Virginia Hospital Escapes - A comparison of the venues surrounding Virginia hospitals

IBM Capstone Project

Author Micah C. Gray

1. Introduction

If you ever spent a night at a hospital, either as a visitor or staff member, you were probably grateful for any contact with the outside world. You might also have appreciated having venues nearby for food, magazines, flowers, or just a path with some fresh air. This report compares hospitals in the state of Virginia with respect to their surrounding venues, with the intent that your next hospital visit in Virginia is a little more freeing.

This analysis draws upon location data obtained from Foursquare.com in order to explore the diversity of venues surrounding the 200-plus hospitals in Virginia. I will attempt to discover the ten most common venue types surrounding each hospital. I will also cluster hospitals based on the types of venues that are nearby. Lastly, I will filter out hospitals with little or few options for nearby food, pharmacies, or nature walks.

2. Data

2.1 Data Sources

Data about the hospitals in Virginia, including the geocoordinates, city, and state, were obtained from http://www.lat-long.com/ while data about the venues surrounding the hospitals were obtained from Foursquare.com using a 1000 meter radius from each hospital's geolocation.

2.2 Data Cleaning

My first step was to obtain the hospital data. I performed a search of Virginia Hospitals on lat-long.com and got results in the form of a table that contained hospital name, feature type (hospital), county, and state. I was able to copy the table and paste it into an Excel spreadsheet. Latitutde and Longitude for each hospital were obtained one at a time and copied individually to new columns in the Excel spreadsheet. This was a manageable task given the size of my data (271 hospitals). Next, I saved the spreadsheet and uploaded it to my jupyter notebook in IBM's Watson studio as a pandas dataframe.

This is a sample of the hospital data:

	Name	Feature Type	County	State	Latitude	Longitude
0	A B Adams Convalescent Center	Hospital	Emporia (city)	VA	36.685705	-77.537758
1	A D Williams Memorial Clinic	Hospital	Richmond	VA	37.539869	-77.430261
2	Access Emergency Hospital	Hospital	Fairfax	VA	38.965666	-77.356930
3	Albemarle County Health Department	Hospital	Charlottesville (city)	VA	38.042083	-78.482789
4	Alexander W Terrell Memorial Infirmary	Hospital	Lynchburg (city)	VA	37.438475	-79.172247

Since I was only concerned with operating hospitals, my next step was to remove historical hospitals by dropping rows that had "historical" in the Name field. I also made sure that there were no rows with missing geocoordinates. I was left with 247 rows in my table. Upon further inspection I noticed that my data contained psychiatric and mental hospitals, nursing homes, a dental clinic, and a veterinary hospital. All but the veterinary hospital were labelled as ‘Hospital’ in the Feature Type. I decided to update the ’Feature Type’ for the other facility types. I then used the pandas ‘groupby’ method to return the number of each type of hospital.

	Name	County	State	Latitude	Longitude
Feature Type					
Dental Clinic	1	1	1	1	1
Hospital	219	219	219	219	219
Nursing Home	15	15	15	15	15
Psychiatric Hospital	11	11	11	11	11
Veterinary Hospital	1	1	1	1	1

I then used the Foursquare API to get the surrounding venue data for all five hospital types in my dataset. I stored the data in a Pandas DataFrame and associated each venue with its respective hospital. I discarded venue features that were not important for my study, keeping venue names, latitude and longitude, distance, and category.

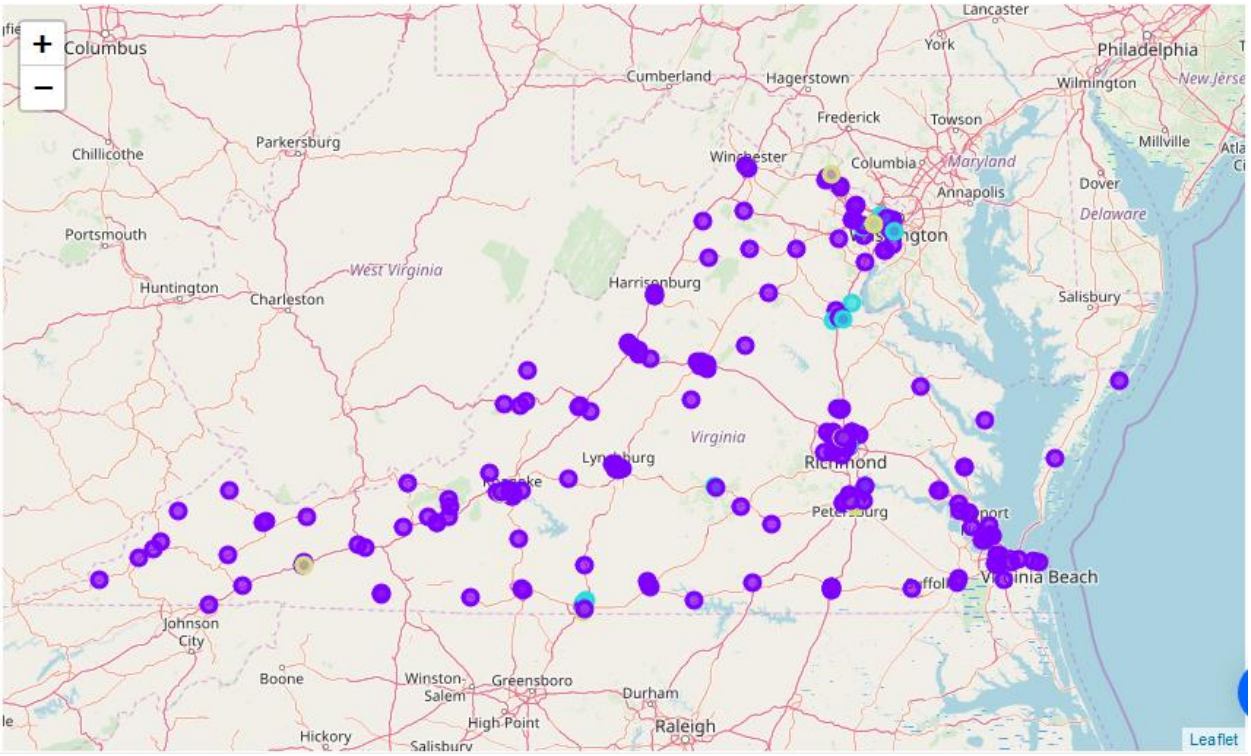
Here is a sample of the resulting data:

	Venue	Venue Lat	Venue Lng	Meters from Hospital	Category	Hospital
1	Greensville County Courthouse	36.686604	-77.542302	417.0	Courthouse	A B Adams Convalescent Center
2	New Century Hospice - Emporia	36.685332	-77.543467	511.0	Medical Center	A B Adams Convalescent Center
3	Peggy Malone - State Farm Insurance Agent	36.693774	-77.538012	898.0	Office	A B Adams Convalescent Center
4	Nationwide Insurance: Radke & Associates LLC	36.684614	-77.543248	504.0	Insurance Office	A B Adams Convalescent Center
5	Calvary Baptist Church	36.693578	-77.542884	988.0	Church	A B Adams Convalescent Center

3. Methodology

3.1 Exploratory Data Analysis

Before getting the venue data from Foursquare I wanted to see where all the hospitals were on a map of Virginia. I plotted the hospitals using Folium and colored them by hospital type.



After obtaining the Foursquare data, I began exploring my data using the Pandas' built-in methods head(), groupby(), and info(). I quickly found that I had up to 100 venues for each hospital, which I felt would be enough to provide some useful statistics and plots. I also found that the distance data was of type 'object', and I wanted it to be 'integers' so I could more easily perform statistical analysis with it. I further found that some of the venues did not have a category specified. I planned to rely on the category field heavily for my analysis, so I made an effort to provide a category name for venues that were easy to categorize, and I deleted the remaining venues with no category. I do not believe deleting those venues had much impact on my analysis because the venues were unusual and obscure, (like Telpage, Piano Lessons, Black Shiny Building, and Qahtani's Home), and not ones that I would consider an escape for hospital patrons.

Narrowing my Focus to Northern Virginia

Due to the hospital markers overlapping and bunching around areas of dense population I decided to narrow my focus to just the hospitals in Northern Virginia. I began by setting geographical boundaries for my hospitals, then I filtered out the hospitals that were not in my boundaries. I then got Foursquare venue data for just those hospitals. I cleaned up the data as described in the paragraph above. Here is a sample of the data once it was ready for statistical analysis:

	Venue	Venue Lat	Venue Lng	Meters from Hospital	Category	Hospital
0	Dr. Michael Joseph Horwath M.D.	38.965462	-77.356857	23	Doctor's Office	Access Emergency Hospital
1	Inova Endocrinology	38.965462	-77.356857	23	Medical Center	Access Emergency Hospital
2	Inova Emergency Room - Reston/Herndon	38.966410	-77.356660	86	Emergency Room	Access Emergency Hospital
3	Office Depot	38.965969	-77.355501	128	Paper / Office Supplies Store	Access Emergency Hospital
4	Harris Teeter	38.965674	-77.354899	175	Supermarket	Access Emergency Hospital
5	United Bank	38.966136	-77.356012	95	Bank	Access Emergency Hospital
6	Elizabeth Arden Red Door Spa	38.964875	-77.355643	142	Spa	Access Emergency Hospital
7	Long & Foster	38.967057	-77.357499	162	Building	Access Emergency Hospital
8	Bank of America	38.966636	-77.357936	138	Bank	Access Emergency Hospital
9	Reston Human Services Building	38.965335	-77.358850	170	Government Building	Access Emergency Hospital
10	Banfield Pet Hospital	38.963664	-77.355629	249	Veterinarian	Access Emergency Hospital

3.2 Statistical Analysis

I used Pandas' describe() method on my data to see some preliminary statistics, and from the results (shown below) I saw that the max distance was not within the 1000 meter radius that I had specified.

	lat	lng	distance
count	2419.000000	2419.000000	2419.000000
mean	38.838887	-77.254302	380.169905
std	0.076981	0.168454	1517.897249
min	38.247773	-77.815900	4.000000
25%	38.818695	-77.364938	129.000000
50%	38.859374	-77.227763	236.000000
75%	38.883730	-77.110007	444.000000
max	38.979113	-77.065195	68761.000000

As a result, I deleted any venues with a distance greater than 1000 meters.

Next, I found that there were 282 unique venue categories represented in my northern Virginia data. I decided to eliminate venues with categories that I did not consider "escapes". I dropped the following venues from my data: Road, Intersection, Doctor's Office, Medical Center, Hospital, Hospital Ward, Dentist's Office, Allergy Clinic, Hospital Imaging Center, Maternity Clinic, and Emergency Room. I also chose to remove venues that I considered were beyond a moderate walking distance of 300 meters.

3.3 Machine Learning

My purpose in doing this study was to group hospitals in Virginia according to the surrounding venues within walking distance that offer some type of escape for hospital patrons and visitors. I know how to use K-Means to form clusters based on distance, and I know that the algorithm can also work with non-distance type data, so I decided to give K-Means a try. I needed to decide what values to feed into K-Means to represent each venue. First I thought of giving each venue type a value based on the quantity venues belonging to that type around a hospital. To prepare the data, I used one-hot encoding to get a binary value for the relationship between each venue's type and the hospital associated with it. I then grouped the values by hospital to obtain

the mean number of venues of each type per hospital. The resulting dataframe allowed me to print out the top 10 most frequent venue types for each hospital in northern Virginia. Here is a sample of those results:

```
----Burke Medical Center----
                  venue  freq
0                  Bank  0.09
1                School  0.05
2  Fast Food Restaurant  0.04
3          Camera Store  0.04
4    Salon / Barbershop  0.04
5           Gas Station  0.04
6         Grocery Store  0.02
7    Recycling Facility  0.02
8         Shopping Mall  0.02
9              Car Wash  0.02


----Circle Terrace Hospital----
                  venue  freq
0         Music School  0.25
1  Housing Development  0.25
2                 Pool  0.25
3    Elementary School  0.25
4  American Restaurant  0.00
5               Office  0.00
6               Mosque  0.00
7          Music Store  0.00
8          Music Venue  0.00
9           Nail Salon  0.00
```

You can see from the frequency column above that Circle Terrace Hospital has only four types of venues nearby, with an equal number of each type. My next step was to cluster the hospitals with respect to the frequency values. I used K-Means and selected an arbitrary value (5) for k. The resulting clusters had a disproportional number of hospitals in them.

Additionally, when I looked at the top 30 venue types in each class, I noticed a large variety of venues in the clusters with multiple hospitals, and a much smaller variety of venues in the clusters with one hospital. Additionally, the venue types represented were not very helpful in determining a cluster that was better supplied with escapes. Consequently, I decided to try K-Means using values based on the distance of the different venue types. I only kept the closest venue when two or more venues near a hospital were of the same type. Here is a sample of the data organized by distance of venue type:

	Hospital	Category	Meters from Hospital
0	Access Emergency Hospital	Bank	97
1	Access Emergency Hospital	Paper / Office Supplies Store	128
2	Access Emergency Hospital	Bar	130
3	Access Emergency Hospital	Building	136
4	Access Emergency Hospital	Government Venue	136
5	Access Emergency Hospital	Hardware Store	138
6	Access Emergency Hospital	Mobile Phone Shop	141
7	Access Emergency Hospital	Salon	142
8	Access Emergency Hospital	Spa	142
9	Access Emergency Hospital	Government Building	144

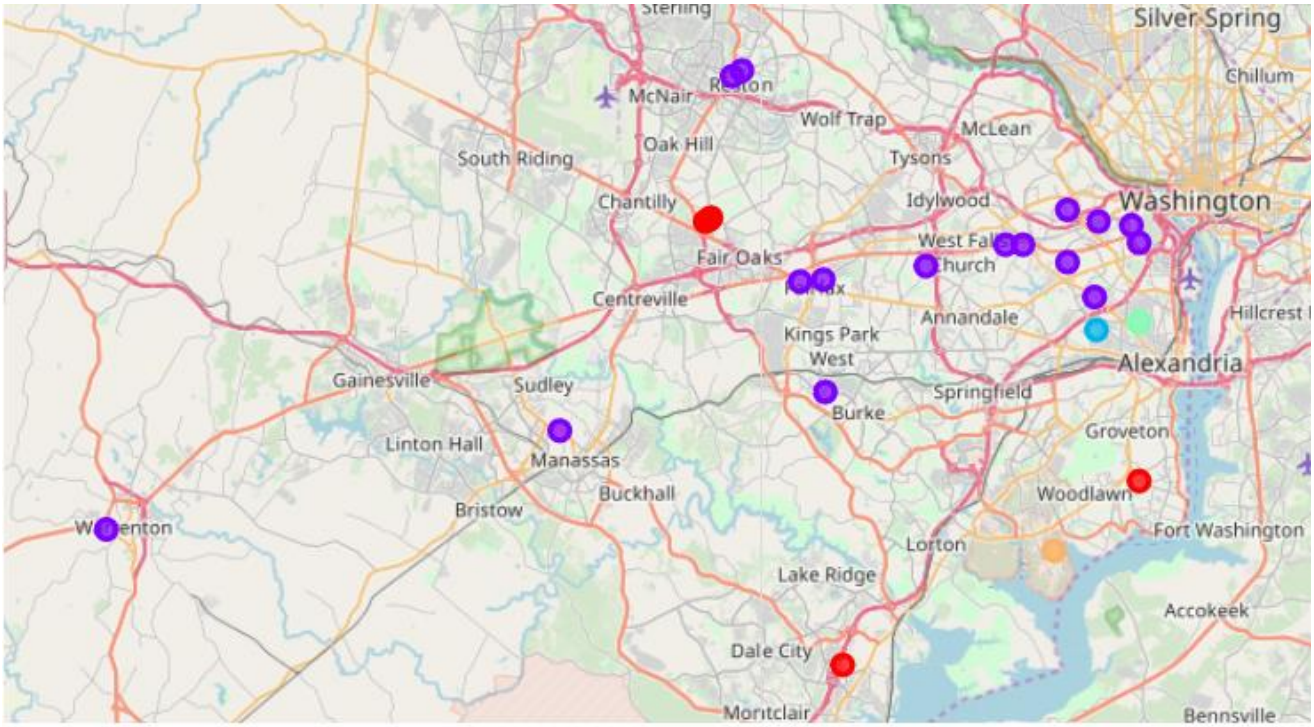
To make the processing simpler, I sorted the distances into bins from 0 to 5, with the closest venues represented by a 5, and the furthest venue types represented with a 1. The 0 value was reserved for venue types that were not found near that hospital. Then I pivoted the table using Hospital as the index, which resulted in 25 rows and 194 columns, with each column representing a venue type. Finally, I clustered the hospitals by providing the distance bin values to K-Means. Like my previous clusters, the resulting clusters had disproportionate numbers of hospitals in them. I looked at the kinds of venues in each of the two groups of clusters and came up with a name for each cluster.

Clusters Based on Frequency		
Cluster ID	Number of Hospitals	Names of Clusters
0	16	Mostly Office Buildings
1	1	Food Court, Coffee Shop, Beer Garden
2	1	Elem. School and Pool
3	1	Entertainment and Playground
4	6	Cafés and Buildings

Clusters Based on Distance Bin		
Cluster ID	Number of Hospitals	Names of Clusters
0	2	Variety Mid Dist
1	5	Variety Farther Dist
2	1	Variety and Leisure Close
3	14	Pharm with Farther Variety
4	3	Café, Pharm, and Med

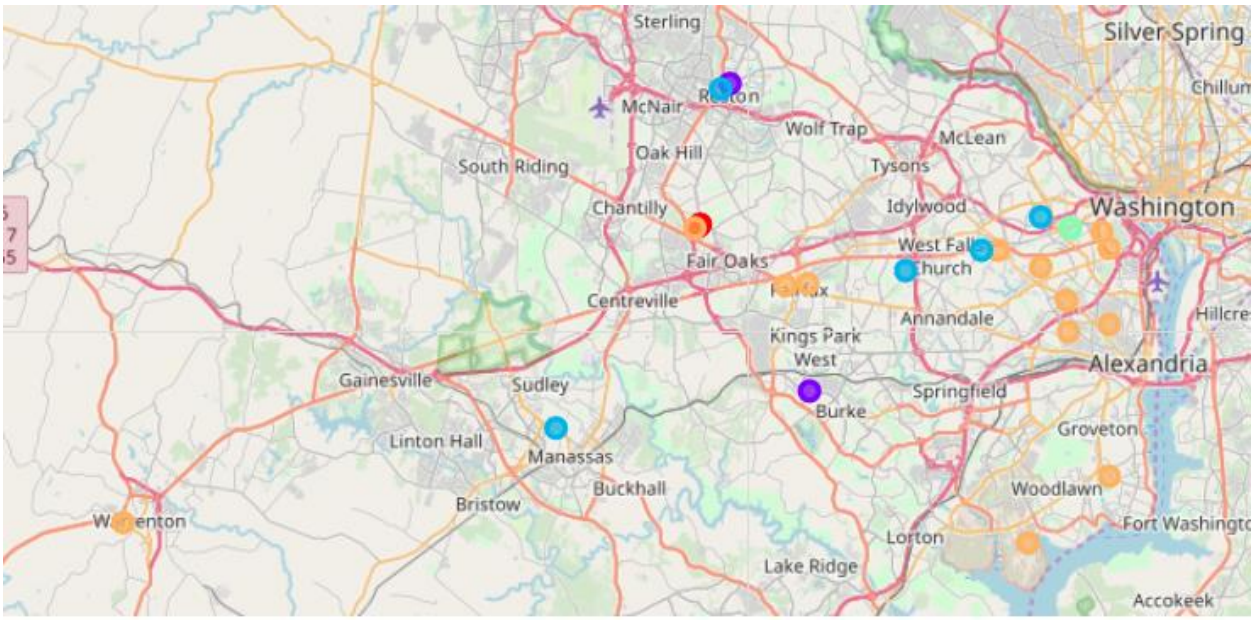
These clusters could be slightly helpful to hospital patrons, but do not offer the kind of information that I was hoping for.

The map below is a plot of the hospitals in northern Virginia classified by the frequency of venue type surrounding them.



Legend
Purple - Mostly Office Buildings
Lt Blue - Food Court, Coffee Shop, Beer Garden
Aqua Grn - Elem. School and Pool
Orange - Entertainment and Playground
Red - Cafes and Buildings

The following map is a plot of the hospitals in northern Virginia classified by the distance of the closest venue per venue type.




```
Legend
Purple - Variety Mid Dist
Lt Blue - Variety Farther Distance
Aqua Grn - Variety and Leisure Close
Orange - Pharm with Farther Variety
Red - Cafe, Pharm and Med
```

3.4 Trying a New Approach

Because these clusters didn't really capture the information that I feel is most helpful to hospital patrons, I created new clusters. This time, instead of relying on a machine learning, I manually created an algorithm to cluster the hospitals. I believe that the most frequently visited venues by hospital patrons are fast food restaurants, parks, and convenience stores. I am assuming that patients in the hospital would not need a pharmacy until they are released, but a pharmacy can also serve as a convenience store, so I will include it with convenience stores in my list of target venues. I will group the hospitals in the following clusters:

1 - Park, Fast Food, and Convenience Store
2 - Fast Food and Park
3 - Fast Food and Convenience Store
4 - Park and Convenience Store
5 - Park
6 - Fast Food
7 - Convenience Store
8 - None of the three

Rather than just focus on the hospitals in northern Virginia, I applied my clusters to all 218 general hospitals in Virginia. I had to grab new venue data for hospitals, avoiding nursing homes, psychiatric hospitals, and other hospital types, then I cleaned up the venue data just like I did for venues and hospitals in northern Virginia. This time I limited the venue data to venues within a 500-meter radius from the hospital.

To make my clusters, I first transformed my data with onehot encoding, making the venue categories my column names. Then I got the mean frequency of each venue type per hospital by grouping by hospital. With my data now ready for clustering, I began to define my clusters.

3.4 Defining the Clusters

I made a list of keywords that would help me identify venue types that belonged in one of my three target venue types: 'Walking', 'Nature', 'Fast Food', 'Burger', 'Sandwich', 'Convenience', 'Grocery', 'Pharmacy', 'Pizza'.

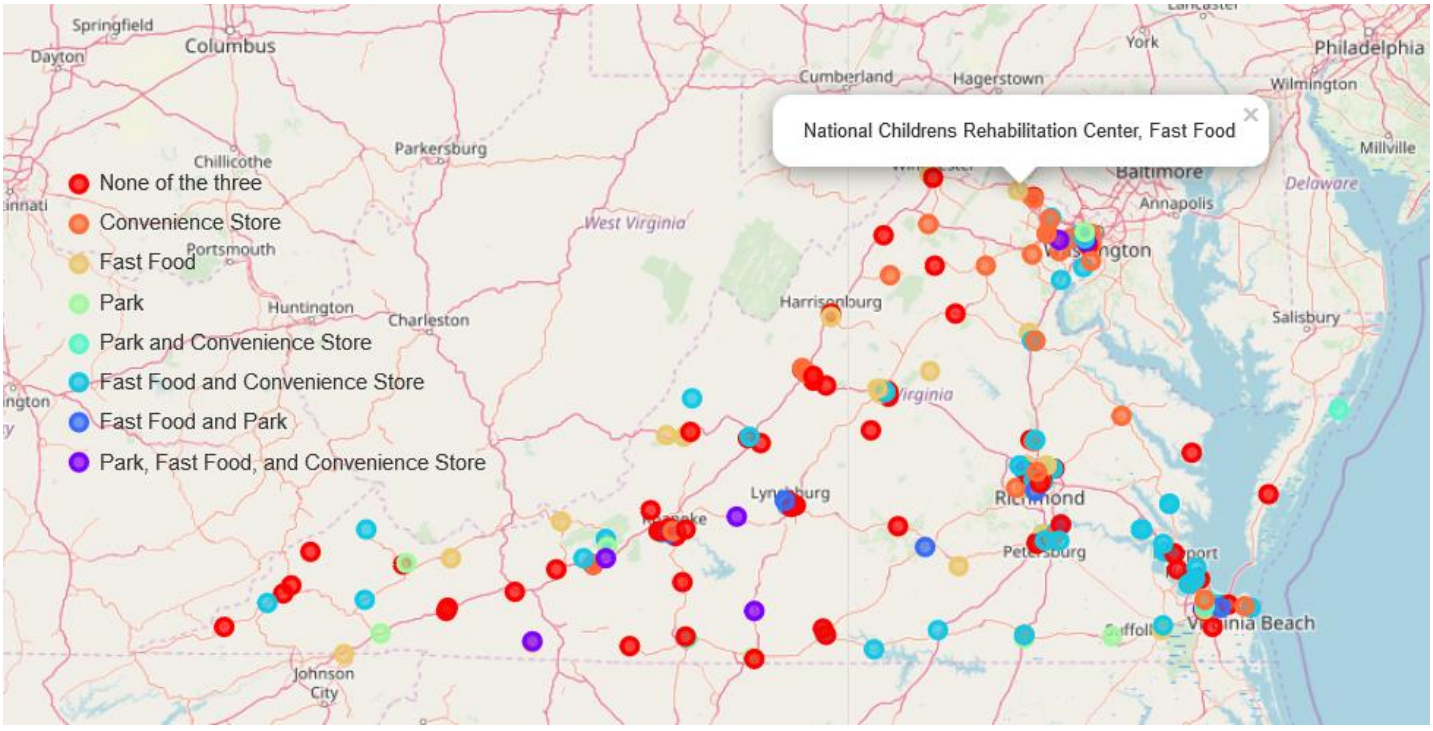
I used the list to identify the column names (venue types) to use in my clustering algorithm. I settled on the columns shown below.

	Target Hospital	Burger Joint	Convenience Store	Fast Food Restaurant	Grocery Store	Organic Grocery	Park	Pharmacy	Pizza Place	Sandwich Place
0	A B Adams Convalescent Center	0.00000	0.00000	0.000000	0.000000	0.0	0.052632	0.052632	0.000000	0.000000
1	A D Williams Memorial Clinic	0.00000	0.00000	0.020619	0.000000	0.0	0.000000	0.010309	0.000000	0.010309
2	Access Emergency Hospital	0.00000	0.00000	0.010753	0.010753	0.0	0.000000	0.010753	0.032258	0.010753
3	Albemarle County Health Department	0.00000	0.00000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
4	Alexander W Terrell Memorial Infirmary	0.00000	0.00000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
5	Alleghany Memorial Hospital	0.00000	0.00000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.025641
6	Alleghany Regional Hospital	0.00000	0.00000	0.037037	0.000000	0.0	0.000000	0.000000	0.000000	0.000000

Next I iterated through the dataframe three times to create three lists of Boolean values – one list for each of my target venue types: Fast Food, Park, and Convenience Store. Finally, I created a column to store the new cluster IDs and used the Boolean masks to sort the hospitals into one of the eight categories. After completing the clustering, I mapped the clusters using Folium.

4. Results

The map below is a plot of the hospitals in all of Virginia according to clusters that I manually created.



The table below shows the number of hospitals grouped in each cluster. I arranged the Cluster IDs in descending order to match the legend on the plot.

Clusters Based on Target Venues		
Cluster ID	Number of Hospitals	Names of Clusters
8	68	None of the three
7	34	Convenience Store
6	27	Fast Food
5	11	Park
4	7	Park and Convenience Store
3	51	Fast Food and Convenience Store
2	11	Fast Food and Park
1	9	Park, Fast Food, and Convenience Store

As you can see in the table above, 68 of the 218 hospitals had none of the three target venue types, while 9 of the hospitals had all three. Parks were the most rare venue type of the three that I compared, and convenience stores were the most common.

5. Discussion

The manually created clusters are much more useful than the ones generated by machine learning (with the K-Means algorithm). I can use the plot to quickly identify hospitals in Virginia that have the nearby venues that most value. These results, however fascinating, are limited in their accuracy and usefulness. First, this study does not take into account gift shops or floral shops, which may also be valued by hospital patrons. Second, the results do not account for walking trails or garden paths around or near the hospitals, which may be more desirable, and likely more common, than parks. Third, the results are only as accurate as the Foursquare data is complete and accurate. As you read in the beginning of my Methodology section, some of the venues did not have a category specified, and it is possible that I removed some venues that could have changed the results. Other venues could have been categorized in a way that was inaccurate or that was not captured by my key word search.

6. Conclusion

This was a fun and educational challenge that yielded some interesting and somewhat useful results. I would propose that the best way to choose a hospital without actually visiting it would be by word of mouth and reputation, and that the most important factors for choosing a hospital should be the quality of the services rendered and care available for a person's particular medical needs. After that, it could be useful and fun to see which hospitals have surrounding venues that offer scarce amenities or an escape from the long, lonely hours spent in a hospital wing. This study could be revised to compare the venues surrounding other hospital types, such as psychiatric hospitals or nursing homes. The algorithms used in this study could also be adapted to finding other venue types simply by changing the keyword searches and filters.