An Analysis of Postal Zones in Montgomery County, Maryland

Capstone Project - The Battle of Neighborhoods

Melissa Qi

---

1. Introduction

Montgomery County is the most populous county in the U.S. state of Maryland, located adjacent to Washington, D.C. It is an important business and research center. Most of the county's residents live in unincorporated locales, of which the most urban are Silver Spring and Bethesda, although the incorporated cities of Rockville and Gaithersburg are also large population centers, as are many smaller but significant places (Wikipedia).

Eric is a biological scientist who is moving to Maryland from California due to the relocation of his job. The Montgomery County is a good choice because of the comfortable distance for Eric commutes to work. However, Eric is not familiar with the neighborhoods and house market as a newcomer from western coast. After an interview with Eric and his family, I learned that Eric is living with his wife and two teenage boys. And they expected to live in an urban neighborhood with grocery stores, gyms, favorite restaurants and theater for a convenient life. And park is also important for Eric's family to enjoy the weekend time. In addition, their budget for this new house in Montgomery is 1 million.

To help Eric start the house hunting easily, I designed this project to let him know the house market and common feathers in each postal zone of Montgomery county quickly. To explore the housing market and living environment, I collected the geographic data, common venue data, and average housing values of each zip code. A then, I utilized machine learning algorithm to conduct data exploratory analysis and clustering analysis, and used the Folium library to visualize the postal zones in Montgomery county and their emerging clusters. Finally, I got three most promising postal zones for Eric's family. I think the project can assist other people who have the similar background and needs as Eric to find a good neighborhood in Montgomery county.

2. Data Preparation

In this section, I introduced the data source and data collection. To develop the model, I collected geographic data, venue category data and housing price data of Montgomery county through web scraping and downloading. To create the data frame used for the analysis, I also did some data cleaning and data modification which will be shown in the coding scripts.

2.1 Description of Data and Data Source

In order to segment and explore the neighborhoods, I essentially need a dataset that contains all zip codes in Montgomery county, MD as well as the latitude and longitude coordinates of each postal area. And the data comes from the following four data source.

2.1.1 Zip Codes and City Names

The data is gathered through web scraping from https://www.zip-codes.com/county/md-montgomery.asp. And this data provides the Zip codes information

2.1.2 Demographic Data and Housing Data

The data is gathered through web scraping from https://www.zip-codes.com/county/md-montgomery.asp and the nested link for each zip code. Some modification and data cleaning will be processed in this section. This data can locate the position of each Zip Code in Montgomery county.

- **Latitude and Longitude.** The coordinates can locate each postal area and combine with venues data to explore the venue information
- **Average House Price**. The average house price of each postal area will be considered by Eric to make final decision.
- **Zip Codes**

2.1.3 Montgomery County's GeoJSON with Zip Codes and other information

The GeoJSON file of Montgomery is downloaded from https://data.montgomerycountymd.gov/Technology/Geographic-data-Zip-Codes-Shape-File-/vz4m-d8ee. The data provides the map layer with zip codes in Montgomery County, MD.

2.1.4 Venues

The data is gathered from web scraping through Foursquare API. The Foursquare location data will provide venues information and categories

2.2 Data Collection and Processing

I collected geographic data, venue category data and housing price data of Montgomery county through web scraping and downloading. During the data collection process, I did some data re-pulling and cleaning to correct some mistakes in data. As I investigated the HTML scripts, I found the demographic data in 10 postal areas are in different rows of the data. I re-pulled the Latitude and Longitude information for them and corrected the mistakes in table.

2.2.1 Zip Codes, Coordinates and Housing Values

The following table contains the Zip codes, cities, coordinates and average housing value and was used to locate postal zones and analyze housing price in Montgomery county.

ZipCode	City	State	Latitude	Longitude	Avg_House_Val
20812	GLEN ECHO	MD	38.968422	-77.14235	898000
20814	BETHESDA	MD	39.004804	-77.102477	759100
20815	CHEVY CHAS	MD	38.984212	-77.079106	985700
20816	BETHESDA	MD	38.95397	-77.135482	939700
20817	BETHESDA	MD	39.003314	-77.159528	89330

Table 1. Zip Codes, Coordinates and Housing Values

Figure 1 is the histogram of average house value in each zip code. And the red line is Eric's budget of $1 million for his new house, which is higher than the average house price in any postal zone. There, house price budget is not a constraint in selecting a neighborhood.

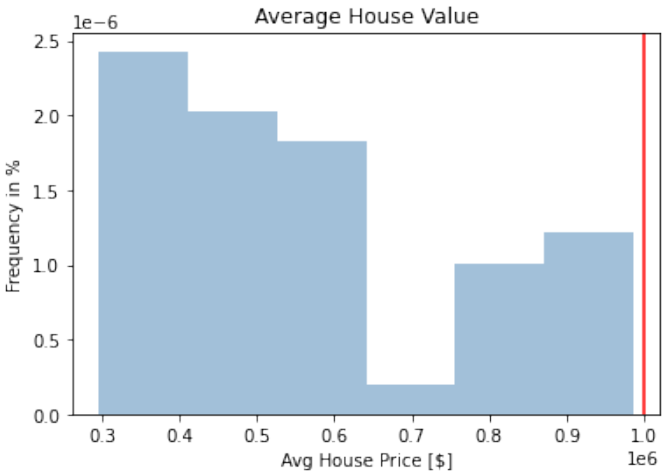

Figure 1. Average House Value Histogram

Figure 2 is the map of average house price in each zip code. And we can see the zip code when clicking the postal zone, such as 20814 in the figure. As we can see from the map, the average house value was divided into 5 groups from low to high. And the black zone is the zip codes we dropped from data collection section due to the $0 value of average house price. It is possible that there is no population or commercial houses in these areas. And the layer of postal zone came from GeoJSON file.

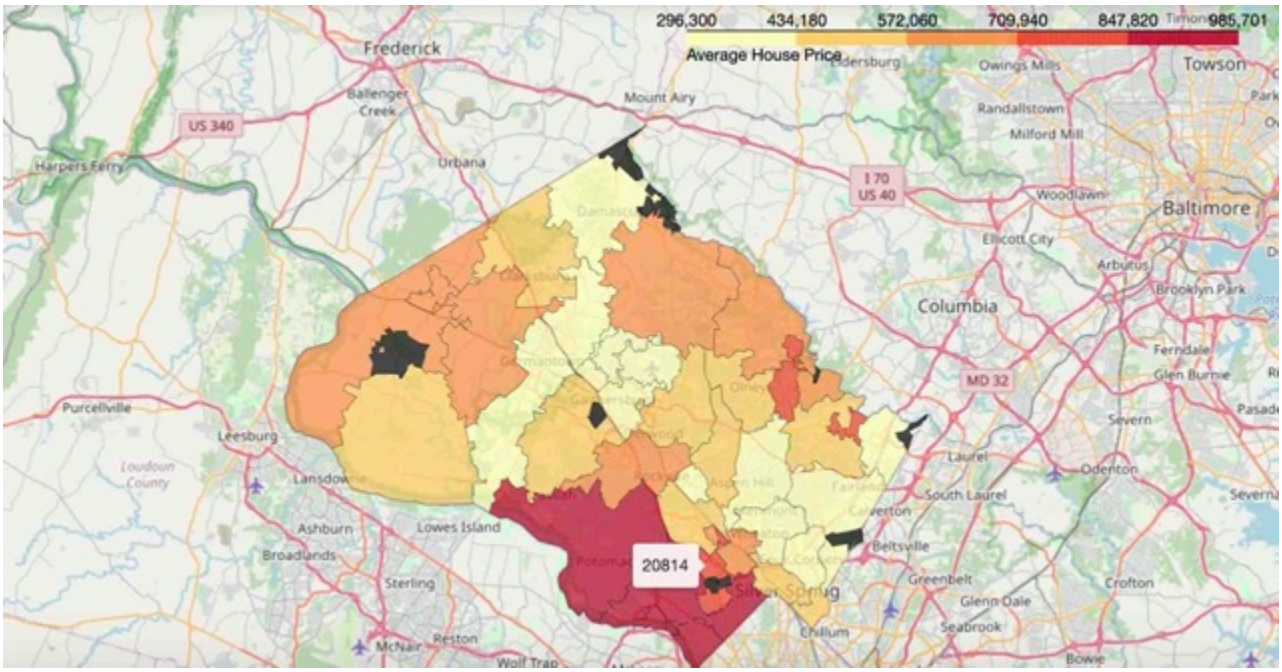

Figure 2. Map of Average House Value

2.2.2 Common Venues Data

Table 2 presents venues' demographic information and categories in Montgomery county.

ZipCode	Venue	Venue Latitude	Venue Longitude	Venue Category
20812	Adventure The	38.965779	-77.137865	Theater
20812	Glen Echo Par	38.967584	-77.139559	Park
20812	Glen Echo Spa	38.965571	-77.138895	Rock Club
20812	Chataqua Locl	38.964482	-77.13829	Trail
20812	Clara Barton N	38.968019	-77.139413	History Museum

Table 2. Montgomery Venues and Venue Categories

The blue circles on figure 3 present the venue coordinates in Montgomery county and the venues include categories such as Parks, Grocery Stores, Restaurants, etc.

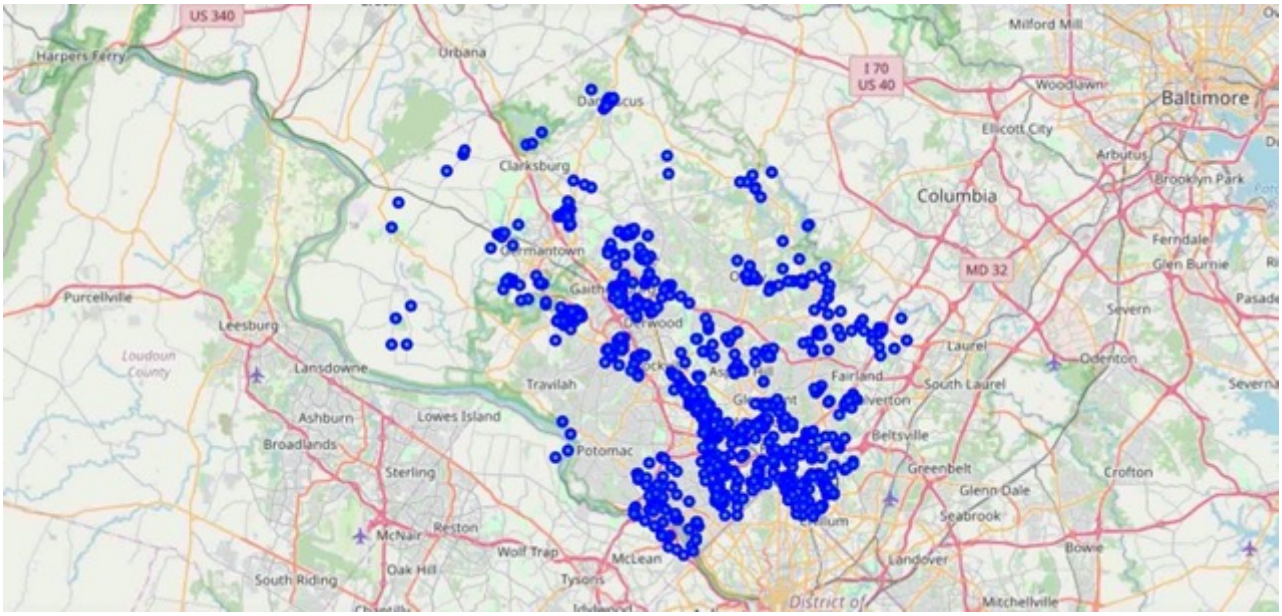

Figure 3. Map of Venues in Montgomery county

3. Methodology

I started with an exploratory data analysis to investigate the postal areas finding out the area with most of Eric's preferred venues categories. And then, I used the K-means clustering to segment the postal areas to select the neighborhoods similar to the one above.

3.1 Exploratory Data Analysis

3.1.1 Analyze each postal area

In this section, I generated the table of top 10 common venue categories of each postal zone to give Eric a review of each neighborhood.

ZipCode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
20812	Trail	History Museum	Park	National Park	Theater	Playground	Fast Food Restaurant	Bridge	Sandwich Place	Rock Club
20814	American Restaurant	Park	Pizza Place	Gym	Hotel	Intersection	Sandwich Place	Mexican Restaurant	Fast Food Restaurant	Diner
20815	Park	Mexican Restaurant	Trail	Grocery Store	Pizza Place	Bakery	Golf Course	Seafood Restaurant	Salad Place	Portuguese Restaurant
20816	Trail	Park	History Museum	Sandwich Place	Theater	Japanese Restaurant	Fast Food Restaurant	Martial Arts School	Bookstore	Stables
20817	Intersection	Golf Course	Pool	Tennis Court	Trail	Garden Center	Plaza	Spiritual Center	Baseball Field	Ice Cream Shop

Table 3. Top 10 Common Venue Categories in Each Zip Code

The map in figure 4 shows the distribution of venues in each zip code.

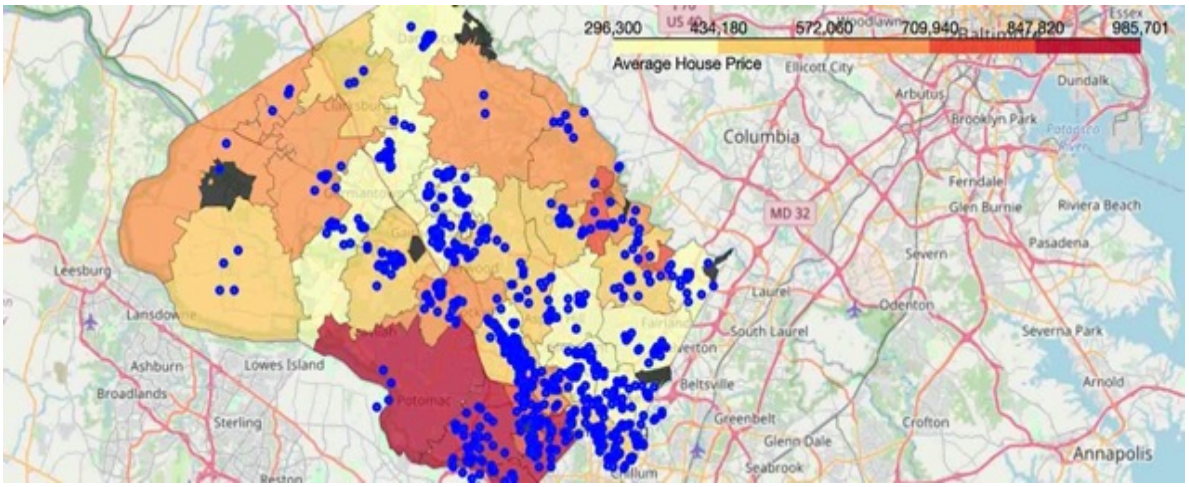

Figure 4. Venues Location in Each Zip Code

3.1.2 Analyze Areas with Preferred Categories

Eric is living with his wife and two teenage boys. They would like to live in an urban area with grocery stores, parks, gyms, favorite restaurants and theaters which can provides a convenient life. In addition, the family like to spend weekend time in parks. The following is the list of venue categories Eric's family prefers and I explored the data with these venue categories:

- Grocery Store
- Park
- Gym / Fitness Center
- Pizza Place, Mexican Restaurant, Italian Restaurant
- Theater

The map in figure 5 shows family's preferred venue categories and average house price in each zip code. And we can find most circles in city Bethesda, Chevy Chase and Silver Spring. And the house price in red area is closer to one million.

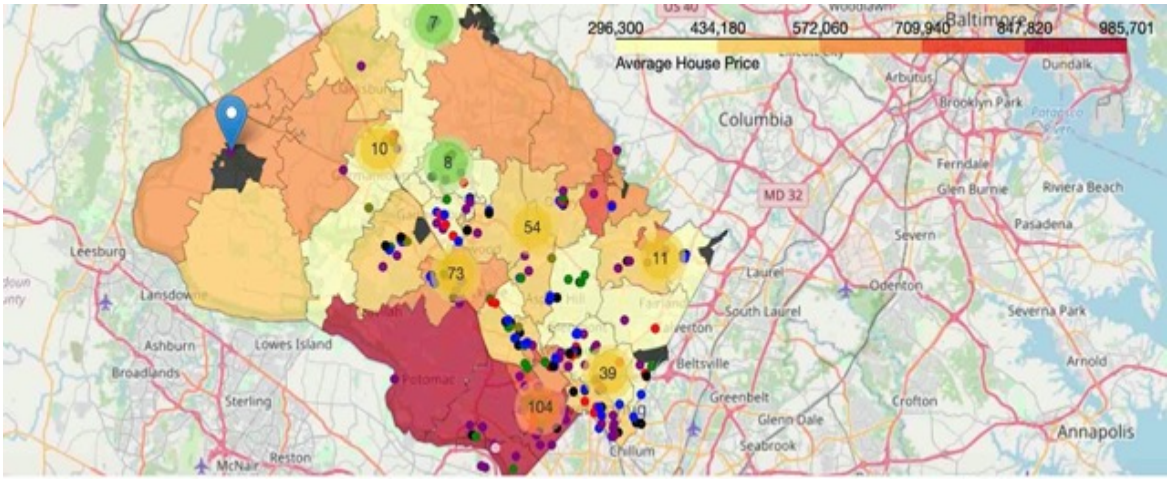

Figure 5. Preferred Venues with House Price in Each Zip Code

Table 3 includes the number of Eric's each preferred venue category and a total number of preferred categories. From the top 5 postal zones, the 20815 has the greatest number of preferred categories Eric's family expect for and the house price is closest to their budget.

ZipCode	City	Avg_House_Val	Grocery Store	Gym / Fitness Center	Italian Restaurant	Mexican Restaurant	Park	Pizza Place	Theater	Total PreCate
20815	CHEVY CHASE	985700	3	2	1	4	6	3	1	20
20910	SILVER SPRING	556000	3	1	1	5	4	3	0	17
20902	SILVER SPRING	392100	3	1	1	2	5	5	0	17
20912	TAKOMA PARK	518400	3	0	0	1	5	6	0	15
20814	BETHESDA	759100	1	1	1	2	6	3	1	15

Table 4. Top 5 Postal Zone with Most Preferred Categories

Therefore, 20815 is a good choice selected for Eric and his family from the preferred categories analysis.

3.2 K-means Clustering Analysis

K-means clustering analysis is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data.

3.2.1 Elbow Method

To determine the optimal K value for K-means clustering analysis, I used Elbow method. From the table 4, I decided the optimal k value is 5.

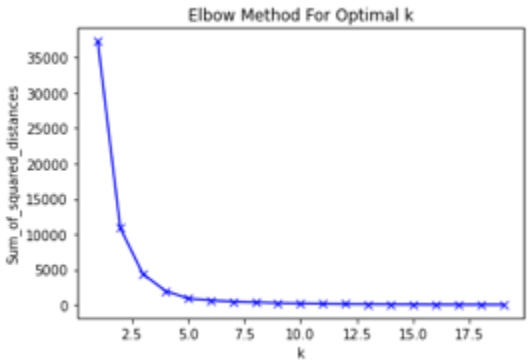

Figure 6. Elbow Method for Optimal K

3.2.2 K-means Cluster Results

There are 5 clusters resulting from the K-means clustering analysis. And figure 6 is a map presents the 5 clusters in Montgomery county.

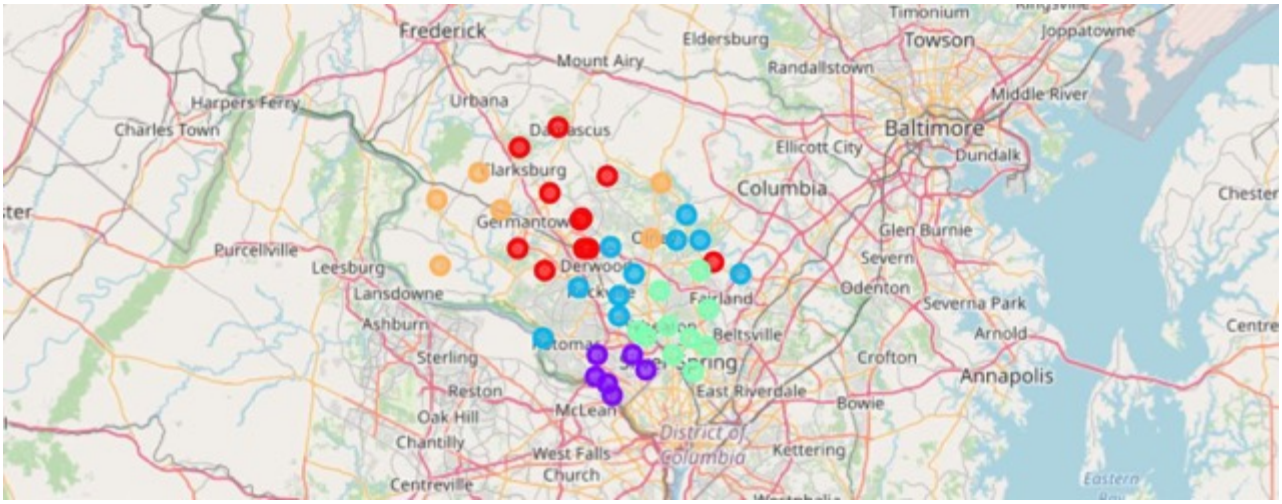

Figure 7. K-means Cluster Results

4. Results

The following cluster includes postal area 20815 which is the zone selected from preferred categories analysis section. There are another four zip codes in this cluster. To learn more about the venue categories of these zip codes, I added the number of preferred categories and total number to these zip codes in Table 5.

ZipCode	City	Avg_House_Val	Grocery Store	Gym / Fitness Center	Italian Restaurant	Mexican Restaurant	Park	Pizza Place	Theater	Total PreCate
20815	CHEVY CHASE	985700	3	2	1	4	6	3	1	20
20814	BETHESDA	759100	1	1	1	2	6	3	1	15
20812	GLEN ECHO	898000	1	1	1	1	2	1	2	9
20816	BETHESDA	939700	0	1	0	0	4	0	2	7
20818	CABIN JOHN	884900	1	1	1	1	2	1	0	7

Table 5. Cluster Results

From the clustering analysis, I found two more postal zones, 20814 and 20812, that can satisfy Eric's all expectations. After the analysis, zip codes 20815, 20814 and 20812 are all good choices for Eric and his family which have all of their preferred categories and the house price close to the budget.

5. Discussion and Recommendation

After the preferred category analysis and the K-means clustering analysis, I finally recommend postal zones 20815, 20814 and 20812 for Eric and his family to move in. Firstly, after the preferred category analysis, I selected the top one zone 20815 as the first choice. It contains most number of categories that Eric likes. And through the K-clustering analysis, I got another 4 zones similar to 20815. They all contain the family's preferred venue categories and are all located in urban areas. These five zones are very similar in living environment and housing price. They are all convenient for a family to live in. In addition, the house prices are all close to their one million dollars budget. From the list of five zones, I chose three of them including 20815, 20814 and 20812 since they contain all preferred categories the family expect for and dropped the other two since they have more than one category empty.

6. Conclusion

This analysis provides a good start for Eric and his family to know more about Montgomery county in Maryland. And the next step for them is to take a look at the environment around them. This analysis can only provide the results that satisfy their expectation for the venues, living environment and average housing prices, but didn't give more information such as the house for sales. Therefore, the project can help narrow down Eric's choice, and it is open to add more requirements, information to improve it. It will be helpful for people new to Montgomery county to start their search for new home.