**Statistical Learning Lab**

**Assignment – 1**

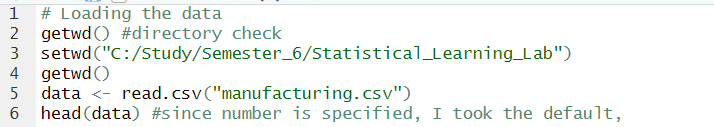
**Linear Regression Assignment**

**Name – Semanti Ghosh**

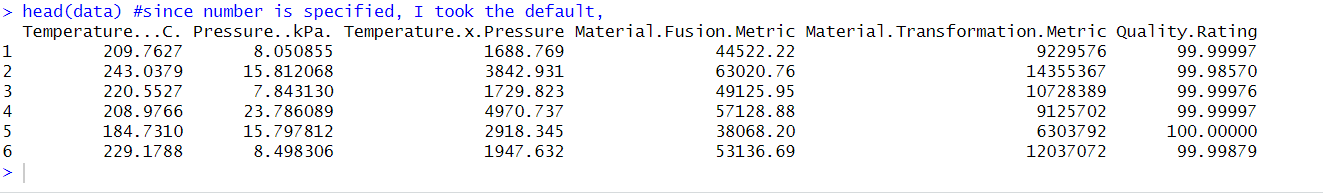
**Roll No – 22IM10036**

1. **Loading the dataset – “manufacturing.csv”**

Code for loading the data

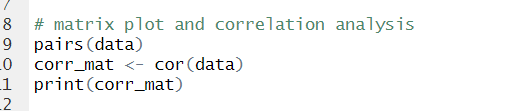


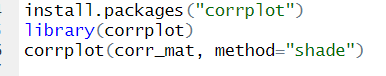
Data printed

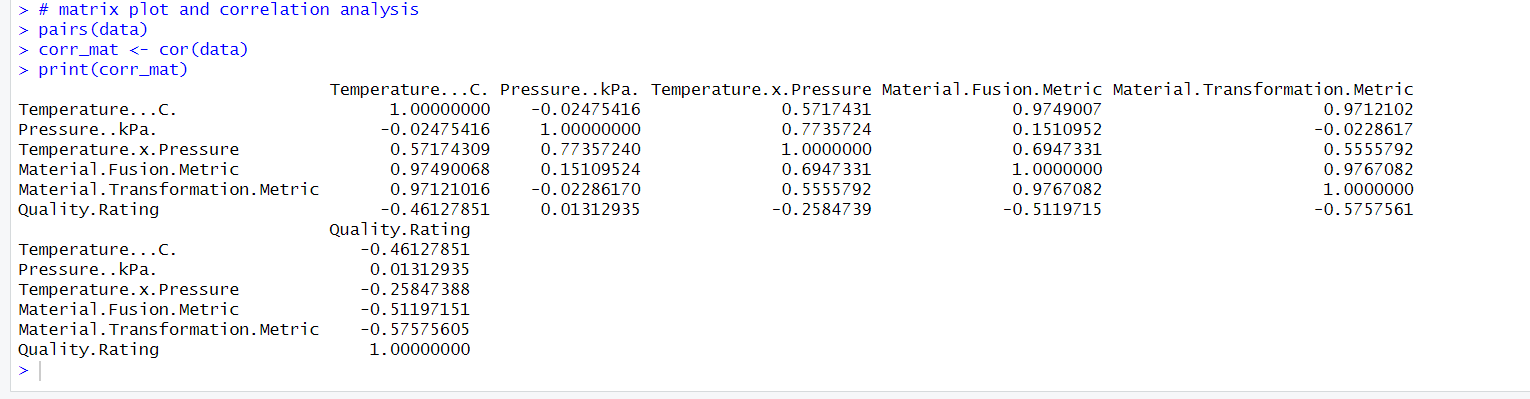


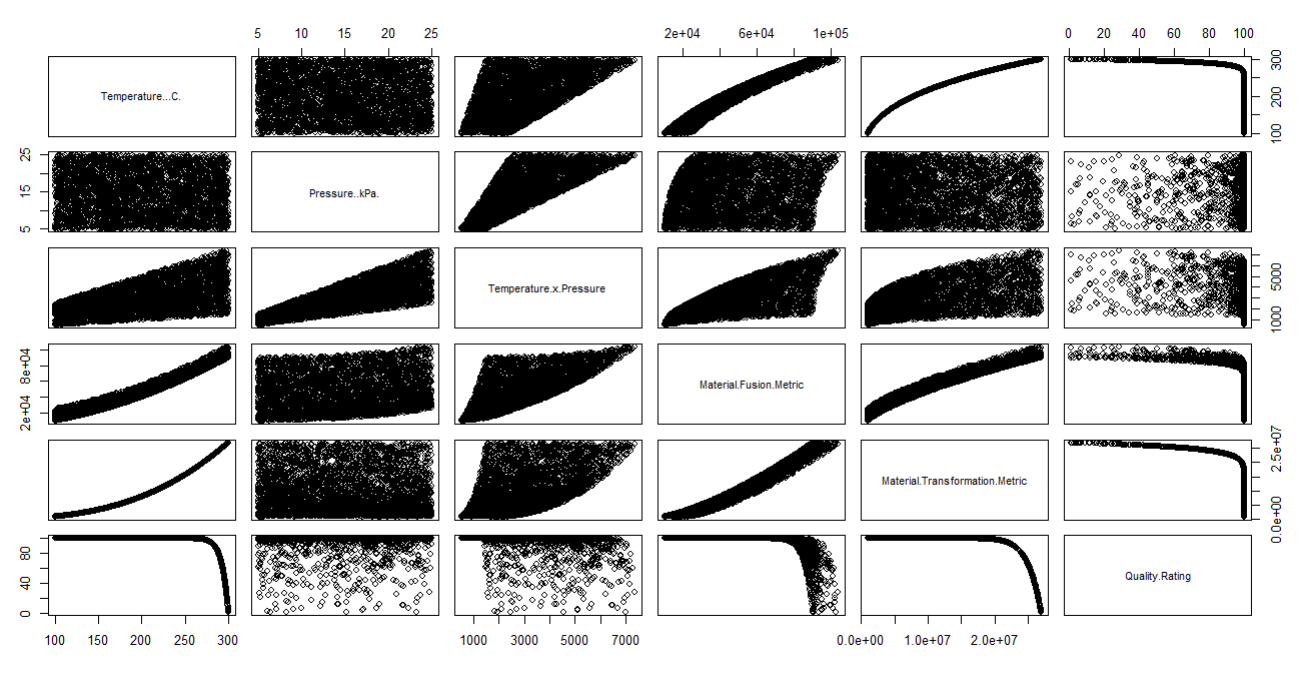
1. **Matrix plot and correlation analysis**

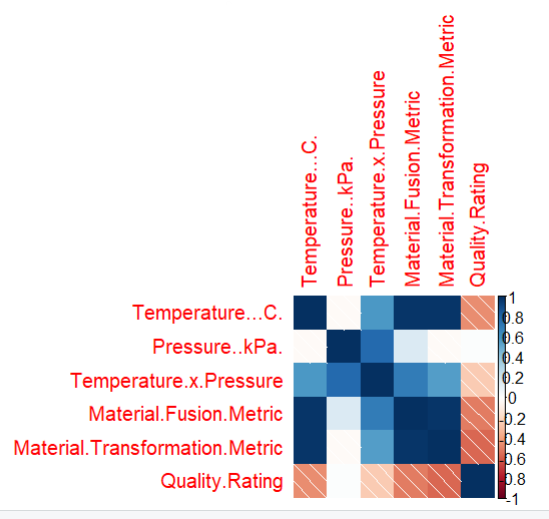
Code snippet





Output



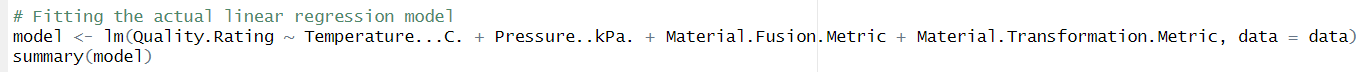
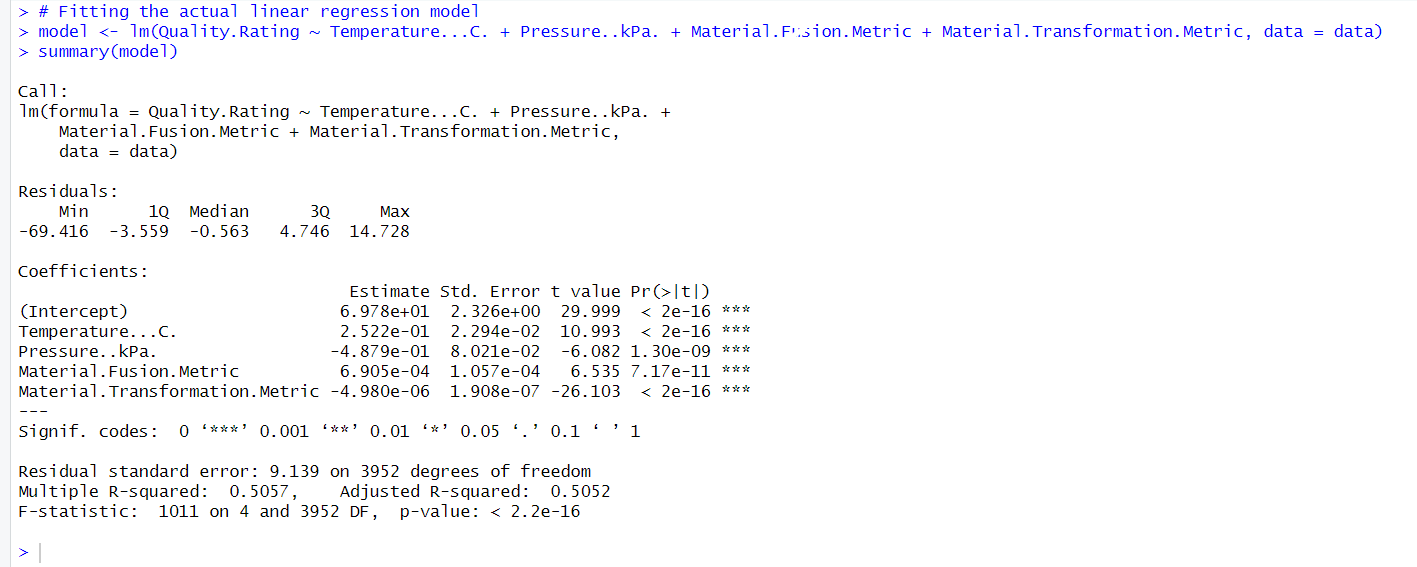


From the above plots, the factors that are correlated are:

* Temperature and material fusion metric
* Temperature and material transformation metric

Also, we can see there are factors even negatively correlated with each other (E.g. – quality rating and material transformation metric). There are some factors that are independent (E.g. – pressure and temperature).

1. **Fitting a linear regression model excluding the interaction term**

Code snippetModel summary 

From the values obtained from linear regression, we can see that the p values for the four parameters used in linear regression are less than 0.001. So, we can say that all four parameters are highly significant.

1. **Interpreting R2 and R2 adjusted**

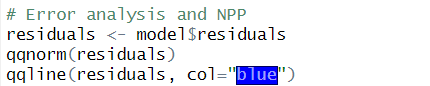
Over here, we got **Multiple R-squared = 0.5057** and **Adjusted R-squared = 0.5052**

This is acceptable because the value of R-squared is above 0.3. A value of R2 closer to 1 would have been preferred. R2 value of 0.5057 means that approximately 50.57% of the variation in the output (Quality.Rating) can be explained by the model.

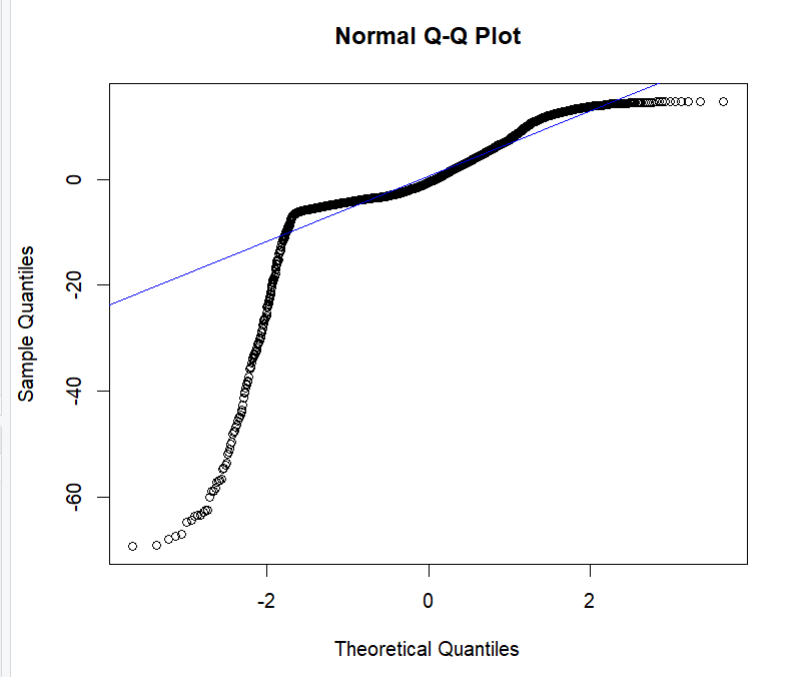
Also, the value of R2 adjusted is very slightly less than the R2 value. Since R2 adjusted accounts for the number of predictors in the model, this may suggest that some predictors do not contribute significantly. However, this shouldn’t be a problem since the difference is very less.

1. **Residual analysis and normal probability plot of residuals**

Code Snippet



**Normal Probability Plot**



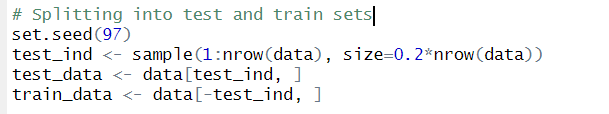
The basic assumption was that the error will follow normal distribution, however on performing NPP, we see that the errors (residuals) do not actually follow normal distribution.

Since our model was based on the above assumption, the model is not adequate.

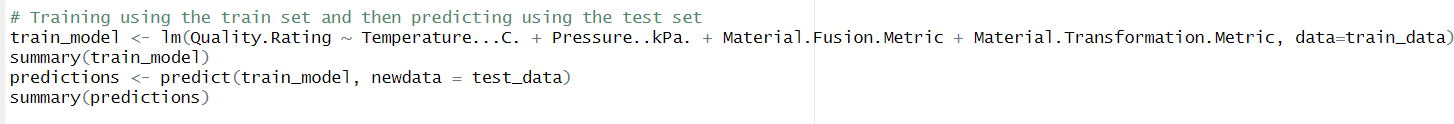
1. **Dividing into training and testing sets, performing linear regression then calculating the RMSE**

Code snippets

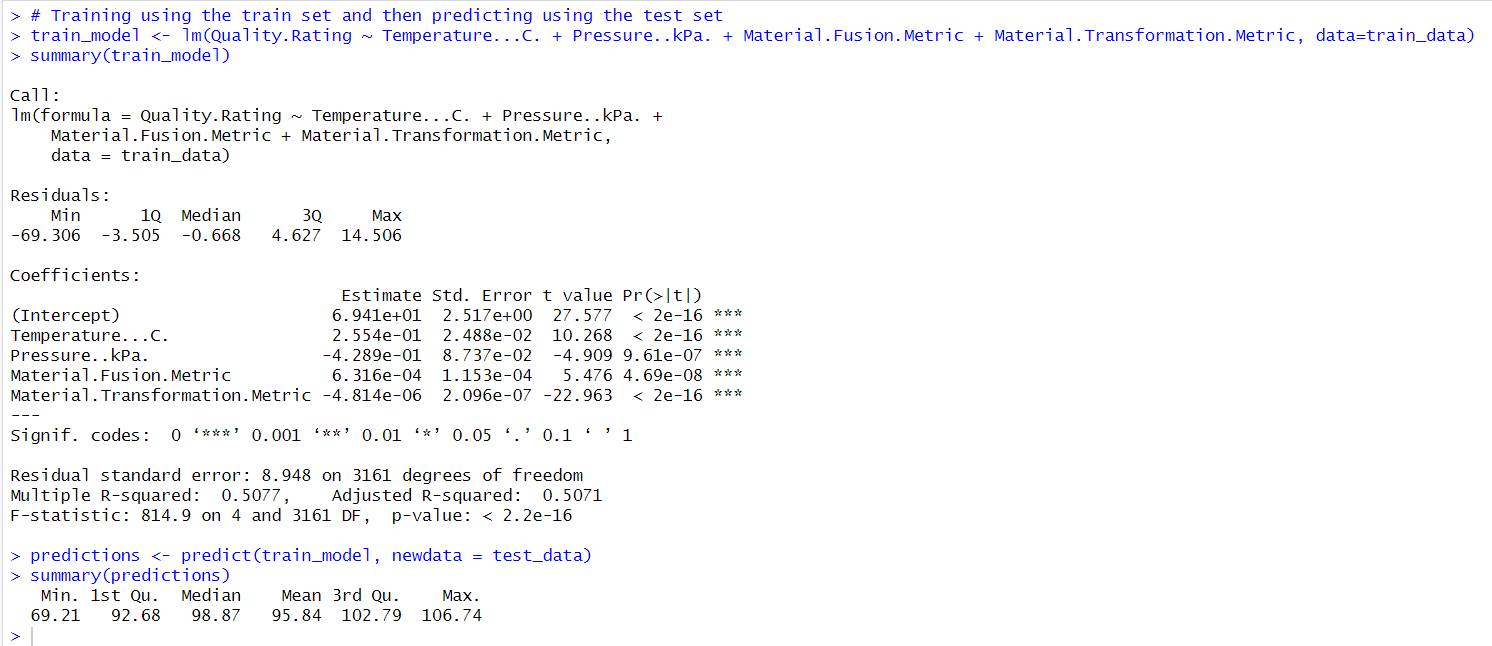
**Dividing into training and test sets**



**Training the linear regression model**

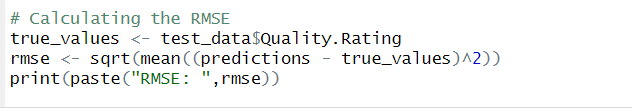


**Summary of predictions obtained**

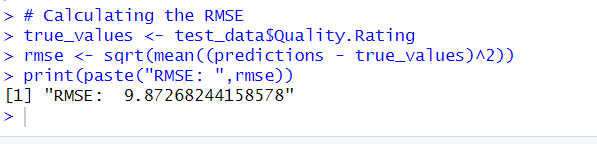


**Calculating RMSE of test data**

Code snippet



Output



**Root Mean Square Error on test data = 9.87268244158578**