Linear Regression Final Project - Diamonds

Wan-Chun Liao, Daniel Tinoco, Cho Hsun Yang

Abstract

What makes one diamond more expensive than the other? The team plans to approach this question using linear regression.

Linear regression is a linear approach to modeling the relationship between a response and predictor(s). We applied this modeling process on a Diamond dataset to create a model that infers the price based on the physical properties (predictors) of each diamond. The final model also suggests which predictors are most significant in predicting the price.

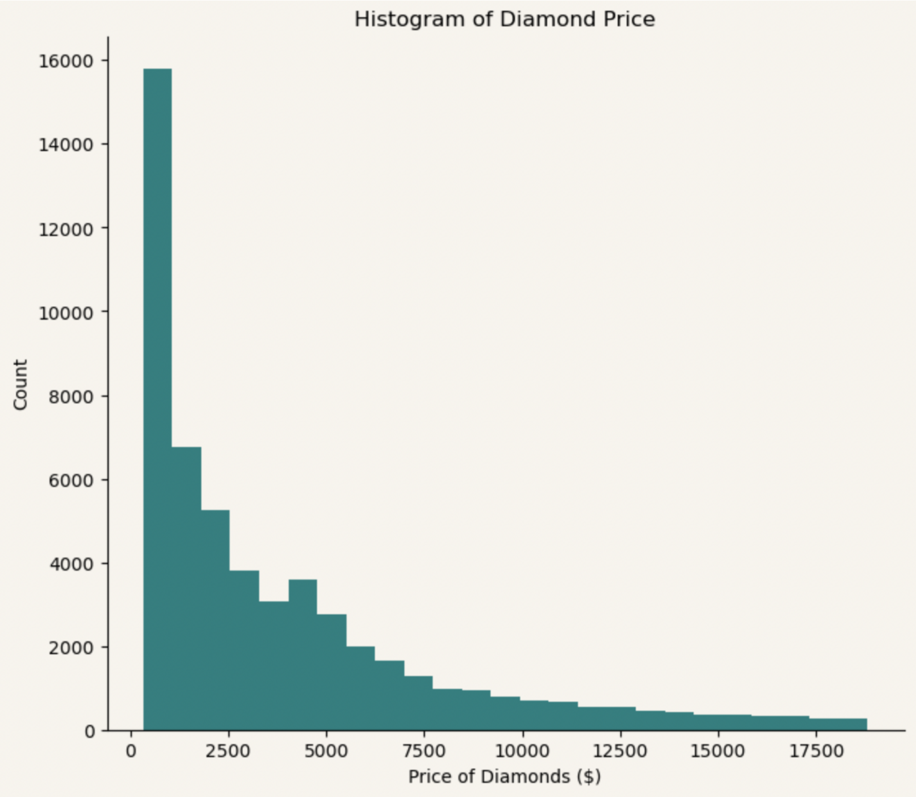
Dataset Description

The selected dataset for our project is Diamond Dataset, which contains observations about diamonds and their corresponding prices. A total of 53490 instances and 10 predictors are included in the dataset, and the linear regression task is to model the relationship between a response variable which is diamond price and potential predictors variables. Details of each attribute are as follows:

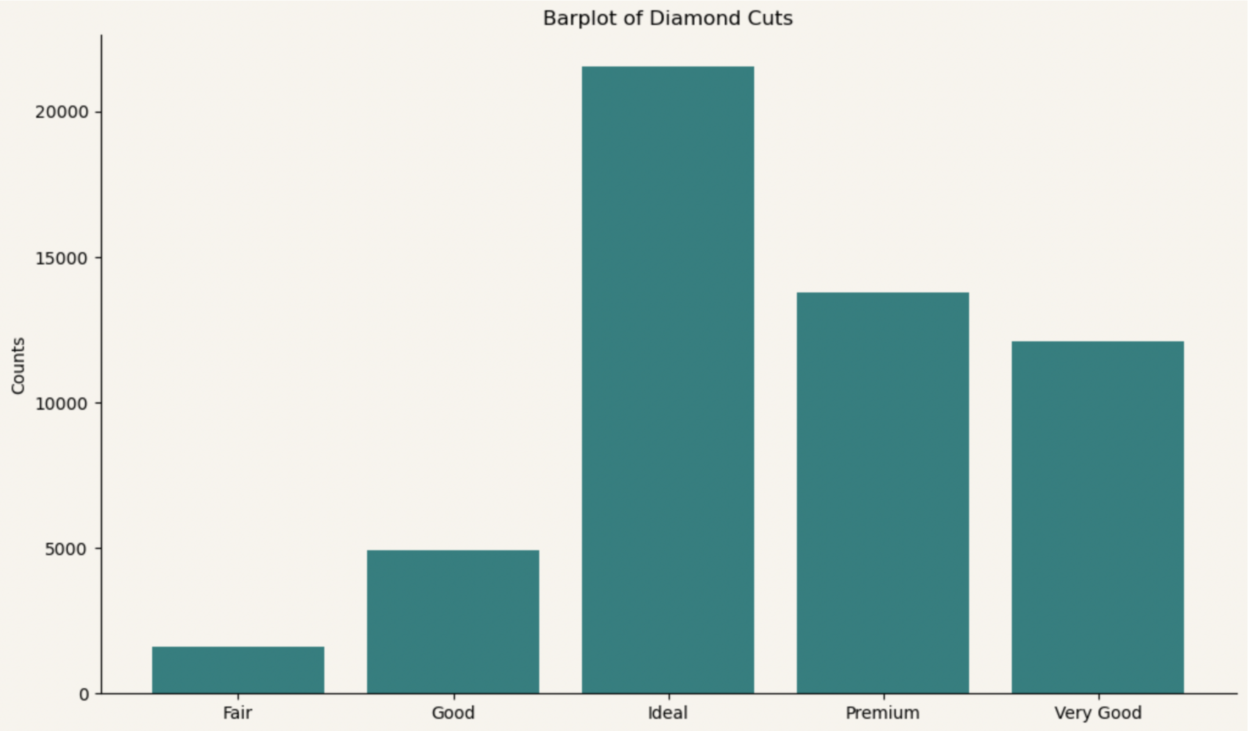
1. price: in US dollars ($326 - $18,823)
2. carat: weight of the diamond (0.2--5.01)
3. cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
4. color: diamond color, from J (worst) to D (best)
5. clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
6. x: length in mm (0--10.74)
7. y: width in mm (0--58.9)
8. z: depth in mm (0--31.8)
9. depth: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
10. table: width of top of diamond relative to the widest point (43--95)

After investigating the dataset, notably, there is no missing value in all the instances.

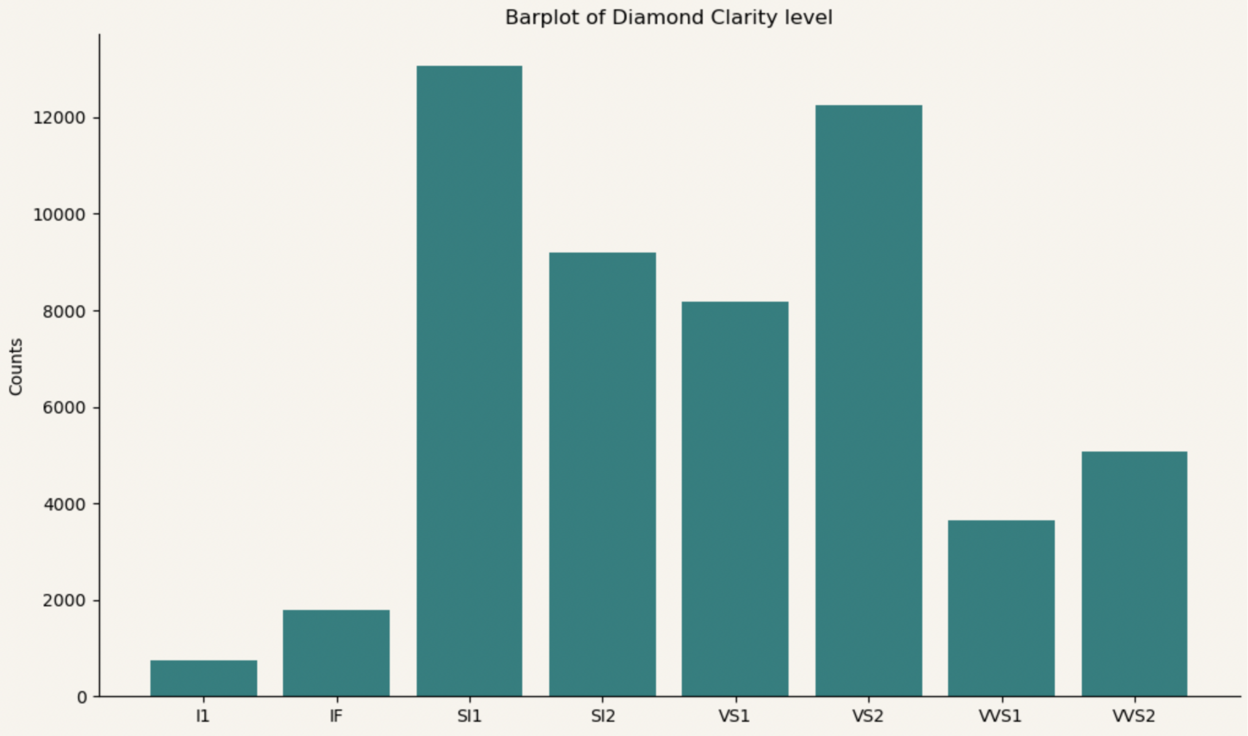
Below is the histogram of the response, showing that price is roughly exponentially distributed; hence, it is not normally distributed, which will come to affect our process for determining the final model.



Below is the bar plot of diamond cuts, where we note the significant disparity between certain categories of diamonds.



Below is the bar plot of diamond clarity levels, where we once again note an uneven distribution.



Data Preprocessing

The dataset we collected is marked in the correct units and contains no missing values. While we take note of the distributions for price, cut and clarity, we will move on for now and seek to address them as we go.

Initial Modeling

In this section, the team has created the first iteration of the linear regression model using all the predictors contained within:

Full model: Price ~ carat + cut + color + clarity + depth + table + x + y + z

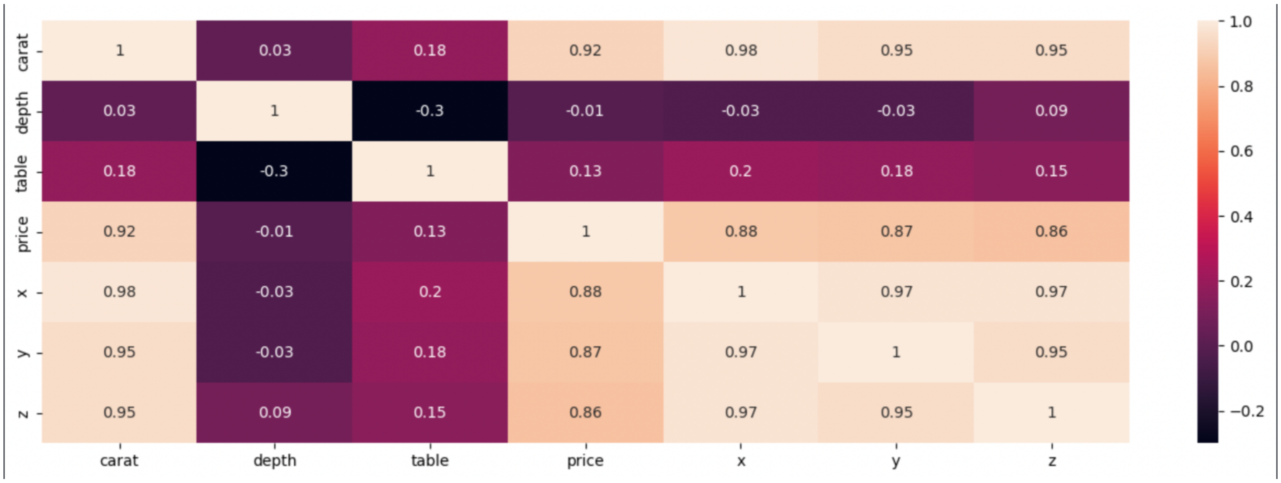
Model Diagnostic

Model diagnostic is an important step in building the model, to identify any problems that can hurt the model performance and inferences. The issues can be generally divided into data structure issues and linear regression model assumption violations.

Multicollinearity

1. **Plot - heat map of pair-wise correlation between predictors**

The first step is to check if there is multicollinearity between the predictors. We accomplished this by plotting out a **heat map** between all predictors, as seen below:



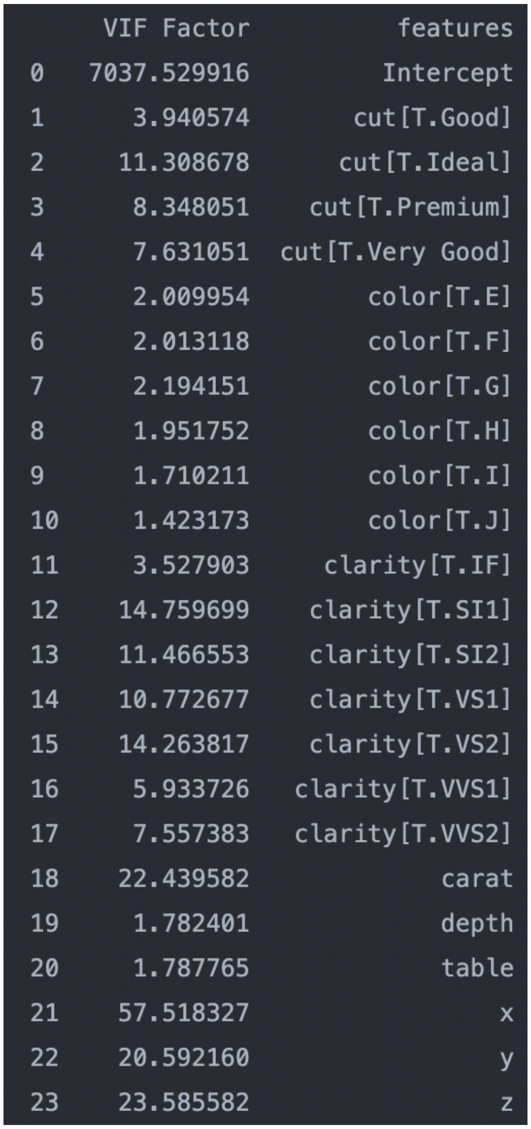
	carat	depth	table	price	x	y	z
carat	1	0.03	0.18	0.92	0.98	0.95	0.95
depth	0.03	1	-0.3	-0.01	-0.03	-0.03	0.09
table	0.18	-0.3	1	0.13	0.2	0.18	0.15
price	0.92	-0.01	0.13	1	0.88	0.87	0.86
x	0.98	-0.03	0.2	0.88	1	0.97	0.97
y	0.95	-0.03	0.18	0.87	0.97	1	0.95
z	0.95	0.09	0.15	0.86	0.97	0.95	1

The higher the number on the scale, the higher the correlation between the 2 predictors. We can see here that x, y, z, and carat are highly correlated with each other.

2. Variance Inflation Factor(VIF)

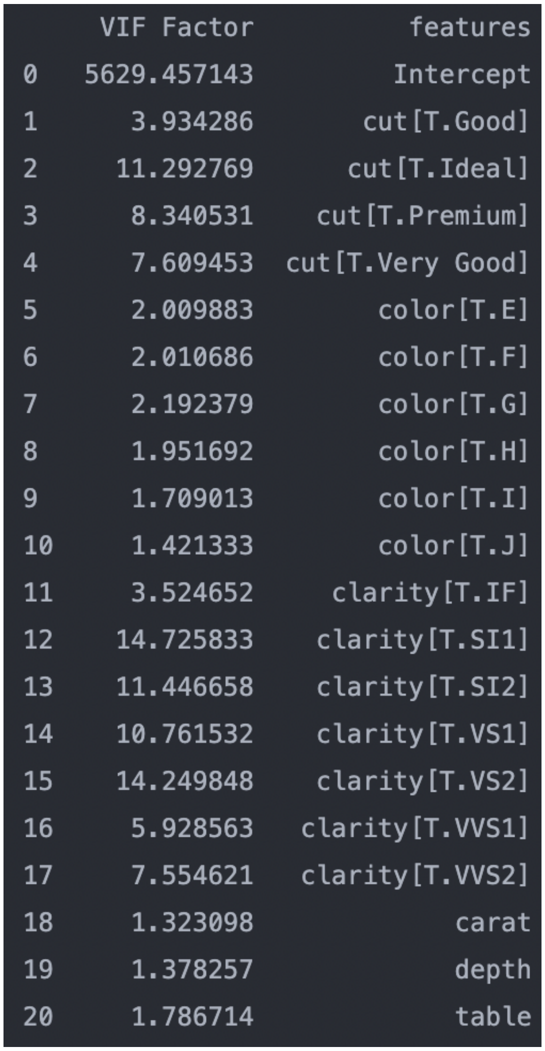
VIF measures how much the variance is inflated in the coefficient estimates. So, we checked with the VIF value. It should be noted that if 1 < VIF <=4, the predictors have a light correlation with other predictor(s) in the model. And if 4 < VIF <= 10, the predictors have moderate correlation. Notably, the predictors of x, y, z, and carat having the highest value(>=10), showing the highly impacted by the multicollinearity. So, we might suggest the categorical variable is problematic.

We also notice that we have high VIF for the intercept: it could be caused by two reasons. First: it’s the data scale; second: it could be due to any categorical predictor being highly imbalanced.



```
     VIF Factor          features
0   7037.529916         Intercept
1      3.940574      cut[T.Good]
2     11.308678     cut[T.Ideal]
3      8.348051   cut[T.Premium]
4      7.631051 cut[T.Very Good]
5      2.009954        color[T.E]
6      2.013118        color[T.F]
7      2.194151        color[T.G]
8      1.951752        color[T.H]
9      1.710211        color[T.I]
10     1.423173        color[T.J]
11     3.527903     clarity[T.IF]
12    14.759699    clarity[T.SI1]
13    11.466553    clarity[T.SI2]
14    10.772677    clarity[T.VS1]
15    14.263817    clarity[T.VS2]
16     5.933726   clarity[T.VVS1]
17     7.557383   clarity[T.VVS2]
18    22.439582             carat
19     1.782401             depth
20     1.787765             table
21    57.518327                 x
22    20.592160                 y
23    23.585582                 z
```

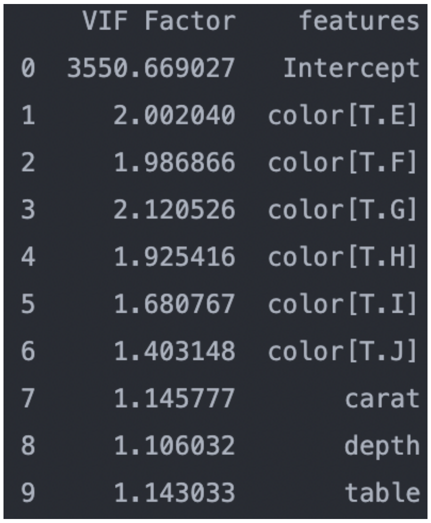
To solve this, we removed the x, y, z predictors since depth is a measurement consisting of these 3 values already. The result looks much better with no more high correlation as seen below. The high values in clarity and cut present are likely due to the data imbalance problem we detailed in the earlier section. We will remove these 2 predictors from our analysis.



```
     VIF Factor          features
0   5629.457143         Intercept
1      3.934286       cut[T.Good]
2     11.292769      cut[T.Ideal]
3      8.340531    cut[T.Premium]
4      7.609453  cut[T.Very Good]
5      2.009883        color[T.E]
6      2.010686        color[T.F]
7      2.192379        color[T.G]
8      1.951692        color[T.H]
9      1.709013        color[T.I]
10     1.421333        color[T.J]
11     3.524652     clarity[T.IF]
12    14.725833    clarity[T.SI1]
13    11.446658    clarity[T.SI2]
14    10.761532    clarity[T.VS1]
15    14.249848    clarity[T.VS2]
16     5.928563   clarity[T.VVS1]
17     7.554621   clarity[T.VVS2]
18     1.323098             carat
19     1.378257             depth
20     1.786714             table
```

After removing clarity and cut, the VIF of the intercept decreased a bit more from 5629 to 3551.

The VIF values of all other predictors are within the light range as well.



```
    VIF Factor    features
0  3550.669027   Intercept
1     2.002040  color[T.E]
2     1.986866  color[T.F]
3     2.120526  color[T.G]
4     1.925416  color[T.H]
5     1.680767  color[T.I]
6     1.403148  color[T.J]
7     1.145777       carat
8     1.106032       depth
9     1.143033       table
```

Our new base model will be:

Price ~ carat + color + depth + table

Check for Influential Points

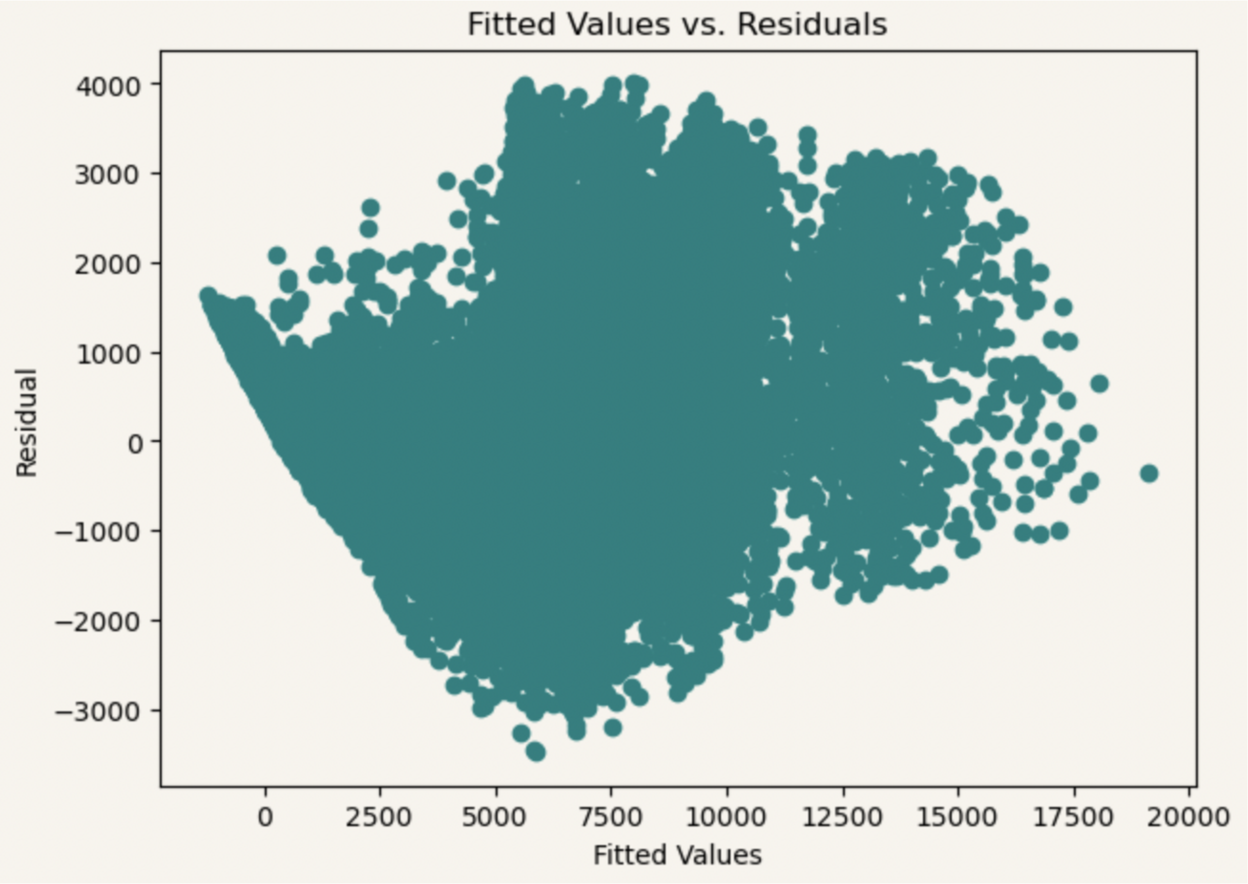
Influential points are points that are outliers where the response value is very far from the other points and fitted values. These points can also be points with high leverage, which can significantly alter the model estimation, leading to very different models with and without these influential points.

One way to determine influential points is to run the model with and without each point, and see how big the difference is between the 2 different models. If the difference is larger than 4 divided by the total number of points, then we can classify it as an influential point.

From the 53490 observations, 3422 of them are flagged as influential points, meaning around 6.34% of the points are influential points. We will remove them for the cleaned final model and compare them with the final model with the influential points to see the difference in performance in the final model section.

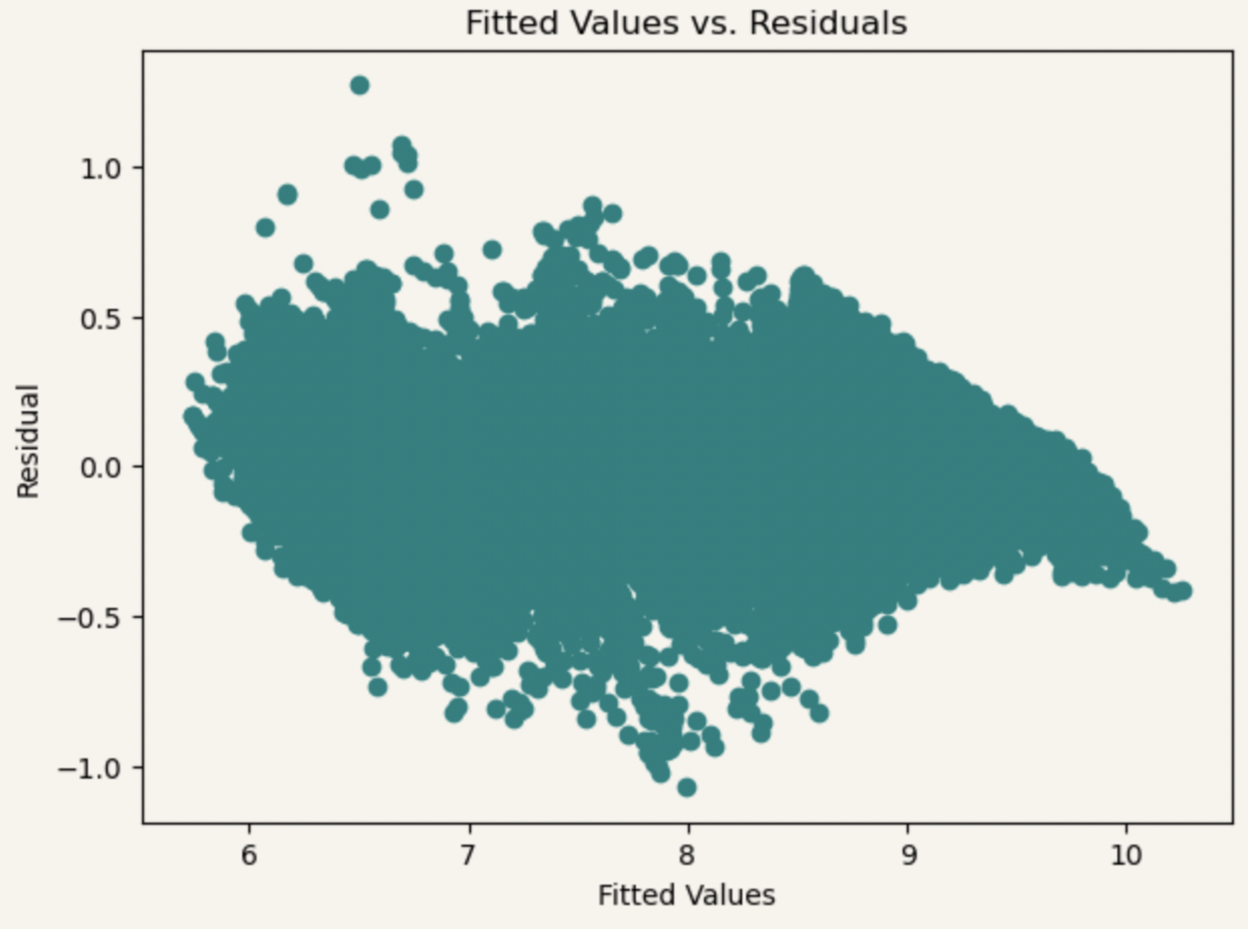
Examine for Heteroscedasticity

In Ordinary Least Squares Estimates for regression, the variance of errors should be constant. Heteroscedasticity happens when this assumption is violated.



While the residuals above are almost evenly distributed around the center, there are certain outlines of a shape to this graph that suggest our assumption of constant variance may be violated. In fact, when we run the Breusch-Pagan test for heteroscedasticity, we find that our model rejects the null hypothesis, or assumption, that our residuals have constant variance. We can see that there is a trend of increase in variance as we move from 0~5000 fitted values, this illustrates that the residuals don’t have constant variance.

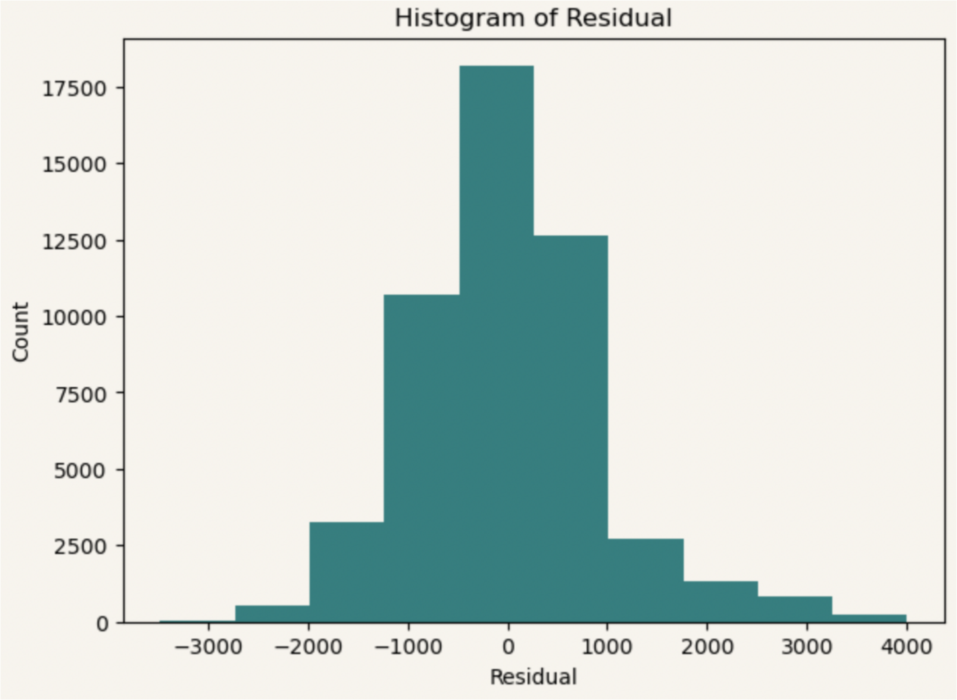
To address this, we applied a log transformation on the response price and the numerical predictor carat and are left with the following residual plot.



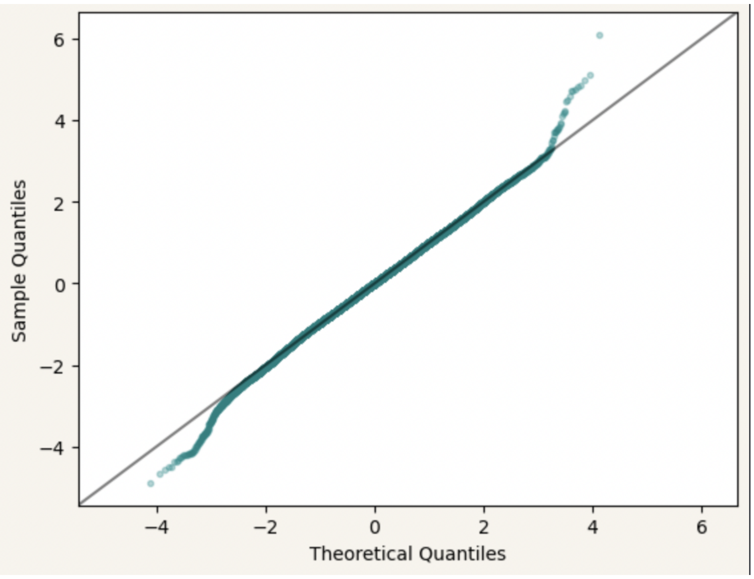
So far, this application of log transformations gives us the best graph for the residuals.

Verify Normality

Another assumption from the model we must observe is the normality of the residuals.



We can further check the normality assumption by viewing a QQ-plot for the residuals.

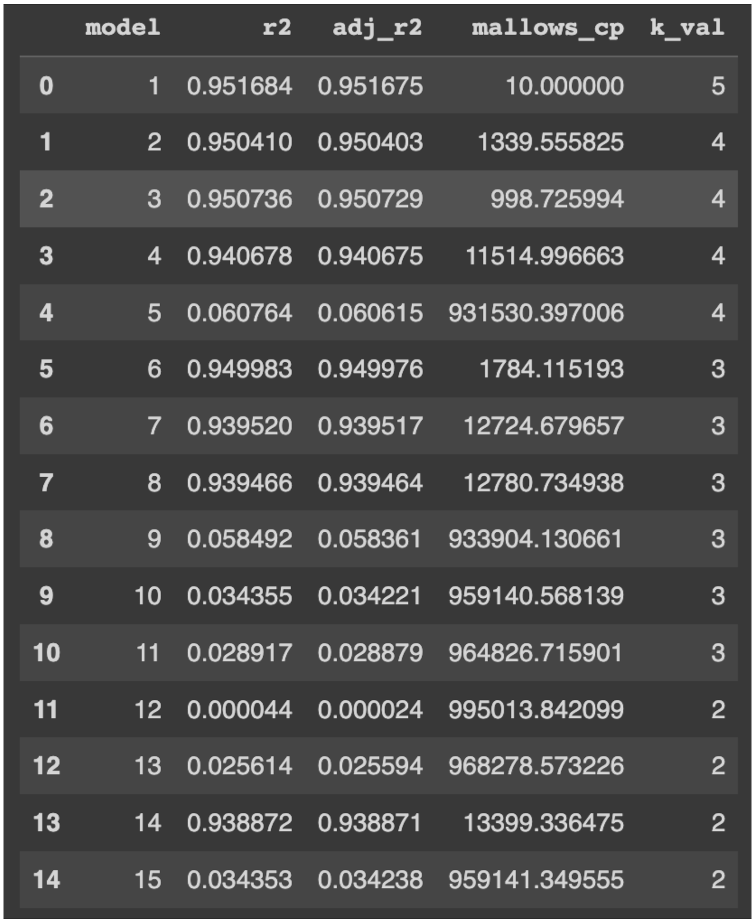


Most notably, we see that the tail ends of the residuals do not align with the rest of the graph, thus our assumption about the normality is still violated.

Model Selection

Now that we have an idea for what predictors we want to include in our model, we next want to identify them by importance with regards to their statistical significance as well as their prediction performance. To do this, we can begin with a null model containing no predictors apart from the coefficient and then add predictors in a series of steps that maintains their significance as mentioned above. The measurements we will use are Mallow's Cp, AIC, BIC and to a lesser extent R-squared and adjusted R-squared.

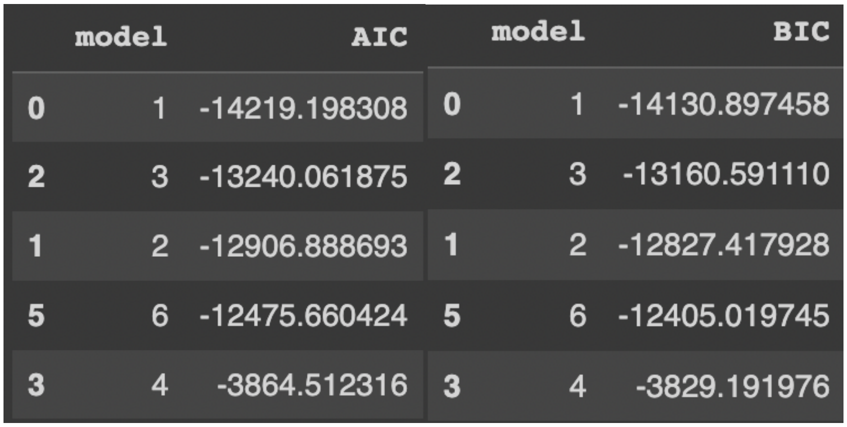
R-squared, Adjusted R-squared, and Mallow's Cp



	model	r2	adj_r2	mallows_cp	k_val
0	1	0.951684	0.951675	10.000000	5
1	2	0.950410	0.950403	1339.555825	4
2	3	0.950736	0.950729	998.725994	4
3	4	0.940678	0.940675	11514.996663	4
4	5	0.060764	0.060615	931530.397006	4
5	6	0.949983	0.949976	1784.115193	3
6	7	0.939520	0.939517	12724.679657	3
7	8	0.939466	0.939464	12780.734938	3
8	9	0.058492	0.058361	933904.130661	3
9	10	0.034355	0.034221	959140.568139	3
10	11	0.028917	0.028879	964826.715901	3
11	12	0.000044	0.000024	995013.842099	2
12	13	0.025614	0.025594	968278.573226	2
13	14	0.938872	0.938871	13399.336475	2
14	15	0.034353	0.034238	959141.349555	2

Using Mallow's Cp, we find that the model with carat, color, depth, and table are all significant in determining the model.

Stepwise AIC and BIC



	model	AIC
0	1	-14219.198308
2	3	-13240.061875
1	2	-12906.888693
5	6	-12475.660424
3	4	-3864.512316

	model	BIC
0	1	-14130.897458
2	3	-13160.591110
1	2	-12827.417928
5	6	-12405.019745
3	4	-3829.191976

From the AIC and BIC table, we find that model 1 has the smallest AIC and BIC values, which is the same model chosen by using the Mallow’s Cp method. Thus, we determine that the best model for us to use is the following:

Log(price) ~ log(carat) + color + depth + table

Final Model

$$\log(price_i) = \hat{\beta}_0 + \hat{\beta}_{carat}\log(X_{i,carat}) + \hat{\beta}_{color}X_{i,color} + \hat{\beta}_{depth}X_{i,depth} + \hat{\beta}_{table}X_{i,table}$$

We take into consideration two versions of this model: one where we include the 3422 influential points found in the original model, and another where we exclude them to improve the normality and variance of the residuals. For the model containing the influential points, we have an adjusted R-squared of 0.947, and for the model without them, we have an adjusted R-squared of 0.952, meaning that in both cases, the goodness of fit of the model is quite high, likely due to the amount of data points. Moreover, the p-values for both cases are ~0, meaning the predictors are significant. Lastly the model with influential points has a Jarque-Bera test score of ~3252, while the model without the influential points has a Jarque-Bera score of ~300.

Potential Problems

While the final model has a high final R-squared value and meaningful predictors that are relevant to the model, we note that it still fails the Jarque-Bera test. This means that there is still significant heteroscedasticity in our model that cannot be further addressed.

Moreover, we also take note of the disparity between sampling among the different categorical variables, perhaps skewing the effects of any results for those observations compared to the few observations for less frequent categorical variables.

Conclusion

We find that the model ‘log(Price) ~ log(Carat) + Color + Depth + Table’ is the best model for determining the price of a diamond. We also consider the two versions of this model to highlight the differences that occur when we keep or discard the influential points found at the beginning of our process. When we exclude the influential data points from our model, we see a reduction in the heteroscedasticity problem of the residuals; however, ultimately, the sample size of this data set is still substantial enough for the statistical tests of the predictors to hold true.