MSDS 601- Linear Regression
Final Project

Nicolas Decavel-Bueff, Jiahui Jin, Teddy Mefford

December 2, 2020

Research Statement

Our goal for this research project is to develop a model that, given certain metrics about a house, can accurately predict the sale price of that house. We will be primarily using multiple linear regression (MLR) techniques to model the data, however, we will briefly explore some other techniques as well. Developing a successful MLR model requires us to identify the best predictor variables, deal with influential points, and explore other model diagnostics that could potentially be impacting our model.

Description of the Dataset

For our analysis, we chose a Housing dataset from Kaggle that has an initial size of 1460 rows and 80 columns, where each row represents a house that was sold, and each column describing the house. We will be designing our model to optimally choose predictors from the column variables to best predict the SalePrice. A complete description of the data variables are given in appendix 1.

Exploratory Data Analysis and a Preliminary Model

To begin to explore the dataset and get some benchmarks for how well a MLR model will work, we choose several variables in the model that seem like they could be significant predictors for SalePrice. Upon an initial investigation of several numeric predictors (Fig 1), it is clear that a linear relationship does exist in the data between these predictors and the target value, SalePrice. These plots also indicate that heteroscedascity is present as can be seen by the widening in the spread of data points for increase in the predictors: GrLivArea, LotArea, and TotalBsmtSF. Heteroscedascity will be explored in more depth below. These initial plots also show a number of outlying points that fall far below the line we would expect to get from a linear fit to the data. Fitting an ordinary least squares regression model to the six numeric variables in (Fig 1-left), we get an initial adjusted R^2 of 0.748.

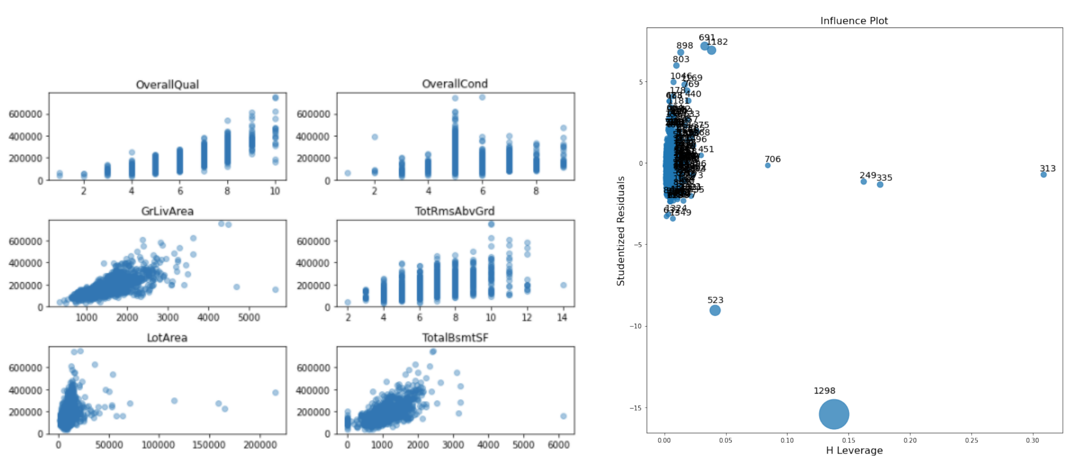


Figure 1: Inital investigation of linearity in data against SalePrice and model summary

To identify the outliers (Fig. 1-left), we look at the data points for which LotArea is greater than 100,000 and TotalBsmtSF greater than 5000 to get point 1298, 249, 313, 335 and 706. Upon further investigation of these points they seem reasonable enough. In particular, they could be very large land properties, with small homes, in low population areas. One point in particular is slightly suspicious, as the total basement is approximately 300ft by 300ft which is the size of two football fields and the house is only 6000 square feet. Further investigation with the model's influence plot (Fig 1-right) clearly shows that these points significantly influence our model.

We now extend this model to include three categorical variables to obtain the model:

$$SalePrice = \beta_0 + \beta_1 * TotRmsAbvGrd + \beta_2 * GrLivArea + \beta_3 * OverallCond + \beta_4 * OverallQual + \beta_5 * TotalBsmtSF + \beta_6 * LotArea + \beta_7 * C(Neighborhood) + \beta_8 * C(ExterQual) + \beta_9 * C(BsmtQual) \tag{1}$$

Removing the most influential data points and fitting the model, our adjusted R^2 increases from 0.815 to 0.847.

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.820
Model:	OLS	**Adj. R-squared:**	0.815
Method:	Least Squares	**F-statistic:**	185.6
Date:	Thu, 03 Dec 2020	**Prob (F-statistic):**	0.00
Time:	02:44:31	**Log-Likelihood:**	-16854.
No. Observations:	1423	**AIC:**	3.378e+04
Df Residuals:	1388	**BIC:**	3.396e+04
Df Model:	34		
Covariance Type:	nonrobust		

TotRmsAbvGrd	-376.7752	1046.006	-0.360	0.719	-2428.699	1675.149
GrLivArea	55.2692	3.751	14.735	0.000	47.911	62.627
OverallCond	6076.5804	924.074	6.576	0.000	4263.847	7889.313
OverallQual	1.372e+04	1252.426	10.959	0.000	1.13e+04	1.62e+04

Omnibus:	352.601	**Durbin-Watson:**	1.915
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	18832.749
Skew:	-0.166	**Prob(JB):**	0.00
Kurtosis:	20.819	**Cond. No.**	7.82e+04

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.850
Model:	OLS	**Adj. R-squared:**	0.847
Method:	Least Squares	**F-statistic:**	230.7
Date:	Thu, 03 Dec 2020	**Prob (F-statistic):**	0.00
Time:	02:48:02	**Log-Likelihood:**	-16638.
No. Observations:	1416	**AIC:**	3.335e+04
Df Residuals:	1381	**BIC:**	3.353e+04
Df Model:	34		
Covariance Type:	nonrobust		

TotRmsAbvGrd	-3687.6349	979.743	-3.764	0.000	-5609.581	-1765.689
GrLivArea	73.6054	3.623	20.314	0.000	66.497	80.713
OverallCond	5839.6598	842.829	6.929	0.000	4186.296	7493.023
OverallQual	1.326e+04	1143.583	11.593	0.000	1.1e+04	1.55e+04

Omnibus:	327.769	**Durbin-Watson:**	1.896
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2573.926
Skew:	0.853	**Prob(JB):**	0.00
Kurtosis:	9.381	**Cond. No.**	7.75e+04

Figure 2: Comparison of model with and without 6 most significant points (categorical variables not shown due to large number

Next, let's look at collinearity between the numeric predictors in this model. From the correlation matrix below, we see that we do have significant correlation between GrLivArea and TotRmsAbvGrd. This is in conjunction with the partial anova test (appendix 2) flagging GrLivArea as insignificant. In general, multicollinearity could create issues with our model coefficient estimates swinging wildly and being very sensitive to small changes in the model. Overall this will weaken the statistical power of our model so we would want to remove this multicollinearity if we were to proceed further with this model. However, there are many more predictors in the dataset that have not yet been explored so we will go into a deeper investigation of optimizing our predictors in the next section

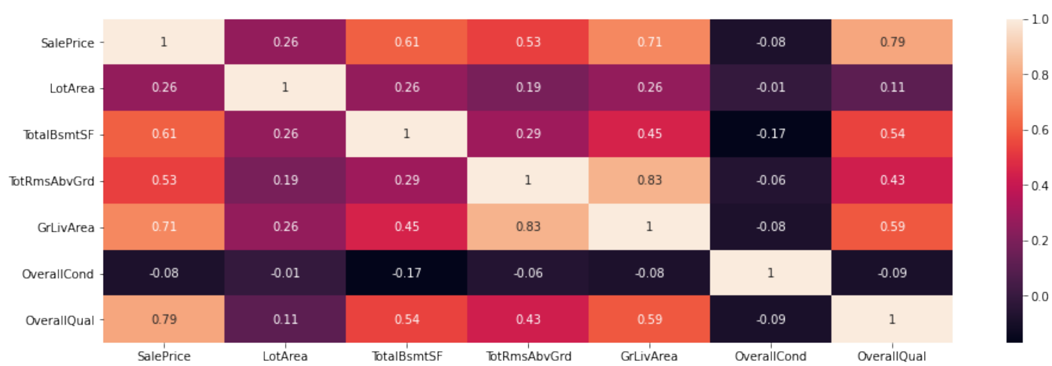


Figure 3: Multicollinearity in the Data

As a last diagnostic on our preliminary model let's look at the normality of the residuals. From Figure 4, we see that our previous model's residuals have heavy tails. We can also observe that eliminating the largest influential points begins to improve the model's kurtosis, but it is still present.

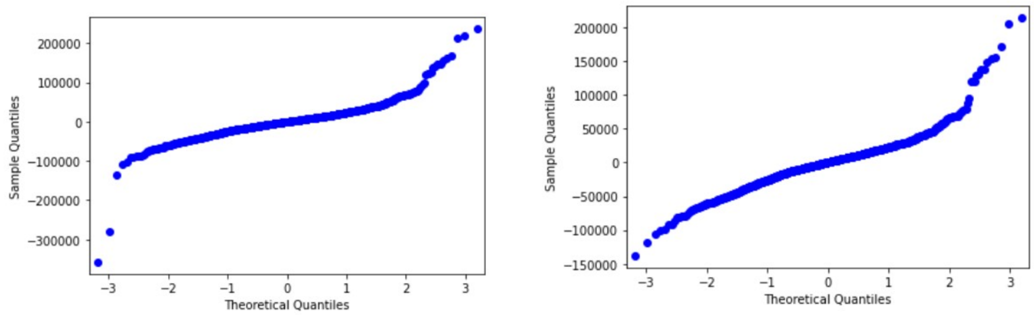


Figure 4: QQ Plots before (left) and after (right) removal of 5 most influential points

With our initial investigation done, we want to extend our model to include all of the possible predictors and to use an algorithm to identify which predictors may give us the best fit. We run a script to compare the adjusted R^2 for a variety of predictor combinations in order to find the most promising candidates.

Feature Engineering

After having our initial model, we decided to perform a more in-depth exploratory analysis on a particular categorical variable in order to decrease the number of parameters it introduced in the model. In particular, we looked at Neighborhood, which has 25 different categories, leading to a total of 24 parameters in our model.

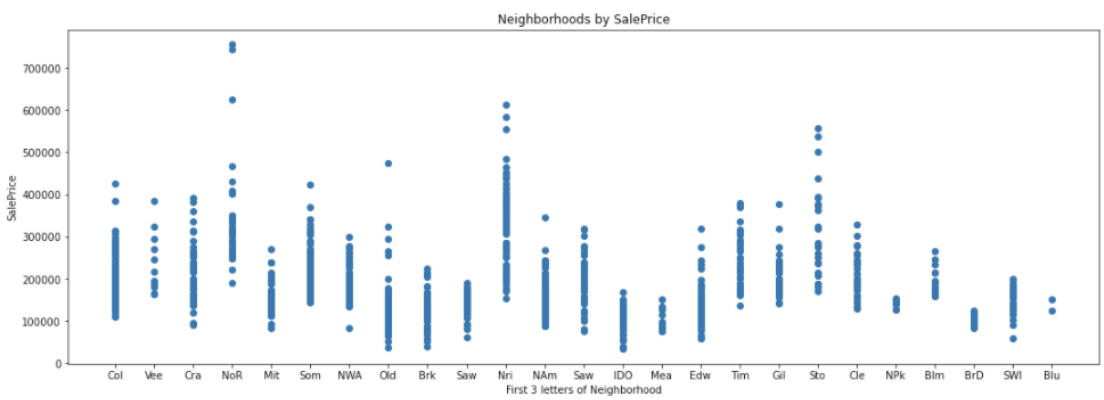


Figure 5: Visual representation of neighborhood

In an attempt to transform this categorical variable into a numerical one, we took the mean SalePrice of each neighborhood, sorted it, and then assigned them values from 1 to 25. This helped us retain some of the information given by Neighborhood while decreasing the degrees of freedom it introduced in the model.

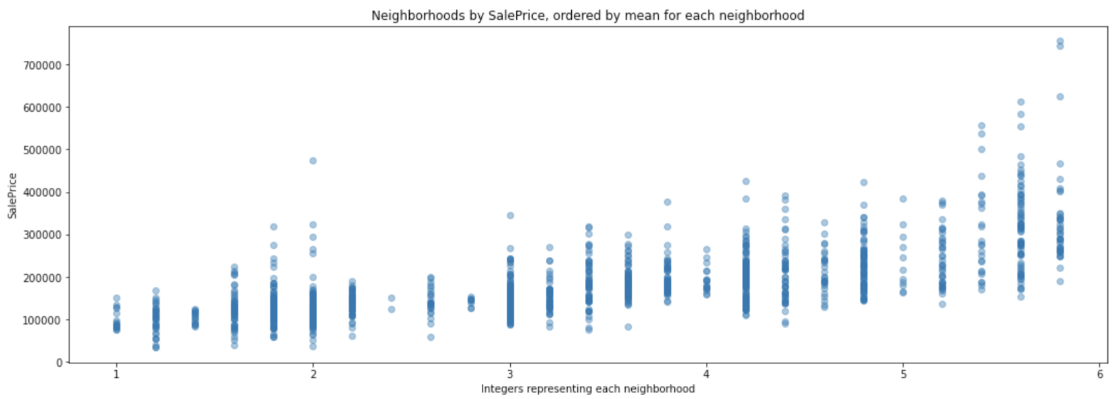


Figure 6: Linear transformation of neighborhood

As we can see in the above transformation, we are able to retain some of the Neighborhood data even as a numerical variable. Note that we will convert SalePrice to the log(SalePrice) which will improve upon the above linear relationship.

Log-Linear Decision

Next, we decided to explore the potential heteroskedasticity in our original model. Plotting the model's fitted values against the residuals, we notice a funnel-like shape (left):

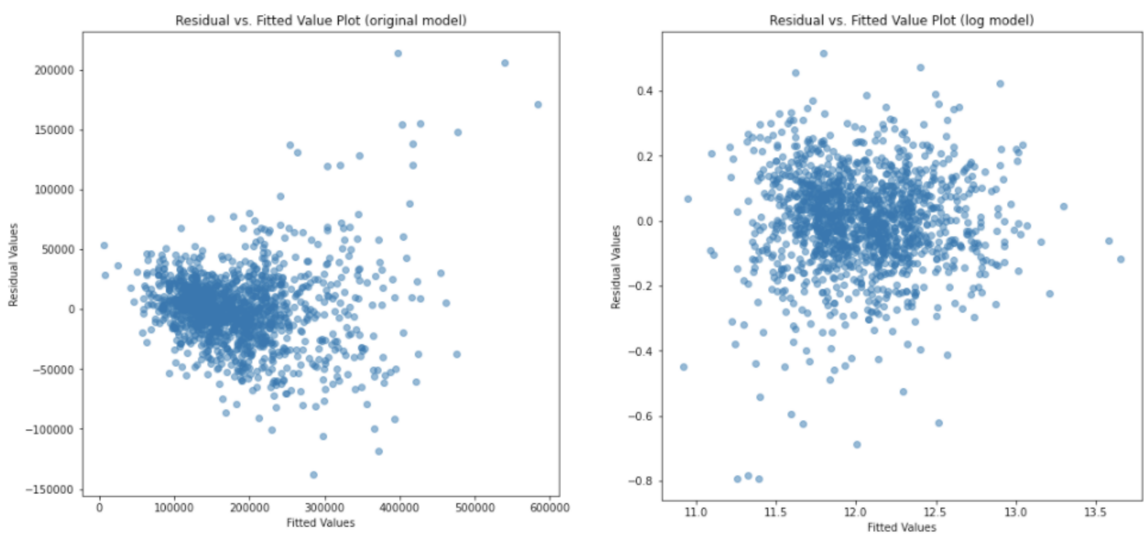


Figure 7: Fitted vs residual plot to show heteroscedasticity

This, in conjunction with a failed Breusch-Pagan test, led us to the conclusion that we should take the log of our dependent variable. As we can see on the right, the fitted value vs. residual plot looks much more random.

Refined Model Selection

At this point, our next goal was to create a script that would choose the best model by continuously adding the next best parameter, as measured by adjusted r-squared. This script would stop when either our adjusted r-squared could not be improved by adding another variable, or when adding the next "best" variable would cause our model to have less than 30 observations per parameter.

After iterating through our script, we came up with a better model than our previous one, with an initial adjusted r-squared of 0.907.

OLS Regression Results

Dep. Variable:	np.log(SalePrice)	**R-squared:**	0.910
Model:	OLS	**Adj. R-squared:**	0.907
Method:	Least Squares	**F-statistic:**	303.1
Date:	Wed, 02 Dec 2020	**Prob (F-statistic):**	0.00
Time:	14:05:06	**Log-Likelihood:**	1025.0
No. Observations:	1460	**AIC:**	-1954.
Df Residuals:	1412	**BIC:**	-1700.
Df Model:	47		
Covariance Type:	nonrobust		

Omnibus:	751.231	**Durbin-Watson:**	1.955
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	22191.309
Skew:	-1.808	**Prob(JB):**	0.00
Kurtosis:	21.754	**Cond. No.**	6.21e+05

Figure 8: Refined model: influential points still included

Through our Jarque-Bera test, Skew and Kurtosis statistics, we can quickly tell that influential points might exist and affect normality. But overall, we don't think the non-normality problem in our model is serious because we have a large sample size. According to the Central Limit Theorem, since we have over 30 sample points per predictor, we can assume that the non-normality problem will not affect the model performance too much. The next step in improving our model is to test for both multicollinearity and remove any influential points.

Multicollinearity and Influential Points

We investigate multicollinearity in our model using both VIF and a correlation table (see Fig 3). We used Variance Inflation Factors (VIF) to test for multicollinearity. As we can see with the below result, none are significant (greater than 10). Note that, when working with categorical variables, we can afford to keep the variable in our model if at least one of its categories have a low VIF.

original model (with influential)

	VIF Factor	features
0	29798.967970	Intercept
1	1.906010	C(KitchenQual)[T.Fa]
2	4.960513	C(KitchenQual)[T.Gd]
3	6.887653	C(KitchenQual)[T.TA]
4	8.189724	C(MSZoning)[T.FV]
5	2.734077	C(MSZoning)[T.RH]
6	26.851703	C(MSZoning)[T.RL]
7	20.526618	C(MSZoning)[T.RM]
8	1.351019	C(PoolArea)[T.480]
9	1.105159	C(PoolArea)[T.512]
10	1.013351	C(PoolArea)[T.519]
11	1.042531	C(PoolArea)[T.555]
12	1.025214	C(PoolArea)[T.576]
13	1.009178	C(PoolArea)[T.648]
14	1.035931	C(PoolArea)[T.738]
15	1.978522	C(BsmtFullBath)[T.1]
16	1.348269	C(BsmtFullBath)[T.2]
17	1.042365	C(BsmtFullBath)[T.3]
18	6.142806	C(OverallCond)[T.2]
19	26.620396	C(OverallCond)[T.3]
20	58.314028	C(OverallCond)[T.4]
21	380.296848	C(OverallCond)[T.5]
22	220.683393	C(OverallCond)[T.6]
23	186.412526	C(OverallCond)[T.7]
24	73.273448	C(OverallCond)[T.8]
25	24.128636	C(OverallCond)[T.9]
26	1.367123	C(MSSubClass_linear)[T.1]
27	1.204138	C(MSSubClass_linear)[T.2]
28	1.571012	C(MSSubClass_linear)[T.3]
29	2.466565	C(MSSubClass_linear)[T.4]
30	2.953417	C(MSSubClass_linear)[T.5]
31	3.410592	C(MSSubClass_linear)[T.6]
32	1.525572	C(MSSubClass_linear)[T.7]
33	1.095524	C(MSSubClass_linear)[T.8]
34	2.288559	C(MSSubClass_linear)[T.9]
35	2.488687	C(MSSubClass_linear)[T.10]
36	8.932346	C(MSSubClass_linear)[T.11]
37	1.557142	C(MSSubClass_linear)[T.12]
38	3.301544	C(MSSubClass_linear)[T.13]
39	9.643356	C(MSSubClass_linear)[T.14]
40	2.372420	Q("BsmtUnfSF")
41	1.495601	Q("Fireplaces")
42	4.453882	Q("TotalBsmtSF")
43	6.952324	Q("YearBuilt")
44	2.039073	Q("GarageCars")
45	3.311815	Q("Neighborhood_linear")
46	4.254945	Q("GrLivArea")
47	3.910792	Q("OverallQual")

Figure 9: VIF of final model

After confirming that none of our variables have serious multicollinearity, we move on to removing influential points. This is in an effort to improve our model's predictions. The below model shows the result of removing all influential points, as indicated by their externally studentized residual and cook's distance.

OLS Regression Results

Dep. Variable:	np.log(SalePrice)	**R-squared:**	0.940
Model:	OLS	**Adj. R-squared:**	0.938
Method:	Least Squares	**F-statistic:**	476.4
Date:	Wed, 02 Dec 2020	**Prob (F-statistic):**	0.00
Time:	14:55:21	**Log-Likelihood:**	1339.1
No. Observations:	1412	**AIC:**	-2586.
Df Residuals:	1366	**BIC:**	-2345.
Df Model:	45		
Covariance Type:	nonrobust		

Omnibus:	13.609	**Durbin-Watson:**	1.977
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	19.495
Skew:	-0.078	**Prob(JB):**	5.85e-05
Kurtosis:	3.554	**Cond. No.**	4.99e+05

Figure 10: Final regression results

At this point, we can see that our model without influential points has improved upon the previous model's adjusted r-squared, skewness, kurtosis, and almost every other metric. Note that removing the outliers keeps our least-squares method unbiased. As mentioned previously, we used a model that fits the log of the SalePrice to the data to deal heteroskedascity.

Best Subsets Analysis

With our initial investigation done we want to extend our model to include all of the possible predictors and to use a best subsets analysis to identify which predictors will give us the best fit. We run a script to calculate the adjusted R^2 and Mallow's C_p value for all combinations of predictors and sort to find the most promising candidates. At this point, we then get the AIC and BIC of each promising candidate and compare for the one that minimizes AIC. We notice that our best model, which minimizes both AIC and BIC is our full model.

Model Evaluation

To evaluate our model, beyond the adjusted r-squared given by the regression results, we decided to use Kaggle's complimentary testing dataset for predicting SalePrice on the housing data. After submitting our model's results, we got a root mean-squared error of 0.14791. Note that we had to create a new model for instances in which our model returned NaN. This would occur because our model would use a variable that the row in the testing dataset would not have.

Metric

Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

2746	**Nicolas Decavel**	0.14791	3	2d

Figure 11: Results from Kaggle

Model Extension

Since our dataset showed considerable heteroscedasticity, we decided to try a robust regression. To do so we specify the categorical variables and create a new dataset using one hot encoding. For this analysis we will not remove any influential points. Now we can train the Theil-Sen Regressor and measure it's performance. Using a Thiel-Sen regression on the preliminary model that we used with 7 predictors, we get an adjusted R^2 of 0.802 which is a decrease from the value we got using linear regression, 0.815. This is not unexpected since Thiel Sen does not weigh the outliers as heavily as in linear regression. Thus, when calculating the adjusted R^2 for the training data set, the outliers will have a larger effect on the Thiel-Sen errors and it will appear that the model performs worse. However, on a different test data set where the outliers are not present, we expect that the Thiel-Sen model may outperform the linear regression. We cannot directly calculate R^2 since the SalePrice data is not directly available but we can submit both models to Kaggle to compare performance. The Thiel-Sen model has a performance score of 0.17110 and the linear regression model has a score of 0.20404 which means that the Thiel-Sen model is outperforming the linear model when influential points are included.

Summary

In this analysis we began with a simple linear model using several quantitative and categorical variable and saw good performance in predicting the sale prices of homes. By extending the model to include all possible predictors and using the algorithm we described above we further increased our model performance while dealing with influential points, multicollinearity, and heteroscedasticity. This model had a high adjusted-R2 and the lowest AIC and BIC among all our candidate models. In the future, we can continue improving our model performance by applying other Machine Learning techniques to further improve our testing accuracy.

Appendices

Appendix 1: Data Description

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet

- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

Appendix 2: Anova Table For Preliminary Model

	df float64	sum_sq float64	mean_sq float64	F float64	PR(>F) float64
C(Neighborhood)	24	4860179279058.795	202507469960.7831	173.9715238859436	0
C(ExterQual)	3	862628886716.1084	287542962238.7028	247.02440523815244	1.9897759515752585e-128
C(BsmtQual)	3	377423724650.116	125807908216.70534	108.07993163709818	6.883638858836087e-63
TotRmsAbvGrd	1	618989269077.4692	618989269077.4692	531.7656007025712	7.122856773030337e-100
GrLivArea	1	409860687854.75055	409860687854.75055	352.10596656429766	3.428874351999046e-70
OverallCond	1	77442853423.30031	77442853423.30031	66.5301444274447	7.637411267815526e-16
OverallQual	1	139790012933.71204	139790012933.71204	120.09177527536774	7.345161737729388e-27
Residual	1388	1615668829169.102	1164026533.9835029	nan	nan

Appendix 3: Regression for Final Model

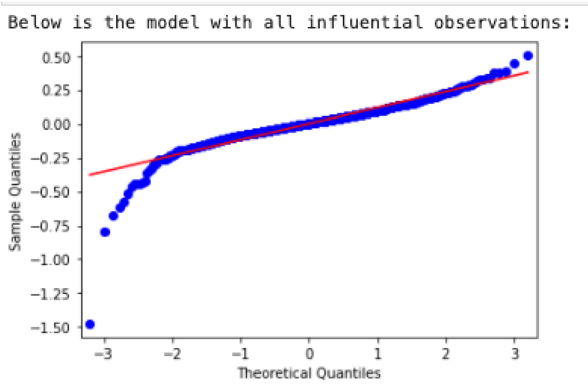
OLS Regression Results

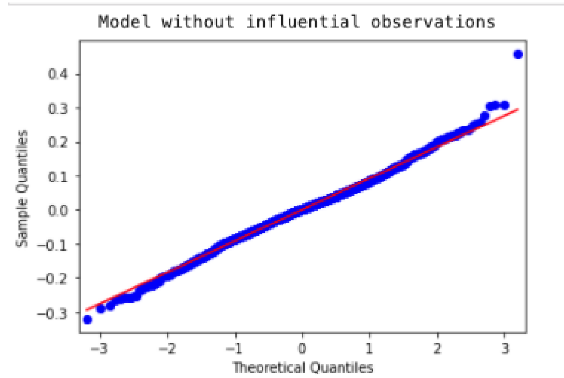
Dep. Variable:	np.log(SalePrice)	**R-squared:**	0.943
Model:	OLS	**Adj. R-squared:**	0.941
Method:	Least Squares	**F-statistic:**	479.5
Date:	Wed, 02 Dec 2020	**Prob (F-statistic):**	0.00
Time:	19:37:29	**Log-Likelihood:**	1372.9
No. Observations:	1411	**AIC:**	-2650.
Df Residuals:	1363	**BIC:**	-2398.
Df Model:	47		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3677	0.438	3.120	0.002	0.508	2.228
C(KitchenQual)[T.Fa]	-0.1007	0.023	-4.459	0.000	-0.145	-0.056
C(KitchenQual)[T.Gd]	-0.0805	0.012	-6.962	0.000	-0.103	-0.058
C(KitchenQual)[T.TA]	-0.0871	0.013	-6.478	0.000	-0.113	-0.061
C(Functional)[T.Maj2]	-0.1754	0.052	-3.390	0.001	-0.277	-0.074
C(Functional)[T.Min1]	0.0453	0.033	1.378	0.168	-0.019	0.110
C(Functional)[T.Min2]	0.0724	0.033	2.227	0.026	0.009	0.136
C(Functional)[T.Mod]	-0.0052	0.042	-0.123	0.902	-0.088	0.078
C(Functional)[T.Sev]	-0.4703	0.104	-4.534	0.000	-0.674	-0.267
C(Functional)[T.Typ]	0.0976	0.028	3.475	0.001	0.043	0.153
C(BldgType)[T.2fmCon]	-0.0207	0.019	-1.101	0.271	-0.057	0.016
C(BldgType)[T.Duplex]	-0.1113	0.016	-7.077	0.000	-0.142	-0.080
C(BldgType)[T.Twnhs]	-0.1257	0.016	-7.968	0.000	-0.157	-0.095
C(BldgType)[T.TwnhsE]	-0.0519	0.010	-4.981	0.000	-0.072	-0.031
C(MSZoning)[T.FV]	0.4349	0.057	7.671	0.000	0.324	0.546
C(MSZoning)[T.RH]	0.3373	0.060	5.592	0.000	0.219	0.456
C(MSZoning)[T.RL]	0.3962	0.055	7.184	0.000	0.288	0.504
C(MSZoning)[T.RM]	0.3589	0.055	6.499	0.000	0.251	0.467
C(RoofMatl)[T.CompShg]	2.6762	0.110	24.378	0.000	2.461	2.892
C(RoofMatl)[T.Membran]	2.8604	0.149	19.189	0.000	2.568	3.153
C(RoofMatl)[T.Metal]	2.7893	0.146	19.127	0.000	2.503	3.075
C(RoofMatl)[T.Roll]	2.7036	0.145	18.627	0.000	2.419	2.988
C(RoofMatl)[T.Tar&Grv]	2.7564	0.115	23.988	0.000	2.531	2.982
C(RoofMatl)[T.WdShake]	2.6168	0.116	22.540	0.000	2.389	2.844
C(RoofMatl)[T.WdShngl]	2.7442	0.114	24.167	0.000	2.521	2.967
C(OverallCond)[T.2]	0.0448	0.114	0.393	0.694	-0.178	0.268
C(OverallCond)[T.3]	-0.1408	0.100	-1.404	0.160	-0.337	0.056
C(OverallCond)[T.4]	0.0120	0.100	0.120	0.904	-0.184	0.208
C(OverallCond)[T.5]	0.0708	0.100	0.710	0.478	-0.125	0.266
C(OverallCond)[T.6]	0.1054	0.100	1.057	0.291	-0.090	0.301
C(OverallCond)[T.7]	0.1467	0.100	1.472	0.141	-0.049	0.342
C(OverallCond)[T.8]	0.1560	0.100	1.555	0.120	-0.041	0.353
C(OverallCond)[T.9]	0.2004	0.102	1.960	0.050	-0.000	0.401
C(GarageCars)[T.1]	0.0894	0.013	6.791	0.000	0.064	0.115
C(GarageCars)[T.2]	0.1305	0.013	9.784	0.000	0.104	0.157
C(GarageCars)[T.3]	0.1983	0.017	11.665	0.000	0.165	0.232
C(GarageCars)[T.4]	0.2550	0.044	5.789	0.000	0.169	0.341
C(BsmtFullBath)[T.1]	0.0219	0.007	3.118	0.002	0.008	0.036
C(BsmtFullBath)[T.2]	0.1501	0.030	4.963	0.000	0.091	0.209
C(BsmtFullBath)[T.3]	0.4389	0.095	4.603	0.000	0.252	0.626
Q("YearRemodAdd")	0.0013	0.000	6.534	0.000	0.001	0.002
Q("BsmtUnfSF")	-6.429e-05	8.63e-06	-7.449	0.000	-8.12e-05	-4.74e-05
Q("Fireplaces")	0.0402	0.005	8.492	0.000	0.031	0.049
Q("TotalBsmtSF")	0.0002	9.85e-06	18.782	0.000	0.000	0.000
Q("YearBuilt")	0.0019	0.000	10.942	0.000	0.002	0.002
Q("Neighborhood_linear")	0.0069	0.001	9.875	0.000	0.006	0.008
Q("GrLivArea")	0.0003	7.53e-06	35.446	0.000	0.000	0.000
Q("OverallQual")	0.0547	0.004	15.494	0.000	0.048	0.062

Omnibus:	8.786	**Durbin-Watson:**	1.964
Prob(Omnibus):	0.012	**Jarque-Bera (JB):**	11.591
Skew:	-0.047	**Prob(JB):**	0.00304
Kurtosis:	3.434	**Cond. No.**	6.33e+05

Appendix 4: QQ plots with and without influential pointsl





Work Distribution

Group Members	Nicolas	Jiahui	Teddy
Proportion of Work	1/3	1/3	1/3
List of Work	Discussion of EDA Feature engineering Modeling script prioritizing adjusted r-squared Discussion of Modeling Model selection and verification Discussion of model selection results Forecasting and analyzing the forecasting results Write the report	Discussion of EDA Discussion of Modeling Discussion of model selection and verification Model Diagnostics Discussion of model selection results Write the report	EDA and initial model building Model Diagnostics on preliminary model Model Extension Discussion of EDA Description of Data Discussion of Model Extension Wrote/ formatted the report into latex