AMOD5240H - DataSet Analysis

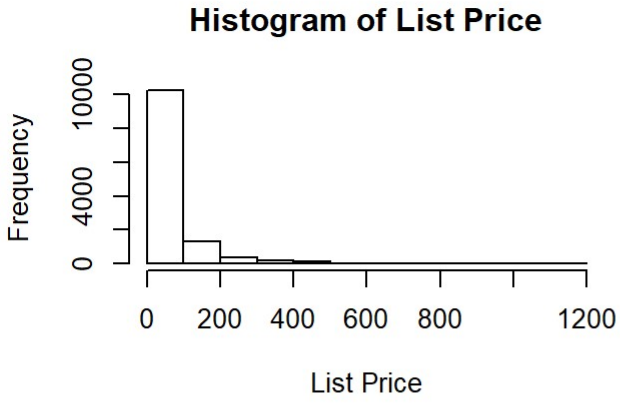
Suraj Suresh Sajjan

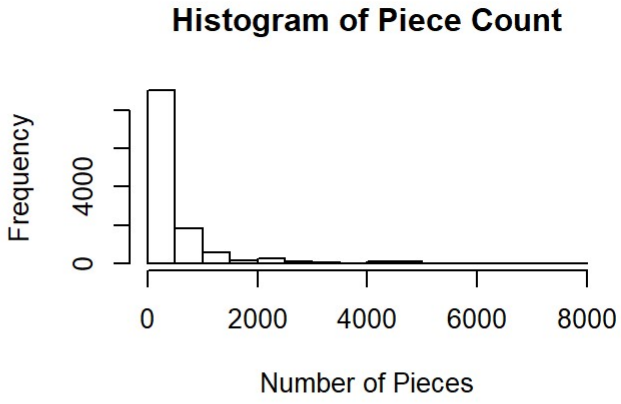
December 9, 2018

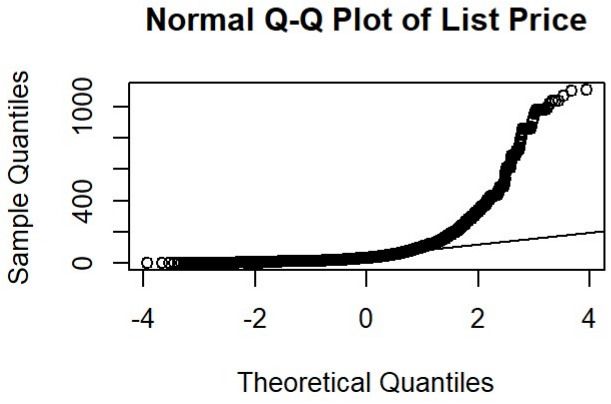
Description of the data-set:

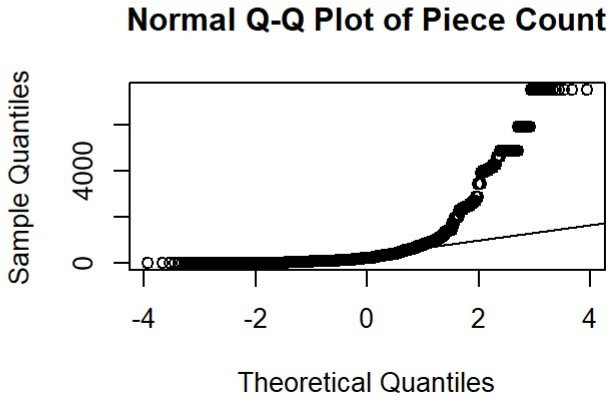
This is a Lego Sets data-set which was retrived from the kaggle datasets. The data-set contains 14 atributes and 10466 records after the omittion of rows that contain missing data points. Each lego-blocks set is identified with a unique product ID and set-name. There is a brief description of the toy set. For each toy set, the number of reviews it received, the number of pieces of lego-blocks, the price of the toy-set is also provided.
Each lego-toy set is further categorised based on the country of origin, the difficulty rating, the age group it is suitable for, the theme name and star ratings.

Appropriateness for statistical analysis:





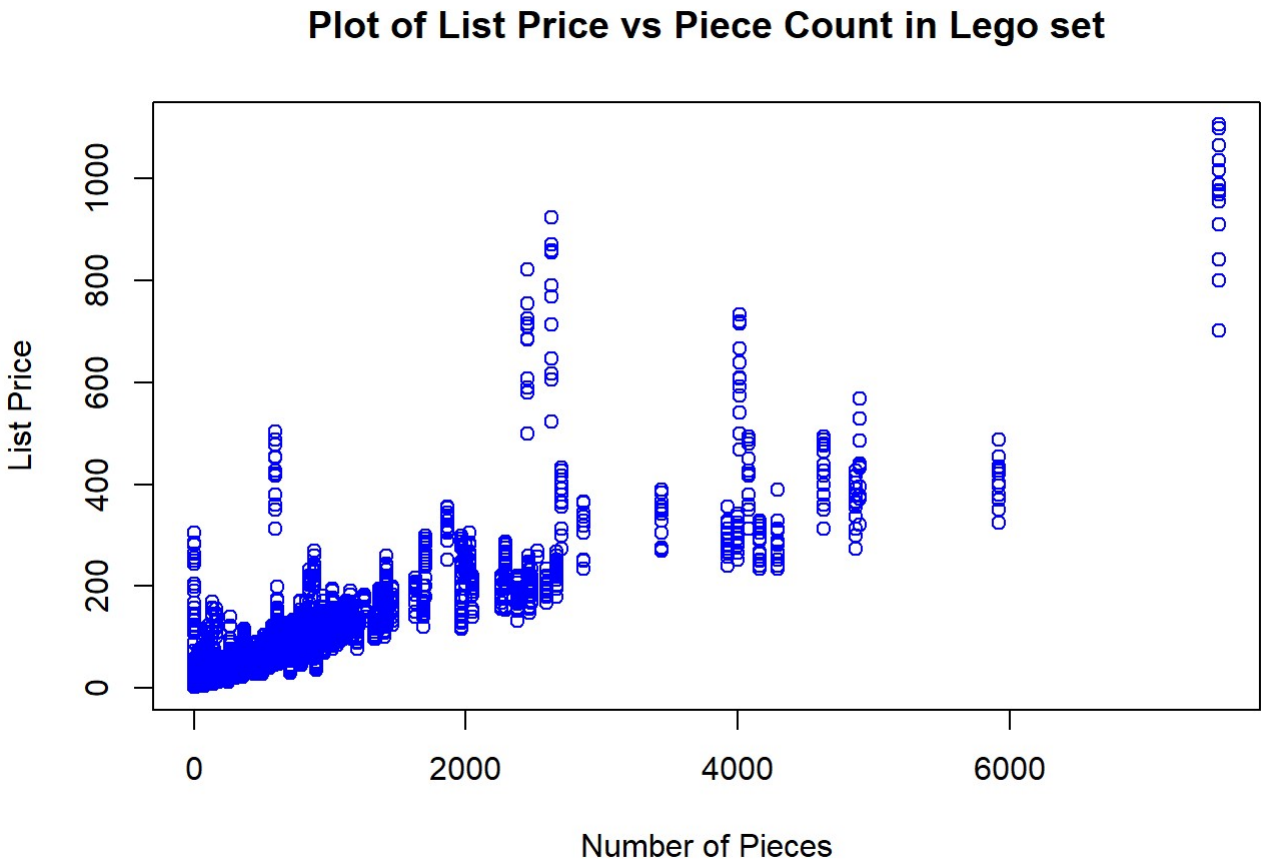




- From the QQnorm plots, we observe that the data points dont exactly follow the normality line.
- From the Histograms, we observe that the data is significantly right skewed.
  Hence the data does not seem to be collected at random. It certainly is a sample from the population as it contains only information on few of the lego-set toys.

Two Questions of interest:

1. Are the number of pieces in the lego set related to the price of the lego set?
2. Would subsetting by limiting the country to a single country and limiting the rating to “greater than 4.5” provide more clarity to the relation?

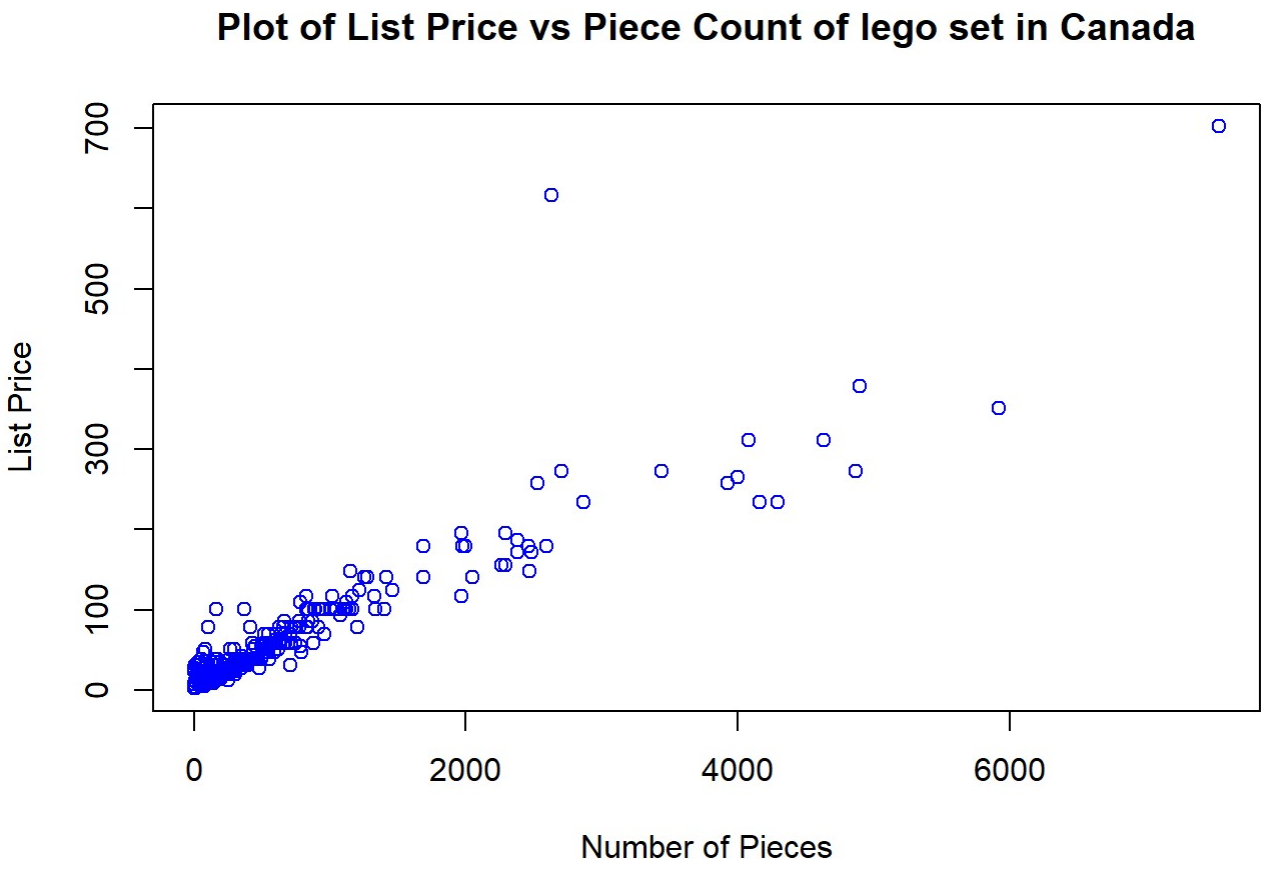


From the scatterplot, we observe that there seems to be a moderate strength between the points of number of pieces and the price of the lego set, with a positive trend. Since we are not sure if the price is measured in a single currency, and because the actual price of the product may vary across different countries, based on the forex exchange value, I chose to limit the lego-sets taken into consideration to only those toys available in Canada. I would also like to check the linearity of the data at a subset of data where the Lego-block sets have an above average rating of 4.5.

Data summary using R:

```
## The number of rows in the subset is: 462
```

```
## The correlation value is: 0.9270017
```



In the above plot, the subset of data including only the lego set sold in Canada with a star rating above 4.5 has been plotted. The subset contains 462 rows. We can observe from the plot that the points appear to be more correlated. Also the linear positive trend can be observed better. We can also observe 2 outliers in the plot.

The correlation value lies between -1 to 1. A value near -1 or 1 determines there is a good linear correlation between data. A value close to 0 determines a weak linear correlation between data. Looking at the correlation value of 0.928, we observe that the data is fairly linear.

Here x is independent variabe - piece count, y is dependent variable list price.

Data analysis using R:

```
lego_CA_model <- lm(y ~ x)
summary(lego_CA_model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -96.35  -8.00  -4.32   5.78 411.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.68608    1.42739   7.486 3.64e-13 ***
## x            0.07369    0.00139  53.011  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.38 on 460 degrees of freedom
## Multiple R-squared:  0.8593, Adjusted R-squared:  0.859
## F-statistic:  2810 on 1 and 460 DF,  p-value: < 2.2e-16
```

```
## The minimum piece count in the dataset is: 1 The maximum piece count in the datase
t is: 7541
```

Let us formulate Null and Alternate Hypothesis for the above condition.

Null Hypothesis: The slope of the line is zero, i.e., the price of the lego set is independent of the number of pieces.
Alternate Hypothesis: There exists a relation between the the price of the lego set and the number of pieces in it.

The intercept of the line is found to be 10.69. The value of slope is found to be 0.073.

The equation of the regression line is found to be List_Price = 0.073(Piece Count) + 10.69.
Where, List Price is the dependent or response variable and Piece Count is the independent or predictor variable.

Looking at the range of piece count in the above dataset, we see that the min piece count is 1 and the max piece count is 7541. To avoid extrapolation, we should not try using the above linear equation for values outside the range. For example, if the piece count is set to zero, the list price would be 10.69, which would not make sense.

The co-efficient t-value is the measure of the number of standard deviations our coefficient estimate is, from 0. If the value is far away from zero, we would reject null hypothesis.

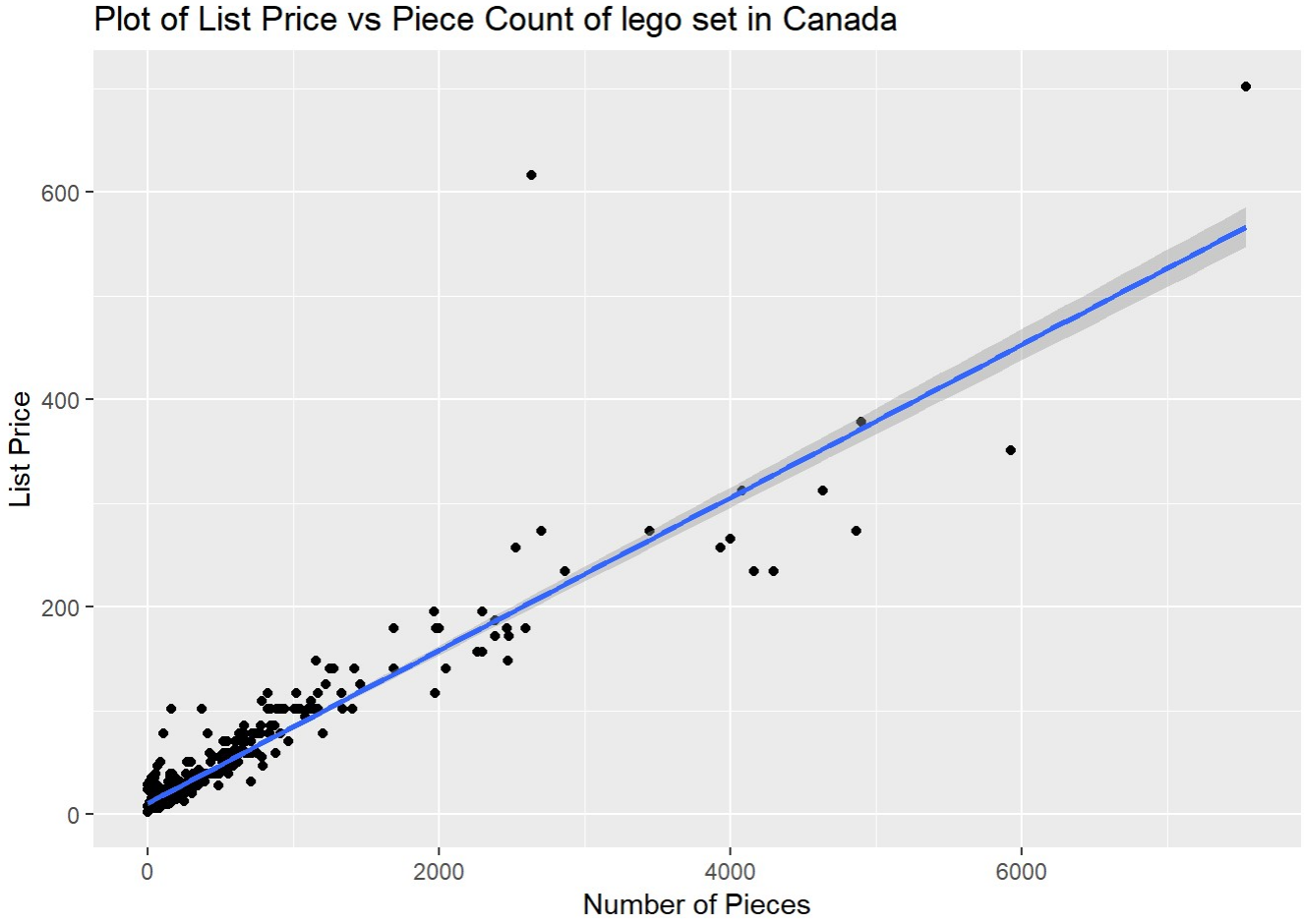
From the coefficient p-value of the slope x, we observe a value of 2.2 X 10^-6 which is much smaller than alpha of 0.05. Hence we reject the null hypothesis.

Residual standard error measures the quality of the linear equation we formulated. It denotes the average amount the response will deviate from the true regression line. The value here is 26.38.

The R-squared value provides a mesure of how well the model fits the actual data. The value close to 1 denotes a good fit and a value close to 0 determines a bad fit. In our case, the value is 0.86.

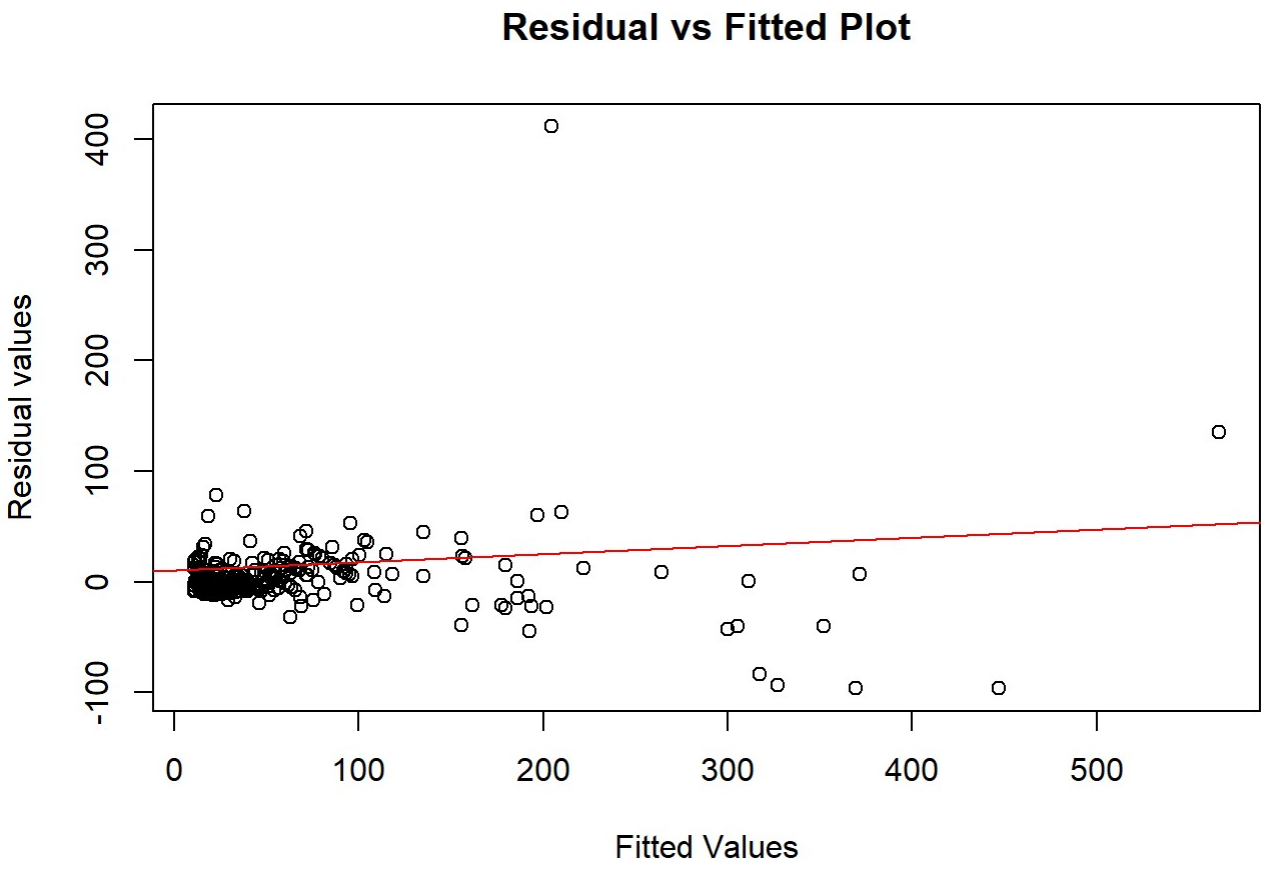
The F-statistic is another good measure of relation between the dependent and independent variables. The further the value of the F-statistic from 0, the better the relation. The F-statistic value here is 2810.

Now that we have a linear model, let us include the regression line in the scatterplot we created earlier.

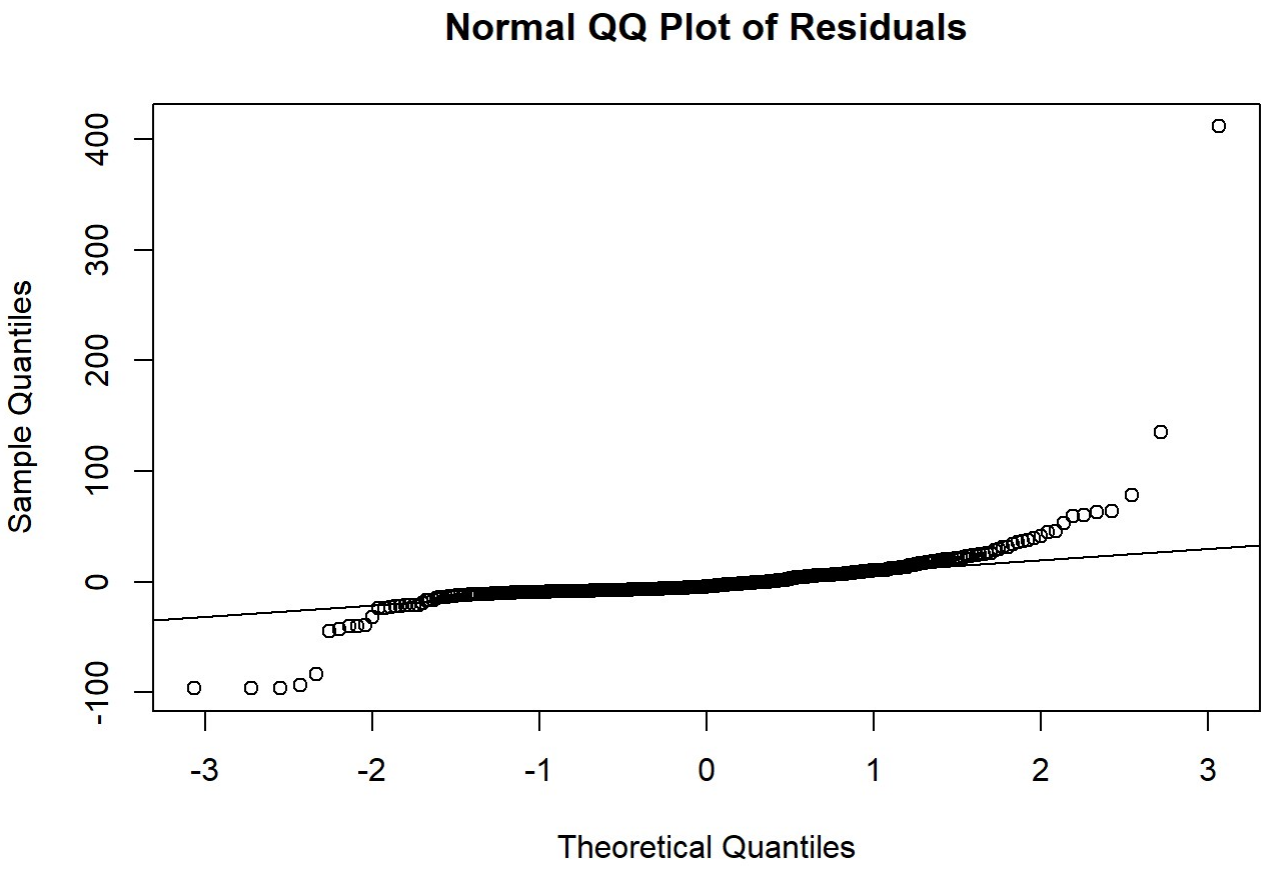


We can observe that the 2 outliers in the plot weakly influence the line of regression. Due to the presence of more outliers in the lower section of the regression line, their influence is reduced.

Now let us plot the linear model and analyse the graphical plots.



From the Residual vs fitted plot, we observe that the points are scattered in a random manner and the line is almost parallel to the X-axis.



From the Normal QQ Plot, we observe that some of the residuals follow the line closely, there are a couple of outliers. The residuals appear approximately normally distributed.

Conclusion:

In conclusion, keeping in mind that “causation is not co-relation”, we can state that the number of pieces in the lego-set is one of the factors that seems to affect the pricing of the lego-set with a positive co-relation. There exists a positve trend.

References:

[1] https://www.kaggle.com/mterzolo/lego-sets (https://www.kaggle.com/mterzolo/lego-sets)