Statistical Learning - Assignment 1

Gabriele Durante
gabridur@uio.no

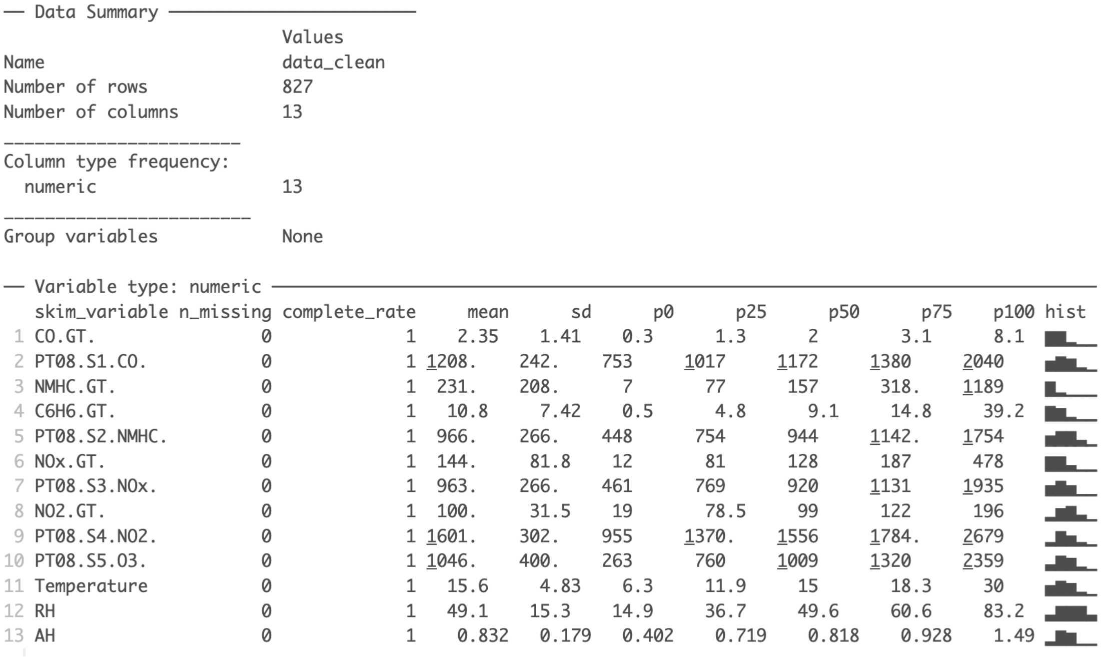
September 18, 2024

1 Problem 1

1.1 Summary Statistics Table

This dataset includes 9,358 hourly averaged readings from five metal oxide chemical sensors located in an air quality monitoring device in a heavily polluted urban area of an Italian city. The data covers the period from March 2004 to February 2005 and features concentrations of CO, non-methanic hydrocarbons, benzene, total nitrogen oxides (NOx), and nitrogen dioxide (NO2), as measured by a co-located certified analyzer. The dataset also shows signs of sensor cross-sensitivity and drift, which may affect the accuracy of concentration estimates. Researchers can use this dataset to enhance predictive models for air pollution levels and to deepen the understanding of air quality in polluted urban environments. For the analysis, date and time variables were removed as it was not an objective of this investigation. After an appropriate pre-processing (removal of NA's value and transformation chr into int variables, in the dataset NA's = 200) of the data, we obtain the following summary statistics table.

For more details and to explore the dataset, you can visit the official webpage at the following link: Air Quality Dataset or the paper: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario.



```
── Data Summary ────────────────────────
                           Values
Name                       data_clean
Number of rows             827
Number of columns          13
_______________________
Column type frequency:
  numeric                  13
________________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────────────────────────
   skim_variable n_missing complete_rate     mean      sd      p0      p25      p50      p75     p100 hist
 1 CO.GT.                0             1    2.35    1.41    0.3      1.3      2        3.1      8.1
 2 PT08.S1.CO.           0             1 1208.    242.    753     1017     1172     1380     2040
 3 NMHC.GT.              0             1  231.    208.      7       77      157      318.    1189
 4 C6H6.GT.              0             1   10.8     7.42    0.5      4.8      9.1     14.8     39.2
 5 PT08.S2.NMHC.         0             1  966.    266.    448      754      944     1142.    1754
 6 NOx.GT.               0             1  144.     81.8    12       81      128      187      478
 7 PT08.S3.NOx.          0             1  963.    266.    461      769      920     1131     1935
 8 NO2.GT.               0             1  100.     31.5    19       78.5     99      122      196
 9 PT08.S4.NO2.          0             1 1601.    302.    955     1370.    1556     1784.    2679
10 PT08.S5.O3.           0             1 1046.    400.    263      760     1009     1320     2359
11 Temperature           0             1   15.6     4.83    6.3     11.9     15       18.3     30
12 RH                    0             1   49.1    15.3    14.9     36.7     49.6     60.6     83.2
13 AH                    0             1    0.832   0.179   0.402    0.719    0.818    0.928    1.49
```

Figure 1: Statistical table (skim)

Something intersting in the target variable, the carbon monoxide levels (CO.GT), is that have a wide range, from as low as 0.3 to a high of 8.1, with a mean of 2.35, suggesting significant variability. This may reveal relationships between pollutant concentrations and environmental factors like temperature and humidity.

2 Problem 2

2.1 Bad Data Visualization

We will now look at examples of bad plots and why they are considered such. In Figure 2 and Figure 3 we can observe barplots that are heavily influenced by the presence of NA's. In this case, the NA's in the dataset are labelled with the value -200. If they are not purposely removed, they alter the graph and do not allow a clear visualisation of the data. In Figure 4 and Figure 5, on the other hand, we can observe two pie plots representing two numerical variables. This way of representing the data is not optimal for numerical data as the visualisation is not at all clear. Figure 6 represents the boxplot of the variable PT08.S3(NOx). This type of graph can be very useful for identifying outliers and observing the distribution of data. In this situation, the reference scale is not shown, making the data uninterpretable. In addition, the outliers were removed but this was not indicated in any way. Figure 7 and Figure 8 are scatter plots where the dots are too large and there are null values that do not allow the data to be displayed correctly. in Figure 9, the axes are manually scaled, making the representation misleading. Most of the information is not displayed and, if decontextualised, could be used to distort the reality of the data. Finally, Figure 10 presents a histogram that uses few bins, not allowing the distribution of data to be correctly visualised.

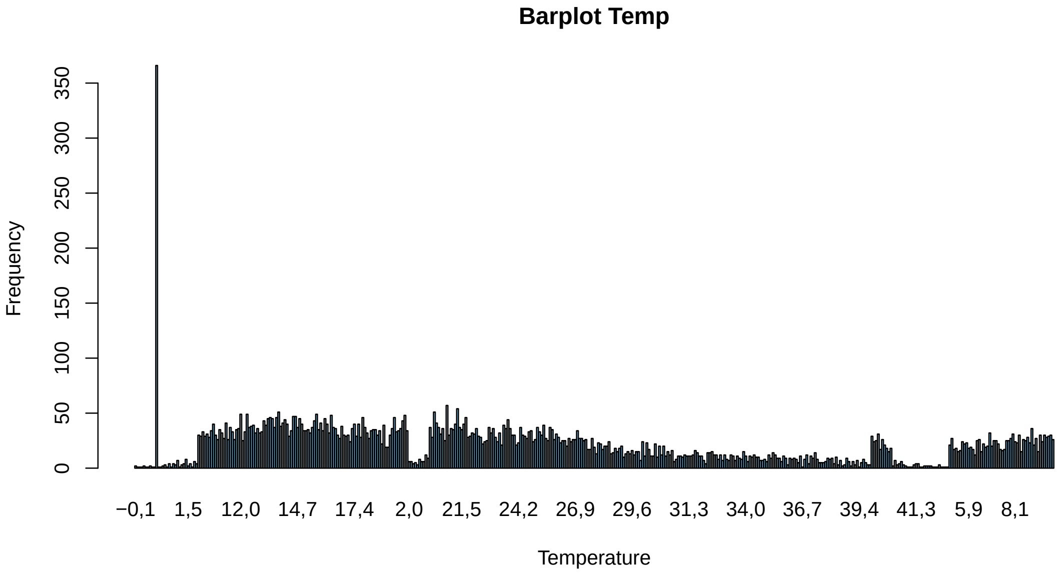


Figure 2: Barplot of Temperature (chr) variable, with NA's

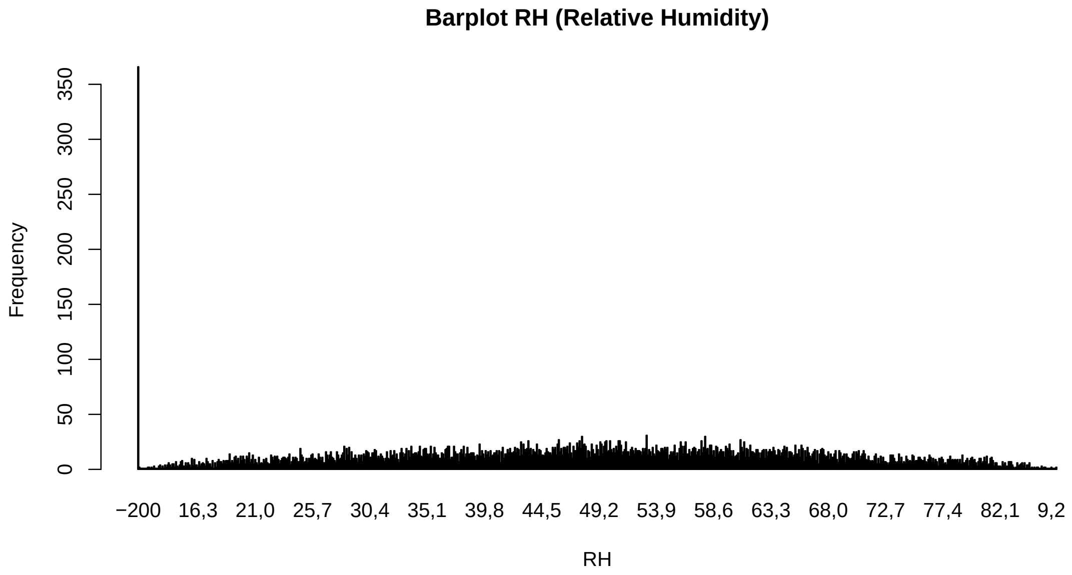


Figure 3: Barplot of Relative Humidity (chr) variable, with NA's

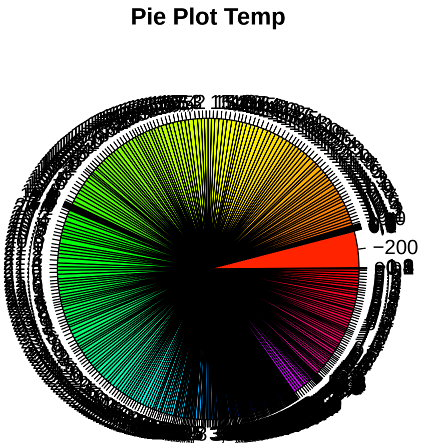


Figure 4: Pie Plot of Temperature variable, with NA's

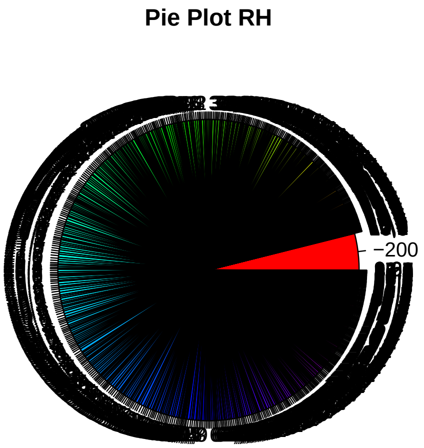


Figure 5: Pie Plot of Relative Humidity variable, with NA's

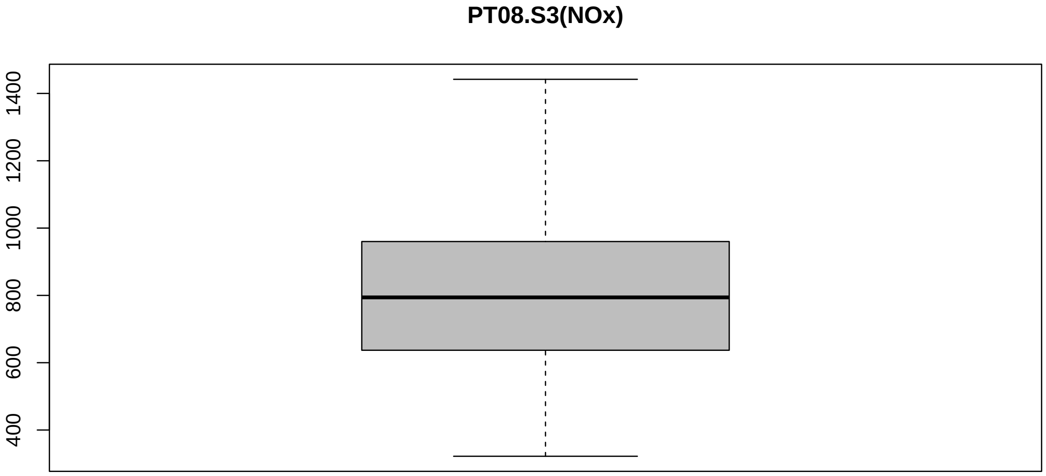


Figure 6: Boxplot PT08.S3(NOx), but we remove outliers only in the plot

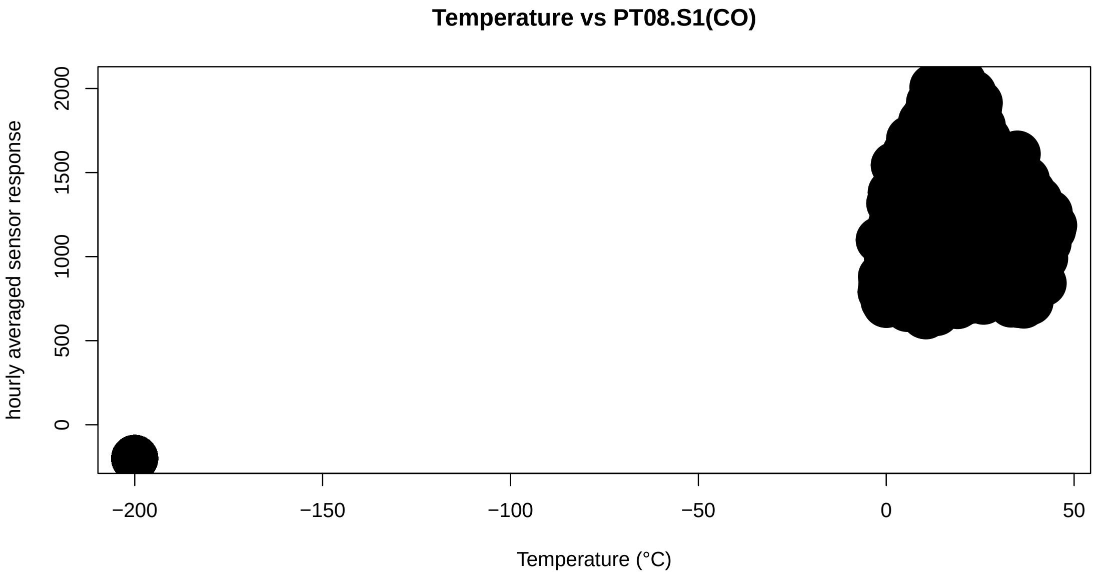


Figure 7: Dumb scatterplot Temperature vs PT08.S1(CO), with NA's

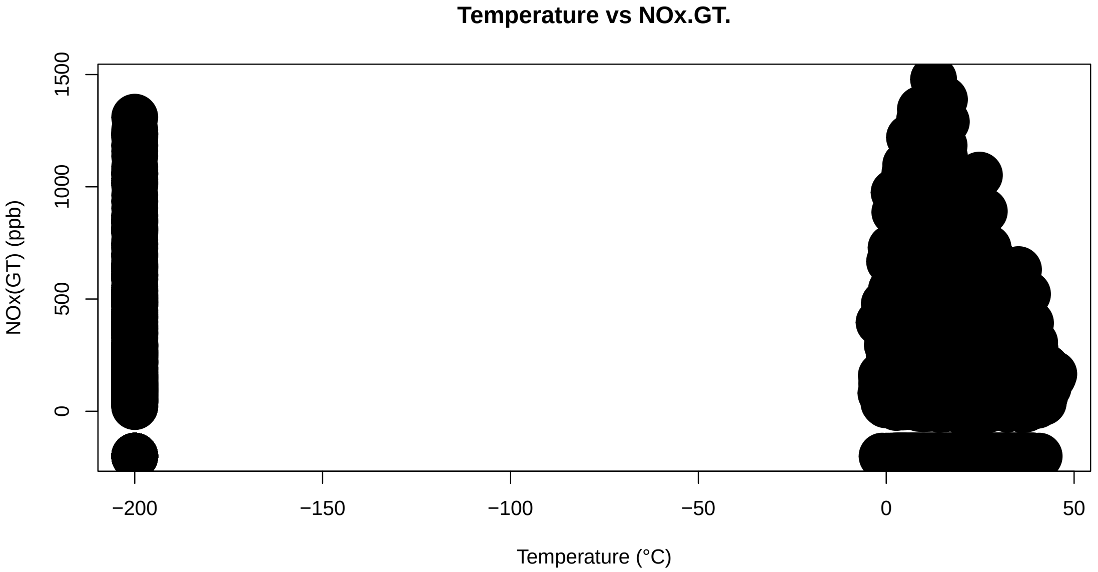


Figure 8: Dumb scatterplot Temperature vs NOx.GT., with NA's

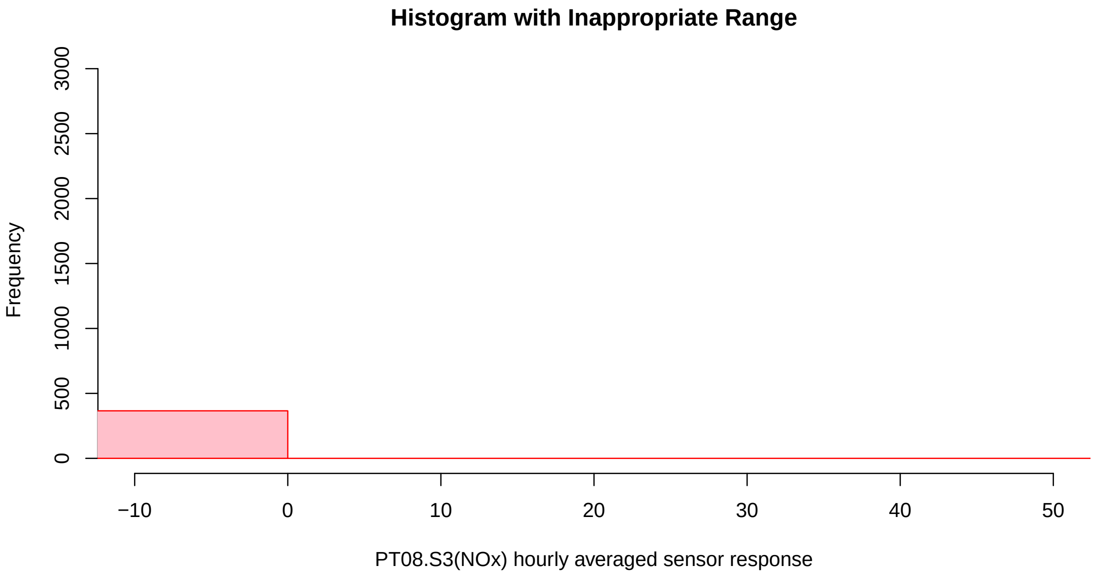


Figure 9: Histogram with (xlim) arbitrary range that may not fit the data

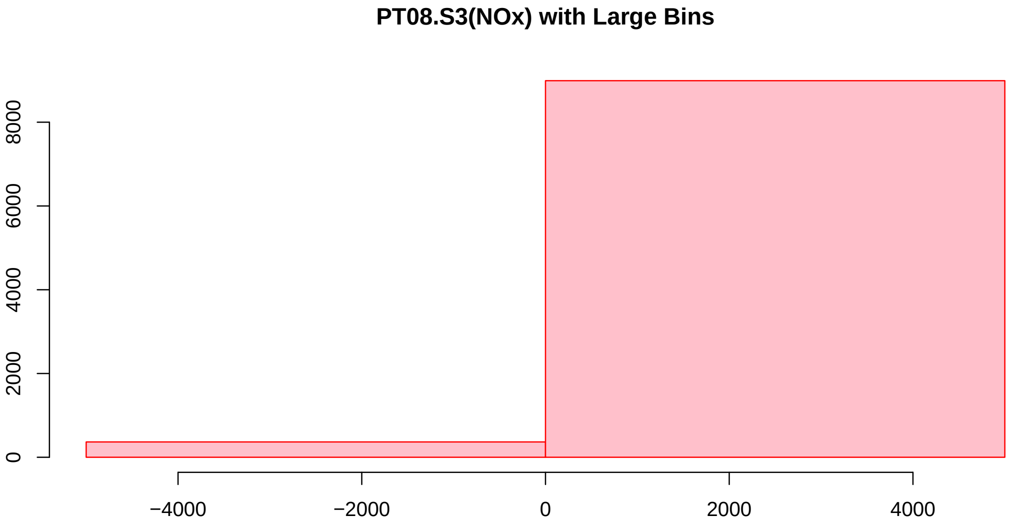


Figure 10: Histogram with large bins

3 Problem 3

3.1 Good Data Visualization

The first substantial change to the graphs was made to the pie charts (e.g. Figure 11), where the categorical variables were grouped in percentage ranges to allow an immediate and effective visualisation of the distribution of character values. Following the cleaning of the null data, the scatter plots were optimised to be more intuitive for investigating possible relationships between variables, as can be seen in Figure 12 and Figure 13. By using clean data and optimising the scale of the parameters, the histograms were also optimised, allowing the distribution of individual features to be observed in full (Figure 14 and Figure 15). Finally, the boxplot was also improved, leaving the outliers in the original visualisation and rotating the scale (Figure 16).

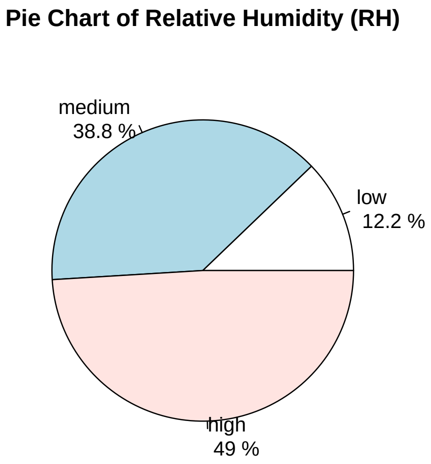


Figure 11: Pie Plot with Range in chr variables

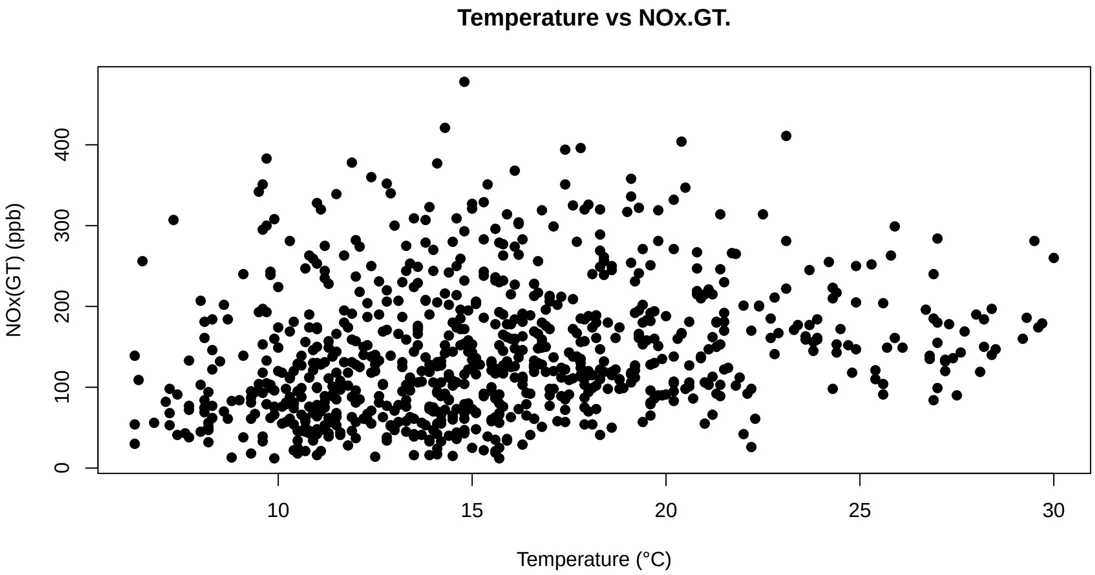


Figure 12: Scatterplot Temperature vs NOx(GT)

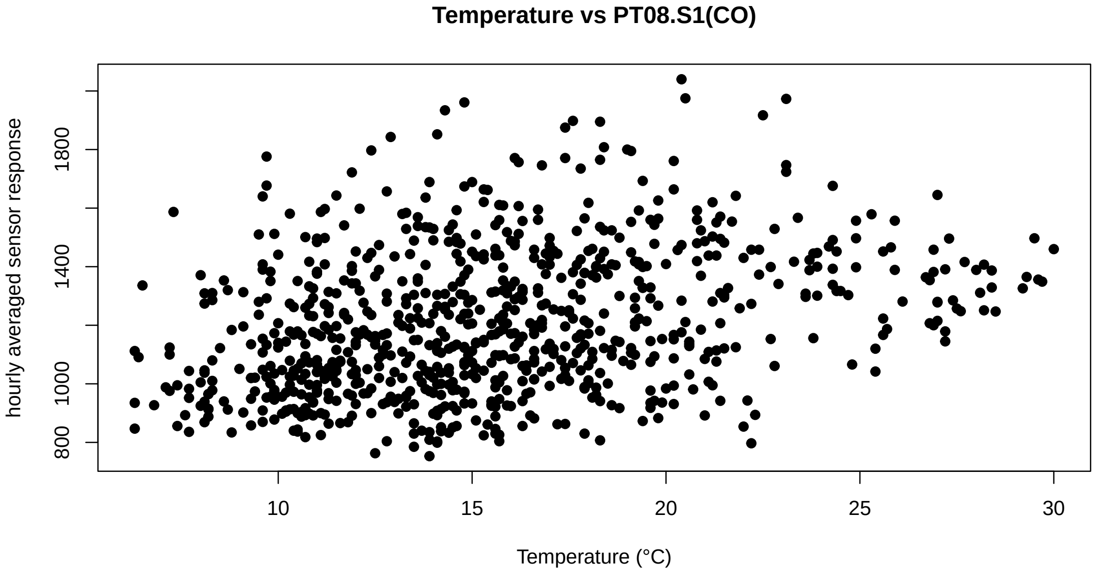


Figure 13: Scatterplot Temperature vs PT08.S1(CO)

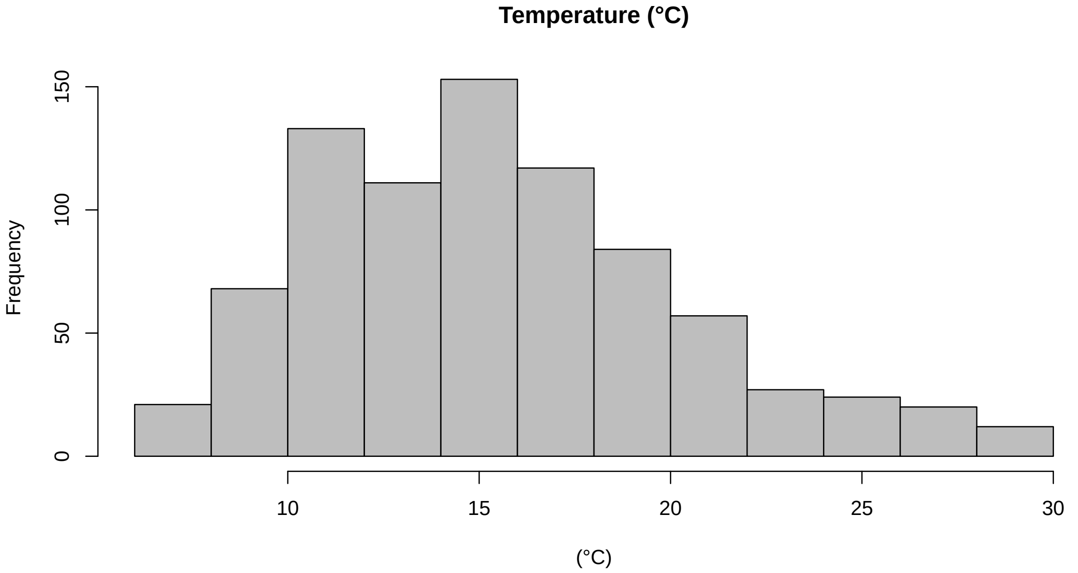


Figure 14: Temperature distrib. Histogram

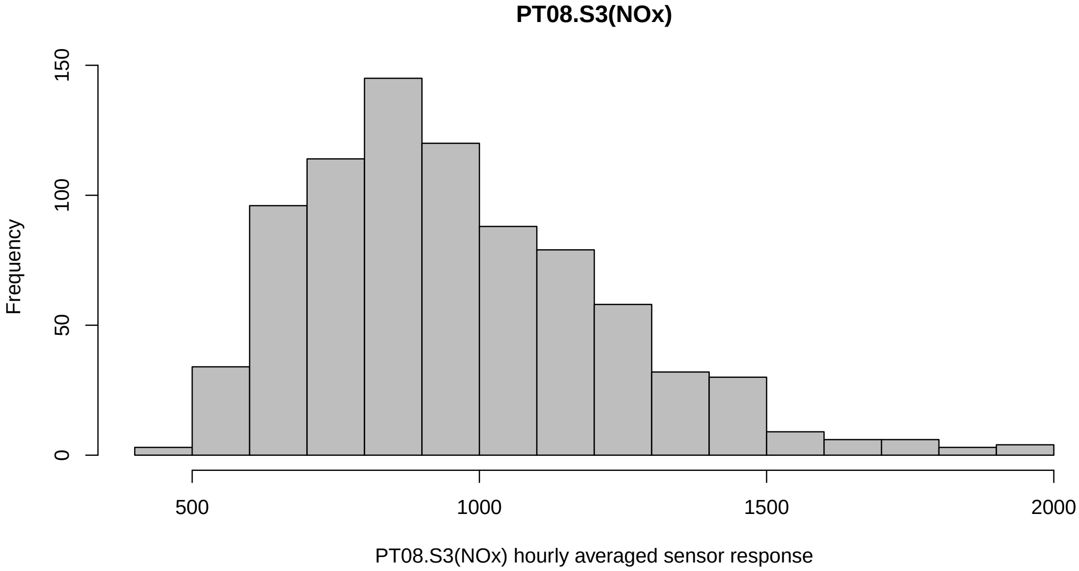


Figure 15: PT08.S3(NOx) distrib. Histogram

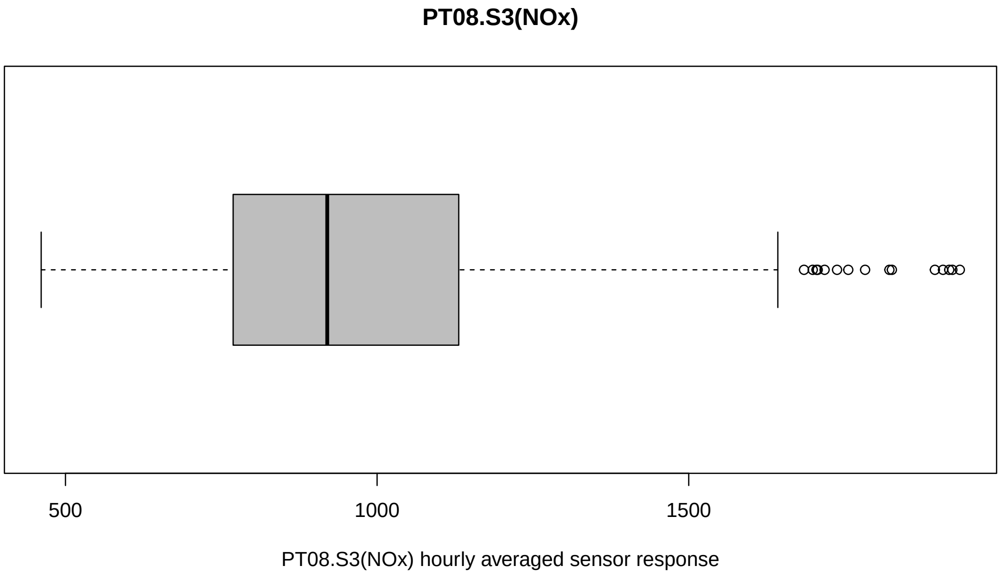
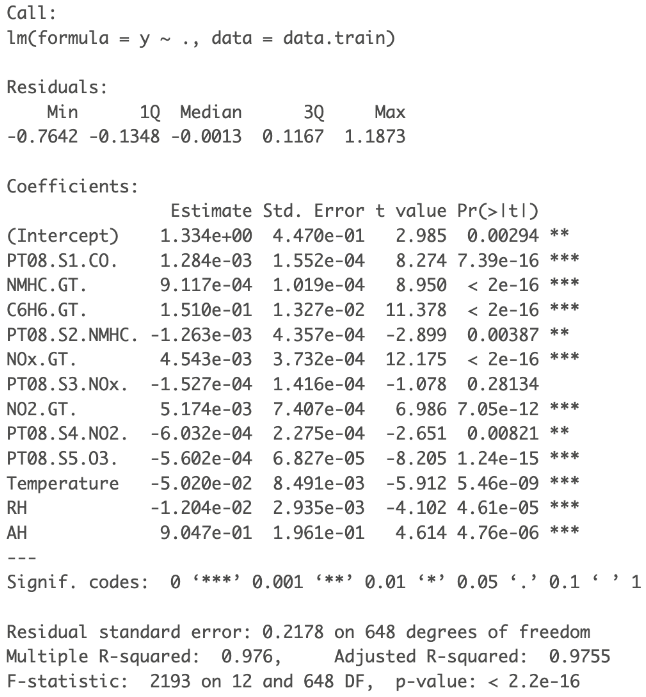


Figure 16: PT08.S3(NOx) boxplot

4 Problem 4

4.1 Simple analysis

The linear regression model was fitted and the performance was evaluated, yielding a RMSE of 0.25294, indicating a moderate level of predictive accuracy on the dataset.



```
Call:
lm(formula = y ~ ., data = data.train)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7642 -0.1348 -0.0013  0.1167  1.1873 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.334e+00  4.470e-01   2.985  0.00294 ** 
PT08.S1.CO.    1.284e-03  1.552e-04   8.274 7.39e-16 ***
NMHC.GT.       9.117e-04  1.019e-04   8.950  < 2e-16 ***
C6H6.GT.       1.510e-01  1.327e-02  11.378  < 2e-16 ***
PT08.S2.NMHC. -1.263e-03  4.357e-04  -2.899  0.00387 ** 
NOx.GT.        4.543e-03  3.732e-04  12.175  < 2e-16 ***
PT08.S3.NOx.  -1.527e-04  1.416e-04  -1.078  0.28134    
NO2.GT.        5.174e-03  7.407e-04   6.986 7.05e-12 ***
PT08.S4.NO2.  -6.032e-04  2.275e-04  -2.651  0.00821 ** 
PT08.S5.O3.   -5.602e-04  6.827e-05  -8.205 1.24e-15 ***
Temperature   -5.020e-02  8.491e-03  -5.912 5.46e-09 ***
RH            -1.204e-02  2.935e-03  -4.102 4.61e-05 ***
AH             9.047e-01  1.961e-01   4.614 4.76e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2178 on 648 degrees of freedom
Multiple R-squared:  0.976,	Adjusted R-squared:  0.9755 
F-statistic:  2193 on 12 and 648 DF,  p-value: < 2.2e-16
```

Figure 17: Model summary of the Regression

As we can se from Figure 17 Most predictors are statistically significant at the 0.05 level, with p-values less than 0.05. These include PT08.S1.CO., NMHC.GT., C6H6.GT., PT08.S2.NMHC., NOx.GT., NO2.GT., PT08.S4.NO2., PT08.S5.O3., Temperature, RH (Relative Humidity), and AH (Absolute Humidity). Notably, many of these variables have extremely small p-values (e.g., ¡ 2e-16), indicating a strong relationship with the target variable. The variable PT08.S3.NOx. has a relatively high p-value (0.28134), indicating that it is not a significant predictor in the model. The model's R-squared value is 0.976. The adjusted R-squared, which accounts for the number of predictors, is 0.9755, indicating that the model is not overfitting despite the number of predictors. The residual standard error is 0.2178, this error seems relatively low, suggesting that the model has a good fit to the training data. The RMSE value of 0.25294 further indicates the model's predictive accuracy.

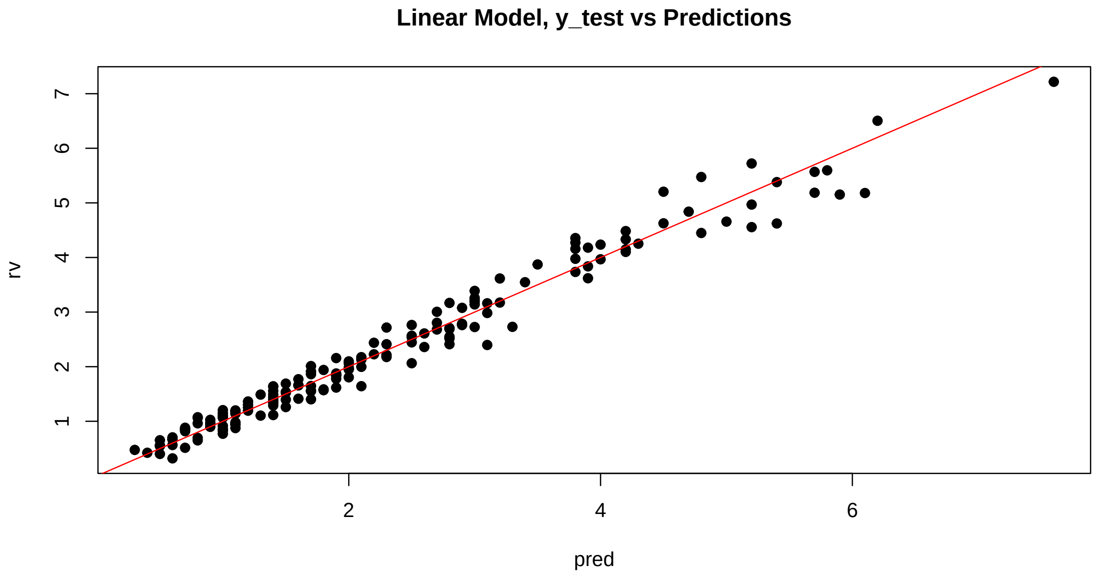
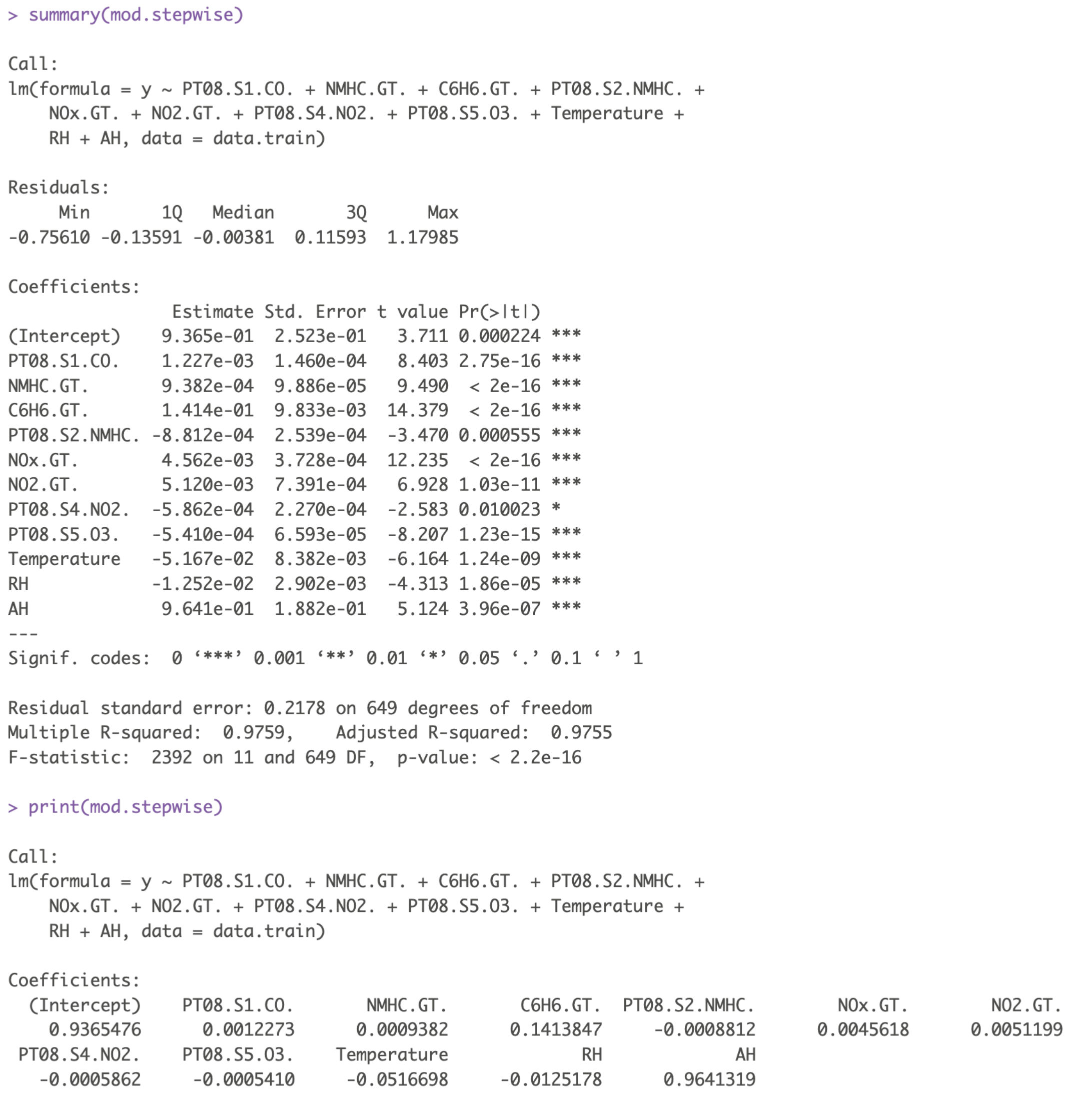


Figure 18: Regression predictions

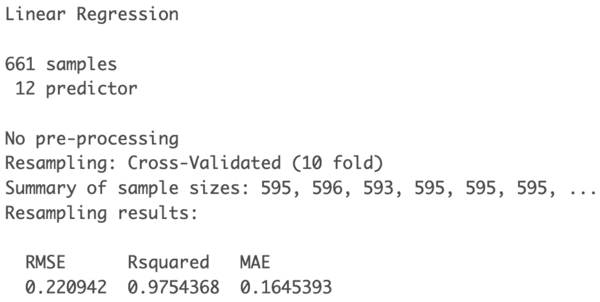
Figure 18 shows the graphical representation of the predicted values and the test values. In addition to simple linear regression, we now perform model selection to look at the covariates.



```
> summary(mod.stepwise)

Call:
lm(formula = y ~ PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. + 
    NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + Temperature + 
    RH + AH, data = data.train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.75610 -0.13591 -0.00381  0.11593  1.17985 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    9.365e-01  2.523e-01   3.711 0.000224 ***
PT08.S1.CO.    1.227e-03  1.460e-04   8.403 2.75e-16 ***
NMHC.GT.       9.382e-04  9.886e-05   9.490  < 2e-16 ***
C6H6.GT.       1.414e-01  9.833e-03  14.379  < 2e-16 ***
PT08.S2.NMHC. -8.812e-04  2.539e-04  -3.470 0.000555 ***
NOx.GT.        4.562e-03  3.728e-04  12.235  < 2e-16 ***
NO2.GT.        5.120e-03  7.391e-04   6.928 1.03e-11 ***
PT08.S4.NO2.  -5.862e-04  2.270e-04  -2.583 0.010023 *  
PT08.S5.O3.   -5.410e-04  6.593e-05  -8.207 1.23e-15 ***
Temperature   -5.167e-02  8.382e-03  -6.164 1.24e-09 ***
RH            -1.252e-02  2.902e-03  -4.313 1.86e-05 ***
AH             9.641e-01  1.882e-01   5.124 3.96e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2178 on 649 degrees of freedom
Multiple R-squared:  0.9759,	Adjusted R-squared:  0.9755 
F-statistic:  2392 on 11 and 649 DF,  p-value: < 2.2e-16

> print(mod.stepwise)

Call:
lm(formula = y ~ PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. + 
    NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + Temperature + 
    RH + AH, data = data.train)

Coefficients:
  (Intercept)    PT08.S1.CO.       NMHC.GT.       C6H6.GT.  PT08.S2.NMHC.        NOx.GT.        NO2.GT.  
    0.9365476      0.0012273      0.0009382      0.1413847     -0.0008812      0.0045618      0.0051199  
 PT08.S4.NO2.    PT08.S5.O3.    Temperature             RH             AH  
   -0.0005862     -0.0005410     -0.0516698     -0.0125178      0.9641319  
```

Figure 19: Model selection: Stepwise

Using the stepwise method (Figure 19) to find a model that balances the complexity and explanatory power of the data, we obtain a model similar to that of simple regression with the removal of a variable (PT08.S3(NOx)) that had already been identified as not statistically significant.



```
Linear Regression

661 samples
 12 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 595, 596, 593, 595, 595, 595, ...
Resampling results:

  RMSE      Rsquared   MAE
  0.220942  0.9754368  0.1645393
```

Figure 20: Model selection: Cross Validation

Through the cross-validation performed in Figure 20 the model appears to have a very good predictive ability on the basis of cross-validation, but it is important to bear in mind the risk of overfitting and to evaluate the model further on a separate test dataset.

4.2 Conclusions

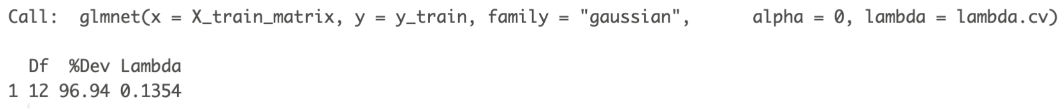
The linear regression model appears to be adequate for this dataset. This conclusion is supported by the high values of $\mathbf{R^2}$ and adjusted $\mathbf{R^2}$, indicating that the model explains a significant portion of the variance in the response variable. Additionally, the statistical significance of the predictors and the low residual standard error suggest that the model is both appropriate and sufficiently accurate in capturing the relationships within the data. However, the high $\mathbf{R^2}$ value raises a potential concern for overfitting. Nevertheless, cross-validation results demonstrate good performance on unseen data, suggesting that the risk of overfitting is appropriately managed. We can say that the linear regression model seems suitable for this dataset, owing to its strong explanatory power and the statistical significance of the included variables. If we look at the **RMSE** of our linear model, in general it's very small. This means that, on average, the model predictions are close to the actual observed values, indicating good predictive accuracy in an absolute sense.

5 Additional models

After training a linear regression model, it is advisable to evaluate Lasso and Ridge regression as well, primarily due to their regularization properties. Linear regression can suffer from issues like overfitting and multicollinearity. Lasso and Ridge address these issues by introducing a penalty to the regression objective. Both methods help in controlling the bias-variance trade-off by tuning a regularization parameter, leading to models with enhanced generalization capability. Those models can yield more robust and parsimonious models, especially in high-dimensional settings.

5.1 Ridge

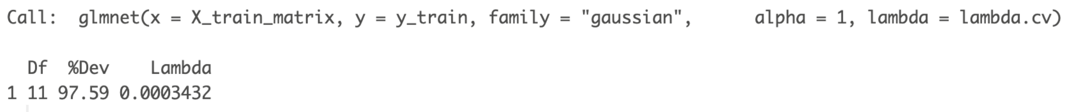
The Ridge regression was performed using the glmnet function. The lambda value here is relatively low that implies that the regularization penalty (which shrinks the coefficients towards zero) is mild. This basically means that the model does not need much regularization to fit the data well. the predictors might not be suffering significantly from issues like multicollinearity or overfitting.



```
Call:  glmnet(x = X_train_matrix, y = y_train, family = "gaussian",      alpha = 0, lambda = lambda.cv) 

  Df  %Dev Lambda
1 12 96.94 0.1354
```

5.2 Lasso

The lambda value in the Lasso model is even lower. This tiny value suggests that the model requires minimal regularization to achieve optimal performance. The predictors included by Lasso might already have a strong relationship with the response variable, and the regularization serves only to fine-tune their influence slightly



```
Call:  glmnet(x = X_train_matrix, y = y_train, family = "gaussian",      alpha = 1, lambda = lambda.cv) 

  Df  %Dev    Lambda
1 11 97.59 0.0003432
```

5.3 Lambda approaches zero

5.3.1 1. Ridge Regression

Objective function Ridge regression:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

When $\lambda \to 0$:

$$\lim_{\lambda \to 0} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

This is the objective function of the OLS regression, which minimizes the sum of squared residuals:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

Therefore, when $\lambda \to 0$, Ridge regression reduces to the OLS regression.

5.3.2 Lasso Regression

Objective function Lasso regression:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \|\beta_j\| \right\}.$$

When $\lambda \to 0$:

$$\lim_{\lambda \to 0} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

Again, this is the objective function of the OLS regression:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

Therefore, when $\lambda \to 0$, Lasso regression also reduces to the OLS regression.

5.4 Conclusions

The fact that both models have low lambda values implies that the dataset's linear relationships are already relatively well-behaved. For an ordinary linear regression model (without any regularization), this suggests that the predictors might not be overly collinear, and the model does not suffer much from overfitting. Essentially, the underlying linear model fits the data quite well even without regularization.