Assessing Home Values

By: Vladimir Antasiuk & Nick Zelada

Project for STA6244

Table of Contents

Goal

House prices are a crucial indicator of the economic condition, and it is also a point of great interest for buyers and sellers. Both parties are interested in the factors that affect the price. From one side, the buyers are interested in evaluating the house price to understand what they can afford and what factors affect the price. The correct understanding of house price drivers will help a buyer find the house with suitable parameters according to their budget. At the same time, it is evident that buyers do not want to overpay. So, they need to know the house's fair price. On the seller side (private or corporate), they need to estimate the asset's value they possess. It will help forecast potential future revenue, estimate the market conjecture (prices, competitors, consumers), elaborate the efficient promotional campaign, etc.

The buyers are interested in selling the house for the higher price; however, it is evident that the price still should be relevant and does not put customers off. There are a lot of scientific papers devoted to the problem of the prediction of house prices. Most of them are using multiple linear regression or some type of Machine Learning Techniques. For instance, the paper by Quang Truong et al. (2019) Examines and compares the accuracy of such Machine Learning methods as Random Forest, XGBoost, and LightGBM. Hybrid Regression and Stacked Generalization Regression are two machine learning techniques predicting house prices using the "Housing Price in Beijing" dataset. The dataset contains more than 300,000 data, with 26 variables representing housing prices traded between 2009 and 2018. Some of the variables included are features such as the house's age, the area of the house and number of bedrooms, etc. The Stacked Generalization Regression gave the smallest RMSLE on the test dataset (0.16350) (Quang, 2020).

Dubin (1998) uses OLS regression to predict house prices based on different features using the data from multiple listings from Baltimore and Maryland (1978). The paper also aims to ignore the existing correlations between neighboring houses' prices and suggest solving this issue by estimating the regression coefficients using the maximum likelihood method and kriging. (Robin, 1998)

However, a more interesting example is the Zillow company. Today Zillow is probably the most famous property selling website, accumulating millions of buyers, sellers, agents, advertisers, and people interested in real estate. The company started as a property listing website, but the management quickly realized that the most valuable asset is data. The company accumulated a vast amount of information from thousands of official sources and now provides comprehensive information about the listed properties:

- Lot area, number of bedrooms, bathrooms, building type, etc.
- Historical prices of different types of properties in different areas and regions of the US.
- The district's ethnicity composition as the percentage and the crime rate (number of crimes and their nature).
- The number of schools and the quality of education (up to the average score). The comfort level of the area (infrastructure development in the form of hospitals/shops/cafes, accessibility on foot/by car) and many more.

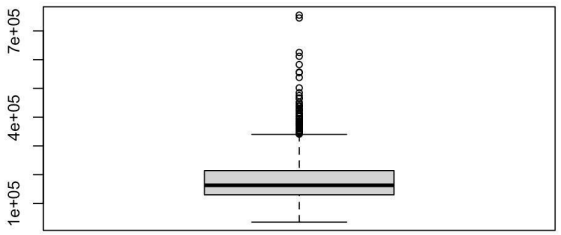
After processing this data, Zillow created its formula, ‘Zestimate,’ and registered it as a brand. The company promotes this formula as the fairest algorithm for a house price estimation in the US, and this formula has a significant influence on the pricing at the US real estate market. (Zillow, 2020)

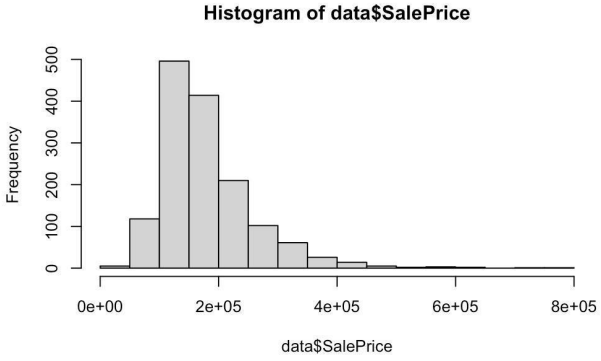
Our project's primary purpose is to implement the knowledge, instruments, and techniques learned in class in solving real-world problems. In particular, we want to build a linear regression model that gives accurate predictions of the house prices based on the dataset. Additionally, to see if there is any correlation based on the variables and how they play in the sales price.

Dataset Description and Data Exploration

The dataset we will use for this project is the "Ames Housing Data" created by Dean De Cock. As a part of his work as a professor, he used the famous Boston Housing Data Set for his regression class (Carnegie Mellon University, 1980). However, this dataset only contains 506 observations and 14 variables and is quite outdated: the original dataset came from the 70s, resulting in data not corresponding to today's modern real estate market. To obtain a more relevant dataset, Dean De Cock, in collaboration with the students of Iowa State StatCom and the Ames City Assessor's Office, created the new Ames Housing data set (Dean, 2011). The data describes the sale of individual residential property in Ames, Iowa, from 2006 to 2010. The original dataset contains 2930 observations and many explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous). Involving in assessing home values in such as lot size in square feet, number of fireplaces, the pool area in square feet, number of kitchens, kitchen quality, type of roof, overall material, finish quality, overall condition rating, the property's sale price in dollars, the building class, the general zoning classification, proximity to the main road or railroad, type of dwelling, year build, etc. The dataset available at Kaggle contains 2920 observations equally divided into the training and test set where the test does not include the values of target variables. More information can be found at the Kaggle or in the article that describes this dataset (Kaggle, 2020).

Below we provide boxplot, histogram and some summary statistics regarding the SalePrice from the train set





Source code used:

```
> fivenum(data$SalePrice)
  34900 129950 163000 214000 755000
> mean(data$SalePrice)
 180921.2
> library(moments)
> kurtosis(data$SalePrice) = 9.509812
> skewness(data$SalePrice) = 1.880941
```

The mean is greater than the median, while the higher tail is higher than the lower tail. This results for a right skewed, a highly right skewed since it is 1.88 and leptokurtic kurtosis being 9.5

The Implementation of the Project

For the project, some of the implementations used were literature review and real-world applications. We also did data exploration such as sample statistics, graphs, normality tests, checking for missing values from the train data. We used the best subset selection to choose the linear regression model, evaluate the training model, and apply it to a test set.

Literature Review and Real-World Applications

During this project, we read the paper by Quang Truong et al. (2019) Examines and compare the accuracy of such Machine Learning methods as Random Forest, XGBoost, and LightGBM. Hybrid Regression and Stacked Generalization Regression are two machine learning techniques predicting house prices using the “Housing Price in Beijing” dataset. We took a look at Dubin (1998) usage on OLS regression to predict house prices based on different features using the data from multiple listings from Baltimore and Maryland (1978). Lastly, Zillow Company, to check out how they use factors to determine pricing of a house.

Discussion of Methods and Instruments

The methods used in the project is the Multiple OLS Linear Regression. Linear Regression usage will allow us to make accurate predictions on house prices from the dataset and see any correlation. To evaluate the precision of the model, we chose adjusted R-squared and MSE as metrics. We used the Shapiro-Wilk test for the normality test, where Ho is that data comes from a normal distribution. To choose the best subset for linear regression, we use the

'Leaps' library in R. We used AIC criteria to pick the best subset. The model with the smallest AIC was chosen.

The null hypothesis here is: Contributing factors of a house do not affect the price of houses.
The alternative hypothesis here is: Contributing factors of a house affect the price of houses.

The Best Subset Selection

With the high number of variables in the dataset and some issues with the variables, we could not run the subset selection on the full dataset because different functions for subset selection in R gave us an error. A high correlation between some variables could have caused it. Thus, we had to manually construct our regression model, taking the problem's logic into account. So, by repeatedly changing/adding/subtracting regressors, we found the model consisting of 27 variables and accuracy of about 84% (adjusted R-squared). We also cleaned the data by using na.omit.

Here is the code when we first started to do it manually with the output.

```
> model1=lm(SalePrice~., data=na.omit(data_best_subset_sel))
> summary(model1)
Residual standard error: 31280 on 1345 degrees of freedom
Multiple R-squared:  0.8557, Adjusted R-squared:  0.8443
F-statistic: 75.26 on 106 and 1345 DF,  p-value: < 2.2e-16
```

From there we used the Leaps Library to use AIC to help us find a regression model, here is the code for it.

```
> data_subset_sel_Clean=na.omit(data_subset_sel)
> full.tr = lm(SalePrice~., data=data_subset_sel_Clean) #full model
> Intercept.tr=lm(SalePrice~1, data=data_subset_sel_Clean) #the smallest model, intercept only
>step.tr_both=stepAIC(Intercept.tr,scope=list(upper=full.tr,lower=Intercept.tr),direction="both")
> step.tr_backward=stepAIC(full.tr,direction="backward")
>step.tr_forward=stepAIC(Intercept.tr,scope=list(upper=full.tr,lower=Intercept.tr),direction="forward")
> summary(step.tr_both)
> summary(step.tr_forward)
```

From the three models developed, “Both” and “Forward” gave us the most accurate model, then from having it constructed manually. Below you will see the source code “Both.”

```
> summary(step.tr_both)

Call:
lm(formula = SalePrice ~ OverallQual + GrLivArea + Neighborhood +
RoofMatl + HouseStyle + ExterQual + BldgType + BsmtFinType1 + YearBuilt +
OverallCond + LotArea + Fireplaces + PoolArea + MasVnrType + MasVnrArea +
Condition1 + LotConfig + Foundation + LandSlope + BsmtFinType2 +
LowQualFinSF+YearRemodAdd +Exterior1st, data = data_subset_sel_Clean)

Residuals:
Min      1Q  Median      3Q     Max
-331718  -12123       0   10925  203533

Coefficients:
Estimate  Std. Error      t value Pr(>|t|)
(Intercept)         -1.610e+06  1.711e+05  -9.410  < 2e-16 ***
OverallQual          9.947e+03  1.124e+03   8.847  < 2e-16 ***
GrLivArea            6.907e+01  2.958e+00  23.350  < 2e-16 ***
NeighborhoodBlueste -6.449e+03  2.195e+04  -0.294 0.768964
NeighborhoodBrDale   6.139e+03  1.159e+04   0.530 0.596455
NeighborhoodBrkSide -1.333e+04  9.707e+03  -1.373 0.169989
NeighborhoodClearCr -2.078e+04  1.026e+04  -2.027 0.042887 *
NeighborhoodCollgCr -1.971e+04  8.043e+03  -2.451 0.014391 *
NeighborhoodCrawfor -8.516e+02  9.376e+03  -0.091 0.927644
NeighborhoodEdwards -2.756e+04  8.825e+03  -3.124 0.001826 **
NeighborhoodGilbert -2.560e+04  8.623e+03  -2.969 0.003039 **
NeighborhoodIDOTRR  -2.440e+04  1.021e+04  -2.390 0.017001 *
NeighborhoodMeadowV -8.154e+03  1.196e+04  -0.682 0.495359
NeighborhoodMitchel -2.605e+04  9.039e+03  -2.881 0.004023 **
NeighborhoodNAmes   -2.472e+04  8.563e+03  -2.887 0.003954 **
NeighborhoodNoRidge  1.709e+04  9.391e+03   1.820 0.068974 .
NeighborhoodNPkVill  1.187e+04  1.283e+04   0.925 0.354951
NeighborhoodNridgHt  2.448e+04  8.446e+03   2.898 0.003814 **
NeighborhoodNWAmes  -3.000e+04  8.869e+03  -3.382 0.000739 ***
NeighborhoodOldTown -2.506e+04  9.319e+03  -2.689 0.007266 **
NeighborhoodSawyer  -2.040e+04  8.992e+03  -2.268 0.023469 *
NeighborhoodSawyerW -1.824e+04  8.760e+03  -2.082 0.037522 *
NeighborhoodSomerst -1.708e+03  8.364e+03  -0.204 0.838250
NeighborhoodStoneBr  4.367e+04  9.400e+03   4.645 3.73e-06 ***
NeighborhoodSWISU   -2.703e+04  1.065e+04  -2.538 0.011270 *
NeighborhoodTimber  -9.703e+03  9.115e+03  -1.065 0.287272
NeighborhoodVeenker  5.133e+03  1.183e+04   0.434 0.664471
```

```
RoofMatlCompShg      5.193e+05  3.204e+04  16.207  < 2e-16 ***
RoofMatlMembran      5.627e+05  4.549e+04  12.368  < 2e-16 ***
RoofMatlMetal        5.712e+05  4.529e+04  12.612  < 2e-16 ***
RoofMatlRoll         5.035e+05  4.322e+04  11.650  < 2e-16 ***
RoofMatlTar&Grv      5.060e+05  3.308e+04  15.296  < 2e-16 ***
RoofMatlWdShake      5.260e+05  3.510e+04  14.983  < 2e-16 ***
RoofMatlWdShngl      6.123e+05  3.380e+04  18.115  < 2e-16 ***
HouseStyle1.5Unf     1.407e+04  8.155e+03   1.726 0.084649 .
HouseStyle1Story     1.545e+04  3.283e+03   4.706 2.79e-06 ***
HouseStyle2.5Fin    -1.496e+04  1.299e+04  -1.151 0.249834
HouseStyle2.5Unf    -1.223e+04  9.225e+03  -1.326 0.185201
HouseStyle2Story    -4.587e+03  3.217e+03  -1.426 0.154130
HouseStyleSFoyer     1.874e+04  6.219e+03   3.013 0.002632 **
HouseStyleSLvl       3.658e+03  4.863e+03   0.752 0.452100
ExterQualFa         -3.847e+04  1.111e+04  -3.464 0.000550 ***
ExterQualGd         -4.389e+04  5.048e+03  -8.694  < 2e-16 ***
ExterQualTA         -4.464e+04  5.662e+03  -7.884 6.62e-15 ***
BldgType2fmCon      -9.390e+03  5.611e+03  -1.674 0.094428 .
BldgTypeDuplex      -1.743e+04  5.333e+03  -3.269 0.001107 **
BldgTypeTwnhs       -4.409e+04  5.841e+03  -7.549 8.19e-14 ***
BldgTypeTwnhsE      -3.168e+04  3.781e+03  -8.379  < 2e-16 ***
BsmtFinType1BLQ     -3.575e+01  3.171e+03  -0.011 0.991005
BsmtFinType1GLQ      8.492e+03  2.891e+03   2.938 0.003364 **
BsmtFinType1LwQ     -8.721e+03  4.168e+03  -2.093 0.036574 *
BsmtFinType1Rec     -3.678e+03  3.406e+03  -1.080 0.280377
BsmtFinType1Unf     -1.029e+04  2.762e+03  -3.726 0.000203 ***
YearBuilt            4.486e+02  7.270e+01   6.172 8.99e-10 ***
OverallCond          5.754e+03  9.032e+02   6.371 2.60e-10 ***
LotArea              6.668e-01  1.076e-01   6.195 7.78e-10 ***
Fireplaces           5.921e+03  1.516e+03   3.905 9.90e-05 ***
PoolArea             1.018e+02  2.041e+01   4.986 6.99e-07 ***
MasVnrTypeBrkFace    1.418e+04  7.723e+03   1.836 0.066612 .
MasVnrTypeNone       1.937e+04  7.766e+03   2.494 0.012757 *
MasVnrTypeStone      2.664e+04  8.174e+03   3.259 0.001147 **
MasVnrArea           2.875e+01  6.637e+00   4.332 1.59e-05 ***
Condition1Feedr      8.537e+02  5.589e+03   0.153 0.878621
Condition1Norm       9.756e+03  4.601e+03   2.120 0.034161 *
Condition1PosA       4.224e+03  1.138e+04   0.371 0.710648
Condition1PosN      -1.559e+04  8.116e+03  -1.920 0.055016 .
Condition1RRAe      -1.831e+04  1.036e+04  -1.768 0.077330 .
Condition1RRAn       7.562e+03  7.443e+03   1.016 0.309844
Condition1RRNe      -5.644e+03  2.084e+04  -0.271 0.786610
Condition1RRNn       4.121e+03  1.410e+04   0.292 0.770109
LotConfigCulDSac     1.023e+04  3.665e+03   2.792 0.005322 **
```

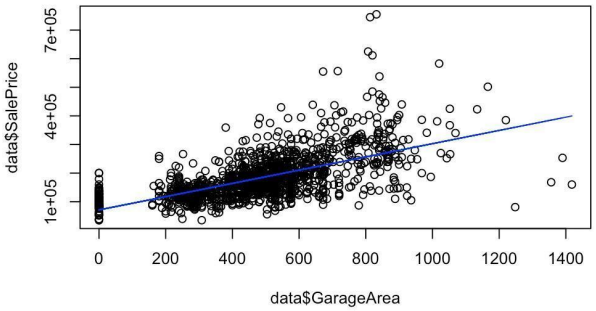
```
LotConfigFR2         -5.387e+03  4.640e+03  -1.161 0.245897
LotConfigFR3         -1.045e+04  1.464e+04  -0.714 0.475419
LotConfigInside      -1.385e+03  2.030e+03  -0.682 0.495293
FoundationCBlock      4.886e+02  3.506e+03   0.139 0.889169
FoundationPConc       6.979e+03  3.920e+03   1.781 0.075193 .
FoundationStone      -5.074e+03  1.201e+04  -0.422 0.672761
FoundationWood       -2.761e+04  1.690e+04  -1.634 0.102595
LandSlopeMod          7.259e+03  3.934e+03   1.845 0.065202 .
LandSlopeSev         -2.421e+04  1.127e+04  -2.149 0.031852 *
BsmtFinType2BLQ      -2.312e+04  8.334e+03  -2.775 0.005604 **
BsmtFinType2GLQ      -8.980e+03  1.074e+04  -0.836 0.403045
BsmtFinType2LwQ      -2.595e+04  8.104e+03  -3.202 0.001396 **
BsmtFinType2Rec      -2.436e+04  7.823e+03  -3.114 0.001885 **
BsmtFinType2Unf      -1.949e+04  6.829e+03  -2.854 0.004384 **
LowQualFinSF         -3.937e+01  1.981e+01  -1.988 0.047066 *
YearRemodAdd          1.134e+02  5.932e+01   1.913 0.056026 .
Exterior1stBrkComm   -5.299e+04  2.891e+04  -1.833 0.067027 .
Exterior1stBrkFace    1.406e+04  8.443e+03   1.665 0.096131 .
Exterior1stCBlock     1.894e+03  3.112e+04   0.061 0.951482
Exterior1stCemntBd   -2.445e+02  8.749e+03  -0.028 0.977712
Exterior1stHdBoard   -6.963e+03  7.570e+03  -0.920 0.357864
Exterior1stImStucc   -1.664e+04  2.925e+04  -0.569 0.569483
Exterior1stMetalSd    1.886e+03  7.335e+03   0.257 0.797104
Exterior1stPlywood   -5.108e+03  7.945e+03  -0.643 0.520362
Exterior1stStone     -2.790e+04  2.306e+04  -1.210 0.226546
Exterior1stStucco     7.339e+01  9.418e+03   0.008 0.993783
Exterior1stVinylSd   -1.659e+03  7.470e+03  -0.222 0.824266
Exterior1stWd Sdng    4.635e+02  7.331e+03   0.063 0.949594
Exterior1stWdShing   -4.206e+03  9.226e+03  -0.456 0.648568
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27730 on 1314 degrees of freedom
Multiple R-squared:  0.886,   Adjusted R-squared:  0.8775
F-statistic: 103.2 on 99 and 1314 DF,  p-value: < 2.2e-16
```

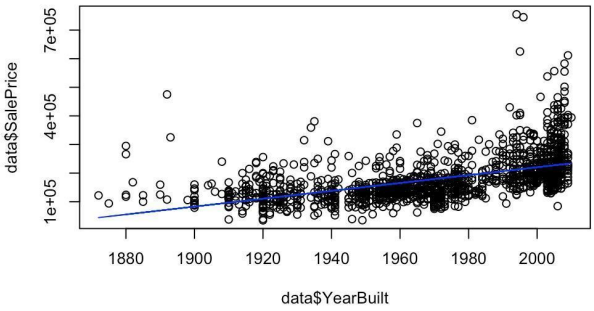
Which lead us to the final model that we work with.

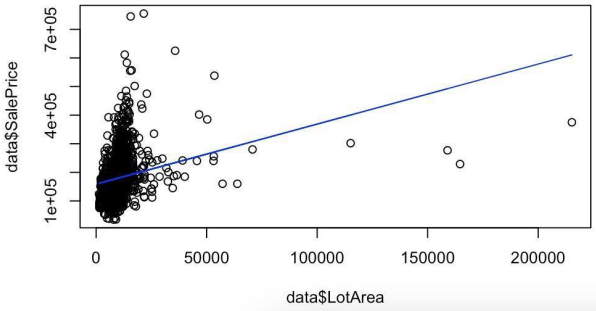
```
lm(formula = SalePrice ~ OverallQual + GrLivArea + Neighborhood +
RoofMatl + HouseStyle + ExterQual + BldgType + BsmtFinType1 + YearBuilt +
OverallCond + LotArea + Fireplaces + PoolArea + MasVnrType + MasVnrArea +
```

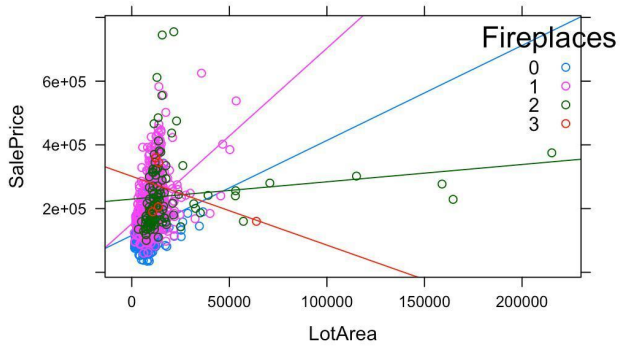
```
    Condition1 + LotConfig + Foundation + LandSlope + BsmtFinType2 +
    LowQualFinSF + YearRemodAdd + Exterior1st, data = data_subset_sel_Clean)

1.  err.tr=mean((data_subset_sel_Clean$SalePrice-Y_pr_tr)^2)
2.  > err.tr
3.   714564136
```

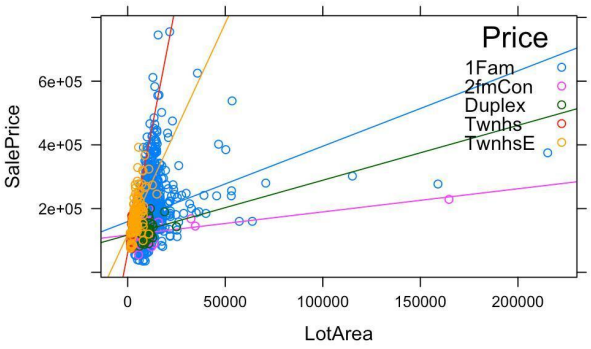
We also look for Mean Square Error in which we got a value of 714564136.

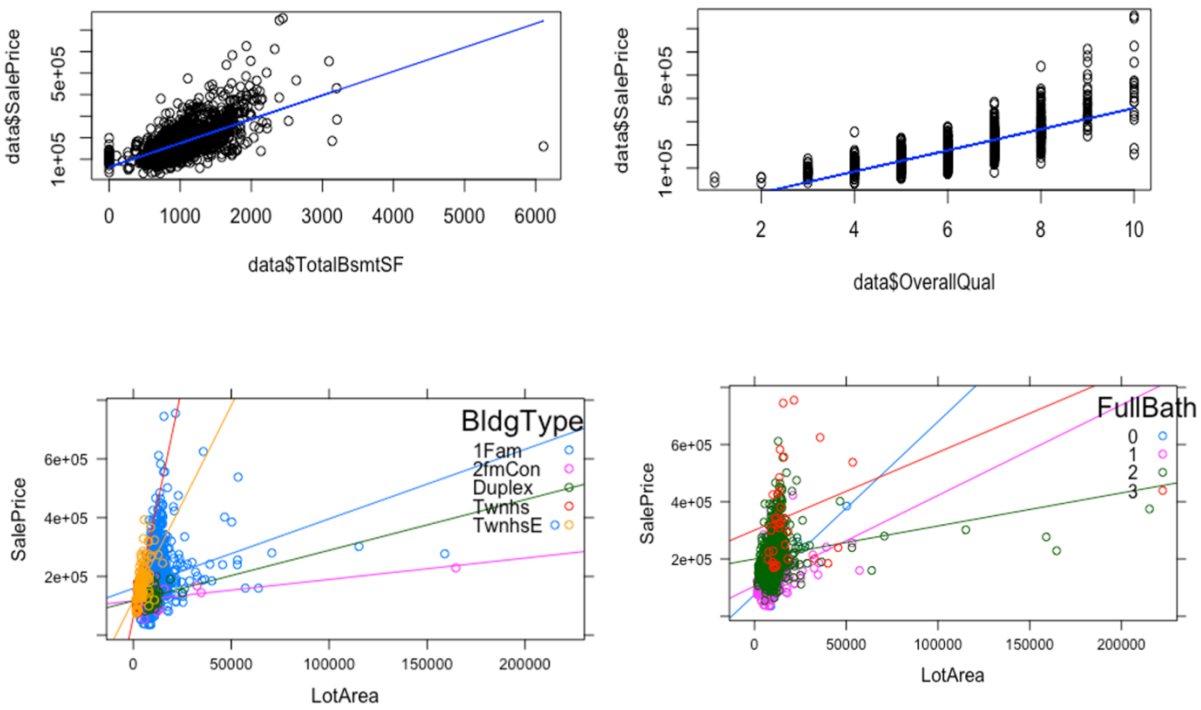












In the first three sets of graphs, we see a positive correlation with price and GarageArea, YearBuilt, and LotArea. While the next two sets of graphs, we are looking at SalePrice with LotArea, but on the left graph, we see they are categorized by the number of fireplaces in the house and how the price changes. In the next set, we see the positive correlation of TotalBsmtSF and OverallQual with Saleprice. While lastly, we have SalePrice with LotArea, but on the left graph, we categorized the building type and how it differs in price. Simultaneously, the right graph we see is categorized by the number of FullBath and how it differs from price.

Prediction

Unfortunately, this dataset came from the Kaggle competition, so we do not have the target variable's actual values in the test set. However, if we took a look at the predicted values, it looks like the predicted values look reasonable. When running the prediction, the values given to us tell us prices that would have been in a real-world situation when it came to the sale price of a home.

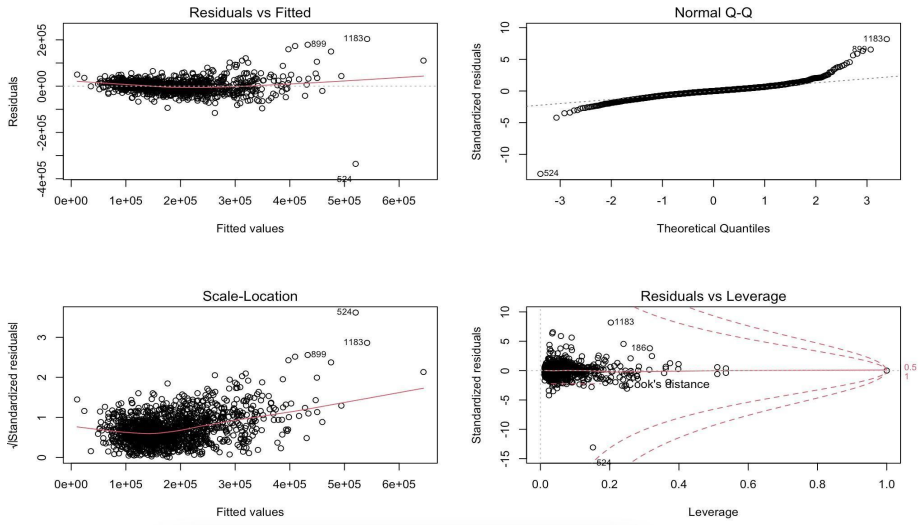
```
1. vars_sub_select_exper_test=c("OverallQual","LotArea","OverallCond","YearB
   uilt","BldgType","YearRemodAdd","RoofStyle","PoolArea","MiscVal","Firepla
   ces","Neighborhood","EnclosedPorch","TotRmsAbvGrd","GrLivArea","LowQualFi
   nSF","ExterQual","MasVnrArea","MasVnrType","Exterior1st","Exterior2nd","L
   otShape","LotConfig", "LandSlope","HouseStyle","Condition1",
   "BsmtCond","BsmtFinType2","ExterCond","HeatingQC","RoofMatl","Foundation"
   ,"LandSlope","BsmtFinType1")
2. data_test=test_set[vars_sub_select_exper_test]
```

```
data_test=na.omit(data_test)
data_test=data_test[!(data_test$Exterior1st=='AsphShn'),]
test_pr=predict(model_final_tr,newdata=data_test)
 head(test_pr)
        1         2         3         4         5         6
 99492.49 159887.89 185212.82 192451.68 220808.66 169451.69
fivenum(test_pr)
  4426.625 127542.799 162134.147 218400.246 610017.138
> mean(test_pr)
  179839.9
```

Shapiro Test

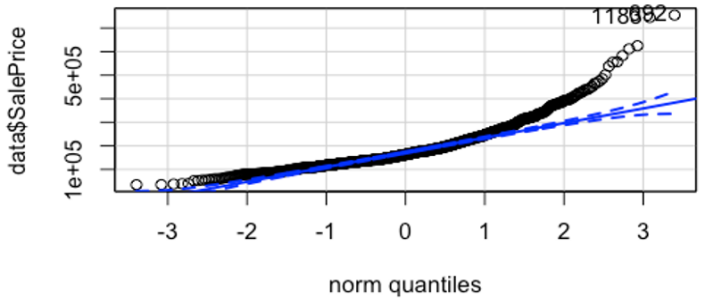
We decided to test the normality of the distribution. We used the residuals of the model we worked with; we used the Shapiro test to check. We ended up getting W being 0.84879 while the p-value being less than 2.2e-16. This means that the data were not distributed normally and rejecting the Ho hypothesis.

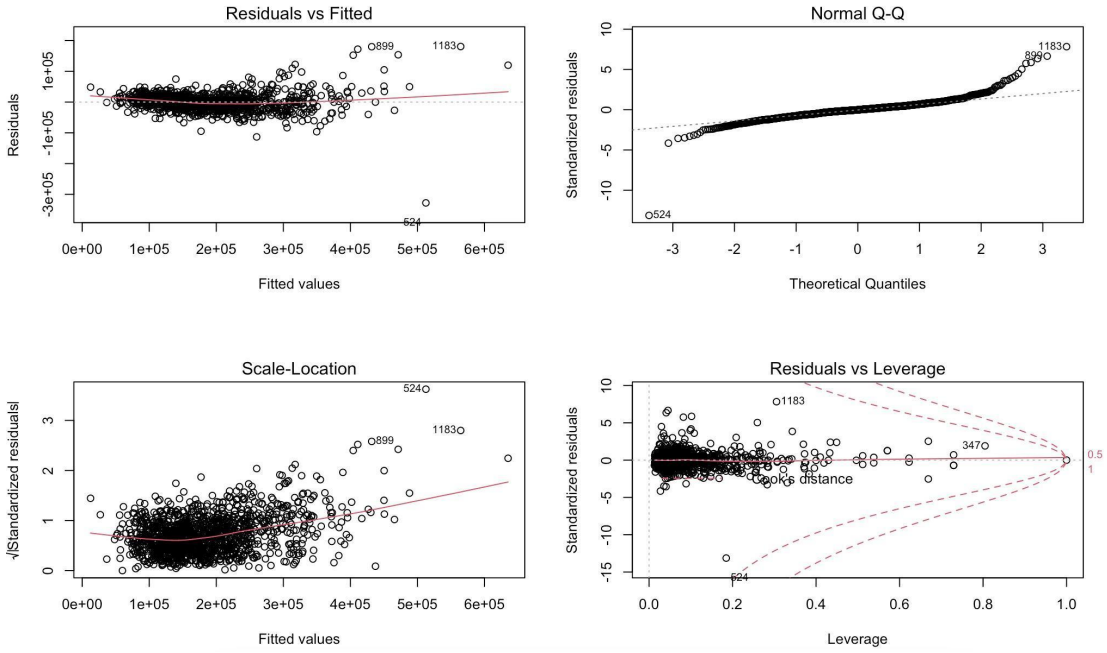
```
err=residuals(model_final_tr)
> shapiro.test(err)
  Shapiro-Wilk normality test
data:  err
W = 0.84879, p-value < 2.2e-16
```



We also ran a Shapiro test on the subset that we picked for the project, in which we noticed the sample set was also distributed normally. By having W=0.84908 and a P-value of less than 2.2e-16, meaning that the data was not indeed distributed normally and rejecting the Ho hypothesis.

```
1. Shapiro-Wilk normality test
2.
3. data:  err
4. W = 0.84908, p-value < 2.2e-16
```





Conclusion

To conclude, we can see that our model's variables have a significant influence on the target variable (Ho for F-test was rejected), and independent variables explain about 88% of the variation in the target variable. The data (dependent variable) is not normally distributed. There is a significant group of outliers, especially to the QQ plot's right side. That can be why we get such high MSE. The initial dataset contains 79 explanatory variables. Using logic and subset selection techniques, we were able to find a linear regression model that uses just 23 predictors

and explains about 88 percent of the variation in the target variable (Adjusted R-squared) with MSE 714564136.

Reference:

Carnegie Mellon University, 1980, lib.stat.cmu.edu/datasets/boston.

De Cook, Dean. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." *Journal of Statistics Education*, 2011, jse.amstat.org/v19n3/decock.pdf.

Dublin, Robin. "Predicting House Prices Using Multiple Listings Data." *ResearchGate*, 1998, www.researchgate.net/publication/5151497_Predicting_House_Prices_Using_Multiple_Listings_Data.

Truong, Quang, et al. "Housing Price Prediction via Improved Machine Learning Techniques." *Procedia Computer Science*, Elsevier, 27 July 2020, www.sciencedirect.com/science/article/pii/S1877050920316318.

"Real Estate, Apartments, Mortgages & Home Values." *Zillow*, www.zillow.com/.

"House Prices: Advanced Regression Techniques." *Kaggle*, www.kaggle.com/c/house-prices-advanced-regression-techniques.

"Normality Test in R." *STHDA*, www.sthda.com/english/wiki/normality-test-in-r.