"Where to live in Washington metropolitan area"

Coursera - "IBM Data Science Professional Certificate"
Specialization Capstone Project

Claudio Ferrao

March 2019

Contents

1. Business Problem

1.1. Problem definition

The process of finding a new place to live can be a daunting experience, especially if you do not know much about the area you want to move, and/or if the area is too broad to research. This process can also be very time consuming and costly, especially if you are moving from a different city, state, country.

The analysis performed with this project is general and can be applied to any other city/county. In this project we are assuming that the area the family wants to search a place to live is Washington Metropolitan area. The Washington metropolitan area is the metropolitan area centered on Washington, D.C., the capital of the United States. The area includes all of the federal district and parts of the U.S. states of Maryland and Virginia, along with a small portion of West Virginia.

1.2. Target audience

In this notebook we propose a way of shortening this search process by using data about the most popular venues of each Washington Metropolitan area county to help find their "venue profile". There are two group of people interested in this type of analysis:

It helps real estate companies knowing what the client requirements (venue profile) are. This will help limit the search area to the preferred venues of clients, and therefore the need of company real estate resources and time.

For the families that want to move or relocate to this area, this type of analysis can speed up the process of searching a place to live saving time and money. It will also help families to take a more informed decision

1.3. Client Venue requirements

This is something that can change based on interviewing the family that is looking for a place to live. For this exercise the venues that are considered important to this family are:

- Grocery stores
- Gym
- Restaurants
- Cafe
- Parks

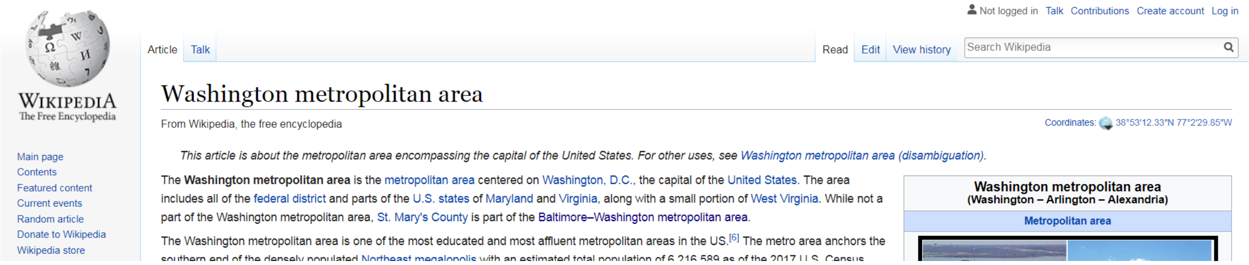
2. Data to support this project

2.1. Data description

In order to tackle the problem of defining the "venue profiles" of neighborhoods in a county, the following data was used:

- Web scraping with BeautifulSoup to obtain a list of all counties from the Washington Metropolitan area.

The Wiki page Washington metropolitan area https://en.wikipedia.org/wiki/Washington_metropolitan_area



Washington metropolitan area

From Wikipedia, the free encyclopedia

Coordinates: 38°53′12.33″N 77°2′29.85″W

This article is about the metropolitan area encompassing the capital of the United States. For other uses, see Washington metropolitan area (disambiguation).

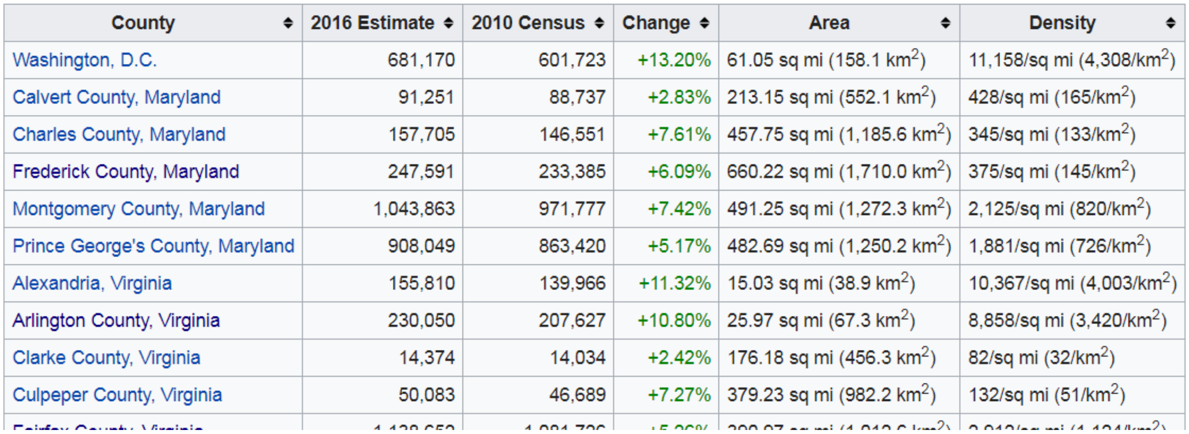
The **Washington metropolitan area** is the metropolitan area centered on Washington, D.C., the capital of the United States. The area includes all of the federal district and parts of the U.S. states of Maryland and Virginia, along with a small portion of West Virginia. While not a part of the Washington metropolitan area, St. Mary's County is part of the Baltimore–Washington metropolitan area.

The Washington metropolitan area is one of the most educated and most affluent metropolitan areas in the US.[6] The metro area anchors the southern end of the densely populated Northeast megalopolis with an estimated total population of 6,216,589 as of the 2017 U.S. Census

Washington metropolitan area
(Washington – Arlington – Alexandria)

Metropolitan area

has a table with:



County	2016 Estimate	2010 Census	Change	Area	Density
Washington, D.C.	681,170	601,723	+13.20%	61.05 sq mi (158.1 km^2)	11,158/sq mi (4,308/km^2)
Calvert County, Maryland	91,251	88,737	+2.83%	213.15 sq mi (552.1 km^2)	428/sq mi (165/km^2)
Charles County, Maryland	157,705	146,551	+7.61%	457.75 sq mi (1,185.6 km^2)	345/sq mi (133/km^2)
Frederick County, Maryland	247,591	233,385	+6.09%	660.22 sq mi (1,710.0 km^2)	375/sq mi (145/km^2)
Montgomery County, Maryland	1,043,863	971,777	+7.42%	491.25 sq mi (1,272.3 km^2)	2,125/sq mi (820/km^2)
Prince George's County, Maryland	908,049	863,420	+5.17%	482.69 sq mi (1,250.2 km^2)	1,881/sq mi (726/km^2)
Alexandria, Virginia	155,810	139,966	+11.32%	15.03 sq mi (38.9 km^2)	10,367/sq mi (4,003/km^2)
Arlington County, Virginia	230,050	207,627	+10.80%	25.97 sq mi (67.3 km^2)	8,858/sq mi (3,420/km^2)
Clarke County, Virginia	14,374	14,034	+2.42%	176.18 sq mi (456.3 km^2)	82/sq mi (32/km^2)
Culpeper County, Virginia	50,083	46,689	+7.27%	379.23 sq mi (982.2 km^2)	132/sq mi (51/km^2)
Fairfax County, Virginia	1,138,652	1,081,726	+5.26%	390.97 sq mi (1,012.6 km^2)	2,912/sq mi (1,124/km^2)

- a list of all the counties,
- 2016 estimate population,
- 2010 census population,
- Percent change,
- area size, and
- density.

Only the first colums with the list of counties is relevant to this project.

- Nominatim from geopy.geocoders for geocoding the County names and get their coordinates.

With Nominatim we will be able to concatenate the list of counties from the Washington metropolitan area with their coordinates.

The list of counties have counties from two states - VA and MD. Nominatim was executed two times to get the coordinates of counties from our list:

- one for Virginia state counties, and
- one for Maryland state counties

- Foursquare API to get all venues in each County.

With Foursquare we will get the top 100 venues that are within a radius of 10000 meters of each county.

3. Methodology

The methods used in this work were:

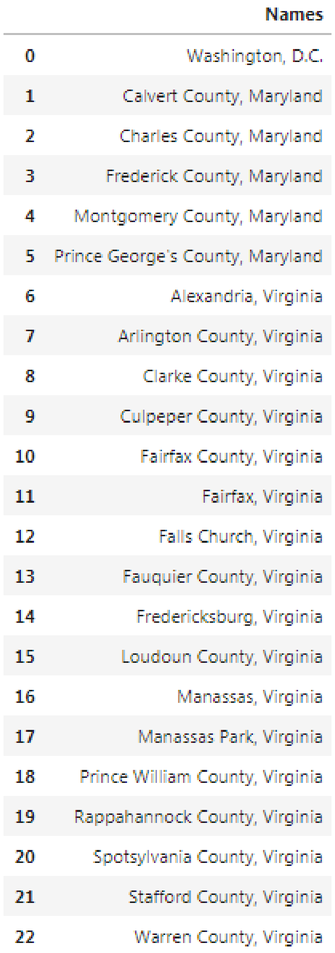
1. Web Scraping with the BeautifulSoup library
2. Geocoding with Nominatim from geopy.geocoders
3. Data acquisition from Foursquare's API
4. Feature reduction by considering most common venue categories
5. Machine learning: k-Means clustering because it is the most simple clustering algorithm and it was capable of meeting the proposed objective

4. Results

4.1. Web scraping

The Wikipedia page (https://en.wikipedia.org/wiki/Washington_metropolitan_area) contains a list of all counties from Washington Metropolitan in a table format.

BeautifulSoup is used to scrape a Wikipedia page and extract the elements of its index/table of contents. It returns a list with the text for each of the elements.



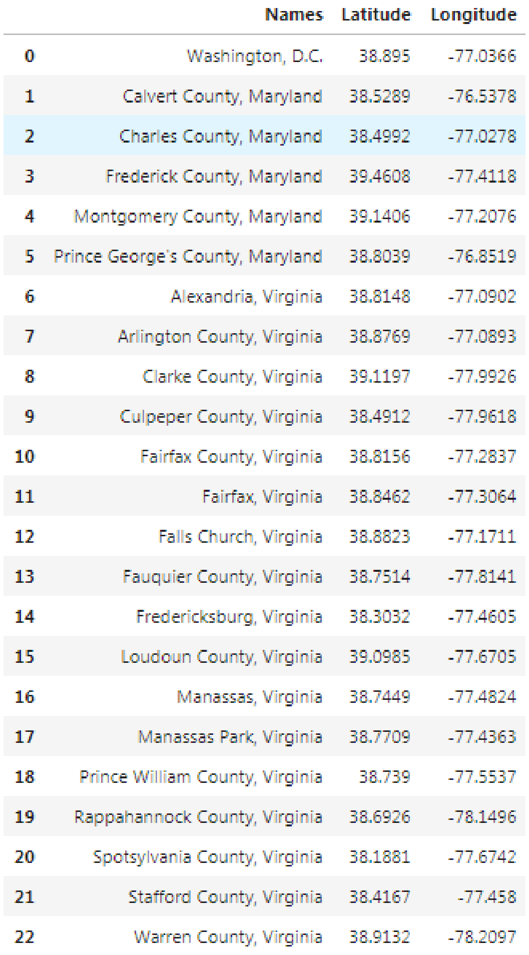
	Names
0	Washington, D.C.
1	Calvert County, Maryland
2	Charles County, Maryland
3	Frederick County, Maryland
4	Montgomery County, Maryland
5	Prince George's County, Maryland
6	Alexandria, Virginia
7	Arlington County, Virginia
8	Clarke County, Virginia
9	Culpeper County, Virginia
10	Fairfax County, Virginia
11	Fairfax, Virginia
12	Falls Church, Virginia
13	Fauquier County, Virginia
14	Fredericksburg, Virginia
15	Loudoun County, Virginia
16	Manassas, Virginia
17	Manassas Park, Virginia
18	Prince William County, Virginia
19	Rappahannock County, Virginia
20	Spotsylvania County, Virginia
21	Stafford County, Virginia
22	Warren County, Virginia

4.2. Geocoding

The Washington metropolitan Area is the metropolitan area centered on Washington, D.C., the capital of the United States. The area includes all of the federal district and parts of the U.S. states of Maryland and Virginia, along with a small portion of West Virginia.

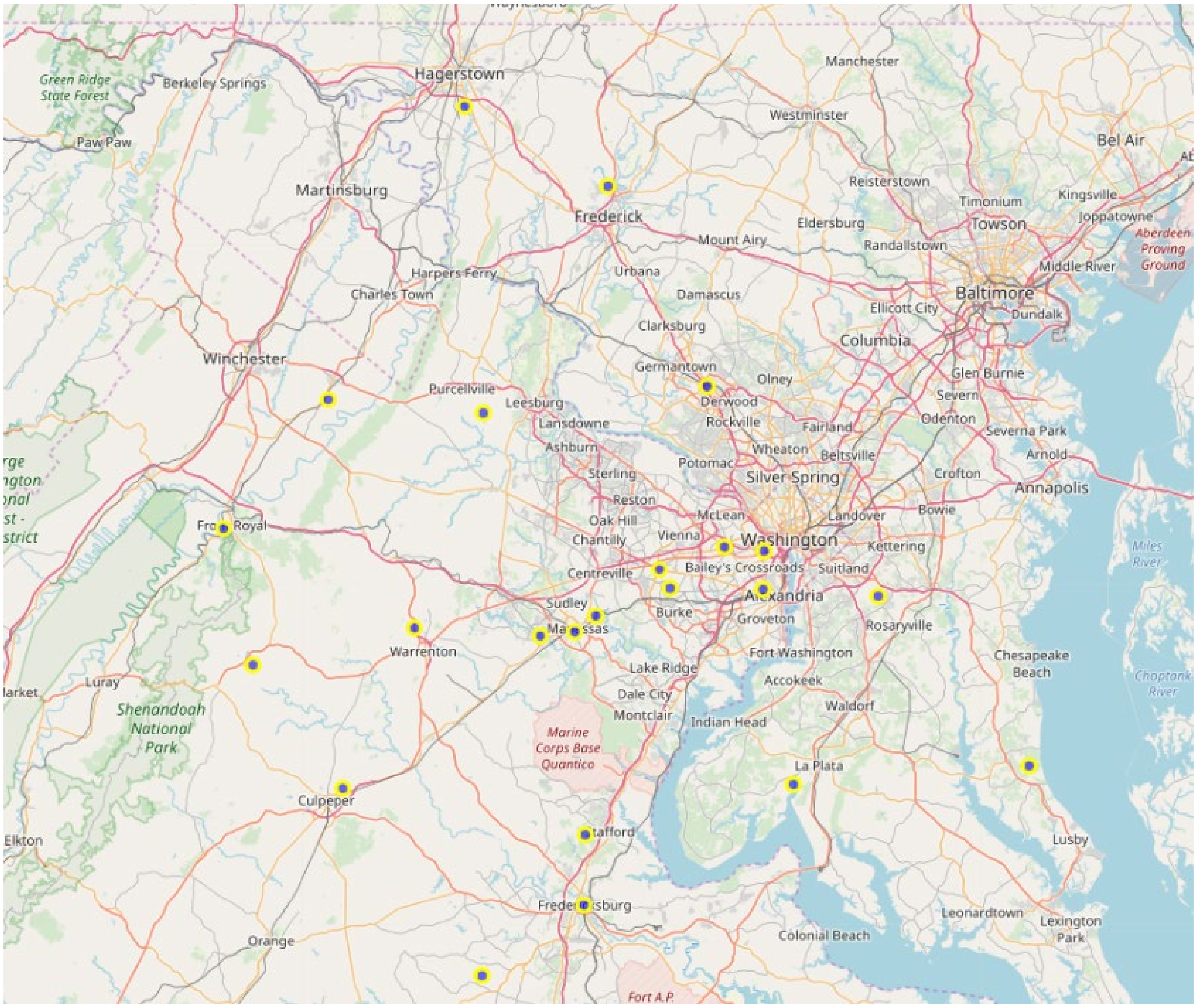
The Nominatim from geopy.geociders is used to obtain the coordinates of each county the geocoding feature of the Nominatim from geopy.geociders.

We geocoded all 23 counties.



	Names	Latitude	Longitude
0	Washington, D.C.	38.895	-77.0366
1	Calvert County, Maryland	38.5289	-76.5378
2	Charles County, Maryland	38.4992	-77.0278
3	Frederick County, Maryland	39.4608	-77.4118
4	Montgomery County, Maryland	39.1406	-77.2076
5	Prince George's County, Maryland	38.8039	-76.8519
6	Alexandria, Virginia	38.8148	-77.0902
7	Arlington County, Virginia	38.8769	-77.0893
8	Clarke County, Virginia	39.1197	-77.9926
9	Culpeper County, Virginia	38.4912	-77.9618
10	Fairfax County, Virginia	38.8156	-77.2837
11	Fairfax, Virginia	38.8462	-77.3064
12	Falls Church, Virginia	38.8823	-77.1711
13	Fauquier County, Virginia	38.7514	-77.8141
14	Fredericksburg, Virginia	38.3032	-77.4605
15	Loudoun County, Virginia	39.0985	-77.6705
16	Manassas, Virginia	38.7449	-77.4824
17	Manassas Park, Virginia	38.7709	-77.4363
18	Prince William County, Virginia	38.739	-77.5537
19	Rappahannock County, Virginia	38.6926	-78.1496
20	Spotsylvania County, Virginia	38.1881	-77.6742
21	Stafford County, Virginia	38.4167	-77.458
22	Warren County, Virginia	38.9132	-78.2097

4.3. Visualization of Counties on a map



4.4. Getting venue information

We will get data about venues around each County (from the Washington Metropolitan area) using the Foursquare API.

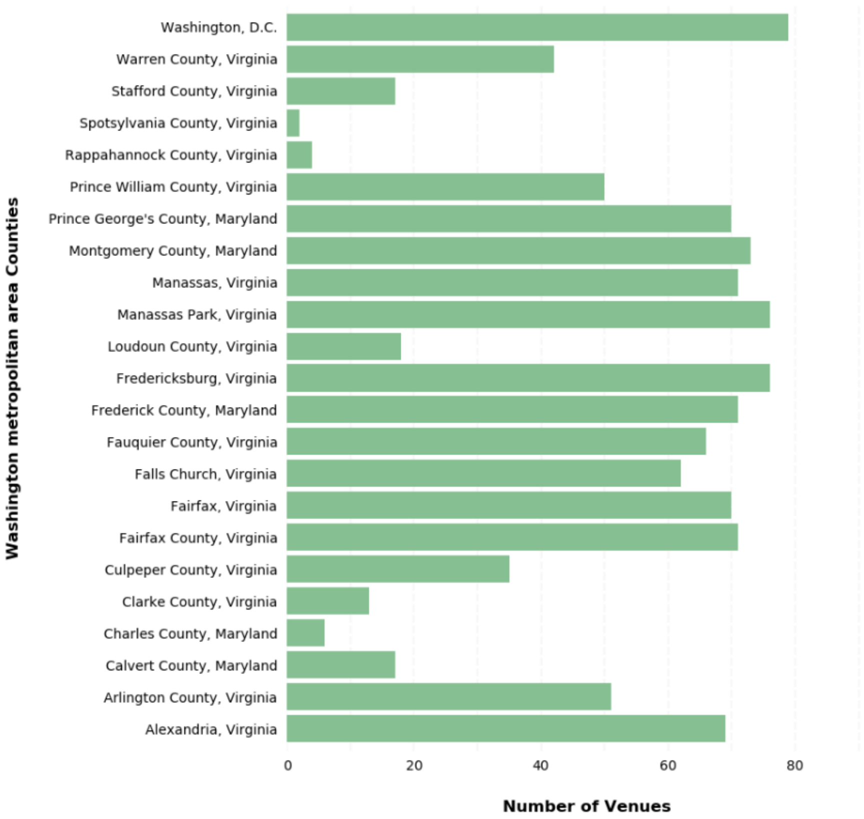
4.5. How are the categories distributed among the venues?

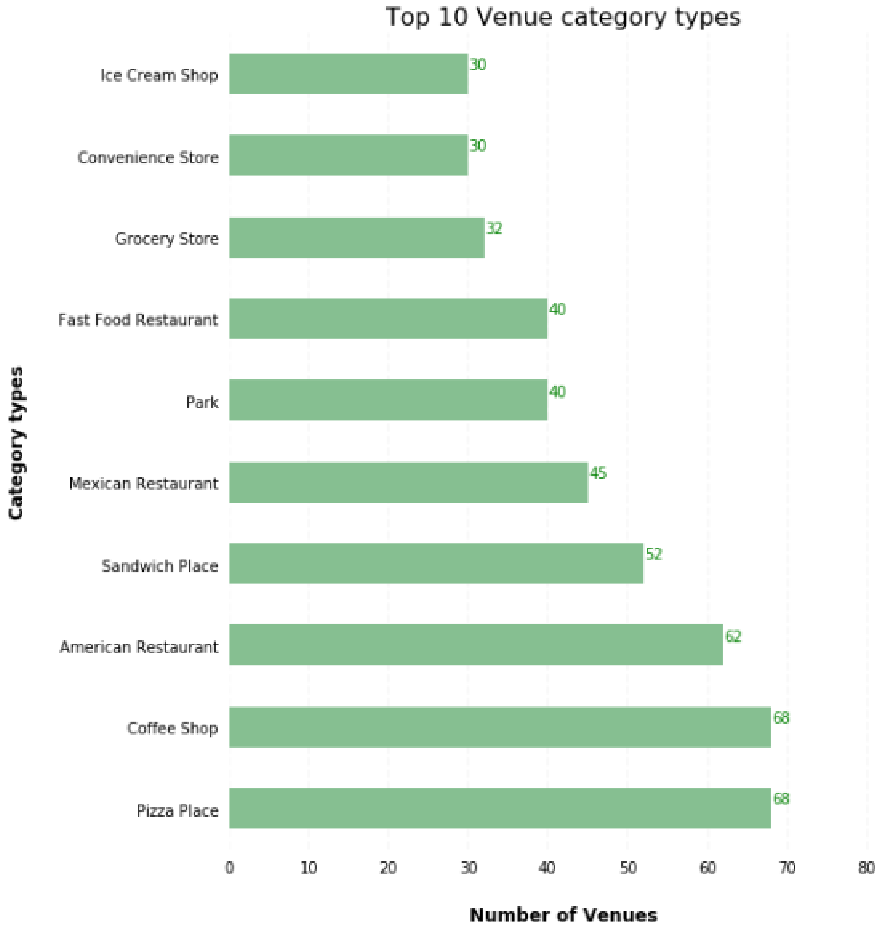
There are a total number of 235 categories.

There are several unique categories, but many of them have very few representative venues.
These low-populated categories can act as noise in our feature space and make clustering harder.
All categories which have less venues and that represent less than 0.5% of the total number of venues were dropped.

We are left with 51 categories for our analysis.

Number of venues by County



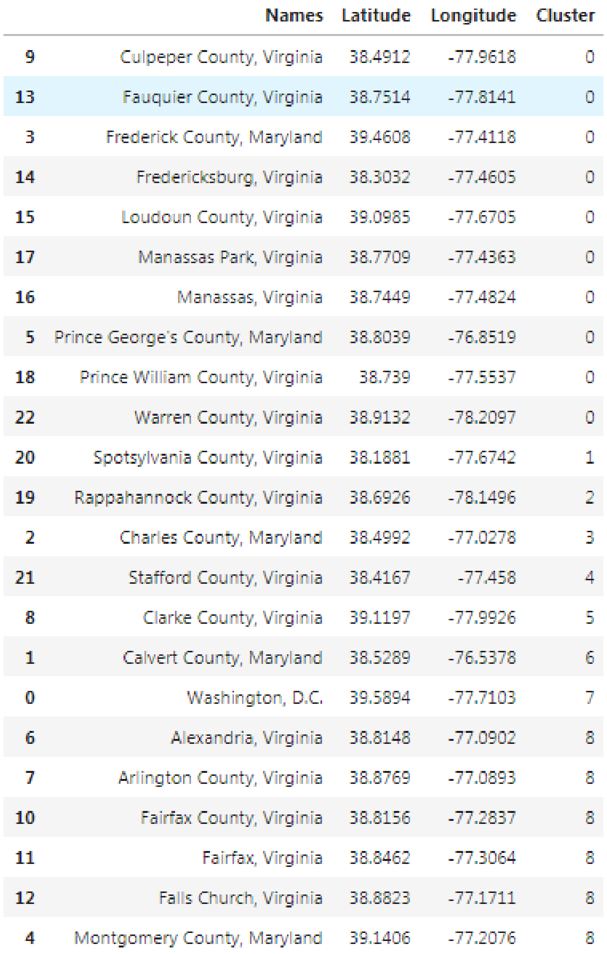


4.6. Clustering

Now we will use k-Means clustering in order to find "venue profiles" for the Washington Metropolitan area counties.
First let's find the optimum value for k, which is the number of clusters.

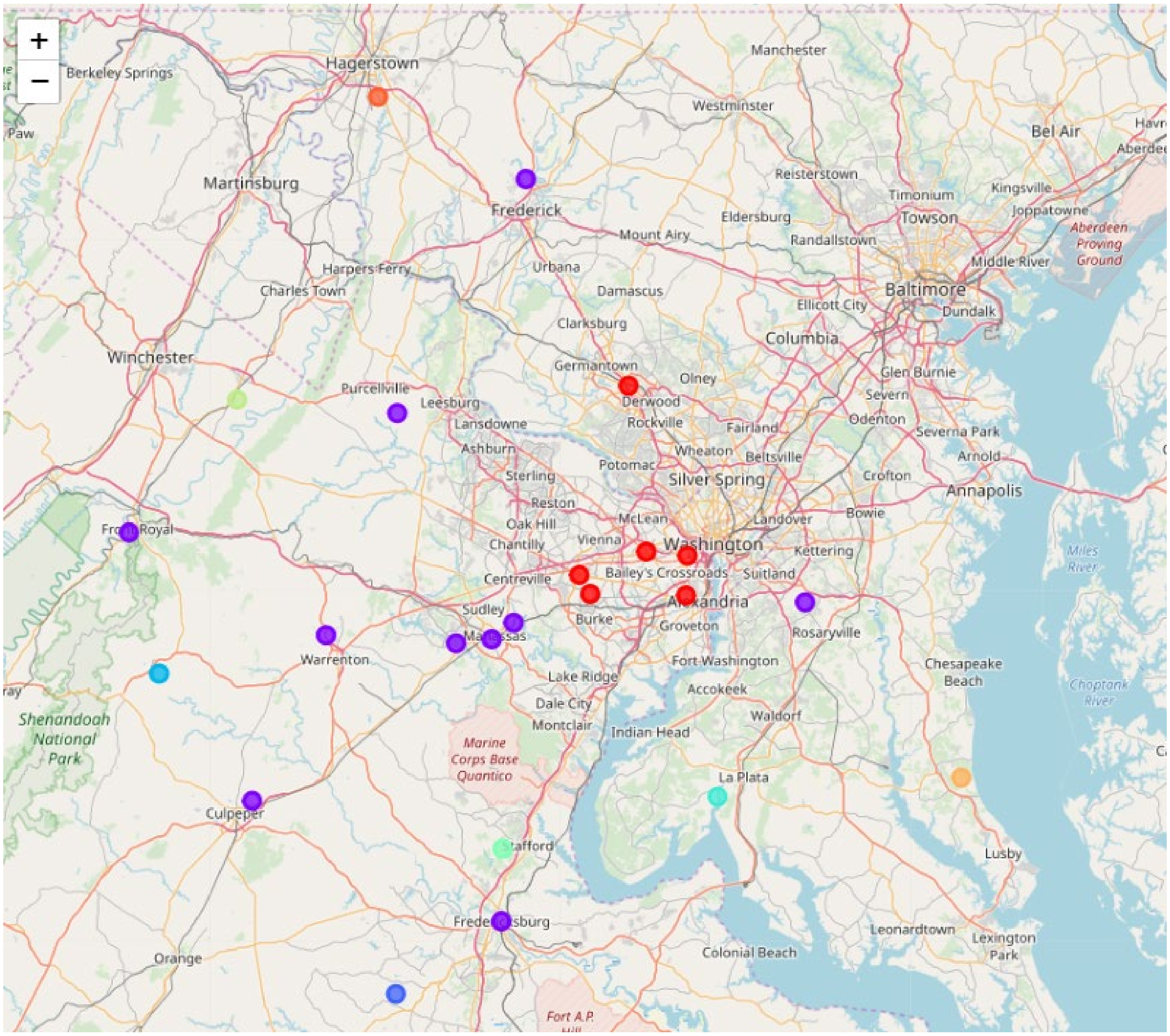
The majority of counties are focused on one cluster. That will not help with reduction of area that a professional want to pick for live. For that reason the number of klsuter is increased to 9. It seems that there is another break there.

Merging original file with Cluster classification (sorted by Cluster and County name)

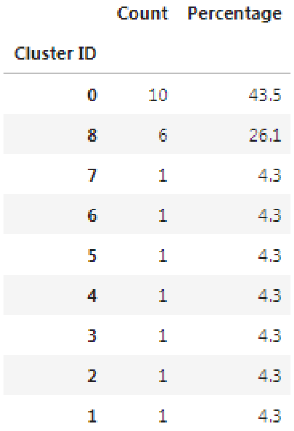


	Names	Latitude	Longitude	Cluster
9	Culpeper County, Virginia	38.4912	-77.9618	0
13	Fauquier County, Virginia	38.7514	-77.8141	0
3	Frederick County, Maryland	39.4608	-77.4118	0
14	Fredericksburg, Virginia	38.3032	-77.4605	0
15	Loudoun County, Virginia	39.0985	-77.6705	0
17	Manassas Park, Virginia	38.7709	-77.4363	0
16	Manassas, Virginia	38.7449	-77.4824	0
5	Prince George's County, Maryland	38.8039	-76.8519	0
18	Prince William County, Virginia	38.739	-77.5537	0
22	Warren County, Virginia	38.9132	-78.2097	0
20	Spotsylvania County, Virginia	38.1881	-77.6742	1
19	Rappahannock County, Virginia	38.6926	-78.1496	2
2	Charles County, Maryland	38.4992	-77.0278	3
21	Stafford County, Virginia	38.4167	-77.458	4
8	Clarke County, Virginia	39.1197	-77.9926	5
1	Calvert County, Maryland	38.5289	-76.5378	6
0	Washington, D.C.	39.5894	-77.7103	7
6	Alexandria, Virginia	38.8148	-77.0902	8
7	Arlington County, Virginia	38.8769	-77.0893	8
10	Fairfax County, Virginia	38.8156	-77.2837	8
11	Fairfax, Virginia	38.8462	-77.3064	8
12	Falls Church, Virginia	38.8823	-77.1711	8
4	Montgomery County, Maryland	39.1406	-77.2076	8

4.7. Mapping counties with cluster color classification

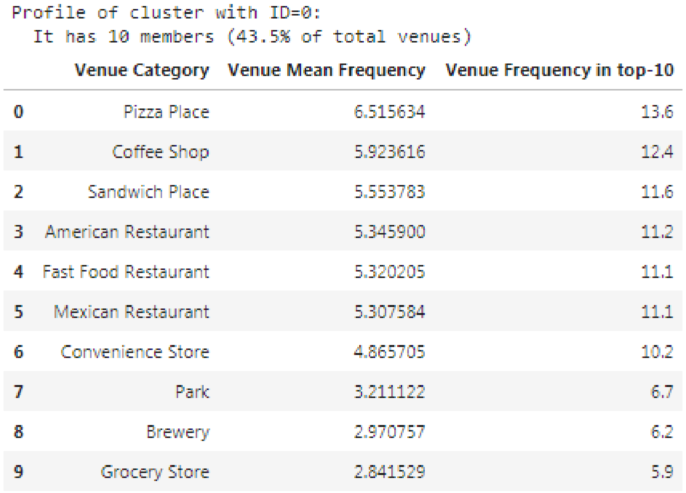


How are these clusters populated?



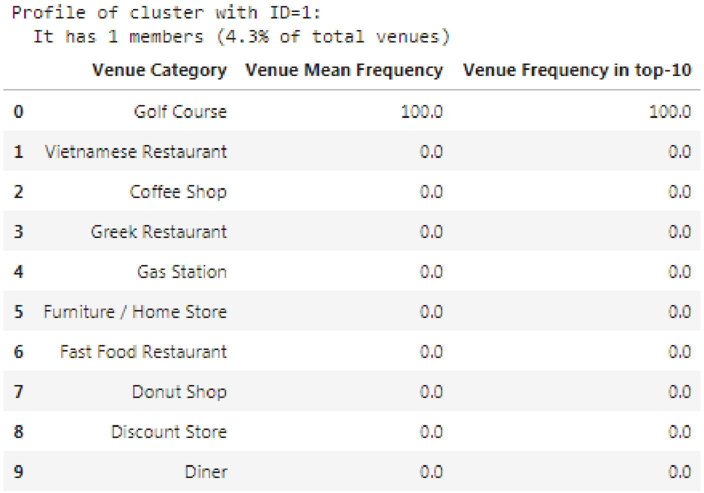
	Count	Percentage
Cluster ID		
0	10	43.5
8	6	26.1
7	1	4.3
6	1	4.3
5	1	4.3
4	1	4.3
3	1	4.3
2	1	4.3
1	1	4.3

4.8. How are the clusters different in terms of venue categories?



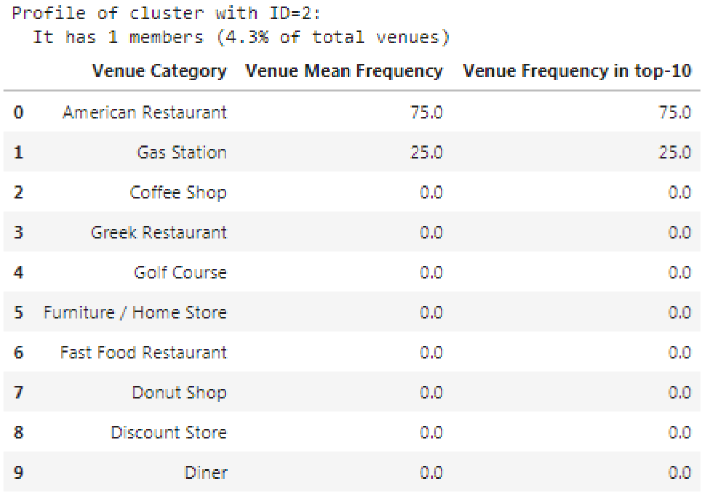
```
Profile of cluster with ID=0:
  It has 10 members (43.5% of total venues)
```

	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	Pizza Place	6.515634	13.6
1	Coffee Shop	5.923616	12.4
2	Sandwich Place	5.553783	11.6
3	American Restaurant	5.345900	11.2
4	Fast Food Restaurant	5.320205	11.1
5	Mexican Restaurant	5.307584	11.1
6	Convenience Store	4.865705	10.2
7	Park	3.211122	6.7
8	Brewery	2.970757	6.2
9	Grocery Store	2.841529	5.9



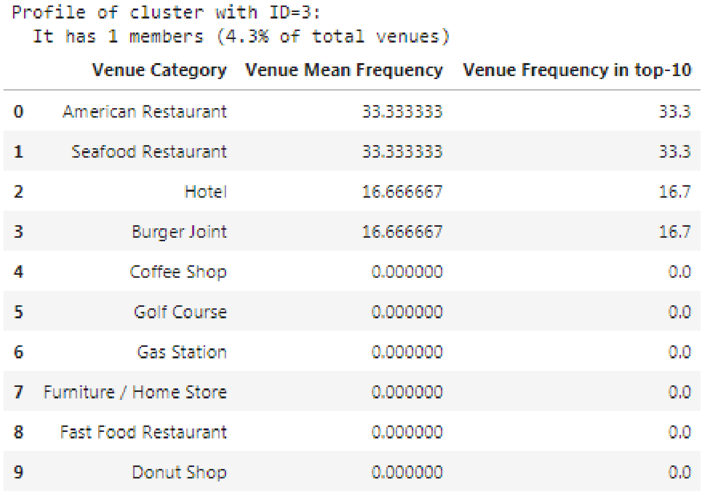
```
Profile of cluster with ID=1:
  It has 1 members (4.3% of total venues)
```

	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	Golf Course	100.0	100.0
1	Vietnamese Restaurant	0.0	0.0
2	Coffee Shop	0.0	0.0
3	Greek Restaurant	0.0	0.0
4	Gas Station	0.0	0.0
5	Furniture / Home Store	0.0	0.0
6	Fast Food Restaurant	0.0	0.0
7	Donut Shop	0.0	0.0
8	Discount Store	0.0	0.0
9	Diner	0.0	0.0



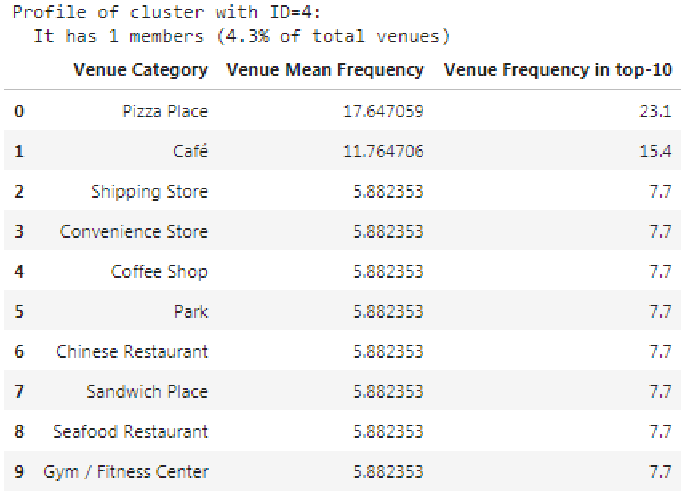
```
Profile of cluster with ID=2:
  It has 1 members (4.3% of total venues)
```

	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	American Restaurant	75.0	75.0
1	Gas Station	25.0	25.0
2	Coffee Shop	0.0	0.0
3	Greek Restaurant	0.0	0.0
4	Golf Course	0.0	0.0
5	Furniture / Home Store	0.0	0.0
6	Fast Food Restaurant	0.0	0.0
7	Donut Shop	0.0	0.0
8	Discount Store	0.0	0.0
9	Diner	0.0	0.0



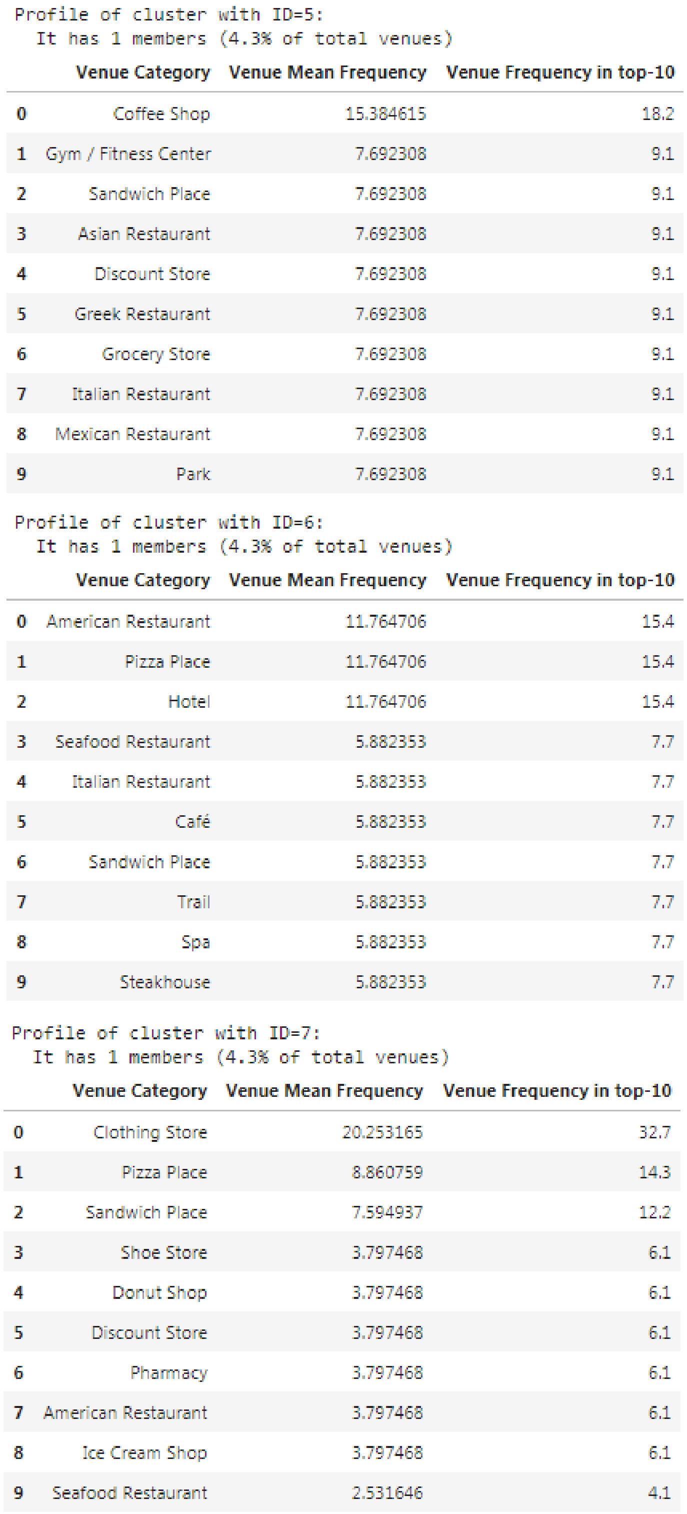
```
Profile of cluster with ID=3:
  It has 1 members (4.3% of total venues)
```

	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	American Restaurant	33.333333	33.3
1	Seafood Restaurant	33.333333	33.3
2	Hotel	16.666667	16.7
3	Burger Joint	16.666667	16.7
4	Coffee Shop	0.000000	0.0
5	Golf Course	0.000000	0.0
6	Gas Station	0.000000	0.0
7	Furniture / Home Store	0.000000	0.0
8	Fast Food Restaurant	0.000000	0.0
9	Donut Shop	0.000000	0.0



```
Profile of cluster with ID=4:
  It has 1 members (4.3% of total venues)
```

	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	Pizza Place	17.647059	23.1
1	Café	11.764706	15.4
2	Shipping Store	5.882353	7.7
3	Convenience Store	5.882353	7.7
4	Coffee Shop	5.882353	7.7
5	Park	5.882353	7.7
6	Chinese Restaurant	5.882353	7.7
7	Sandwich Place	5.882353	7.7
8	Seafood Restaurant	5.882353	7.7
9	Gym / Fitness Center	5.882353	7.7



```
Profile of cluster with ID=5:
  It has 1 members (4.3% of total venues)
```

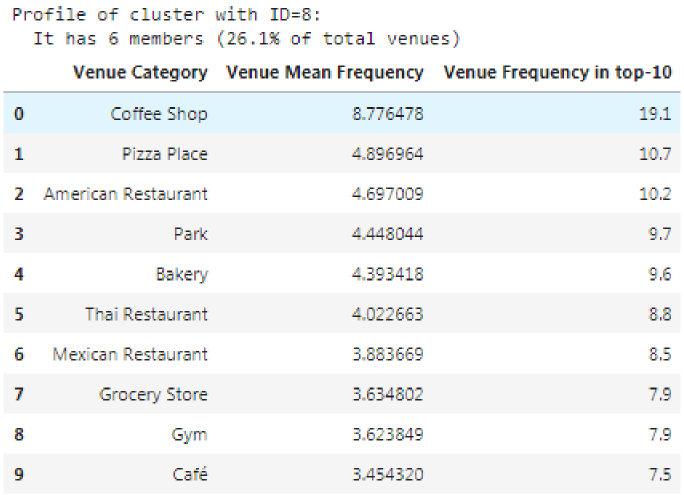
	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	Coffee Shop	15.384615	18.2
1	Gym / Fitness Center	7.692308	9.1
2	Sandwich Place	7.692308	9.1
3	Asian Restaurant	7.692308	9.1
4	Discount Store	7.692308	9.1
5	Greek Restaurant	7.692308	9.1
6	Grocery Store	7.692308	9.1
7	Italian Restaurant	7.692308	9.1
8	Mexican Restaurant	7.692308	9.1
9	Park	7.692308	9.1

```
Profile of cluster with ID=6:
  It has 1 members (4.3% of total venues)
```

	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	American Restaurant	11.764706	15.4
1	Pizza Place	11.764706	15.4
2	Hotel	11.764706	15.4
3	Seafood Restaurant	5.882353	7.7
4	Italian Restaurant	5.882353	7.7
5	Café	5.882353	7.7
6	Sandwich Place	5.882353	7.7
7	Trail	5.882353	7.7
8	Spa	5.882353	7.7
9	Steakhouse	5.882353	7.7

```
Profile of cluster with ID=7:
  It has 1 members (4.3% of total venues)
```

	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	Clothing Store	20.253165	32.7
1	Pizza Place	8.860759	14.3
2	Sandwich Place	7.594937	12.2
3	Shoe Store	3.797468	6.1
4	Donut Shop	3.797468	6.1
5	Discount Store	3.797468	6.1
6	Pharmacy	3.797468	6.1
7	American Restaurant	3.797468	6.1
8	Ice Cream Shop	3.797468	6.1
9	Seafood Restaurant	2.531646	4.1



```
Profile of cluster with ID=8:
  It has 6 members (26.1% of total venues)
```

	Venue Category	Venue Mean Frequency	Venue Frequency in top-10
0	Coffee Shop	8.776478	19.1
1	Pizza Place	4.896964	10.7
2	American Restaurant	4.697009	10.2
3	Park	4.448044	9.7
4	Bakery	4.393418	9.6
5	Thai Restaurant	4.022663	8.8
6	Mexican Restaurant	3.883669	8.5
7	Grocery Store	3.634802	7.9
8	Gym	3.623849	7.9
9	Café	3.454320	7.5

We successfully clustered the Counties of Washington metropolitan area into 8 clusters with distinct "venue profiles".

5. Discussion

The objective of this analysis was to shorten the search process of finding a place to live in the Washington Metropolitan area by analyzing the most popular "venue profiles" for clusters of counties from the Washington metropolitan area.

The ten most common venue categories of each cluster that we identified are enough to reveal differences between them and allow us to imagine which "client profile" fits the best with each cluster.

Family venue requirements:

- Grocery stores,
- Gym
- Restaurants
- Cafe
- Parks

Resume of cluster venues and matches from the family venue requirements.

Cluster 0 - Have 4 of the 5 family requirements: Grocery stores, Restaurants, Cafe, and Parks

Cluster 1 - Have 2 of the 5 family requirements: Restaurants and Cafe

Cluster 2 - Have 2 of the 5 family requirements: Restaurants and Cafe

Cluster 3 - Have 2 of the 5 family requirements: Restaurants and Cafe

Cluster 4 - Have 4 of the 5 family requirements: Gym, Restaurants, Cafe, and Parks

Cluster 5 - Have 4 of the 5 family requirements: Gym, Restaurants, Cafe, and Parks

Cluster 6 - Have 2 of the 5 family requirements: Restaurants and Cafe

Cluster 7 - Have 1 of the 5 family requirements: Restaurants

Cluster 8 - Have 5 of the 5 family requirements: Grocery stores, Gym, Restaurants, Cafe, and Parks

The Cluster that better represent the family requirements is Cluster 8.

6. Conclusion

The analysis showed here, albeit simple, successfully identified 9 clusters of counties of the Washington Metropolitan area with different venues profiles that can be mapped to different "client requirements".

The cluster that best matches the client's "venue requirements" is cluster 8. It still has too many Counties, a new clustering can be performed considering only the elements of the chosen cluster. This can be done several times until there's a reduced area that eventual could mean one or a couple of counties left as the best match.

Future enhancements to this analysis to help families choose an area to live:

- New clustering to reduce search area
- include office locations of the major companies established in this area.
- include types of public transportation to the map
- include school locations and ranking to the map

Other enhancements:

- use the profession profiles of the family, and scrap results of type of profession in Google jobs (https://careers.google.com/jobs/) and map it to map.