**A Comparison of Locations in the Washington DC Metro Area**

A Report for the Applied Data Science Capstone course on Coursera by IBM

**Introduction/Business Problem**

Company ABC's office is located in Alexandria, VA, just outside of Washington DC. The company is happy with the location of its office, however now that their office lease is ending and the company has recently grown in size, they are in a position where they need to leave their current office space. While looking for new office space, they do not want to be limited by only what is available in Alexandria, so they want to look into what nearby areas would be similar as other options. These locations need to be within the Washington DC Metro area, to ensure that current employees will not need to relocate to be close to the office.

Company ABC’s requirements are as follows:

(1) They want the area to be similar to Alexandria, VA.

(2) They are limiting locations to the DC Metro area (or the DMV, DC-Maryland-Virginia).

In order to solve this problem, we will leverage information about the types of locations in different areas around the DC Metro area to determine which locations will fit Company ABC's needs.

**Data**

In order to determine locations that Company ABC may want to consider, we will limit our search to counties in the DC Metro area. A list of counties in this area has been found [here](https://en.wikipedia.org/wiki/Washington_metropolitan_area#Composition). Additionally, we will need the geographic coordinates of each of these counties in order to leverage location data. We use [this table](https://en.wikipedia.org/wiki/User:Michael_J/County_table) of US counties, which includes coordinates, to get this information. This information will be combined into a single dataframe that can be used to get more specific location data. We will use the Foursquare API to find nearby venues to each of the county coordinates we have obtained. This will be used to solve our initial problem by determining which counties are similar in terms of types of venues.

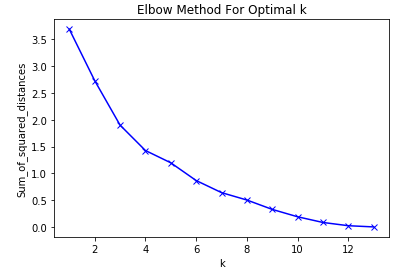
**Methodology**

**Data Manipulation**

To compare our locations, we took the following steps. First, we used the Foursquare API to obtain information on a maximum of 100 venues nearby to each county’s coordinates. Using one hot encoding, we manipulated the venue data so that each line represents a single county and location, with all location types as possible columns. This dataframe was then grouped by county, so that each row contains a county and each venue type column contains the proportion of venues that fit that category in that county. From this point, we determined the top 10 most common venue types for each county. Counties were removed from the dataset if the Foursquare API found no venue data.

**Analysis**

Using the data on the ten most common types of venue for each county, we will begin our analysis. Because we want to determine which counties are similar to Alexandria, VA, we need to cluster our data. To do this, we will use the k-means clustering algorithm. This method requires dividing the data into k clusters, which means we need to determine the optimal k value. We used the Elbow Method for determining our k-value, choosing the k-value where the graph bends at an “elbow.” Based on our graph (Figure 1), we chose a k-value of 5.



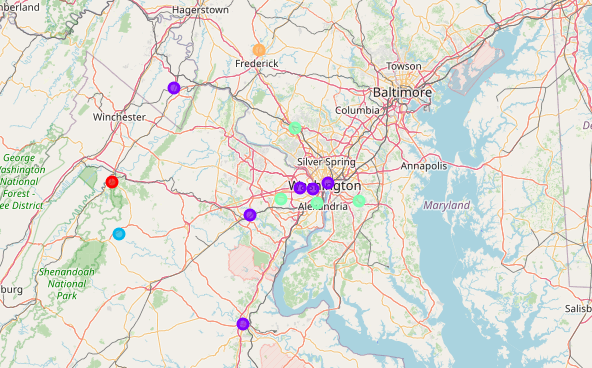
*Figure 1:* A graph of the Elbow Method for determining the optimal k value for k-means clustering

**Results**

Using a k-value of 5, we performed the k-means clustering algorithm to divide our data into five clusters. Figure 2 shows a map of the area with the counties marked, with different colors marking the clusters. We see 5 distinct clusters of differing sizes; the content of these clusters can be seen in Table 1.

|  |  |
| --- | --- |
| **County** | **Cluster Label** |
| Arlington, VA | 1 |
| Fairfax, VA | 3 |
| Rappahannock, VA | 2 |
| Warren, VA | 0 |
| Alexandria, VA | 3 |
| Falls Church, VA | 1 |
| Fredericksburg, VA | 1 |
| Manassas Park, VA | 1 |
| Frederick, MD | 4 |
| Montgomery, MD | 3 |
| Prince George’s, MD | 3 |
| District of Columbia, DC | 1 |
| Jefferson, WV | 1 |

*Table 1:* County clusters



*Figure 2:* Results of k-means clustering on counties

**Conclusion**

Company ABC asked us to determine possible locations for their new office space with the following two requirements: (1) the location should be similar to Alexandria, VA and (2) locations should be limited to the Washington DC Metro area. By getting venue data for the different counties in the specific area, we were able to compare the counties and perform a clustering algorithm to find what counties are similar to Alexandria, VA.

We had a total of 5 clusters. Cluster 3 (see Table 1) contained Alexandria, VA; Fairfax, VA; Montgomery, MD; and Prince’s George, MD. The counties in Cluster 3 fit the client’s requirements for location. Therefore, our recommendation to Company ABC is to limit their search to the counties in Cluster 3.