

Assessing Home Values
By: Vladimir Antasiuk & Nick Zelada
Project for STA6244

Table of Contents

<i>Goal.....</i>	<i>3</i>
<i>Dataset Description and Data Exploration</i>	<i>4</i>
<i>The Implementation of the Project</i>	<i>5</i>
<i>Literature Review and Real-World Applications</i>	<i>5</i>
<i>Discussion of Methods and Instruments.....</i>	<i>6</i>
<i>The Best Subset Selection.....</i>	<i>6</i>
<i>Prediction</i>	<i>11</i>
<i>Shapiro Test.....</i>	<i>12</i>
<i>Conclusion.....</i>	<i>13</i>
<i>Reference:.....</i>	<i>15</i>

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

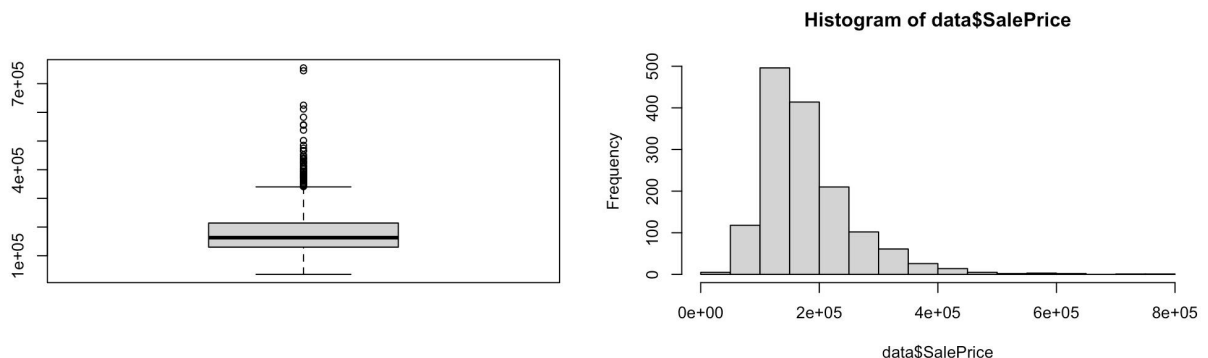
US, and this formula has a significant influence on the pricing at the US real estate market. (Zillow, 2020)

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) with some missing values. 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.

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



Source code used:

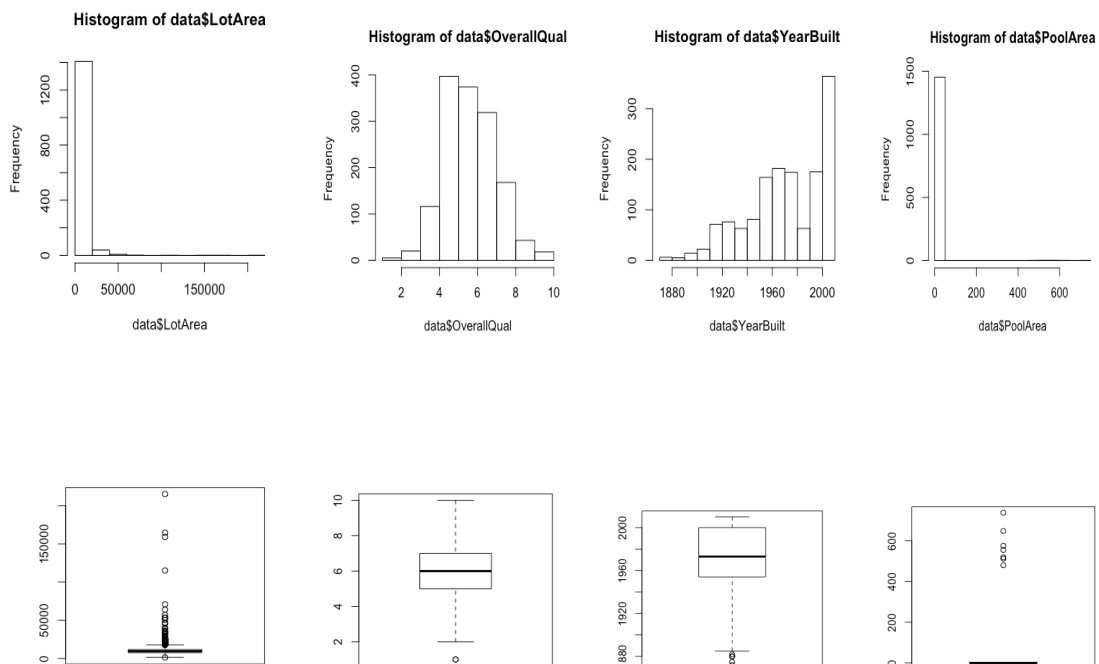
```

1. > fivenum(data$SalePrice)
2.   34900 129950 163000 214000 755000
3. > mean(data$SalePrice)
4. 180921.2
5. > library(moments)
6. > kurtosis(data$SalePrice) = 9.509812
7. > skewness(data$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

Also, below we are doing graphical representation of some of explanatory variables (higher row is histogram and below is corresponding boxplot (LotArea, OverallQual, YearBuilt, PoolArea)).



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 H_0 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.

For the F-test for the whole model the null hypothesis here is: contributing factors of a house do not affect the price of houses (all the coefficients are equal to 0). The alternative hypothesis here is: contributing factors of a house affect the price of houses (at least one coefficient is different from 0). For the coefficients of the model, we used t-test where H_0 is that coefficient is equal to 0, while H_a supposes that the coefficient is different from 0. The significance level chosen is 5%.

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, different formats of the data, missing values 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 from missing value by using `na.omit`.

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

```
1. > model1=lm(SalePrice~., data=na.omit(data_best_subset_sel))
2. > summary(model1)
3. Residual standard error: 31280 on 1345 degrees of freedom
4. Multiple R-squared: 0.8557, Adjusted R-squared: 0.8443
5. 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.

```

1. > data_subset_sel_Clean=na.omit(data_subset_sel)
2. > full.tr = lm(SalePrice~., data=data_subset_sel_Clean) #full model
3. > Intercept.tr=lm(SalePrice~1, data=data_subset_sel_Clean) #the
   smallest model, intercept only
4. > step.tr_both=stepAIC(Intercept.tr,scope=list(upper=full.tr,lower=Inter
   cept.tr),direction="both")
5. > step.tr_backward=stepAIC(full.tr,direction="backward")
6. > step.tr_forward=stepAIC(Intercept.tr,scope=list(upper=full.tr,lower=In
   tercept.tr),direction="forward")
7. > summary(step.tr_both)
8. > 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.”

```

1. > summary(step.tr_both)
2.
3. Call:
4. 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)
5.
6. Residuals:
7. Min      1Q  Median      3Q      Max
8. -331718 -12123      0    10925  203533
9.
10. Coefficients:
11. Estimate      Std. Error    t value Pr(>|t|)
12. (Intercept)    -1.610e+06  1.711e+05  -9.410 < 2e-16 ***
13. OverallQual      9.947e+03  1.124e+03   8.847 < 2e-16 ***
14. GrLivArea       6.907e+01  2.958e+00  23.350 < 2e-16 ***
15. NeighborhoodBlueste -6.449e+03  2.195e+04  -0.294 0.768964
16. NeighborhoodBrDale  6.139e+03  1.159e+04   0.530 0.596455
17. NeighborhoodBrkSide -1.333e+04  9.707e+03  -1.373 0.169989
18. NeighborhoodClearCr -2.078e+04  1.026e+04  -2.027 0.042887 *
19. NeighborhoodCollgCr -1.971e+04  8.043e+03  -2.451 0.014391 *
20. NeighborhoodCrawfor -8.516e+02  9.376e+03  -0.091 0.927644
21. NeighborhoodEdwards -2.756e+04  8.825e+03  -3.124 0.001826 **
22. NeighborhoodGilbert -2.560e+04  8.623e+03  -2.969 0.003039 **
23. NeighborhoodIDOTRR  -2.440e+04  1.021e+04  -2.390 0.017001 *
24. NeighborhoodMeadowV -8.154e+03  1.196e+04  -0.682 0.495359
25. NeighborhoodMitchel -2.605e+04  9.039e+03  -2.881 0.004023 **
26. NeighborhoodNAMES  -2.472e+04  8.563e+03  -2.887 0.003954 **
27. NeighborhoodNoRidge  1.709e+04  9.391e+03   1.820 0.068974 .
28. NeighborhoodNPkVill  1.187e+04  1.283e+04   0.925 0.354951
29. NeighborhoodNridgHt  2.448e+04  8.446e+03  2.898 0.003814 **
30. NeighborhoodNWAmes  -3.000e+04  8.869e+03  -3.382 0.000739 ***
31. NeighborhoodOldTown -2.506e+04  9.319e+03  -2.689 0.007266 **

```

32.	NeighborhoodSawyer	-2.040e+04	8.992e+03	-2.268	0.023469	*
33.	NeighborhoodSawyerW	-1.824e+04	8.760e+03	-2.082	0.037522	*
34.	NeighborhoodSomerst	-1.708e+03	8.364e+03	-0.204	0.838250	
35.	NeighborhoodStoneBr	4.367e+04	9.400e+03	4.645	3.73e-06	***
36.	NeighborhoodSWISU	-2.703e+04	1.065e+04	-2.538	0.011270	*
37.	NeighborhoodTimber	-9.703e+03	9.115e+03	-1.065	0.287272	
38.	NeighborhoodVeenker	5.133e+03	1.183e+04	0.434	0.664471	
39.	RoofMatlCompShg	5.193e+05	3.204e+04	16.207	< 2e-16	***
40.	RoofMatlMembran	5.627e+05	4.549e+04	12.368	< 2e-16	***
41.	RoofMatlMetal	5.712e+05	4.529e+04	12.612	< 2e-16	***
42.	RoofMatlRoll	5.035e+05	4.322e+04	11.650	< 2e-16	***
43.	RoofMatlTar&Grv	5.060e+05	3.308e+04	15.296	< 2e-16	***
44.	RoofMatlWdShake	5.260e+05	3.510e+04	14.983	< 2e-16	***
45.	RoofMatlWdShngl	6.123e+05	3.380e+04	18.115	< 2e-16	***
46.	HouseStyle1.5Unf	1.407e+04	8.155e+03	1.726	0.084649	.
47.	HouseStyle1Story	1.545e+04	3.283e+03	4.706	2.79e-06	***
48.	HouseStyle2.5Fin	-1.496e+04	1.299e+04	-1.151	0.249834	
49.	HouseStyle2.5Unf	-1.223e+04	9.225e+03	-1.326	0.185201	
50.	HouseStyle2Story	-4.587e+03	3.217e+03	-1.426	0.154130	
51.	HouseStyleSFoyer	1.874e+04	6.219e+03	3.013	0.002632	**
52.	HouseStyleSLvl	3.658e+03	4.863e+03	0.752	0.452100	
53.	ExterQualFa	-3.847e+04	1.111e+04	-3.464	0.000550	***
54.	ExterQualGd	-4.389e+04	5.048e+03	-8.694	< 2e-16	***
55.	ExterQualTA	-4.464e+04	5.662e+03	-7.884	6.62e-15	***
56.	BldgType2fmCon	-9.390e+03	5.611e+03	-1.674	0.094428	.
57.	BldgTypeDuplex	-1.743e+04	5.333e+03	-3.269	0.001107	**
58.	BldgTypeTwnhs	-4.409e+04	5.841e+03	-7.549	8.19e-14	***
59.	BldgTypeTwnhsE	-3.168e+04	3.781e+03	-8.379	< 2e-16	***
60.	BsmtFinTypelBLQ	-3.575e+01	3.171e+03	-0.011	0.991005	
61.	BsmtFinTypelGLQ	8.492e+03	2.891e+03	2.938	0.003364	**
62.	BsmtFinTypelLwQ	-8.721e+03	4.168e+03	-2.093	0.036574	*
63.	BsmtFinTypelRec	-3.678e+03	3.406e+03	-1.080	0.280377	
64.	BsmtFinTypelUnf	-1.029e+04	2.762e+03	-3.726	0.000203	***
65.	YearBuilt	4.486e+02	7.270e+01	6.172	8.99e-10	***
66.	OverallCond	5.754e+03	9.032e+02	6.371	2.60e-10	***
67.	LotArea	6.668e-01	1.076e-01	6.195	7.78e-10	***
68.	Fireplaces	5.921e+03	1.516e+03	3.905	9.90e-05	***
69.	PoolArea	1.018e+02	2.041e+01	4.986	6.99e-07	***
70.	MasVnrTypeBrkFace	1.418e+04	7.723e+03	1.836	0.066612	.
71.	MasVnrTypeNone	1.937e+04	7.766e+03	2.494	0.012757	*
72.	MasVnrTypeStone	2.664e+04	8.174e+03	3.259	0.001147	**
73.	MasVnrArea	2.875e+01	6.637e+00	4.332	1.59e-05	***
74.	ConditionlFeedr	8.537e+02	5.589e+03	0.153	0.878621	
75.	ConditionlNorm	9.756e+03	4.601e+03	2.120	0.034161	*
76.	ConditionlPosA	4.224e+03	1.138e+04	0.371	0.710648	
77.	ConditionlPosN	-1.559e+04	8.116e+03	-1.920	0.055016	.
78.	ConditionlRR Ae	-1.831e+04	1.036e+04	-1.768	0.077330	.
79.	ConditionlRRAn	7.562e+03	7.443e+03	1.016	0.309844	
80.	ConditionlRRNe	-5.644e+03	2.084e+04	-0.271	0.786610	
81.	ConditionlRRNn	4.121e+03	1.410e+04	0.292	0.770109	
82.	LotConfigCulDSac	1.023e+04	3.665e+03	2.792	0.005322	**
83.	LotConfigFR2	-5.387e+03	4.640e+03	-1.161	0.245897	
84.	LotConfigFR3	-1.045e+04	1.464e+04	-0.714	0.475419	
85.	LotConfigInside	-1.385e+03	2.030e+03	-0.682	0.495293	
86.	FoundationCBlock	4.886e+02	3.506e+03	0.139	0.889169	
87.	FoundationPConc	6.979e+03	3.920e+03	1.781	0.075193	.
88.	FoundationStone	-5.074e+03	1.201e+04	-0.422	0.672761	


```

89. FoundationWood -2.761e+04 1.690e+04 -1.634 0.102595
90. LandSlopeMod 7.259e+03 3.934e+03 1.845 0.065202 .
91. LandSlopeSev -2.421e+04 1.127e+04 -2.149 0.031852 *
92. BsmtFinType2BLQ -2.312e+04 8.334e+03 -2.775 0.005604 **
93. BsmtFinType2GLQ -8.980e+03 1.074e+04 -0.836 0.403045
94. BsmtFinType2LwQ -2.595e+04 8.104e+03 -3.202 0.001396 **
95. BsmtFinType2Rec -2.436e+04 7.823e+03 -3.114 0.001885 **
96. BsmtFinType2Unf -1.949e+04 6.829e+03 -2.854 0.004384 **
97. LowQualFinSF -3.937e+01 1.981e+01 -1.988 0.047066 *
98. YearRemodAdd 1.134e+02 5.932e+01 1.913 0.056026 .
99. Exterior1stBrkComm -5.299e+04 2.891e+04 -1.833 0.067027 .
100. Exterior1stBrkFace 1.406e+04 8.443e+03 1.665 0.096131 .
101. Exterior1stCBlock 1.894e+03 3.112e+04 0.061 0.951482
102. Exterior1stCemntBd -2.445e+02 8.749e+03 -0.028 0.977712
103. Exterior1stHdBoard -6.963e+03 7.570e+03 -0.920 0.357864
104. Exterior1stImStucc -1.664e+04 2.925e+04 -0.569 0.569483
105. Exterior1stMetalSd 1.886e+03 7.335e+03 0.257 0.797104
106. Exterior1stPlywood -5.108e+03 7.945e+03 -0.643 0.520362
107. Exterior1stStone -2.790e+04 2.306e+04 -1.210 0.226546
108. Exterior1stStucco 7.339e+01 9.418e+03 0.008 0.993783
109. Exterior1stVinylSd -1.659e+03 7.470e+03 -0.222 0.824266
110. Exterior1stWd Sdng 4.635e+02 7.331e+03 0.063 0.949594
111. Exterior1stWdShing -4.206e+03 9.226e+03 -0.456 0.648568
112. ---
113. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
114.
115. Residual standard error: 27730 on 1314 degrees of freedom
116. Multiple R-squared: 0.886, Adjusted R-squared: 0.8775
117. F-statistic: 103.2 on 99 and 1314 DF, p-value: < 2.2e-16

118. err.tr=mean((data_subset_sel_Clean$SalePrice-Y_pr_tr)^2)
119. > err.tr
120. 714564136

```

Regarding the results of the regression, we need to say that for F-test (the H_0 : contributing factors of a house do not affect the price of houses (all the coefficients are equal to 0), the H_a : contributing factors of a house affect the price of houses (at least one coefficient is different from 0, $\alpha=0.05$) the F-statistic is 103.2 on 99 and 1314 DF, p-value < 2.2e-16. Since p-value is less than alpha H_0 is rejected and we can conclude that at least one of the model coefficients is different from 0 and contributing factors (in total) affect the house price, i.e. model works. The proposed model explains about 88.6% of variation in target variable (Multiple R-squared) and about 87.75% after the adjustment for the large number of regressors in the model (Adjusted R-squared). Residual standard error is 27730 on 1314 degrees of freedom. Based on the p-value a lot of coefficients turned out to be statistically significant at different level of significance (p-value < alpha hence H_0 saying that coefficient is equal to 0 is rejected and we can conclude that coefficient is different from 0 at alpha % level of significance). The MSE of the model is 714564136.

We also try to run log-linear regression with the same variables (log-log requires too much time to process the data) and it helped us to improve Multiple R-squared and Adjusted R-squared a little bit.

```

121. Copy_data_subset_sel_Clean$SalePrice=log(Copy_data_subset_sel_Clean$SalePrice)
122. log_model_tr=lm(SalePrice~.,data=Copy_data_subset_sel_Clean)

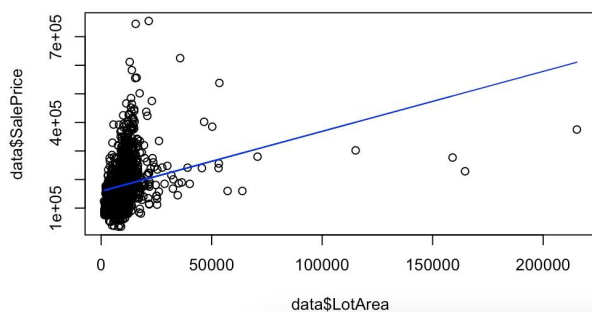
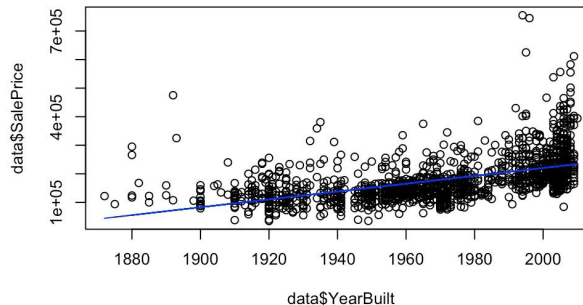
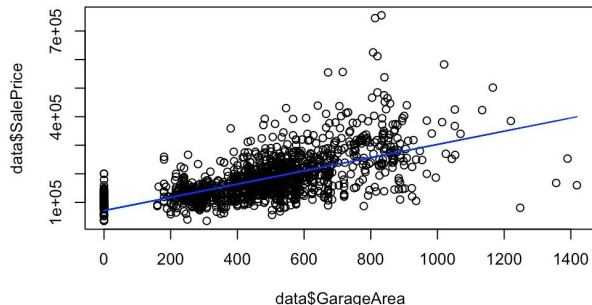
```

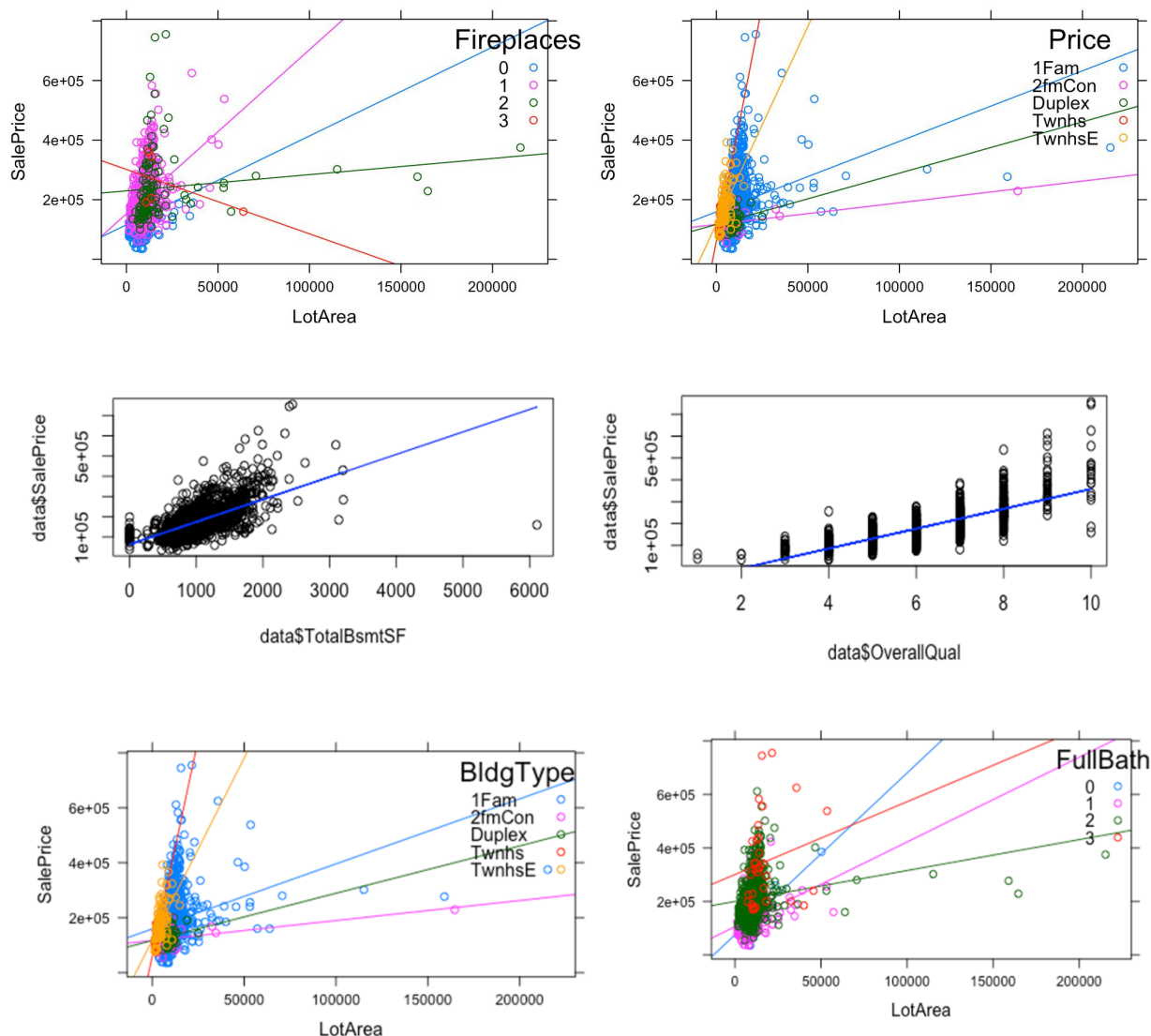
```

123. summary(log_model_tr)
124.
125. Residual standard error: 0.1285 on 1278 degrees of freedom
126. Multiple R-squared:  0.9036,    Adjusted R-squared:  0.8935
127. F-statistic: 88.77 on 135 and 1278 DF,  p-value: < 2.2e-16

```

Below are some graphical representations explaining correlation between dependent variable and independent ones. 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.





So, our final model is:

```
1. 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)
```

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", "YearBuilt", "BldgType", "YearRemodAdd", "RoofStyle", "PoolArea", "MiscVal", "Fireplaces", "Neighborhood", "EnclosedPorch", "TotRmsAbvGrd", "GrLivArea", "LowQualFinSF", "ExterQual", "MasVnrArea", "MasVnrType", "Exterior1st", "Exterior2nd", "LotShape", "LotConfig", "LandSlope", "HouseStyle", "Condition1", "BsmtCond", "BsmtFinType2", "ExterCond", "HeatingQC", "RoofMatl", "Foundation", "LandSlope", "BsmtFinType1")
2. data_test=test_set[vars_sub_select_exper_test]
3. data_test=na.omit(data_test)
4. data_test=data_test[!(data_test$Exterior1st=='AsphShn'),]
5. test_pr=predict(model_final_tr,newdata=data_test)
6. head(test_pr)
7.      1      2      3      4      5      6
8. 99492.49 159887.89 185212.82 192451.68 220808.66 169451.69
9. fivenum(test_pr)
10.  4426.625 127542.799 162134.147 218400.246 610017.138
11. > mean(test_pr)
12. 179839.9

```

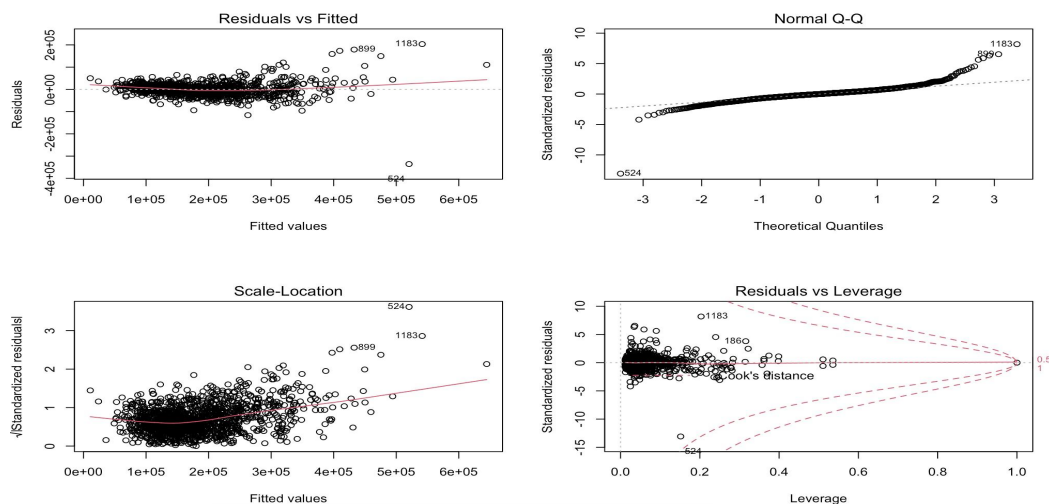
Shapiro Test

To test for the normality of the distribution we used the residuals of the model and run Shapiro Test. $W=0.84879$, p -value being less than $2.2e-16$ ($\alpha=0.05$), H_0 hypothesis (data comes from normal distribution) is rejected, so there is no evidence that data came from normal distribution at 5% level of significance.

```

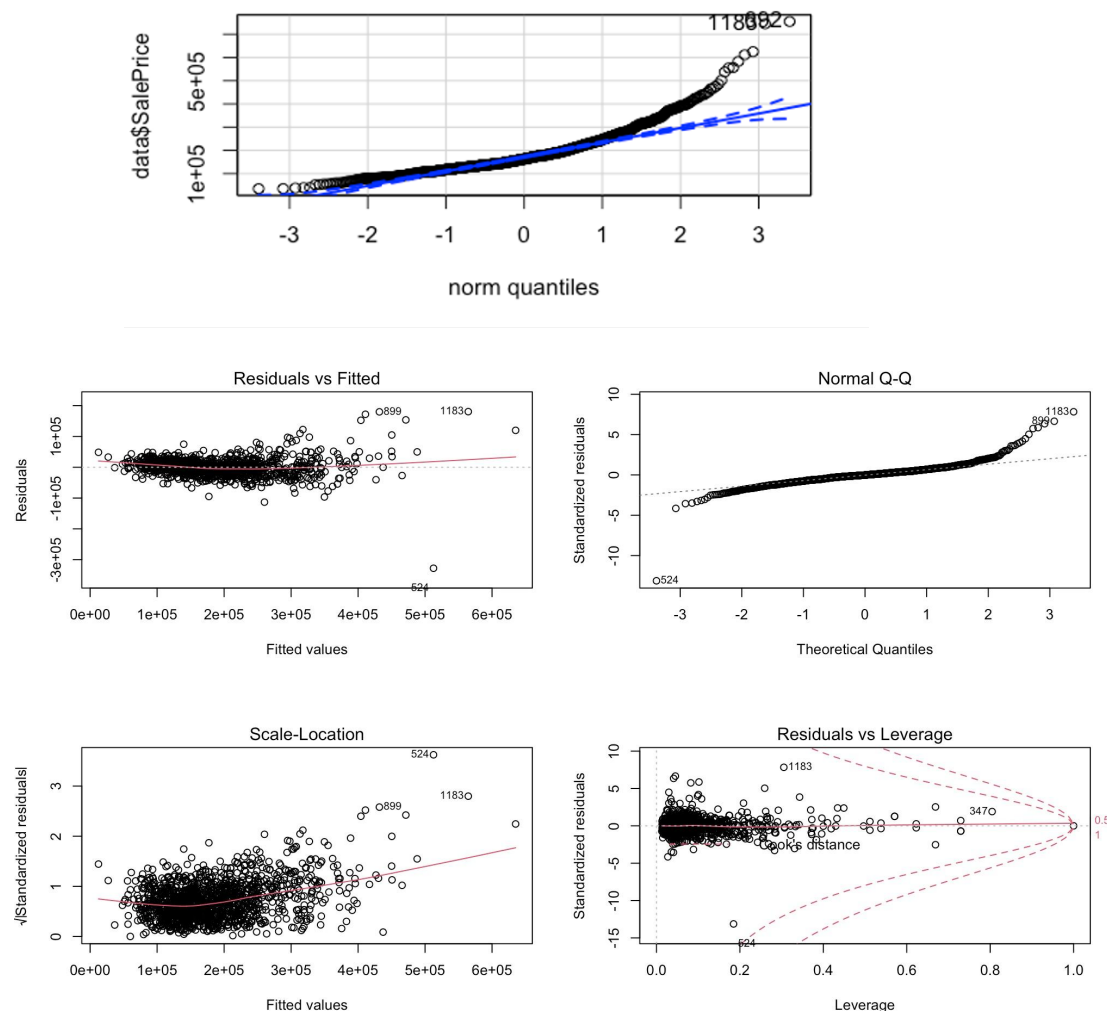
1. err=residuals(model_final_tr)
2. > shapiro.test(err)
3. Shapiro-Wilk normality test
4. data:  err
5. W = 0.84879, p-value < 2.2e-16

```



We also ran a Shapiro test on the subset that we picked for the project, in which we noticed the sample set was also not 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 H_0 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 (H_0 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 after correction for the large amount of regressors in the model (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.