BSAN 450 Take Home Exam

The data in the file named AmesHousing.csv is a data set that contains information from the Ames Assessor’s Office about individual residential properties sold in Ames, Iowa from 2006 to 2010. The purpose of this project is to develop a model for the Sale Price of a home in Ames based on the other variables in the data set using the statistical methods we have covered up to this point in the class. You are to complete this project on your own without anyone’s help. If you have questions about the project you can ask Professor Hillmer but you cannot ask anyone else.

The description of the variables is given at the end of this document.

For this project you need to do the following:

1) Develop a model to predict the variable SalePrice based on the other variables in the data set.

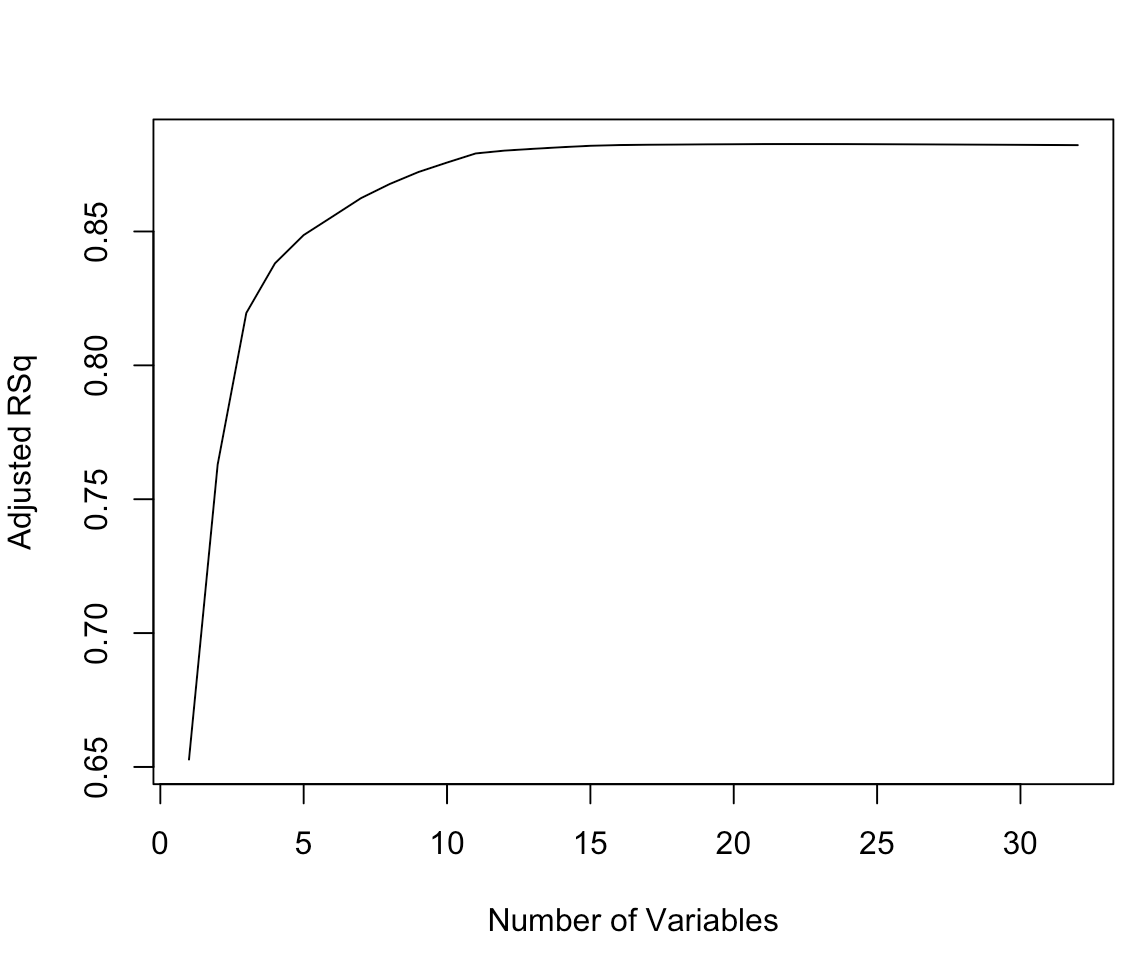
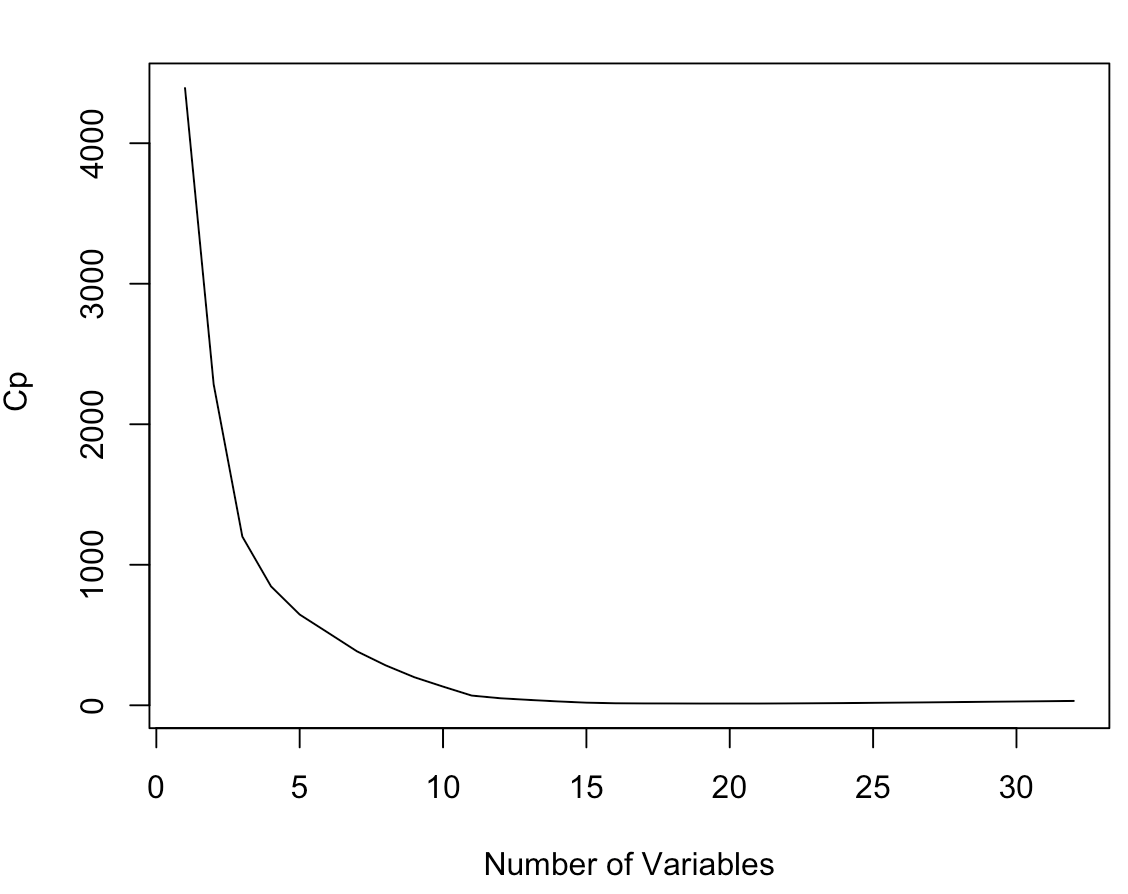
a) For each step of your development, clearly describe what you are doing and your reasons for taking this step. If you do not describe your reasoning you will lose points because I cannot read your mind. I need to know the rationale for what you are doing in developing the model.

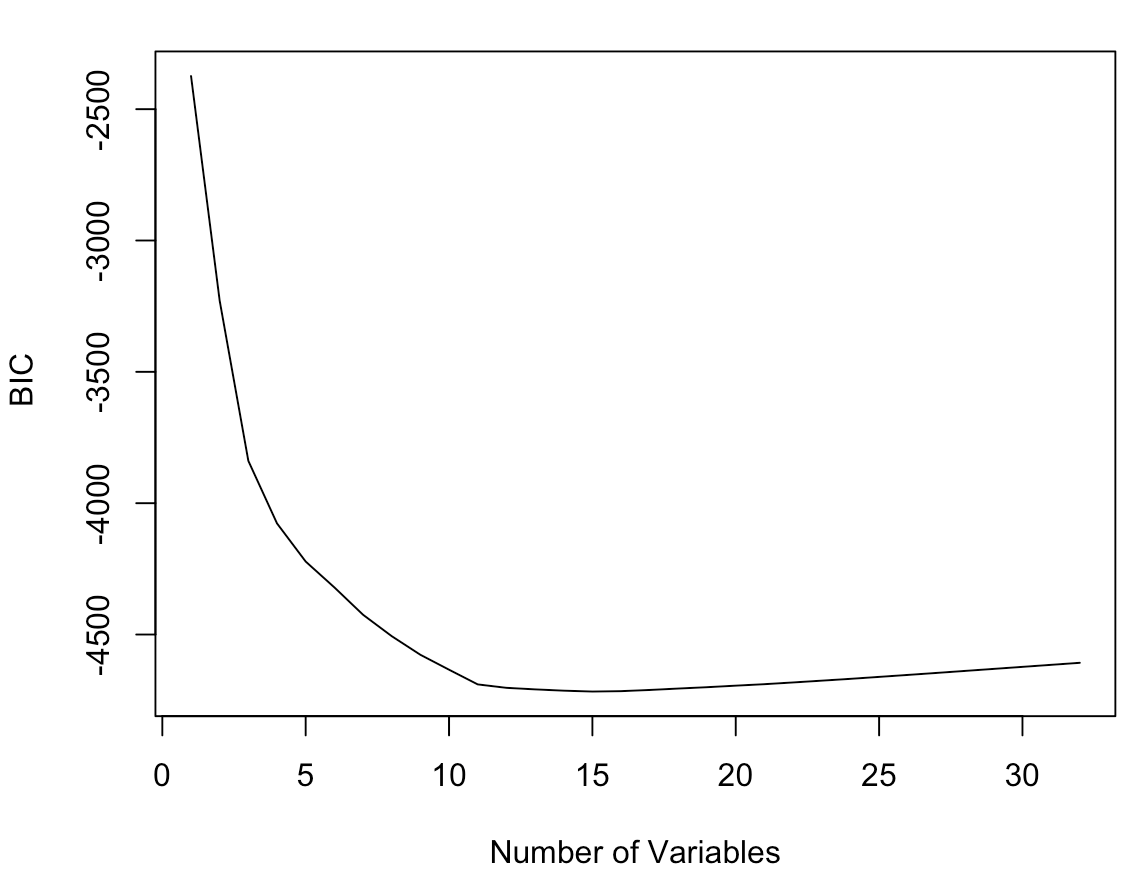
**The first step I took in the process of creating my model was to remove rows with omit data values as this will cause issues when do initial model fitting creation.**

**The next step I took in creating my model was to create a training data set and a testing data set. I made it so that my training data set had 1800 row which is approximately 80% of the data row to allow for an accurate training set, while still have enough data to compare the model using both data sets. This step is important because it will allow for me to test if the model I fit using the training data is a good fit.**

**After setting the data sets up, I started doing analysis to determine how many variables should be included in the model. In order to do this, I first chose to determine how many continuous variables should be included in the model. To determine how many continuous variables to include I did best subsets regression. Below are the R graphs as well as a table of values for the BIC, CP, and ADJR2.**

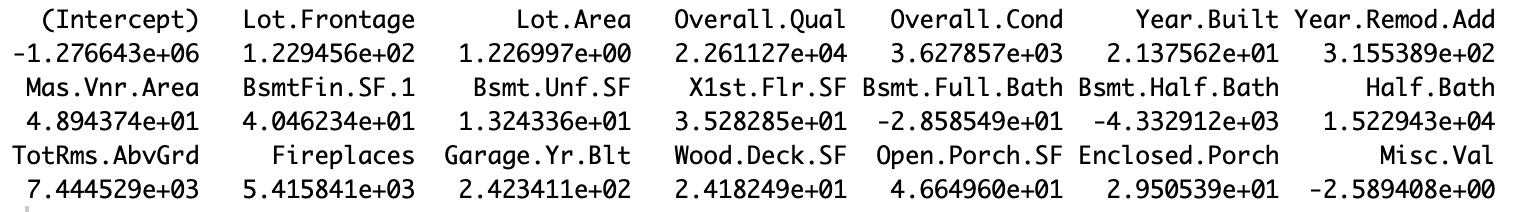
|  |  |  |  |
| --- | --- | --- | --- |
| NUM | BIC | CP | ADJ R^2 |
| 12 | -4702.886 | 50.04934 | 0.8801955 |
| 13 | -4708.785 | 38.32382 | 0.8808630 |
| 14 | -4713.723 | 27.64137 | 0.8814763 |
| 15 | -4717.039 | 18.64150 | 0.8820017 |
| 16 | -4715.548 | 14.45609 | 0.8822744 |
| 17 | -4711.031 | 13.28069 | 0.8823889 |
| 18 | -4705.598 | 13.01532 | 0.8824556 |
| 19 | -4700.482 | 12.43921 | 0.8825387 |
| 20 | -4694.641 | 12.58195 | 0.8825841 |



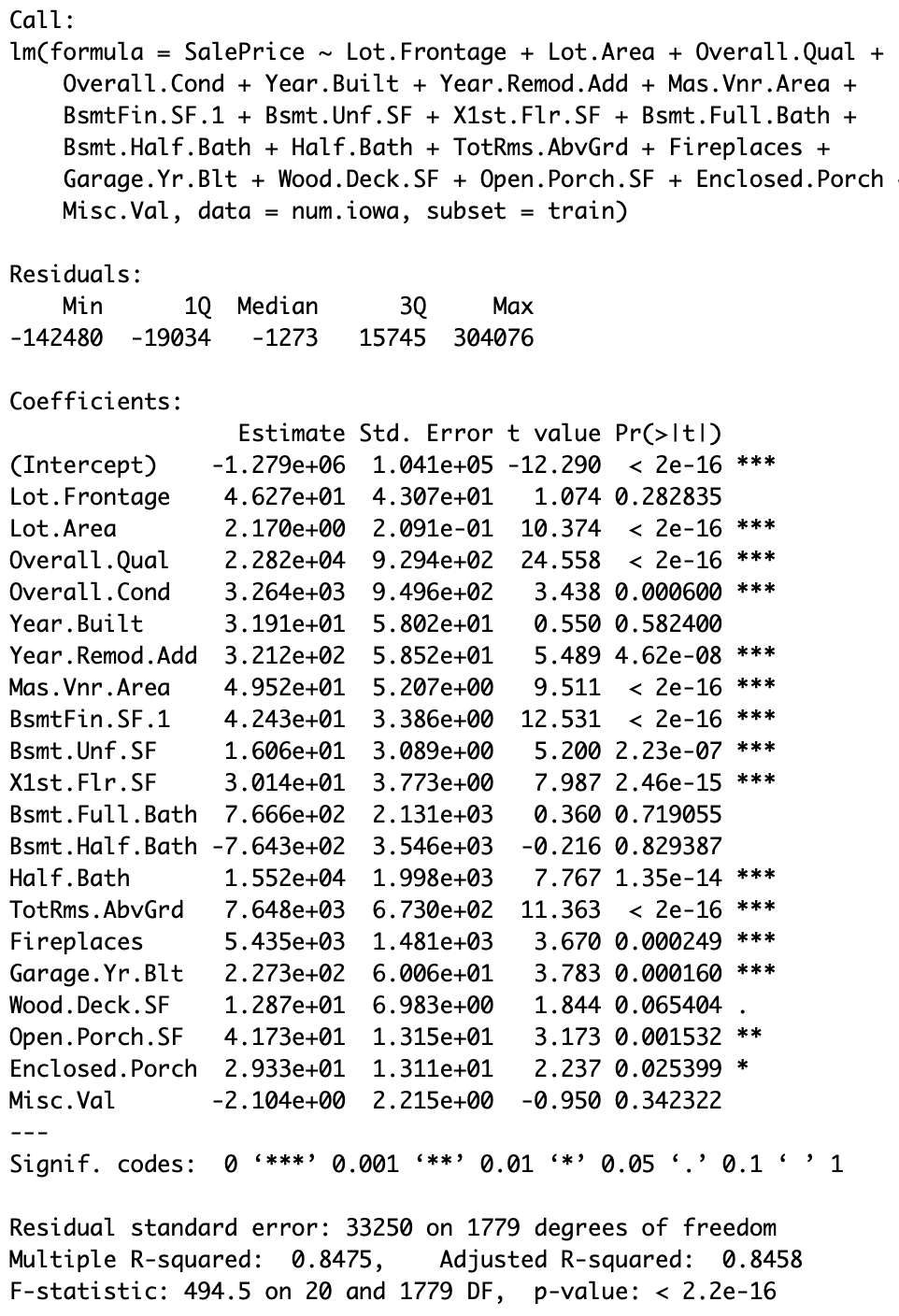
**The interval of acceptable number of variables appears to be between 10 and 25.**

**After running best subsets regression looking at the BIC, Cp, and R^2, it appears that 20 variables would be a good fit, because it is allowing for very small BIC and CP values while still remaining relatively high for the adjusted R^2 value. I also chose 20 to be cautious because I can always run the model and remove variables that are not significant by looking at the p-values.**

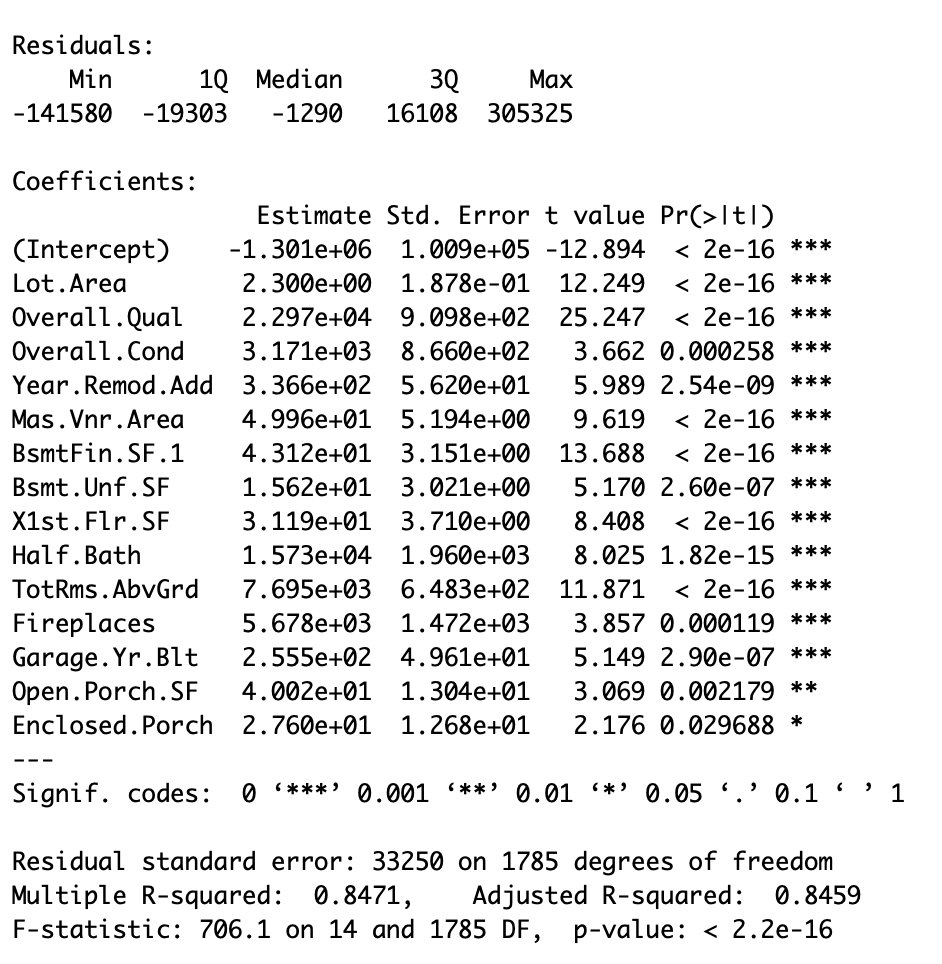
****

**Looking at the coefficients above I can determine which continuous variables should be included in my model.**

**I then proceeded to fit a linear model to the data with the variables I determined above. The summary is below:**

****

**Looking at the model above, there are some variables that are not statistically significant in the model. Thus I will remove Lot.Frontage, Year.Built, Bsmt.Full.Bath, Bsmt.Half.Bath, Wood.Deack.SF, and Misc.Val. Removing these values then brings the number of continuous variable to 14 which is still within an interval that works well based on the best subsets regression done previously. The summary of the new model removing the aforementioned variables is below:**

****

**This model looks much better as all the variables are statistically significant.**

**The next step I took in determining a model was to start adding discrete variables to the model. The method I used to do this was to create two models: one that is the model with only continuous variables, and one that is the model with one new variable added. I then ran the anova command to compare the two model and if the anova command returned a small p-value then I added that variable to the first model, because a small p-value indicates a statistical difference between the two model, indicating that the new variable to model2 is impactful and should be included. If the p-value from the anova command was large, then I did not use that variable or add it to the model. I did this for all of the discrete variable in the data set. If there was a small p-value but it was larger than .05 I still included it in the model, knowing that I would remove any unnecessary variables looking at the summary of the linear model fit.**

**The summary of the fit after adding potential discrete variable is below:**

Residuals:

Min 1Q Median 3Q Max

-142252 -13705 -267 12649 262221

Coefficients: (2 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7.121e+05 1.210e+05 -5.884 4.81e-09 \*\*\*

Lot.Area 1.852e+00 1.867e-01 9.919 < 2e-16 \*\*\*

Overall.Qual 1.164e+04 9.053e+02 12.862 < 2e-16 \*\*\*

Overall.Cond 4.392e+03 7.771e+02 5.651 1.87e-08 \*\*\*

Year.Remod.Add 9.897e+01 5.202e+01 1.903 0.057248 .

Mas.Vnr.Area 5.224e+01 5.670e+00 9.213 < 2e-16 \*\*\*

BsmtFin.SF.1 4.337e+01 3.836e+00 11.305 < 2e-16 \*\*\*

Bsmt.Unf.SF 1.989e+01 3.834e+00 5.189 2.37e-07 \*\*\*

X1st.Flr.SF 5.407e+01 4.442e+00 12.172 < 2e-16 \*\*\*

Half.Bath 7.236e+03 1.812e+03 3.994 6.77e-05 \*\*\*

TotRms.AbvGrd 2.019e+03 7.414e+02 2.723 0.006533 \*\*

Fireplaces 8.008e+03 2.477e+03 3.232 0.001251 \*\*

Garage.Yr.Blt 2.358e+02 4.603e+01 5.122 3.37e-07 \*\*\*

Open.Porch.SF 1.414e+01 1.070e+01 1.321 0.186717

Enclosed.Porch 3.104e+00 1.063e+01 0.292 0.770353

Lot.ShapeIR2 1.318e+03 4.894e+03 0.269 0.787721

Lot.ShapeIR3 -3.815e+03 9.616e+03 -0.397 0.691624

Lot.ShapeReg 3.324e+01 1.625e+03 0.020 0.983684

Land.ContourHLS 1.546e+04 4.730e+03 3.268 0.001104 \*\*

Land.ContourLow -6.629e+03 6.947e+03 -0.954 0.340099

Land.ContourLvl 1.323e+03 3.695e+03 0.358 0.720378

Lot.ConfigCulDSac 1.037e+04 3.747e+03 2.767 0.005713 \*\*

Lot.ConfigFR2 -2.402e+03 4.538e+03 -0.529 0.596698

Lot.ConfigFR3 -1.591e+04 9.719e+03 -1.637 0.101909

Lot.ConfigInside 1.291e+03 1.739e+03 0.743 0.457825

Land.SlopeMod 3.969e+03 3.755e+03 1.057 0.290555

Land.SlopeSev -3.303e+04 1.333e+04 -2.477 0.013330 \*

Condition.1Feedr 1.140e+04 4.631e+03 2.462 0.013925 \*

Condition.1Norm 1.690e+04 3.795e+03 4.452 9.06e-06 \*\*\*

Condition.1PosA 3.300e+04 8.821e+03 3.741 0.000190 \*\*\*

Condition.1PosN 1.408e+04 6.998e+03 2.012 0.044421 \*

Condition.1RRAe -1.084e+03 7.269e+03 -0.149 0.881459

Condition.1RRAn 6.763e+03 5.843e+03 1.157 0.247283

Condition.2Norm 4.381e+03 7.696e+03 0.569 0.569235

Condition.2Pos 1.577e+04 1.442e+04 1.094 0.274161

Bldg.Type2fmCon -1.206e+04 5.283e+03 -2.283 0.022582 \*

Bldg.TypeDuplex -2.199e+04 4.609e+03 -4.770 2.00e-06 \*\*\*

Bldg.TypeTwnhs -1.553e+04 3.984e+03 -3.898 0.000101 \*\*\*

Bldg.TypeTwnhsE -9.745e+03 2.815e+03 -3.462 0.000549 \*\*\*

House.Style1.5Unf -1.189e+04 7.973e+03 -1.492 0.135962

House.Style1Story -1.863e+04 2.703e+03 -6.892 7.73e-12 \*\*\*

House.Style2.5Unf 1.365e+04 7.989e+03 1.708 0.087769 .

House.Style2Story 1.698e+04 2.809e+03 6.044 1.84e-09 \*\*\*

House.StyleSFoyer -1.313e+04 5.272e+03 -2.490 0.012875 \*

House.StyleSLvl -1.397e+04 4.331e+03 -3.226 0.001278 \*\*

Roof.StyleGable 1.514e+04 1.038e+04 1.459 0.144725

Roof.StyleGambrel 1.375e+04 1.267e+04 1.085 0.277873

Roof.StyleHip 1.484e+04 1.043e+04 1.422 0.155122

Roof.StyleOther 2.471e+04 1.423e+04 1.736 0.082688 .

Exterior.1stBrkFace 1.683e+04 1.061e+04 1.586 0.112932

Exterior.1stCemntBd 1.161e+03 1.736e+04 0.067 0.946707

Exterior.1stHdBoard 5.581e+03 1.055e+04 0.529 0.596887

Exterior.1stMetalSd 1.116e+04 1.149e+04 0.971 0.331531

Exterior.1stStucco -8.050e+03 1.140e+04 -0.706 0.480269

Exterior.1stVinylSd -2.024e+03 1.188e+04 -0.170 0.864738

Exterior.1stWood 2.988e+03 1.002e+04 0.298 0.765561

Exterior.2ndBrkFace -2.506e+03 1.084e+04 -0.231 0.817258

Exterior.2ndCmentBd 7.036e+03 1.753e+04 0.401 0.688140

Exterior.2ndHdBoard -5.105e+03 1.057e+04 -0.483 0.629020

Exterior.2ndMetalSd -4.232e+03 1.146e+04 -0.369 0.711948

Exterior.2ndPlywood -2.864e+03 1.014e+04 -0.283 0.777558

Exterior.2ndStucco 1.496e+04 1.092e+04 1.370 0.170860

Exterior.2ndVinylSd 4.253e+03 1.190e+04 0.357 0.720817

Exterior.2ndwood -2.564e+02 1.010e+04 -0.025 0.979749

Mas.Vnr.TypeBrkFace 6.264e+03 7.850e+03 0.798 0.425011

Mas.Vnr.TypeNone 1.856e+04 7.851e+03 2.364 0.018215 \*

Mas.Vnr.TypeStone 1.867e+04 8.100e+03 2.305 0.021313 \*

Exter.QualFa -2.580e+04 9.905e+03 -2.604 0.009283 \*\*

Exter.QualGd -2.259e+04 4.415e+03 -5.117 3.45e-07 \*\*\*

Exter.QualTA -3.065e+04 5.046e+03 -6.074 1.54e-09 \*\*\*

FoundationCBlock 3.329e+01 2.634e+03 0.013 0.989917

FoundationOther -1.273e+04 8.934e+03 -1.425 0.154420

FoundationPConc 2.985e+03 3.064e+03 0.974 0.330140

FoundationSlab 4.100e+03 9.733e+03 0.421 0.673677

Bsmt.QualFa -2.724e+04 5.220e+03 -5.219 2.02e-07 \*\*\*

Bsmt.QualGd -2.392e+04 3.012e+03 -7.941 3.62e-15 \*\*\*

Bsmt.QualNone -3.249e+04 1.040e+04 -3.125 0.001806 \*\*

Bsmt.QualTA -2.792e+04 3.815e+03 -7.318 3.88e-13 \*\*\*

Bsmt.ExposureGd 6.763e+03 2.856e+03 2.368 0.017996 \*

Bsmt.ExposureMn -5.674e+03 2.877e+03 -1.972 0.048723 \*

Bsmt.ExposureNo -4.938e+03 2.106e+03 -2.345 0.019153 \*

Bsmt.ExposureNone NA NA NA NA

BsmtFin.Type.2BLQ -1.069e+04 6.227e+03 -1.716 0.086280 .

BsmtFin.Type.2GLQ 1.264e+04 7.562e+03 1.671 0.094822 .

BsmtFin.Type.2LwQ -1.424e+04 5.983e+03 -2.379 0.017450 \*

BsmtFin.Type.2None NA NA NA NA

BsmtFin.Type.2Rec -1.491e+04 5.775e+03 -2.582 0.009906 \*\*

BsmtFin.Type.2Unf -2.022e+04 5.030e+03 -4.020 6.09e-05 \*\*\*

Heating.QCFa -7.292e+03 4.259e+03 -1.712 0.087016 .

Heating.QCGd -1.032e+03 2.012e+03 -0.513 0.607935

Heating.QCTA -4.945e+03 1.954e+03 -2.531 0.011463 \*

Kitchen.QualFa -1.586e+04 5.967e+03 -2.659 0.007918 \*\*

Kitchen.QualGd -1.593e+04 3.304e+03 -4.822 1.55e-06 \*\*\*

Kitchen.QualTA -1.919e+04 3.750e+03 -5.117 3.46e-07 \*\*\*

FunctionalMaj2 -1.015e+04 1.570e+04 -0.646 0.518266

FunctionalMin1 1.262e+04 9.023e+03 1.399 0.162092

FunctionalMin2 1.256e+04 8.920e+03 1.408 0.159417

FunctionalMod 1.082e+04 9.985e+03 1.083 0.278767

FunctionalTyp 2.601e+04 7.957e+03 3.269 0.001101 \*\*

Fireplace.QuFa -1.466e+04 6.228e+03 -2.354 0.018672 \*

Fireplace.QuGd -1.258e+04 4.881e+03 -2.578 0.010022 \*

Fireplace.QuNone -6.439e+03 5.721e+03 -1.126 0.260526

Fireplace.QuPo -1.315e+04 7.267e+03 -1.810 0.070545 .

Fireplace.QuTA -9.494e+03 5.065e+03 -1.875 0.061024 .

Garage.TypeAttchd 1.426e+04 6.799e+03 2.098 0.036066 \*

Garage.TypeBasment 1.202e+04 9.432e+03 1.275 0.202561

Garage.TypeBuiltIn 3.286e+04 7.343e+03 4.475 8.17e-06 \*\*\*

Garage.TypeCarPort 1.165e+04 1.075e+04 1.083 0.278998

Garage.TypeDetchd 1.715e+04 6.876e+03 2.495 0.012707 \*

Sale.TypeCon 5.143e+03 6.458e+03 0.796 0.425894

Sale.TypeNew 1.711e+04 4.617e+03 3.706 0.000217 \*\*\*

Sale.TypeOth 3.634e+04 1.572e+04 2.312 0.020884 \*

Sale.TypeWD 1.141e+04 9.630e+03 1.184 0.236385

Sale.TypeWD 5.766e+03 3.928e+03 1.468 0.142307

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25940 on 1688 degrees of freedom

Multiple R-squared: 0.912, Adjusted R-squared: 0.9062

F-statistic: 157.6 on 111 and 1688 DF, p-value: < 2.2e-16

**Looking at the summary there was some variable that were not statistically significant (p-val > 0.05). I then removed those variables from the model and reran the fit. The summary of the new fit is below:**

Residuals:

Min 1Q Median 3Q Max

-142455 -13733 -524 12980 274312

Coefficients: (2 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.548e+05 1.117e+05 -5.864 5.41e-09 \*\*\*

Lot.Area 1.749e+00 1.789e-01 9.774 < 2e-16 \*\*\*

Overall.Qual 1.167e+04 8.917e+02 13.088 < 2e-16 \*\*\*

Overall.Cond 4.143e+03 7.516e+02 5.512 4.08e-08 \*\*\*

Year.Remod.Add 8.584e+01 5.060e+01 1.697 0.089956 .

Mas.Vnr.Area 5.113e+01 5.631e+00 9.079 < 2e-16 \*\*\*

BsmtFin.SF.1 4.567e+01 3.759e+00 12.150 < 2e-16 \*\*\*

Bsmt.Unf.SF 2.168e+01 3.769e+00 5.752 1.04e-08 \*\*\*

X1st.Flr.SF 5.434e+01 4.368e+00 12.439 < 2e-16 \*\*\*

Half.Bath 8.146e+03 1.773e+03 4.596 4.63e-06 \*\*\*

TotRms.AbvGrd 1.848e+03 7.380e+02 2.504 0.012389 \*

Fireplaces 8.532e+03 2.453e+03 3.479 0.000516 \*\*\*

Garage.Yr.Blt 2.333e+02 4.329e+01 5.390 8.00e-08 \*\*\*

Open.Porch.SF 1.224e+01 1.062e+01 1.153 0.249055

Enclosed.Porch 9.794e+00 1.040e+01 0.942 0.346272

Land.ContourHLS 1.491e+04 4.675e+03 3.189 0.001455 \*\*

Land.ContourLow -8.802e+03 6.855e+03 -1.284 0.199257

Land.ContourLvl 4.123e+02 3.652e+03 0.113 0.910109

Lot.ConfigCulDSac 1.049e+04 3.597e+03 2.917 0.003581 \*\*

Lot.ConfigFR2 -2.890e+03 4.482e+03 -0.645 0.519149

Lot.ConfigFR3 -1.737e+04 9.654e+03 -1.800 0.072083 .

Lot.ConfigInside 1.247e+03 1.726e+03 0.723 0.469851

Land.SlopeMod 3.553e+03 3.693e+03 0.962 0.336208

Land.SlopeSev -3.792e+04 1.268e+04 -2.990 0.002830 \*\*

Condition.1Feedr 1.091e+04 4.578e+03 2.382 0.017313 \*

Condition.1Norm 1.617e+04 3.753e+03 4.308 1.74e-05 \*\*\*

Condition.1PosA 3.634e+04 8.335e+03 4.360 1.38e-05 \*\*\*

Condition.1PosN 1.506e+04 6.922e+03 2.175 0.029753 \*

Condition.1RRAe -2.317e+03 7.252e+03 -0.319 0.749402

Condition.1RRAn 5.695e+03 5.513e+03 1.033 0.301709

Bldg.Type2fmCon -1.176e+04 5.271e+03 -2.230 0.025862 \*

Bldg.TypeDuplex -2.286e+04 4.544e+03 -5.031 5.38e-07 \*\*\*

Bldg.TypeTwnhs -1.604e+04 3.848e+03 -4.169 3.21e-05 \*\*\*

Bldg.TypeTwnhsE -8.585e+03 2.700e+03 -3.180 0.001501 \*\*

House.Style1.5Unf -1.244e+04 7.966e+03 -1.561 0.118643

House.Style1Story -1.896e+04 2.625e+03 -7.223 7.63e-13 \*\*\*

House.Style2.5Unf 1.544e+04 7.861e+03 1.965 0.049627 \*

House.Style2Story 1.683e+04 2.761e+03 6.095 1.35e-09 \*\*\*

House.StyleSFoyer -1.262e+04 5.246e+03 -2.405 0.016262 \*

House.StyleSLvl -1.365e+04 4.274e+03 -3.194 0.001430 \*\*

Mas.Vnr.TypeBrkFace 7.886e+03 7.832e+03 1.007 0.314131

Mas.Vnr.TypeNone 2.095e+04 7.818e+03 2.680 0.007440 \*\*

Mas.Vnr.TypeStone 2.007e+04 8.079e+03 2.484 0.013082 \*

Exter.QualFa -2.820e+04 9.787e+03 -2.881 0.004014 \*\*

Exter.QualGd -2.408e+04 4.293e+03 -5.610 2.35e-08 \*\*\*

Exter.QualTA -3.284e+04 4.915e+03 -6.681 3.20e-11 \*\*\*

Bsmt.QualFa -2.919e+04 5.133e+03 -5.687 1.52e-08 \*\*\*

Bsmt.QualGd -2.417e+04 2.975e+03 -8.124 8.55e-16 \*\*\*

Bsmt.QualNone -2.740e+04 8.189e+03 -3.346 0.000838 \*\*\*

Bsmt.QualTA -2.828e+04 3.728e+03 -7.584 5.46e-14 \*\*\*

Bsmt.ExposureGd 6.679e+03 2.848e+03 2.345 0.019133 \*

Bsmt.ExposureMn -4.769e+03 2.862e+03 -1.666 0.095838 .

Bsmt.ExposureNo -4.326e+03 2.094e+03 -2.066 0.039015 \*

Bsmt.ExposureNone NA NA NA NA

BsmtFin.Type.2BLQ -9.771e+03 6.204e+03 -1.575 0.115462

BsmtFin.Type.2GLQ 1.692e+04 7.501e+03 2.255 0.024235 \*

BsmtFin.Type.2LwQ -1.430e+04 5.968e+03 -2.395 0.016705 \*

BsmtFin.Type.2None NA NA NA NA

BsmtFin.Type.2Rec -1.363e+04 5.753e+03 -2.369 0.017950 \*

BsmtFin.Type.2Unf -1.952e+04 5.009e+03 -3.897 0.000101 \*\*\*

Heating.QCFa -6.655e+03 4.200e+03 -1.585 0.113223

Heating.QCGd -1.820e+03 1.980e+03 -0.920 0.357906

Heating.QCTA -5.811e+03 1.870e+03 -3.108 0.001915 \*\*

Kitchen.QualFa -1.639e+04 5.890e+03 -2.782 0.005459 \*\*

Kitchen.QualGd -1.602e+04 3.282e+03 -4.880 1.16e-06 \*\*\*

Kitchen.QualTA -1.980e+04 3.727e+03 -5.312 1.23e-07 \*\*\*

FunctionalMaj2 -8.384e+03 1.555e+04 -0.539 0.589913

FunctionalMin1 1.146e+04 8.977e+03 1.277 0.201819

FunctionalMin2 1.051e+04 8.833e+03 1.190 0.234361

FunctionalMod 1.025e+04 9.901e+03 1.035 0.300867

FunctionalTyp 2.505e+04 7.910e+03 3.167 0.001570 \*\*

Fireplace.QuFa -1.555e+04 6.194e+03 -2.510 0.012171 \*

Fireplace.QuGd -1.205e+04 4.857e+03 -2.480 0.013223 \*

Fireplace.QuNone -5.674e+03 5.682e+03 -0.999 0.318138

Fireplace.QuPo -1.505e+04 7.252e+03 -2.075 0.038156 \*

Fireplace.QuTA -9.940e+03 5.043e+03 -1.971 0.048889 \*

Garage.TypeAttchd 1.346e+04 6.743e+03 1.995 0.046152 \*

Garage.TypeBasment 1.178e+04 9.305e+03 1.266 0.205691

Garage.TypeBuiltIn 3.290e+04 7.272e+03 4.524 6.47e-06 \*\*\*

Garage.TypeCarPort 1.462e+04 1.056e+04 1.385 0.166331

Garage.TypeDetchd 1.688e+04 6.815e+03 2.476 0.013367 \*

Sale.TypeCon 6.040e+03 6.427e+03 0.940 0.347451

Sale.TypeNew 1.811e+04 4.606e+03 3.931 8.81e-05 \*\*\*

Sale.TypeOth 3.927e+04 1.574e+04 2.495 0.012683 \*

Sale.TypeWD 1.259e+04 9.611e+03 1.310 0.190262

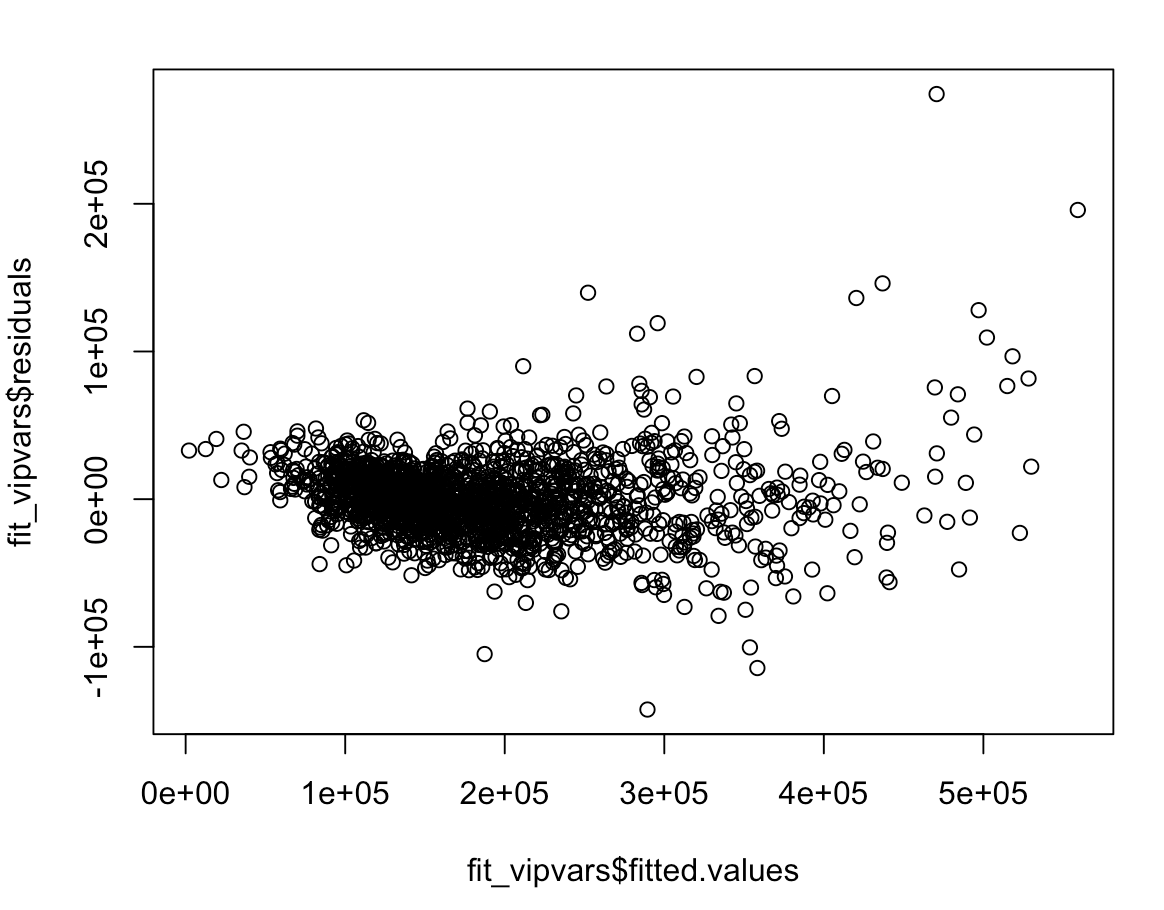
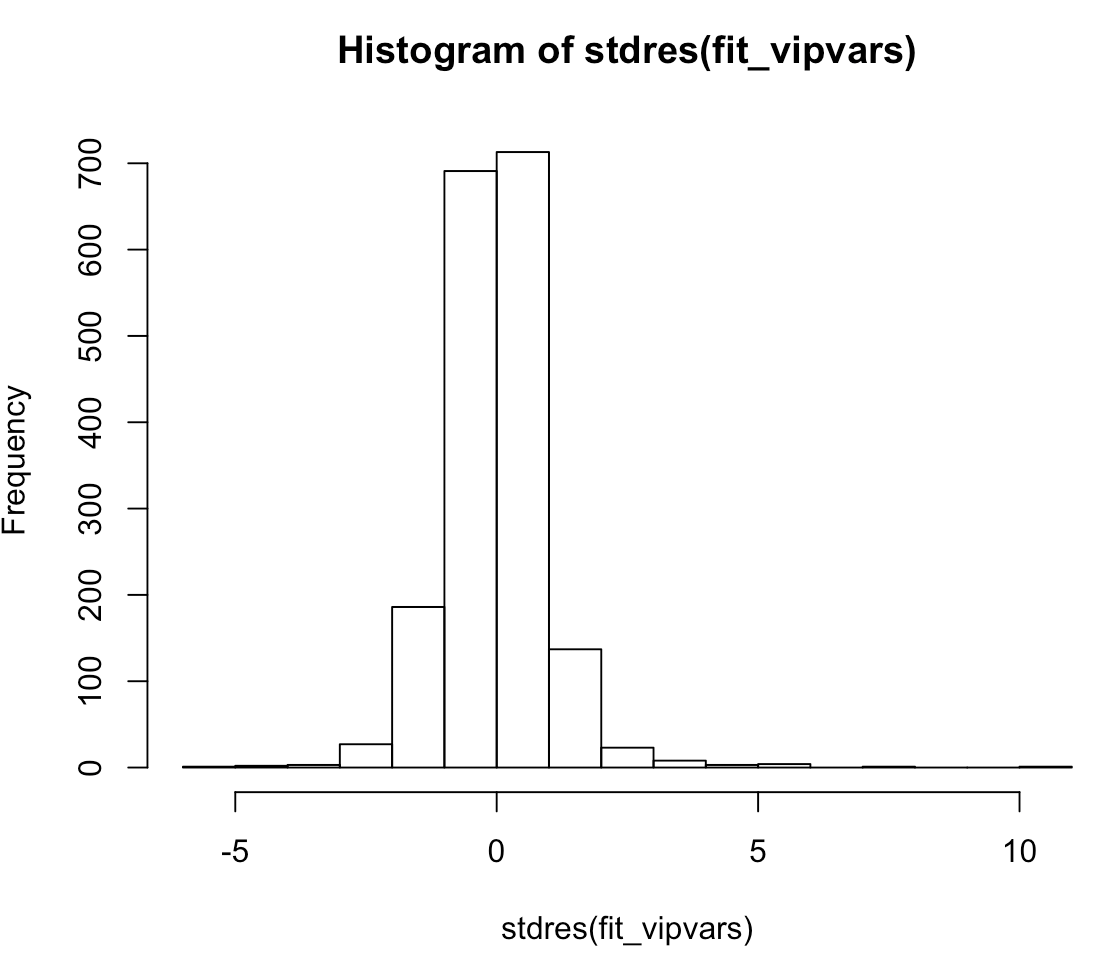
Sale.TypeWD 6.439e+03 3.923e+03 1.641 0.100921

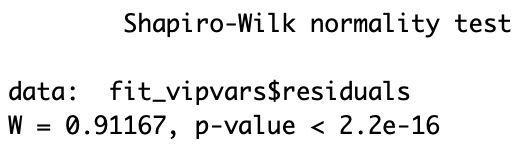
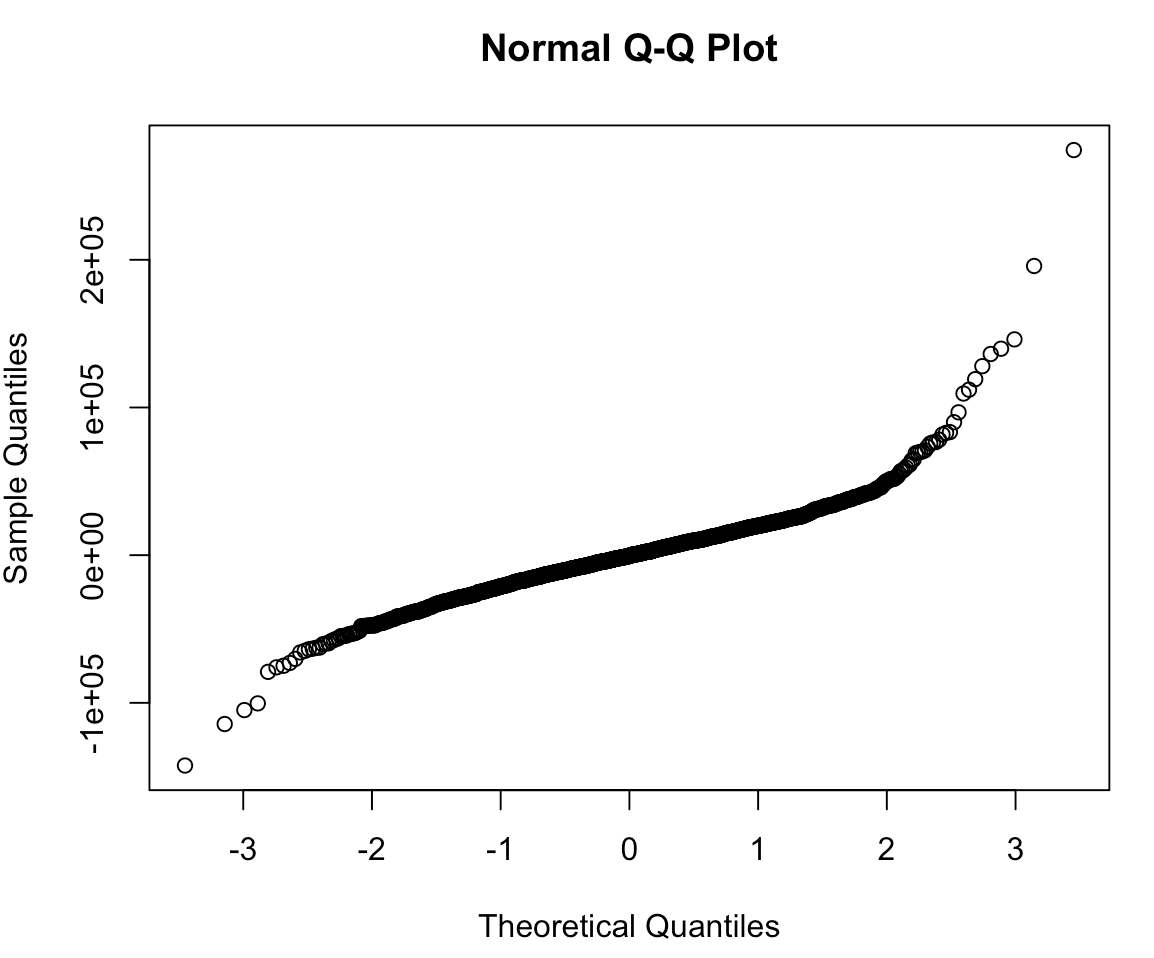
Residual standard error: 26050 on 1716 degrees of freedom

Multiple R-squared: 0.9098, Adjusted R-squared: 0.9054

F-statistic: 208.5 on 83 and 1716 DF, p-value: < 2.2e-16

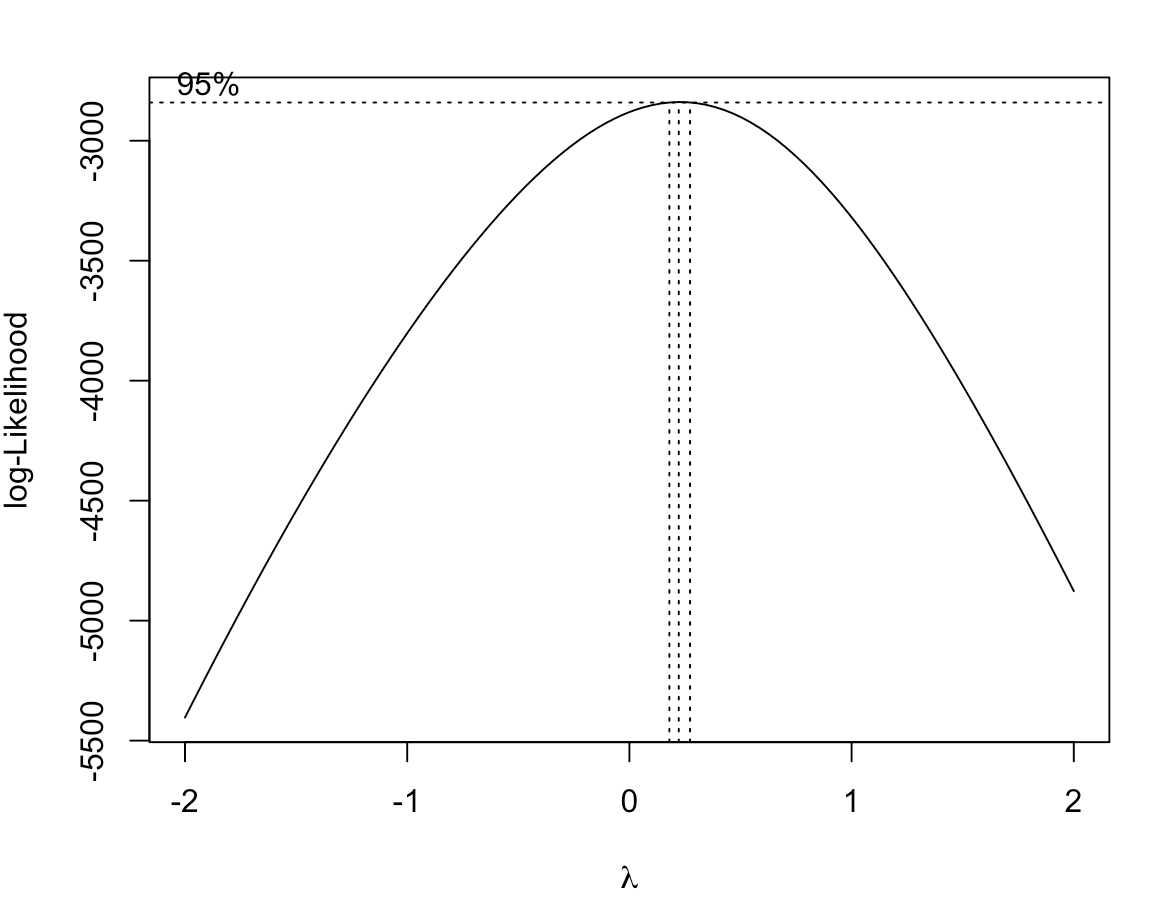
**I did not remove Open.Porch and Enclosed.Porch from the model since they were significant variables in previous iterations of my model and I wanted to look at diagnostics before removing any more variables from my model. Below are the diagnostics from the model above:**

**** ****

****

**The diagnostics suggest an issue. In the graph of the residuals versus fitted values appears to have an increasing variance. The issue indicated to me that a transformation was needed for the y-variable. In order to find the transformation that works best I did a boxcox. The diagnostics also suggest that there are outliers in the data which I determined from the histogram of the standard residual. The shapiro-wilk test also indicates that the model is not following a normal distribution. I will first apply the transformation.**

**The boxcox of the previous model is given below:**

****

**The boxcox indicates that lambda = 0 would work best. Lambda = 0 correlates to doing a log transformation of the y variable (SalePrice).**

**The summary of the new model with the log transformation of the y variable is below:**

Residuals:

Min 1Q Median 3Q Max

-0.80099 -0.06288 0.00141 0.06856 0.41444

Coefficients: (2 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.414e+00 5.128e-01 10.559 < 2e-16 \*\*\*

Lot.Area 7.669e-06 8.216e-07 9.334 < 2e-16 \*\*\*

Overall.Qual 7.373e-02 4.095e-03 18.006 < 2e-16 \*\*\*

Overall.Cond 3.769e-02 3.451e-03 10.921 < 2e-16 \*\*\*

Year.Remod.Add 7.911e-04 2.323e-04 3.405 0.000677 \*\*\*

Mas.Vnr.Area 2.872e-05 2.586e-05 1.111 0.266820

BsmtFin.SF.1 1.832e-04 1.726e-05 10.614 < 2e-16 \*\*\*

Bsmt.Unf.SF 7.498e-05 1.731e-05 4.332 1.56e-05 \*\*\*

X1st.Flr.SF 2.661e-04 2.006e-05 13.265 < 2e-16 \*\*\*

Half.Bath 4.129e-02 8.140e-03 5.073 4.35e-07 \*\*\*

TotRms.AbvGrd 1.530e-02 3.389e-03 4.516 6.73e-06 \*\*\*

Fireplaces 2.180e-02 1.126e-02 1.935 0.053112 .

Garage.Yr.Blt 1.841e-03 1.988e-04 9.262 < 2e-16 \*\*\*

Open.Porch.SF 1.502e-04 4.875e-05 3.081 0.002093 \*\*

Enclosed.Porch 1.110e-04 4.774e-05 2.325 0.020174 \*

Land.ContourHLS 6.648e-02 2.147e-02 3.097 0.001985 \*\*

Land.ContourLow -2.157e-02 3.148e-02 -0.685 0.493335

Land.ContourLvl 1.615e-02 1.677e-02 0.963 0.335545

Lot.ConfigCulDSac 1.790e-02 1.652e-02 1.084 0.278592

Lot.ConfigFR2 -2.910e-03 2.058e-02 -0.141 0.887563

Lot.ConfigFR3 -6.677e-02 4.433e-02 -1.506 0.132216

Lot.ConfigInside 3.161e-03 7.924e-03 0.399 0.690022

Land.SlopeMod 2.657e-02 1.696e-02 1.567 0.117343

Land.SlopeSev -1.468e-01 5.823e-02 -2.521 0.011778 \*

Condition.1Feedr 4.863e-02 2.102e-02 2.313 0.020849 \*

Condition.1Norm 8.708e-02 1.723e-02 5.053 4.82e-07 \*\*\*

Condition.1PosA 1.463e-01 3.828e-02 3.822 0.000137 \*\*\*

Condition.1PosN 8.419e-02 3.178e-02 2.649 0.008156 \*\*

Condition.1RRAe -7.930e-03 3.330e-02 -0.238 0.811799

Condition.1RRAn 4.412e-02 2.531e-02 1.743 0.081572 .

Bldg.Type2fmCon -2.106e-02 2.421e-02 -0.870 0.384327

Bldg.TypeDuplex -4.113e-02 2.087e-02 -1.971 0.048868 \*

Bldg.TypeTwnhs -9.368e-02 1.767e-02 -5.302 1.29e-07 \*\*\*

Bldg.TypeTwnhsE -3.060e-02 1.240e-02 -2.468 0.013665 \*

House.Style1.5Unf -1.005e-01 3.658e-02 -2.747 0.006076 \*\*

House.Style1Story -9.373e-02 1.205e-02 -7.777 1.28e-14 \*\*\*

House.Style2.5Unf 4.414e-02 3.610e-02 1.223 0.221605

House.Style2Story 5.448e-02 1.268e-02 4.296 1.83e-05 \*\*\*

House.StyleSFoyer -5.808e-02 2.409e-02 -2.411 0.016012 \*

House.StyleSLvl -4.478e-02 1.963e-02 -2.282 0.022632 \*

Mas.Vnr.TypeBrkFace 3.865e-02 3.596e-02 1.075 0.282667

Mas.Vnr.TypeNone 4.405e-02 3.590e-02 1.227 0.219955

Mas.Vnr.TypeStone 6.712e-02 3.710e-02 1.809 0.070603 .

Exter.QualFa -9.357e-02 4.494e-02 -2.082 0.037500 \*

Exter.QualGd -9.075e-03 1.971e-02 -0.460 0.645284

Exter.QualTA -3.340e-02 2.257e-02 -1.480 0.139101

Bsmt.QualFa -1.168e-01 2.357e-02 -4.956 7.92e-07 \*\*\*

Bsmt.QualGd -3.717e-02 1.366e-02 -2.721 0.006583 \*\*

Bsmt.QualNone -1.395e-01 3.761e-02 -3.711 0.000213 \*\*\*

Bsmt.QualTA -7.644e-02 1.712e-02 -4.464 8.55e-06 \*\*\*

Bsmt.ExposureGd 1.202e-02 1.308e-02 0.919 0.358301

Bsmt.ExposureMn -4.767e-03 1.314e-02 -0.363 0.716863

Bsmt.ExposureNo 2.387e-03 9.616e-03 0.248 0.803990

Bsmt.ExposureNone NA NA NA NA

BsmtFin.Type.2BLQ -4.950e-02 2.849e-02 -1.737 0.082495 .

BsmtFin.Type.2GLQ 7.527e-02 3.444e-02 2.185 0.028991 \*

BsmtFin.Type.2LwQ -4.969e-02 2.740e-02 -1.813 0.069973 .

BsmtFin.Type.2None NA NA NA NA

BsmtFin.Type.2Rec -4.619e-02 2.642e-02 -1.748 0.080580 .

BsmtFin.Type.2Unf -8.902e-02 2.300e-02 -3.870 0.000113 \*\*\*

Heating.QCFa -6.645e-02 1.928e-02 -3.446 0.000583 \*\*\*

Heating.QCGd -1.309e-02 9.090e-03 -1.440 0.150050

Heating.QCTA -5.028e-02 8.585e-03 -5.857 5.64e-09 \*\*\*

Kitchen.QualFa -7.320e-02 2.705e-02 -2.706 0.006867 \*\*

Kitchen.QualGd -1.769e-02 1.507e-02 -1.174 0.240524

Kitchen.QualTA -4.102e-02 1.711e-02 -2.397 0.016646 \*

FunctionalMaj2 -3.387e-02 7.142e-02 -0.474 0.635432

FunctionalMin1 7.873e-02 4.122e-02 1.910 0.056309 .

FunctionalMin2 7.666e-02 4.056e-02 1.890 0.058937 .

FunctionalMod 4.223e-02 4.547e-02 0.929 0.353120

FunctionalTyp 1.281e-01 3.632e-02 3.526 0.000432 \*\*\*

Fireplace.QuFa -1.421e-02 2.844e-02 -0.500 0.617344

Fireplace.QuGd 2.763e-03 2.230e-02 0.124 0.901431

Fireplace.QuNone -2.691e-02 2.609e-02 -1.031 0.302563

Fireplace.QuPo -3.294e-02 3.330e-02 -0.989 0.322725

Fireplace.QuTA -2.601e-03 2.316e-02 -0.112 0.910585

Garage.TypeAttchd 4.724e-02 3.096e-02 1.525 0.127333

Garage.TypeBasment -1.332e-02 4.273e-02 -0.312 0.755373

Garage.TypeBuiltIn 9.971e-02 3.339e-02 2.986 0.002866 \*\*

Garage.TypeCarPort -4.346e-02 4.849e-02 -0.896 0.370204

Garage.TypeDetchd 2.162e-02 3.129e-02 0.691 0.489716

Sale.TypeCon 8.639e-03 2.951e-02 0.293 0.769778

Sale.TypeNew 9.158e-02 2.115e-02 4.329 1.58e-05 \*\*\*

Sale.TypeOth 2.040e-01 7.227e-02 2.823 0.004814 \*\*

Sale.TypeWD 6.195e-02 4.414e-02 1.404 0.160625

Sale.TypeWD 5.284e-02 1.802e-02 2.933 0.003403 \*\*

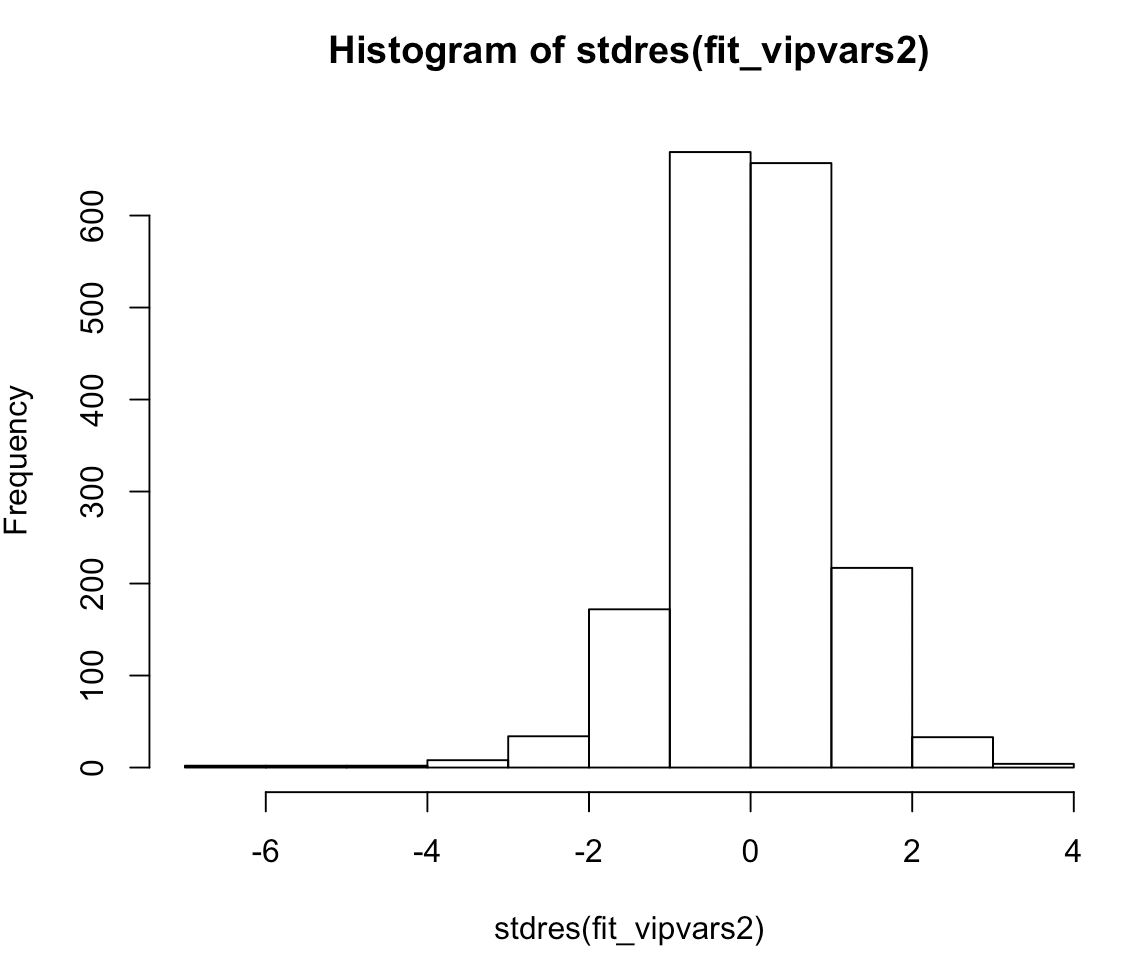
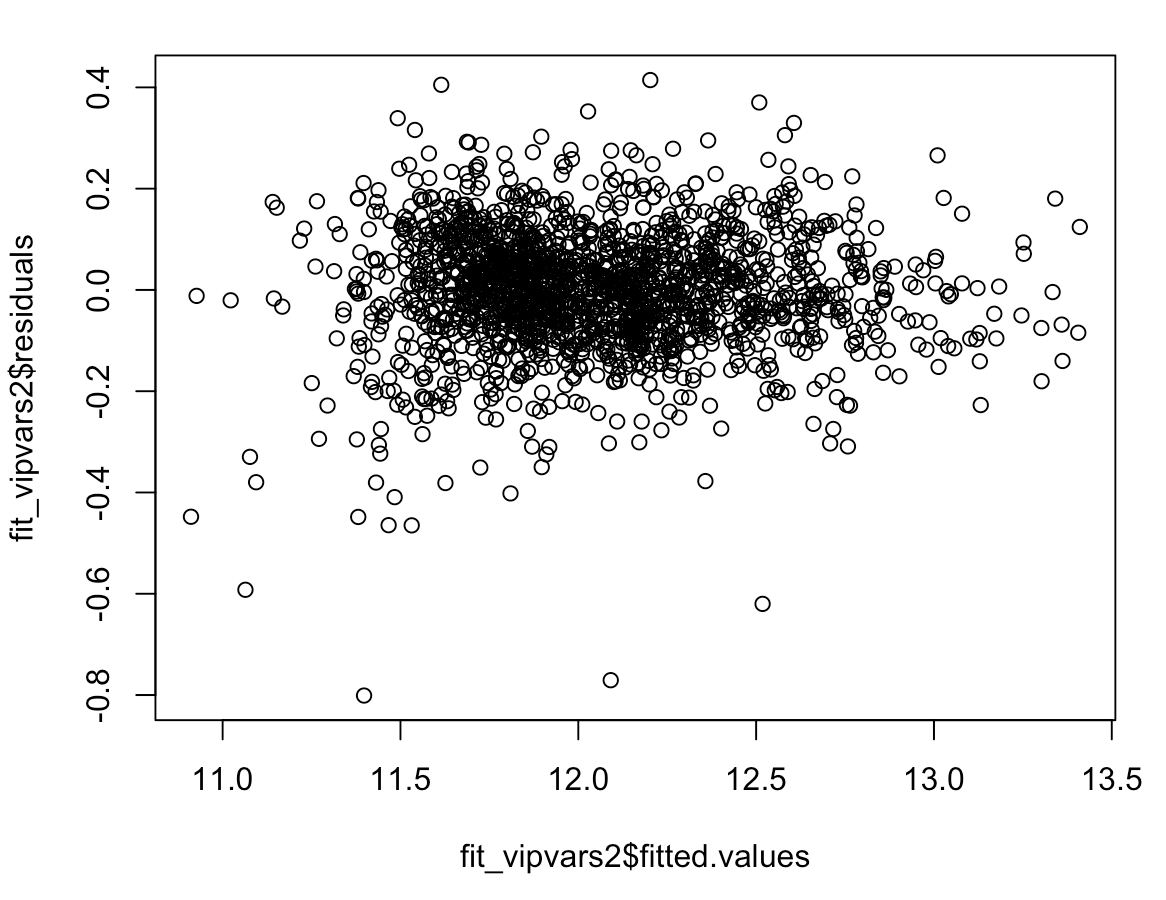
---

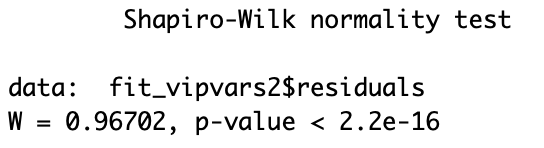
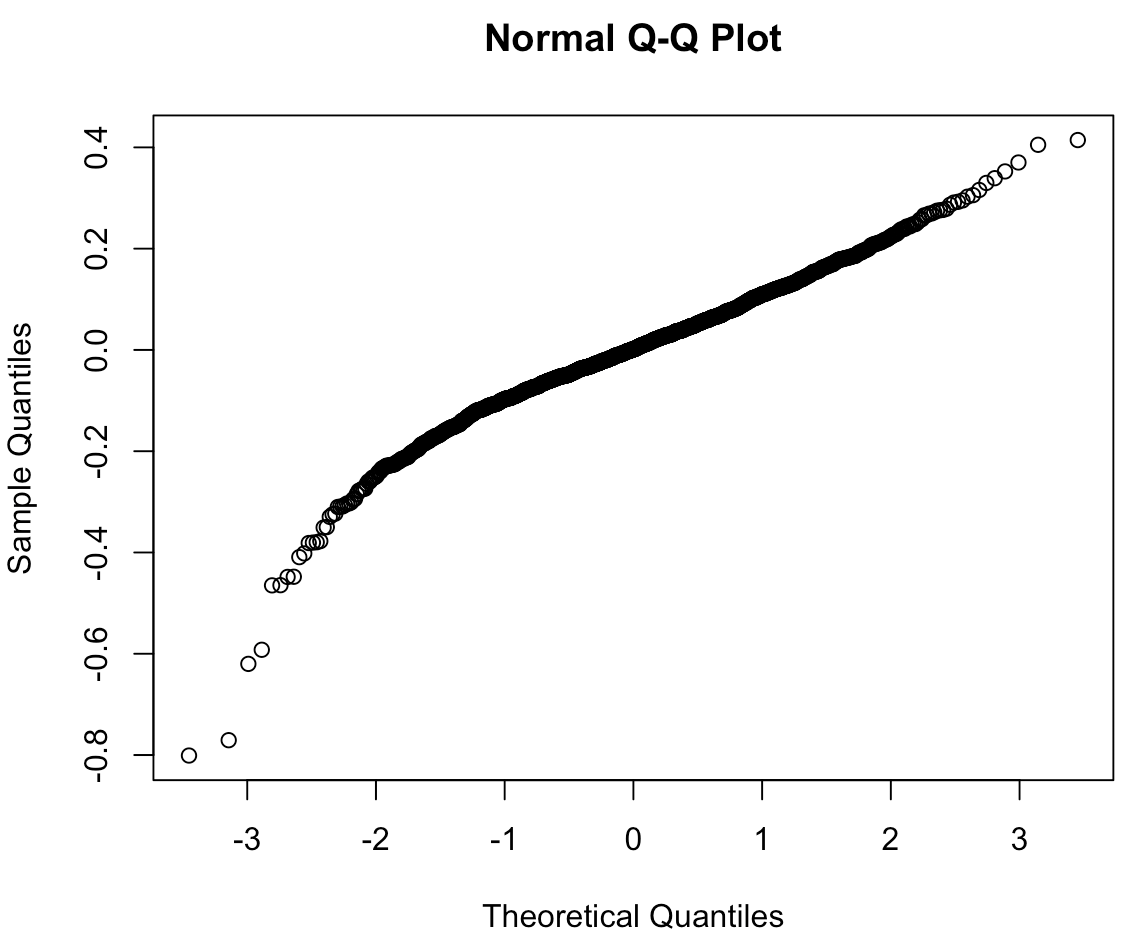
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1196 on 1716 degrees of freedom

Multiple R-squared: 0.9163, Adjusted R-squared: 0.9123

F-statistic: 226.5 on 83 and 1716 DF, p-value: < 2.2e-16

****

****

**The new diagnostics look better. The variance in the residual vs fitted graph looks constant. The histogram is still suggesting outliers and the QQ Norm plot is looking more linear. The shaprio-wilk test is still indicating that the data is not a normal distribution.**

**I noticed in the model that there is some variable that have large p-values and thus are not statistically significant. I removed those variables and the new model is below with new diagnostics. The resulting model has 13 continuous variables and 14 discrete variables.**

lm(formula = logPrice ~ Lot.Area + Overall.Qual + Overall.Cond +

Year.Remod.Add + BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF +

Half.Bath + TotRms.AbvGrd + Fireplaces + Garage.Yr.Blt +

Open.Porch.SF + Enclosed.Porch + Land.Contour + Land.Slope +

Condition.1 + Bldg.Type + House.Style + Mas.Vnr.Type + Exter.Qual +

Bsmt.Qual + BsmtFin.Type.2 + Heating.QC + Kitchen.Qual +

Functional + Garage.Type + Sale.Type, data = iowa, subset = train)

Residuals:

Min 1Q Median 3Q Max

-0.79694 -0.06453 0.00217 0.07046 0.40782

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.451e+00 5.054e-01 10.786 < 2e-16 \*\*\*

Lot.Area 7.918e-06 7.981e-07 9.920 < 2e-16 \*\*\*

Overall.Qual 7.467e-02 4.049e-03 18.439 < 2e-16 \*\*\*

Overall.Cond 3.748e-02 3.444e-03 10.883 < 2e-16 \*\*\*

Year.Remod.Add 7.788e-04 2.317e-04 3.361 0.000794 \*\*\*

BsmtFin.SF.1 1.833e-04 1.699e-05 10.790 < 2e-16 \*\*\*

Bsmt.Unf.SF 7.447e-05 1.720e-05 4.330 1.57e-05 \*\*\*

X1st.Flr.SF 2.694e-04 1.985e-05 13.567 < 2e-16 \*\*\*

Half.Bath 4.197e-02 8.095e-03 5.186 2.41e-07 \*\*\*

TotRms.AbvGrd 1.583e-02 3.356e-03 4.717 2.59e-06 \*\*\*

Fireplaces 3.896e-02 5.525e-03 7.051 2.55e-12 \*\*\*

Garage.Yr.Blt 1.824e-03 1.976e-04 9.228 < 2e-16 \*\*\*

Open.Porch.SF 1.488e-04 4.842e-05 3.074 0.002146 \*\*

Enclosed.Porch 1.064e-04 4.742e-05 2.245 0.024926 \*

Land.ContourHLS 7.008e-02 2.116e-02 3.311 0.000947 \*\*\*

Land.ContourLow -2.639e-02 3.122e-02 -0.845 0.398056

Land.ContourLvl 1.493e-02 1.672e-02 0.893 0.371831

Land.SlopeMod 2.948e-02 1.658e-02 1.777 0.075676 .

Land.SlopeSev -1.510e-01 5.788e-02 -2.609 0.009156 \*\*

Condition.1Feedr 4.526e-02 2.087e-02 2.168 0.030292 \*

Condition.1Norm 8.694e-02 1.705e-02 5.099 3.80e-07 \*\*\*

Condition.1PosA 1.437e-01 3.795e-02 3.786 0.000158 \*\*\*

Condition.1PosN 8.441e-02 3.156e-02 2.675 0.007551 \*\*

Condition.1RRAe -7.712e-03 3.305e-02 -0.233 0.815494

Condition.1RRAn 4.275e-02 2.515e-02 1.700 0.089337 .

Bldg.Type2fmCon -2.695e-02 2.405e-02 -1.121 0.262457

Bldg.TypeDuplex -4.240e-02 2.071e-02 -2.048 0.040758 \*

Bldg.TypeTwnhs -9.039e-02 1.743e-02 -5.186 2.40e-07 \*\*\*

Bldg.TypeTwnhsE -2.875e-02 1.229e-02 -2.340 0.019397 \*

House.Style1.5Unf -1.024e-01 3.652e-02 -2.804 0.005110 \*\*

House.Style1Story -9.602e-02 1.200e-02 -8.005 2.18e-15 \*\*\*

House.Style2.5Unf 3.923e-02 3.601e-02 1.089 0.276133

House.Style2Story 5.299e-02 1.255e-02 4.223 2.54e-05 \*\*\*

House.StyleSFoyer -6.159e-02 2.310e-02 -2.666 0.007746 \*\*

House.StyleSLvl -4.847e-02 1.906e-02 -2.543 0.011083 \*

Mas.Vnr.TypeBrkFace 3.887e-02 3.588e-02 1.083 0.278854

Mas.Vnr.TypeNone 3.730e-02 3.558e-02 1.048 0.294628

Mas.Vnr.TypeStone 6.521e-02 3.703e-02 1.761 0.078418 .

Exter.QualFa -1.078e-01 4.439e-02 -2.429 0.015233 \*

Exter.QualGd -1.480e-02 1.923e-02 -0.769 0.441752

Exter.QualTA -4.051e-02 2.209e-02 -1.834 0.066832 .

Bsmt.QualFa -1.217e-01 2.343e-02 -5.195 2.29e-07 \*\*\*

Bsmt.QualGd -3.941e-02 1.341e-02 -2.938 0.003345 \*\*

Bsmt.QualNone -1.443e-01 3.640e-02 -3.965 7.65e-05 \*\*\*

Bsmt.QualTA -7.873e-02 1.688e-02 -4.665 3.33e-06 \*\*\*

BsmtFin.Type.2BLQ -5.053e-02 2.831e-02 -1.785 0.074478 .

BsmtFin.Type.2GLQ 7.338e-02 3.425e-02 2.142 0.032302 \*

BsmtFin.Type.2LwQ -5.224e-02 2.724e-02 -1.918 0.055335 .

BsmtFin.Type.2None NA NA NA NA

BsmtFin.Type.2Rec -4.795e-02 2.632e-02 -1.822 0.068641 .

BsmtFin.Type.2Unf -8.905e-02 2.283e-02 -3.901 9.96e-05 \*\*\*

Heating.QCFa -6.563e-02 1.922e-02 -3.415 0.000652 \*\*\*

Heating.QCGd -1.312e-02 9.060e-03 -1.449 0.147653

Heating.QCTA -4.993e-02 8.540e-03 -5.847 5.97e-09 \*\*\*

Kitchen.QualFa -7.208e-02 2.685e-02 -2.685 0.007329 \*\*

Kitchen.QualGd -1.797e-02 1.488e-02 -1.208 0.227366

Kitchen.QualTA -4.108e-02 1.690e-02 -2.431 0.015167 \*

FunctionalMaj2 -3.483e-02 7.122e-02 -0.489 0.624901

FunctionalMin1 8.068e-02 4.107e-02 1.964 0.049643 \*

FunctionalMin2 7.916e-02 4.039e-02 1.960 0.050182 .

FunctionalMod 3.925e-02 4.526e-02 0.867 0.385927

FunctionalTyp 1.302e-01 3.613e-02 3.602 0.000325 \*\*\*

Garage.TypeAttchd 5.301e-02 3.075e-02 1.724 0.084850 .

Garage.TypeBasment -9.264e-03 4.260e-02 -0.217 0.827882

Garage.TypeBuiltIn 1.081e-01 3.318e-02 3.259 0.001140 \*\*

Garage.TypeCarPort -3.664e-02 4.836e-02 -0.758 0.448812

Garage.TypeDetchd 2.746e-02 3.108e-02 0.883 0.377098

Sale.TypeCon 1.090e-02 2.938e-02 0.371 0.710721

Sale.TypeNew 9.313e-02 2.103e-02 4.428 1.01e-05 \*\*\*

Sale.TypeOth 2.056e-01 7.212e-02 2.851 0.004406 \*\*

Sale.TypeWD 6.432e-02 4.400e-02 1.462 0.144011

Sale.TypeWD 5.334e-02 1.795e-02 2.972 0.003004 \*\*

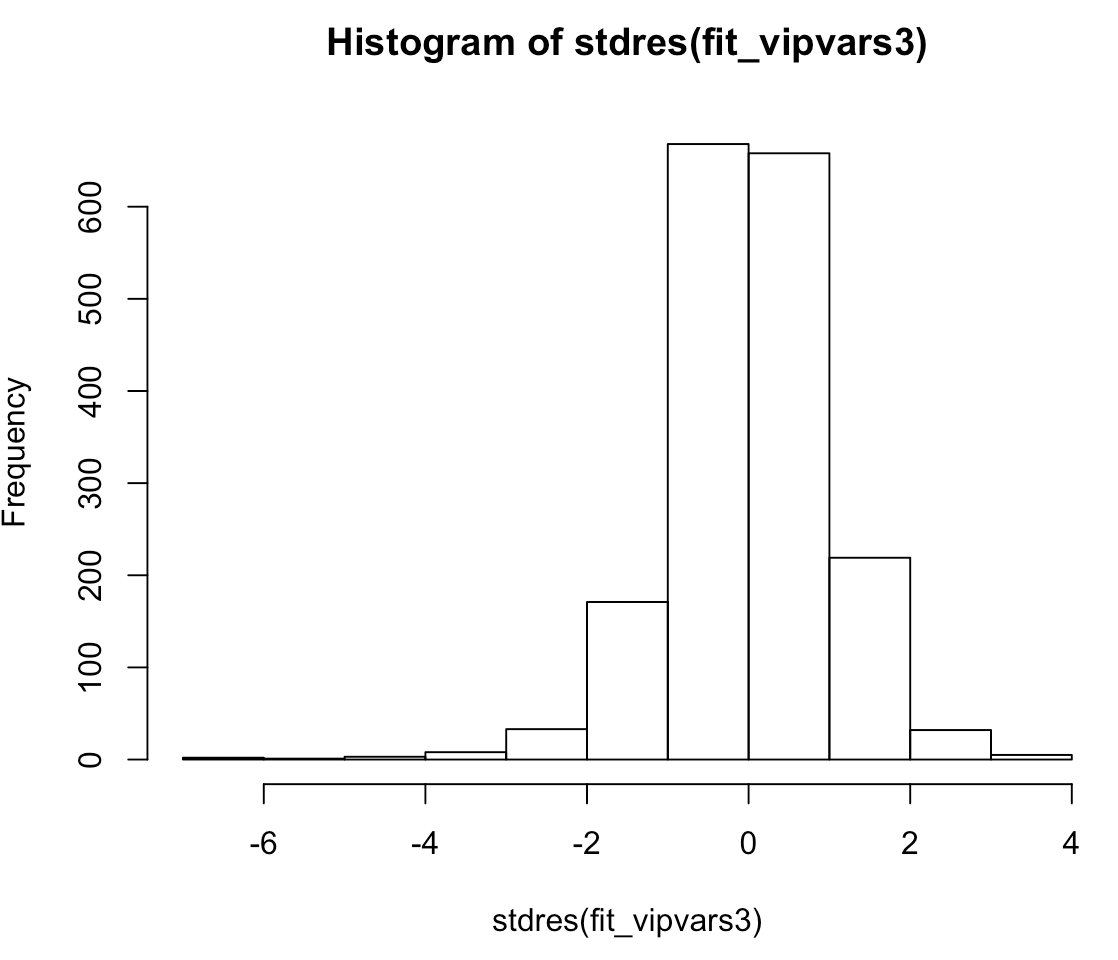
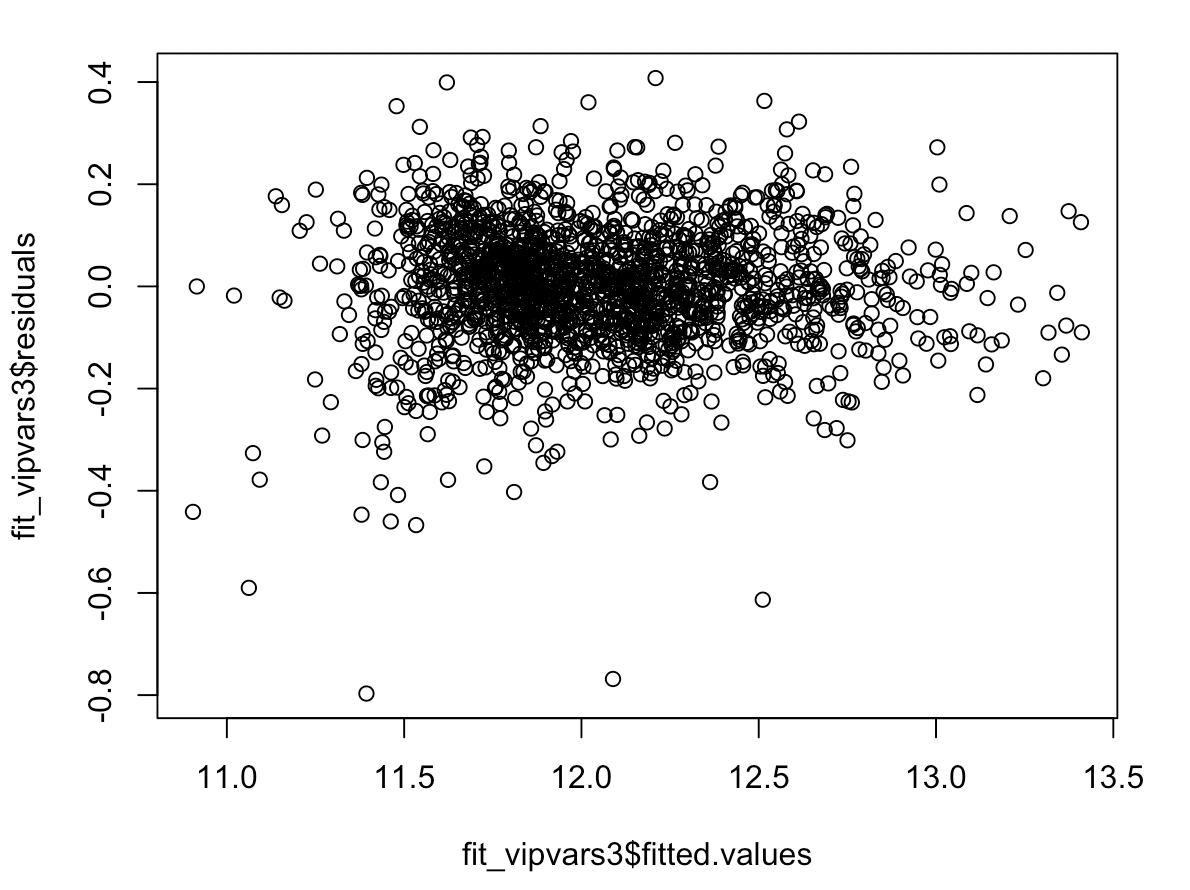
---

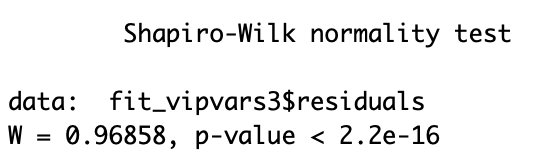
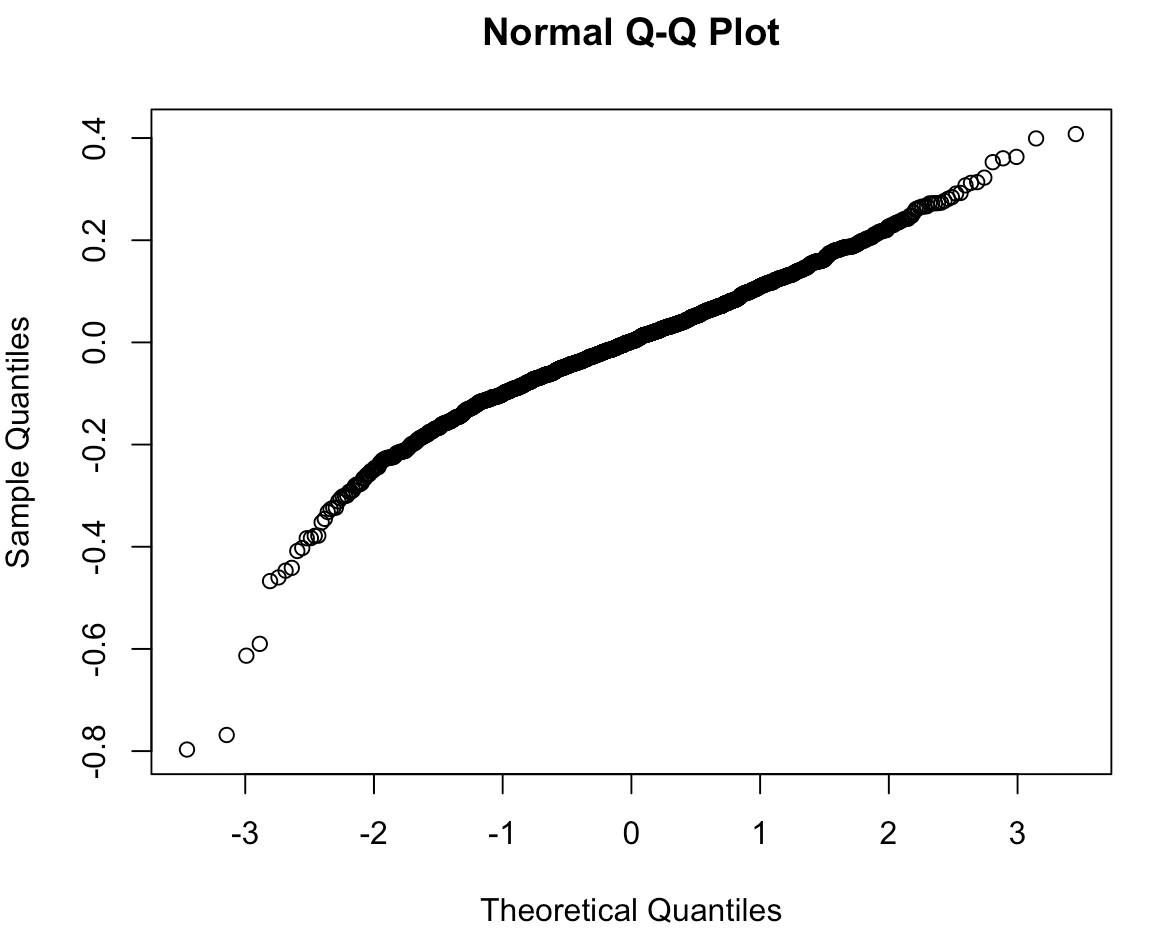
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1196 on 1729 degrees of freedom

Multiple R-squared: 0.9158, Adjusted R-squared: 0.9124

F-statistic: 268.5 on 70 and 1729 DF, p-value: < 2.2e-16





**The diagnostics look as they did before which is still indicating that there are outliers that need to be removed from the data set.**

**Before I remove any outliers, I am going to test my training model against my testing data set and see how they compare.**

Residuals:

Min 1Q Median 3Q Max

-0.46786 -0.06041 0.00136 0.06425 0.32542

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.598e+00 1.006e+00 6.558 1.75e-10 \*\*\*

Lot.Area 1.052e-05 1.792e-06 5.870 9.36e-09 \*\*\*

Overall.Qual 8.248e-02 8.313e-03 9.922 < 2e-16 \*\*\*

Overall.Cond 3.560e-02 7.091e-03 5.021 7.86e-07 \*\*\*

Year.Remod.Add 2.362e-04 4.464e-04 0.529 0.597067

BsmtFin.SF.1 1.378e-04 3.845e-05 3.584 0.000382 \*\*\*

Bsmt.Unf.SF 5.248e-05 3.823e-05 1.373 0.170669

X1st.Flr.SF 3.077e-04 4.351e-05 7.073 7.11e-12 \*\*\*

Half.Bath 4.478e-02 1.578e-02 2.838 0.004778 \*\*

TotRms.AbvGrd 6.061e-03 6.643e-03 0.912 0.362113

Fireplaces 3.643e-02 1.108e-02 3.289 0.001099 \*\*

Garage.Yr.Blt 1.980e-03 3.693e-04 5.360 1.43e-07 \*\*\*

Open.Porch.SF 1.761e-04 9.867e-05 1.785 0.075113 .

Enclosed.Porch -1.073e-04 1.177e-04 -0.912 0.362559

Land.ContourHLS 6.142e-03 3.852e-02 0.159 0.873399

Land.ContourLow -5.659e-02 6.953e-02 -0.814 0.416228

Land.ContourLvl -1.813e-04 2.928e-02 -0.006 0.995063

Land.SlopeMod 3.857e-02 3.595e-02 1.073 0.284061

Land.SlopeSev -1.733e+00 3.755e-01 -4.615 5.34e-06 \*\*\*

Condition.1Feedr 3.969e-02 3.798e-02 1.045 0.296632

Condition.1Norm 8.845e-02 2.902e-02 3.048 0.002463 \*\*

Condition.1PosA 4.312e-02 8.076e-02 0.534 0.593732

Condition.1PosN 1.161e-01 7.452e-02 1.558 0.120004

Condition.1RRAe 1.040e-01 6.810e-02 1.527 0.127670

Condition.1RRAn 1.612e-02 5.206e-02 0.310 0.756995

Bldg.Type2fmCon -4.958e-02 3.943e-02 -1.257 0.209409

Bldg.TypeDuplex -3.762e-02 4.302e-02 -0.874 0.382395

Bldg.TypeTwnhs -8.292e-02 3.378e-02 -2.455 0.014521 \*

Bldg.TypeTwnhsE -1.425e-02 2.555e-02 -0.558 0.577293

House.Style1.5Unf -1.445e-01 8.955e-02 -1.613 0.107542

House.Style1Story -1.500e-01 2.466e-02 -6.084 2.81e-09 \*\*\*

House.Style2.5Unf 4.821e-03 6.877e-02 0.070 0.944142

House.Style2Story -5.701e-03 2.477e-02 -0.230 0.818056

House.StyleSFoyer -9.161e-02 5.204e-02 -1.760 0.079134 .

House.StyleSLvl -1.184e-01 3.341e-02 -3.543 0.000444 \*\*\*

Mas.Vnr.TypeBrkFace 8.464e-02 7.135e-02 1.186 0.236236

Mas.Vnr.TypeNone 6.371e-02 7.078e-02 0.900 0.368624

Mas.Vnr.TypeStone 1.023e-01 7.399e-02 1.383 0.167393

Exter.QualFa -1.114e-01 7.251e-02 -1.537 0.125116

Exter.QualGd 3.303e-02 3.942e-02 0.838 0.402619

Exter.QualTA 1.261e-02 4.535e-02 0.278 0.781216

Bsmt.QualFa -1.741e-01 5.751e-02 -3.027 0.002636 \*\*

Bsmt.QualGd -3.788e-02 2.836e-02 -1.336 0.182371

Bsmt.QualNone -2.202e-01 8.282e-02 -2.659 0.008155 \*\*

Bsmt.QualTA -8.361e-02 3.351e-02 -2.495 0.013018 \*

BsmtFin.Type.2BLQ -1.844e-01 7.375e-02 -2.501 0.012807 \*

BsmtFin.Type.2GLQ -3.722e-02 7.775e-02 -0.479 0.632440

BsmtFin.Type.2LwQ -7.351e-02 7.110e-02 -1.034 0.301786

BsmtFin.Type.2None NA NA NA NA

BsmtFin.Type.2Rec -3.686e-02 6.839e-02 -0.539 0.590192

BsmtFin.Type.2Unf -1.048e-01 6.653e-02 -1.574 0.116192

Heating.QCFa -4.642e-02 3.824e-02 -1.214 0.225443

Heating.QCGd -1.929e-02 1.812e-02 -1.065 0.287540

Heating.QCTA -1.975e-02 1.719e-02 -1.149 0.251326

Kitchen.QualFa -6.016e-02 5.727e-02 -1.050 0.294147

Kitchen.QualGd -5.647e-02 3.211e-02 -1.759 0.079391 .

Kitchen.QualTA -6.296e-02 3.637e-02 -1.731 0.084227 .

FunctionalMaj2 -3.352e-01 1.710e-01 -1.961 0.050637 .

FunctionalMin1 -1.432e-01 1.516e-01 -0.944 0.345621

FunctionalMin2 -1.979e-01 1.536e-01 -1.289 0.198327

FunctionalMod -2.197e-01 1.565e-01 -1.404 0.161204

FunctionalTyp -1.427e-01 1.467e-01 -0.973 0.331311

Garage.TypeAttchd -3.541e-02 1.259e-01 -0.281 0.778716

Garage.TypeBasment -1.089e-01 1.347e-01 -0.808 0.419390

Garage.TypeBuiltIn 2.556e-02 1.280e-01 0.200 0.841869

Garage.TypeCarPort -3.185e-01 1.794e-01 -1.775 0.076683 .

Garage.TypeDetchd -7.730e-02 1.256e-01 -0.615 0.538588

Sale.TypeCon 7.830e-02 5.038e-02 1.554 0.120959

Sale.TypeNew 9.406e-02 3.739e-02 2.515 0.012292 \*

Sale.TypeWD 1.664e-01 9.197e-02 1.809 0.071170 .

Sale.TypeWD 5.198e-02 2.985e-02 1.742 0.082377 .

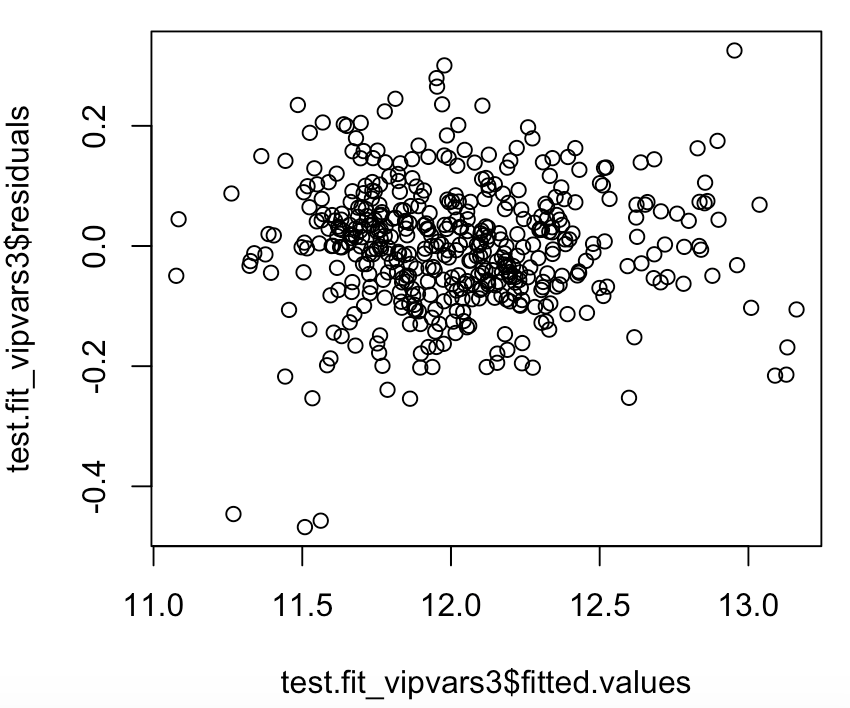
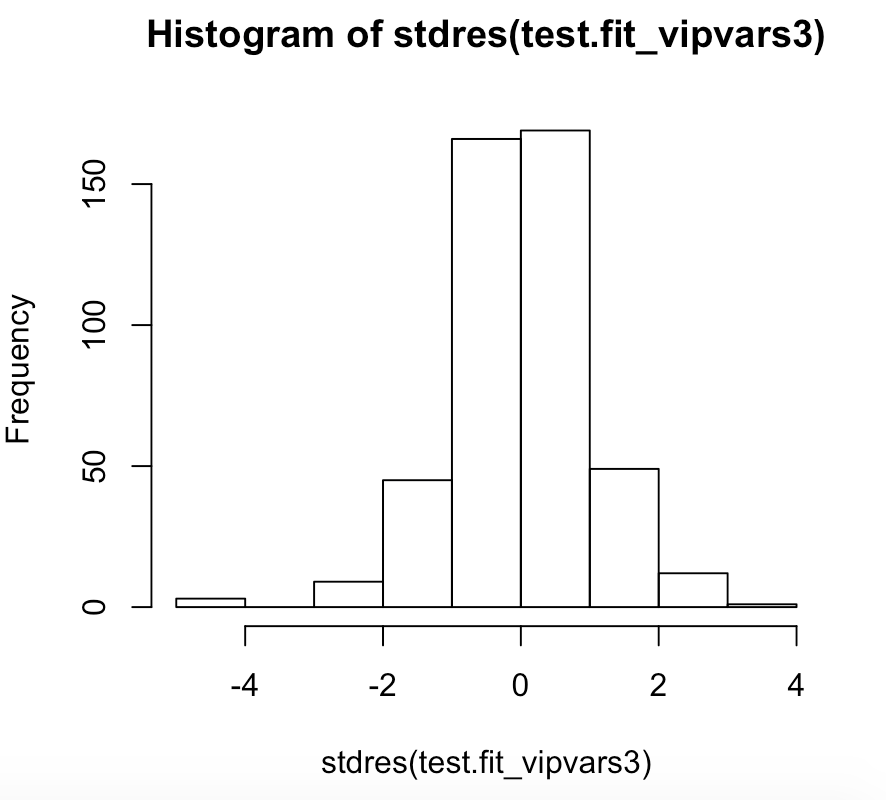
---

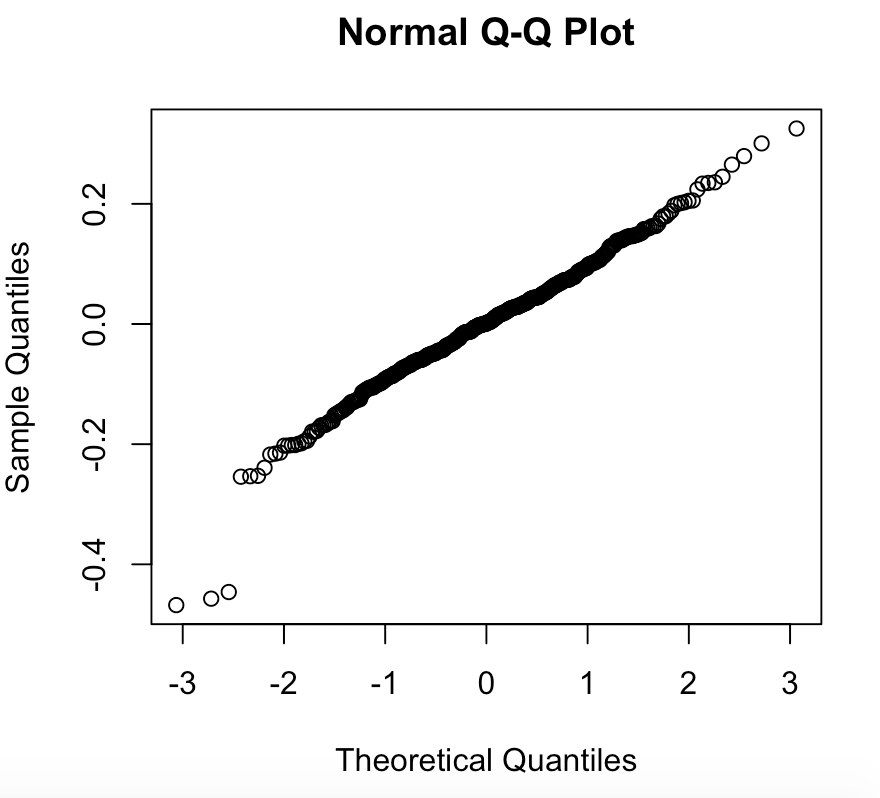
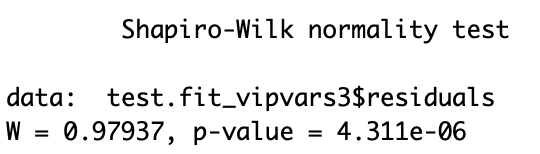
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1138 on 388 degrees of freedom

Multiple R-squared: 0.9231, Adjusted R-squared: 0.9094

F-statistic: 67.48 on 69 and 388 DF, p-value: < 2.2e-16

**The diagnostics of the testing data set indicates that there are outliers in the dataset which can be seen in the histogram of the standard residuals. It also appears that the data might not follow a normal distribution as indicated by the low p-value in the shapairo-wilk test.**

|  |  |  |
| --- | --- | --- |
|  | **Training Set** | **Testing Set** |
| **Size** | 1800 | 458 |
| **Residual std error** | 0.1196 on 1729 degrees of freedom | 0.1138 on 388 degrees of freedom |
| **R-squared** | 0.9158 | 0.9231 |

**The coefficients of the variables look similar. The R-squared values and the residual standard error are very similar. Year.Remod.Add, Bsmt.Unf.SF, TotRms.AbvGrd, Enclosed.Porch, Land.Contour, Mas.Vnr.Type, Exter.Qual, Heating.QC all have p-values above 0.05. Since some of these values were 0.20 or above, I removed Year.Remod.Add, TotRms.AbvGrd, Enclosed.Porch, Land.Contour, Mas.Vnr.Type, Heating.QC from the model. Doing so resulted in the following Residual std error and r-squared for the testing and training sets:**

|  |  |  |
| --- | --- | --- |
|  | **Training Set** | **Testing Set** |
| **Residual std error** | 0.1231 on 1741 degrees of freedom | 0.1137 on 400 degrees of freedom |
| **R-squared** | 0.91 | 0.9209 |

**Since both of these errors and r-squared values are similar to each other and similar to the model before removing those variables, this indicates that the model is a good fit for the testing and training data sets. All of the variables that are in the model now are statistically significant (p-values less than 0.05), which also indicates a good model.**

**Now that I have a model that works well for both my training and testing dataset, I tested the model on the entire dataset including any rows that were previously omitted due to missing data. I also took the advice from below and included “na.action=na.exclude” to ensure that R would calculate the outliers of the model correctly since there are missing values in the dataset.**

Residuals:

Min 1Q Median 3Q Max

-0.81370 -0.06540 0.00438 0.07170 0.48584

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.719e+00 3.026e-01 22.202 < 2e-16 \*\*\*

Lot.Area 3.472e-06 3.834e-07 9.056 < 2e-16 \*\*\*

Overall.Qual 7.763e-02 3.294e-03 23.571 < 2e-16 \*\*\*

Overall.Cond 3.868e-02 2.560e-03 15.113 < 2e-16 \*\*\*

BsmtFin.SF.1 1.650e-04 1.320e-05 12.498 < 2e-16 \*\*\*

Bsmt.Unf.SF 6.475e-05 1.339e-05 4.836 1.40e-06 \*\*\*

X1st.Flr.SF 3.162e-04 1.374e-05 23.007 < 2e-16 \*\*\*

Half.Bath 3.305e-02 6.403e-03 5.162 2.62e-07 \*\*\*

Fireplaces 4.087e-02 4.448e-03 9.189 < 2e-16 \*\*\*

Garage.Yr.Blt 2.076e-03 1.500e-04 13.836 < 2e-16 \*\*\*

Open.Porch.SF 1.484e-04 3.937e-05 3.770 0.000167 \*\*\*

Land.SlopeMod 4.552e-02 1.186e-02 3.838 0.000127 \*\*\*

Land.SlopeSev -9.781e-02 3.753e-02 -2.606 0.009212 \*\*

Condition.1Feedr 4.173e-02 1.747e-02 2.389 0.016945 \*

Condition.1Norm 8.887e-02 1.448e-02 6.135 9.74e-10 \*\*\*

Condition.1PosA 1.090e-01 3.133e-02 3.480 0.000509 \*\*\*

Condition.1PosN 9.269e-02 2.477e-02 3.742 0.000186 \*\*\*

Condition.1RRAe 1.867e-02 2.646e-02 0.706 0.480400

Condition.1RRAn 3.808e-02 2.150e-02 1.771 0.076623 .

Bldg.Type2fmCon -4.044e-02 1.984e-02 -2.038 0.041646 \*

Bldg.TypeDuplex -3.604e-02 1.606e-02 -2.244 0.024925 \*

Bldg.TypeTwnhs -1.306e-01 1.404e-02 -9.301 < 2e-16 \*\*\*

Bldg.TypeTwnhsE -6.307e-02 9.460e-03 -6.667 3.15e-11 \*\*\*

House.Style1.5Unf -1.428e-01 3.268e-02 -4.369 1.30e-05 \*\*\*

House.Style1Story -1.230e-01 9.355e-03 -13.152 < 2e-16 \*\*\*

House.Style2.5Unf 7.846e-02 2.864e-02 2.739 0.006200 \*\*

House.Style2Story 6.286e-02 1.010e-02 6.226 5.52e-10 \*\*\*

House.StyleSFoyer -8.984e-02 1.820e-02 -4.936 8.44e-07 \*\*\*

House.StyleSLvl -7.552e-02 1.423e-02 -5.306 1.21e-07 \*\*\*

Exter.QualFa -1.674e-01 3.514e-02 -4.765 1.99e-06 \*\*\*

Exter.QualGd -1.023e-02 1.644e-02 -0.622 0.533742

Exter.QualTA -4.714e-02 1.848e-02 -2.551 0.010793 \*

Bsmt.QualFa -1.621e-01 2.041e-02 -7.946 2.80e-15 \*\*\*

Bsmt.QualGd -6.391e-02 1.105e-02 -5.784 8.13e-09 \*\*\*

Bsmt.QualNone -2.083e-01 2.875e-02 -7.246 5.59e-13 \*\*\*

Bsmt.QualTA -1.097e-01 1.345e-02 -8.154 5.33e-16 \*\*\*

BsmtFin.Type.2BLQ -7.756e-02 2.291e-02 -3.385 0.000722 \*\*\*

BsmtFin.Type.2GLQ 4.198e-02 2.812e-02 1.493 0.135620

BsmtFin.Type.2LwQ -6.488e-02 2.172e-02 -2.988 0.002836 \*\*

BsmtFin.Type.2None NA NA NA NA

BsmtFin.Type.2Rec -6.307e-02 2.110e-02 -2.989 0.002822 \*\*

BsmtFin.Type.2Unf -9.777e-02 1.844e-02 -5.301 1.25e-07 \*\*\*

Kitchen.QualFa -9.624e-02 2.307e-02 -4.171 3.12e-05 \*\*\*

Kitchen.QualGd -2.894e-02 1.268e-02 -2.281 0.022596 \*

Kitchen.QualTA -6.585e-02 1.408e-02 -4.676 3.08e-06 \*\*\*

FunctionalMaj2 -4.555e-02 5.651e-02 -0.806 0.420249

FunctionalMin1 9.542e-02 3.509e-02 2.719 0.006584 \*\*

FunctionalMin2 8.971e-02 3.503e-02 2.561 0.010483 \*

FunctionalMod 7.495e-02 3.823e-02 1.960 0.050060 .

FunctionalTyp 1.289e-01 3.156e-02 4.085 4.53e-05 \*\*\*

Garage.TypeAttchd 1.485e-02 2.657e-02 0.559 0.576309

Garage.TypeBasment -2.725e-02 3.398e-02 -0.802 0.422656

Garage.TypeBuiltIn 8.133e-02 2.830e-02 2.874 0.004087 \*\*

Garage.TypeCarPort -7.686e-02 4.232e-02 -1.816 0.069490 .

Garage.TypeDetchd -2.081e-02 2.676e-02 -0.778 0.436895

Sale.TypeCon 3.954e-02 2.391e-02 1.654 0.098315 .

Sale.TypeNew 1.174e-01 1.683e-02 6.975 3.84e-12 \*\*\*

Sale.TypeOth 1.893e-01 6.316e-02 2.997 0.002748 \*\*

Sale.TypeWD 8.911e-02 3.684e-02 2.419 0.015629 \*

Sale.TypeWD 6.322e-02 1.396e-02 4.530 6.17e-06 \*\*\*

---

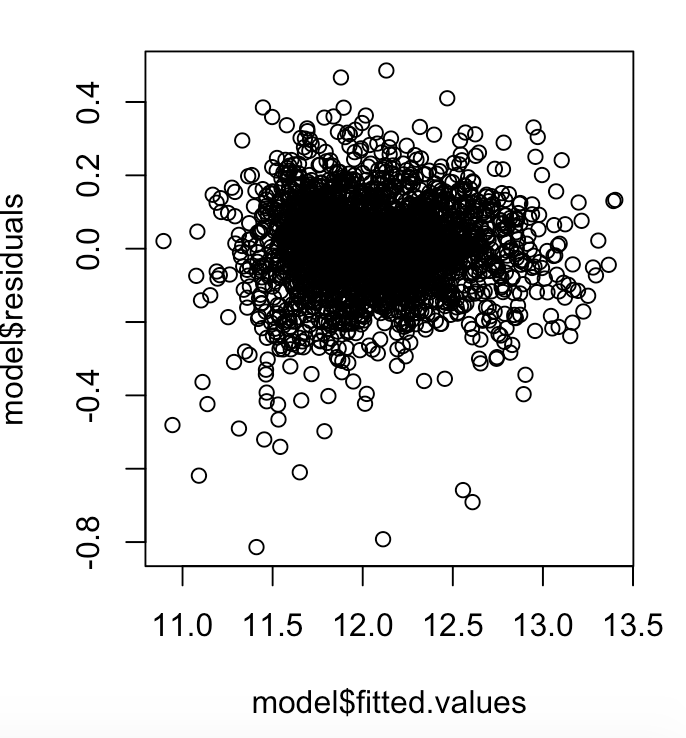
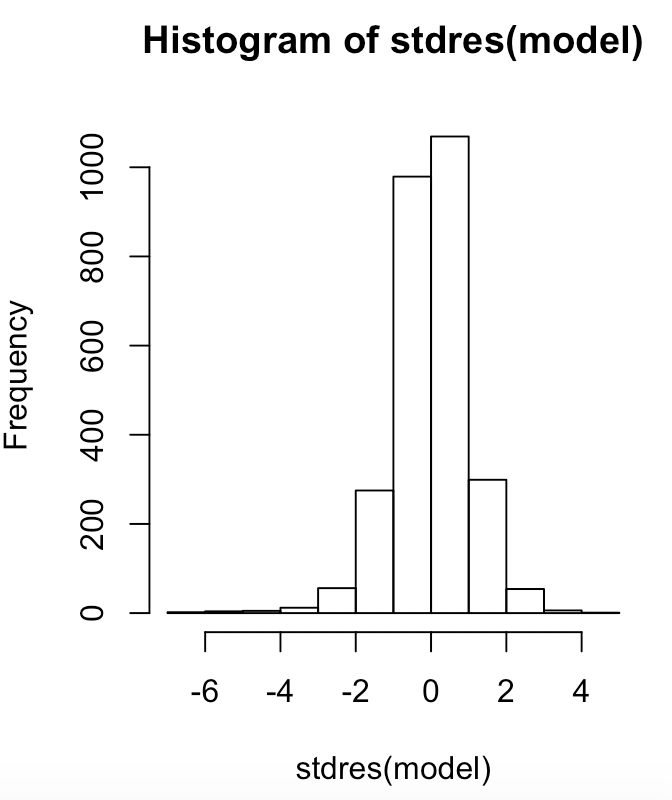
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

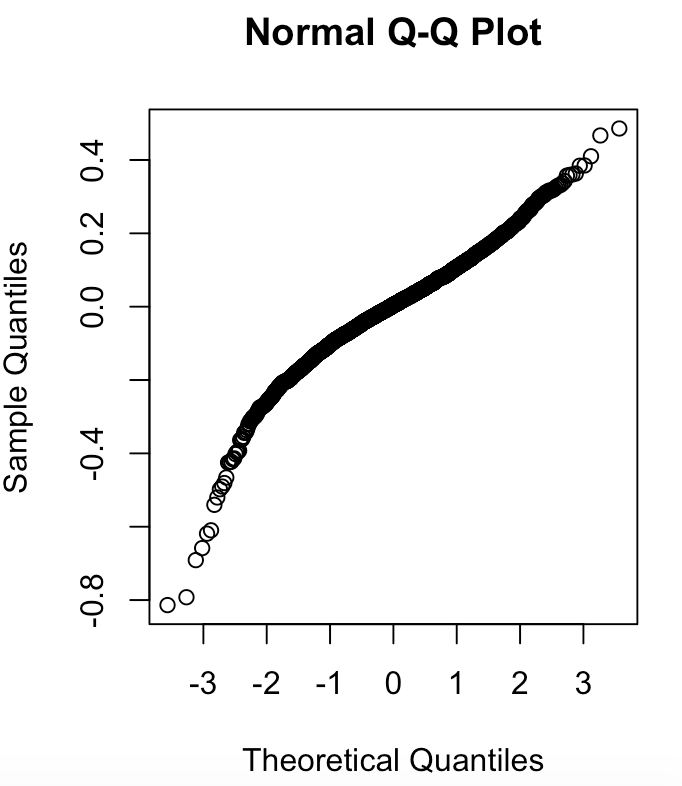
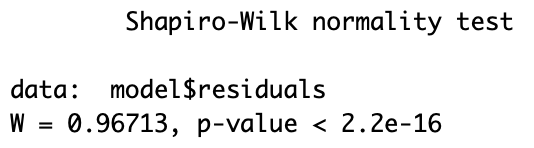
Residual standard error: 0.1224 on 2703 degrees of freedom

(163 observations deleted due to missingness)

Multiple R-squared: 0.9011, Adjusted R-squared: 0.899

F-statistic: 424.7 on 58 and 2703 DF, p-value: < 2.2e-16

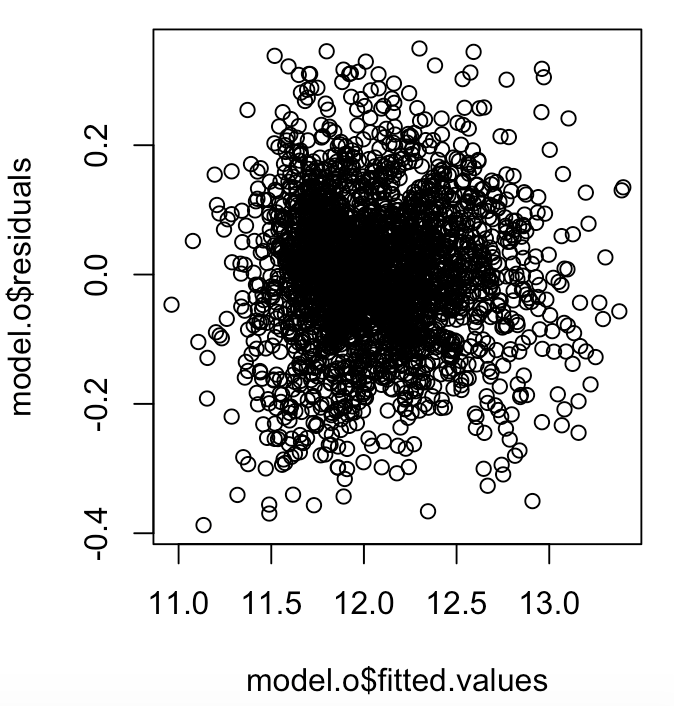
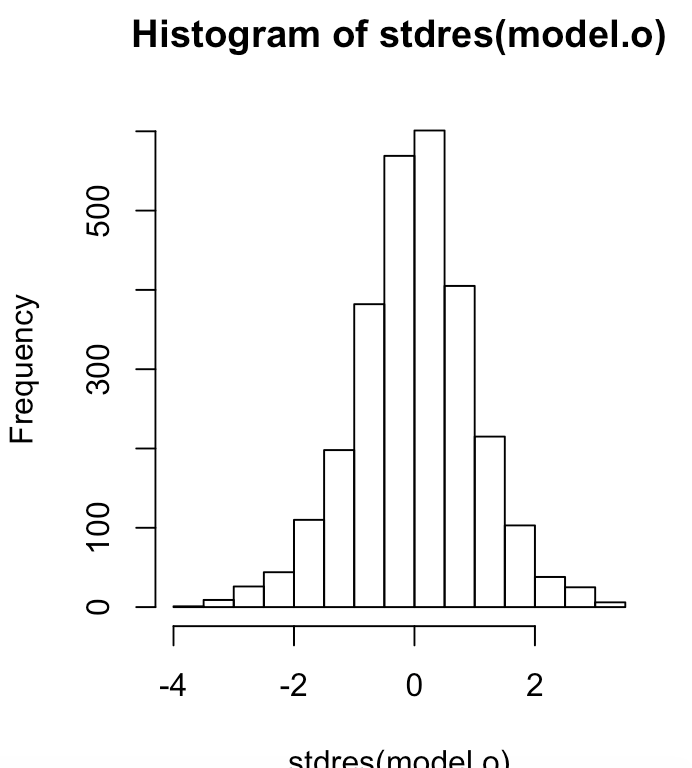
 

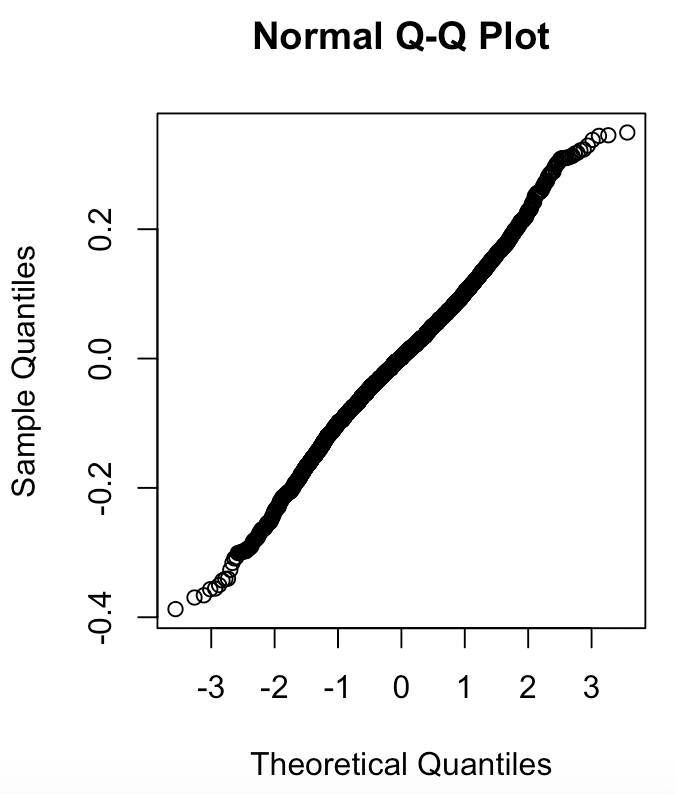
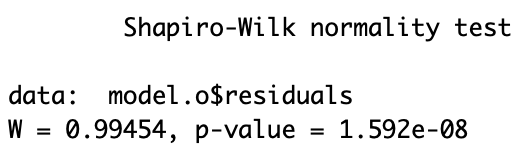
 

**The model looks good. It has relatively low std residual error as well as having a relatively high r-squared value. The diangostics still suggest that there are outliers in the data that need to be removed.**

**In order to remove the outliers from the data set I added an additional column to the original dataset called sres which is the standard residual from the model I fit above. This then allowed for me to create a new subset of the dataset where the standard residual was larger than the absolute value of 3, meaning more than 3 standard deviations. I then ran a new model using the new dataset without outliers.**

**Rerunning the model with the new dataset resulted in these diagnostic graphs:**

**** ****

**** ****

**There appears to still be outliers in the data as can be seen in the histogram by values outside of the interval [-3, 3]. In order to fix this issue, I repeated the same process I just did two more times in order to not have any outliers in the dataset for the model. The resulting model and diagnostics are below:**

Residuals:

Min 1Q Median 3Q Max

-0.316020 -0.063973 0.002373 0.065721 0.311369

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.730e+00 2.760e-01 24.384 < 2e-16 \*\*\*

Lot.Area 6.763e-06 5.138e-07 13.163 < 2e-16 \*\*\*

Overall.Qual 7.443e-02 2.900e-03 25.660 < 2e-16 \*\*\*

Overall.Cond 3.347e-02 2.261e-03 14.806 < 2e-16 \*\*\*

BsmtFin.SF.1 1.394e-04 1.176e-05 11.848 < 2e-16 \*\*\*

Bsmt.Unf.SF 4.479e-05 1.191e-05 3.761 0.000173 \*\*\*

X1st.Flr.SF 3.310e-04 1.253e-05 26.417 < 2e-16 \*\*\*

Half.Bath 2.513e-02 5.591e-03 4.494 7.28e-06 \*\*\*

Fireplaces 3.484e-02 3.873e-03 8.996 < 2e-16 \*\*\*

Garage.Yr.Blt 2.055e-03 1.368e-04 15.026 < 2e-16 \*\*\*

Open.Porch.SF 1.311e-04 3.428e-05 3.823 0.000135 \*\*\*

Land.SlopeMod 3.154e-02 1.036e-02 3.045 0.002352 \*\*

Land.SlopeSev -2.651e-02 3.469e-02 -0.764 0.444760

Condition.1Feedr 5.085e-02 1.515e-02 3.356 0.000801 \*\*\*

Condition.1Norm 9.303e-02 1.255e-02 7.415 1.63e-13 \*\*\*

Condition.1PosA 1.251e-01 2.753e-02 4.543 5.79e-06 \*\*\*

Condition.1PosN 9.333e-02 2.134e-02 4.374 1.27e-05 \*\*\*

Condition.1RRAe 1.570e-02 2.278e-02 0.689 0.490703

Condition.1RRAn 4.313e-02 1.864e-02 2.314 0.020731 \*

Bldg.Type2fmCon -3.405e-02 1.744e-02 -1.952 0.051010 .

Bldg.TypeDuplex -4.440e-02 1.401e-02 -3.169 0.001548 \*\*

Bldg.TypeTwnhs -1.133e-01 1.240e-02 -9.136 < 2e-16 \*\*\*

Bldg.TypeTwnhsE -5.310e-02 8.513e-03 -6.237 5.16e-10 \*\*\*

House.Style1.5Unf -1.463e-01 2.812e-02 -5.203 2.11e-07 \*\*\*

House.Style1Story -1.235e-01 8.225e-03 -15.017 < 2e-16 \*\*\*

House.Style2.5Unf 3.088e-02 2.591e-02 1.192 0.233439

House.Style2Story 6.708e-02 8.878e-03 7.555 5.72e-14 \*\*\*

House.StyleSFoyer -9.779e-02 1.586e-02 -6.167 8.05e-10 \*\*\*

House.StyleSLvl -7.963e-02 1.239e-02 -6.429 1.52e-10 \*\*\*

Exter.QualFa -1.023e-01 3.167e-02 -3.231 0.001249 \*\*

Exter.QualGd -7.582e-03 1.419e-02 -0.534 0.593111

Exter.QualTA -4.730e-02 1.599e-02 -2.958 0.003119 \*\*

Bsmt.QualFa -1.937e-01 1.785e-02 -10.850 < 2e-16 \*\*\*

Bsmt.QualGd -6.825e-02 9.561e-03 -7.139 1.21e-12 \*\*\*

Bsmt.QualNone -2.314e-01 2.526e-02 -9.159 < 2e-16 \*\*\*

Bsmt.QualTA -1.125e-01 1.172e-02 -9.601 < 2e-16 \*\*\*

BsmtFin.Type.2BLQ -4.856e-02 2.001e-02 -2.427 0.015281 \*

BsmtFin.Type.2GLQ 3.631e-02 2.427e-02 1.496 0.134747

BsmtFin.Type.2LwQ -5.335e-02 1.880e-02 -2.838 0.004580 \*\*

BsmtFin.Type.2None NA NA NA NA

BsmtFin.Type.2Rec -5.510e-02 1.828e-02 -3.013 0.002608 \*\*

BsmtFin.Type.2Unf -7.517e-02 1.611e-02 -4.667 3.21e-06 \*\*\*

Kitchen.QualFa -1.009e-01 2.033e-02 -4.964 7.35e-07 \*\*\*

Kitchen.QualGd -4.076e-02 1.105e-02 -3.688 0.000231 \*\*\*

Kitchen.QualTA -7.333e-02 1.228e-02 -5.973 2.64e-09 \*\*\*

FunctionalMaj2 -2.793e-02 4.860e-02 -0.575 0.565505

FunctionalMin1 1.112e-01 3.035e-02 3.663 0.000254 \*\*\*

FunctionalMin2 1.079e-01 3.035e-02 3.555 0.000385 \*\*\*

FunctionalMod 8.918e-02 3.306e-02 2.698 0.007021 \*\*

FunctionalTyp 1.597e-01 2.719e-02 5.871 4.87e-09 \*\*\*

Garage.TypeAttchd 4.060e-02 2.294e-02 1.770 0.076894 .

Garage.TypeBasment 5.907e-03 2.947e-02 0.200 0.841127

Garage.TypeBuiltIn 9.779e-02 2.444e-02 4.001 6.48e-05 \*\*\*

Garage.TypeCarPort -6.282e-02 3.644e-02 -1.724 0.084846 .

Garage.TypeDetchd 1.109e-02 2.313e-02 0.479 0.631827

Sale.TypeCon 2.476e-02 2.099e-02 1.179 0.238458

Sale.TypeNew 1.086e-01 1.467e-02 7.408 1.72e-13 \*\*\*

Sale.TypeOth 7.244e-02 6.210e-02 1.167 0.243488

Sale.TypeWD 7.573e-02 3.172e-02 2.388 0.017021 \*

Sale.TypeWD 5.612e-02 1.219e-02 4.602 4.37e-06 \*\*\*

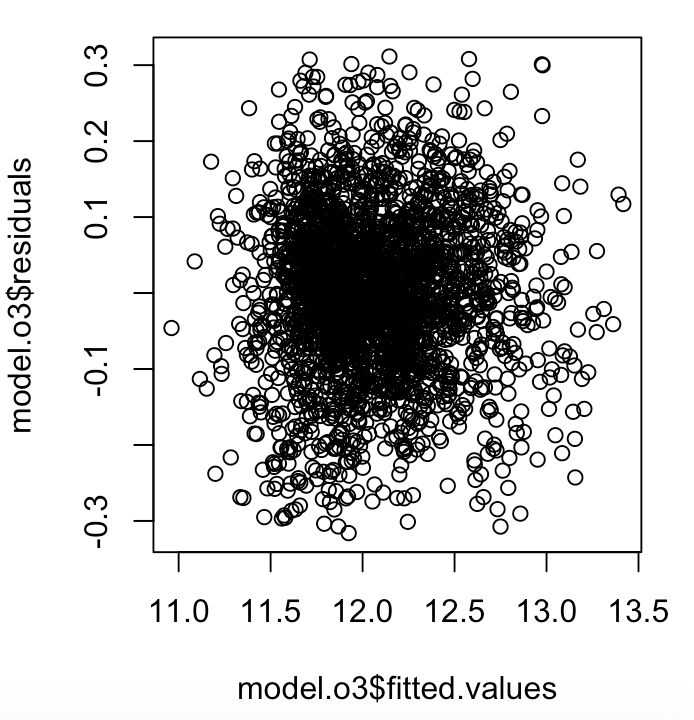
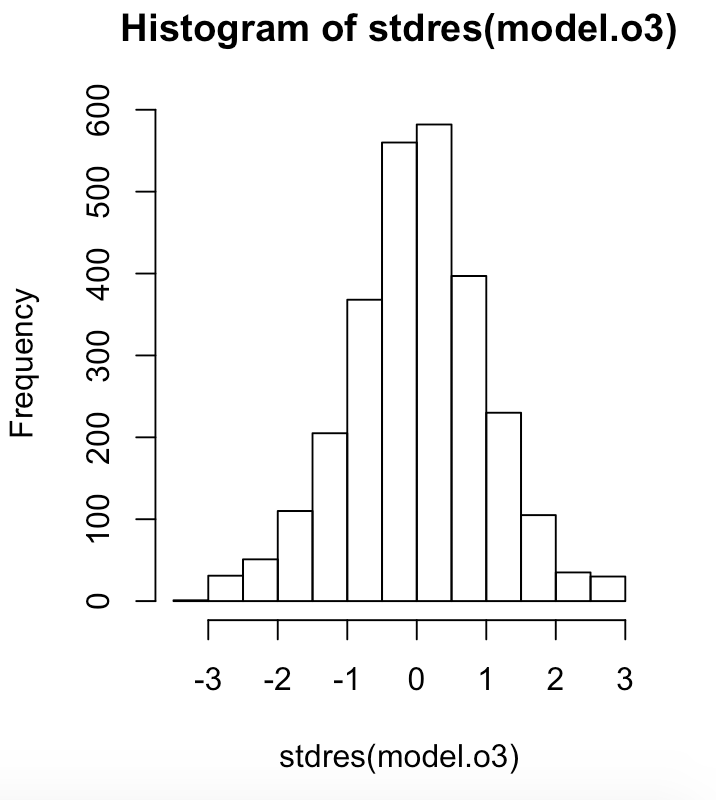
---

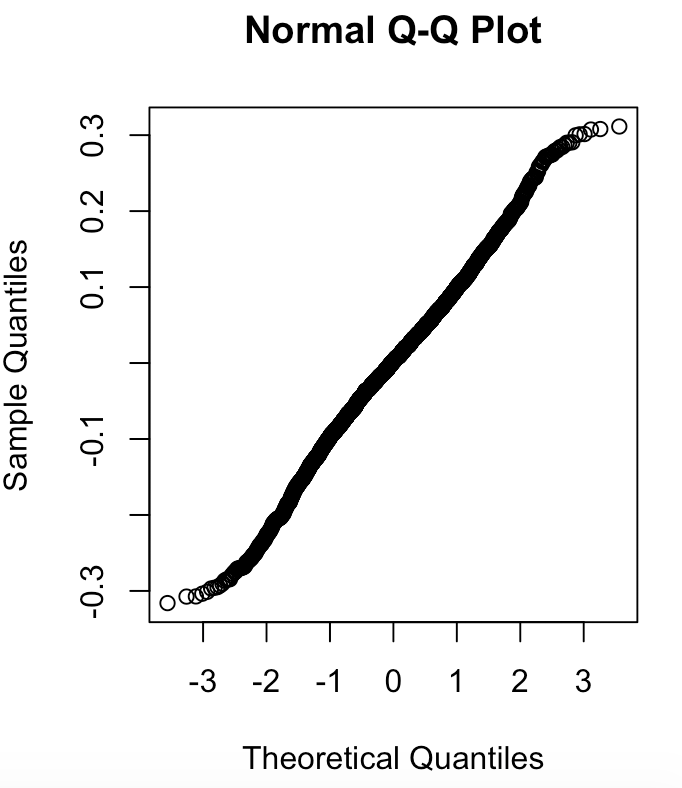
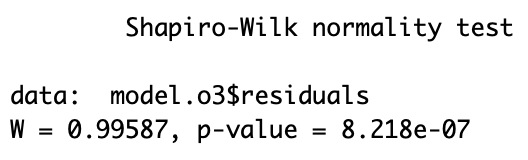
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1052 on 2646 degrees of freedom

Multiple R-squared: 0.9225, Adjusted R-squared: 0.9208

F-statistic: 542.7 on 58 and 2646 DF, p-value: < 2.2e-16

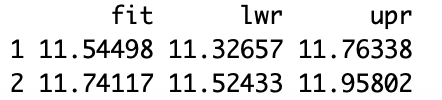
 

**All the diagnostics for the model look good. Except, the shapiro-wilk test is still indicating that the data does not follow a normal distribution because the p-value is less than 0.05.**

b) When appropriate provide output from R to justify the conclusions you are making. I do not need to see everything you do in R, but I need enough output to follow your reasoning.

2) After you have developed a model, predict the SalePrice for the 2 houses that are in the file AmesHousing\_predict.csv. Compute a prediction along with 95% prediction intervals for the SalePrice.

**For the 2 houses in the file the 95% prediction intervals for log(SalePrice) is below:**



**The prediction intervals for the SalePrice of the 2 houses are:**

|  |  |  |  |
| --- | --- | --- | --- |
| House | Fit | Lower | Upr |
| 1 | 103257.382 | 82997.8514 | 128460.918 |
| 2 | 125639.252 | 101146.981 | 156063.772 |

**Commented R code for reference:**

library(leaps)

library(MASS)

#SETTING UP THE DATA AND CLEANING

#Read in data as iowa

iowa\_full = read.csv('AmesHousing.csv')

#omit missing values

iowa = na.omit(iowa\_full)

#create a training set and a testing set

set.seed(1)

train=sample(2258,1800)

test=(c(1:2258)[-train])

#FIND OUT HOW MANY VARIABLE

#split the data to only have numeric vars

num.iowa=data.frame(iowa[,c(2,3,14:17,23,31,33:35,40:49,51,53,56,58,59,63:68,70,71,74)])

fit\_num = lm(SalePrice~., data = num.iowa, subset=train)

summary(fit\_num)

#test how many variables should be included

fit\_num.full = regsubsets(SalePrice~., data=num.iowa, nvmax=35)

summary(fit\_num.full)

reg.summary = summary(fit\_num.full)

reg.summary$adjr2

plot(reg.summary$adjr2, xlab="Number of Variables", ylab = "Adjusted RSq", type = "l")

reg.summary$cp

plot(reg.summary$cp, xlab = "Number of Variables", ylab = " Cp", type ="l")

reg.summary$bic

plot(reg.summary$bic, xlab = "Number of Variables", ylab=" BIC", type = "l")

#find the variables that should be included for a X variable model

coef(fit\_num.full, 20)

fit\_num.part = lm(SalePrice~Lot.Frontage+Lot.Area + Overall.Qual + Overall.Cond + Year.Built + Year.Remod.Add+

Mas.Vnr.Area + BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Bsmt.Full.Bath +

Bsmt.Half.Bath + Half.Bath + TotRms.AbvGrd + Fireplaces + Garage.Yr.Blt + Wood.Deck.SF+

Open.Porch.SF + Enclosed.Porch + Misc.Val, data = num.iowa, subset= train)

summary(fit\_num.part)

#adjust the model based off the pvalues from prev output

fit\_num.part1 = lm(SalePrice~Lot.Area + Overall.Qual + Overall.Cond + Year.Remod.Add+

Mas.Vnr.Area + BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath + TotRms.AbvGrd +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF + Enclosed.Porch,

data = num.iowa, subset= train)

summary(fit\_num.part1)

#Add categorical variables to the model one at a time run anova command to decide if to add

#if close mentality is to add and can remove later if needed

model1 = lm(SalePrice~Lot.Area + Overall.Qual + Overall.Cond + Year.Remod.Add+

Mas.Vnr.Area + BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath + TotRms.AbvGrd +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF + Enclosed.Porch +Lot.Shape + Land.Contour + Lot.Config +

Land.Slope+ Condition.1 + Condition.2+ Bldg.Type+ House.Style+ Roof.Style+ Exterior.1st+Exterior.2nd+

Mas.Vnr.Type+ Exter.Qual+Foundation+Bsmt.Qual+Bsmt.Exposure+ BsmtFin.Type.2+ Heating.QC+Kitchen.Qual+

Functional+ Fireplace.Qu+Garage.Type+Sale.Type,

data = iowa, subset= train)

model2 = lm(SalePrice~Lot.Area + Overall.Qual + Overall.Cond + Year.Remod.Add+

Mas.Vnr.Area + BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath + TotRms.AbvGrd +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF + Enclosed.Porch + Lot.Shape + Land.Contour + Lot.Config +

Land.Slope+ Condition.1 + Condition.2 + Bldg.Type + House.Style + Roof.Style + Exterior.1st + Exterior.2nd+

Mas.Vnr.Type+ Exter.Qual +Foundation +Bsmt.Qual + Bsmt.Exposure + BsmtFin.Type.2 + Heating.QC+ Kitchen.Qual+

Functional+ Fireplace.Qu+ Garage.Type+Sale.Type,

data = iowa, subset= train)

anova(model1, model2)

#model with only statistically significant variables

fit\_vipvars = lm(SalePrice~Lot.Area + Overall.Qual + Overall.Cond + Year.Remod.Add+

Mas.Vnr.Area + BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath + TotRms.AbvGrd +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF + Enclosed.Porch + Lot.Shape + Land.Contour + Lot.Config +

Land.Slope+ Condition.1 + Condition.2 + Bldg.Type + House.Style + Roof.Style + Exterior.1st + Exterior.2nd+

Mas.Vnr.Type+ Exter.Qual +Foundation +Bsmt.Qual + Bsmt.Exposure + BsmtFin.Type.2 + Heating.QC+ Kitchen.Qual+

Functional+ Fireplace.Qu+ Garage.Type+Sale.Type,

data = iowa, subset= train)

summary(fit\_vipvars)

#after looking at p-values remove some vars from the model

fit\_vipvars = lm(SalePrice~Lot.Area + Overall.Qual + Overall.Cond + Year.Remod.Add+

Mas.Vnr.Area + BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath + TotRms.AbvGrd +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF + Enclosed.Porch + Land.Contour + Lot.Config +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

Mas.Vnr.Type+ Exter.Qual +Bsmt.Qual + Bsmt.Exposure + BsmtFin.Type.2 + Heating.QC+ Kitchen.Qual+

Functional+ Fireplace.Qu+ Garage.Type+Sale.Type,

data = iowa\_noOut3)

summary(fit\_vipvars)

#Look for issues in res v fitted and stdres

plot(fit\_vipvars$residuals~fit\_vipvars$fitted.values)

hist(stdres(fit\_vipvars))

qqnorm(fit\_vipvars$residuals)

shapiro.test(fit\_vipvars$residuals)

#test to see how tranformation would look since residual looks cornicopia shaped

boxcox(fit\_vipvars)

iowa$logPrice = log(iowa$SalePrice)

fit\_vipvars2 = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond + Year.Remod.Add+

Mas.Vnr.Area + BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath + TotRms.AbvGrd +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF + Enclosed.Porch + Land.Contour + Lot.Config +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

Mas.Vnr.Type+ Exter.Qual +Bsmt.Qual + Bsmt.Exposure + BsmtFin.Type.2 + Heating.QC+ Kitchen.Qual+

Functional+ Fireplace.Qu+ Garage.Type+Sale.Type,

data = iowa, subset= train)

summary(fit\_vipvars2)

plot(fit\_vipvars2$residuals~fit\_vipvars2$fitted.values)

hist(stdres(fit\_vipvars2))

qqnorm(fit\_vipvars2$residuals)

shapiro.test(fit\_vipvars2$residuals)

#removed some variable that are not significant

fit\_vipvars3 = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond + Year.Remod.Add+

BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath + TotRms.AbvGrd +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF + Enclosed.Porch + Land.Contour +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

Mas.Vnr.Type+ Exter.Qual +Bsmt.Qual + BsmtFin.Type.2 + Heating.QC+ Kitchen.Qual+

Functional + Garage.Type+Sale.Type,

data = iowa, subset= train)

summary(fit\_vipvars3)

#run diagnostics on the model

plot(fit\_vipvars3$residuals~fit\_vipvars3$fitted.values)

hist(stdres(fit\_vipvars3))

qqnorm(fit\_vipvars3$residuals)

shapiro.test(fit\_vipvars3$residuals)

#use testing data to see if model is a good fit

#also removed variables one at a time that were not statistically significant

test.fit\_vipvars3 = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond +

BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

+ Exter.Qual +Bsmt.Qual + BsmtFin.Type.2 + Kitchen.Qual+

Functional + Garage.Type+Sale.Type,

data = iowa, subset= train)

summary(test.fit\_vipvars3)

plot(test.fit\_vipvars3$residuals~test.fit\_vipvars3$fitted.values)

hist(stdres(test.fit\_vipvars3))

qqnorm(test.fit\_vipvars3$residuals)

shapiro.test(test.fit\_vipvars3$residuals)

#combine all the data together and test the model using entire data set

iowa\_full$logPrice = log(iowa\_full$SalePrice)

model = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond +

BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

+ Exter.Qual +Bsmt.Qual + BsmtFin.Type.2 + Kitchen.Qual+

Functional + Garage.Type+Sale.Type,

data = iowa\_full, na.action=na.exclude )

summary(model)

plot(model$residuals~model$fitted.values)

hist(stdres(model))

qqnorm(model$residuals)

shapiro.test(model$residuals)

#show and remove outliers

#add standard residual column to data set

iowa\_full$sres = stdres(model)

#show outliers using standard residual column

subset(iowa\_full,model$sres< -3)

subset(iowa\_full,model$sres> 3)

#modify dataset to exclude outliers

iowa\_noOut=subset(iowa\_full,abs(iowa\_full$sres) < 3)

#rerun dianogstics to see if problems are fixed

model.o = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond +

BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

+ Exter.Qual +Bsmt.Qual + BsmtFin.Type.2 + Kitchen.Qual+

Functional + Garage.Type+Sale.Type,

data = iowa\_noOut, na.action=na.exclude )

summary(model.o)

plot(model.o$residuals~model.o$fitted.values)

hist(stdres(model.o))

qqnorm(model.o$residuals)

shapiro.test(model.o$residuals)

#data still has outliers

#redo the above process but using the new dataset

iowa\_noOut$sres = stdres(model.o)

subset(iowa\_noOut,model.o$sres< -3)

subset(iowa\_noOut,model.o$sres> 3)

iowa\_noOut1=subset(iowa\_noOut,abs(iowa\_noOut$sres) < 3)

model.o1 = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond +

BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

+ Exter.Qual +Bsmt.Qual + BsmtFin.Type.2 + Kitchen.Qual+

Functional + Garage.Type+Sale.Type,

data = iowa\_noOut1, na.action=na.exclude )

summary(model.o1)

plot(model.o1$residuals~model.o1$fitted.values)

hist(stdres(model.o1))

qqnorm(model.o1$residuals)

shapiro.test(model.o1$residuals)

iowa\_noOut1$sres = stdres(model.o1)

#data still has outliers

#redo the above process but using the new dataset

subset(iowa\_noOut1,model.o1$sres < -3)

subset(iowa\_noOut1,model.o1$sres > 3)

iowa\_noOut2=subset(iowa\_noOut1,abs(iowa\_noOut1$sres) < 3)

model.o2 = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond +

BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

+ Exter.Qual +Bsmt.Qual + BsmtFin.Type.2 + Kitchen.Qual+

Functional + Garage.Type+Sale.Type,

data = iowa\_noOut2, na.action=na.exclude )

summary(model.o2)

plot(model.o2$residuals~model.o2$fitted.values)

hist(stdres(model.o2))

qqnorm(model.o2$residuals)

shapiro.test(model.o2$residuals)

iowa\_noOut2$sres = stdres(model.o2)

#data still has outliers

#redo the above process but using the new dataset

subset(iowa\_noOut2,model.o2$sres < -3)

subset(iowa\_noOut2,model.o2$sres > 3)

iowa\_noOut3=subset(iowa\_noOut2,abs(iowa\_noOut2$sres) < 3)

model.o3 = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond +

BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

+ Exter.Qual +Bsmt.Qual + BsmtFin.Type.2 + Kitchen.Qual+

Functional + Garage.Type+Sale.Type,

data = iowa\_noOut3, na.action=na.exclude )

summary(model.o3)

plot(model.o3$residuals~model.o3$fitted.values)

hist(stdres(model.o3))

qqnorm(model.o3$residuals)

shapiro.test(model.o3$residuals)

##Final Model and prediction

predict = read.csv("AmesHousing\_predict.csv")

model\_final = lm(logPrice~Lot.Area + Overall.Qual + Overall.Cond +

BsmtFin.SF.1 + Bsmt.Unf.SF + X1st.Flr.SF + Half.Bath +

Fireplaces + Garage.Yr.Blt + Open.Porch.SF +

Land.Slope+ Condition.1 + Bldg.Type + House.Style+

+ Exter.Qual +Bsmt.Qual + BsmtFin.Type.2 + Kitchen.Qual+

Functional + Garage.Type+Sale.Type,

data = iowa\_noOut3, na.action=na.exclude )

pred=predict(model\_final,predict,interval="prediction")

pred

Considerations in Grading:

Initial Model Determination: There are many possible independent variables in the data set. You should have a strategy for determining which of these variables you will include in the model and you should clearly describe that strategy so that I can understand your approach. You should provide an appropriate output from R to support the choices you make.

Model Checking and Modification: Once you have estimated a model, describe what you do to verify that the model assumptions are approximately correct and what changes to the model you are making. You should clearly describe why you are making changes in a model. Provide appropriate R output to support your choices.

Dealing with Outliers: In a large data set there are undoubtedly outliers. Describe how you identify any outliers and how you propose to deal with the observations that are outliers.

Model Verification: You should have a method to validate your model on an independent set of data. Describe how you propose to do this and provide an appropriate R output.

There are many possible models that can be derived from this data set, therefore there are many possible correct answers. I am not just interested in the final model, but also in the process your go through to arrive at your model. A significant amount of the grade for the project will be determined by the process you use to come up with your final model. If you have a poor process or do not clearly describe that process your grade will be reduced significantly.

Suggestions on how to approach the analysis of the data:

1) Since you will be doing a lot of things that involve using the data to suggest which variables to include in a model, it is a good idea to divide the data set into a training set and a test set. You should then do all of your analysis to determine a model on the training set without using the test set. After you have come up with a model using the training set you can then use the test set to check to see if all the variables that you have included in your model are needed.

2) The data set has a number of missing values. R has a variety of ways to deal with missing values and sometimes this can cause problems, thus when you are developing a model it is a good idea to remove the observations who have missing values. This seems OK with this data set because there are a large number of observations in the data set so that removing those with missing values will still leave a lot of data. The command to remove the missing values is: newdataframe = na.omit(dataframe) where newdataframe is the name you give to the data frame without the missing values and dataframe is the name of the data frame that has the data with missing values.

After you develop a model, it is a good idea to refit the model using the entire data set (including missing values) because your final model will probably not include some of the inputs which have missing values.

3) This is a data set with a very large number of possible input variables and in addition there are a large number of categorical variables. Because there are a large number of possible input variables, it is natural to try to use some sort of variable selection method like best subsets regression or stepwise regression to help in determining which input variables to choose. It can be a problem using a variable selection method with categorical variables because for a categorical variable with K categories, R will create K-1 indicator variables. This will greatly increase the number of potential input variables and could cause a problem.

One way to proceed is to break the data into 2 groups: the group of numeric variables and the group of categorical (factor) variables and to use a sequential approach to determining which variables to include in the model like that described below.

Step 1: Create a data frame that contains only the numeric inputs and the dependent variable, SalePrice. If you have read the data into a dataframe named ames, then a command to select these variables is:

num.ames=data.frame(ames[,c(2,3,14:17,23,31,33:35,40:49,51,53,56,58,59,63:68,70,71,74)])

This will create a dataframe named num.ames that only contains the variables in the columns listed in the command. These are the columns that contain the numeric variables and the variable SalePrice.

Step 2: Use the new data frame to develop a model that only includes the numeric variables. When you are satisfied that you have a good model for the numeric variables you can add the appropriate categorical variables. (Note, you should have removed the missing values.)

Step 3: You need a way to decide which categorical variables to add to the model you developed in Step 2. One way to do this is to start with the model you have from step 2, call this model 1, and then add one or more categorical variables to this model. Call the model with the added categorical variables model 2. Then the following R commands will perform a hypothesis test of whether or not the variables that are added to model 1 are statistically significant.

model1 = lm( the R expressions for model 1)

model2 = lm(the R expressions for model 2)

anova(model1,model2)

The output from the anova command will give a p-value for testing the null hypothesis that model1 is correct versus the alternative hypothesis that model2 is correct. The smaller the p-value, the more evidence in favor of model2.

You want to find a way to figure out which of the categorical variables are most likely to improve the model you developed in step 2. Once you have decided on which categorical variables are most likely to be helpful, you can add those to the model you developed in step 2 to get a model that includes both numeric and categorical variables.

4) When the data set has missing values, when R computes the residuals for a model, it does not compute the residuals for any data that is missing. Because of that the number of residuals is not the same as the number of variables in the data set. This causes a problem when you are trying to identify the outliers is a set of data that contains missing values. One way to resolve this problem is to include the expression na.action=na.exclude in the regression command. This will tell the program to put missing values in the residuals in places where the original data had missing data. The command would be like the following:

fit=lm(Y~X1+X2 + X3, data=dataframename, na.action=na.exclude)

If you use this command, then the residuals with large values will correspond to the rows of the observations that are associated with those large observations and you will be able to properly identify the observations corresponding to the outliers.