Alisher Siddikov

Assignment 1

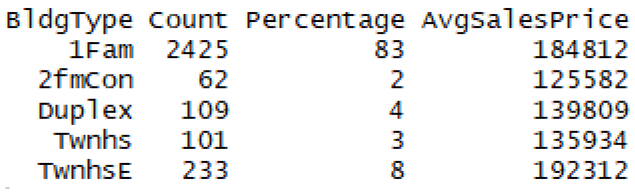
Introduction:

We will explore the housing dataset of Ames, Iowa. The housing data has 2930 observations and 82 variables, and it was obtained from the Ames Assessor's Office which was used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. The type of information contained in the data is similar to what a typical home buyer would want to know before making a purchase. This housing data is an alternative to the Boston Housing dataset.

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. We will need to run a data quality check and examine the data for errors and outliers. We will also need to examine the strength of relationship of target and predicted varaibles.

Results:

Section 1: Sample Definition



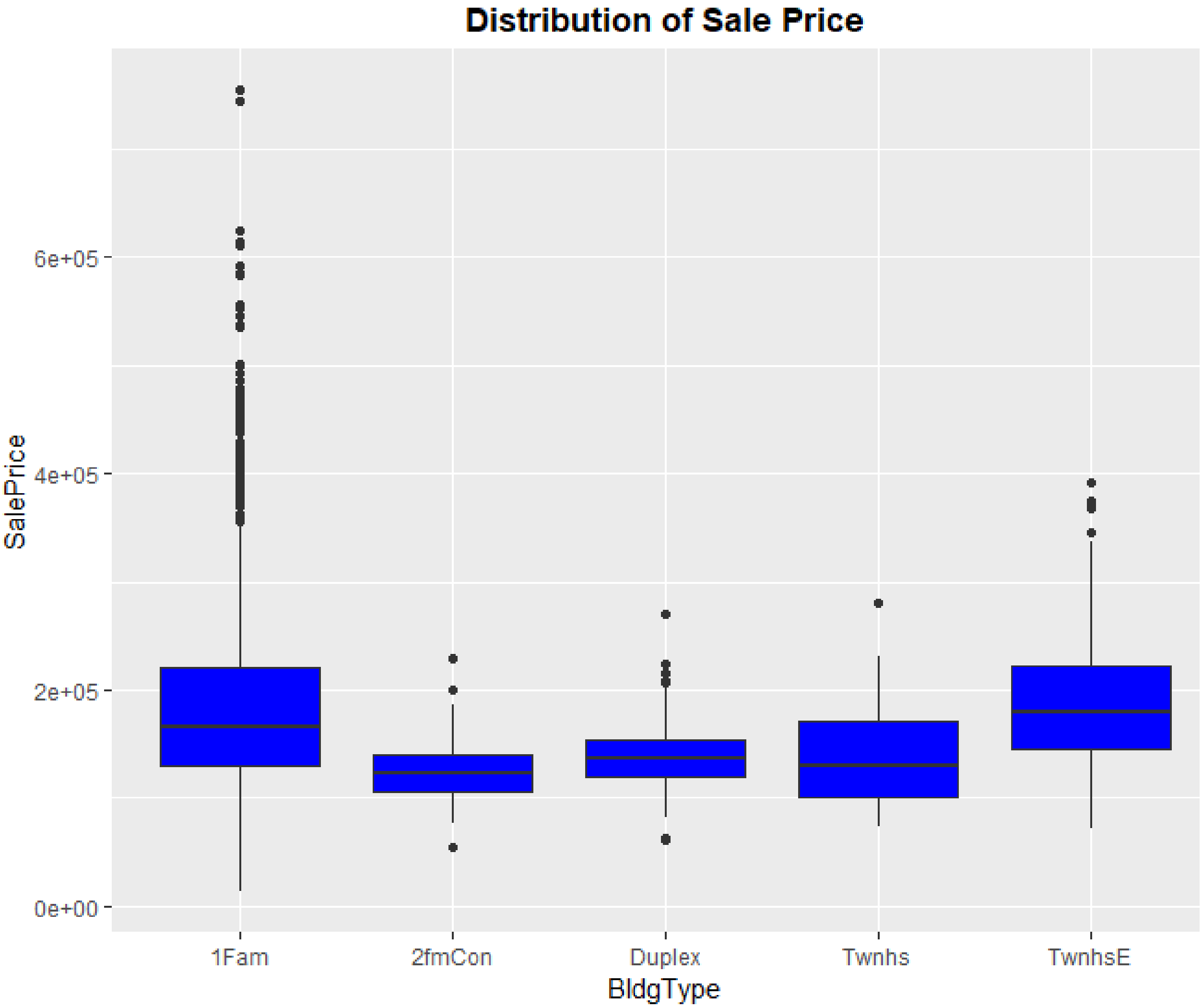
```
BldgType Count Percentage AvgSalesPrice
    1Fam  2425         83        184812
  2fmCon    62          2        125582
  Duplex   109          4        139809
   Twnhs   101          3        135934
  TwnhsE   233          8        192312
```

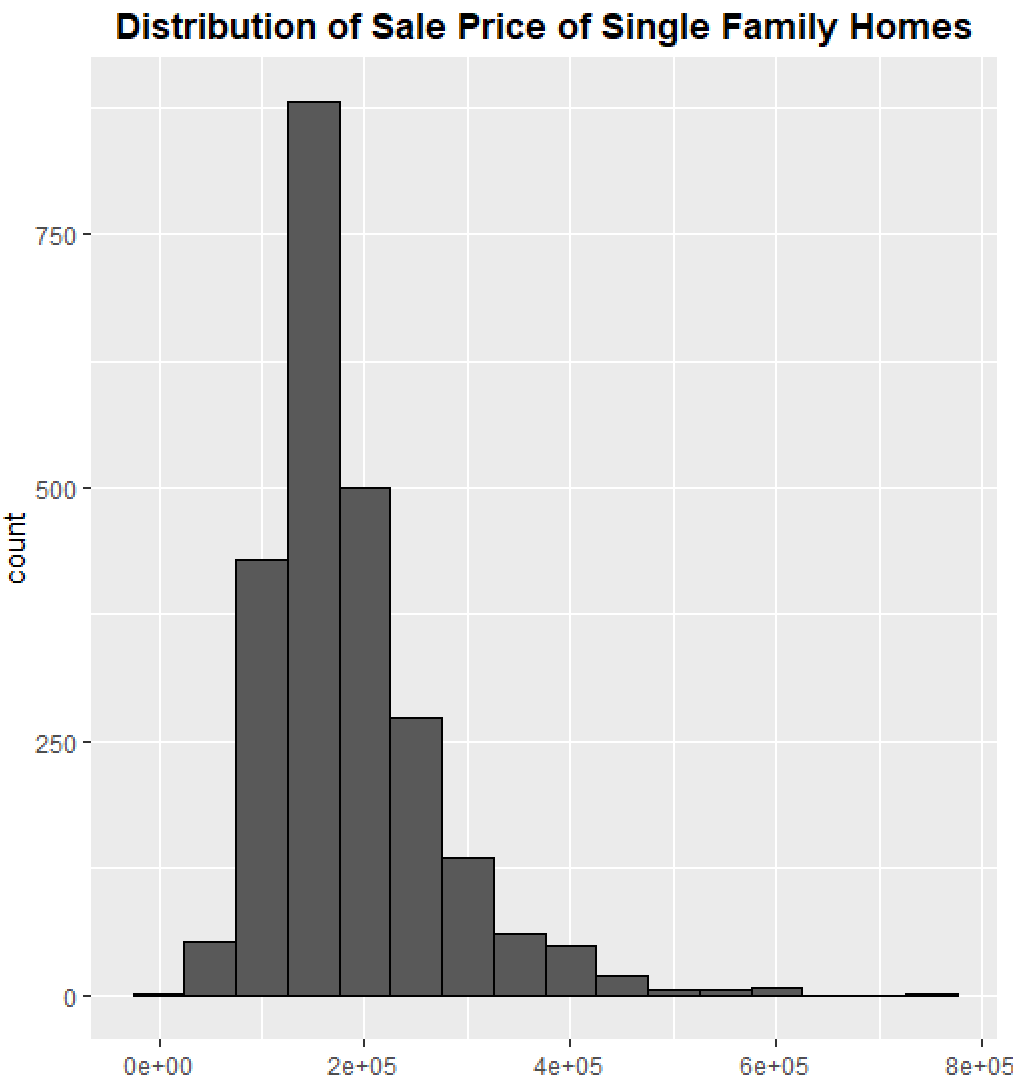
If we look at the type of dwelling, single family homes represent most of the data (83%). We can eliminate two family conversion, duplex, and townhouse dwelling types.

After excluding any homes not of type of single family, the data shape changed. Here are the updated observation numbers:

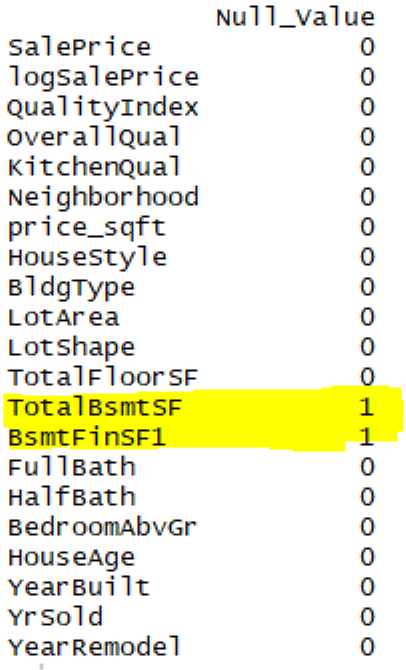
- Updated row numbers: 2,425; 17% of it eliminated
- Updated column numbers: 5 derived new rows were added: total floor square feet, house age, quality index, log sales price, and price per square feet.

Section 2: Data Quality Check





Sales price of single-family home has some outliers and it should be eliminated. There are no negative or missing sales price. Used 3 standard deviation to remove the 39 outliers from the sales price of single-family home. After sub selecting 20 variables, there were no missing data except basements.



```
              Null_Value
SalePrice              0
logSalePrice           0
QualityIndex           0
OverallQual            0
KitchenQual            0
Neighborhood           0
price_sqft             0
HouseStyle             0
BldgType               0
LotArea                0
LotShape               0
TotalFloorSF           0
TotalBsmtSF            1
BsmtFinSF1             1
FullBath               0
HalfBath               0
BedroomAbvGr           0
HouseAge               0
YearBuilt              0
YrSold                 0
YearRemodel            0
```

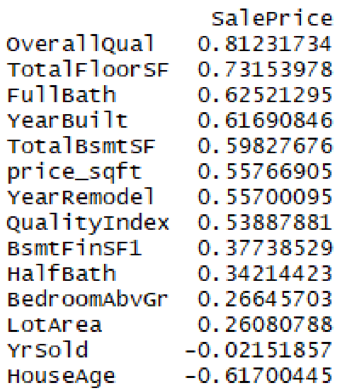
The data quality results with discussion for the twenty selected variables

After excluding NA rows and sub selecting 20 rows, the data shape changed. Here are the updated numbers:

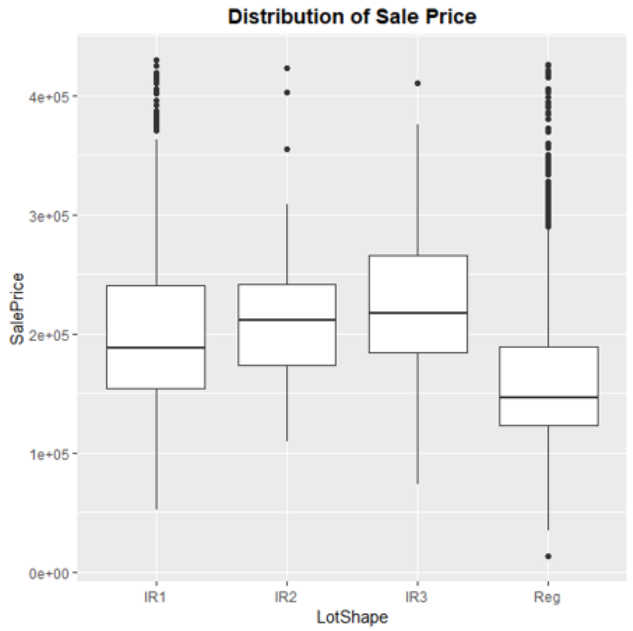
- Updated row numbers: 2,385; one NA row and 39 outliers were eliminated
- Updated column numbers: 20

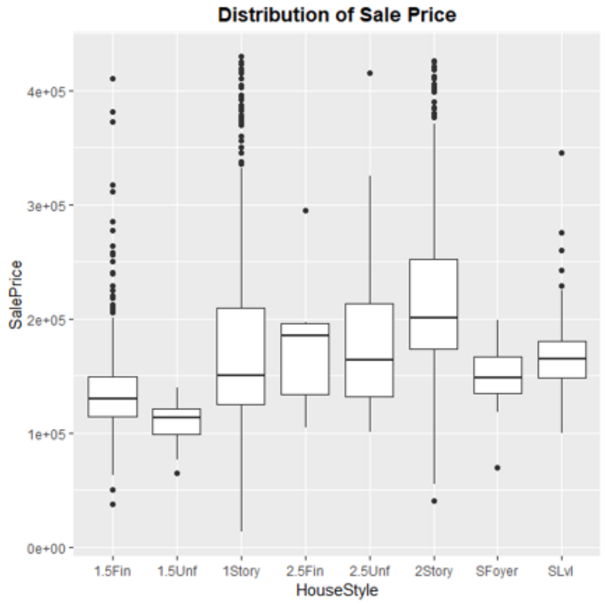
Section 3: Initial Exploratory Data Analysis

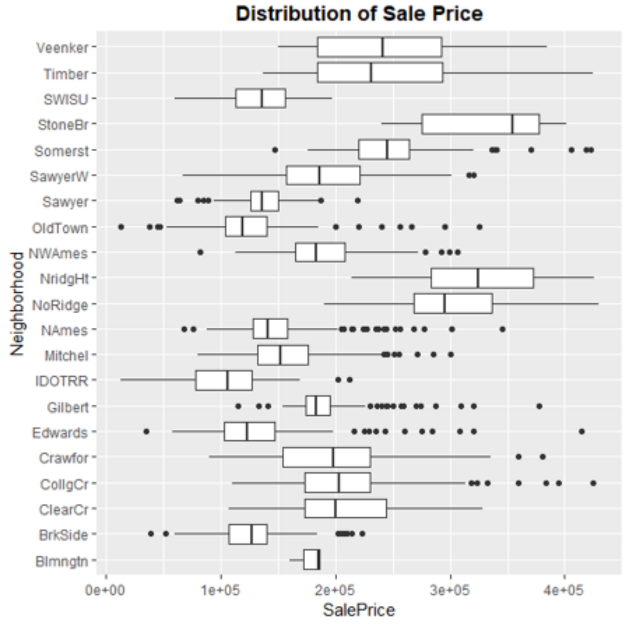
Twenty columns were split into categorial and numerical varaibles. Here is the correlation table between sale price and selected numerical variables. Here are the rest of categorial varaibles: KitchenQual, Neighborhood, HouseStyle, and LotShape. After exploring the data further, we need to select ten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

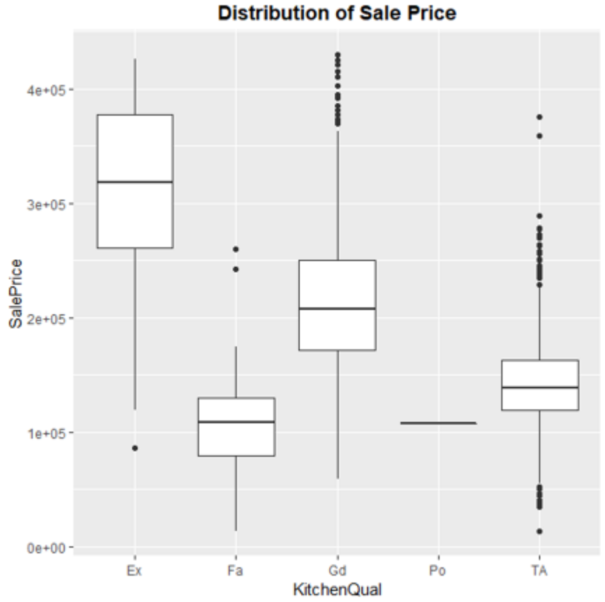


```
                 SalePrice
OverallQual     0.81231734
TotalFloorSF    0.73153978
FullBath        0.62521295
YearBuilt       0.61690846
TotalBsmtSF     0.59827676
price_sqft      0.55766905
YearRemodel     0.55700095
QualityIndex    0.53887881
BsmtFinSF1      0.37738529
HalfBath        0.34214423
BedroomAbvGr    0.26645703
LotArea         0.26080788
YrSold         -0.02151857
HouseAge       -0.61700445
```



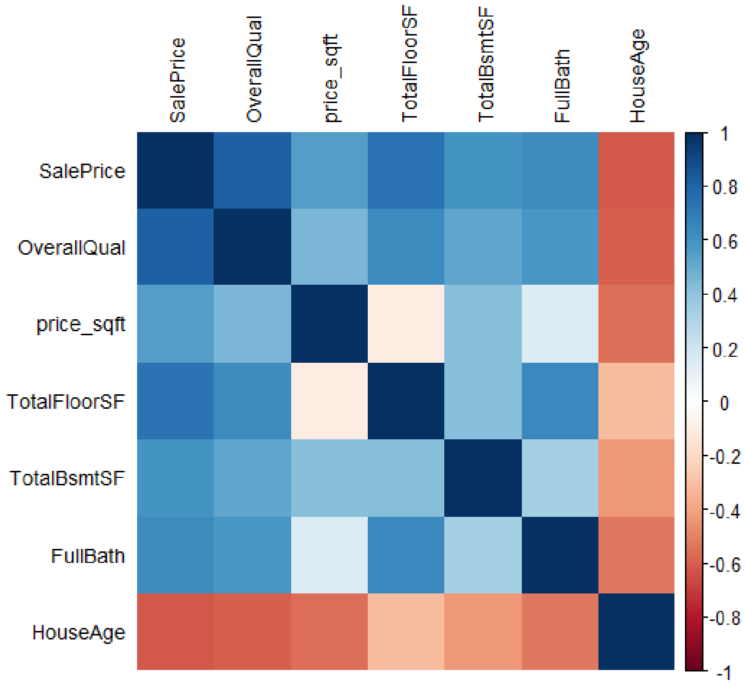


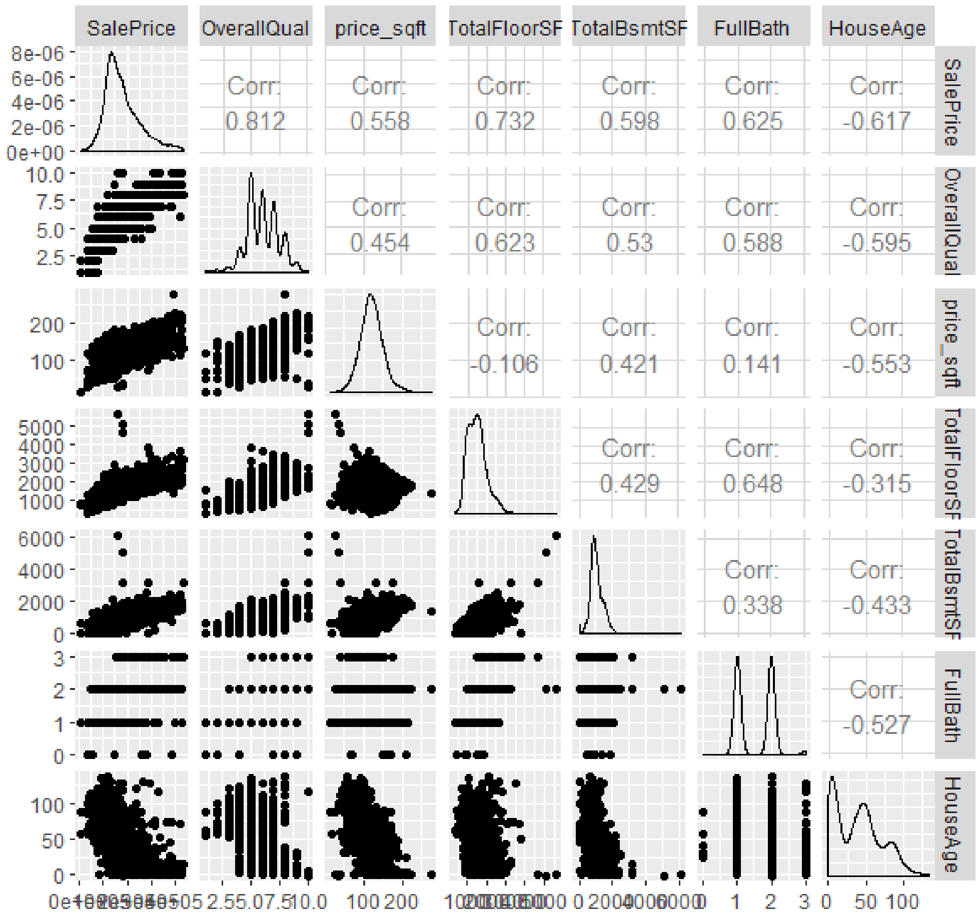




We can see that house style, neighborhood, and kitchen quality gives us sale price indication. We will drop lot shape.

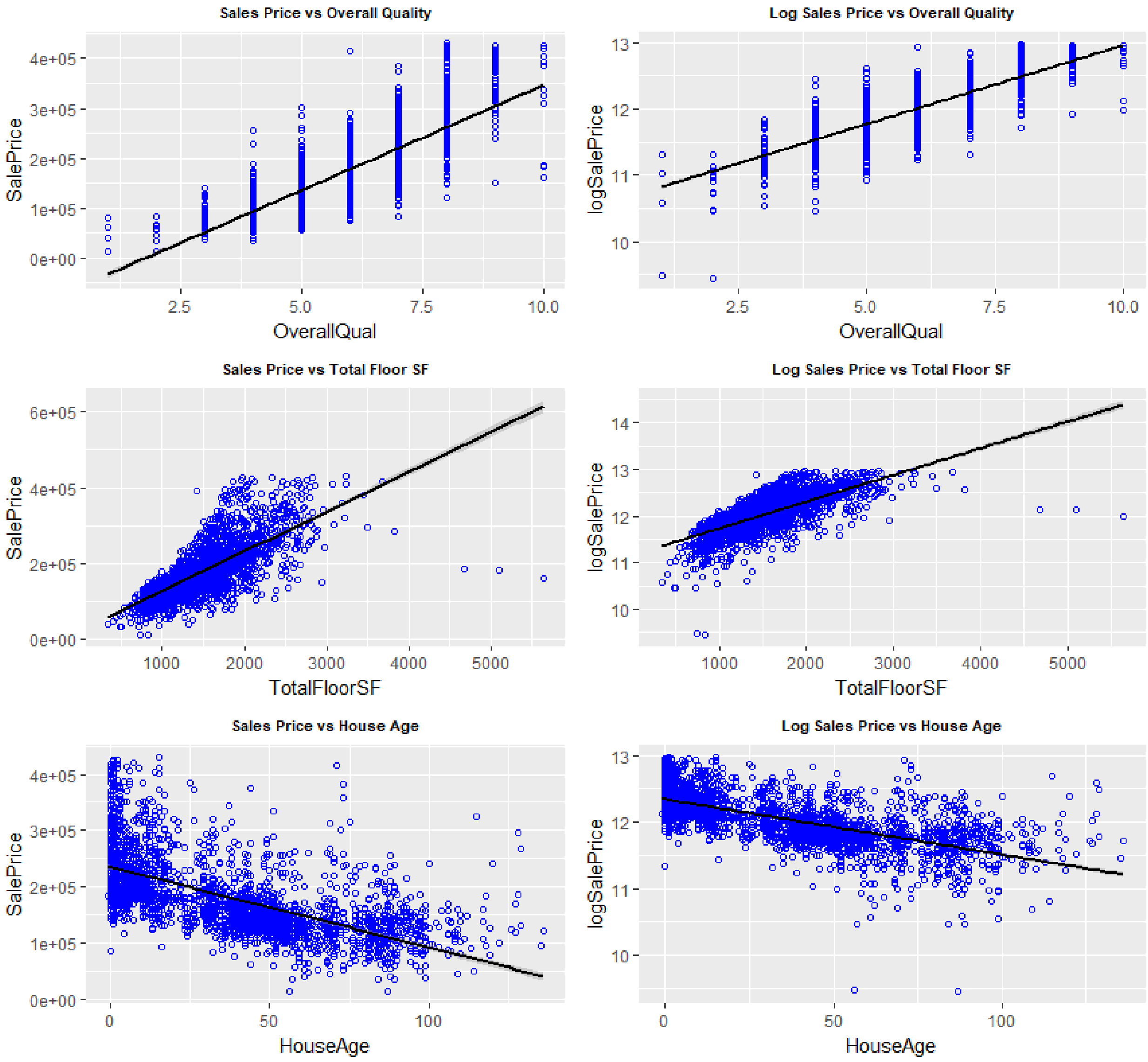
After dropping the low correlated varaibles and lot shape, I ended up with 7 numerical variables and 3 categorical varaibles.





SalePrice
OverallQual
price_sqft
TotalFloorSF
TotalBsmtSF
FullBath
HouseAge
Corr:
0.812
Corr:
0.558
Corr:
0.732
Corr:
0.598
Corr:
0.625
Corr:
-0.617
Corr:
0.454
Corr:
0.623
Corr:
0.53
Corr:
0.588
Corr:
-0.595
Corr:
-0.106
Corr:
0.421
Corr:
0.141
Corr:
-0.553
Corr:
0.429
Corr:
0.648
Corr:
-0.315
Corr:
0.338
Corr:
-0.433
Corr:
-0.527
8e-06
6e-06
4e-06
2e-06
0e+00
10.0
7.5
5.0
2.5
200
100
5000
4000
3000
2000
1000
6000
4000
2000
0
3
2
1
0
100
50
0
2.5
5.0
7.5
10.0
100
200
0
2000
4000
6000
0
1
2
3
0
50
100

Section 4: Exploratory Data Analysis for Modeling



We can see that as overall quality and total floor sqft go up, the price also goes up. Newer houses sold 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; there were one NA row only. 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 best indicators of the property price.