Devin Arrants
May 5, 2020

Clustering D.C. Metro Stations to Anticipate Demand for Growth

I. Introduction

The D.C. Metro is one of the busiest rapid transit systems in the United States, second only to the New York City Subway. As of May 2020, the network includes 91 stations, six lines and serves hundreds of thousands in the surrounding areas of Maryland and Virginia. The Washington Metropolitan Area Transit Authority, WMATA, anticipates an average of one million riders daily by 2030. Due to the increase in population and commuters, the WMATA has been focused on extending service, building new stations, and constructing additional lines to alleviate congestion.

This necessity for growth can either be an obstacle that hurts a city fiscally and does not meet the demands of the citizens, or it can be an opportunity to expand efficiently, thus decreasing congestion, and easing daily commutes as the population grows. Rather than grow blindly, cities can use the growing arsenal of data analytic techniques to anticipate demand.

To address this need, I propose an examination and classification of metro stations in D.C. based on surrounding venues. Leveraging the Foursquare API, we can explore the venue types surrounding each station, thus allowing us to make a model that clusters stations based on their primary usage. City planners can begin to predict the demand of people traveling to and from this station. For example, if the station is in a largely residential neighborhood, it can be presumed that there is a necessity for transit to commercial and professional neighborhoods. Thus, this beginning exploratory analysis of venues surrounding metro stations, can aid in the prediction process of where people will need to travel, in the future.

II. Data

For this analysis, location of each metro station, and the venues surrounding it are needed. The name and geographical location are obtained using the WMATA API. There is additional information, such as: lines served e.g. red, blue, silver, the kind of platform at the station, and the address. This information was not necessary for this initial examination of the metro stations, thus it was dropped. Additionally, if the station served more than one line, it was repeated in the json file. Therefore, duplicate rows were removed for the purposes of this analysis.

Using the latitude and longitude, the Foursquare API is used to determine the classification of surrounding venues. Foursquare separates a venue by ten large classifications that can be further refined. For this initial exploration, the more broad classifications suffice. These classifications include: Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, and Travel & Transport. Every venue within a 500 m radius

of the metro station, a reasonable walking distance without approaching another station, was placed into one of these categories, and then counted, to create a data table (Figure 1) that contains all of the metro stations in DC.

	Station	Lat	Lon	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Metro Center	38.898303	-77.028099	33	22	1	39	36	53	104	6	178	78
1	Farragut North	38.903192	-77.039766	25	32	0	49	41	55	109	5	192	81
2	Dupont Circle	38.909499	-77.043620	20	20	3	86	53	53	128	26	184	67
3	Woodley Park-Zoo/Adams Morgan	38.924999	-77.052648	3	3	0	33	9	14	29	14	24	17
4	Cleveland Park	38.934703	-77.058226	4	1	0	22	4	7	18	11	30	12
5	Van Ness-UDC	38.943620	-77.063511	4	0	1	11	0	13	55	7	21	8

Figure 1. A sample of the categorical counts per station data frame

III. Methodology

Initially, the data is retrieved and cleaned, as previously explained. Using the geographical location of each station, it is possible to visualize the distribution of stations in the D.C. area, as seen below (Figure 2):

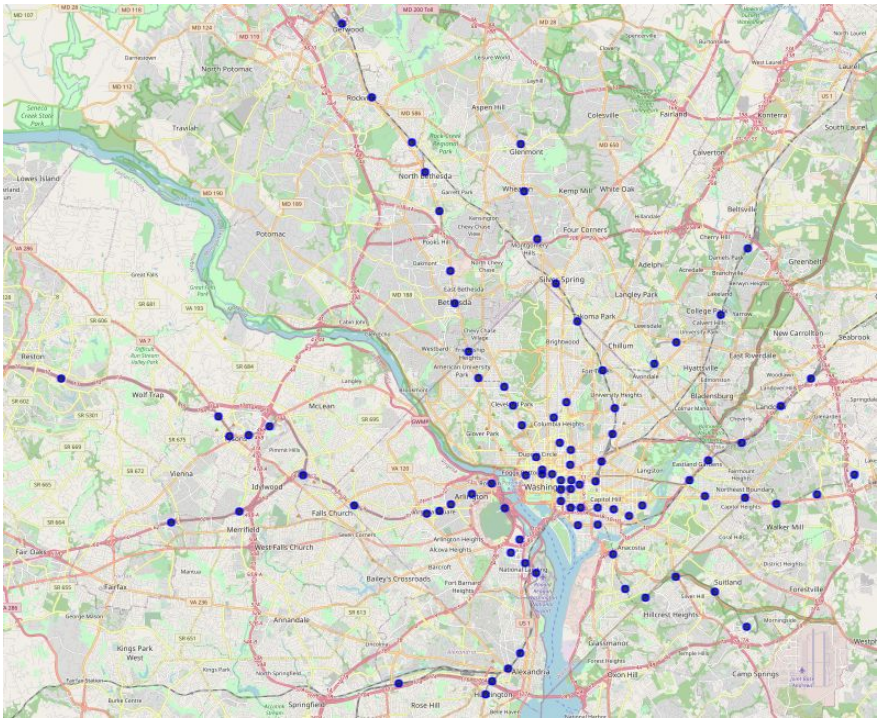


Figure 2. WMATA Metro Stations location

This map is an interactive folium map that can be found on the github attached to this report.

When a station is pressed on, the name and location of the station will pop up in a label.

The greatest concentration of stations lies in D.C. and the immediately surrounding areas. The further away from downtown, the more spread out the metro stations become.

Using the Foursquare explore API, with categoryID to indicate which of the ten broad categories looked at for each station, a data frame including the station name, geographical coordinates, and the total count of each category in a 500 m radius of the station is obtained (Figure 1). The counts were as expected: Foggy Bottom, the location of the George Washington University, has the highest count of college and university related venues at 117; Rosslyn station has the most professional venues in the 500 m radius, and Dupont Circle has the most residential venues. These three stations are in the heart of DC or just outside of it in Virginia. The further away from downtown D.C. the less concentrated the areas are, thus the less venues of each category are found in those areas, like the Vienna station.

In order to explore the data further, a box chart is built to visualize distribution of venue categories (Figure 4). The box plot is overlaid with a swarm plot that shows the actual distribution. From this it is clear that most metro stations have a lower count of venues, however, the few metro stations in highly dense areas skew the mean. The right skew makes for longer right tails on the plots indicating that there is greater variance in the stations with more venues than the median, however, the bulk of stations falls on the lower end of venue counts. Additionally, we can see that shops, professional, and food venues have the highest count in the area surrounding metro stations in the DC, Virginia and Maryland area.

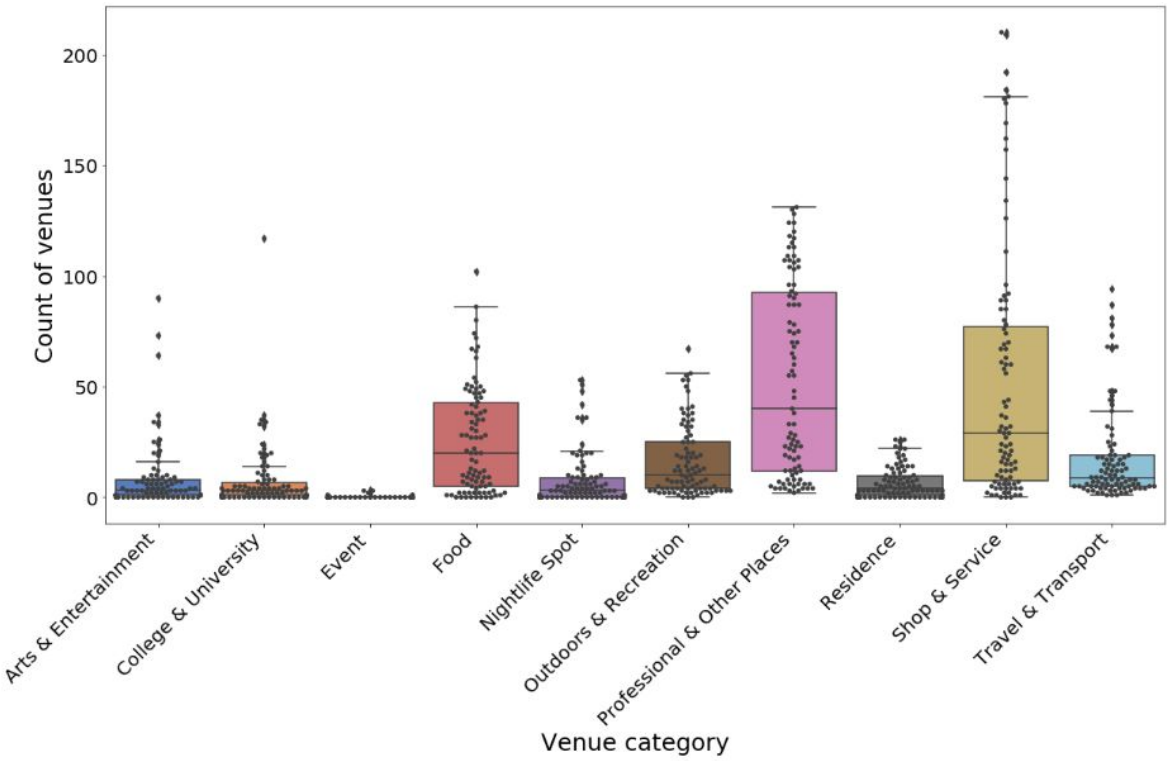
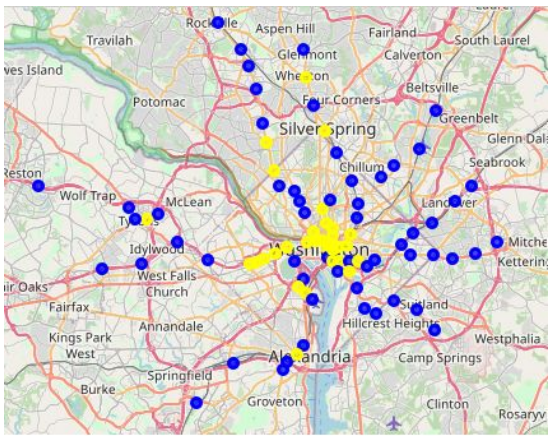


Figure 4. Box Plot Overlaid With a Swarm Plot to Visualize Distribution of Venue Categories

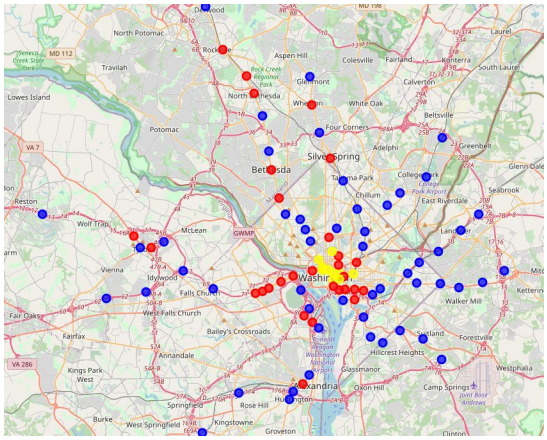
The event category is seemingly insignificant due to such small relative counts, therefore, it was dropped from the venue categories before clustering began. In order to cluster together similar metro stations to classify them, KMeans model from sklearn is imported. This model was chosen over other clustering methods because KMeans is more general-purpose, works best with fewer clusters, and is best with flat geometry. Before fitting the model, I normalized the data using a MinMaxScaler so that each

category's count is between zero and one, because the features have different ranges. The MinMaxScaler will preserve the shape of the original data.

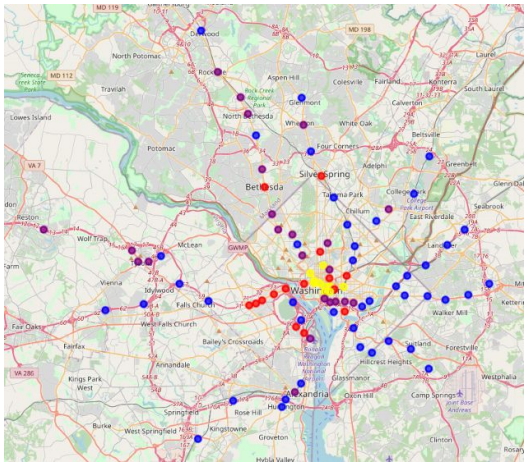
Beginning with two clusters, and working up to five clusters, this is how the various KMeans models clustered the stations:



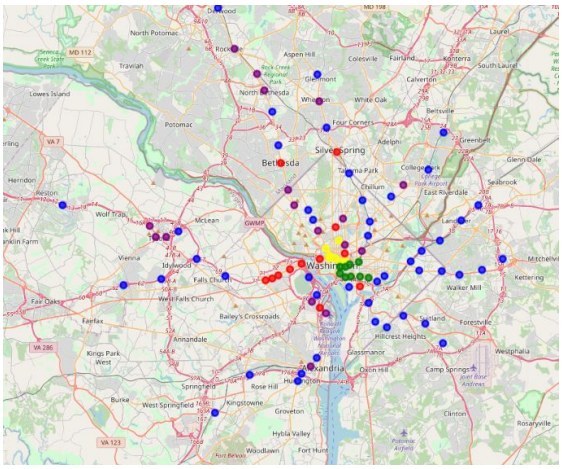
Two Cluster Model



Three Cluster Model



Four Cluster Model



Five Cluster Model

The model using two clusters appears to delineate the downtown areas that have a high density in venues from the more suburban areas with a lower density in venues. The model using three clusters further divides downtown. The yellow represents the immediate downtown, with a high concentration of venues, particularly in the shops and food categories. Four clusters further divide the more urban areas. Beyond four clusters, how the machine is delineating clusters becomes too complex for the purposes of this study. The four cluster model appears to cluster at the most informative level before becoming difficult to comprehend, therefore, four clusters are used in the final analysis (Figure 5).

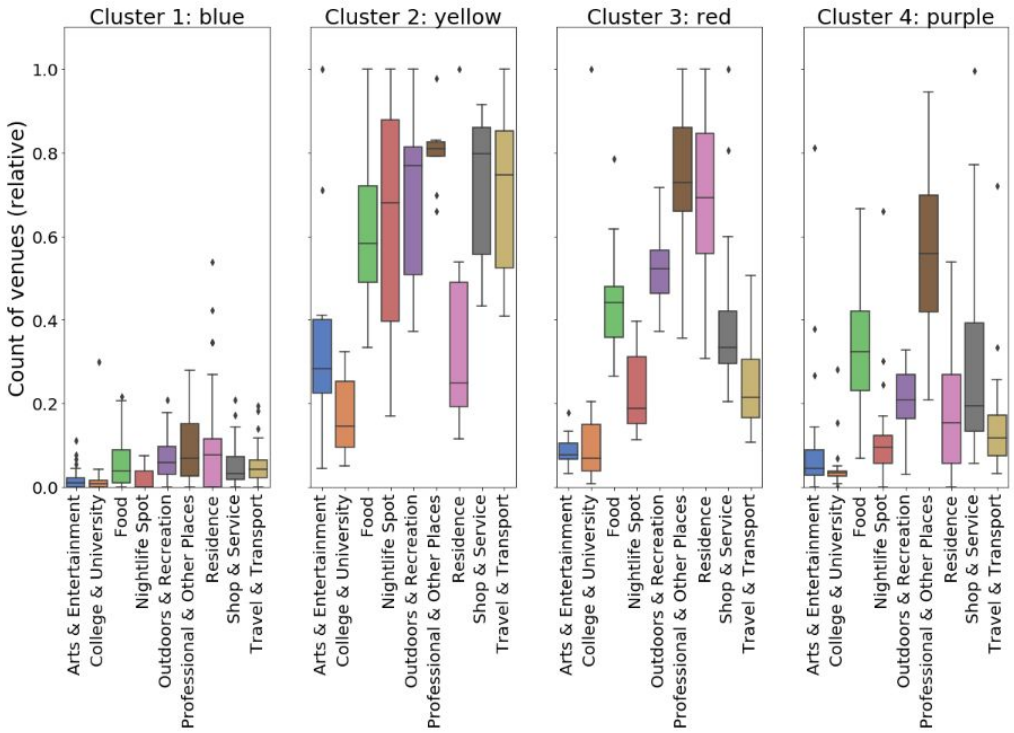


Figure 5. Division of Clusters Visualized with Box Plots

By looking at Figure 5, more in depth conclusions involving what constitutes each cluster can be drawn. These will be discussed in the results section. In addition to the visualized results, I listed the top five venues from the normalized data for each metro station and organized the resulting data frame by the cluster number for the more linguistically inclined minds, like myself.

	Station	Lat	Lon	Cluster	First Most Common Venue	Second Most Common Venue	Third Most Common Venue	Fourth Most Common Venue	Fifth Most Common Venue
45	Eastern Market	38.884124	-76.995334	4	Professional & Other Places	Food	Shop & Service	Outdoors & Recreation	Nightlife Spot
24	Wheaton	39.038558	-77.051098	4	Shop & Service	Food	Professional & Other Places	Residence	Outdoors & Recreation
89	Spring Hill	38.929273	-77.241988	4	Professional & Other Places	Food	Outdoors & Recreation	Residence	Shop & Service
88	Greensboro	38.919749	-77.235192	4	Professional & Other Places	Shop & Service	Food	Residence	Outdoors & Recreation
87	Tysons Corner	38.920056	-77.223314	4	Shop & Service	Professional & Other Places	Food	Nightlife Spot	Outdoors & Recreation
59	Prince George's Plaza	38.965276	-76.956182	4	Shop & Service	Food	Professional & Other Places	Residence	Outdoors & Recreation
57	Georgia Ave-Petworth	38.936077	-77.024728	4	Residence	Professional & Other Places	Outdoors & Recreation	Food	Shop & Service
54	Shaw-Howard U	38.912919	-77.022194	4	Nightlife Spot	Professional & Other Places	Food	Shop & Service	Travel & Transport
44	Capitol South	38.884968	-77.005137	4	Professional & Other Places	Food	Outdoors & Recreation	Nightlife Spot	Travel & Transport
43	Federal Center SW	38.884958	-77.015860	4	Professional & Other Places	Food	Outdoors & Recreation	Arts & Entertainment	Travel & Transport

Figure 6. A Slice of the Data Frame Showing Venue Category Rank for Each Station

IV. Results

Through an examination of the above maps and plots, a clearer understanding of the clusters characteristics arises. Cluster 1, the blue dots, signify regions with few venues surrounding the station. These neighborhoods are less densely populated, with the main demand for the station being residential areas and professional areas. Cluster 1 stations serve the suburban neighborhoods in the D.C. and surrounding areas.

Cluster 2, the yellow dots, lie right in the heart of downtown D.C., thus understandably has the highest relative counts of nightlife, recreation, and food. In fact, cluster 2 has high relative counts for all venue categories except for residential areas. This hints at the fact that there is a demand from these stations to more residential neighborhoods.

Cluster 3, the red dots, have a high concentration of professional venues, but what characterizes this cluster appears to be the even higher concentration of residential venues. The vast majority of the red dots are located just outside of DC. This explains the higher density of residential areas, because the area just outside of D.C. is more affordable and there is more space. Due to this, it would make sense that cluster 2 and cluster 3 be efficiently connected because of demand. Those living in cluster 3 commute to cluster 2 for various venues, but return to cluster 3 for their apartments and homes.

The final cluster, cluster 4, the purple dots, is characterized as an in between cluster. These neighborhoods are not as spread out as cluster 1, but they are not as densely packed with venues as the other clusters. Shops and services appear to be the main common venues for this final cluster.

V. Discussion

Based on this initial examination of the WMATA metro stations, we can begin to understand demand to and from stations. The D.C. metro is actively expanding, thus understanding what the common venues surrounding each station are can allow for efficient expansion that benefits the most people. If a hypothetical new station were to be built, it could be added to the data to determine what cluster it falls into, and thus what the demand to and from that station is.

I propose an additional analysis, involving comparing the D.C. metro to some of the largest, most efficient rapid transit systems around the world, such as: Hong Kong, New York, and the London Underground. Ideally, there would be a second clustering where similar cities with efficient metro systems are grouped together. Then the WMATA can use the similar cities as examples for expansion. Seeing how other, larger cities expanded can aid the WMATA in refining the current D.C. system as population grows.

VI. Conclusion

The purpose of this project is to examine the primary usage of the metro stations in the DMV area in order to understand the demand to and from each station. This information is the beginning step in helping the WMATA determine the next best steps for growth. However, this is simply a starting point for city planners. This analysis allows for those working on expansion to more fully understand what they are working with and what they might need in terms of new stations or new lines. Further data and models must be used to continue to answer the question of how the metro system can expand most efficiently, meeting the demands of the citizens.