Ames Housing OLS Regression Project

Introduction

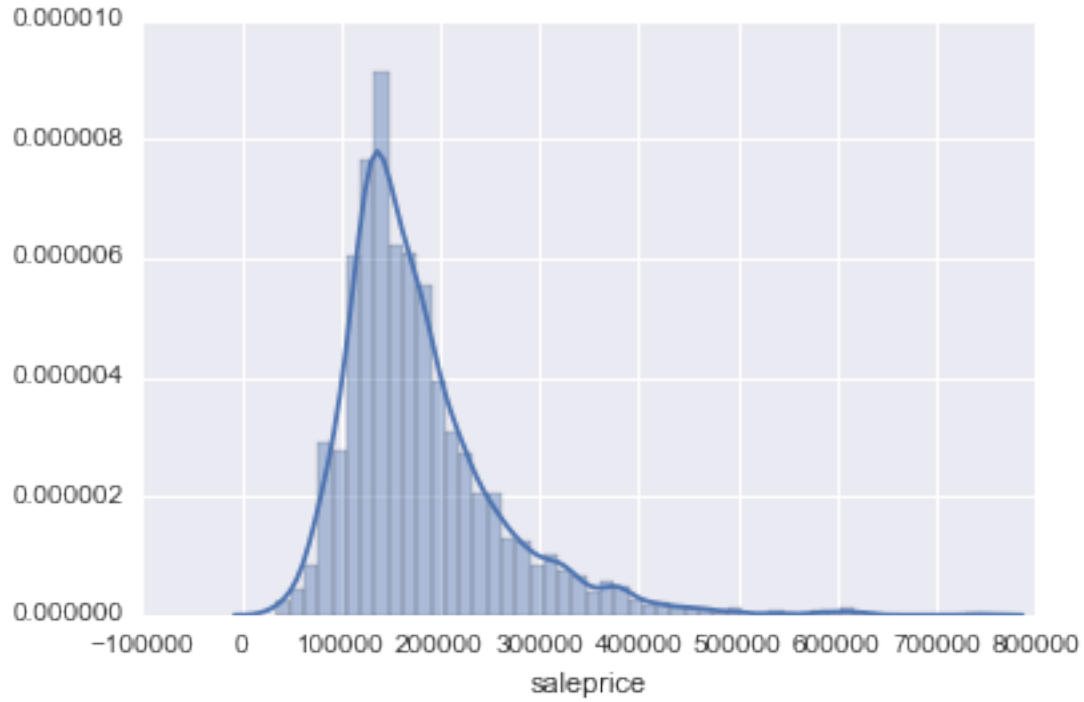
In this document we explore regression techniques to come up with a prediction for the sale price of single family homes in Ames, IOWA using Ames Housing Dataset. This report considered the variables that are commonly used when consumers looking to buy a home, such as living area size, whether home equipped with the basement. The exploratory data analysis that is performed as part of this assignment also validated that these variables are highly correlated with the sale price. This analysis found that the total living area above grade, year built, and mean price per square foot by neighborhood have a positive correlation with sale price. Both simple and multiple regressions models are considered as part of the analysis, and performed a log transformation of the response variable. A multiple linear regression with a log transformation of SalePrice produced the best model for predicting SalePrice. Further exploration is required to consider the impact of the Neighborhood and also the elimination of outliers could improve the future models.

Data Exploration

The dataset explored in this analysis contains information from the Ames Assessor's Office and contains about 80 variables that were directly related to the property sales. Without going deep into details about each and every variable, that these variables focus on the quality and the quantity of many physical attributes of the house. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property, e.g. how many square feet of living space in the dwelling? Because the dataset includes wide range of properties, subset of properties is defined to focus on when predicting a sale price. For example, model to predict the commercial property needs to consider different variables and techniques than the single family house.

In this analysis we are interested in the sale price of a typical single family residence sold in Ames, Iowa. The Exploratory data analysis in this documents considers the SalePrice as the response variable, which indicates the sale price of the property. As we are trying to predict the SalePrice, let's look at the descriptive statistics summary and the histogram of the variable.

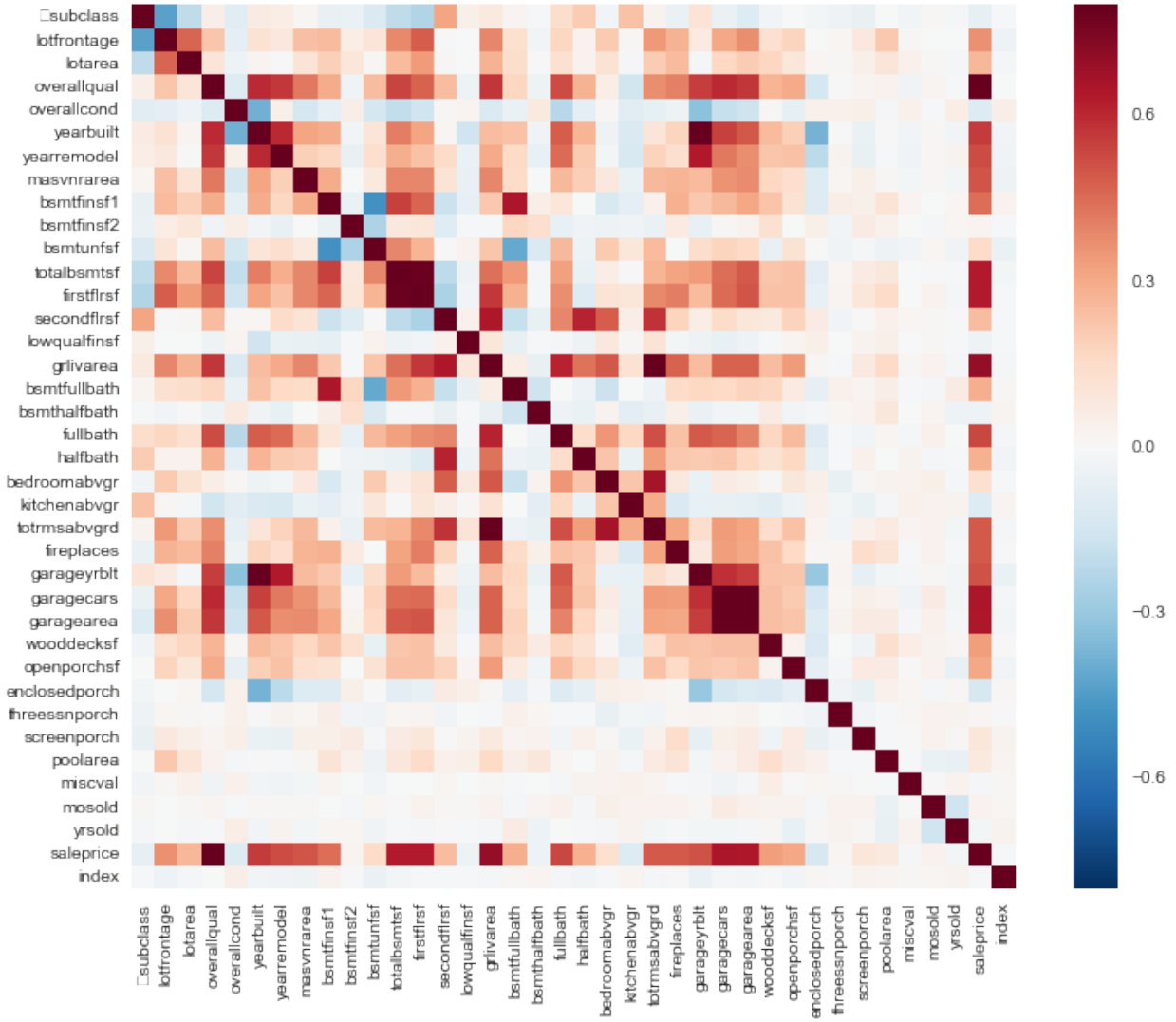
```
count      2039.000000
mean     179368.827857
std       78982.943661
min       34900.000000
25%      128500.000000
50%      160000.000000
75%      210125.000000
max      745000.000000
Name: saleprice, dtype: float64
```



Because the minimum value is greater than zero, we can be confident enough to build the model upon. From the histogram, SalePrice data Deviate from the normal distribution and has appreciable positive skewness and it also Show peakedness.

Next, let's look at the correlation between the SalePrice and the other variables. Heatmap is the best way to do this.

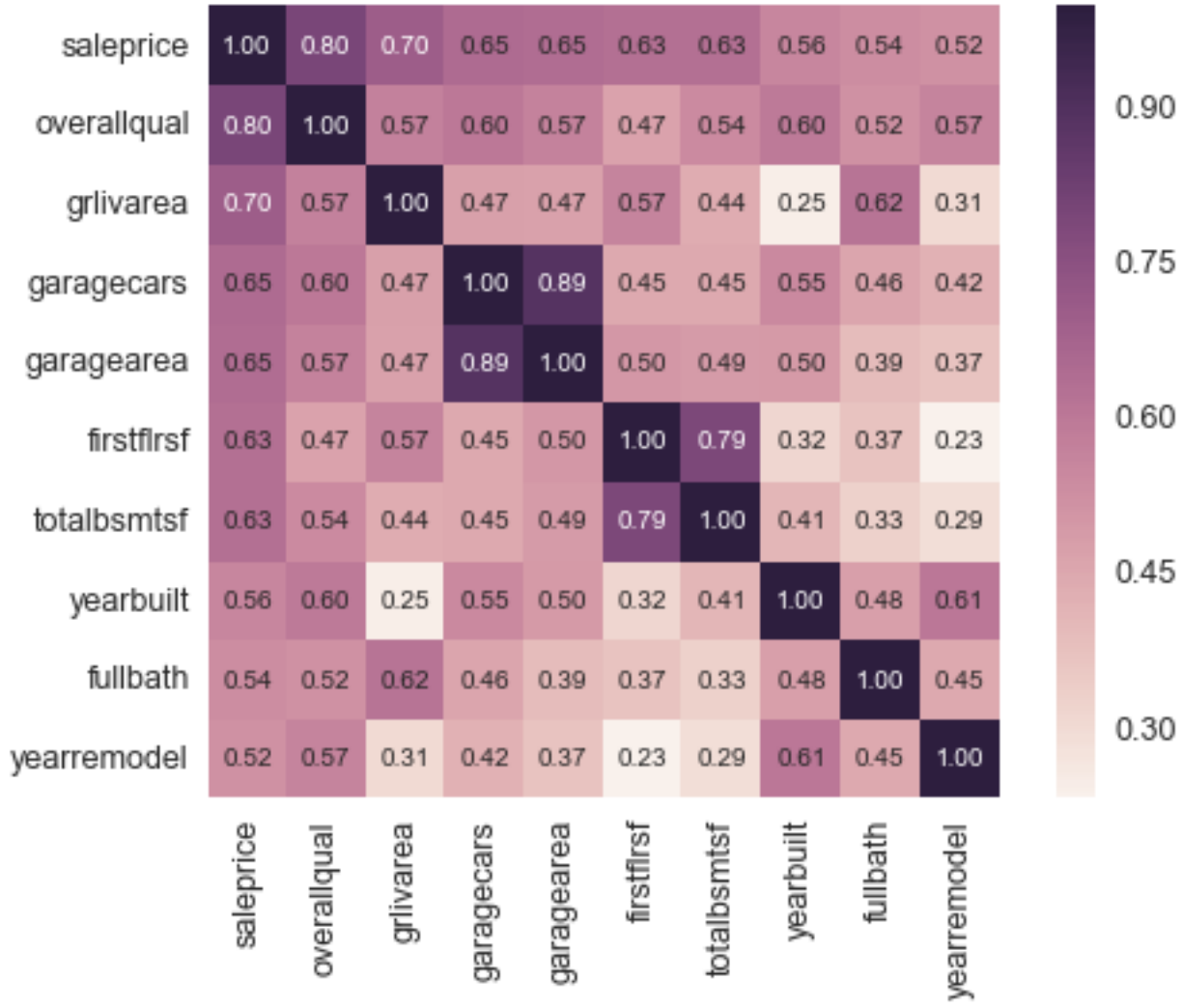
Figure 1: Correlation matrix



At first glance, there are two red colored squares (2x2) that are interesting. The first one refers to the 'TotalBsmtSF' and '1stFlrSF' variables, and the second one refers to the 'Garage*Cars*' and GarageArea variables. Both cases show how significant the correlation is between these variables. From this we can conclude that they give out same information and it is a case of multicollinearity. From the same heatmap, we can derive that 'GrLivArea', 'TotalBsmtSF', and 'OverallQual' are highly correlated with the SalePrice. Also there are some other variables that should be taken into account.

Next, we will take a look at the variables that are most correlated with the SalePrice.

Figure 2: 'SalePrice' correlation matrix

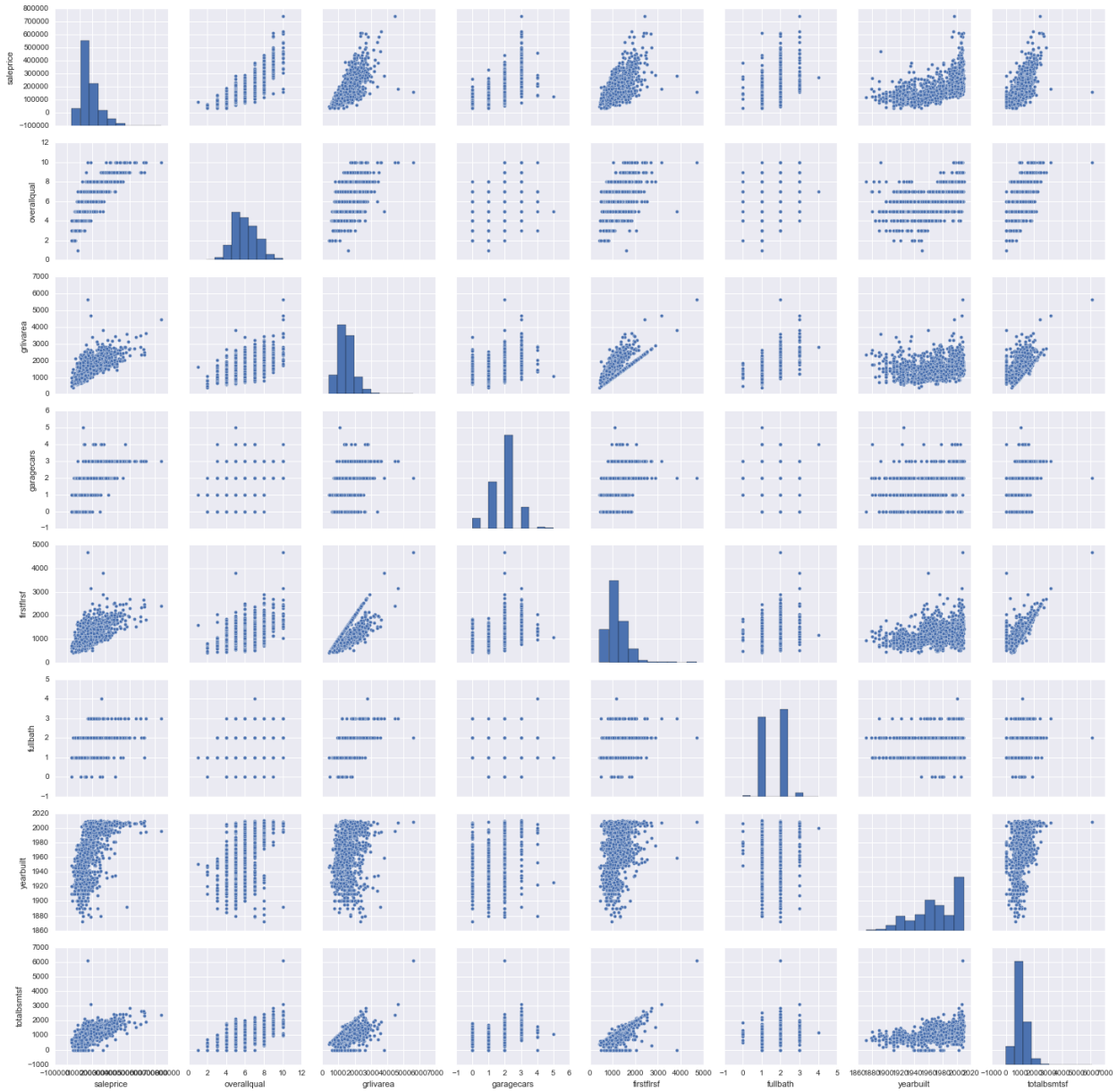


We can consider the following observations from the above heatmap.

- 'OverallQual', 'GrLivArea' and 'TotalBsmtSF' are strongly correlated with 'SalePrice'.
- GarageCars' and 'GarageArea' are also some of the most strongly correlated variables, this is because the number of cars fit in the garage is a direct consequence of the garage area. We just need one of these variables as they both provide same information.
- Also there is a strong relationship between the FirstFlrSF and the TotalBsmtSF, so we are going to consider just the TotalBsmtSF.
- And the final one we are going to consider is the 'YearBuilt' as it has stronger linear relationship with SalePrice than the rest.

Also from the Scatter plots shown below, we can observe an interesting fact about the SalePrice and the YearBuilt. The prices are going up and stays above the limit as the year's progresses.

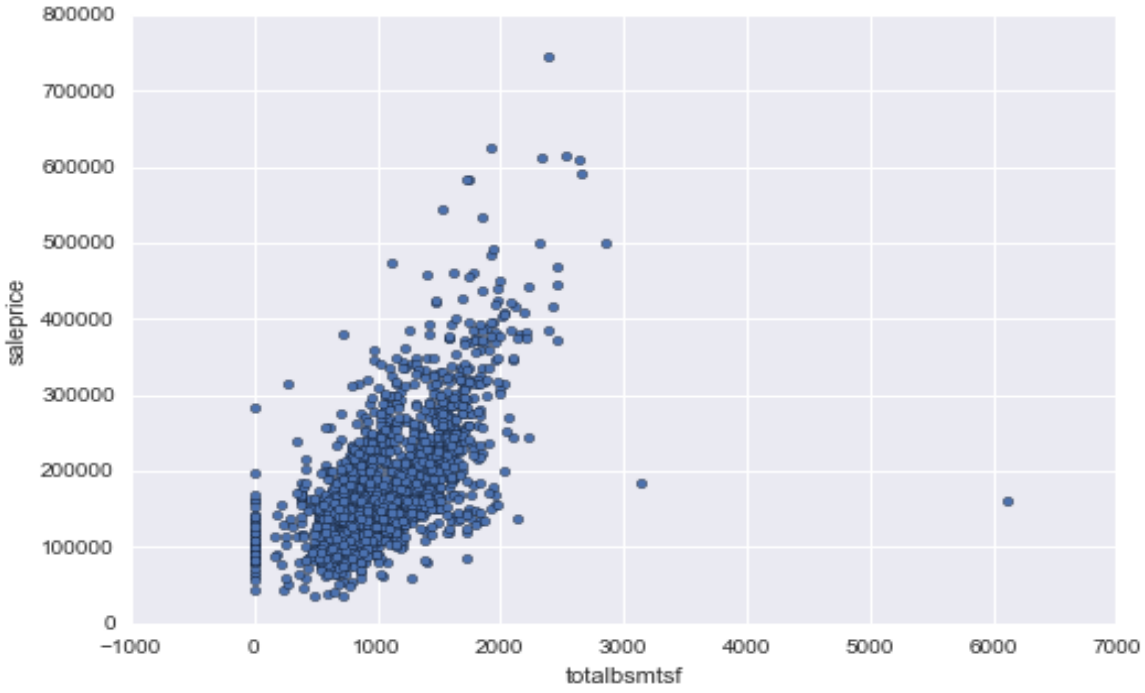
Figure 3: Scatter plot between 'SalePrice' and Correlated vriables



Now let's look at the one on one relationship between SalePrice and the variables discussed above.

The scatter plot shown below shows the strong linear relationship between 'TotalBsmtSF' and the "SalePrice". Also we notice that there are lot of missing values that are set to zero.

Figure 4: Scatter Plot between 'SalePrice' and 'TotalBsmtSF'



Next, let's look at the scatter plot between SalePrice and the GrLivArea. As expected they have the strong linear relationship. Also there are couple of outliers that we need to consider during the data preparation.

Figure 5: Scatter Plot between 'SalePrice' and 'GrlivArea'

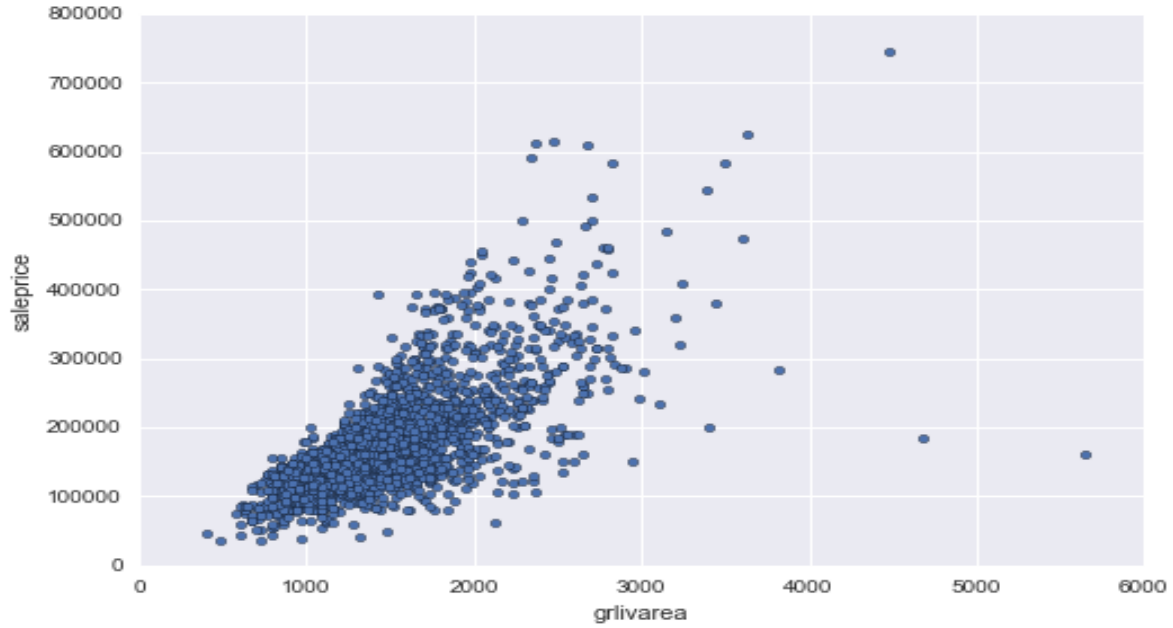


Figure 6: Box Plot between 'OverallQual' and 'SalePrice'

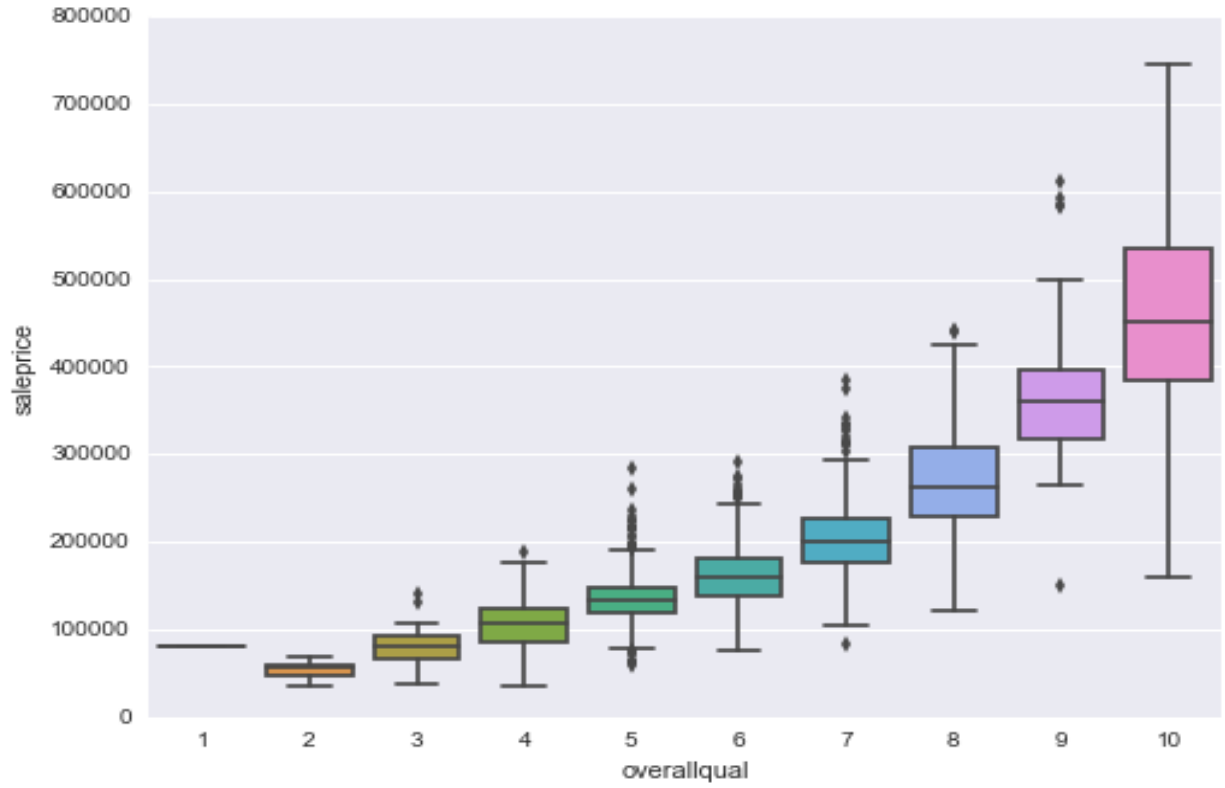


Figure 6. shows the linear relationship between the 'SalePrice' and the 'OverallQual'. One observation to note here is that prices stay up as the year's progress.

Data Preparation

Missing Data:

Before building the model on the dataset, first we need to make sure that data is prepared for it. First thing we look at the missing data within the dataset. It is important that we ensure that missing data process is not biased and hiding any important details. As part of the process, any variable that has more than 15% missing data will be dropped from the dataset. Filling up these huge chunks of missing data can lead to the wrong projections.

From the table below, we see that 'PoolQC', 'MiscFeature', 'Alley', 'Fence', 'FireplaceQU', and 'LotFrontAge' are missing majority data. From the Exploratory data analysis we performed earlier, none of these variables have stronger linearity with the SalePrice. So dropping these variables should not affect our model. Moreover, these variables are stronger candidates for the outliers. Next, all the 'Garage' related variables seem to be missing same amount of data and it is possible that all these refer to the same set of observations. Since most important information regarding the Garages is expressed by 'GarageCars', and as we discussed in the previous section that 'GarageCars' (also as shown in ***Figure 2)*** has strong linear relationship with the 'SalePrice', we can drop rest of these 'Garage' related variables and just consider the 'GarageCars' for our model. We will apply the same principle to the rest of 'bsmt' related variables. As they have just over 2% missing data, we will consider the 'TotalBsmtSF' for our model and drop the rest.

Remaining two variables 'MasVnrArea' and 'MasVnrType', they have strong correlation with 'YearBuilt' and 'OveralQual' (as shown in ***Figure 2)***. Thus, we will not lose information if we delete these two variables.

	Total	Percent
poolqc	2028	0.994605
miscfeature	1963	0.962727
alley	1893	0.928396
fence	1639	0.803825
fireplacequ	1012	0.496322
lotfrontage	325	0.159392
garagecond	119	0.058362

garageyrblt	119	0.058362
garagefinish	119	0.058362
garagequal	119	0.058362
garagetype	118	0.057872
bsmtexposure	56	0.027464
bsmtfintype2	54	0.026484
bsmtcond	54	0.026484
bsmtqual	54	0.026484
bsmtfintype1	54	0.026484
masvnrarea	15	0.007357
masvnrtype	15	0.007357

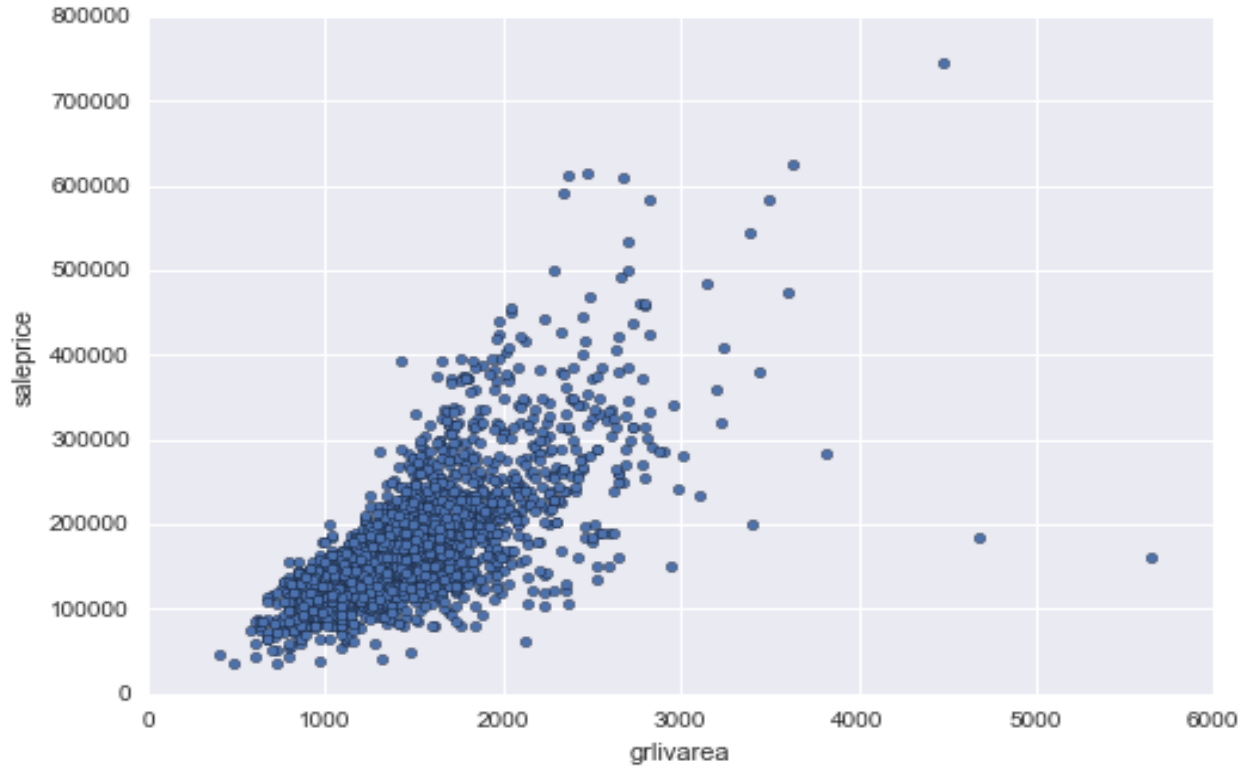
Outliers:

Next we will standardize the data to detect the outliers, meaning will convert the data values to have mean of 0 and standard deviation of 1. Lower range of the distribution have similar values and closer to 0. However, higher range of values are far from 0 and we definitely need to looking into the values that resulting in 7.

```
outer range (low) of the distribution:
[[-1.83009867]
 [-1.8248942 ]
 [-1.79210984]
 [-1.76551765]
 [-1.71486587]
 [-1.70220292]
 [-1.6832085 ]
 [-1.63888819]
 [-1.6135623 ]
 [-1.60723083]]

outer range (high) of the distribution:
[[ 4.50264057]
 [ 4.63210653]
 [ 5.10961356]
 [ 5.12945639]
 [ 5.21919869]
 [ 5.45236151]
 [ 5.47334401]
 [ 5.51567624]
 [ 5.6423057 ]
 [ 7.16185921]]
```

Figure 7: Scatter plot between 'GrLivArea' and 'SalePrice'



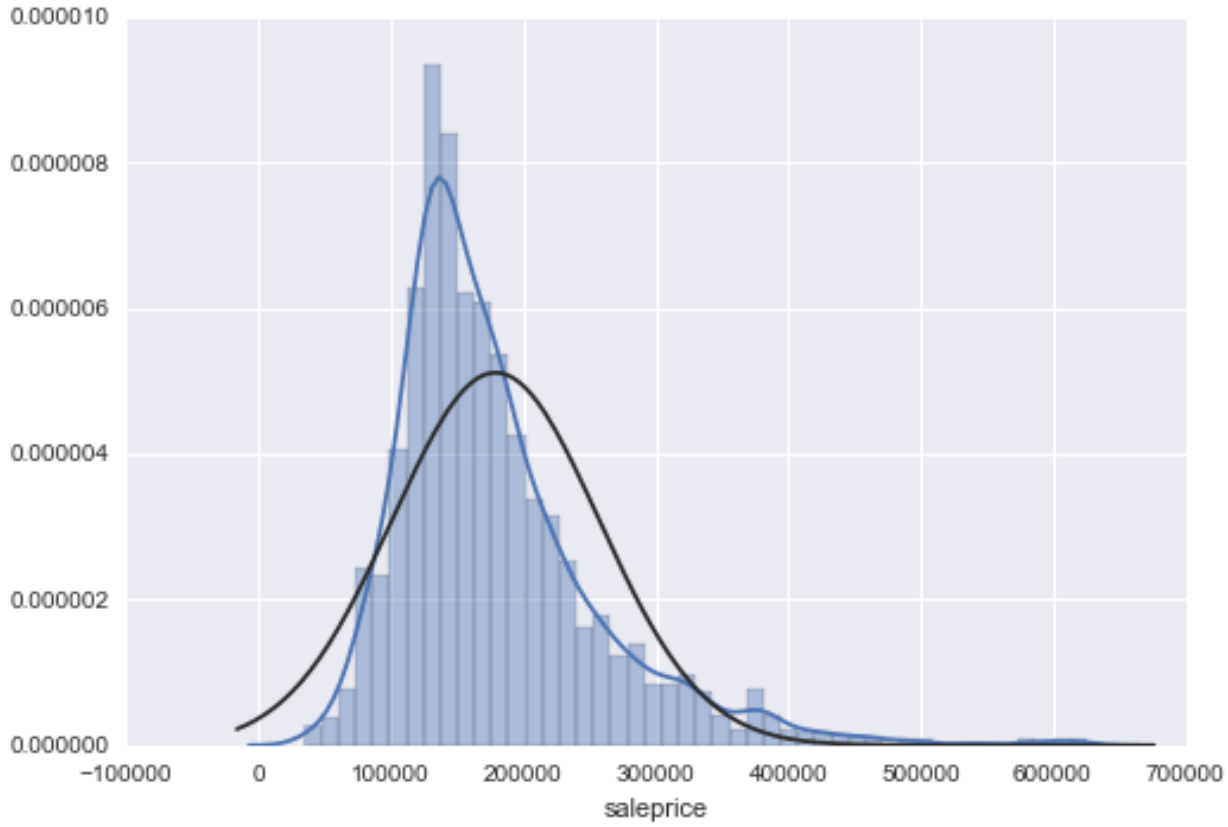
From the above scatter plot, we can see that there are two values with bigger 'GrLivArea' seems to be out of range. At least they are not following the trend here as they have high value of 'GrLivArea' but low in the price range. So we call these two as outliers and delete them. The observation in the top of the plot (750k+) seem to be following the trend, so we will keep that one.

Transformation:

The histogram shown below (Figure 8) for the 'SalePrice' shows that the distribution is not normal. It shows that there is peakedness and also there is a positive skewness in the distribution. A simple log transformation can solve this problem. We will follow the same guideline to 'GrLivArea' and 'TotalBsmtSF'.

One thing note here is that, as we seen in Figure 4, 'TotalBsmtSF' has lot of 0 values in its data. So simple log transformation does not work on this variable. We will have to divide this data as the ones with 'has basement' and the other category without. And then we apply the log transformation on the data that has the basement.

Figure 8: Histogram for 'SalePrice'



Models

Four different models are fitted to come up with an appropriate solution for predicting the 'SalePrice'. Both simple linear regression and the multiple linear regression techiniques are applied.

Model 1: GrLivArea

The results from OLS regression on 'GrLivArea' and 'SalePrice'. The simple linear regression model has an r-squared value of 0.518, which suggests that 'GrLivArea' has the positive correlation with 'SalePrice'. We can reject the null hypothesis that there is no relationship between GrLivArea and SalePrice.

OLS Regression Results

Dep. Variable:	saleprice	R-squared:	0.518
Model:	OLS	Adj. R-squared:	0.518
Method:	Least Squares	F-statistic:	2190.
Date:	Sun, 01 Oct 2017	Prob (F-statistic):	0.00
Time:	18:16:12	Log-Likelihood:	-273.21
No. Observations:	2036	AIC:	550.4
Df Residuals:	2034	BIC:	561.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.5039	0.139	39.522	0.000	5.231	5.777
grlivarea	0.8979	0.019	46.795	0.000	0.860	0.935

Model 2: YearBuilt

This model was performed on 'YearBuilt'. The simple linear regression model has r-squared 0.385, which suggests YearBuilt has a positive correlation with SalePrice, though not very strong.

OLS Regression Results

Dep. Variable:	saleprice	R-squared:	0.385
Model:	OLS	Adj. R-squared:	0.385
Method:	Least Squares	F-statistic:	1273.
Date:	Sun, 01 Oct 2017	Prob (F-statistic):	6.05e-217
Time:	18:20:52	Log-Likelihood:	-522.31
No. Observations:	2036	AIC:	1049.
Df Residuals:	2034	BIC:	1060.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.9981	0.449	-8.908	0.000	-4.878	-3.118
yearbuilt	0.0081	0.000	35.679	0.000	0.008	0.009

The linear model (below) appears to have a strong linear relationship with SalePrice, with most values clustered around the model. There seem to be larger residuals after 1980.

Model 3: Multiple Linear Regression – GrLivArea and YearBuilt

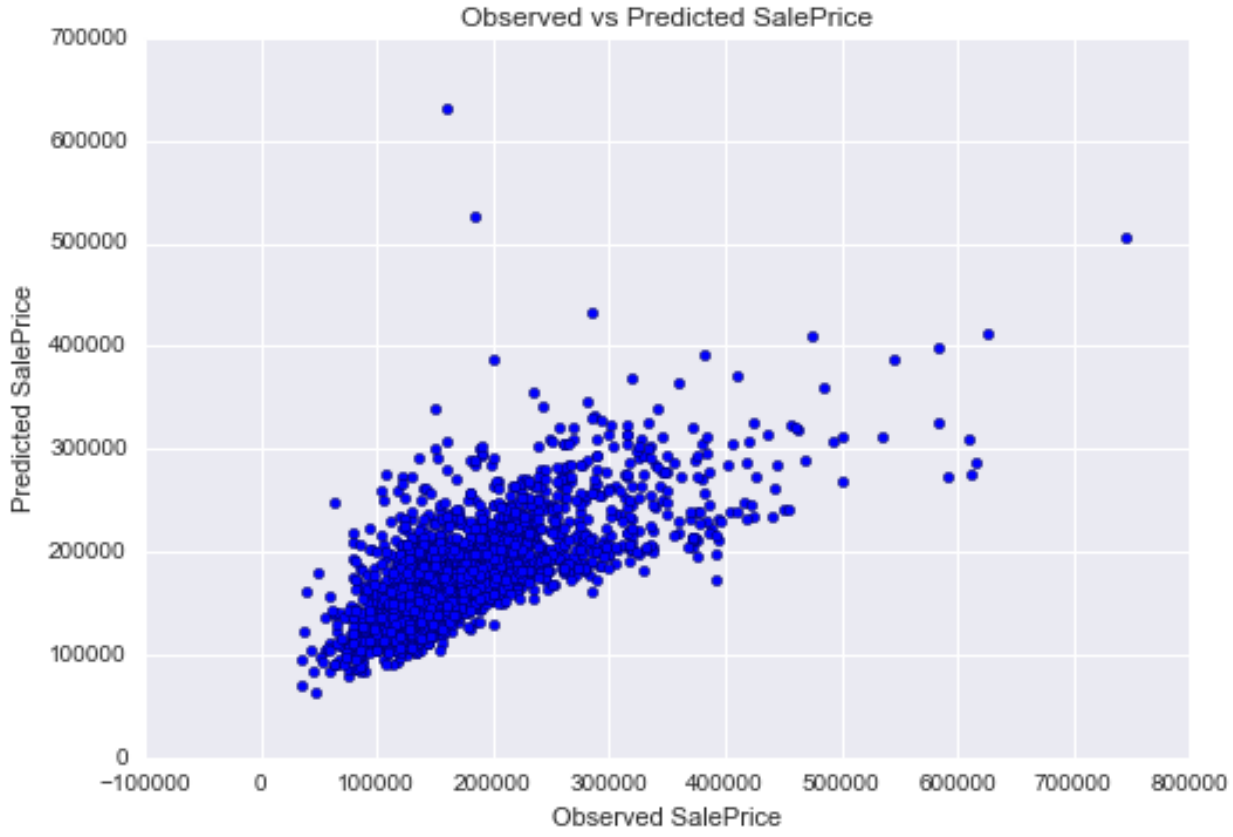
As part of third model, both 'GrLivArea' and 'YearBuilt' are considered for multiple linear regression. The multiple linear regression model has a r-squared of 0.918, which suggests there is a strong positive correlation between GrLivArea and YearBuilt to SalePrice. This correlation coefficient is higher than that of the simple linear models of GrLivArea to SalePrice or YearBuilt to SalePrice, but this is expected because the correlation coefficient usually gets larger when more variables are included. For this reason, we look at the adjusted r-squared value which adjusts for the number of predictors in the model, however in this case both has the same value. Both coefficients have a p-value less than 0.05, which means we can reject the null hypothesis that there is no relationship between GrLivArea and YearBuilt and SalePrice.

OLS Regression Results

Dep. Variable:	saleprice	R-squared:	0.918
Model:	OLS	Adj. R-squared:	0.918
Method:	Least Squares	F-statistic:	1.140e+04
Date:	Sun, 01 Oct 2017	Prob (F-statistic):	0.00
Time:	18:40:33	Log-Likelihood:	-25190.
No. Observations:	2039	AIC:	5.038e+04
Df Residuals:	2037	BIC:	5.040e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
grlivarea	108.6333	2.531	42.923	0.000	103.670	113.597
yearbuilt	9.3833	2.011	4.666	0.000	5.440	13.327

Lastly, a plot of the observed vs predicted SalePrice values shows a stronger linear relationship than our simple linear models. The model is less accurate after a SalePrice of 400,000, which may be because we don't have enough observations that have a SalePrice above 400,000 to make an accurate prediction at that range.



Model 4: GrLivArea, GarageArea, YearBuilt, OverallQual and TotalBsmtSF

The final model was built using all the five variables that were highlighted as part of the EDA. R-squared value of 0.964 shows that linearity between all five variables and the SalePrice is strong.

Selected Model

Model 4 has chosen as the final product. This model producing the better adjusted r-squared value. Also P value and coefficients indicate that there is a strong correlation between all five variables and the SalePrice.

OLS Regression Results

Dep. Variable:	saleprice	R-squared:	0.964
Model:	OLS	Adj. R-squared:	0.964
Method:	Least Squares	F-statistic:	1.078e+04
Date:	Sun, 01 Oct 2017	Prob (F-statistic):	0.00
Time:	19:21:16	Log-Likelihood:	-24361.
No. Observations:	2039	AIC:	4.873e+04
Df Residuals:	2034	BIC:	4.876e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
grlivarea	45.4203	2.101	21.619	0.000	41.300	49.541
yearbuilt	-52.7102	1.967	-26.803	0.000	-56.567	-48.853
overallqual	2.487e+04	839.613	29.621	0.000	2.32e+04	2.65e+04
totalbsmtsf	33.6320	2.369	14.198	0.000	28.986	38.278
garagearea	63.2926	4.893	12.935	0.000	53.696	72.889

Model Formula

The following model formula is based on the selected model (Model 4).

```
p_saleprice = 45.420312 * grlivarea -
              52.710160 * yearbuilt +
              24869.822809 * overallqual +
              33.632031 * totalbsmtsf +
              63.292645 * garagearea
```

Conclusion

To summarize the findings GrLivArea, YearBuilt, GarageArea, OverallQual and TotalBsmtSF appear to have the positive relationship with SalePrice and we reject the null hypothesis that they have no relationship with the SalePrice. A log transformation of the response variable validated that the assumption of normally distributed residuals for linear regression. Our models consistently performed better at SalePrices below 300,000, and an expanded dataset of high SalePrices could lessen the heterodasticity present in many of our models. In general, there are improvements that can be made to fix slight imbalances or abnormal features in our data, including additional techniques to remove outliers, and future analysis should explore more regression techniques, including log transformations, to refine the model.