Alisher Siddikov

Assignment 1

Introduction:

We will explore the housing dataset of Ames, Iowa. The data was obtained from the Ames Assessor's Office which was used in computing assessed values for individual residential properties sold in Ames, IA. We will use the same data and work on predicting the home prices. The type of information contained in the data is similar to what a typical home buyer would want to know before making a purchase. We need to explore features of homes and use the relevant ones to predict the future home sale price. This housing data is an alternative to the Boston Housing dataset.

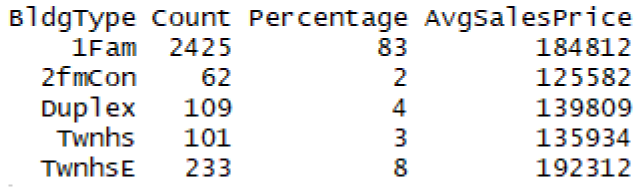
Data Survey:

The housing data has 2930 observations and 82 variables. This data contains property features sold in Ames from 2006 to 2010. There are a target varaible which is home sale price and 81 variables containing property characteristics such as property age, number of rooms, number of additional features, space size, quality, style, and etc. As we explore and analyze the features, we need to down select the relevant ones. We also need to run a data quality check and examine the data for missing data, errors, and outliers. We will also need to examine the strength of relationship of target and predicted varaibles.

This is the first assignment which will be used for further data explorations, analysis, and predictions in next five assignments. So, this assignment is a foundation and we need to explore and understand our data well.

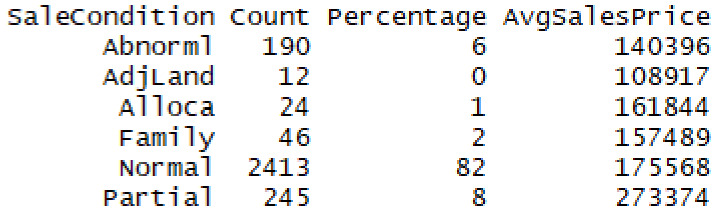
Section 1: Sample Definition

We need to define a normal/typical residential homes and sale conditions.



```
BldgType Count Percentage AvgSalesPrice
    1Fam  2425         83        184812
  2fmCon    62          2        125582
  Duplex   109          4        139809
   Twnhs   101          3        135934
  TwnhsE   233          8        192312
```

If we look at the dwelling (building) type, the single-family homes (1Fam) represent most of the data (83%). We can exclude the rest of the dwelling types such as two-family conversion, duplex, and townhouses.



```
SaleCondition Count Percentage AvgSalesPrice
      Abnorml   190          6        140396
      AdjLand    12          0        108917
       Alloca    24          1        161844
       Family    46          2        157489
       Normal  2413         82        175568
      Partial   245          8        273374
```

If we look at the sale condition, the normal condition represents most of the data (82%). We can drop the rest of the sale conditions such as abnormal sale (trade, foreclosure, short sale), adjoining land purchase, allocation (two linked properties with separate deeds, typically condo with a garage unit), sale between family members, and partial purchase (home was not completed when last assessed and it is associated with new homes).

Here are the derived/calculated columns:

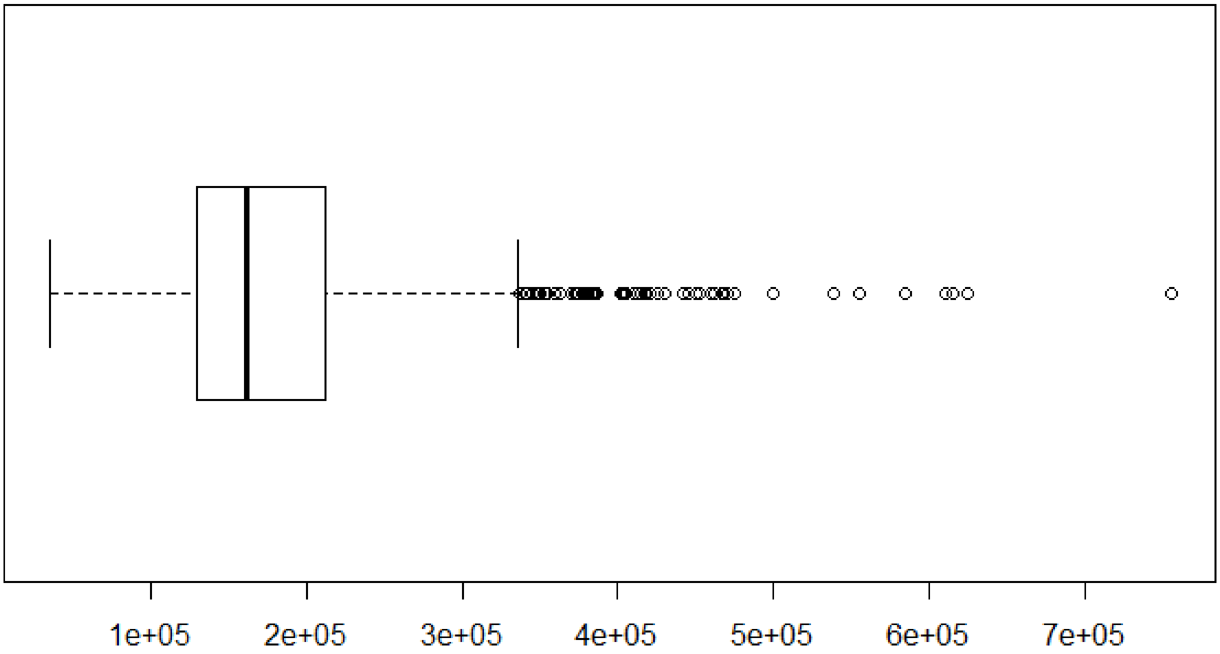
- mydata$TotalFloorSF <- mydata$FirstFlrSF + mydata$SecondFlrSF
- mydata$HouseAge <- mydata$YrSold - mydata$YearBuilt
- mydata$QualityIndex <- mydata$OverallQual * mydata$OverallCond
- mydata$logSalePrice <- log(mydata$SalePrice)
- mydata$price_sqft <- mydata$SalePrice/mydata$TotalFloorSF

The data shape or dimension changed after excluding homes that were not a single-family home and not a normal sale. Here are the updated data dimentions:

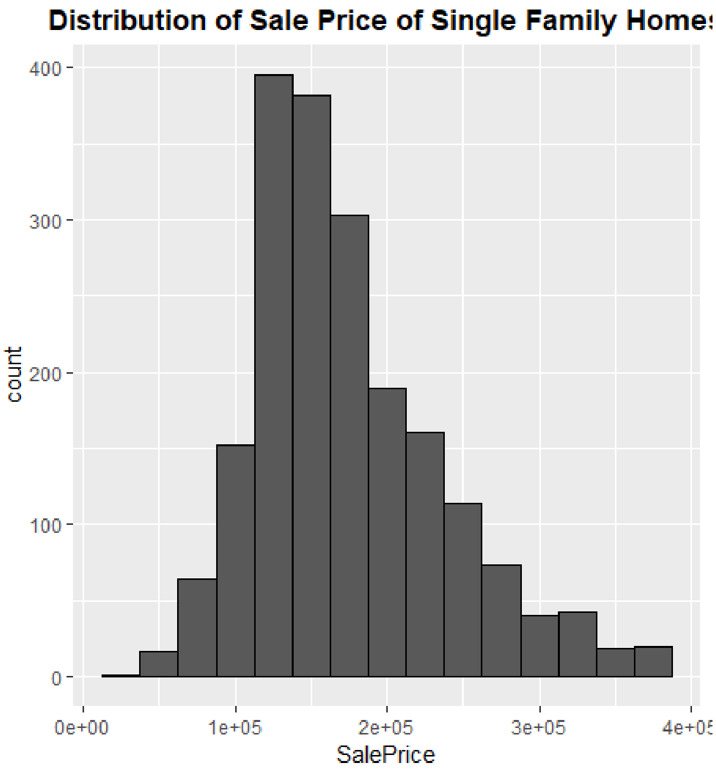
- Updated row numbers: 2,002 rows; about 1/3 of data was eliminated and they were 505 dwelling type homes and 423 sale conditions.
- Updated column numbers: 87 columns; 5 derived new columns were added: total floor square feet, house age, quality index, log sales price, and price per square feet.

Section 2: Data Quality Check

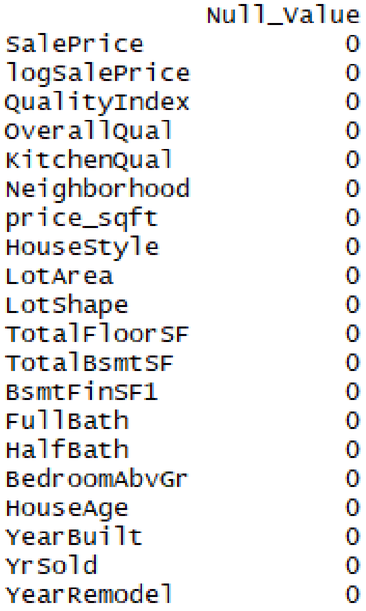
We need to look at our target variable to see the errors or outliers. Any sales after $700, 000 should be an error or an extreme outlier. This is right skewed sales data. The outliers are influencing the average sale price. The average sale price is about $180k and median is about $160k. The lowest sold house is $35k and the highest one is $755k.



Sales price of single-family home has some outliers and it should be eliminated. There are no negative or missing sales price. I used 3 standard deviation formula to remove the 33 outliers from the sales price of a single-family home with normal sales condition.



Here is the 20 relevant variables and they don’t have any missing data.



```
             Null_Value
SalePrice             0
logSalePrice          0
QualityIndex          0
OverallQual           0
KitchenQual           0
Neighborhood          0
price_sqft            0
HouseStyle            0
LotArea               0
LotShape              0
TotalFloorSF          0
TotalBsmtSF           0
BsmtFinSF1            0
FullBath              0
HalfBath              0
BedroomAbvGr          0
HouseAge              0
YearBuilt             0
YrSold                0
YearRemodel           0
```

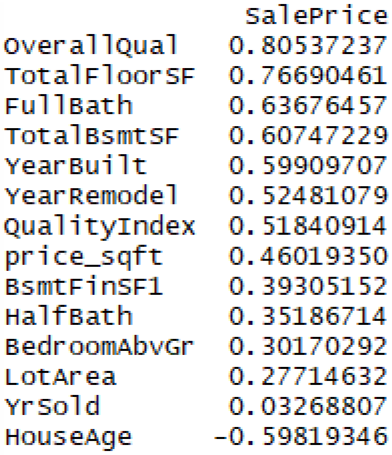
The data shape or dimension changed after excluding sales price outliers and selecting relevant 20 variables. Here are the updated data dimentions:

- Updated row numbers: 1969; 33 outliers were eliminated
- Updated column numbers: 20

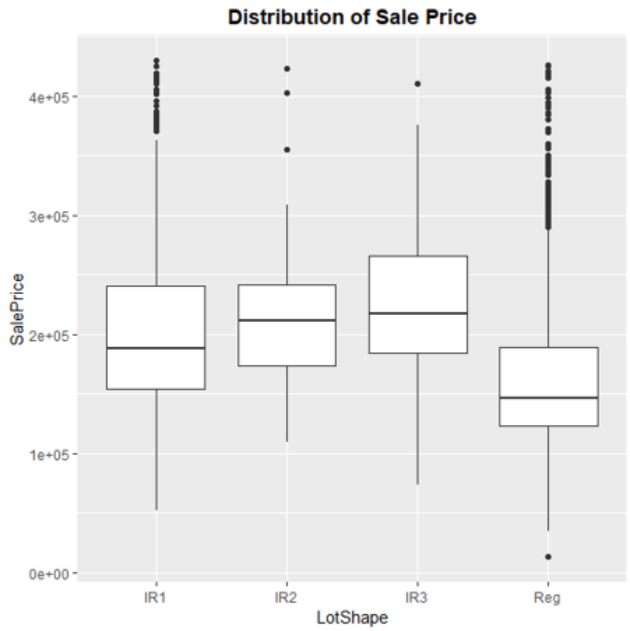
Section 3: Initial Exploratory Data Analysis

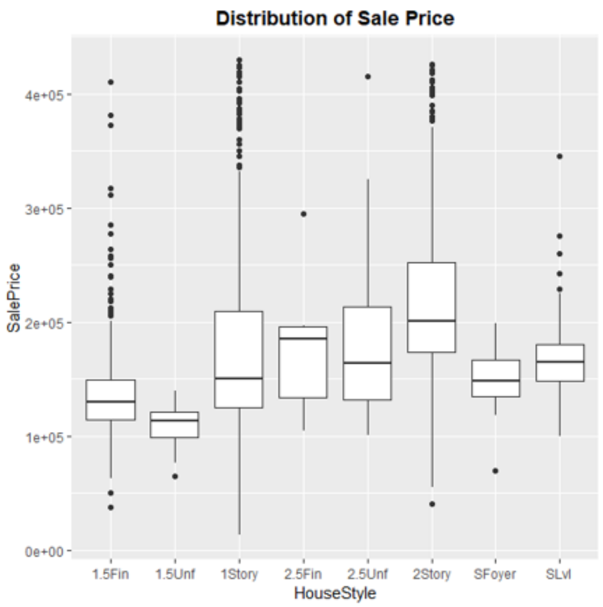
Twenty columns were split into categorial and numerical varaibles. Here is the correlation table between sale price and selected numerical variables. Here are the rest of categorial varaibles: KitchenQual, Neighborhood, HouseStyle, and LotShape. After exploring the data further, we need to select ten varaibles out of twenty.

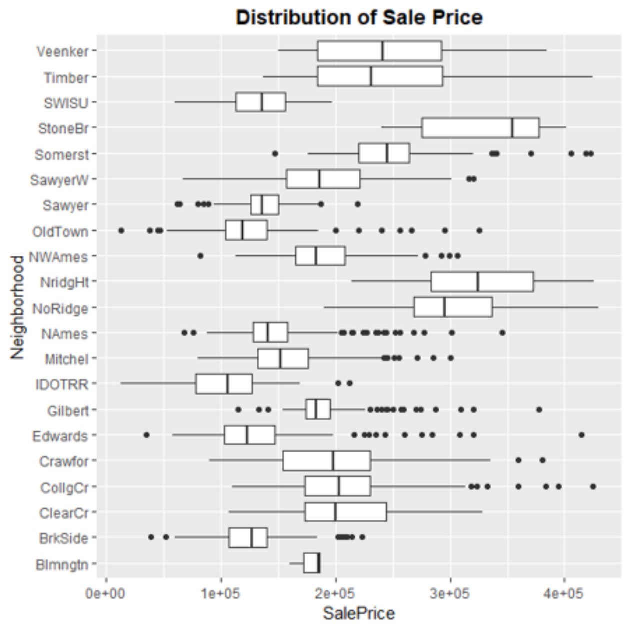
Here are the correlation values between sales price and numerical home characteristics. We can see that overall quality, total floor square footage, number of full baths, total basement square footage, year build and remodeled, and quality index show the highest positive correlation value, as well as house age the highest negative correlation value.

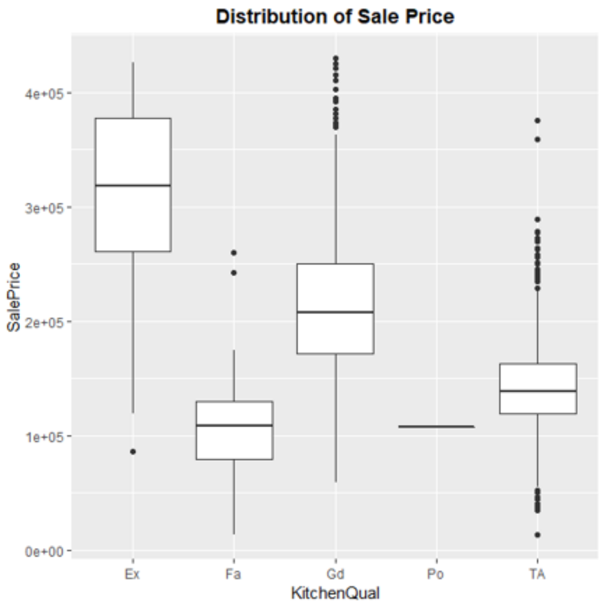


```
               SalePrice
OverallQual   0.80537237
TotalFloorSF  0.76690461
FullBath      0.63676457
TotalBsmtSF   0.60747229
YearBuilt     0.59909707
YearRemodel   0.52481079
QualityIndex  0.51840914
price_sqft    0.46019350
BsmtFinSF1    0.39305152
HalfBath      0.35186714
BedroomAbvGr  0.30170292
LotArea       0.27714632
YrSold        0.03268807
HouseAge     -0.59819346
```



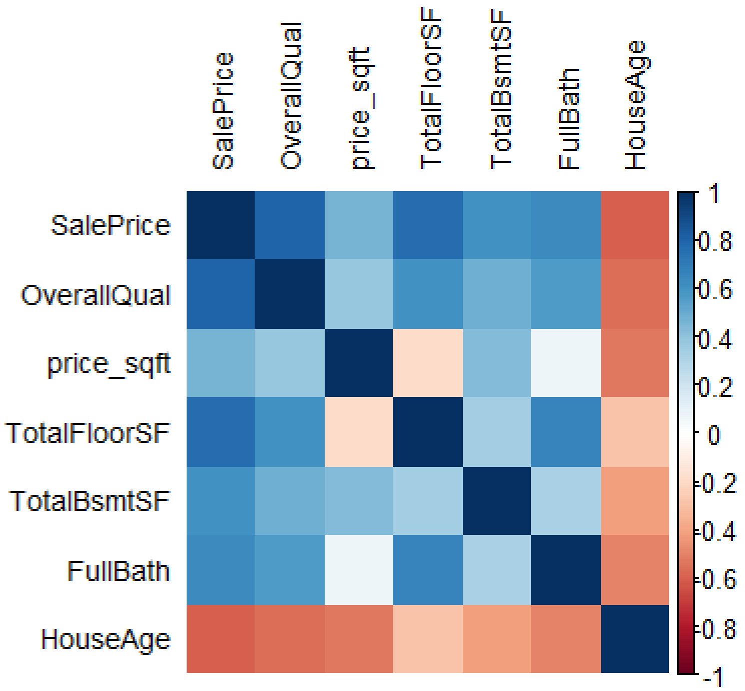


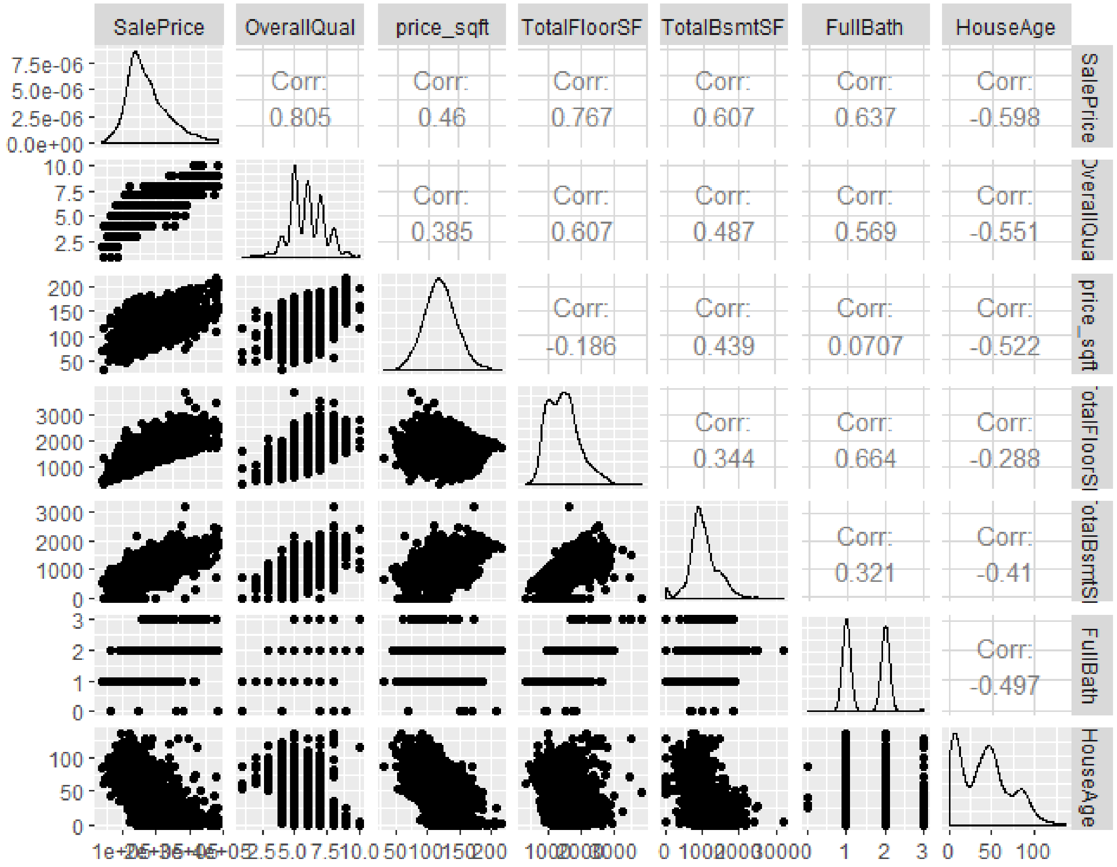




We can see that house style, neighborhood, and kitchen quality gives us sale price indication. Lot shape does not give us much information. We will drop lot shape.

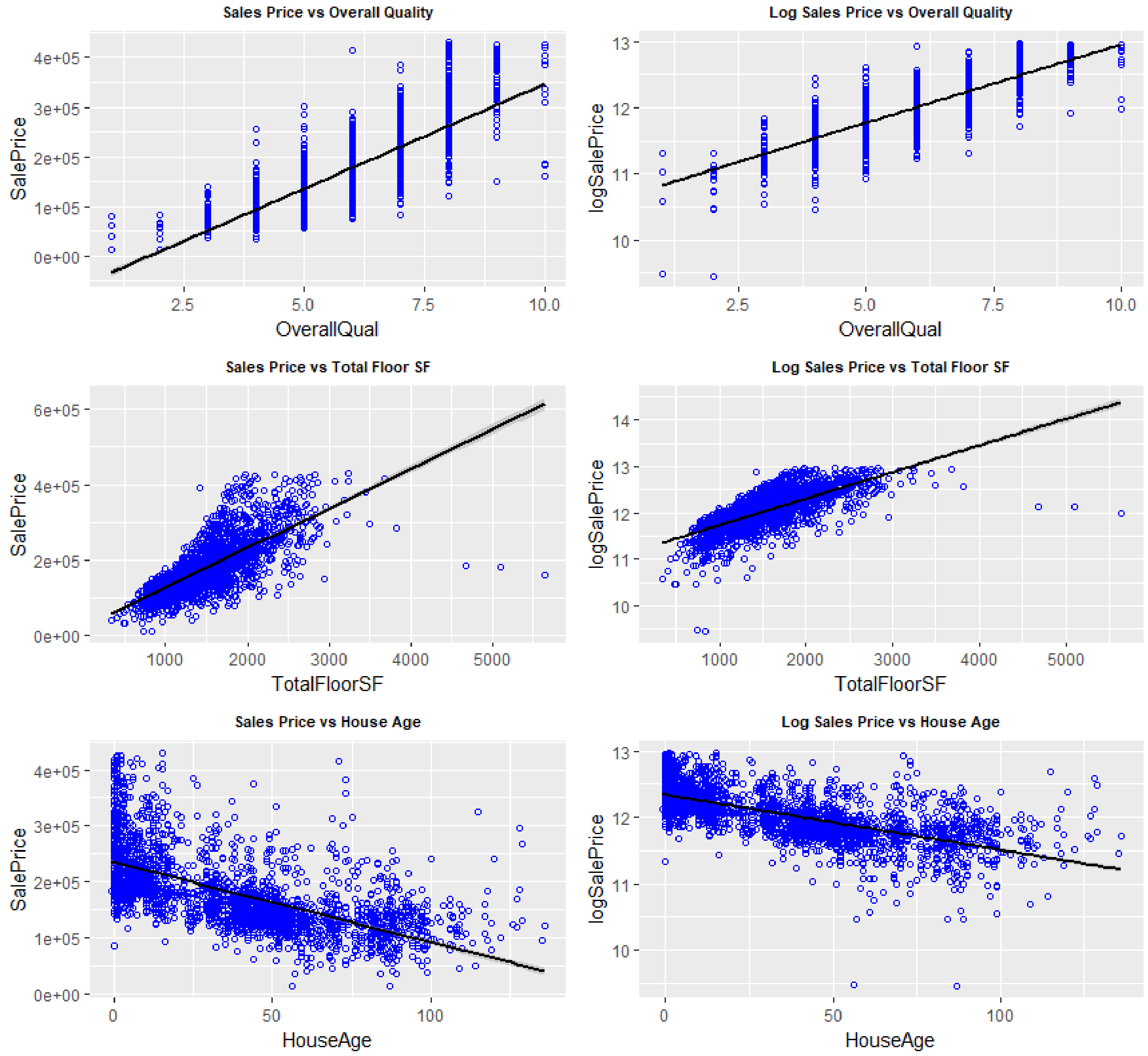
After dropping the low correlated varaibles and lot shape, I ended up with 7 numerical variables and 3 categorical varaibles.


SalePrice
OverallQual
price_sqft
TotalFloorSF
TotalBsmtSF
FullBath
HouseAge
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


SalePrice
OverallQual
price_sqft
TotalFloorSF
TotalBsmtSF
FullBath
HouseAge
Corr:
0.805
Corr:
0.46
Corr:
0.767
Corr:
0.607
Corr:
0.637
Corr:
-0.598
Corr:
0.385
Corr:
0.607
Corr:
0.487
Corr:
0.569
Corr:
-0.551
Corr:
-0.186
Corr:
0.439
Corr:
0.0707
Corr:
-0.522
Corr:
0.344
Corr:
0.664
Corr:
-0.288
Corr:
0.321
Corr:
-0.41
Corr:
-0.497
7.5e-06
5.0e-06
2.5e-06
0.0e+00
10.0
7.5
5.0
2.5
200
150
100
50
3000
2000
1000
3000
2000
1000
0
3
2
1
0
100
50
0
1e+05 2e+05 3e+05 4e+05 5e+05
2.5 5.0 7.5 10.0
50 100 150 200
1000 2000 3000
0 1000 2000 3000 0000
1 2 3
0 50 100

Section 4: Exploratory Data Analysis for Modeling

We need to select the top 3 varaibles which can predict the sales price of homes. Overall quality, total floor square footage, and the age of property relate well with sales price and they should be the top 3 predictors. We can see the trend and relationship visually below. The left side compares top three varaibles with sales price and left compares them with transformed log sales price.


Sales Price vs Overall Quality

Log Sales Price vs Overall Quality

Sales Price vs Total Floor SF

Log Sales Price vs Total Floor SF

Sales Price vs House Age

Log Sales Price vs House Age

We can see that as overall quality and total floor sqft go up, the price also goes up. As the age of house goes down, the price goes up. This means that newer houses sell higher price than the older houses.

Section 5: Summary/Conclusions

We defined sample size, performed data quality check on selected features, removed price outliers and cleaned data, and performed the basic data exploration. The dataset is relatively clean. With data exploration we were able to eliminate noise and narrow down the predictors of homes. After performing data exploration, we can conclude that there is relationship between property price and predictors. Our goal is to predict the sales price of the property by given data. Overall quality of single-family homes, total floor square footage, and property age are the top three indicators of the property price.