IBM DATA SCIENCE CAPSTONE PROJECT

Housing Analysis of Rockville, MD

Hanh Ta

August 4, 2020

Table of Contents

1. Introduction

1.1. Background

Rockville is one of Maryland's oldest towns, with its origin dating back to Colonial America. During Revolutionary times, Rockville was known as Hungerford's Tavern, the name of its most familiar business and landmark. Rockville has grown very rapidly since its founding in 1803, from a leisurely, agriculturally-oriented county seat to a cosmopolitan city of many neighborhoods. Located approximately 14 miles north of Washington D.C., Rockville is home to a well-educated diverse population and serves as an employment center for national biomed corporations, the federal government and county government. Restaurants and shops are clustered around Rockville and its surrounding neighborhoods, with multitude of parks and recreation centers.

Due to its rapid growth, the city's population is changing in some dramatic ways, which drive the housing needs of residents and workers including growing demand for rental housing, lower-cost housing and housing that can accommodate multi-generational households. The population of single-person households, including young people starting out in their careers has increased substantially. This population of young adults includes a significant number of people who may be heading towards marriage, children, and potentially, homeownership.

1.2. Problem

Buying a home is an emotional process and a major commitment. One can find his/her home through a local real estate agent or via Multiple Listing Service (MLS) listings websites nowadays. Oftentimes it is challenging to have a clear set of preferences as first-time home buyers. As a result, this project aims to provide a big picture of the housing market (i.e. price range, home conditions, etc.) of Rockville and the surrounding neighborhoods.

1.3. Interest

The interests of this project are followed:

- Analyzing the distribution of single-home housing in Rockville and its nearby neighborhoods
- Determining the common features that drive housing cost
- Exploring the most popular venues around the areas of interest

2. Data Acquisition and Cleaning

2.1. Data sources

Multiple MLS-based listing websites such as Zillow provide real estate APIs but these databases are not open-source and can only be accessed with an agent's credentials. I was able to download current property listings and sold properties data via Redfin for different neighborhoods; however, these are fixed datasets which were updated as of July 1st, 2020. These datasets were then used to analyze the distribution and explore most common housing features. For further exploration of venues, I utilized Foursquare location data to obtain a list of popular venues and their locations. The original datasets consist of 27 features which are described in Table 1.

Table 1. List of Features and Descriptions

Feature	Description
SALE TYPE	Type of sale
SOLD DATE	Date when the property was sold
PROPERTY TYPE	Type of property
ADDRESS	Address of property
CITY	City of property
STATE OR PROVINCE	State of property
ZIP OR POSTAL CODE	Zip code of property
PRICE	Listing price of property
BEDS	Number of beds of property
BATHS	Number of baths of property
LOCATION	Location of property
SQUARE FEET	Area of property in square feet
LOT SIZE	Lot size of property
YEAR BUILT	Year when the property was built
DAYS ON MARKET	Number of days the property was on market
$/SQUARE FEET	Price per square feet of property
HOA/MONTH	Monthly homeowners association fee
STATUS	Status of listing
NEXT OPEN HOUSE START TIME	Open house start time
NEXT OPEN HOUSE END TIME	Open house end time
URL (SEE http://www.redfin.com/buy-a-home/comparative-market-analysis FOR INFO ON PRICING)	Link to property listing
SOURCE	Source of listing
MLS#	Multiple Listing Service identification number
FAVORITE	Y/N whether property was marked as favorite
INTERESTED	Y/N whether property was marked as interested
LATITUDE	Coordinate of property

LONGITUDE	Coordinate of property

Figure 1. Example of dataset.

	SALE TYPE	SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	STATE OR PROVINCE	ZIP OR POSTAL CODE	PRICE	BEDS	BATHS	...	STATUS	NEXT OPEN HOUSE START TIME	NEXT OPEN HOUSE END TIME	URL (SEE http://www. home/comparative-ma
0	MLS Listing	NaN	Single Family Residential	9100 Chesley Knoll Ct	Gaithersburg	MD	20879	420000	5.0	3.0	...	Active	NaN	NaN	http://www.redfin.com/MD/
1	PAST SALE	April-13-2020	Single Family Residential	7810 Warfield Rd	Gaithersburg	MD	20886	845000	6.0	6.5	...	Sold	NaN	NaN	http://www.redfin.com/MD/
2	PAST SALE	NaN	Single Family Residential	12136 Pawnee Dr	Gaithersburg	MD	20878	570000	NaN	2.5	...	NaN	NaN	NaN	http://www.redfin.com/MD/C

2.2. Data cleaning

Data downloaded for different neighborhoods were combined into one table. After merging, there were 1090 samples and 27 features in the data. It is necessary to check for duplicates and remove duplicate listings based on the property's unique address. There are 172 duplicate listings removed from the dataframe. Some of the column names carry special characters such as $,/, and #. In the next step, the columns were renamed as needed.

2.2.1. Feature Selection

Some attributes that were not necessarily used for future analysis and needed to be dropped from the table. For example, in this project the table is retrieved solely of single-family residential data; therefore, *PROPERTY TYPE* attribute is redundant. Additionally, I eliminated other attributes to simplify the table and discard missing values, such as:

- SOLD DATE
- NEXT OPEN HOUSE END TIME
- NEXT OPEN HOUSE START TIME
- FAVORITE
- INTERESTED
- LOCATION
- SOURCE
- STATUS
- DAY ON MARKET
- MLS#

Table 2. Feature selection during data cleaning.

Dropped Feature	Reason for Dropping Feature
SOLD DATE, NEXT OPEN HOUSE END TIME, NEXT OPEN HOUSE START TIME, SOURCE, DAY ON MARKET	Unnecessary time data for the scope of this project
LOCATION	Redundant data
MLS#, SOURCE, FAVORITE, INTERESTED	Trivial data

2.2.2. Missing Data

A table was generated to present the total number of missing values, and I also calculated the percentage of null values in each column. There were six features that exhibit missing values, the

top two of the list, *HOA PER MONTH* and *BEDS*, exhibit prevalent missing data of 78% and 27% respectively. In this case, we should not discard missing data because it would reduce a large number of samples in the data. Therefore, these features needed to be handled with different strategies.

First of all, missing values in the *HOA PER MONTH* feature were filled by zero since most of all single-family houses are fee-free. Secondly, features like *BEDS*, *PRICE PER SQFT*, *SQUARE FEET*, and *LOT SIZE* were filled by the sample median. Finally, missing values in *YEAR BUILT* were then filled with the mode of the sample.

Figure 2. Missing values.

	Total	Missing (%)
HOA PER MONTH	716	77.995643
BEDS	249	27.124183
PRICE PER SQFT	8	0.871460
SQUARE FEET	8	0.871460
YEAR BUILT	6	0.653595
LOT SIZE	3	0.326797
LONGITUDE	0	0.000000
STATE OR PROVINCE	0	0.000000
PROPERTY TYPE	0	0.000000
ADDRESS	0	0.000000
CITY	0	0.000000
BATHS	0	0.000000
ZIP OR POSTAL CODE	0	0.000000
PRICE	0	0.000000
LATITUDE	0	0.000000
URL	0	0.000000
SALE TYPE	0	0.000000

3. Exploratory Data Analysis

3.1. Outliers Detection

Before performing further exploratory analysis, we needed to handle outliers that skewed the distribution of some main features. A histogram was plotted to display the distribution of different features shown in figure 3. Features of interest such as *BATHS, LOT SIZE*, and *PRICE* are heavily skewed. Hence, two types of filters were then used to alleviate the asymmetry of the probability distribution.

The interquartile range (IQR) is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 - Q1. A function was created to filter out outliers that lie outside of the IQR range.

Another method used to detect and filter outliers is Six Sigma. Sigma represents the population standard deviation, which is a measure of the variation in a dataset. Statistically, about 99.5% of

data falls within ±3σ from μ (mean). Thus, for fairly normally distributed data, 6σ filter yields a good resultant dataset. Both filters were applied to the data and the outcomes were then compared in Table 3. Overall, filtered dataset yielded after applying IQR filter on feature *LOT SIZE* provided the best results of normal distribution of numerical columns.

Figure 3. Distribution of numerical features.

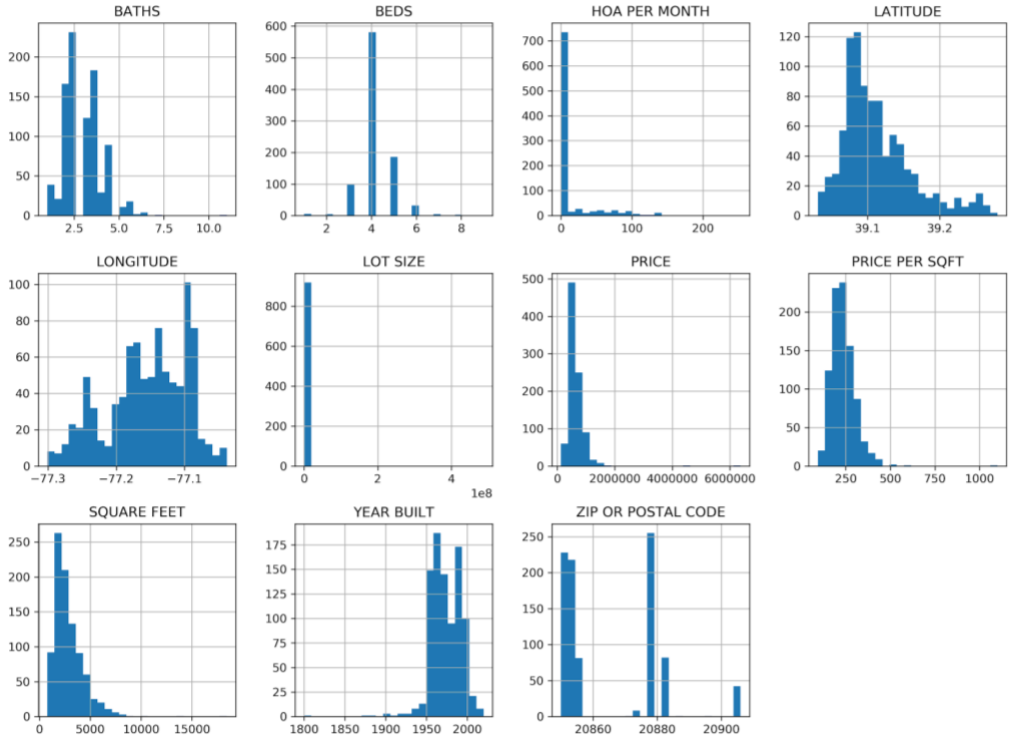


Table 3. Outlier filters comparison.

Filtered Feature	Filter	Outputs
BATHS	Six Sigma	Mean:2.959, Sigma:1.046, 3sigma_range: -0.18:6.1 Number of rows dropped: 6
	IQR	Median:3.0, 25%:2.5, 75%:3.5, IQR:-2.0:10.0 Number of rows dropped: 1
HOA PER MONTH	Six Sigma	Mean:13.776, Sigma:33.219, 3sigma_range: -85.88:113.43 Number of rows dropped: 23
	IQR	Median:0.0, 25%:0.0, 75%:0.0, IQR:0.0:0.0 Number of rows dropped: 918
LOT SIZE	Six Sigma	Mean:564489.243, Sigma:16087103.46, 3sigma_range: -47696821.14:48825799.62 Number of rows dropped: 1
	IQR	Median:11197.0, 25%:8516.75, 75%:20624.5, IQR:-5836.5:52446.0

		Number of rows dropped: 114
PRICE	Six Sigma	Mean:623048.144, Sigma:316768.755, 3sigma_range: -327258.12:1573354.41 Number of rows dropped: 4
	IQR	Median:560000.0, 25%:450000.0, 75%:743750.0, IQR:-340000.0:2047500.0 Number of rows dropped: 2
YEAR BUILT	Six Sigma	Mean:1973.3, Sigma:18.798, 3sigma_range: 1916.91:2029.7 Number of rows dropped: 14
	IQR	Median:1971.0, 25%:1962.0, 75%:1987.0, IQR:-1953.0:5945.0 Number of rows dropped: 6
SQUARE FEET	Six Sigma	Mean:2871.089, Sigma:1508.081, 3sigma_range: -1653.15:7395.33 Number of rows dropped: 10
	IQR	Median:2454.5, 25%:1890.75, 75%:3552.5, IQR:-1327.0:9559.5 Number of rows dropped: 3

Figure 4. Interquartile range (IQR).

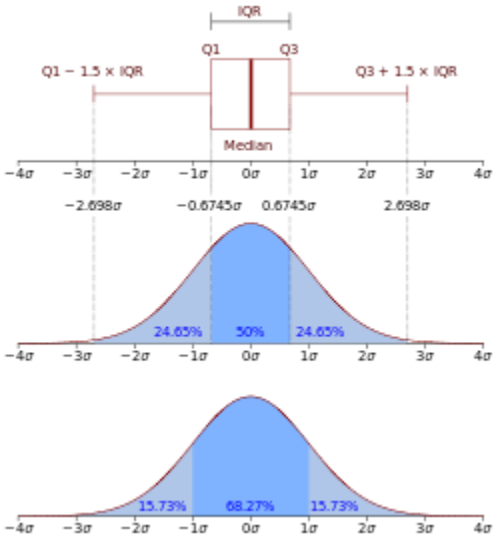


Figure 5. Six sigma distribution.

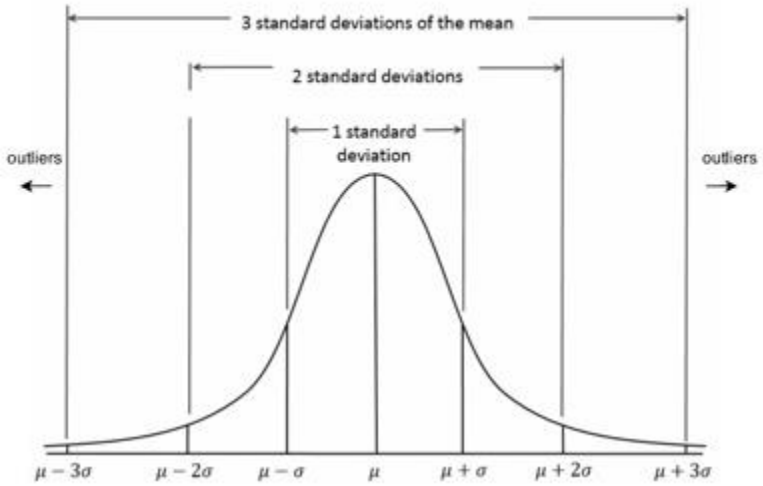
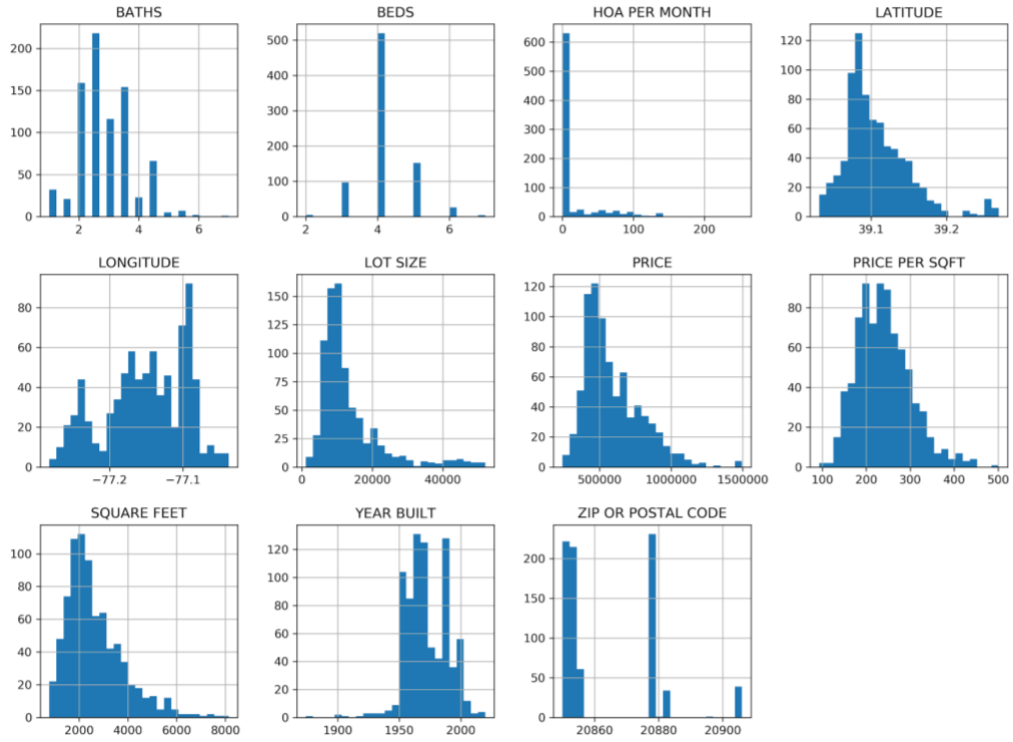


Figure 6. Filtered distribution of numerical features.



3.2. Relationship Between Variables

In this dataset, the majority of property listings are located in Rockville, followed by Gaithersburg, North Potomac, and Silver Spring. Most of them are single family residentials since it is our interested type of property and have on average two and a half bathrooms and four bedrooms.

Looking at the correlation matrix of numeric features, the number of bedrooms, the number of bathrooms, property's square footage, and property's year built demonstrate a strong positive relationship with listing price. The correlation coefficients of those features with listing price are 0.34, 0.69, 0.80, and 0.5 respectively. On the other hand, the price per square feet has a negative relationship with the total listing price besides property's latitude and longitude.

Aside from listing price, the matrix indicates strong positive correlation between property's square footage and number of bathrooms with the correlation coefficient of 0.80. Year built and square footage also yield positive correlation coefficient of 0.59.

Figure 7. Frequency of Neighborhood and Property Type.

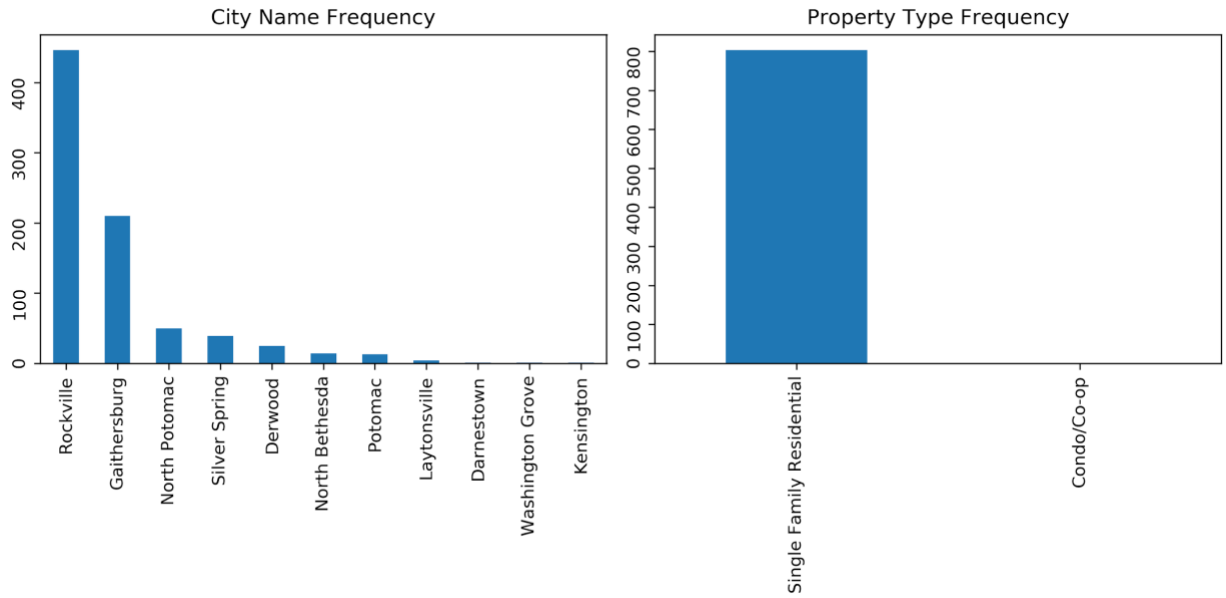
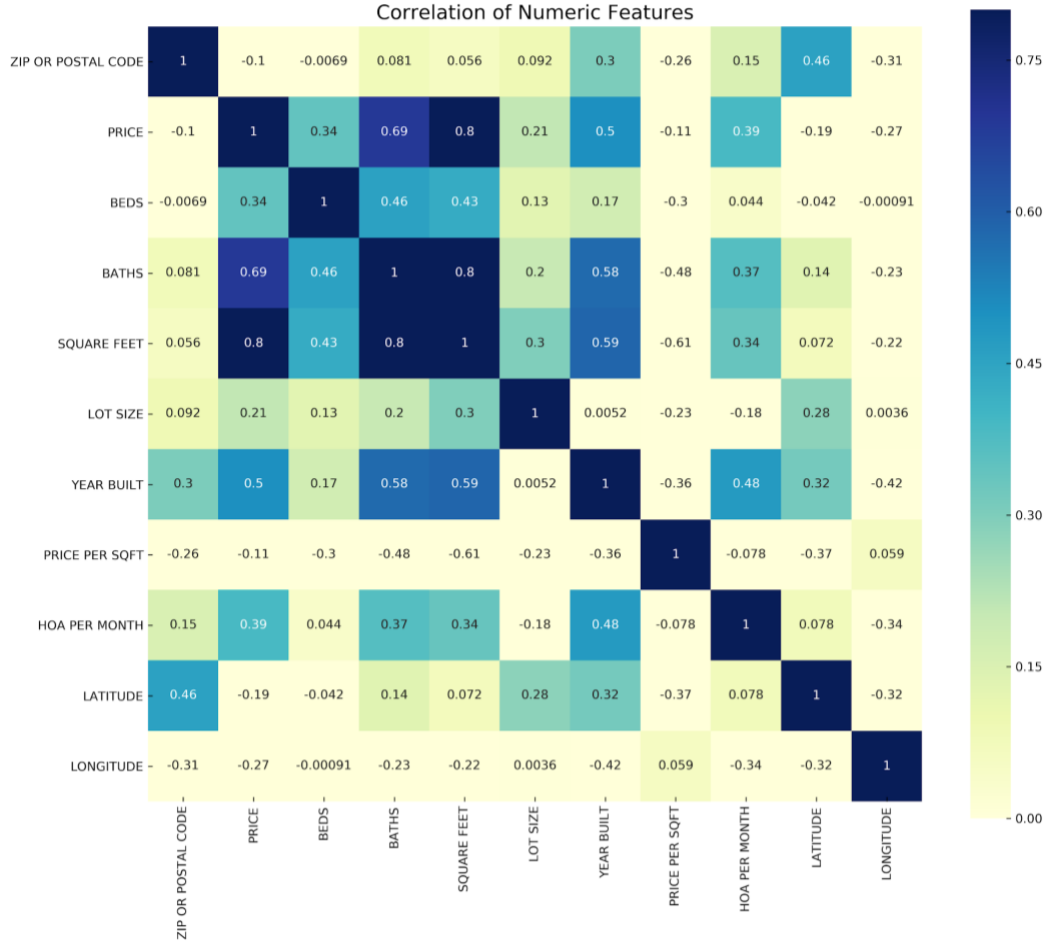


Figure 8. Correlation matrix of Numeric features.



3.3. Price Variation among Neighborhoods

In order to have a better understanding of the sale price, we can compare the range and distribution of the listing price among the neighborhoods. We observed that there is a greater variability in Rockville as well as larger outliers. Rockville has a relatively low average listing price compared to North Bethesda whose average sale price is above $800,000. Average listing prices in Gaithersburg, Rockville, and Silver Spring are almost constant; however, Silver Spring market is not as versatile as Rockville and Gaithersburg market. It is inclusive about Darnestown, Kensington, and Washington Grove since the number of listings in these neighborhoods are significantly small in our dataset.

We observe an interesting insight looking at the price per square feet distribution. Even though North Potomac and Potomac have higher median listing prices, the price per square feet of these neighborhoods is actually lower than Rockville. On average, properties in Rockville are priced at $250 per square foot, higher than Derwood, Gaithersburg, and Silver Spring. In other words, houses in Rockville have better value per square foot of livable area.

Another feature that has a high correlation with listing price is year built. Figure 11 displays the distribution of year built which we can use to compare housing among different neighborhoods. As discussed, Gaithersburg, Silver Spring and Rockville are within the same price range, but Gaithersburg is the “newest” neighborhood with the majority of houses built after 1980. Most of the houses in North Potomac were also built during this time. We observed many historic properties located in Rockville and Gaithersburg with the oldest property built before 1880.

Figure 9. Sale Price among Neighborhoods.

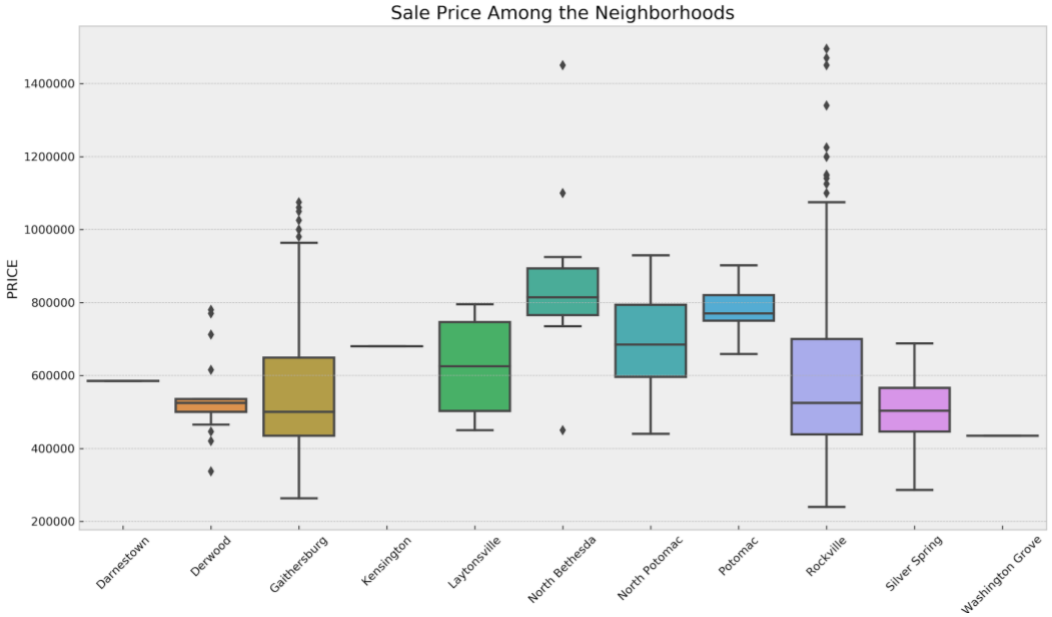


Figure 10. Price Per Square Feet among the Neighborhoods.

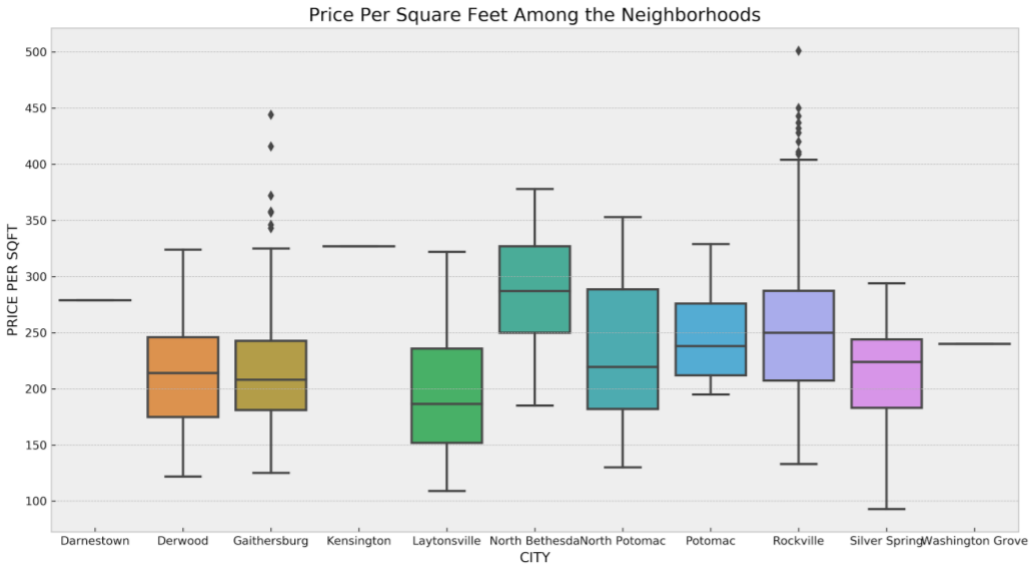
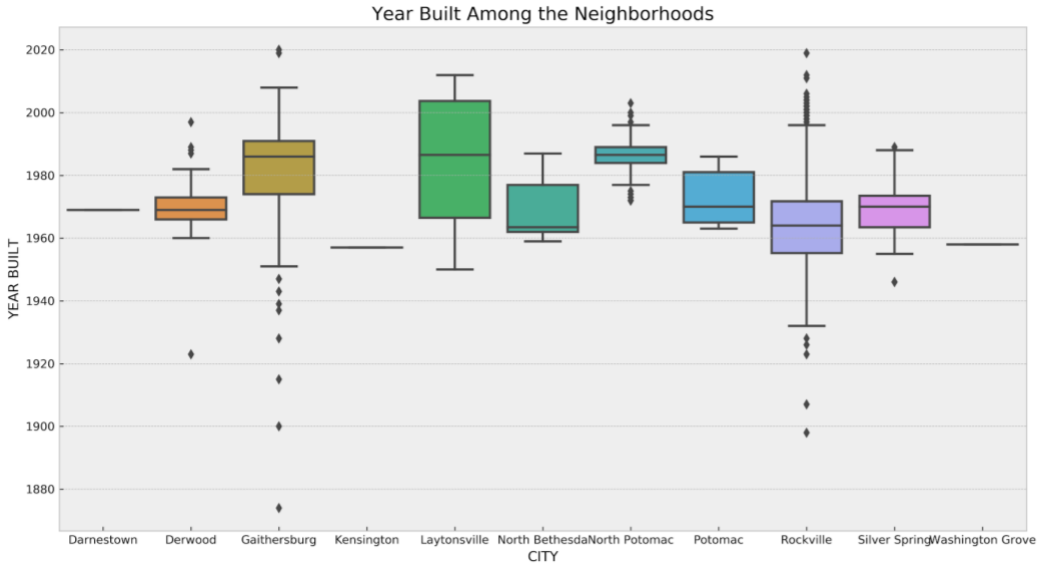


Figure 11. Year Built among Neighborhoods.



4. Spatial Distribution

After learning about the price range among neighborhoods, we can look at the spatial distribution of property listings based on listing price and other correlated features. The sale price was color coded, based on the IQR range, as shown below:

- Blue: Under $444,500
- Green: $444,500 - $533,500
- Yellow: $533,500 - $700,000
- Red: Over $700,000

The majority of “blue” and “green” properties are located in Gaithersburg and southeast of Rockville. It is interesting to observe that the distribution of housing prices is remarkably lower on the east side of the Washington Metropolitan Area Transit Authority (WMATA) metro line. Houses located near main streets and highways as well as greenery areas are more expensive.

Figure 12. Spatial Distribution of Housing Price

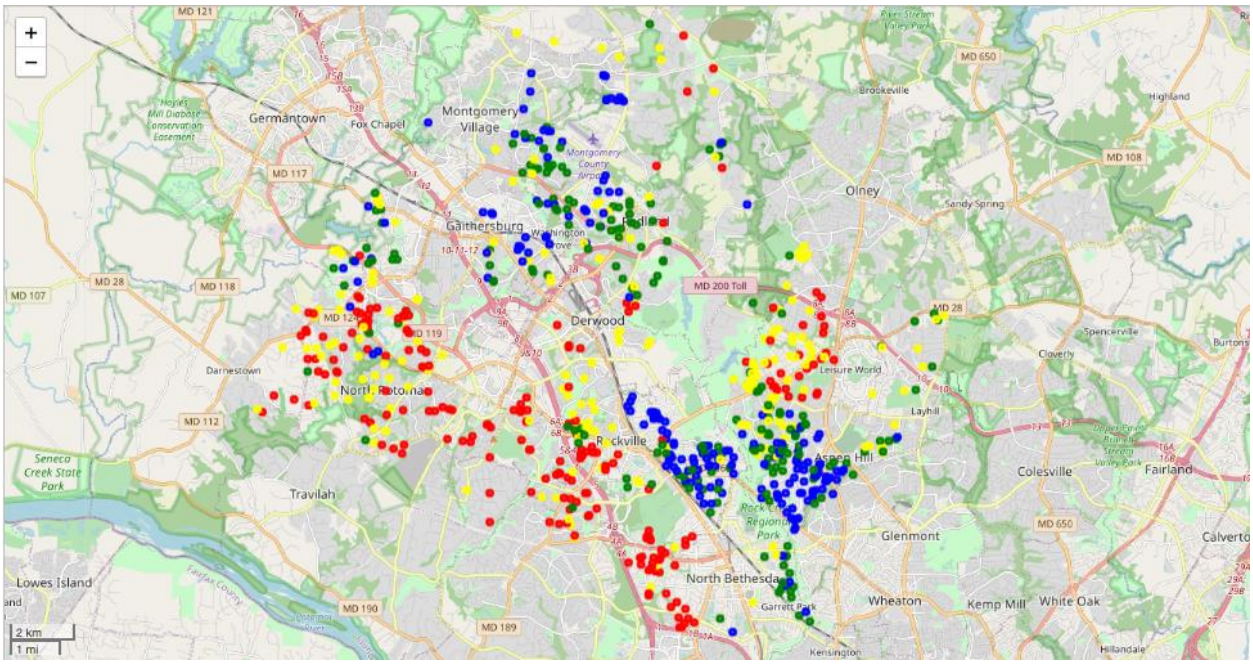


Figure 13. Spatial Distribution of Price Per Square Feet.

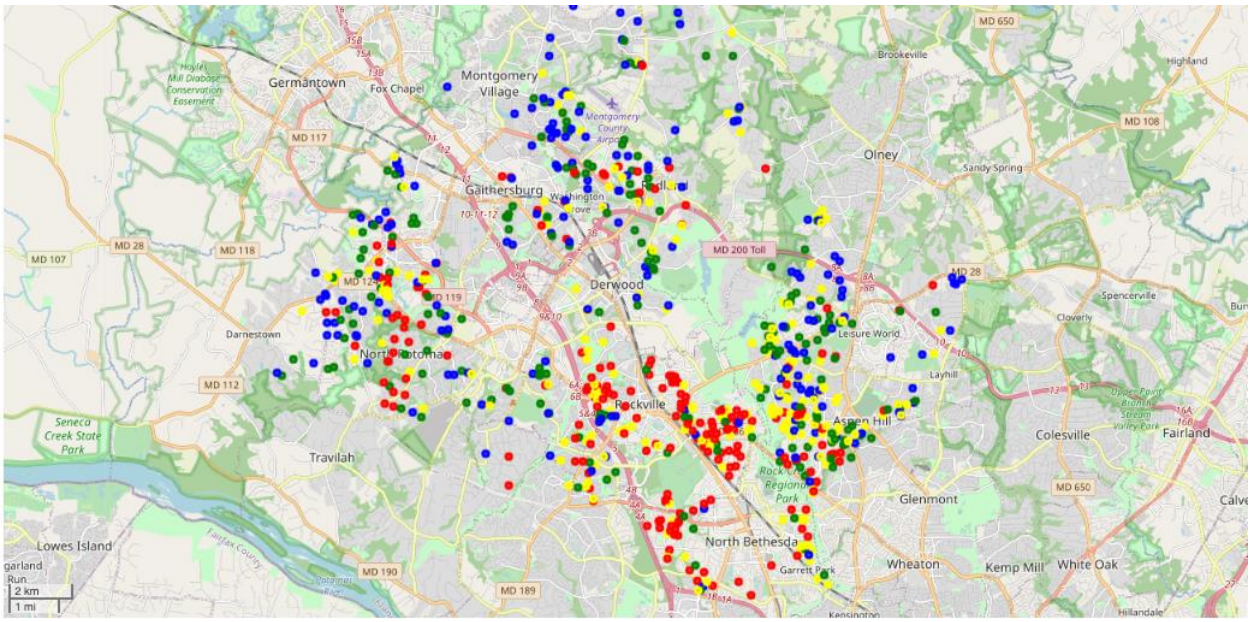
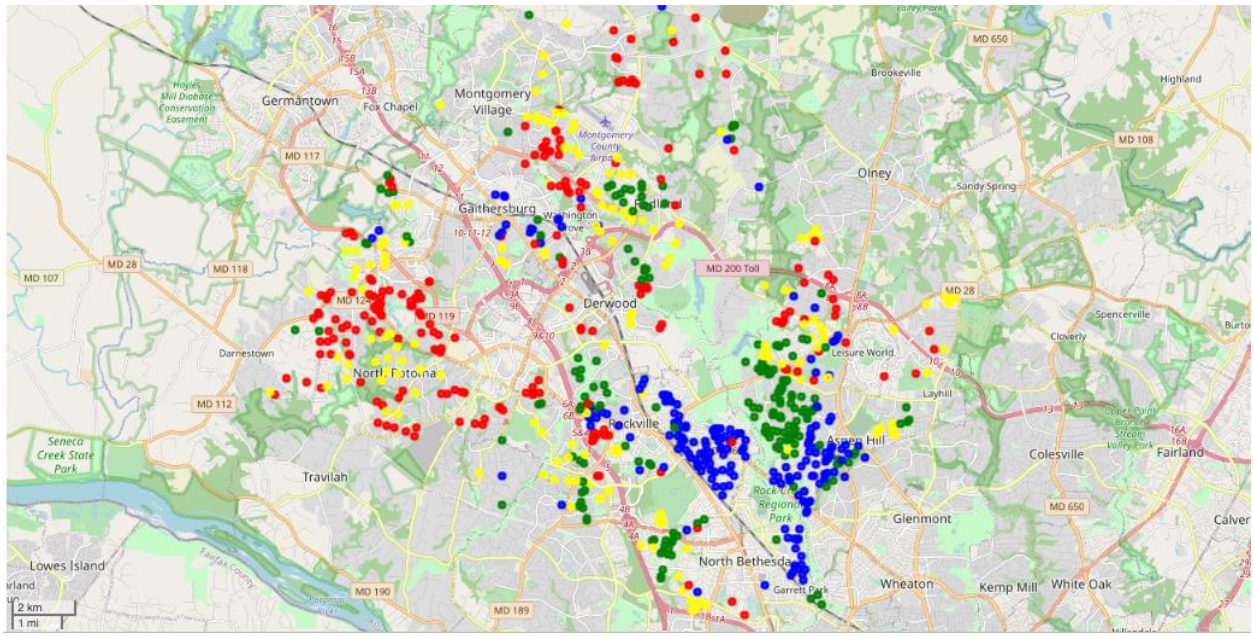


Figure 14. Spatial Distribution of Year Built.



5. Exploration of Nearby Venues

5.1. Venue Data Collection and Exploration

In the next step of the analysis, the cities were explored in greater detail. A list of popular venues in each city was collected via Foursquare API. An URL was constructed to send requests to the API to search for nearby venues, their locations and categories. After arranging the data, we had up to 100 venues for each neighborhood. Venues are collected within a radius of 6000 meters (3.7 miles) from the point of city coordinates. The collected data is shown in Table 4. We could also check for the summary of the number of venues that had been collected for each city.

We used one-hot encoding to explore the venue categories. This created dummy variables for the categories in order to be applied for machine learning. Thus, there were 172 unique categories collected from popular venues in our neighborhoods. In addition, we grouped these venues based on similar categories and calculated the frequency of occurrence of each category. Thus, we were able to construct a table with a list of neighborhoods and their ten most common venues as shown in Table 6.

Table 4. Nearby Venues around Neighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Gaithersburg	39.143441	-77.20137	Sardi's Pollo a la Brasa	39.147017	-77.205992	Latin American Restaurant
1	Gaithersburg	39.143441	-77.20137	La Casita Pupuseria & Cocina C.A.	39.142039	-77.198903	Latin American Restaurant
2	Gaithersburg	39.143441	-77.20137	Old Town Cafe	39.142541	-77.193474	Café
3	Gaithersburg	39.143441	-77.20137	Bohrer Park	39.132643	-77.193522	Park
4	Gaithersburg	39.143441	-77.20137	Tortacos	39.153013	-77.197087	Taco Place
5	Gaithersburg	39.143441	-77.20137	Ixtapalapa Taqueria	39.146487	-77.204229	Taco Place
6	Gaithersburg	39.143441	-77.20137	The Vitamin Shoppe	39.144836	-77.203069	Supplement Shop
7	Gaithersburg	39.143441	-77.20137	ProFIT Club	39.142379	-77.193276	Gym / Fitness Center
8	Gaithersburg	39.143441	-77.20137	Dogfish Head Alehouse	39.141832	-77.217272	Bar
9	Gaithersburg	39.143441	-77.20137	99 Ranch Market	39.149857	-77.205941	Supermarket

Table 5. Number of Collected Venues.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Darnestown	100	100	100	100	100	100
Derwood	100	100	100	100	100	100
Gaithersburg	100	100	100	100	100	100
Kensington	100	100	100	100	100	100
Laytonsville	52	52	52	52	52	52
North Bethesda	100	100	100	100	100	100
North Potomac	100	100	100	100	100	100
Potomac	99	99	99	99	99	99
Rockville	100	100	100	100	100	100
Silver Spring	100	100	100	100	100	100
Washington Grove	100	100	100	100	100	100

Table 6. Most Common Venues for Each Neighborhood.

	Neighborhood	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
0	Darnestown	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
1	Derwood	Gym / Fitness Center	Pizza Place	Park	Donut Shop	Chinese Restaurant	American Restaurant	Mexican Restaurant	Hotel	Ice Cream Shop	Coffee Shop
2	Gaithersburg	Mexican Restaurant	Pizza Place	American Restaurant	Supermarket	Sandwich Place	Café	Coffee Shop	Bar	Gym / Fitness Center	Donut Shop
3	Kensington	Pizza Place	Park	Vietnamese Restaurant	Trail	Bagel Shop	Farmers Market	Grocery Store	Mexican Restaurant	Gym / Fitness Center	Movie Theater
4	Laytonsville	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
5	North Bethesda	Grocery Store	American Restaurant	Pizza Place	Coffee Shop	Bakery	Salad Place	Gym / Fitness Center	Supermarket	Cosmetics Shop	New American Restaurant
6	North Potomac	Pizza Place	Mexican Restaurant	Sushi Restaurant	American Restaurant	Coffee Shop	Bar	Sandwich Place	Gym	Grocery Store	Seafood Restaurant
7	Potomac	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
8	Rockville	Grocery Store	Gym / Fitness Center	American Restaurant	Ice Cream Shop	Donut Shop	Pizza Place	Persian Restaurant	Burger Joint	Café	Salad Place
9	Silver Spring	Grocery Store	Latin American Restaurant	American Restaurant	Bakery	Brewery	Coffee Shop	Trail	Mexican Restaurant	Vietnamese Restaurant	Indian Restaurant
10	Washington Grove	Mexican Restaurant	American Restaurant	Pizza Place	Gym / Fitness Center	Ice Cream Shop	Coffee Shop	Donut Shop	Bar	Supermarket	Park

5.2. Clustering

Once we had the dataset of surrounding venues, we performed clustering. Unsupervised machine learning technique was used based on K-means. For K-means clustering, first we decided on the number of clusters. To avoid the trial and error approach, the silhouette score was used. As shown in Figure 15, the optimal number of clusters is six with the highest silhouette score. In the next step, we ran the K-means clustering algorithm with the parameter of 6 as the number of clusters. Furthermore, we visualized the clusters on the map in Figure 16.

Figure 15. Selecting the Number of Clusters.

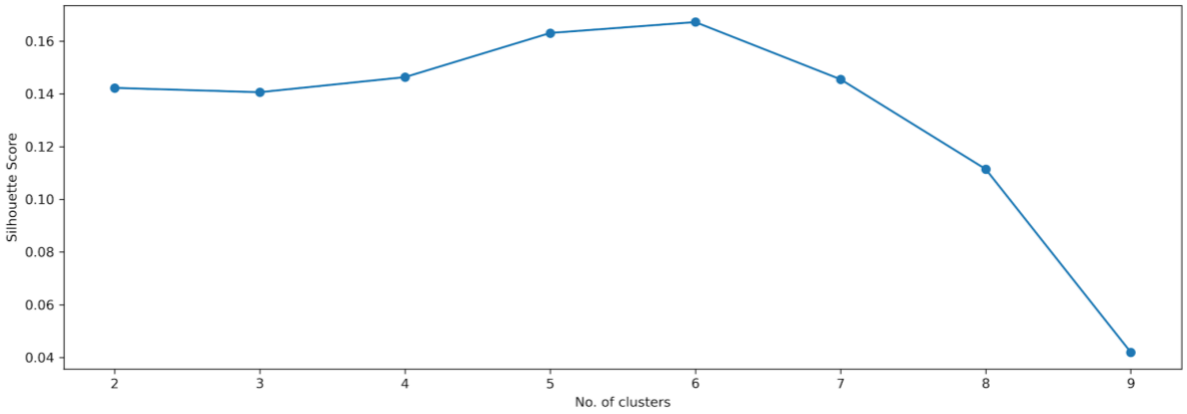
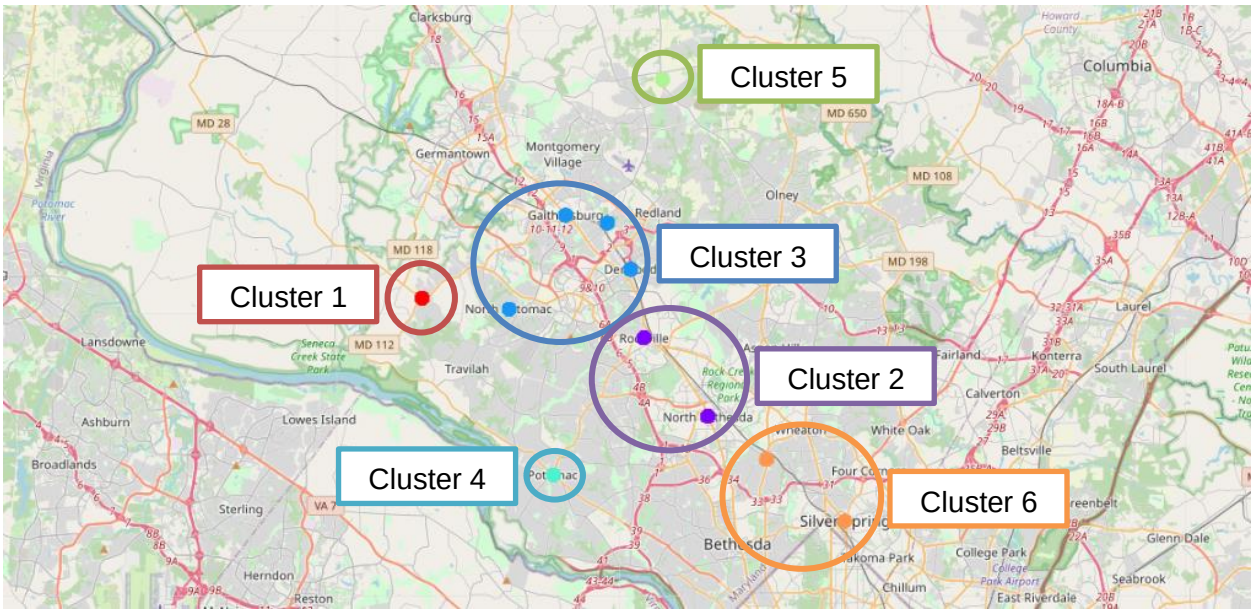


Figure 16. Visualizing Clusters of Venues.



5.3. Cluster Results

5.3.1. Cluster 1 (Red)

	Neighborhood Latitude	Venue Longitude	Venue Category	Cluster Labels	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
252	39.103441	-77.253840	Pizza Place	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
253	39.103441	-77.291616	Grocery Store	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
254	39.103441	-77.267962	Farm	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
255	39.103441	-77.250068	Market	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
256	39.103441	-77.291697	Coffee Shop	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
347	39.103441	-77.314519	Scenic Lookout	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
348	39.103441	-77.248945	Trail	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
349	39.103441	-77.248944	Farm	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
350	39.103441	-77.273340	Grocery Store	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant
351	39.103441	-77.263643	Grocery Store	0	Pizza Place	Grocery Store	Coffee Shop	Chinese Restaurant	Trail	Farm	Fast Food Restaurant	Sandwich Place	Thai Restaurant	Restaurant

100 rows × 14 columns

5.3.2. Cluster 2 (Purple)

	Neighborhood Latitude	Venue Longitude	Venue Category	Cluster Labels	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
452	39.084005	-77.153057	Deli / Bodega	1	Grocery Store	Gym / Fitness Center	American Restaurant	Ice Cream Shop	Donut Shop	Pizza Place	Persian Restaurant	Burger Joint	Café	Salad Place
453	39.084005	-77.150672	Shopping Plaza	1	Grocery Store	Gym / Fitness Center	American Restaurant	Ice Cream Shop	Donut Shop	Pizza Place	Persian Restaurant	Burger Joint	Café	Salad Place
454	39.084005	-77.153266	Bubble Tea Shop	1	Grocery Store	Gym / Fitness Center	American Restaurant	Ice Cream Shop	Donut Shop	Pizza Place	Persian Restaurant	Burger Joint	Café	Salad Place
455	39.084005	-77.150218	Hotel	1	Grocery Store	Gym / Fitness Center	American Restaurant	Ice Cream Shop	Donut Shop	Pizza Place	Persian Restaurant	Burger Joint	Café	Salad Place
456	39.084005	-77.143439	Gym / Fitness Center	1	Grocery Store	Gym / Fitness Center	American Restaurant	Ice Cream Shop	Donut Shop	Pizza Place	Persian Restaurant	Burger Joint	Café	Salad Place
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
847	39.046129	-77.145936	Cosmetics Shop	1	Grocery Store	American Restaurant	Pizza Place	Coffee Shop	Bakery	Salad Place	Gym / Fitness Center	Supermarket	Cosmetics Shop	New American Restaurant
848	39.046129	-77.158043	Gym	1	Grocery Store	American Restaurant	Pizza Place	Coffee Shop	Bakery	Salad Place	Gym / Fitness Center	Supermarket	Cosmetics Shop	New American Restaurant
849	39.046129	-77.140792	Thai Restaurant	1	Grocery Store	American Restaurant	Pizza Place	Coffee Shop	Bakery	Salad Place	Gym / Fitness Center	Supermarket	Cosmetics Shop	New American Restaurant
850	39.046129	-77.146109	Shopping Mall	1	Grocery Store	American Restaurant	Pizza Place	Coffee Shop	Bakery	Salad Place	Gym / Fitness Center	Supermarket	Cosmetics Shop	New American Restaurant
851	39.046129	-77.076768	Coffee Shop	1	Grocery Store	American Restaurant	Pizza Place	Coffee Shop	Bakery	Salad Place	Gym / Fitness Center	Supermarket	Cosmetics Shop	New American Restaurant

200 rows × 14 columns

5.3.3. Cluster 3 (Blue)

	Neighborhood Latitude	Venue Longitude	Venue Category	Cluster Labels	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
0	39.143441	-77.205992	Latin American Restaurant	2	Mexican Restaurant	Pizza Place	American Restaurant	Supermarket	Sandwich Place	Café	Coffee Shop	Bar	Gym / Fitness Center	Donut Shop
1	39.143441	-77.198903	Latin American Restaurant	2	Mexican Restaurant	Pizza Place	American Restaurant	Supermarket	Sandwich Place	Café	Coffee Shop	Bar	Gym / Fitness Center	Donut Shop
2	39.143441	-77.193474	Café	2	Mexican Restaurant	Pizza Place	American Restaurant	Supermarket	Sandwich Place	Café	Coffee Shop	Bar	Gym / Fitness Center	Donut Shop
3	39.143441	-77.193522	Park	2	Mexican Restaurant	Pizza Place	American Restaurant	Supermarket	Sandwich Place	Café	Coffee Shop	Bar	Gym / Fitness Center	Donut Shop
4	39.143441	-77.197087	Taco Place	2	Mexican Restaurant	Pizza Place	American Restaurant	Supermarket	Sandwich Place	Café	Coffee Shop	Bar	Gym / Fitness Center	Donut Shop
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
647	39.139830	-77.204135	Chocolate Shop	2	Mexican Restaurant	American Restaurant	Pizza Place	Gym / Fitness Center	Ice Cream Shop	Coffee Shop	Donut Shop	Bar	Supermarket	Park
648	39.139830	-77.206934	Grocery Store	2	Mexican Restaurant	American Restaurant	Pizza Place	Gym / Fitness Center	Ice Cream Shop	Coffee Shop	Donut Shop	Bar	Supermarket	Park
649	39.139830	-77.196246	Mexican Restaurant	2	Mexican Restaurant	American Restaurant	Pizza Place	Gym / Fitness Center	Ice Cream Shop	Coffee Shop	Donut Shop	Bar	Supermarket	Park
650	39.139830	-77.219293	Hotel	2	Mexican Restaurant	American Restaurant	Pizza Place	Gym / Fitness Center	Ice Cream Shop	Coffee Shop	Donut Shop	Bar	Supermarket	Park
651	39.139830	-77.205204	Pizza Place	2	Mexican Restaurant	American Restaurant	Pizza Place	Gym / Fitness Center	Ice Cream Shop	Coffee Shop	Donut Shop	Bar	Supermarket	Park

400 rows × 14 columns

5.3.4. Cluster 4 (Light Blue)

	Neighborhood Latitude	Venue Longitude	Venue Category	Cluster Labels	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
852	39.017936	-77.211060	Restaurant	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
853	39.017936	-77.209167	Miscellaneous Shop	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
854	39.017936	-77.210538	Burger Joint	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
855	39.017936	-77.209141	Pub	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
856	39.017936	-77.209994	Yoga Studio	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
946	39.017936	-77.168530	Trail	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
947	39.017936	-77.249251	Outdoor Sculpture	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
948	39.017936	-77.178590	Intersection	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
949	39.017936	-77.180790	Trail	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place
950	39.017936	-77.155880	Boutique	3	Trail	Golf Course	Bank	Tennis Court	American Restaurant	Coffee Shop	Burger Joint	Pool	Park	Pizza Place

99 rows × 14 columns

5.3.5. Cluster 5 (Light Green)

	Neighborhood Latitude	Venue Longitude	Venue Category	Cluster Labels	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
200	39.20879	-77.139956	Pizza Place	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
201	39.20879	-77.117286	Golf Course	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
202	39.20879	-77.189259	Liquor Store	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
203	39.20879	-77.171436	Thai Restaurant	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
204	39.20879	-77.188837	Greek Restaurant	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
205	39.20879	-77.171950	Mexican Restaurant	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
206	39.20879	-77.167714	Flower Shop	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
207	39.20879	-77.080256	Brewery	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
208	39.20879	-77.085457	Trail	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
209	39.20879	-77.160089	Motorcycle Shop	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
210	39.20879	-77.152715	Gymnastics Gym	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
211	39.20879	-77.188769	Sushi Restaurant	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
212	39.20879	-77.154183	Mexican Restaurant	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery
213	39.20879	-77.131172	Farm	4	Mexican Restaurant	Golf Course	Park	Farm	Wine Shop	Grocery Store	Greek Restaurant	Video Store	Trail	Brewery

5.3.6. Cluster 6 (Orange)

	Neighborhood Latitude	Venue Longitude	Venue Category	Cluster Labels	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
652	39.025672	-77.074486	Pizza Place	5	Pizza Place	Park	Vietnamese Restaurant	Trail	Bagel Shop	Farmers Market	Grocery Store	Mexican Restaurant	Gym / Fitness Center	Movie Theater
653	39.025672	-77.076768	Coffee Shop	5	Pizza Place	Park	Vietnamese Restaurant	Trail	Bagel Shop	Farmers Market	Grocery Store	Mexican Restaurant	Gym / Fitness Center	Movie Theater
654	39.025672	-77.074350	Chinese Restaurant	5	Pizza Place	Park	Vietnamese Restaurant	Trail	Bagel Shop	Farmers Market	Grocery Store	Mexican Restaurant	Gym / Fitness Center	Movie Theater
655	39.025672	-77.074349	French Restaurant	5	Pizza Place	Park	Vietnamese Restaurant	Trail	Bagel Shop	Farmers Market	Grocery Store	Mexican Restaurant	Gym / Fitness Center	Movie Theater
656	39.025672	-77.072088	Farmers Market	5	Pizza Place	Park	Vietnamese Restaurant	Trail	Bagel Shop	Farmers Market	Grocery Store	Mexican Restaurant	Gym / Fitness Center	Movie Theater
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1046	38.995946	-77.094378	Department Store	5	Grocery Store	Latin American Restaurant	American Restaurant	Bakery	Brewery	Coffee Shop	Trail	Mexican Restaurant	Vietnamese Restaurant	Indian Restaurant
1047	38.995946	-77.012047	Food Truck	5	Grocery Store	Latin American Restaurant	American Restaurant	Bakery	Brewery	Coffee Shop	Trail	Mexican Restaurant	Vietnamese Restaurant	Indian Restaurant
1048	38.995946	-77.092240	Farmers Market	5	Grocery Store	Latin American Restaurant	American Restaurant	Bakery	Brewery	Coffee Shop	Trail	Mexican Restaurant	Vietnamese Restaurant	Indian Restaurant
1049	38.995946	-77.076398	Grocery Store	5	Grocery Store	Latin American Restaurant	American Restaurant	Bakery	Brewery	Coffee Shop	Trail	Mexican Restaurant	Vietnamese Restaurant	Indian Restaurant
1050	38.995946	-77.096350	Hot Dog Joint	5	Grocery Store	Latin American Restaurant	American Restaurant	Bakery	Brewery	Coffee Shop	Trail	Mexican Restaurant	Vietnamese Restaurant	Indian Restaurant

200 rows × 14 columns

6. Conclusions

In this study, single-home housing listing data was collected from Redfin webpage and analyzed in order to answer the following questions:

- How does the distribution of single-home housing in Rockville and its nearby neighborhoods look like?
- What are the common features that drive housing cost?
- What are the most popular venues around the areas of interest?

We built a correlation matrix and plot several distribution graphs to determine the influencing features of housing cost. We also compared listing prices and other characteristics of housing in different neighborhoods. With the coordinate data available in the dataset, we were able to visualize these characteristics on a map and further determine that distance of the property to existing main roads is also an important feature. Furthermore, we collected the surrounding venue data and used machine learning to cluster similar venues among neighborhoods.

7. Future Directions

Even though features discussed in this study are essential to determine the listing price, other features such as crime data and public school ratings also influence the price of a property. For future analysis, these dataset should be collected and analyzed to produce further accurate conclusions.

8. References

https://www.rockvillemd.gov/978/History
https://www.rockvillemd.gov/DocumentCenter/View/18286/Housing-Market-Analysis-and-Needs-Assessment?bidId=
https://en.wikipedia.org/wiki/Interquartile_range

The Jupyter notebooks of the analysis can be found on Github:
https://github.com/hta6/IBM-Capstone-Project-Notebook