhood. That data will be used to segment and cluster the neighborhoods allowing identification of neighborhoods with similar characteristics.

Methodology

After retrieving the data, I performed some data wrangling to get it into an appropriate format for further analysis. I dropped rows with empty values and reformatted some neighborhood names to enable retrieving the geographic coordinates using the Nominatim library. I generated some descriptive statistics to get a general understanding about the distribution of neighborhoods in the 8 wards. Categorical variables were converted to “dummy” variables. I selected k-means clustering to segment and cluster the neighborhoods because it is fast and robust. The “elbow method” was used to determine the optimal number of clusters. The clusters were then manually analyzed to determine the discriminating venue categories that distinguish each cluster.

Results/Discussion

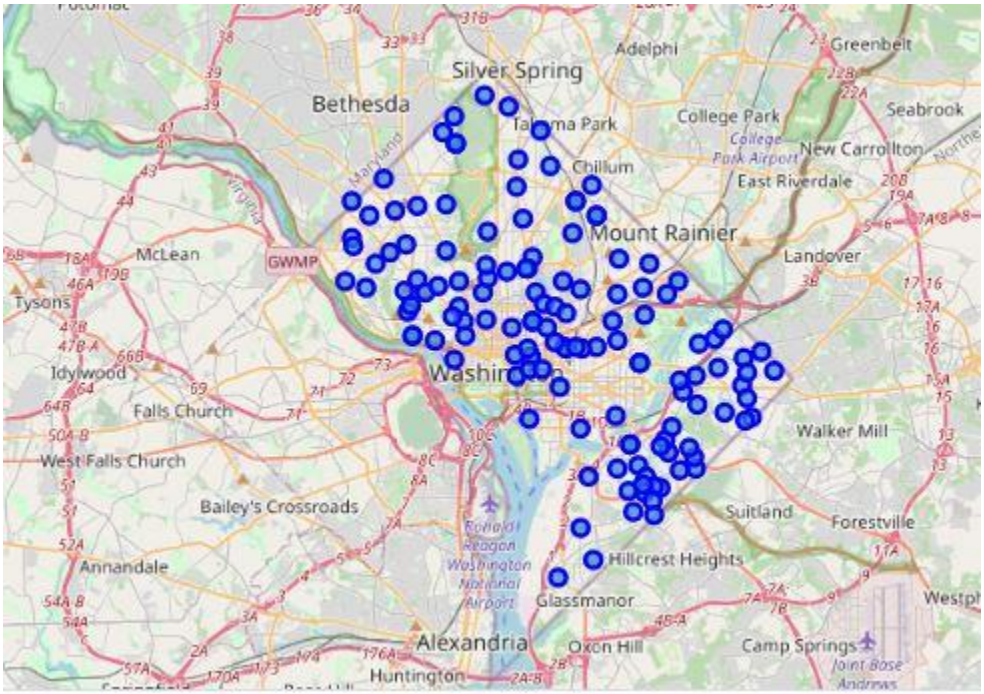

Figure 1. DC Neighborhoods

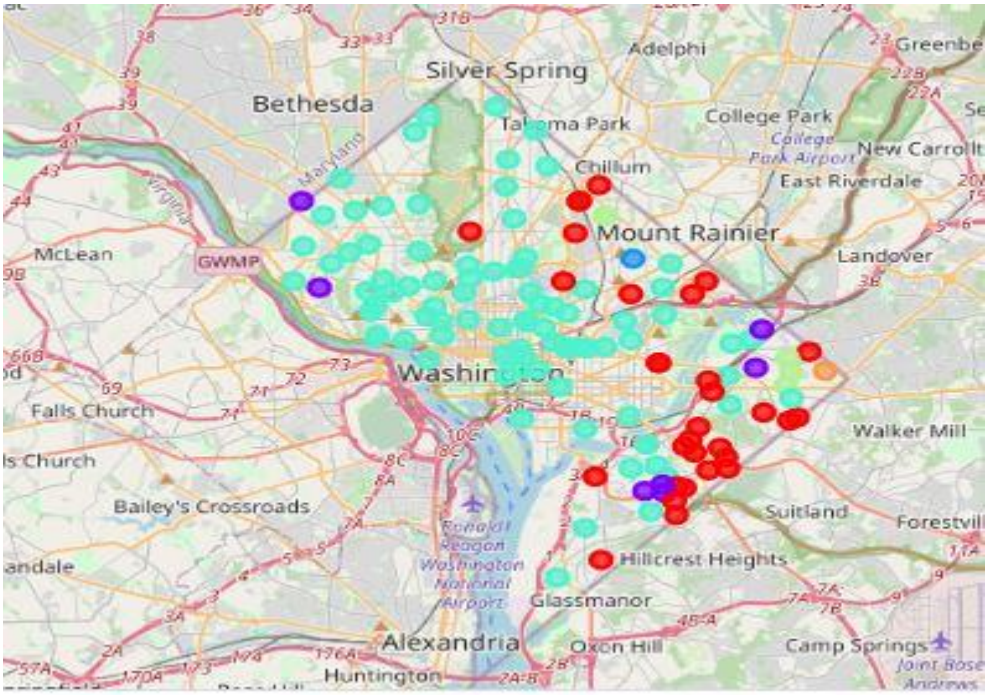

Figure 2. DC Neighborhood Clusters

Most neighborhoods lied in one of the following three clusters:

- A cluster where convenience and liquor stores were the most popular
- A cluster where parks and playgrounds were very popular, and
- A cluster where coffee shops were very prevalent

Three neighborhoods were so unique they fell in their own cluster. The three clusters that most neighborhoods fell into were geographically dispersed, so if someone's current neighborhood was in one of those clusters, they could find similar neighborhoods in different parts of the city.

Conclusion

Unsupervised machine learning techniques, specifically clustering, is a viable technique to segment and cluster neighborhoods to enable identifying similar neighborhoods in different locations throughout Washington, D.C. This information will be very useful for those business professionals who take new job opportunities across the city, but like to live within walking distance of their employer.