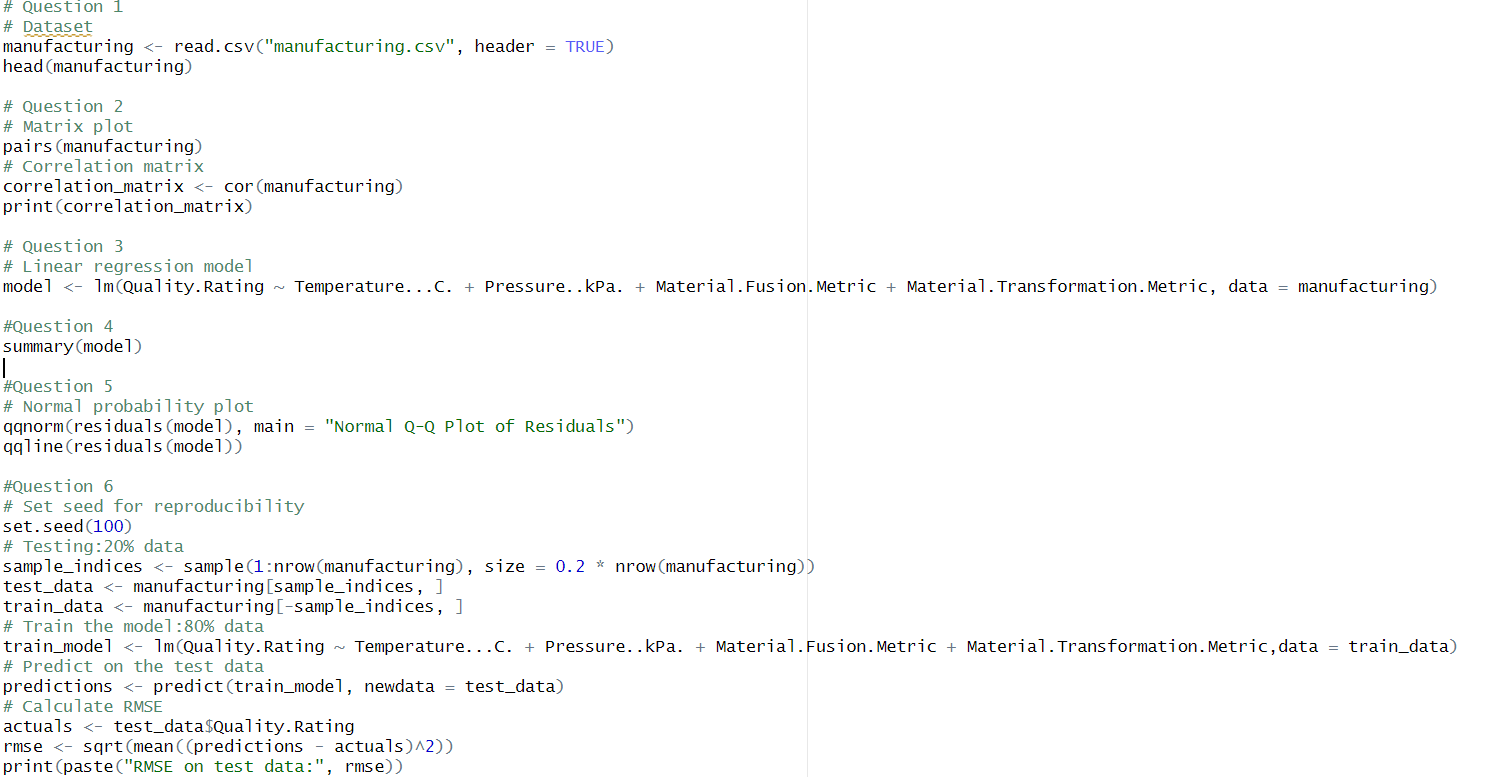
**Statistical Learning Lab**

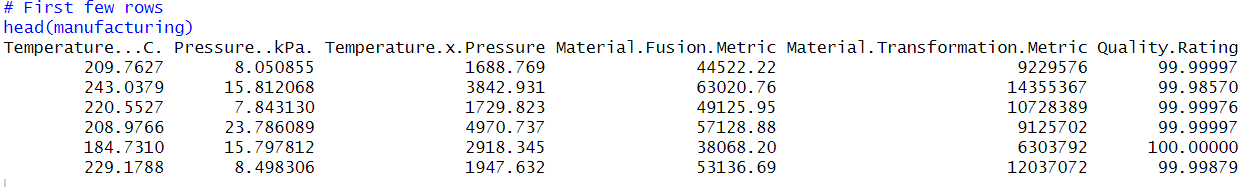
Assignment - 1

Linear Regression Assignment

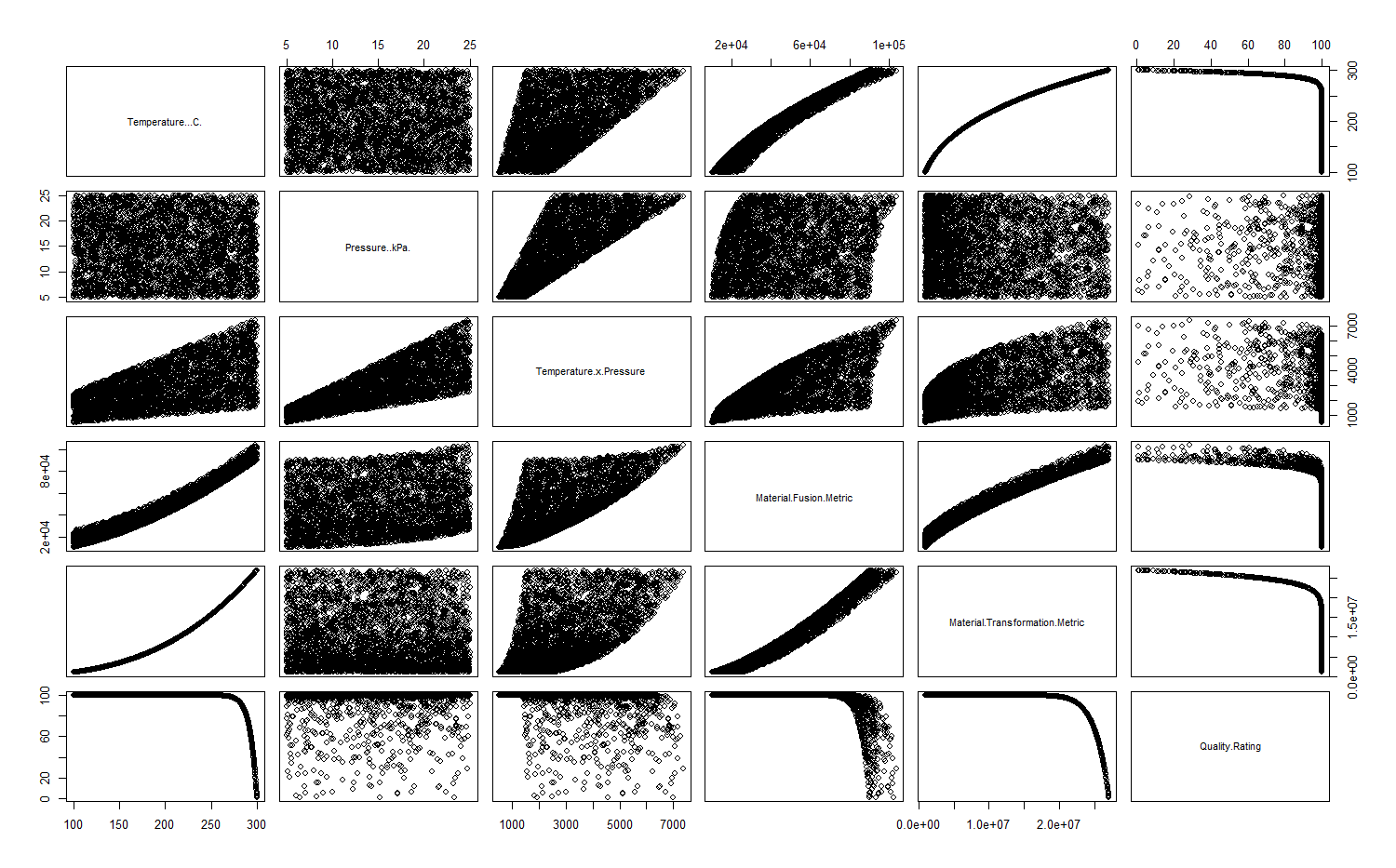
Anurag Singha | 22IM30007 | 20/01/2025

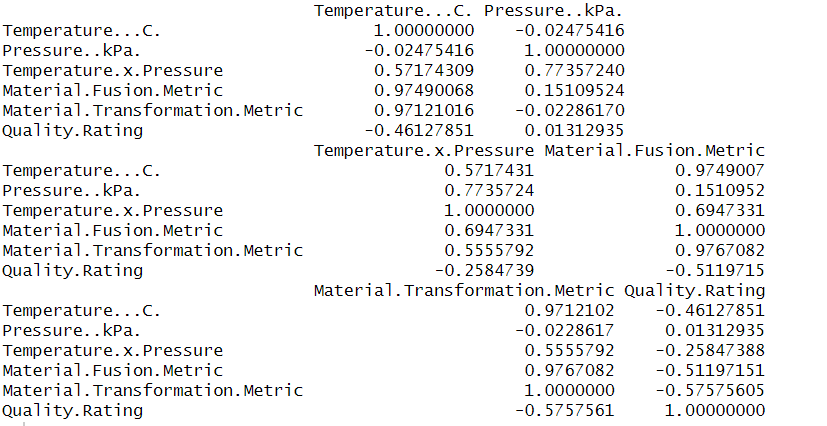


1.Load the dataset “manufacturing.csv”. Display first few rows of the dataset**.**

Ans: 

2.Perform matrix plot and correlation analysis indicate if there is any correlation among the predictors.

Ans: 



Based on the correlation analysis, here are the key findings regarding correlations among predictors:

1)High Correlations (potential multicollinearity concerns):

Temperature and Material Fusion Metric (0.97)

Temperature and Material Transformation Metric (0.97)

Material Fusion Metric and Material Transformation Metric (0.98)

2)Moderate Correlations:

Pressure and Temperature × Pressure interaction (0.77)

Temperature × Pressure and Material Fusion Metric (0.69)

Temperature × Pressure and Material Transformation Metric (0.56)

3)Low Correlations:

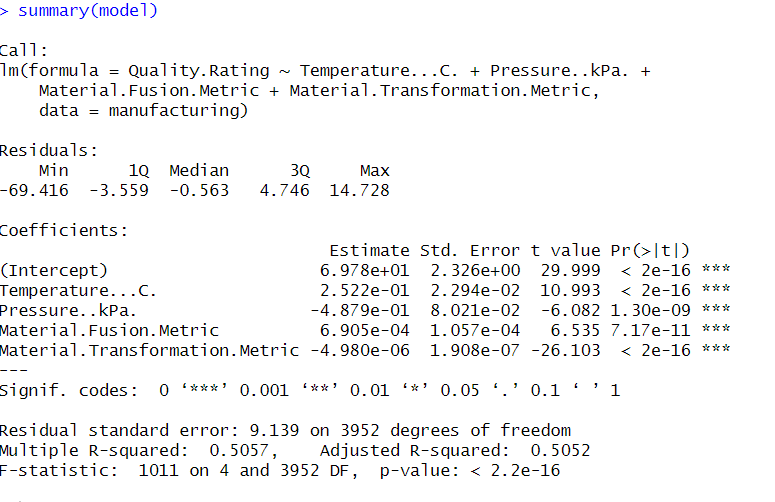
Temperature and Pressure (-0.02)

Pressure and Material Fusion Metric (0.15)

Pressure and Material Transformation Metric (-0.02)

3. Fit a Linear Regression model without the interaction term. From the linear regression summary which factors seem to be significant?

Ans:



All factors in the model are highly significant (p < 0.001) as indicated by the '\*\*\*' significance codes. Specifically:

1)Temperature (Temperature...C.):

t-value: 10.993

p-value: < 2e-16

Highly significant positive effect (coefficient = 0.2522)

2)Pressure (Pressure..kPa.):

t-value: -6.082

p-value: 1.30e-09

Highly significant negative effect (coefficient = -0.4879)

3)Material Fusion Metric:

t-value: 6.535

p-value: 7.17e-11

Highly significant positive effect (coefficient = 0.0006905)

3)Material Transformation Metric:

t-value: -26.103

p-value: < 2e-16

Highly significant negative effect (coefficient = -0.000004980)

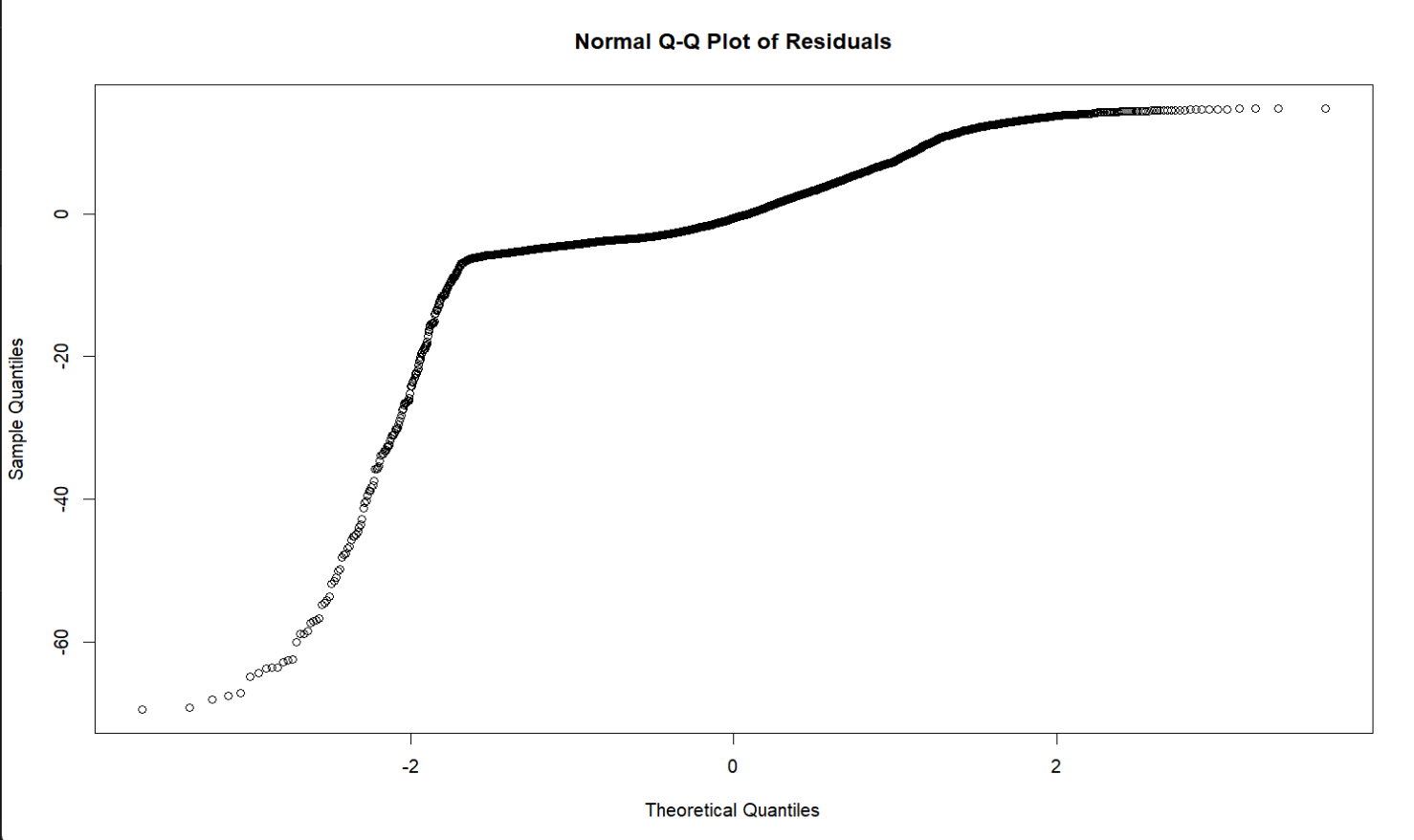
Therefore, all factors included in the model are statistically significant predictors of Quality Rating, suggesting they all contribute meaningfully to explaining variations in the quality outcome.

4.What is your interpretation R-sq and R-sq adjusted?

Ans: An R-squared value of 0.5057 suggests that approximately 50.57% of the variation in Quality Rating is accounted for by the current predictors, leaving 49.43% unexplained. The Adjusted R-squared value of 0.5052, being very close to the R-squared, indicates minimal risk of overfitting and suggests that all predictors contribute meaningfully to the model. While the moderate R-squared values demonstrate that the model has a reasonable level of predictive power, there may still be other significant factors not included in the analysis.

5.Perform normal probability plot of residuals and comment on model adequacy.

Ans:



1. Non-normality issues:

Clear deviation from the diagonal line at both tails

Particularly strong deviation at the lower tail (below -2 theoretical quantiles)

Some deviation at the upper tail (above 2 theoretical quantiles)

1. Pattern analysis:

The central portion (between -2 and 2 theoretical quantiles) follows the line reasonably well

The S-shaped pattern indicates heavy tails in the distribution

The residuals are skewed, especially at the lower end

3)Model adequacy conclusions:

The model violates the normality assumption of linear regression

The systematic deviations suggest the model may not be capturing some important non-linear relationships

Transformations of variables or consideration of non-linear terms might be needed to improve the model fit

6.Randomly sample 20% of the data and keep it as test data. Use rest of the 80% data to train the linear model. What is the RMSE on test data?

Ans:

