## **Crime and Housing Prices in Montgomery County, Maryland**

Arash Asadabadi, Nibret Daba, Natasha Hurwitz, Ashley Turnbull

## Introduction

Our group chose to work on crime and housing market data for Montgomery County, Maryland. Our goal was to see if crime can bring a meaningful influence upon neighborhood housing values, how crime rates vary geographically, and if there is any correlation among various crimes in the county. In order to achieve the targeted goals, the group worked on data acquisition, cleanup, feature engineering, descriptive statistics, model building, and testing.

## Data Acquisition and Cleanup

Our team began by acquiring publicly available datasets which are described in more detail in the Datasets section. One of our biggest challenges was finding a dataset of housing sales records. Metropolitan Regional Information Systems (MRIS) is the company that has intellectual property over this information. Individual real estate agents and companies provide the information, but then, it belongs to the MRIS in exchange for having it listed on many websites used by potential buyers. MRIS would not give access to the information. They suggested contacting a real estate agent to see if one would be willing to help.

After several calls to many agents and several companies, we reached Michael Bystry, Senior Market Research Analyst at Long and Foster. He was happy to help, but was concerned that he was not permitted to share the data. He needed to check with his supervisors, first. It was November 7th, 2016. We were running out of time to obtain this crucial dataset for our project. Without a dataset for housing, our project would not work.

As we awaited a response from Long and Foster, we explored other ways of obtaining the data we needed on housing sales in MoCo. We explored scraping the data from Redfin.com, Zillow.com and Trulia.com, with minimal success. Scraping was going to be time consuming and the freely available Google Chrome add-in, Webscraper, was not able to scrape all of the information that would be needed. As we were scraping the Redfin website, we discovered a button that would allow the user to download all of their homes of interest into a .csv. This was exactly what we needed, although there were problems.

Luckily, we received a call from Michael Bystry at Long and Foster on November 14, 2016. He had a dataset for all housing sales in 2014 in MoCo that he was willing to share. In addition to sending the dataset, he also sent documentation on some of the less intuitive column names and flagged some columns he thought might have unreliable data. Now, we not only had a fairly clean dataset for housing sales, but also a jump start on what was in there. It was a major breakthrough.

Meanwhile, we had a number of datasets that we had combined on zipcode. We had already merged them with the Redfin dataset, so it was easy to replace the Redfin dataset with the Long and Foster one. The descriptions and preparation of those datasets is in Appendix A, Datasets. Most datasets were re-retrieved on 11/2/2016, unless otherwise noted.

## Feature Engineering

In order to get the best possible outcome of our models, it comes necessary to get the most from our data and obviously best algorithms. Therefore, our group created additional features for the crime data. Sixtyseven additional features were identified telling about the type of crime a record is categorized. The new features include ROB FIREARM - STREET, AGG ASSLT FIREARM CITIZEN, BURG FORCE-RES/NIGHT, LARCENY PICK POCKET, PROSTITUTION/VICE-SOLICIT/PANDER, AUTO THEFT, ASSAULT & BATTERY, INTIMIDATION, ARSON, FORGERY, BAD CHECKS, EMBEZZLEMENT, STOLEN PROP-POSSES, VANDALISM, WEAPON POSSESSION, SEX OFFENSE, CDS-MANU-MARIJUANA, CDS-SELL-OPIUM, CDS-SELL-COCAINE, CDS-POSS DRUG, CDS RX FORGERY, FAMILY OFFENSE, JUVENILE RUNAWAY, LIQUOR-UNLAWFUL POSS UNDER 21, DISORDERLY CONDUCT, SUICIDE, HOME IMPROVEMENT VIOLATION, FAIL TO RETURN RENTAL PROPERTY, BOMB THREAT, IMPERSONATING POLICE OFFICER, BLACKMAIL, FAIL PAY BOARD/LDG/FOOD/TAXI/SERVIC, FALSE STATEMENT/REPORT OF CRIME, FIRE CODE VIOLATION, EXPLOSIVE DEVICE, KIDNAPPING, FIREWORKS,ESCAPEE, WELFARE FRAUD, LITTERING/TRASH DUMPING, PORNOGRAPHY, RENTAL CAR VIOLATION, UNAUTH. USE OF MOTOR VEHICLE, SOLICITING/TRADE W/O LIC, ROGUE AND VAGABOND, TRESPASSING, THREATENING/ANNOYING PHONE CALL, HARASSMENT/STALKING, EX PARTE/PROTECT. ORDER VIOL., FUGITIVE FROM JUSTICE(OUT OF STATE), ALL OTHER NON-TRAFFIC CRIME OFFENSES, DRIVING UNDER THE INFLUENCE, SUDDEN DEATH NATURAL, FIRE OTHER, ILL PERSON, DRUNK, POL INFORMATION, MENTAL TRANSPORT, MISSING PERSON, LOST PROPERTY, RECOVERED PROPERTY/MONT. CO., FAMILY TROUBLE, SUSPICIOUS SIT/PERSON/VEHICLE, EMERGENCY SHELTER CARE, ANIMAL OFFENSE - ANIMAL NEGLECT, OTHER, and DWI.

In the process of developing the features the group used the ‘Class’ label in the original dataset to identify similar crimes and designate a search word in the description column and assign TRUE or FALSE in the added feature column. For example the 300 block of class code is for different type of Robbery whereas the 20000 block codes are for burglary and so on. Therefore, for Robbery an excel formula looks for the keyword “ROB” in the description of the record and assign TRUE if condition is satisfied. (=ISNUMBER(SEARCH("ROB", D23))). Based on relevance, the group ended up using only twenty seven of this categories for analysis.

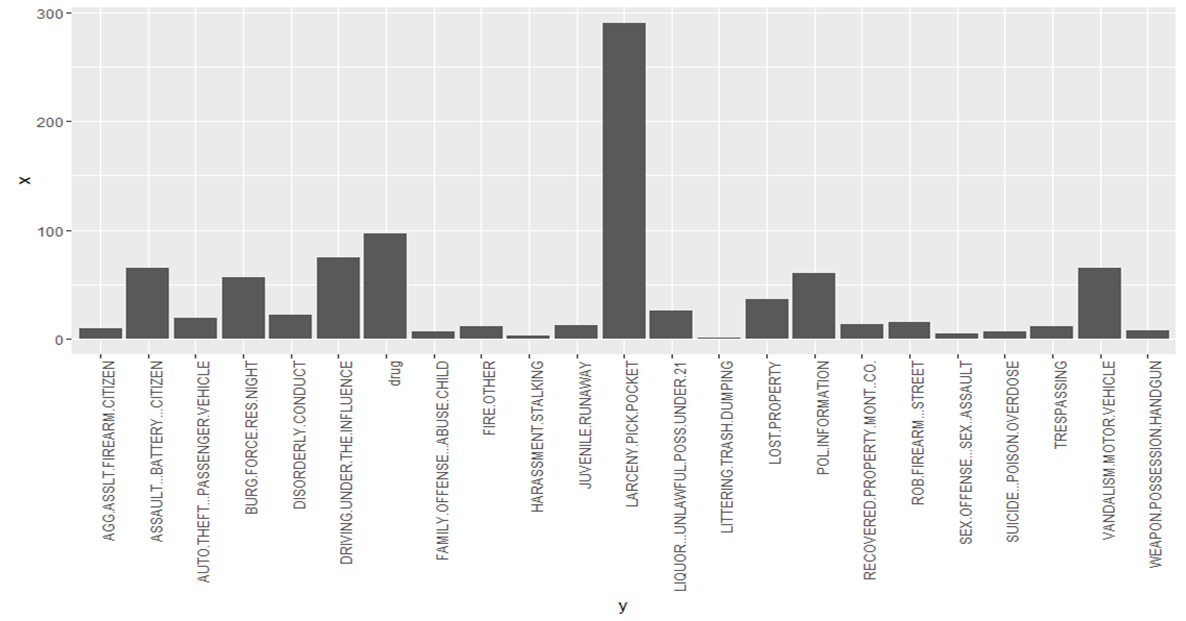
## Descriptive Statistics

### Crime

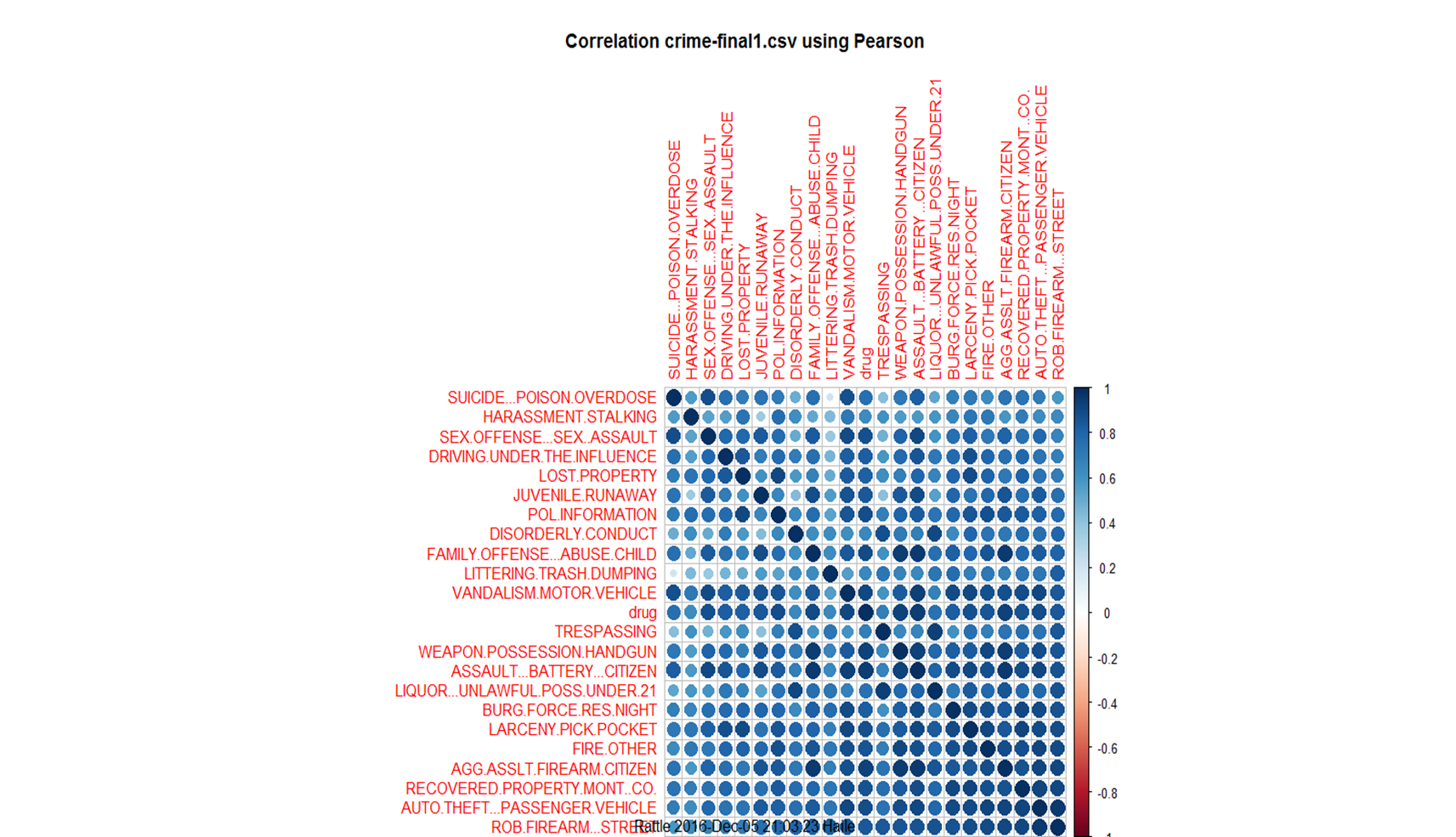
The group worked to see the statistical distribution of the crime dataset. We did a summary analysis of the crime data and here below is the output from r script.

**ROB.FIREARM...STREET AGG.ASSLT.FIREARM.CITIZEN BURG.FORCE.RES.NIGHT Min. : 0.00 Min. : 0.00 Min. : 0.00 1st Qu.: 0.00 1st Qu.: 0.50 1st Qu.: 12.00 Median : 7.00 Median : 4.00 Median : 46.00 Mean :15.02 Mean : 9.14 Mean : 55.91 3rd Qu.:22.50 3rd Qu.:11.50 3rd Qu.: 95.50 Max. :82.00 Max. :38.00 Max. :222.00 LARCENY.PICK.POCKET AUTO.THEFT...PASSENGER.VEHICLE Min. : 0.0 Min. : 0.00 1st Qu.: 32.0 1st Qu.: 2.00 Median : 175.0 Median : 7.00 Mean : 290.6 Mean :18.88 3rd Qu.: 482.5 3rd Qu.:35.50 Max. :1132.0 Max. :79.00 ASSAULT...BATTERY...CITIZEN VANDALISM.MOTOR.VEHICLE Min. : 0.00 Min. : 0.0 1st Qu.: 7.50 1st Qu.: 10.0 Median : 39.00 Median : 49.0 Mean : 64.63 Mean : 65.3 3rd Qu.: 95.00 3rd Qu.:117.0 Max. :222.00 Max. :219.0 WEAPON.POSSESSION.HANDGUN SEX.OFFENSE...SEX..ASSAULT drug Min. : 0.000 Min. : 0.000 Min. : 0.00 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 6.50 Median : 5.000 Median : 3.000 Median : 52.00 Mean : 7.791 Mean : 4.791 Mean : 96.91 3rd Qu.:11.500 3rd Qu.: 8.500 3rd Qu.:152.00 Max. :33.000 Max. :21.000 Max. :396.00 FAMILY.OFFENSE...ABUSE.CHILD JUVENILE.RUNAWAY Min. : 0.000 Min. : 0.0 1st Qu.: 1.000 1st Qu.: 0.0 Median : 5.000 Median : 5.0 Mean : 6.233 Mean :12.3 3rd Qu.: 8.500 3rd Qu.:19.5 Max. :24.000 Max. :56.0 LIQUOR...UNLAWFUL.POSS.UNDER.21 DISORDERLY.CONDUCT Min. : 0.00 Min. : 0.00 1st Qu.: 1.00 1st Qu.: 0.00 Median : 8.00 Median : 6.00 Mean : 26.12 Mean : 22.26 3rd Qu.: 26.50 3rd Qu.: 22.50 Max. :207.00 Max. :189.00 SUICIDE...POISON.OVERDOSE LITTERING.TRASH.DUMPING TRESPASSING Min. : 0.000 Min. : 0.000 Min. : 0.00 1st Qu.: 1.000 1st Qu.: 0.000 1st Qu.: 0.00 Median : 6.000 Median : 0.000 Median : 2.00 Mean : 6.674 Mean : 1.116 Mean : 10.79 3rd Qu.:10.500 3rd Qu.: 1.000 3rd Qu.: 14.00 Page 1**

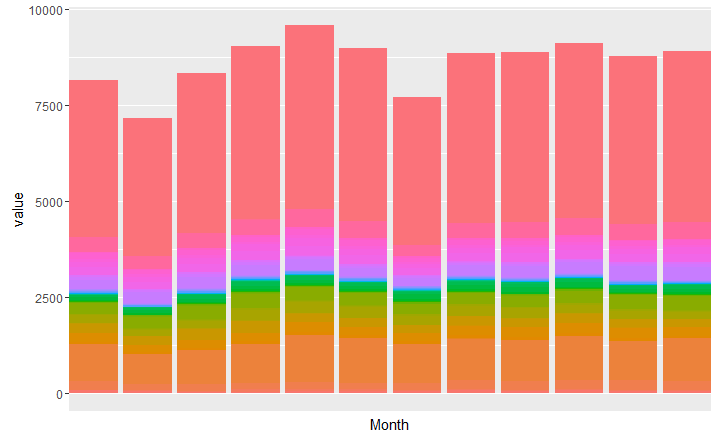
**myfile Max. :27.000 Max. :13.000 Max. :129.00 HARASSMENT.STALKING DRIVING.UNDER.THE.INFLUENCE FIRE.OTHER Min. : 0.000 Min. : 0.00 Min. : 0.00 1st Qu.: 0.000 1st Qu.: 5.00 1st Qu.: 1.00 Median : 2.000 Median : 36.00 Median : 7.00 Mean : 2.465 Mean : 74.91 Mean :11.12 3rd Qu.: 4.000 3rd Qu.:151.00 3rd Qu.:16.50 Max. :10.000 Max. :246.00 Max. :43.00 POL.INFORMATION LOST.PROPERTY RECOVERED.PROPERTY.MONT..CO. Min. : 0.00 Min. : 0.00 Min. : 0.00 1st Qu.: 10.50 1st Qu.: 3.50 1st Qu.: 1.00 Median : 44.00 Median : 17.00 Median : 7.00 Mean : 60.09 Mean : 36.56 Mean :13.28 3rd Qu.: 85.50 3rd Qu.: 57.50 3rd Qu.:21.50 Max. :262.00 Max. :208.00 Max. :91.00**

By way of exporting only the mean out of the output above, we able to see the crime “Larceny Pickpocket” be having a higher mean average compared to the other categories of crime. Same is depicted with the plot below.

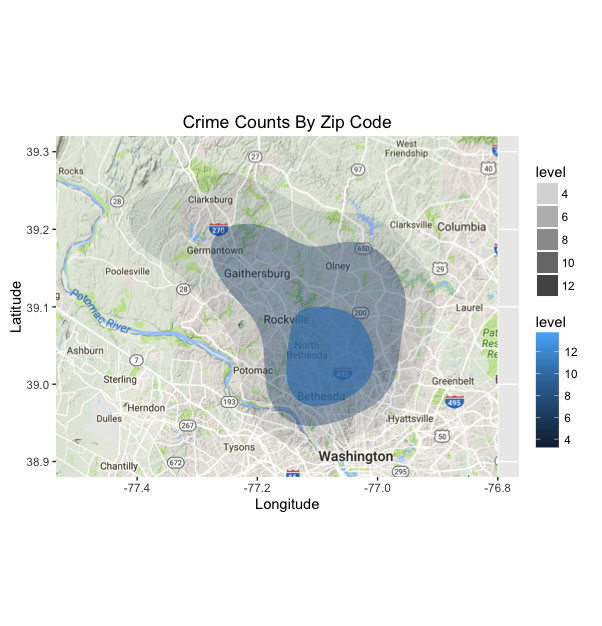
The group also looked into possible correlation between different crime categories and it is possible to find out that most of the crime have a strong positive correlations. Our findings were interesting because the strongest positive correlations happened with Liquor and Disorderly Conduct, Drug and Family Offense/Child Abuse, Juvenile Runaway and Family Offense. Below is the plot showing the same.

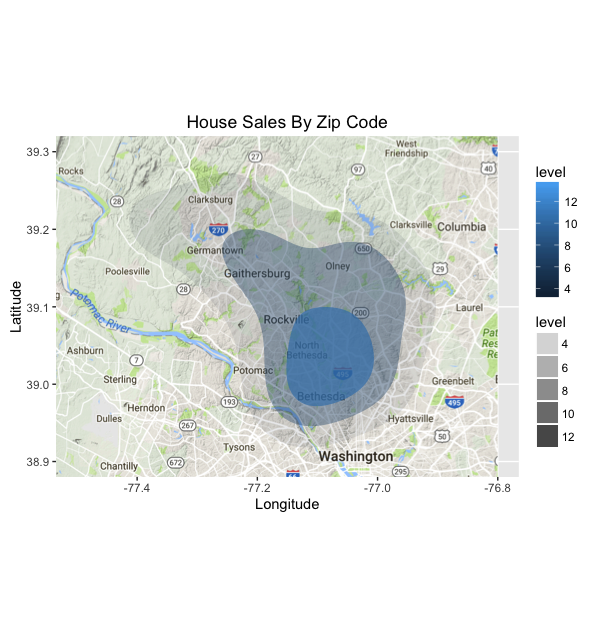


We also looked into trend of crime in a certain time period. Analysis is done to see if there is is any disparity of crime frequency in a certain month of the year. The analysis turned out the crime frequency has almost similar pattern throughout the year.

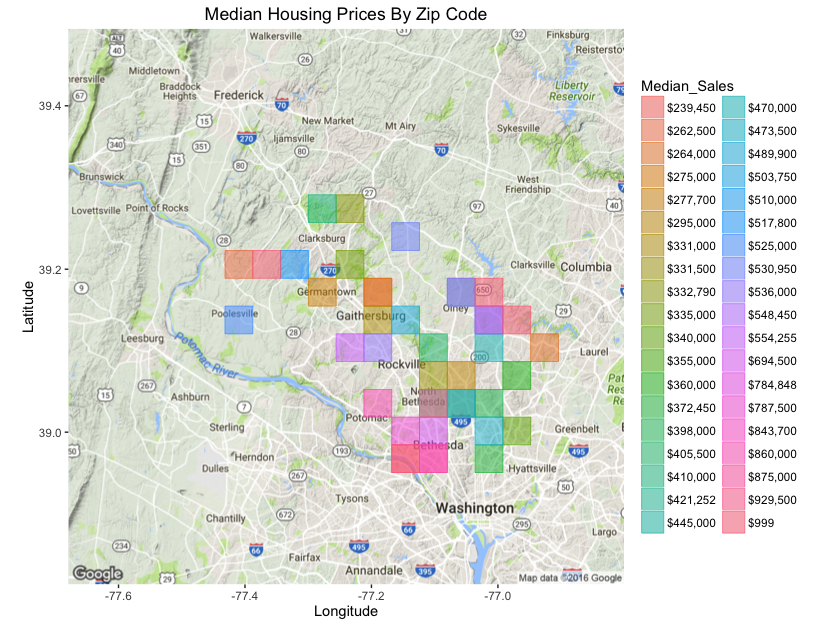


### Maps





We see that the maps for housing sales and crimes counts are very similar. This can be explained by population. The higher the population (by zip code), more crimes occur, more houses exist, which means more housing sales take place, compared to zip codes with lower population.

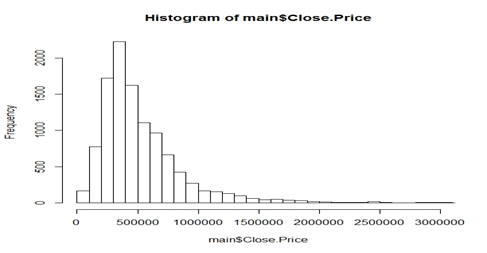


There is a concentrated area of pink on the median housing prices map, which indicates high median housing price. Oranges, greens, and teals indicate lower median housing price (compared to the rest of the scale). We note that the dark blue section, which is common on the crime and housing sale maps, covers an area that is southeast of Rockville and spans though Bethesda and North Bethesda. As we see on the median housing map, this area is almost split by east and west. We see an area of pink around Bethesda, which denotes higher median housing price. The surrounding areas are darker colors, which indicate lower median housing prices. Therefore, even though there are a great amount of housing sales and crime counts in the large area (blue shaded area on crime and housing sale maps), only about half of that has high median housing prices.

## Model Building

### Classification models

The goal of this part is to build different classification models to predict housing class. In order to do this, housing price is categorized into 3 classes: 1-Cheap, 2-Medium and 3-Expensive. The classification is based on the frequency of prices. The distribution of price is shown in Plot (1) . The classes are chosen arbitrarily, but with considering distribution and also based on the general pricing statistics in Maryland.



Plot (1) - Distribution of Housing prices

Cheap:[0,300000), Medium:[300000,650000), Expensive : More than 650000

#### SVM

The first classification model used here is SVM. Feature engineering is used to determine the best predictors which give our model the highest accuracy. The variables are added one by one and the accuracy is calculated in each step which is shown in the following table:

|  |  |
| --- | --- |
| Predictor(s) | Accuracy |
| Base | .509 |
| +number of bedrooms | .632 |
| + number of bathrooms | .663 |
| +type of house | .691 |
| +garage | 0.721 |
| +total crimes | 0.729 |
| +facilities | 0.767 |
| + all the crimes in details(23 predictor) | .842 |

As we can see, all of these variables increase the accuracy in some extend, only total number of crimes does not have any impact on accuracy. Interestingly, if we use the detailed crime rates by categories it would increase the accuracy about 8% which considered a valuable gain.

For Error Analysis we can look at the contingency table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Reference | | |
|  |  | 1 | 2 | 3 |
| Prediction | 1 | 426 | 64 | 0 |
| 2 | 112 | 978 | 101 |
| 3 | 0 | 64 | 423 |

After looking closer to the falsely predicted cases it can be seen that most of those records are very close to boundaries which means they are very close to the adjacent intervals. This is inevitable for classifying a continuous variable but as we saw model is able to predict 85% of the cases truly.

#### Decision Tree

The second model used is Decision Tree. The results is shown in Plot (2). The accuracy of this model is 0.802. The contingency table is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Reference | | |
|  |  | 1 | 2 | 3 |
| Prediction | 1 | 417 | 99 | 0 |
| 2 | 111 | 933 | 121 |
| 3 | 0 | 83 | 403 |

The table is very similar to SVM table as we expected.

#### 

#### 

#### Plot (2)

#### Random Forest model

Random Forest is an improvised decision tree algorithm which can boost the accuracy in some cases. It uses numerous trees with different variables chosen randomly to overcome the overfitting issue in decision trees. In this case the accuracy of the predicted model is 0.856. As we can see there is a good improvement in the model from decision tree model

#### Multinomial Logit model

Since our target variable is not binary we should use multinomial logistic regression rather than the binomial one. Here for each record we have three probabilities which sum up to 1.

u ln() = b02 + b12 \* x1 + b22 \* x2 +…

u ln() = b03 + b13 \* x1 + b23 \* x2 +…

Here, class 1 is the base class and the regression model is based on the ratio of probabilities to base class.

For prediction, the class with maximum probability is assigned to each record. For getting the accuracy cross validation technique is used to get the average accuracy for 30 models. In each iteration the training and testing data is different. The Average model accuracy is 0.82

### Comparison between models

The summary of all results are in the following table . Random Forest model gives us the best result while other models also can predict the test data truly in more than 80% cases. This classification model can be used by agencies to predict the housing value’s class. Our main predictor which is “crime rates” play a significant rule in classifying houses.

|  |  |
| --- | --- |
| Model | Accuracy |
| SVM | 0.842 |
| Multinomial logit | 0.821 |
| Decision tree | 0.802 |
| Random Forest | 0.856 |

### Regression Model to predict price change

Housing price is changed from the day a place is listed in the Market to its closing date. People may be forced to reduce pricing if there is no demand or they may have the option to increase their first price if there are enough interests. The purpose of this part is to predict how much the prices would change based on our available predictors.

Linear regression for this goal. :

Price difference ~ Quarter Date + housing type + log(Original price) + log(median sales) + NO bedrooms + NO bathrooms + Parking type

Price difference = log (original price) – log (closing price)

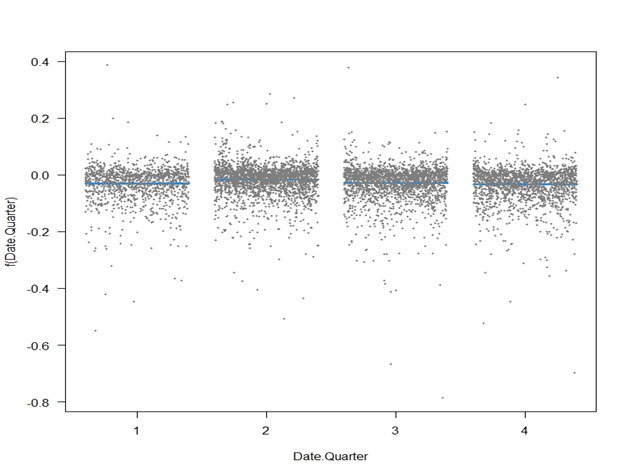
The regression results are in following table. The significant variables are highlighted. Housing price is expected to be increased in spring since there is more moving in that season and demand is higher. Plot (3) shows the price difference for different months.

The coefficient for some types of houses are significantly negative. These are the more special houses which do not have regular demand so usually their prices shrink at the end.

Log(original price) has a negative coefficient which means more expensive houses experience a reduction in their price before being sold.

Overall the RSS is high for the model. These predictors are not enough to predict the price difference. It seems that some other variables may be effective in the decision of reducing or increasing the price by agencies which cannot be captured here.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Std. Error | t value |
| (Intercept) | 0.133097085 | 0.021542071 | 6.178472 |
| community\_facilities\_count | -0.000337599 | 0.000114755 | -2.94192 |
| Number\_of\_Crimes\_2014 | 6.78E-07 | 7.69E-07 | 0.882136 |
| Date.Quarter2 | 0.013789248 | 0.001938061 | 7.11497 |
| Date.Quarter3 | 0.001421499 | 0.002072896 | 0.685755 |
| Date.Quarter4 | -0.004244532 | 0.002134421 | -1.98861 |
| Baths.All | -0.000832803 | 0.0004441 | -1.87526 |
| Bedrooms | -0.001605793 | 0.000627581 | -2.5587 |
| Type.yBack-to-Back | -0.004882025 | 0.013257669 | -0.36824 |
| Type.yDetached | 0.000430829 | 0.005327421 | 0.08087 |
| Type.yDuplex | -0.030844777 | 0.015139725 | -2.03734 |
| Type.yDwelling w/Rental | -0.117618746 | 0.041957228 | -2.8033 |
| Type.yGarden 1-4 Floors | -0.025071309 | 0.005603511 | -4.47421 |
| Type.yHi-Rise 9+ Floors | -0.032067674 | 0.005777762 | -5.55019 |
| Type.yMid-Rise 5-8 Floors | -0.022231607 | 0.00816972 | -2.72122 |
| Type.yMulti-Family | -0.007878051 | 0.021367656 | -0.36869 |
| Type.yOther | -0.044456379 | 0.014092065 | -3.15471 |
| Type.yPatio Home | 0.006835869 | 0.010049284 | 0.680234 |
| Type.yPenthouse | -0.049395065 | 0.041785754 | -1.1821 |
| Type.yQuad | -0.109013997 | 0.058843699 | -1.8526 |
| Type.ySemi-Detached | -0.001298438 | 0.012804427 | -0.10141 |
| Type.yTownhouse | 0.004906017 | 0.00527229 | 0.930529 |
| Has.GarageTRUE | -0.00022548 | 0.001470398 | -0.15335 |
| log(Original.List.Price) | -0.012792814 | 0.00176168 | -7.26171 |
| log(Median\_Sales) | 0.004713641 | 0.000927454 | 5.082342 |



### Plot(3)

### Regression Model to predict the waiting days

Usually sellers should wait some days so their listed housing would be sold. The purpose of this part is to predict how many days a listed property should wait before being sold.

Linear regression for this goal:

Days in Market ~ Quarter Date + housing type + log(Original price) + NO bedrooms + NO bathrooms + Parking type

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Std. Error | t value |
| (Intercept) | -78.7312 | 22.2346 | -3.54093 |
| Date.Quarter2 | -19.1513 | 2.167491 | -8.8357 |
| Date.Quarter3 | -13.8891 | 2.311782 | -6.00797 |
| Date.Quarter4 | 2.516086 | 2.380461 | 1.056974 |
| Baths.All | 8.5505 | 0.924979 | 9.243998 |
| Bedrooms | -5.96735 | 0.975437 | -6.11762 |
| Type.yBack-to-Back | 22.77999 | 14.87564 | 1.531361 |
| Type.yDetached | 22.0243 | 6.120122 | 3.59867 |
| Type.yDuplex | 27.82002 | 16.5533 | 1.680633 |
| Type.yDwelling w/Rental | 33.32 | 47.05715 | 0.708075 |
| Type.yGarden 1-4 Floors | 29.03682 | 6.345849 | 4.575719 |
| Type.yHi-Rise 9+ Floors | 34.9784 | 6.545509 | 5.343878 |
| Type.yHouse of Worship | -36.2948 | 65.8642 | -0.55105 |
| Type.yMid-Rise 5-8 Floors | 35.49886 | 9.211218 | 3.853872 |
| Type.yMulti-Family | -8.1237 | 25.48539 | -0.31876 |
| Type.yOther | 23.98684 | 16.16294 | 1.484064 |
| Type.yPatio Home | 7.937056 | 12.46899 | 0.636544 |
| Type.yPenthouse | 141.9681 | 46.75531 | 3.036405 |
| Type.yQuad | 23.59955 | 46.75041 | 0.504799 |
| Type.ySemi-Detached | 14.54012 | 15.4629 | 0.940323 |
| Type.yTownhouse | 8.469331 | 6.028048 | 1.404987 |
| Has.GarageTRUE | 4.256822 | 1.652013 | 2.576748 |
| log(Original.List.Price) | 8.210767 | 1.73617 | 4.72924 |

The significant variables are highlighted. Houses stay shorter in market during spring and summer because those are hot seasons for housing.

Again, more special houses like penthouses face less demand so they are expected to stay longer in market. For instance a Hi-Rise 9+floor house is expected to stay 35-22=13 more days in market than a detached house.

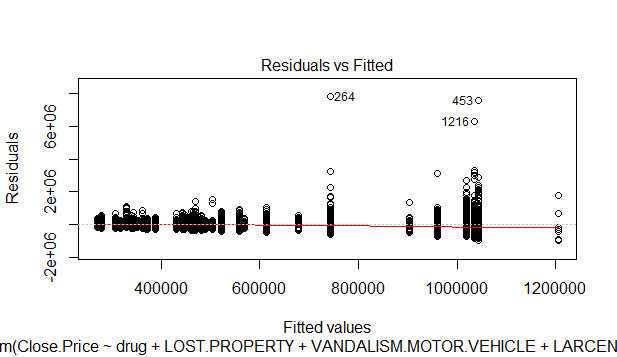
Price is also positively correlated with days in market meaning that if original price is higher the house is expected to stay more days in market.

### Linear Model

The goal of the linear model was to see if we could predict housing prices based on the information in our final dataset. While our model was not able to give an accurate prediction for the housing price, we were able to improve the model. The results of the linear model and components are shown, below.

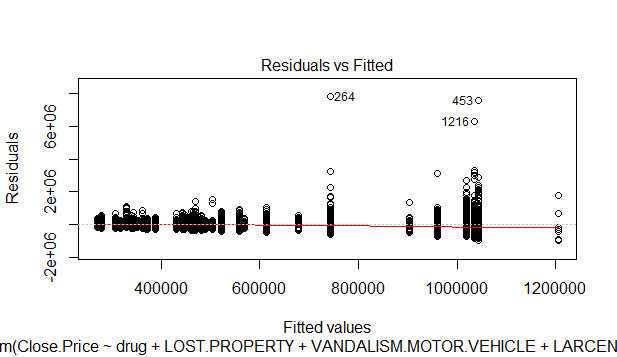
We ran several linear models and used the functions StepAIC and ANOVA to evaluate them. The models are described below and plots are provided. Additional plots will run in the R script.

#### Model 1



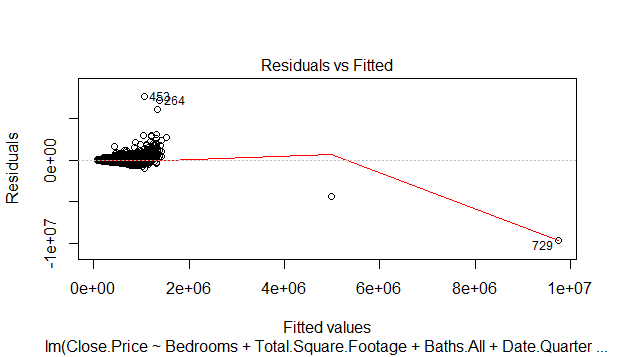
* Model 1 has only the crime data, since crime data was by zipcode and had no other inputs, the model predicted the same price for every house in every zipcode.
  + drug + LOST.PROPERTY + VANDALISM.MOTOR.VEHICLE + LARCENY.PICK.POCKET + Zip.Code + AGG.ASSLT.FIREARM.CITIZEN + BURG.FORCE.RES.NIGHT, data = trainset
  + To improve the model, add factors related to the homes
  + Since this model only contains information by zip code, only zip code is significant. This is a baseline.

#### Model 2



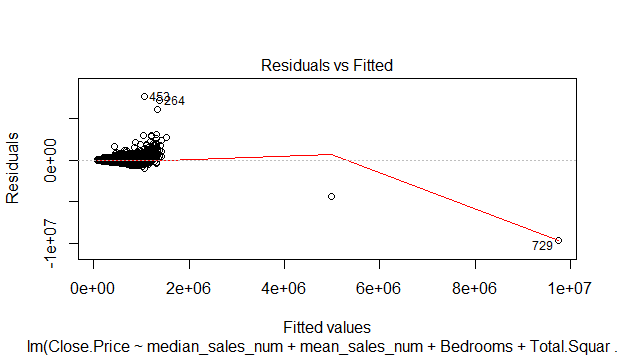
* Model 2 considered the following
  + Close.Price ~ Bedrooms + Total.Square.Footage + Baths.All + Date.Quarter + community\_facilities\_count + Number\_of\_Crimes\_2014 + drug + LOST.PROPERTY + VANDALISM.MOTOR.VEHICLE + LARCENY.PICK.POCKET + Zip.Code + AGG.ASSLT.FIREARM.CITIZEN + BURG.FORCE.RES.NIGHT, data = trainset
  + I ran the AIC and ANOVA. The model found Number\_of\_Crimes\_2014 not significant.

#### Model 3



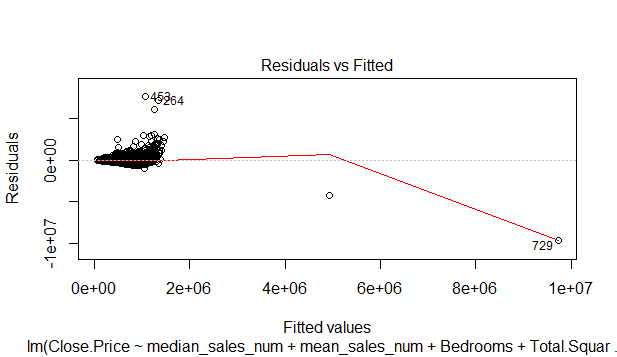
* Model 3 removes Number\_of\_Crimes\_2014
  + Close.Price ~ Bedrooms + Total.Square.Footage + Baths.All + Date.Quarter + community\_facilities\_count + drug + LOST.PROPERTY + VANDALISM.MOTOR.VEHICLE + LARCENY.PICK.POCKET + Zip.Code + AGG.ASSLT.FIREARM.CITIZEN + BURG.FORCE.RES.NIGHT, data = trainset

#### Model 4



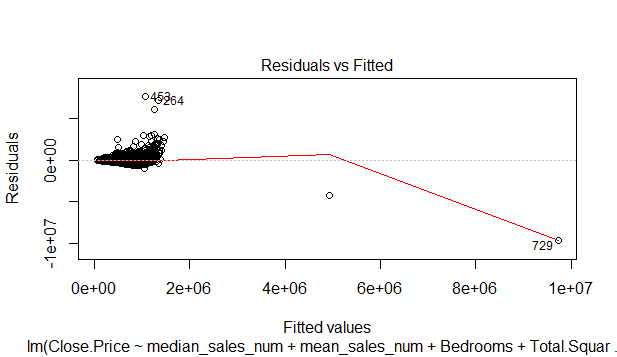
* Model 4 adds median and mean prices for 2014 and uses the crime descriptive statistics to determine the most frequent crimes
  + mylogit <- lm(Close.Price ~ median\_sales\_num + mean\_sales\_num + Bedrooms + Total.Square.Footage + Baths.All + Date.Quarter + community\_facilities\_count + drug + LOST.PROPERTY + VANDALISM.MOTOR.VEHICLE + LARCENY.PICK.POCKET + Zip.Code + ASSAULT...BATTERY...CITIZEN + POL.INFORMATION + DRIVING.UNDER.THE.INFLUENCE + BURG.FORCE.RES.NIGHT, data = trainset)
  + Zipcode and lost property were not significant

#### Model 5



* Final Linear Model - Model 4 with Zipcode and lost property removed
  + mylogit <- lm(Close.Price ~ median\_sales\_num + mean\_sales\_num + Bedrooms + Total.Square.Footage + Baths.All + Date.Quarter + community\_facilities\_count + drug + VANDALISM.MOTOR.VEHICLE + LARCENY.PICK.POCKET + ASSAULT...BATTERY...CITIZEN + POL.INFORMATION + DRIVING.UNDER.THE.INFLUENCE + BURG.FORCE.RES.NIGHT, data = trainset)

#### Model 6

-

* Housing characteristics and crimes with highest counts from descriptive statistics:
  + mylogit <- lm(Close.Price ~ median\_sales\_num + mean\_sales\_num + Bedrooms + Total.Square.Footage + Baths.All + Date.Quarter + community\_facilities\_count + drug + VANDALISM.MOTOR.VEHICLE + LARCENY.PICK.POCKET + ASSAULT...BATTERY...CITIZEN + POL.INFORMATION + DRIVING.UNDER.THE.INFLUENCE + BURG.FORCE.RES.NIGHT, data = trainset)
  + ANOVA final model - Final Model:

Close.Price ~ mean\_sales\_num + Bedrooms + Total.Square.Footage +

Baths.All + community\_facilities\_count + drug + VANDALISM.MOTOR.VEHICLE +

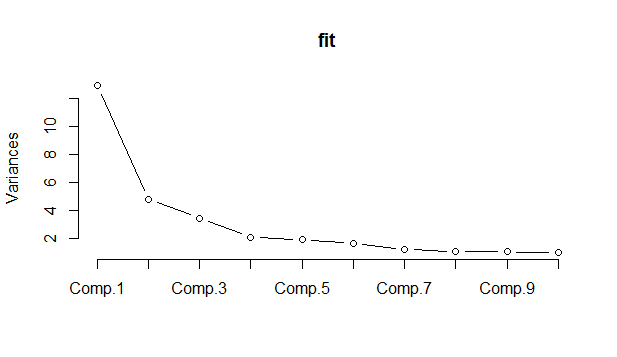
LARCENY.PICK.POCKET + ASSAULT...BATTERY...CITIZEN + DRIVING.UNDER.THE.INFLUENCE

### Principle Component Analysis

Principle Component Analysis (PCA) was used to see if we could narrow down the most influential factors. The linear model with StepAIC and ANOVA was not giving a better model. Using PCA, we were able to narrow in on different factors and create a linear model with only those factors that we then evaluated with AIC and ANOVA.

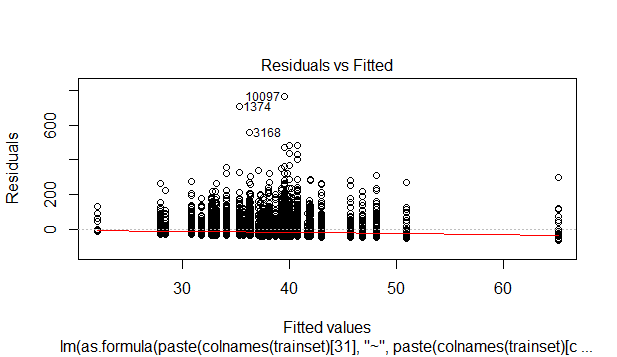
The PCA shows that there are 4 components of more importance than the rest. It also prints a list of the components in order of interest, by index. The plot shows that 10 components are of interest and the remaining are likely less important.

PCA Plot:



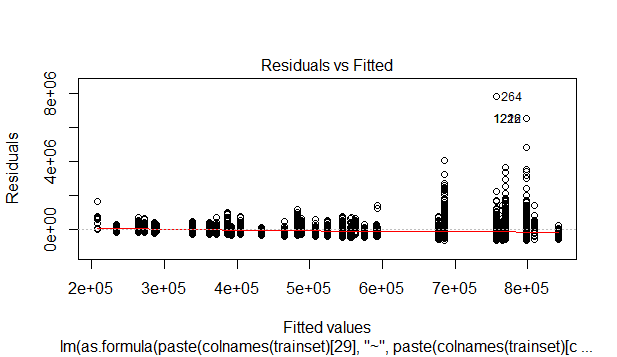
We then used the top factors for the crime data from the PCA to come up with a linear model.

#### Model 7

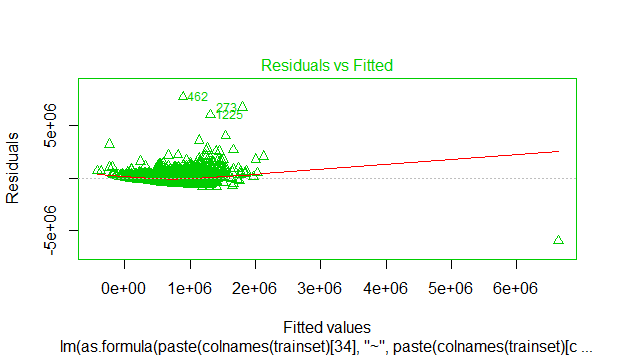
- 

Next we looked at the model with just the top PCA factors.

#### Model 8



#### Model 9



* Model 9 was the best model
  + Initial Model:

Close.Price ~ ROB.FIREARM...STREET + LITTERING.TRASH.DUMPING + BURG.FORCE.RES.NIGHT + AGG.ASSLT.FIREARM.CITIZEN + Baths.All +

Bedrooms + Total.Square.Footage

* + Final Model:

Close.Price ~ ROB.FIREARM...STREET + LITTERING.TRASH.DUMPING +

BURG.FORCE.RES.NIGHT + AGG.ASSLT.FIREARM.CITIZEN + Baths.All +

Bedrooms + Total.Square.Footage

In summary, none of the linear models were able to predict the housing price, accurately. Model 9, which used a combination of PCA and factor data had the best AIC at 227765.1. It is a slight improvement over our baseline AIC of 228178.9, but not by too much. Adding in mean and median sale price for the zip code improved this model to 227477.4.   
 The original project in class involving the Zillow Dataset had historical records on housing sales. Having a baseline price for the home, whether that is by looking at other nearby homes that have sold recently or looking at the last sale price is an important factor in predicting the housing sale price. Without access to that information, our model only had the median and mean housing sales prices in the zip code as a guide. Originally, we thought we would have access to the figures on past sales prices for the homes, but the dataset we had to work with did not have this information. Adding the mean and median sale prices by zip code improved the model, but only slightly.

## APPENDIX A

**Datasets**

Below, the origin of each dataset is listed and notes have been made about how the dataset was cleaned and what the dataset was used for. Datasets were obtained on November 2, 2016 unless otherwise noted.

**MC\_Zip\_Codes\_IRS (**[**http://www.zipcodestogo.com/Maryland/**](http://www.zipcodestogo.com/Maryland/)**, gold access, 8K CSV file)**

* This dataset was filtered by state, then by county. All non-MC records were removed.
* Used to check zip codes for other datasets to verify that the zipcodes were in MC.
* Fields are zip, type (type of zipcode), primary\_city, latitude, longitude, irs\_estimated\_population\_2014. All other fields were removed.

**MC\_IRS (from the MC\_Zip\_Codes\_IRS dataset)**

* Pulls out just the zipcode, type of zipcode, and IRS population from the MC\_Zip\_Codes\_IRS dataset
* Only this data was combined in Most\_Data.csv, the master dataset for zipcode related information.

**Crime\_by\_Zipcode (**[**https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3/data**](https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3/data)**, MC Crime Dataset)**

* We used a dplyr function to count and create a table for the number of crime incidents that occurred between 7/1/2013 and 10/31/2016.
* This dataset contains zip codes and number of incidents.

**Facilities\_by\_Zipcode (**<https://catalog.data.gov/dataset/all-public-facilities-available-for-community-use>**, data.gov Montgomery, 2014 facilities list)**

* This is a dataset of public, community facilities in Montgomery County.
* We used the same dplyr function as in Crime\_by\_Zipcode.
* This dataset contains the zip code and the number of facilities in that zipcode.

**MC\_House\_Sales\_by\_Zipcode\_2014**, <http://planning.maryland.gov/msdc/sale_data/saledata.shtml>, accessed on November 1, 2016)

* The original full title of this dataset is 2014 RESIDENTIAL TOTAL SALES AND SALES BY HOUSE TYPES FOR MARYLAND'S ZIPCODE BY JURISDICTIONS. This title was found in the header of the dataset and was removed.
* This dataset shows the median house sale price by MC zipcode for 2014, according to the Maryland Department of Planning.
* If there were fewer than 3 sales in that zipcode in 2014, it has $999 in the median field, meaning there was not enough data to calculate a mean. This information is in the original dataset.
* Many fields were removed from this dataset in order for it to be easily combined with the other data.
* Zipcode 20896 has one sale in this dataset (with $999 for mean and median because there is only one sale)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

**Most\_Data** (Combined Dataset)

* Created so that we can look back at this file to see what information we have without the housing dataset.
* Combines: Crime, Facilities, IRS population, zipcode type, home sales median and mean from MD Department of Planning.
* School information remains in this dataset, but it is later removed so there is no discussion of the school information, here.
* Created with the R function, merge - merge(x, x1, by = intersect(names(x), names(x1))).
* Rows have been removed.
  + If zipcode type was “PO Box”, then all other fields were N/A (since there is no housing or facilities there). PO Box zip codes and their rows were removed
  + If zipcode type was “unique”, again, most other fields were N/A so these zip codes were removed.
  + Zip code 20839 also had N/A in nearly all fields and was removed.
* 20812, 20838, 20862, 20868 were removed as there are not enough sales in these zipcodes for analysis.
* Number of rows pulled for the Housing\_Sales\_by\_Zip dataset was noted in the column, sales\_pulled. This figure minus one for the header is the number of records per zip code.

**Most\_Data\_2014**

* Created from the Crime\_2014 and Most\_Data datasets.
* Column involving schools have been removed. We decided not to use them in the project because we are focusing on crime.
* Zip codes 2079, 2814, 20012, 20705, 20707, 20724, 20777, 20782, 20783, 20880\*, 20889, 20892, 20899, 20993, 21771, and 46 rows without zipcodes from the Crime data have been removed. These are not residential zip codes. 201 crimes fell into these zip codes. Some are errors (the four digit ones, for example) and have been corrected; some zip codes are not in Montgomery County.
* 20880 was removed. There is only one corresponding sale in the Long\_and\_Foster\_Columns\_Removed dataset which would not have given us enough sales to work with for this zipcode.

**Housing\_Combined\_Dataset**

* The number of observations increased by 2 because there are 2 zipcodes for which there are no properties in Housing\_Sales\_by\_Zip
* 7990 original observations. When duplicate address records were removed, 7975 records remain. There are 40 variables (although some of them are useless.)
* 2359 records would need to be moved if we took out everything with missing values. Worst case scenario is all of those records need to be removed and we still will have 5616 complete records to work with.

**Original\_Long\_and\_Foster**

* This dataset was obtained from Michael Bystry, Senior Market Research Analyst at Long and Foster on 11/14/2016.
* It has more information than our Housing\_Sales\_by\_Zip dataset and more records.
* We should discuss cleaning it and combining it with our other data.

**Long\_and\_Foster\_Columns\_Removed**

* There are 2 records in this dataset that are not in Montgomery County with zipcodes 20707, 20783. They have been removed.
* One sale record for 20880 was removed. There will not be enough data with just one sale to deal with this zipcode.
* To make data analysis easier, columns that would not be used were removed. The following is a list of removed columns:
  + State - all states were Maryland
  + Zip 4 - Although this 4 digit code does link to a more precise area inside of a zipcode, for example a city block, we would not have enough records in each 4 digit zip to add this to our analysis.
  + Advertised Subdivision - not needed as legal subdivision remains
  + Status - all have the “sold” status
  + Baths Half - We kept Baths All
  + Baths Full - We kept Baths All
  + Cooling - type of cooling for the home, irrelevant
  + Dining Kitchen - Description of the dining and kitchen area, irrelevant
  + Farm - Whether or not the property is a farm, all are not farms
  + Fireplaces - number of fireplaces, irrelevant
  + Heating - irrelevant
  + HOA - this HOA fee column is blank if this column is FALSE, redundant
  + Townhouse Type - too detailed, irrelevant
* A column was added, “Has Garage” to indicate if the property has a garage. We are thinking that maybe there are certain crimes where homes that have garages are considered safer by home buyers (would have an increased price over those in the same zipcode without garages, perhaps)
* Type - The type column has 16 levels, which refer to the type of home, the number of each in the dataset is also listed here:
  + 1 Attach/Row Hse 160
  + 2 Back-to-Back 29
  + 3 Detached 6301
  + 4 Duplex 18
  + 5 Dwelling w/Rental 2
  + 6 Garden 1-4 Floors 963
  + 7 Hi-Rise 9+ Floors 694
  + 8 House of Worship 2
  + 9 Mid-Rise 5-8 Floors 114
  + 10 Multi-Family 9
  + 11 Other 23
  + 12 Patio Home 49
  + 13 Penthouse 2
  + 14 Quad 2
  + 15 Semi-Detached 29
  + 16 Townhouse 2497
* “Closedate\_Quarter” was added which breaks down the close date, quarterly:
  + 1 - 1/1/14 - 3/30/14
  + 2 - 4/1/14 - 6/30/14
  + 3 - 7/1/14 - 9/30/14
  + 4 - 10/1/14 - 12/31/14