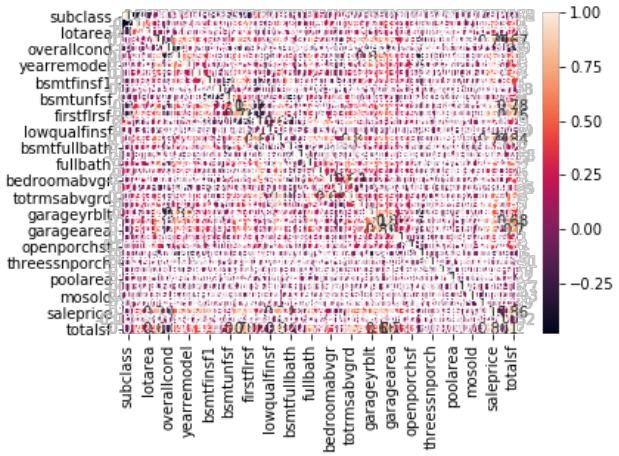
Introduction

The objective of this project is to use housing data to accurately predict the sell price of houses in the Ames, IA market. The Ames training data is a collection of 2039 records and 80 variables, not counting the index. The records are based on the sale of residential property from the years 2006 to 2010. Through Python, I will use Ordinary Least Squares Regression to build a model summary and create a formula. Kaggle will be used to judge the model's accuracy.

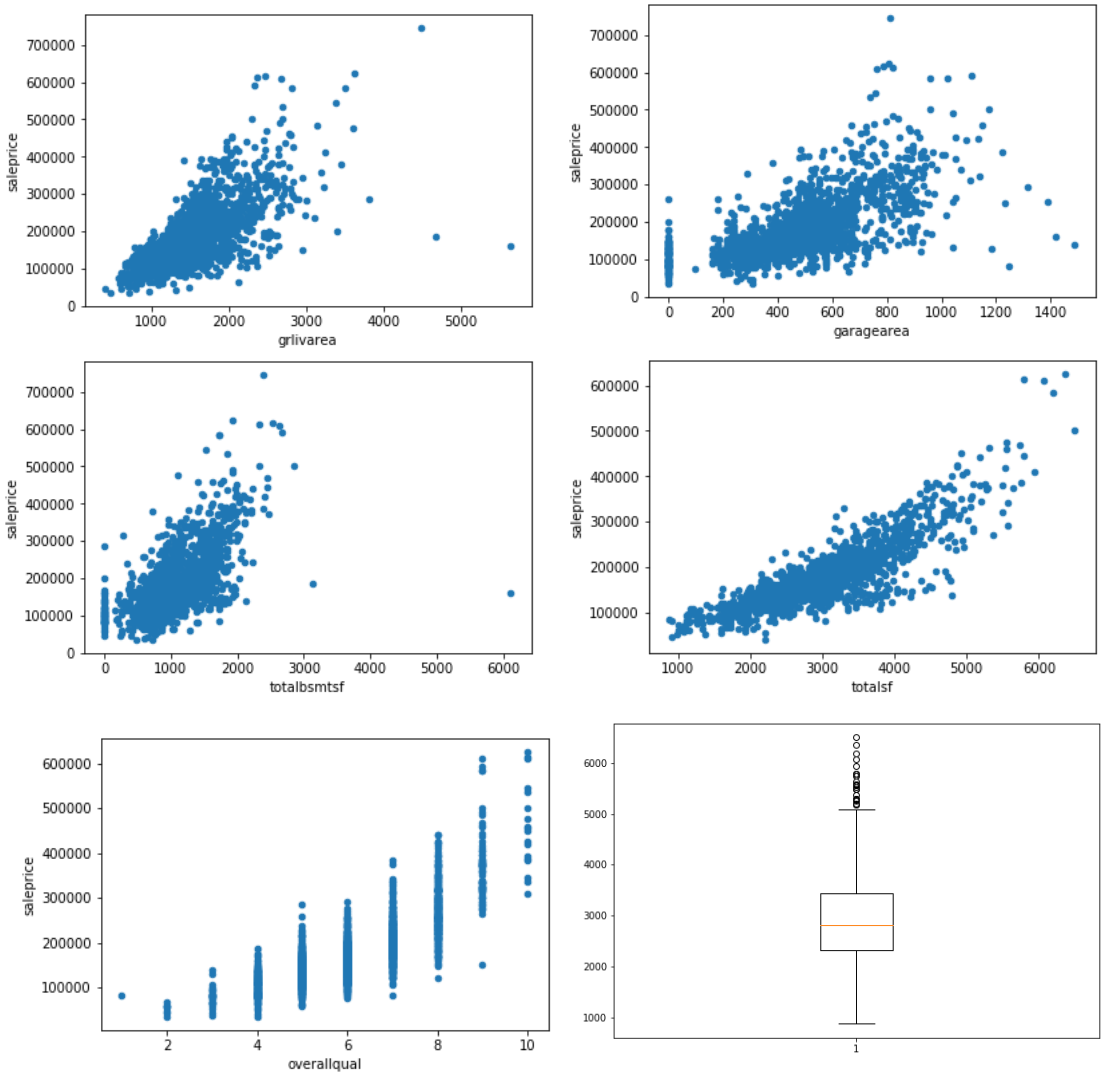
Data Exploration

Exploring the Ames, IA data set took a multitude of approaches in order to get a clear and accurate view of the dataset and what it provided. There were many different extensions of Python that were used in order to get a firm grasp on the intricacies of the set. Some of this included using the shape and describe functions. One of the most helpful parts of the data exploration was building a correlation coefficient table amongst all of the variables to see how they interacted between and among each other. The Seaborn heatmap that I built was also a very good indicator that reinforced the table of values. The gradient shows that the lighter the color gets the stronger the correlation coefficient.

```
saleprice      0.857188
totalsf        0.793854
overallqual    0.738296
grlivarea      0.638423
garagecars     0.637966
firstflrsf     0.636748
totalbsmtsf    0.628665
garagearea     0.857188
totalsf        0.793854
overallqual    0.738296
grlivarea      0.638423
garagecars     0.637966
firstflrsf     0.636748
totalbsmtsf    0.628665
```



I used Python to sort the values in a descending order to bring the highest correlated variables to the top. Above is a slice of the output that Python provided. I used this as my start to building my models. The variables with the highest correlation I used to build charts and then eventually the models. It was at this point in time when I decided to build a new variable called total square feet (totalsf). It is a combination of the grlivarea, totalbsmtsf and garage area variables (I also created Quality per Square foot and Price per Square foot variables for later analysis). These variables were all tested separately as well, but I combined them to one new variable due to their individual correlations to sales price and also R squared scores. Below are the charts individually and also the new variable called totalsf.



The above plots illustrated to me the potential for a linear regression model. I investigated dropping values that appeared to be outliers or missing values by using mean and median or filling in the cell with another calculation. The overallqual plot showed good potential but the R squared scores were not good enough to beat the other single variable regression tests or Kaggle. The box plot above is an example of one of the many ones I built to assess variable distribution. This figure is measuring the distribution of totalsf and I was able to see that most of the potential outliers appeared to be in the properties that had more than 5000 square feet. I used this as one of my tests in the model to improve record selection for the model. In most cases the outcome was negative in terms of correlation coefficient strength and R squared values. Despite this outcome, I still decided to try and submit the csv files to Kaggle for testing of the mean root square error.

Data Preparation

In order to prep the data for a proper investigation and build models a thorough look through the data had to be completed. I used a multitude of exploratory data analysis techniques to accomplish this task. They included looking at various plots and building tables to look at individual variable relationships and correlations. I also looked at the data in various Excel pivot tables. Some of these tables were outputs of csv files from my Jupyter Notebook. The preparation of the data, I believe, goes hand in hand with the exploration of the data. In order to build a clean model, I used my plots and generated tables to find missing values and potential outliers. I also had to adjust the chained assignment option within pandas to procure a table with missing values. I used that along with the mean and median operators to fill in 0 and null cells with a value. Ultimately, those both tended to make my models less accurate and did not improve the mean square error on Kaggle. I did end up using only the normal sale condition for the training dataset. This had a positive effect on my model my reducing the mean square error and also produced a much stronger R squared value. The reason for using the normal sale type was based on the data dictionary and I did not believe that foreclosures and condo prices would be beneficial to predicting the normal sale price of a home. This was also the point in project where I created my new variable totalsf and eliminated some of the totalsf values that were above the range and 7000 square feet. By doing these actions my model score improved in Kaggle. The last fix I did involved the test file. I had to fix a blank cell and put a 0 in for the garage area for a record. I used Python to correct for this. I noticed the error when the scoring file was not putting a value in the cell and Kaggle was not correcting for it.

Selected Models & Model Formula

OLS Regression Results for Sale Price by Total Square Foot

Dep. Variable:	saleprice	R-squared:	0.735
Model:	OLS	Adj. R-squared:	0.735
Method:	Least Squares	F-statistic:	4646.
Date:	Thu, 24 Jan 2019	Prob (F-statistic):	0.00
Time:	01:07:32	Log-Likelihood:	-20002.
No. Observations:	1679	AIC:	4.001e+04
Df Residuals:	1677	BIC:	4.002e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.147e+04	3139.715	-10.023	0.000	-3.76e+04	-2.53e+04
totalsf	70.1116	1.029	68.161	0.000	68.094	72.129

Omnibus:	181.702	Durbin-Watson:	1.968
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1324.506
Skew:	0.196	Prob(JB):	2.44e-288
Kurtosis:	7.334	Cond. No.	1.09e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.09e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

```
Intercept  -31468.319118
totalsf         70.111612
Length: 2, dtype: float64
```

The above model was the best one I built due to its' R squared score and how it scored the best on Kaggle. According to the output, 73% of the variation can be explained by the model. The formula for this model is $\hat{y} = -31{,}468.319 + 70.112x$. The predicted sale price equation has an intercept at -31,468.319 and a slope of 70.112 per square foot. This model also has a slight positive skew. I trust this model the most and like the fact that it takes all of the square feet together from different variables to account for the predicted rate. It takes from grlivarea,bsmtsqft and garage area. That helps limit the single variable risk that could exist.

```
            OLS Regression Results for Sale Price By Overall Quality
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.630
Model:                            OLS   Adj. R-squared:                  0.630
Method:                 Least Squares   F-statistic:                     2858.
Date:                Thu, 24 Jan 2019   Prob (F-statistic):               0.00
Time:                        01:37:10   Log-Likelihood:                -20281.
No. Observations:                1679   AIC:                         4.057e+04
Df Residuals:                    1677   BIC:                         4.058e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -7.552e+04   4780.874    -15.797      0.000   -8.49e+04   -6.61e+04
overallqual  4.172e+04    780.413     53.460      0.000    4.02e+04    4.33e+04
==============================================================================
Omnibus:                      454.796   Durbin-Watson:                   2.078
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2254.977
Skew:                           1.188   Prob(JB):                         0.00
Kurtosis:                       8.156   Cond. No.                         28.9
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Intercept      -75523.846702
overallqual     41720.600420
Length: 2, dtype: float64
```

The second model above was another one I selected due to its' R squared score and how it scored the on Kaggle as well. It is the second-place model for data set. According to the output, 63% of the variation can be explained by the model. The formula for this model is $\hat{y} = -75{,}523.847 + 41720.600x$. The predicted sale price equation has an intercept at -75,523.847 and a slope of 41720.600 overall quality unit. This model also has a positive skew as well. I really like this model but do have some concerns with it. There is an inherent risk to predicting significant negative equity/loss sales on lower quality homes. From a logical perspective this makes sense, but I don't see that being prevalent in the real world. I believe there might have to be additional conditions where some records that could fall into this category would need to be excluded from the prediction.

Non-Selected Models

The below are a sampling of some of the models I did not use. Most of them had very poor R squared scores and too much skew to the positive and negative to be used accurately to predict a value.

```
OLS Regression Results Sale Price by Year Sold
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.230
Date:                Thu, 24 Jan 2019   Prob (F-statistic):              0.268
Time:                        02:02:59   Log-Likelihood:                -25886.
No. Observations:                2039   AIC:                         5.178e+04
Df Residuals:                    2037   BIC:                         5.179e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    3.14e+06   2.67e+06      1.176      0.240    -2.1e+06    8.38e+06
yrsold     -1474.6609   1329.924     -1.109      0.268   -4082.814    1133.492
==============================================================================
Omnibus:                      760.842   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3302.569
Skew:                           1.764   Prob(JB):                         0.00
Kurtosis:                       8.141   Cond. No.                     3.07e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.07e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
Intercept    3.140216e+06
yrsold      -1.474661e+03
dtype: float64
```

The above model exhibited an incredible amount of skew and very low R squared scores. This meant that very little or none or the variability could be explained by the model. The plot for this also exhibited these characteristics.

```
                          OLS Regression Results Sales Price by masverarea
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.159
Model:                            OLS   Adj. R-squared:                  0.158
Method:                 Least Squares   F-statistic:                     381.2
Date:                Thu, 24 Jan 2019   Prob (F-statistic):           6.53e-78
Time:                        02:04:52   Log-Likelihood:                -25523.
No. Observations:                2024   AIC:                         5.105e+04
Df Residuals:                    2022   BIC:                         5.106e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.252e+05   3197.235     39.165      0.000    1.19e+05    1.31e+05
masvnrarea   240.5226     12.319     19.524      0.000     216.363     264.682
==============================================================================
Omnibus:                      468.466   Durbin-Watson:                   1.944
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1513.890
Skew:                           1.146   Prob(JB):                         0.00
Kurtosis:                       6.564   Cond. No.                         515.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Intercept     125221.247907
masvnrarea       240.522644
dtype: float64
```

The above model exhibits the same issues as the one prior. It has very low R-Squared scores and the model has a very low correlation to the actual data.

```
            OLS Regression Results Sales Price by Square Foot Quality
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.010
Method:                 Least Squares   F-statistic:                     22.59
Date:                Thu, 24 Jan 2019   Prob (F-statistic):           2.15e-06
Time:                        02:06:44   Log-Likelihood:                -25875.
No. Observations:                2039   AIC:                         5.175e+04
Df Residuals:                    2037   BIC:                         5.177e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.528e+05   5848.005     26.134      0.000    1.41e+05    1.64e+05
sfqual       106.9749     22.508      4.753      0.000      62.834     151.115
==============================================================================
Omnibus:                      730.542   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3052.847
Skew:                           1.699   Prob(JB):                         0.00
Kurtosis:                       7.938   Cond. No.                         873.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Intercept    152833.021468
sfqual          106.974861
dtype: float64
```

The above model was another attempt by me to make a new variable and combining square foot and quality. The new metric ended up being quality per square foot. I ended up regressing sale price by this new ordinal variable. The outcome was not good. The model didn't score well on Kaggle and the R-Squared scores are poor along with a lot of kurtosis and skew. The result makes sense to me due to combing an ordinal variable with a continuous variable. The arbitrary rating couldn't effectively combine with the measured one.

```
          OLS Regression Results of Sales Price by Price Square Foot
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.385
Model:                            OLS   Adj. R-squared:                  0.385
Method:                 Least Squares   F-statistic:                     1274.
Date:                Thu, 24 Jan 2019   Prob (F-statistic):          3.52e-217
Time:                        02:07:37   Log-Likelihood:                -25391.
No. Observations:                2039   AIC:                         5.079e+04
Df Residuals:                    2037   BIC:                         5.080e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -6039.9572   5372.169     -1.124      0.261   -1.66e+04    4495.560
pricesf     1525.8375     42.744     35.697      0.000    1442.011    1609.664
==============================================================================
Omnibus:                      523.698   Durbin-Watson:                   2.015
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2015.635
Skew:                           1.211   Prob(JB):                         0.00
Kurtosis:                       7.226   Cond. No.                         492.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Intercept   -6039.957194
pricesf      1525.837541
dtype: float64
```

The above model was a better attempt of creating a new variable and using it in regression with sales price. It has less skew and a better R squared. The R squared is not great but is improved over the other rejected models. I believe that could be due to using continuous variables and combing them vs. taking two different types of variables and combing them to make a new metric. I will have to take a further statistical approach to prove this fact though.

Scored Data File
I have included the scored data file as a separate file on the turn in.

Conclusion

I believe this data set was a good challenge both creatively and statistically. In the end the best model was selected based on different tests from plotting to model summary and Kaggle. The winning model with total square feet as a predictor variable had a good enough R squared to explain away most of the variability. Some of the concerns I have is that the data is a decade old and might not reflect to today's market based on economic conditions. The 2006-2010 time frame from which the data was taken was during the housing crisis and Great Recession. It would include the extreme selling prices up until 2008 and then also the crash in 2009-2010. I am unable to prove the validity of this due to only having data from one market. I would have to see how Ames, IA overall housing prices trended due to the rest of the country. A few markets were not hit as bad as others during this time frame. Overall, the data set contained enough information to build multiple models to predict the sales price.