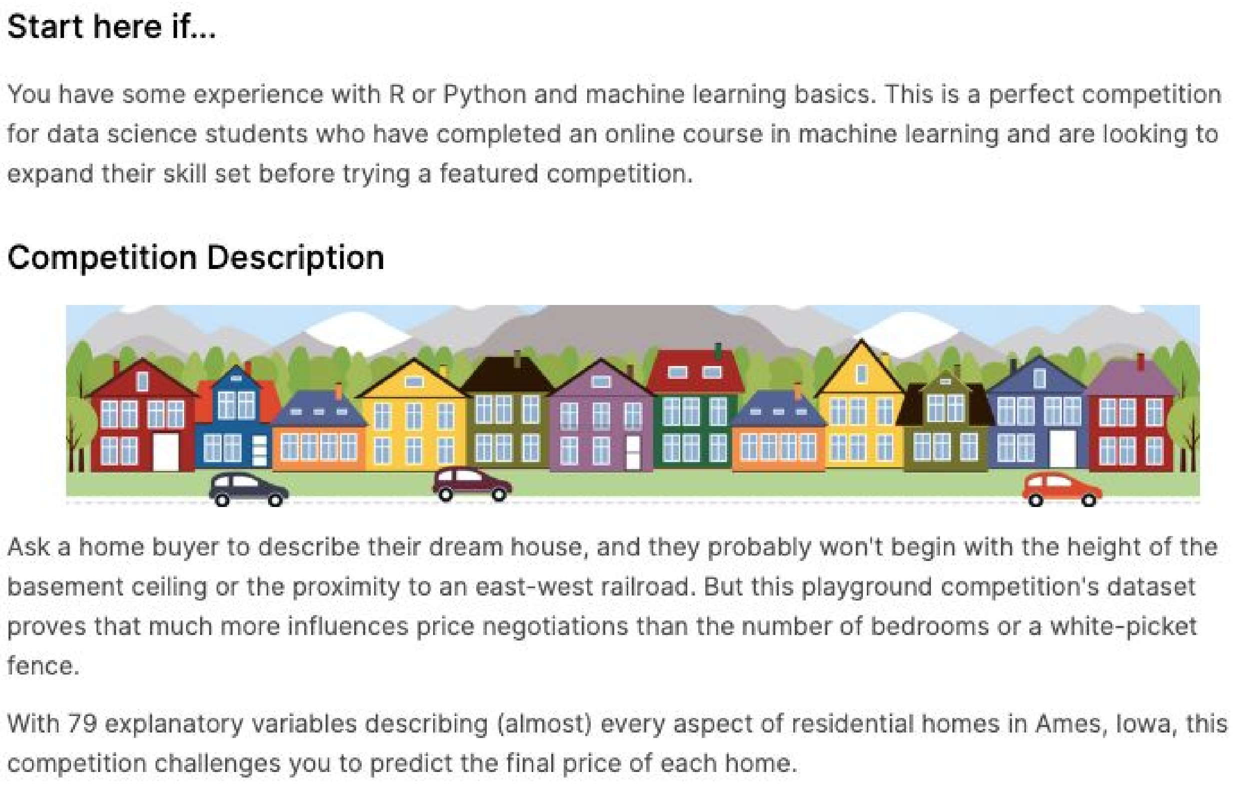
IPD 237

House Prices: Advanced Regression Techniques

STEVEN MCWILLIAMS



Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Abstract

For the final project and in accordance with my proposal, I completed the House Prices: Advanced Regression Techniques competition on Kaggle. The train dataset contains 1460 homes that sold in Ames, IA in 2010. There were 80 variables provided in the train set and 79 variables in the test set (test set excluded actual Sale Prices). My goal was to submit my predictions to Kaggle and score in the top 25%, after testing multiple methods including Linear Regression, KMeans, and Multiple Linear Regression, the test data scored a Root Mean Squared Logarithmic Error (Kaggle's accuracy metric) of 0.26, compared to 0.18 of the

train data. A score of 0.26 ranked in the top 80%. While I did not score as well as I initially planned, I learned a lot about new techniques including preprocessing, KMeans, KFold, and Multiple Linear Regression. I also used scikit more intensively that I have in the past and was introduced to new packages such as yellowbrick.

Packages

The following packages were used in this analysis:

1. pandas: Read the csv file and distribute predictions to new csv file
2. numPy: Statistical analysis
3. sciPy: Calculate R-squared for quick statistical analysis of predictions
4. Pyplot from Matplotlib: Plotting charts
5. SKLearn: Used for statistical analysis
6. Yellowbrick: Created KMeans Elbow chart
7. Math: Needed sqrt to calculate RMSLE
8. Time: Gave program a break between instantiations to keep organized
9. Seaborn: Plotting

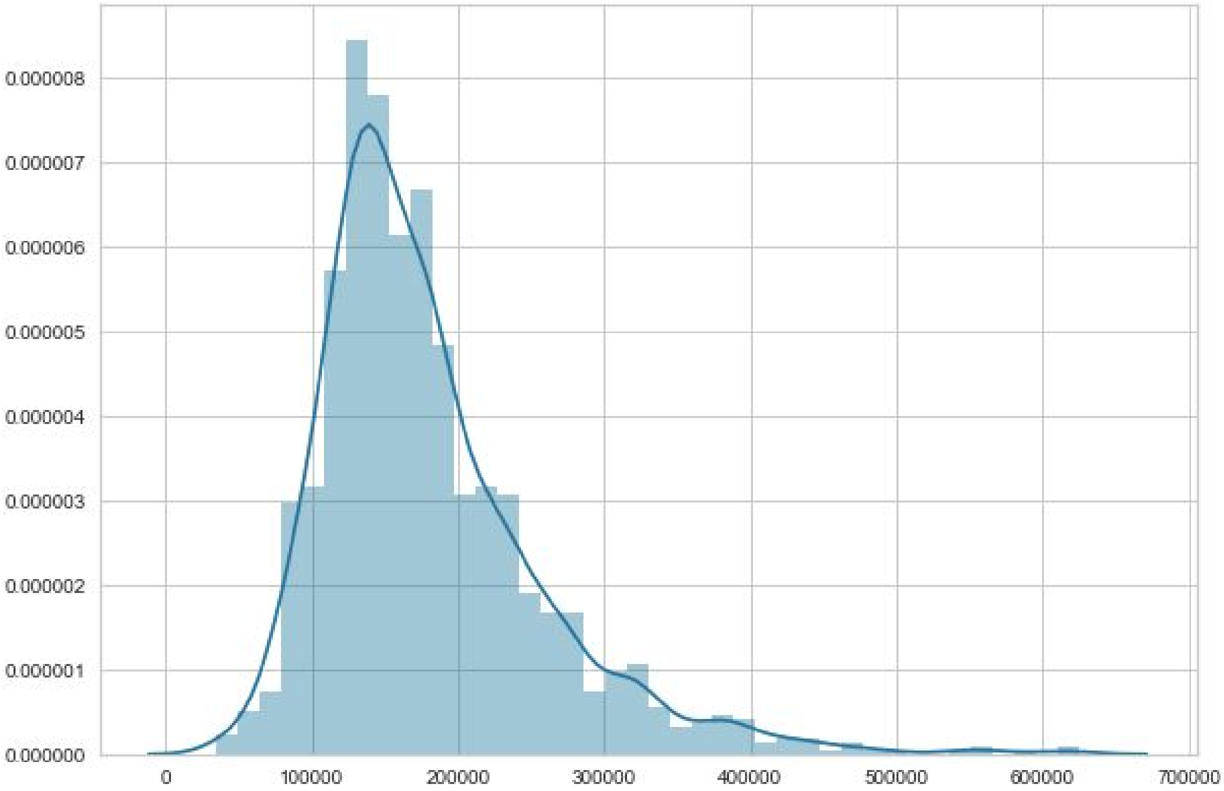
Applied Skills

The following skills learned in class were applied to this program

1. Numpy
2. Pandas
3. List Comprehension
4. Functions
5. Classes
6. Linear Regression
7. KMeans
8. KFold
9. Plotting with Pyplot/Seaborn
10. Indexing Lists

Analysis

Distribution



```
Skewed Distribution:  1.564345548419458
```

Distribution Summary

From the above, distribution is skewed right due to the high count of homes sold between $100,000 and $200,000. From the scipy.skew library, we see that the data is skewed 1.56 to the right, when a score of 0.0 is ideal.

Preprocessing

Preprocessing of the data included dropping Gross Living Area of over 4000 feet, as those homes were gross outliers. I used numpy to drop this data from the train set. I interpolated null values and tested the data, showing zero null values in the train dataset. After testing various square footage restraints, it is clear that 4000 gave the best results (by testing accuracies), as increasing/decreasing the restraint by 500sf reduced accuracy of my final model by a nominal percentage. That does not warrant removal of the data, as the training model should be based on as many data points as possible as not to overfit to the training dataset.

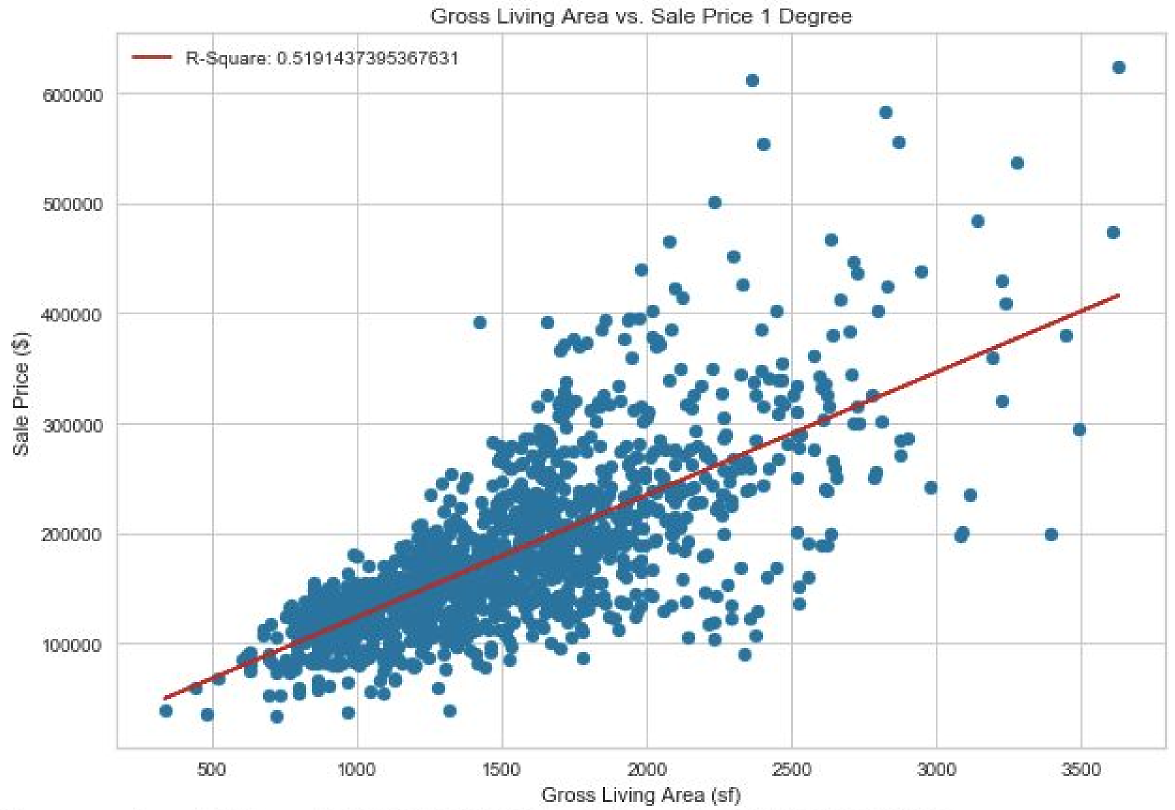
I then applied the above preprocessing techniques to the test dataset, except for the square footage restriction, as I wanted the final test model to utilize all available data.

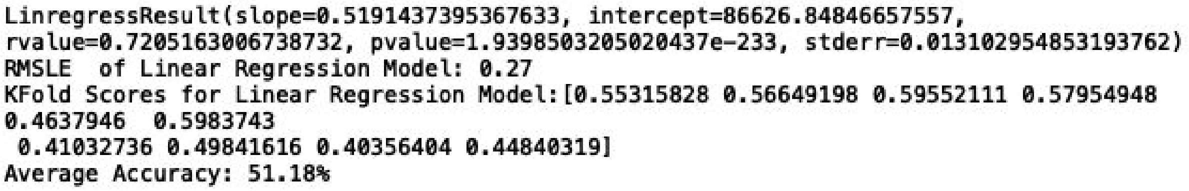
Methodology

I used multiple data science methods to predict house prices including Linear Regression, KMeans, and Multiple Linear Regression. KFolds was used for cross validation and was applied to the Linear Regression and Multiple Linear Regression models to test accuracy.

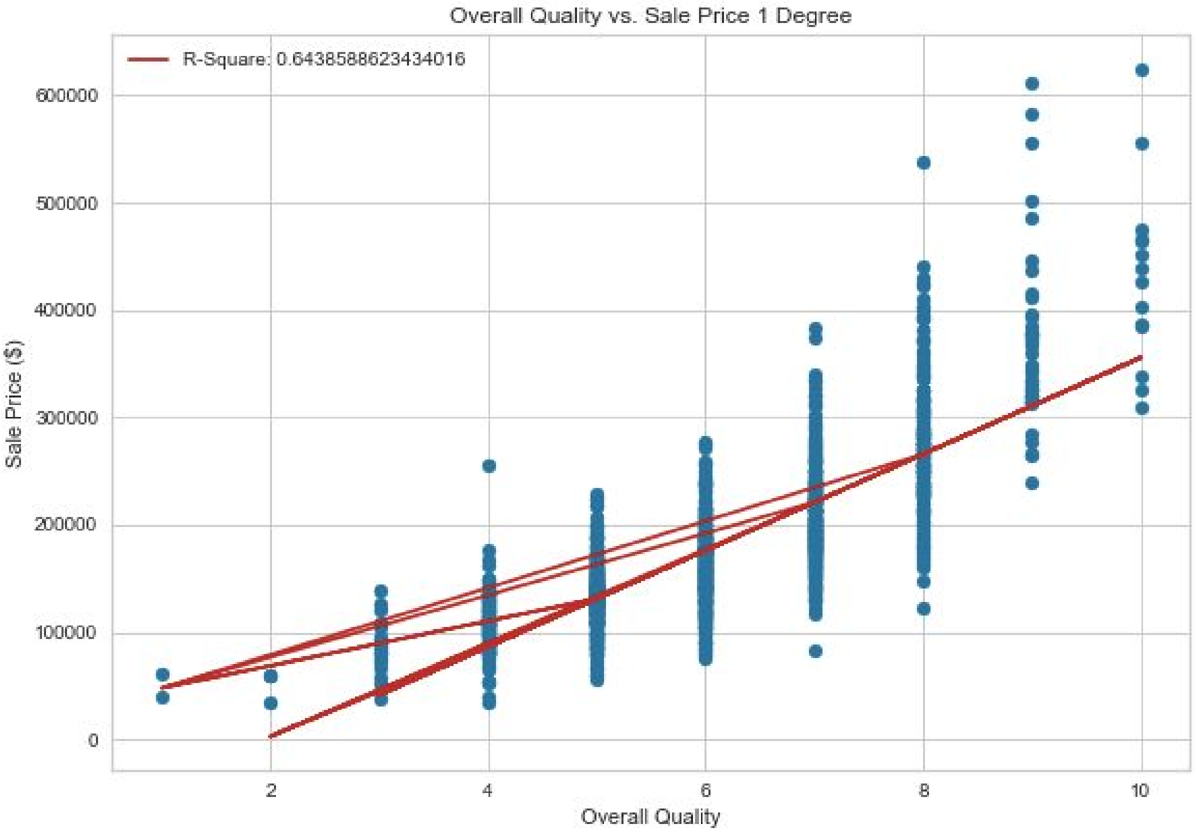
I wanted to focus only on the pertinent variables to house price, as such I sorted the columns based on their correlation to Sale Price. There were only two variables that had over a 70% correlation, Overall Quality and Gross Living Area, which were used in the Linear Regression and KMeans analysis. I then ran checks to identify the most accurate degrees to use by comparing the increase r-squared compared to the prior, along with the K-Fold accuracy score. The K-Fold accuracy is driven by the fitted model, which does not reflect degrees; however, the RMSLE is printed above the KFold Accuracy score, which can be a reference for changes in accuracy over the changes in degrees. Furthermore, R-Squared can be used to judge appropriateness of degrees.

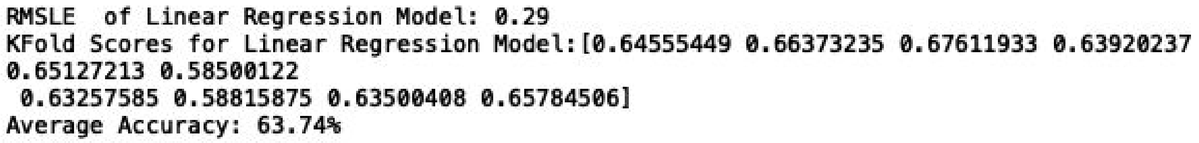
Linear Regression



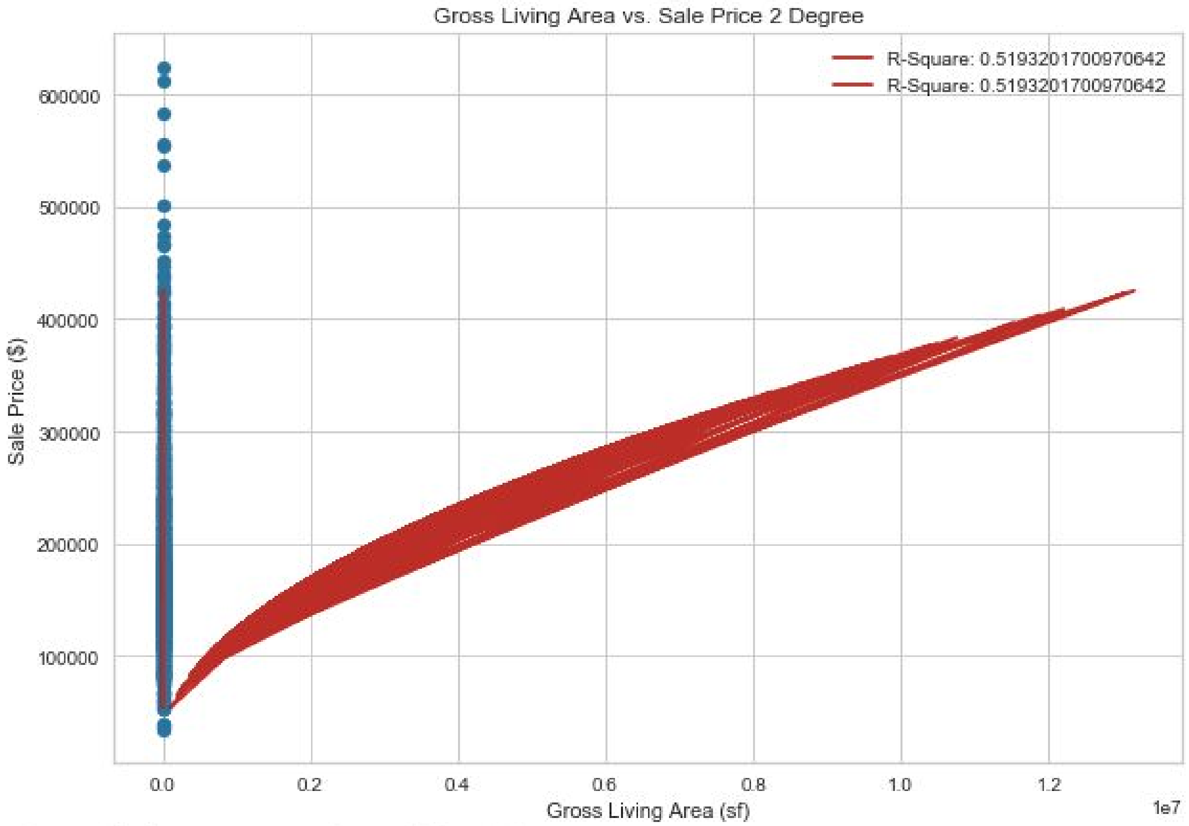


```
LinregressResult(slope=0.5191437395367633, intercept=86626.84846657557,
rvalue=0.7205163006738732, pvalue=1.9398503205020437e-233, stderr=0.013102954853193762)
RMSLE  of Linear Regression Model: 0.27
KFold Scores for Linear Regression Model:[0.55315828 0.56649198 0.59552111 0.57954948
0.4637946  0.5983743
 0.41032736 0.49841616 0.40356404 0.44840319]
Average Accuracy: 51.18%
```

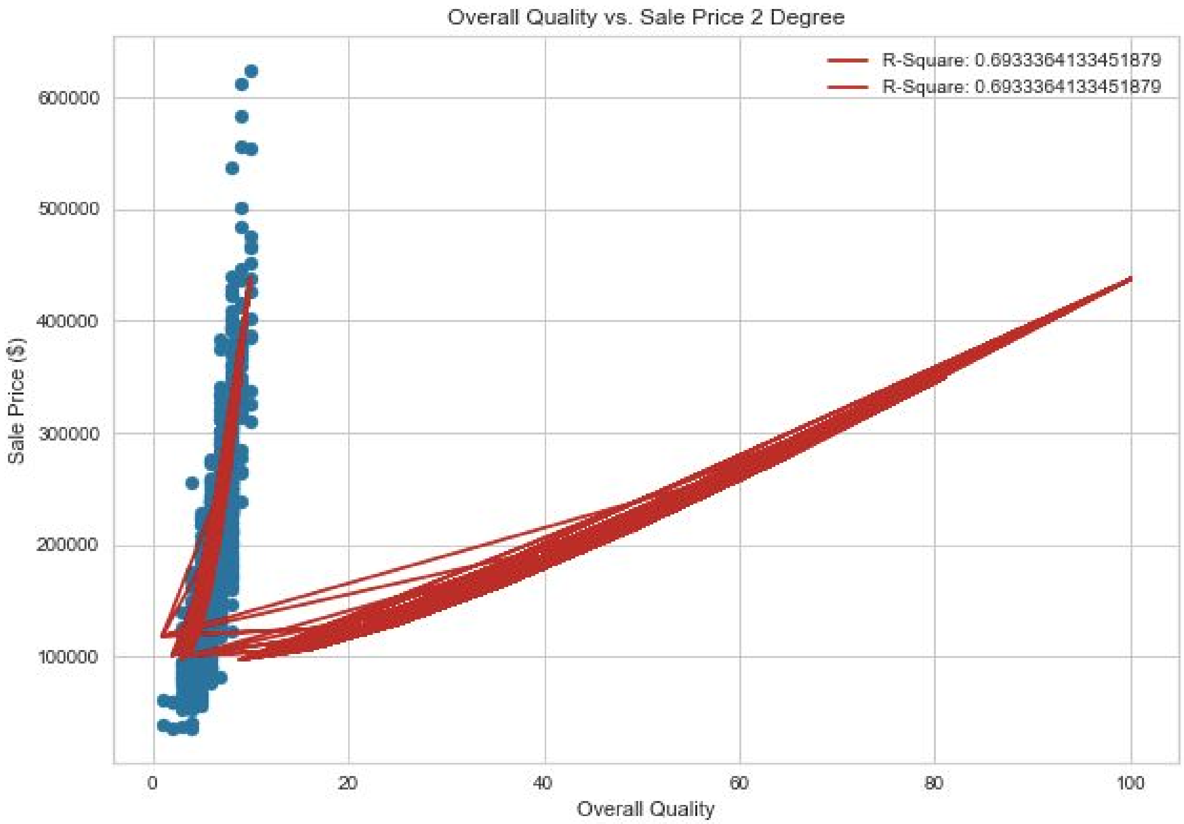




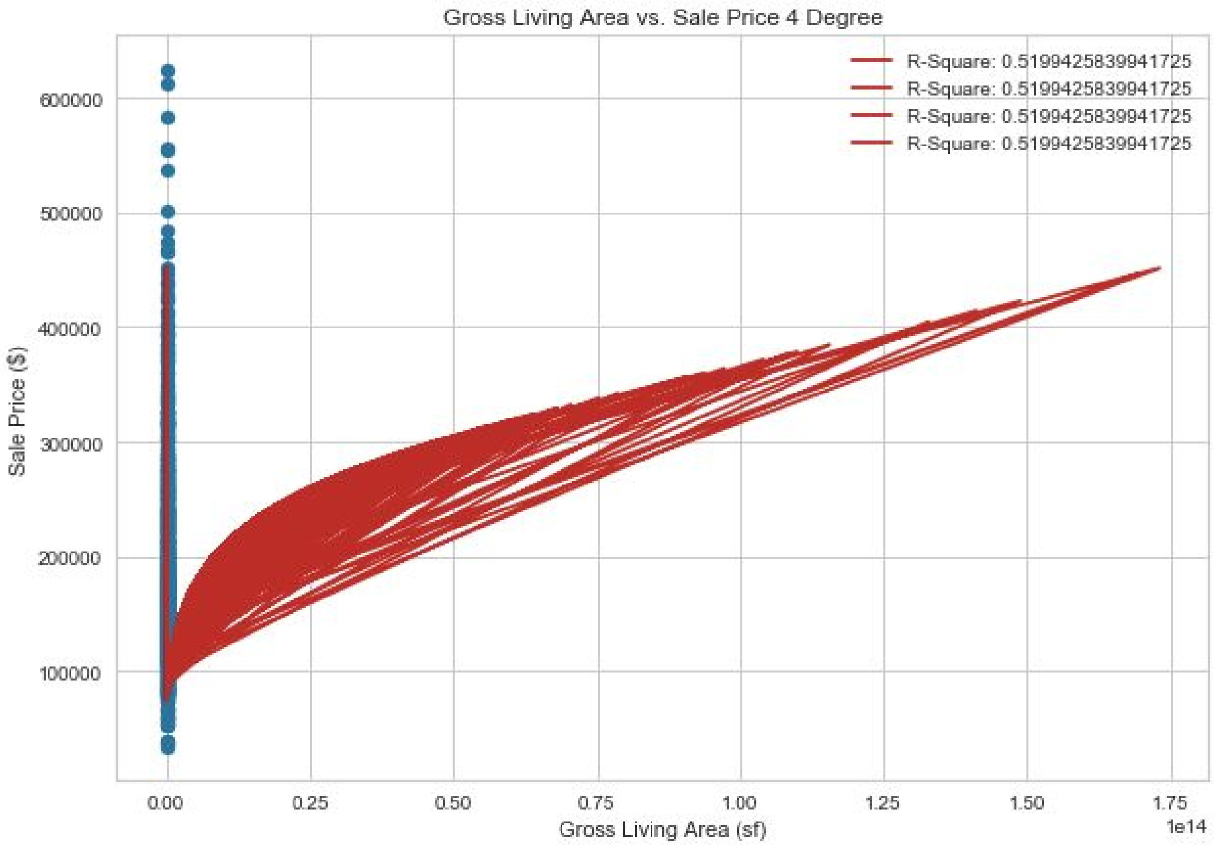
```
RMSLE  of Linear Regression Model: 0.29
KFold Scores for Linear Regression Model:[0.64555449 0.66373235 0.67611933 0.63920237
0.65127213 0.58500122
 0.63257585 0.58815875 0.63500408 0.65784506]
Average Accuracy: 63.74%
```



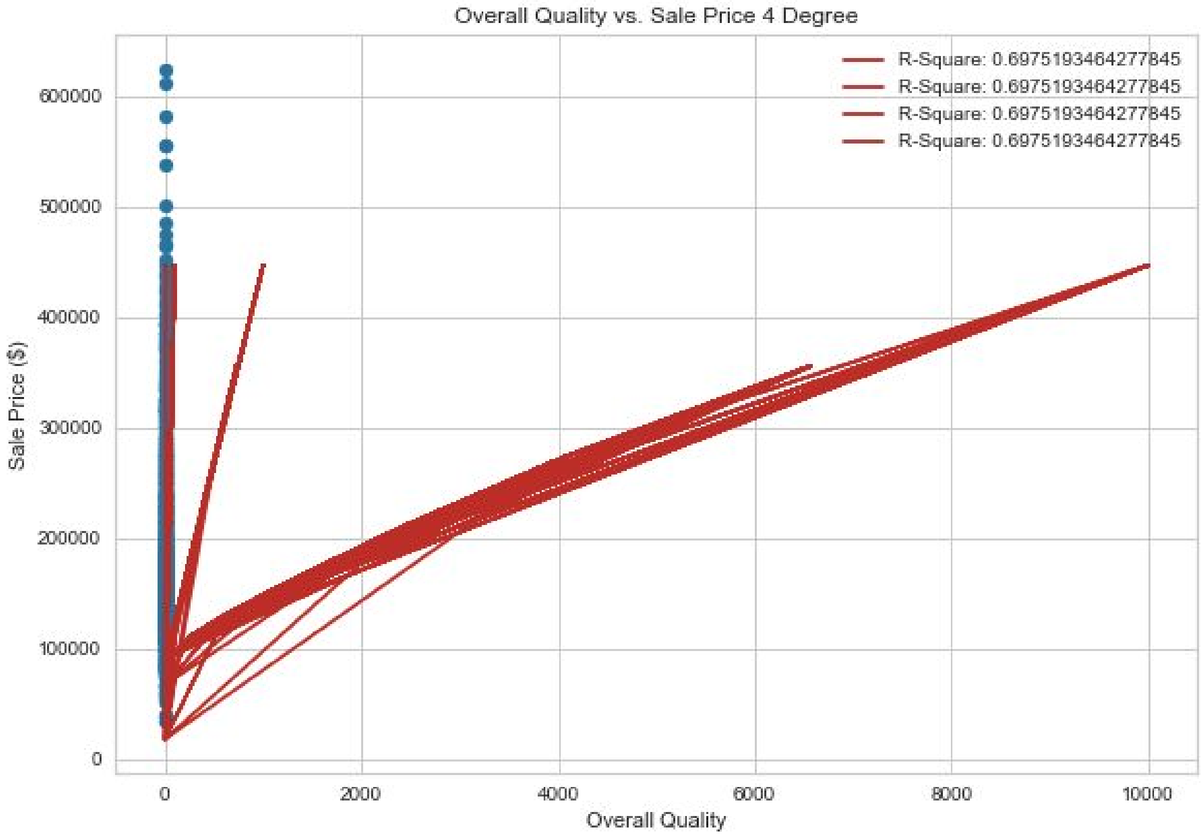
```
RMSLE  of Linear Regression Model: 0.27
KFold Scores for Linear Regression Model:[0.55315828 0.56649198 0.59552111 0.57954948
0.4637946  0.5983743
 0.41032736 0.49841616 0.40356404 0.44840319]
Average Accuracy: 51.18%
```

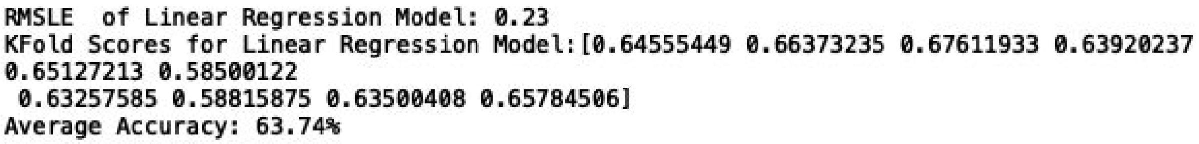


```
LinregressResult(slope=0.6933364133451881, intercept=55245.82341045339,
rvalue=0.8326682492716939, pvalue=0.0, stderr=0.012092633953151391)
RMSLE  of Linear Regression Model: 0.23
KFold Scores for Linear Regression Model:[0.64555449 0.66373235 0.67611933 0.63920237
0.65127213 0.58500122
 0.63257585 0.58815875 0.63500408 0.65784506]
Average Accuracy: 63.74%
```



```
RMSLE  of Linear Regression Model: 0.27
KFold Scores for Linear Regression Model:[0.55315828 0.56649198 0.59552111 0.57954948
0.4637946  0.5983743
 0.41032736 0.49841616 0.40356404 0.44840319]
Average Accuracy: 51.18%
```





```
RMSLE  of Linear Regression Model: 0.23
KFold Scores for Linear Regression Model:[0.64555449 0.66373235 0.67611933 0.63920237
0.65127213 0.58500122
 0.63257585 0.58815875 0.63500408 0.65784506]
Average Accuracy: 63.74%
```

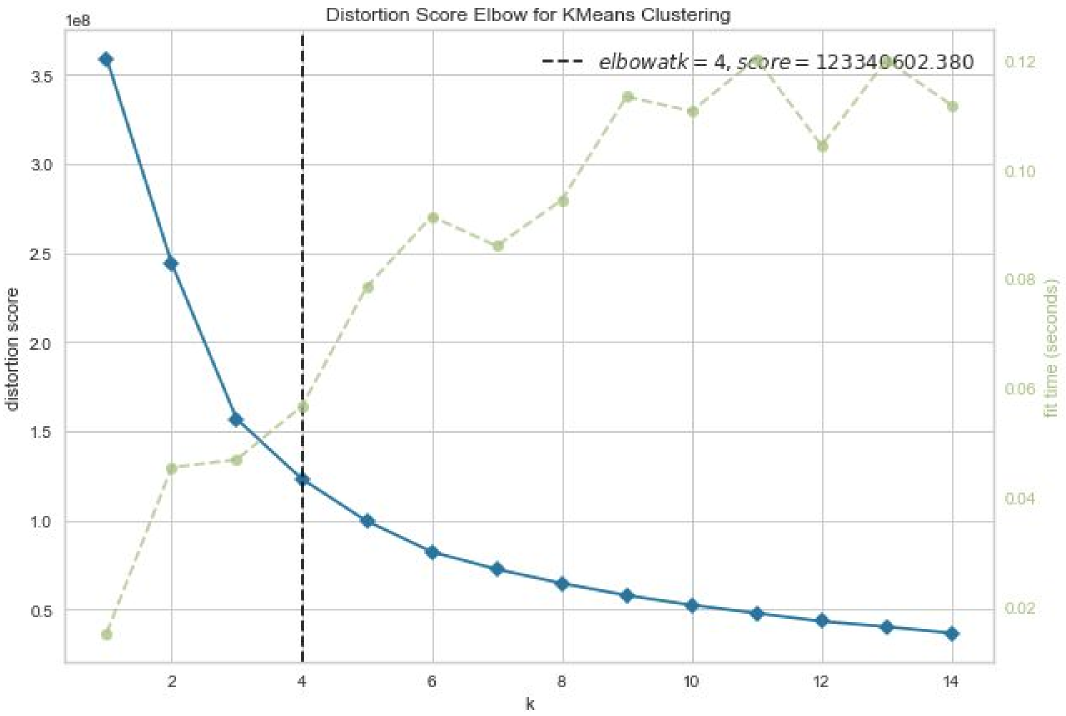
Linear Regression Summary

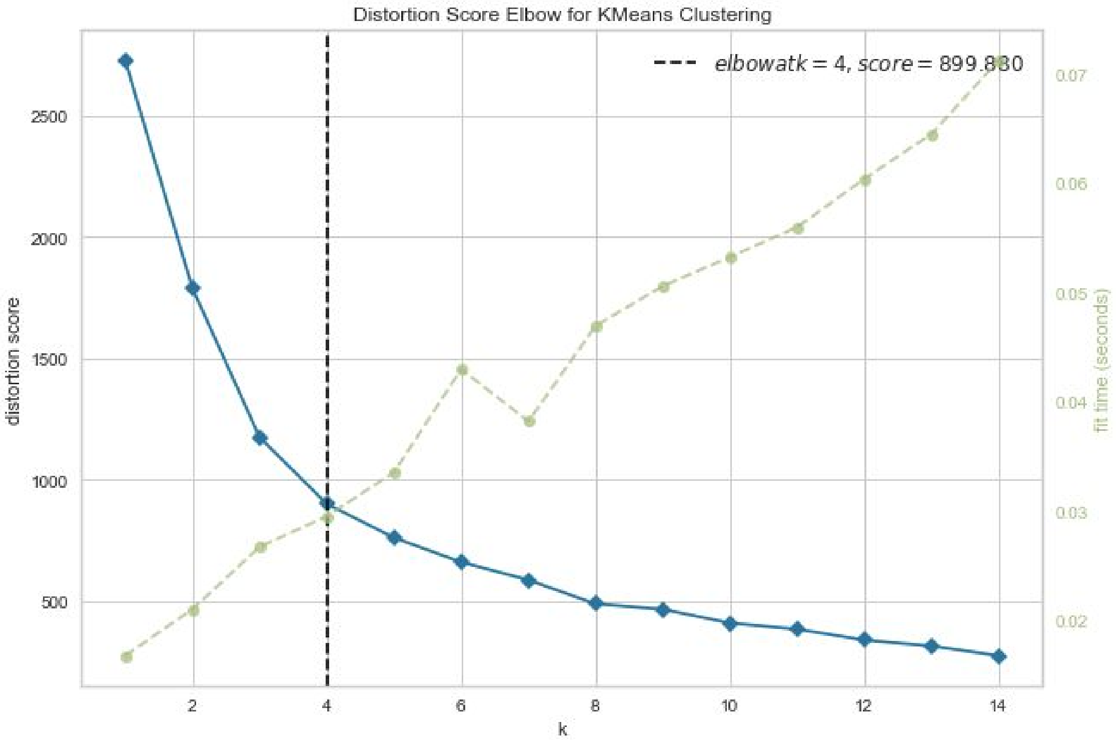
From the above, the charts became distorted when using multiple degrees of linear regression, but the r-squared metric was still calculated correctly. For Overall Quality, the only meaningful r-squared change was from degree one to degree two, an increase from 0.64 to 0.69. Further, for the selected degrees, the statistics are printed to the terminal, with the Overall Quality and Gross Living Area reaching r-values of 0.72 and 0.83, respectively.

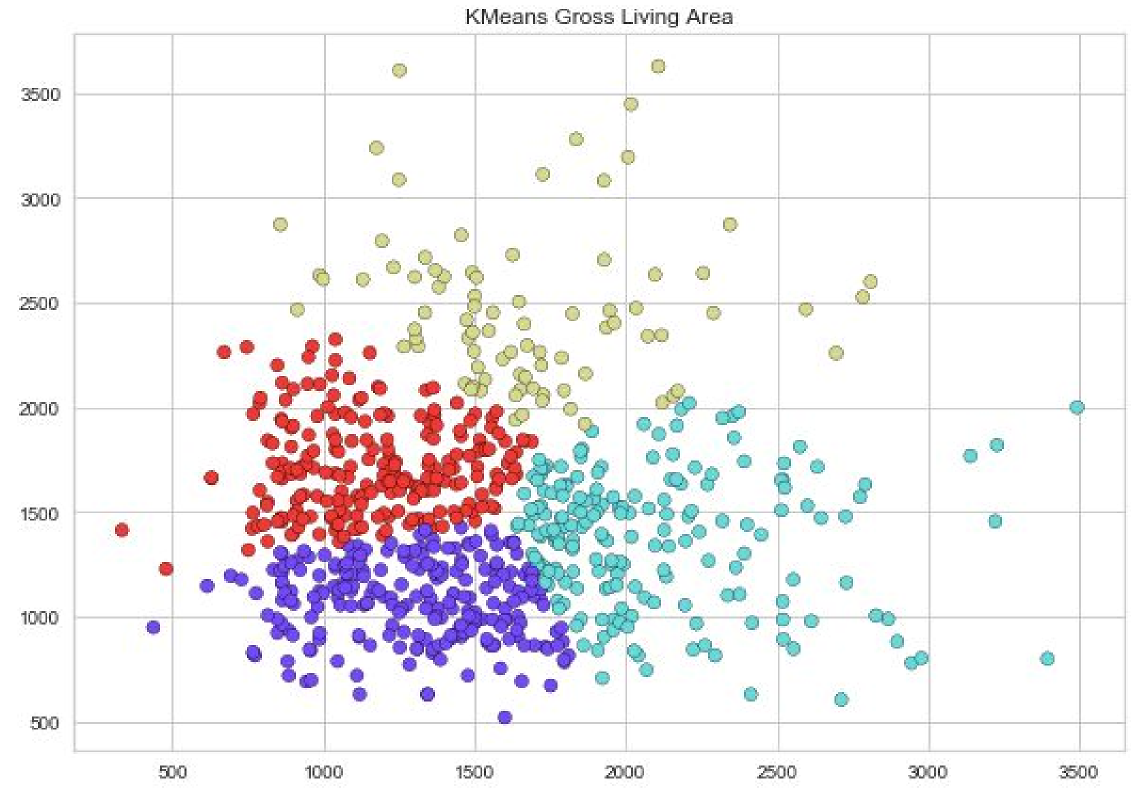
The Gross Living Area r-squared value did not change materially no matter how many degrees were applied, thus I selected to use degree one.

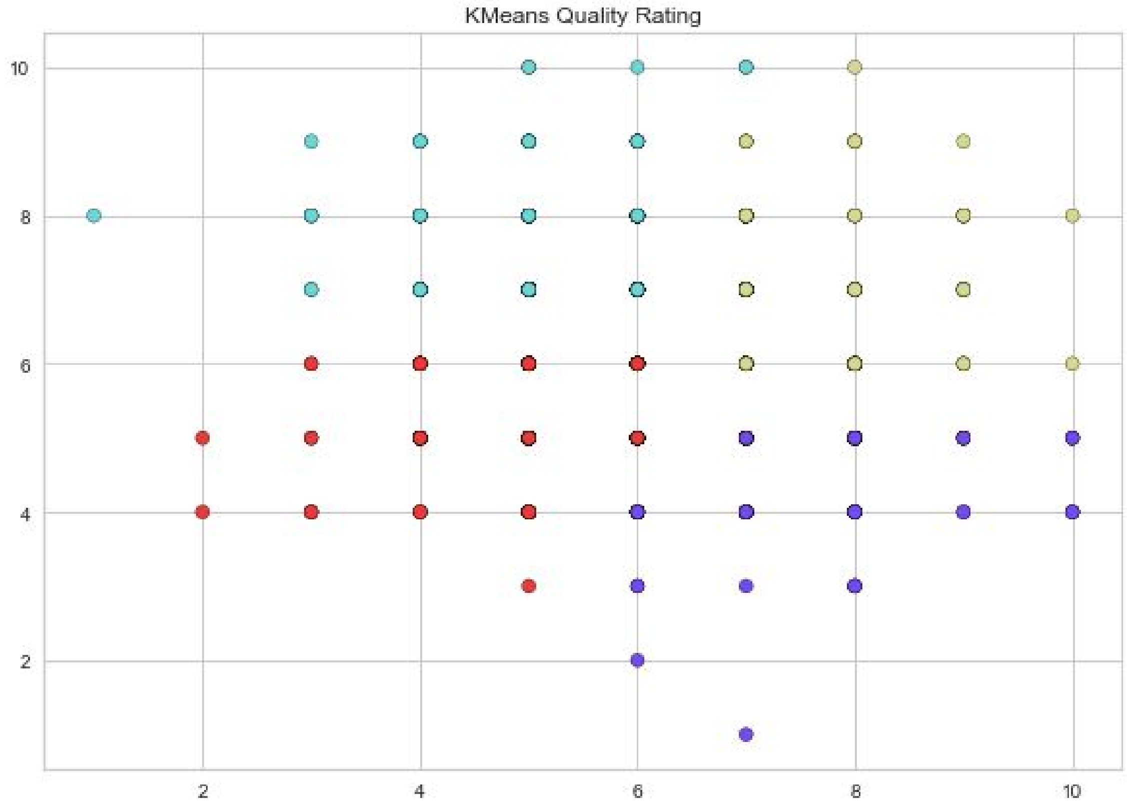
It should be noted that the RMSLE was large for the linear regression models, as such I moved to a different analysis.

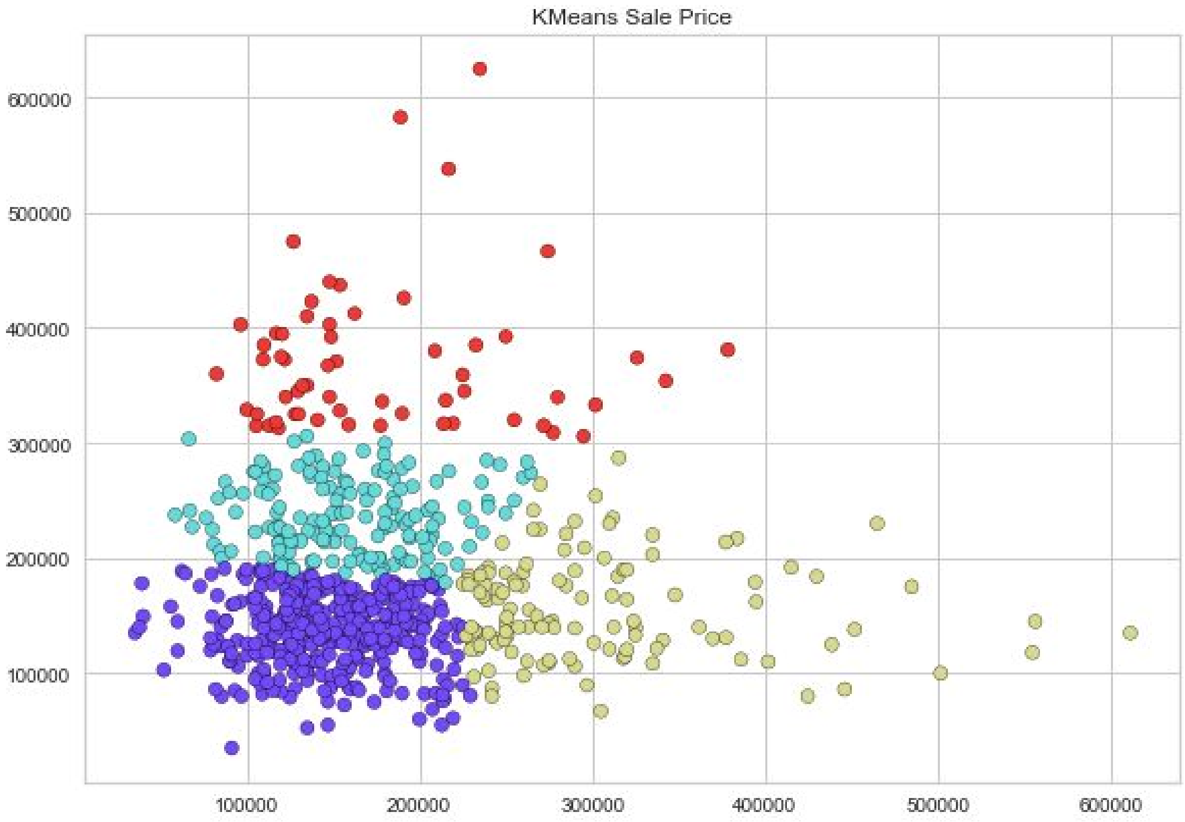
KMeans


Distortion Score Elbow for KMeans Clustering
1e8
elbowatk = 4, score = 123340602.380
distortion score
3.5
3.0
2.5
2.0
1.5
1.0
0.5
2
4
6
8
10
12
14
k
fit time (seconds)
0.12
0.10
0.08
0.06
0.04
0.02


Distortion Score Elbow for KMeans Clustering
elbowatk = 4, score = 899.880
distortion score
2500
2000
1500
1000
500
fit time (seconds)
0.07
0.06
0.05
0.04
0.03
0.02
2
4
6
8
10
12
14
k


KMeans Gross Living Area
3500
3000
2500
2000
1500
1000
500
500
1000
1500
2000
2500
3000
3500

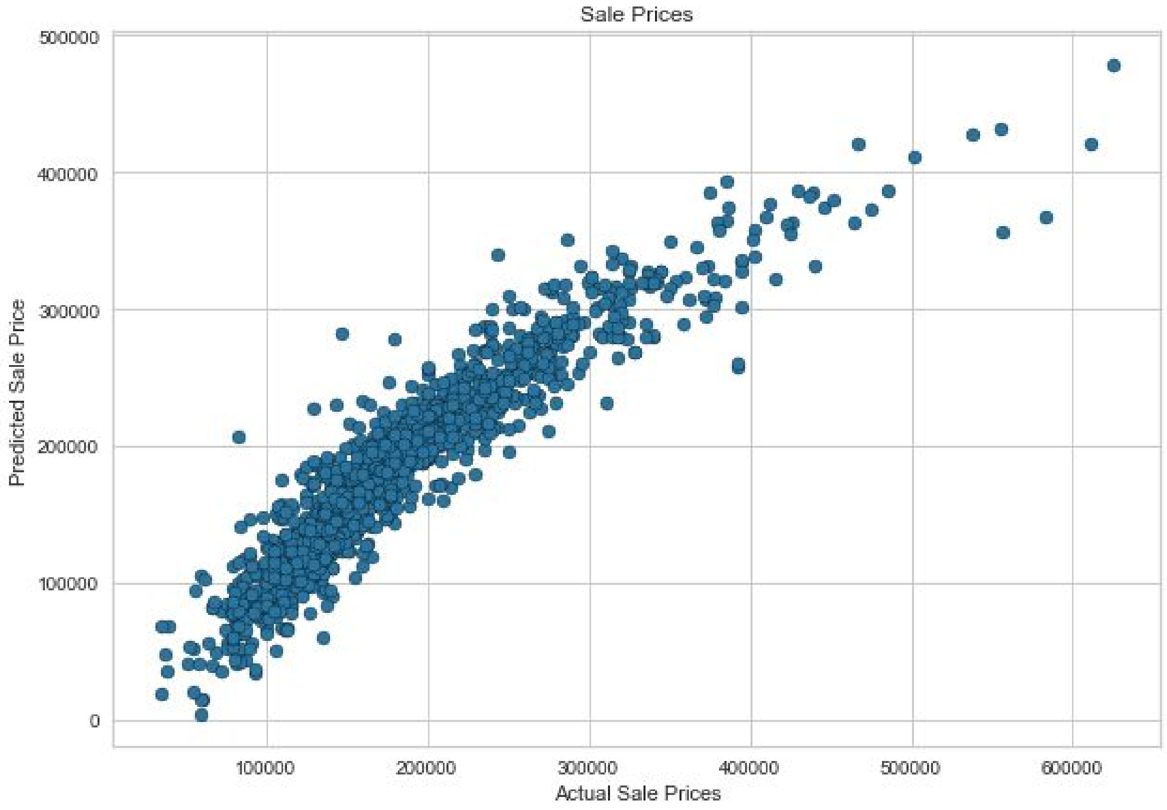

KMeans Quality Rating
10
8
6
4
2
2
4
6
8
10

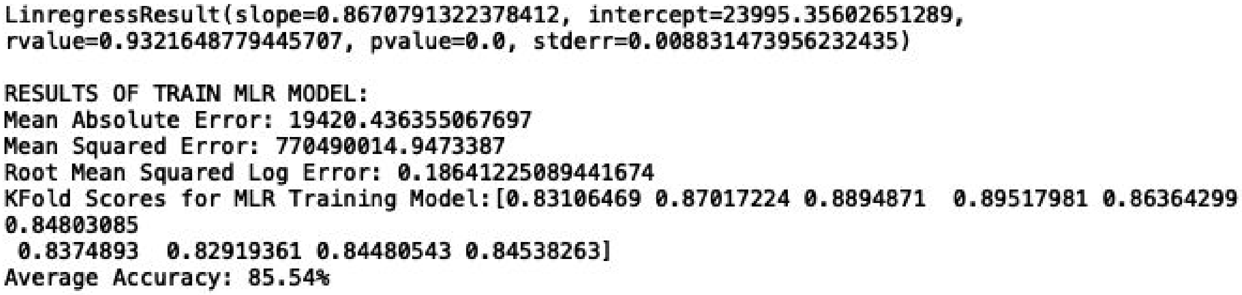


KMeans Summary

The elbow method charts, plotted for both Overall Quality, Gross Living Area, and Sale Price, respectively, indicate that a KMeans of 4 should be used for the analysis. When splitting the data into four clusters, the data shows that any quality score over 6 yields significantly different results than scores lower than six. Further, the Gross Living Area Kmeans shows around 1,750sf that leads to slightly higher sale prices. Finally, the Sale Price chart shows that there are four clear groups with a general line around $225,000. The three KMeans charts indicate that the higher-price homes are correlated with the scores greater than six and with 1,750sf+. While the KMeans charts show readily available information, it would require more in-depth analysis to find the drivers behind the data. It would also require weighing my findings in order to accurately predict sale prices of the test data, which I do not have the skill to do. While KMeans provides meaningful information, the work required to use the information to predict Sale Price is out of the scope of work for this project.

Multiple Linear Regression





```
LinregressResult(slope=0.8670791322378412, intercept=23995.35602651289,
rvalue=0.9321648779445707, pvalue=0.0, stderr=0.008831473956232435)

RESULTS OF TRAIN MLR MODEL:
Mean Absolute Error: 19420.436355067697
Mean Squared Error: 770490014.9473387
Root Mean Squared Log Error: 0.18641225089441674
KFold Scores for MLR Training Model:[0.83106469 0.87017224 0.8894871  0.89517981 0.86364299
0.84803085
 0.8374893  0.82919361 0.84480543 0.84538263]
Average Accuracy: 85.54%
```

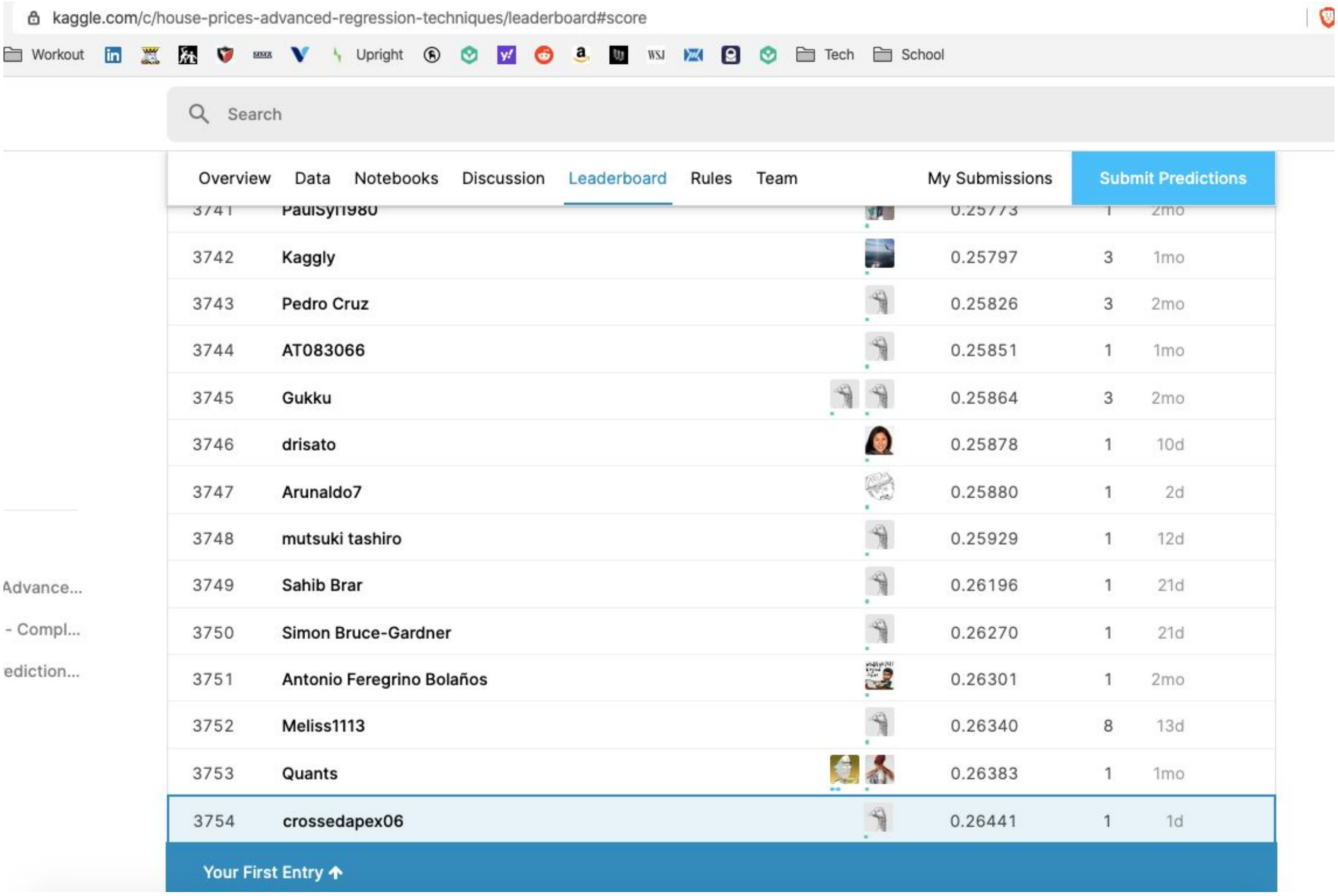
Multiple Linear Regression Summary

The multiple linear regression model uses all data available, except for categorical data. In this case, 36 variables are at play when calculating Sale Price. Because the linear regression models showed low r-squared levels, and low KFold accuracy ranging from 51% to 63%, the multiple linear regression model is the best option at my disposal with an 85.54% KFold accuracy. More importantly for the Kaggle competition, the RMSLE decreased from a best of 0.23 using linear regression to 0.18.

The multiple linear regression model was more accurate than the linear regression models, as it reached an r-value of 0.93, this accuracy is also supported by the p-value of 0.0.The Gross Living Area p-value of 1.93 shows that its not a reliable predictor, which is supported by its r-value.. The Overall Quality regression model received a p-value of 0.0 as well, indicating that the multiple linear regression method and Overall Quality, as a variable, is more reliable than the Gross Living Area variable.

Conclusion

Overall, a RMSLE of 0.18 is fairly accurate, but does not rank well in the Kaggle competition, as I was in the 80th percentile, screenshot below. The test data revealed a score of 0.26 Because the model did not have normal distribution, a logarithmic regression would likely result in a more accurate result. Although, after a lot of research, multiple linear regression was the best method within my skill set. Many other methods were popular including more in-depth feature engineering, residual plots, XGBRegressor, Gradient Boosting Regressor, and Bayesian Ridge Regression; I am excited to learn this in the Machine Learning and Deep Learning Program.



3741	PaulSyn1980	0.25773	1	2mo
3742	Kaggly	0.25797	3	1mo
3743	Pedro Cruz	0.25826	3	2mo
3744	AT083066	0.25851	1	1mo
3745	Gukku	0.25864	3	2mo
3746	drisato	0.25878	1	10d
3747	Arunaldo7	0.25880	1	2d
3748	mutsuki tashiro	0.25929	1	12d
3749	Sahib Brar	0.26196	1	21d
3750	Simon Bruce-Gardner	0.26270	1	21d
3751	Antonio Feregrino Bolaños	0.26301	1	2mo
3752	Meliss1113	0.26340	8	13d
3753	Quants	0.26383	1	1mo
3754	crossedapex06	0.26441	1	1d