UC2SAT101 Final Assessment

Martin Slaastuen

May 2021

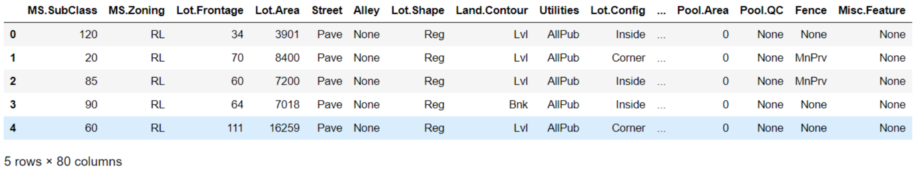
Contents

1 Statistical Data Analysis

1.1 Introduction

The goal of this analysis is to build a model that based on features of houses can predict the sales prices. The dataset used to build the model is from real estate sales in Ames, Iowa between 2006 and 2010.

Before building the model I need to get an overview of our data to see what I have to work with. Figure 1 and figure 2 show us what the data contains and we can start to get an overview of our data.



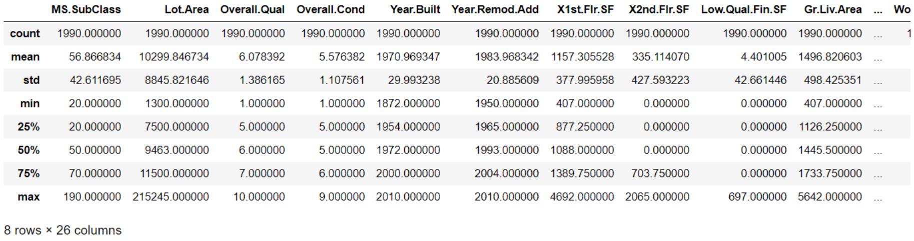
	MS.SubClass	MS.Zoning	Lot.Frontage	Lot.Area	Street	Alley	Lot.Shape	Land.Contour	Utilities	Lot.Config	...	Pool.Area	Pool.QC	Fence	Misc.Feature
0	120	RL	34	3901	Pave	None	Reg	Lvl	AllPub	Inside	...	0	None	None	None
1	20	RL	70	8400	Pave	None	Reg	Lvl	AllPub	Corner	...	0	None	MnPrv	None
2	85	RL	60	7200	Pave	None	Reg	Lvl	AllPub	Inside	...	0	None	MnPrv	None
3	90	RL	64	7018	Pave	None	Reg	Bnk	AllPub	Inside	...	0	None	None	None
4	60	RL	111	16259	Pave	None	Reg	Lvl	AllPub	Corner	...	0	None	None	None

5 rows × 80 columns

Figure 1: First 5 rows from the dataset

1.1.1 Descriptive Statistics

Figure 2 give us an insight of the mean, standard deviation(STD), min and max values for our variables.



	MS.SubClass	Lot.Area	Overall.Qual	Overall.Cond	Year.Built	Year.Remod.Add	X1st.Flr.SF	X2nd.Flr.SF	Low.Qual.Fin.SF	Gr.Liv.Area	...	Wo
count	1990.000000	1990.000000	1990.000000	1990.000000	1990.000000	1990.000000	1990.000000	1990.000000	1990.000000	1990.000000	...	1
mean	56.866834	10299.846734	6.078392	5.576382	1970.969347	1983.968342	1157.305528	335.114070	4.401005	1496.820603	...	
std	42.611695	8845.821646	1.386165	1.107561	29.993238	20.885609	377.995958	427.593223	42.661446	498.425351	...	
min	20.000000	1300.000000	1.000000	1.000000	1872.000000	1950.000000	407.000000	0.000000	0.000000	407.000000	...	
25%	20.000000	7500.000000	5.000000	5.000000	1954.000000	1965.000000	877.250000	0.000000	0.000000	1126.250000	...	
50%	50.000000	9463.000000	6.000000	5.000000	1972.000000	1993.000000	1088.000000	0.000000	0.000000	1445.500000	...	
75%	70.000000	11500.000000	7.000000	6.000000	2000.000000	2004.000000	1389.750000	703.750000	0.000000	1733.750000	...	
max	190.000000	215245.000000	10.000000	9.000000	2010.000000	2010.000000	4692.000000	2065.000000	697.000000	5642.000000	...	

8 rows × 26 columns

Figure 2: Descriptive Statistics of the data

As I am going to build a model that predicts the sales prices wanted to see the distribution plot for the sale prices of the dataset to see what kind of distribution we have and if we have any outliers in our data.

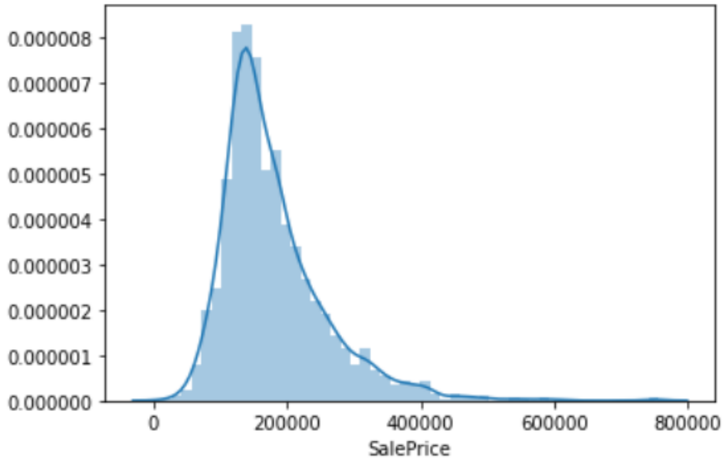


Figure 3: Distribution plot of sales price data

As we can see from figure 3, most of the sales prices range between 100k and 200k, but we have some outliers that you can see to the right of the distribution plot. Outliers can affect the accuracy of my model, so I want to remove any outliers we have within the data. I used the Numpy library in python to identify and remove any outliers and created a new distribution plot. The new distribution plot has a more normalized distribution, but is still a bit skewed to the right. But it is a lot better than the distribution plot in figure 3.

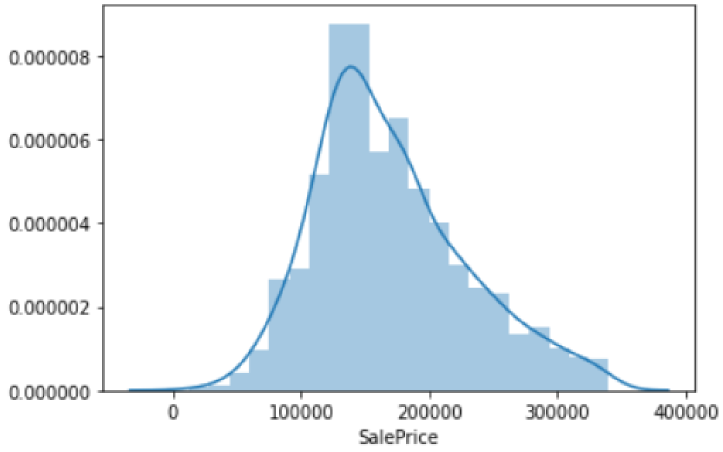


Figure 4: New distribution plot after removing outliers.

1.1.2 Correlation Heatmap

To predict the sales prices, I want to find predictor variables that are highly influential, which means variables that has a high correlation to sales prices within the dataset. To do this I used Seaborn and Python to create a correlation heatmap as you can see in figure 5.

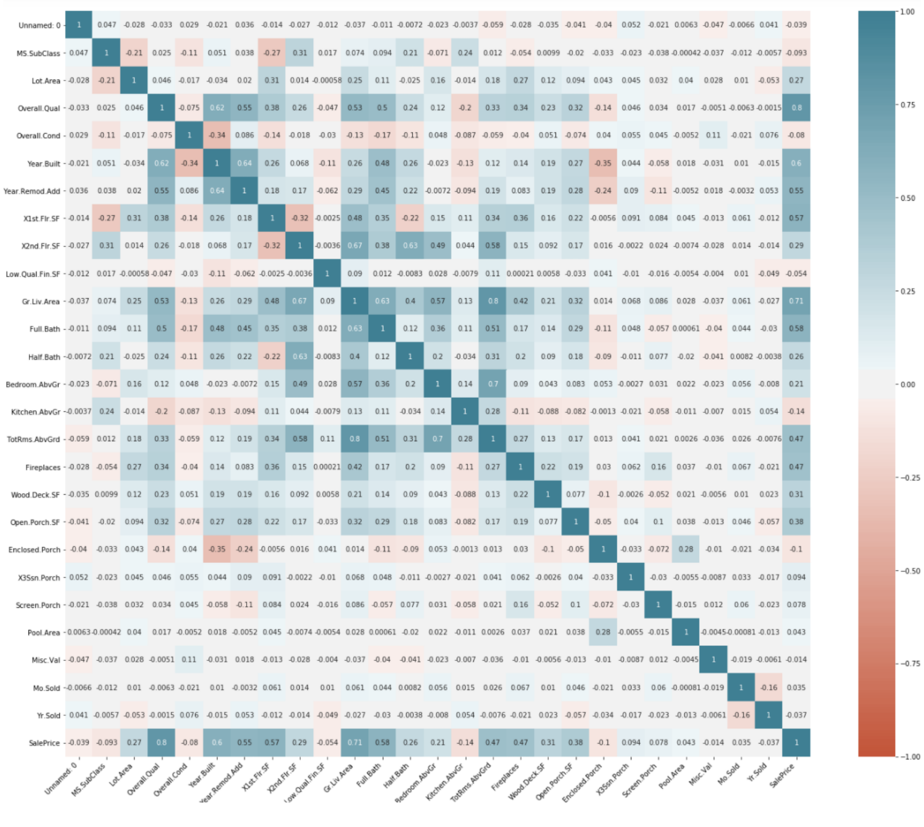


Figure 5: Correlation heatmap of dataset

The correlation coefficient is between -1 and 1. Where -1 is a strong negative correlation, and 1 is a strong positive correlation. 0 is insignificant to none correlation between variables. Here there are several variables that has a fairly high correlation to sales prices, but if we focus on what variables has the highest correlation we can see from the heatmap that Overall.Qual (Overall quality) and Gr.Live.Area (living area square feet) has the highest correlation. Overall.Qual has an pearson correlation of $r = 0.8$, while Gr.Live.Area has an pearson correlation coefficient of $r = 0.71$.Those are the predictor variables I will be focusing on for my regression analysis.

Figure 6 and figure 7 shows scatter plots of the two predictor variables I chose to use for this analysis. I have done this to get a visual representation of our data and it gives us an insight into what kind of regression model we may use to predict the sales price. Looking at the scatter plots, it looks like a linear model may be a good fit. But I have to preform tests on which model is the best fit before anything conclusive can be said.

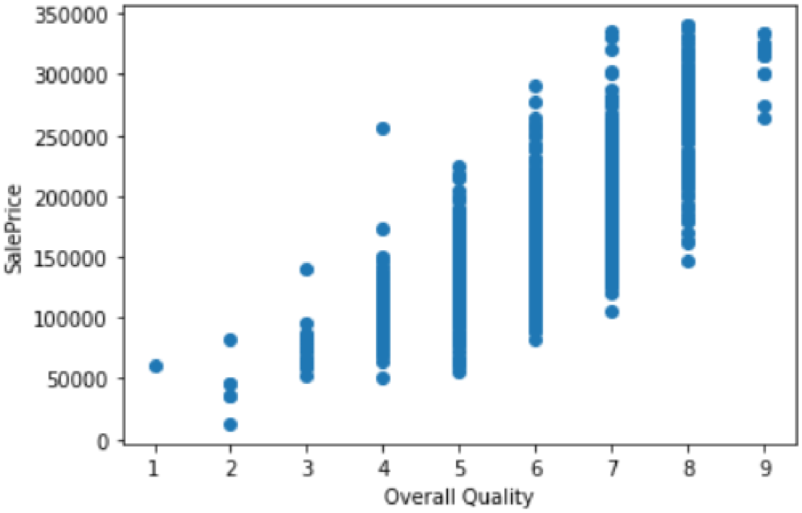


Figure 6: Scatter plot where Sales price is on Y-axis, and Overall quality is on the X-axis

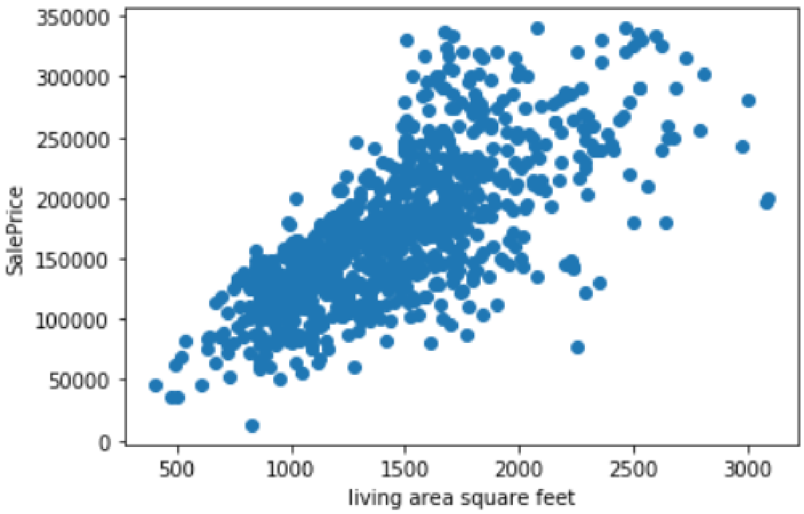


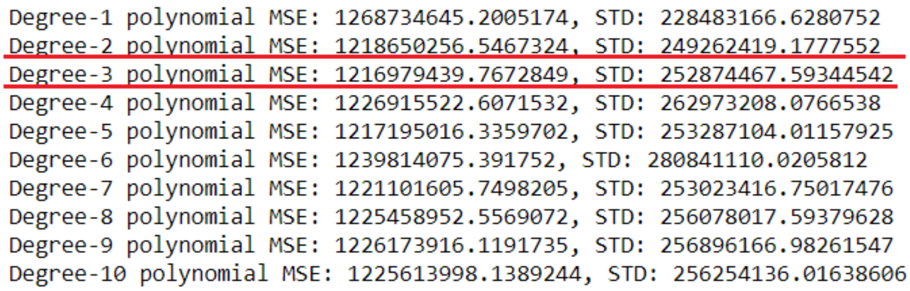
Figure 7: Scatter plot where Sales price is on Y-axis, and Gr.Liv.Area (living area square feet) is on the X-axis

1.2 Regression Analysis

1.2.1 K-Fold Cross Validation

For me to select the correct model, I need to perform some tests, In this analysis I have chose to use the K-Fold cross validation. This method splits the data into equal sized folds and gives each fold the chance to be the validation set while the rest of the data is used as training set. The results I get from the test is the average error over all folds. I have chosen to set $K = 10$. This means that the test creates 10 folds that is used to test the model.

Figure 8 shows the results of the K-Fold cross validation when using the Overall Quality as the predictor variable.

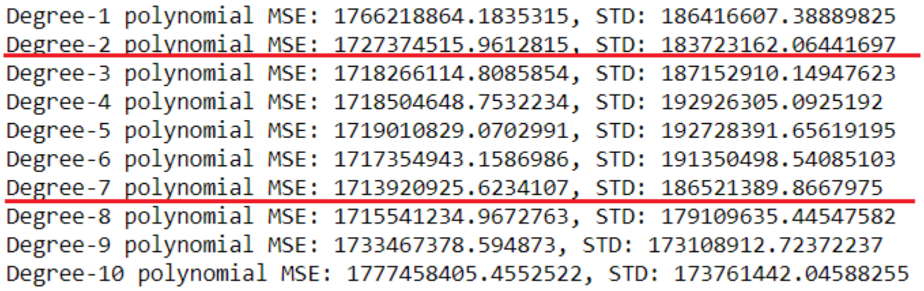


```
Degree-1 polynomial MSE: 1268734645.2005174, STD: 228483166.6280752
Degree-2 polynomial MSE: 1218650256.5467324, STD: 249262419.1777552
Degree-3 polynomial MSE: 1216979439.7672849, STD: 252874467.59344542
Degree-4 polynomial MSE: 1226915522.6071532, STD: 262973208.0766538
Degree-5 polynomial MSE: 1217195016.3359702, STD: 253287104.01157925
Degree-6 polynomial MSE: 1239814075.391752, STD: 280841110.0205812
Degree-7 polynomial MSE: 1221101605.7498205, STD: 253023416.75017476
Degree-8 polynomial MSE: 1225458952.5569072, STD: 256078017.59379628
Degree-9 polynomial MSE: 1226173916.1191735, STD: 256896166.98261547
Degree-10 polynomial MSE: 1225613998.1389244, STD: 256254136.01638606
```

Figure 8: K-Fold cross validation results when using predictor variable Overall Quality

Looking at the results, We can see that a polynomial regression model with a degree of 3 gives the least amount of mean squared error. But as I want my model to understand the general problem and not memorize the problem, I am going to build a polynomial regression model with a degree of 2, also called a quadratic regression model.

I also preformed the K-Fold cross validation when using Gr.Liv.Area(living area square feet) as the predictor variable that you can see in figure 9



```
Degree-1 polynomial MSE: 1766218864.1835315, STD: 186416607.38889825
Degree-2 polynomial MSE: 1727374515.9612815, STD: 183723162.06441697
Degree-3 polynomial MSE: 1718266114.8085854, STD: 187152910.14947623
Degree-4 polynomial MSE: 1718504648.7532234, STD: 192926305.0925192
Degree-5 polynomial MSE: 1719010829.0702991, STD: 192728391.65619195
Degree-6 polynomial MSE: 1717354943.1586986, STD: 191350498.54085103
Degree-7 polynomial MSE: 1713920925.6234107, STD: 186521389.8667975
Degree-8 polynomial MSE: 1715541234.9672763, STD: 179109635.44547582
Degree-9 polynomial MSE: 1733467378.594873, STD: 173108912.72372237
Degree-10 polynomial MSE: 1777458405.4552522, STD: 173761442.04588255
```

Figure 9: K-Fold cross validation results when using predictor variable Gr.Liv.Area(living area square feet)

The results for this test preforms worse than the previous test, with a much higher mean squared error than the previous test. We can also see that a polynomial regression with a degree of 7 gives the best result, this will cause the model to memorize the problem which is problematic when new data is introduced to the model. Based on these two tests, I am going forward and building two quadratic regression models, based on each of the variables so that I can compare them later on.

1.2.2 Building Models

Having figured out what kind of model to use, we can now split our data into a training set and a testing set. We do this so we can train our model with the training set, and then use the test set to test the accuracy of our model. I do this by using the sklearn library in python to randomly split my data 50/50, so I use 50% of the data to train the model, on the other 50% to test my model. After doing this I used the training model to fit my quadratic model and plotted it into a scatter plot to visualize the regression line.

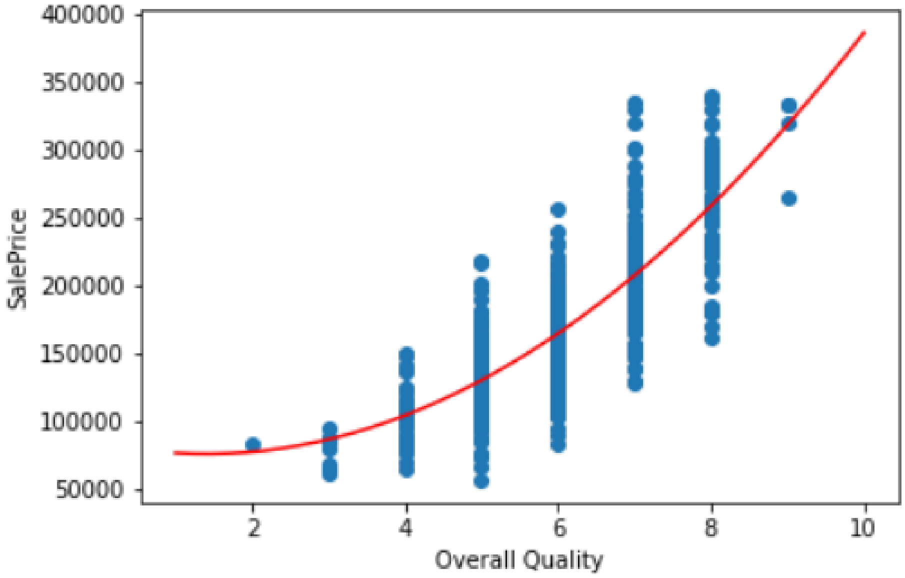
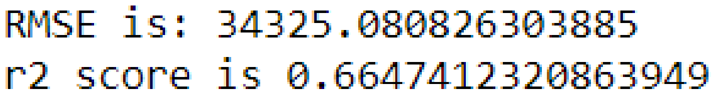


Figure 10: Regression line on training model

After that I computed the r2 score and RMSE (root mean squared error) on the training model to check the accuracy of the model, which you can see in figure 11. The r2 score indicates how well the model fit, the results here this gives us $r2 = 0,66$. Or around 66%.



```
RMSE is: 34325.080826303885
r2 score is 0.6647412320863949
```

Figure 11: r2 score and RMSE for training data

1.2.3 Testing The Models

Before I test the model I plotted the testing data into a scatter plot to get a visual look on the test data with the regression line as seen in Figure 12

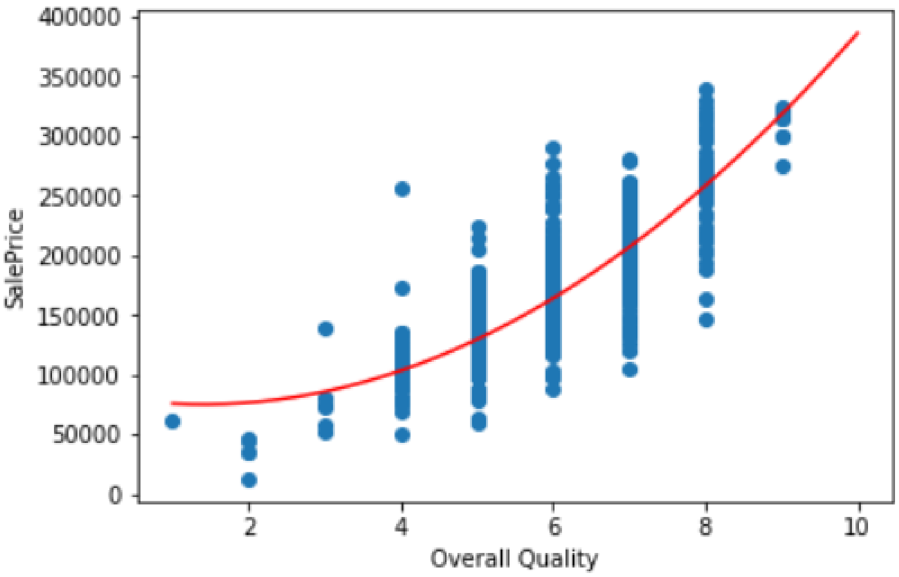
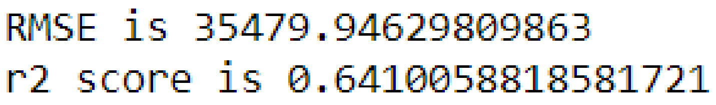


Figure 12: Scatter plot of testing data with regression line

I then computed the RMSE and r2 score when using the testing data.Figure 13 shows the results, Where we got $r2 = 0,64$ and $RMSE = 35479$



```
RMSE is 35479.94629809863
r2 score is 0.6410058818581721
```

Figure 13: r2 score and RMSE for testing data

I used the same process to create a model with the Gr.Liv.Area (living area square feet) as the predictor variable as well. Just to compare the two models. Because I did the exact same process just with another variable, so figure 14 and figure 15 show the results of the analysis.

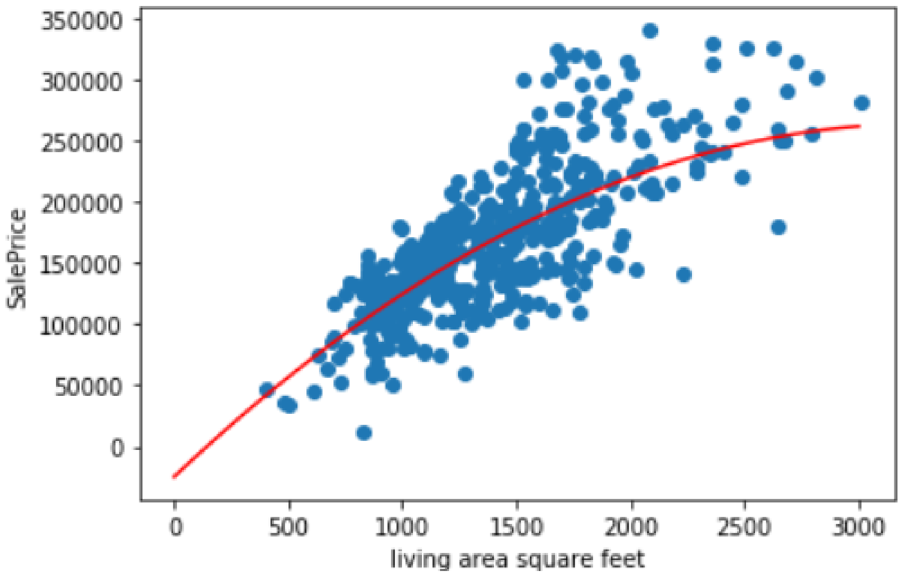
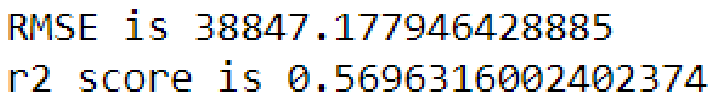


Figure 14: Regression line on test data when using Gr.Liv.Area (living area square feet) as predictor variable.



```
RMSE is 38847.177946428885
r2 score is 0.5696316002402374
```

Figure 15: r2 score and RMSE for testing data when using Gr.Liv.Area (living area square feet) as predictor variable.

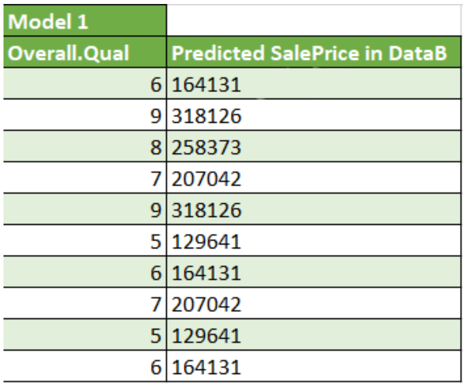
As predicted by the K-Fold cross validation it preforms worse than the other model. With an RMSE of 38847 and $r2 = 0,56$

1.3 Discussion

1.3.1 Predicting sales prices in dataset DataB

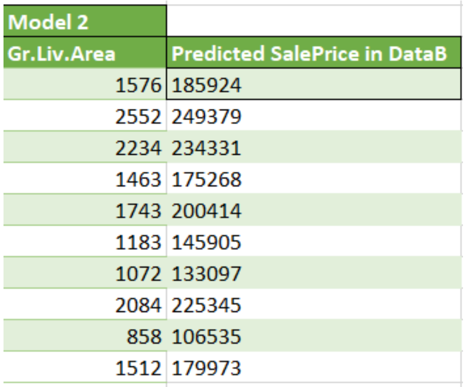
Now that we have both of our models, we can start predicting the sales prices of the houses within dataset **DataB**.
For clarification Model 1, is the model where Overall Quality as predictor variable is used.
Model 2 is the model where Gr.Liv.Area (living area square feet) as predictor variable is used

Figure 16 shows the table of predicted sales prices ± RMSE in dataset **DataB** when using model 1, while figure 17 show the table of predicted sales prices ± RMSE in dataset **DataB** when using model 2.



Model 1	
Overall.Qual	Predicted SalePrice in DataB
6	164131
9	318126
8	258373
7	207042
9	318126
5	129641
6	164131
7	207042
5	129641
6	164131

Figure 16: Model 1 sales prices predictions on houses in dataset DataB



Model 2	
Gr.Liv.Area	Predicted SalePrice in DataB
1576	185924
2552	249379
2234	234331
1463	175268
1743	200414
1183	145905
1072	133097
2084	225345
858	106535
1512	179973

Figure 17: Model 2 sales prices predictions on houses in dataset DataB

1.3.2 Observations

If we take a look at the tables of the predicted sales prices above, I realized that when using living area square feet as the predictor value we get different sales prices for each house. But when using overall quality, we get houses with the same predicted sales price because they have the overall quality. Now this made me wonder, what happens with my models prediction if a small house around 800 square feet, has an overall quality of 9 or 10? My model would predict a price of 318 126 ± RMSE, which sounds very expensive for such a small house. I realize that to make a more accurate model we need to use more predictor variables. By using multiple regression we can build a model that uses both living area square feet and overall quality as predictor variables. We can also include more variables that has a strong correlation to sales prices.

1.4 Conclusion

The analysis shows that model 1, where we use the overall quality as predictor variable is a better fit than model 2 which uses Gr.Liv.Area (living area square feet) as the predictor variable. While an r2 score of $0,64$ is within acceptable limits, The model can be improved by including more predictor variables. The correlation heatmap show us a total of 6 other variables with a pearson correlation coefficient of r over 0,5, which suggest a strong positive correlation that can be used as predictor variables. So further analysis with more predictor variables should be done to build a more accurate model.