PM2.5 dataset

Step 1

Producing a plot and a correlation matrix of the data and analyzing the relationships between the response and predictors and relationships between the predictors themselves.

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
install.packages("corrplot")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(gridExtra)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::combine() masks gridExtra::combine()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data <- read.csv("pm25.csv")
```

Creating a plot between the response and the predictors

```
p1 <- ggplot(data= data) +
  geom_point(mapping=aes(temperature, pm25)) +
  labs(title = "Temperature")

# Create the second subplot
p2 <- ggplot(data= data) +
  geom_point(mapping=aes(pm25, precipitation)) +
  labs(title = "precipitation")


# Create the third subplot
p3 <- ggplot(data= data) +
```

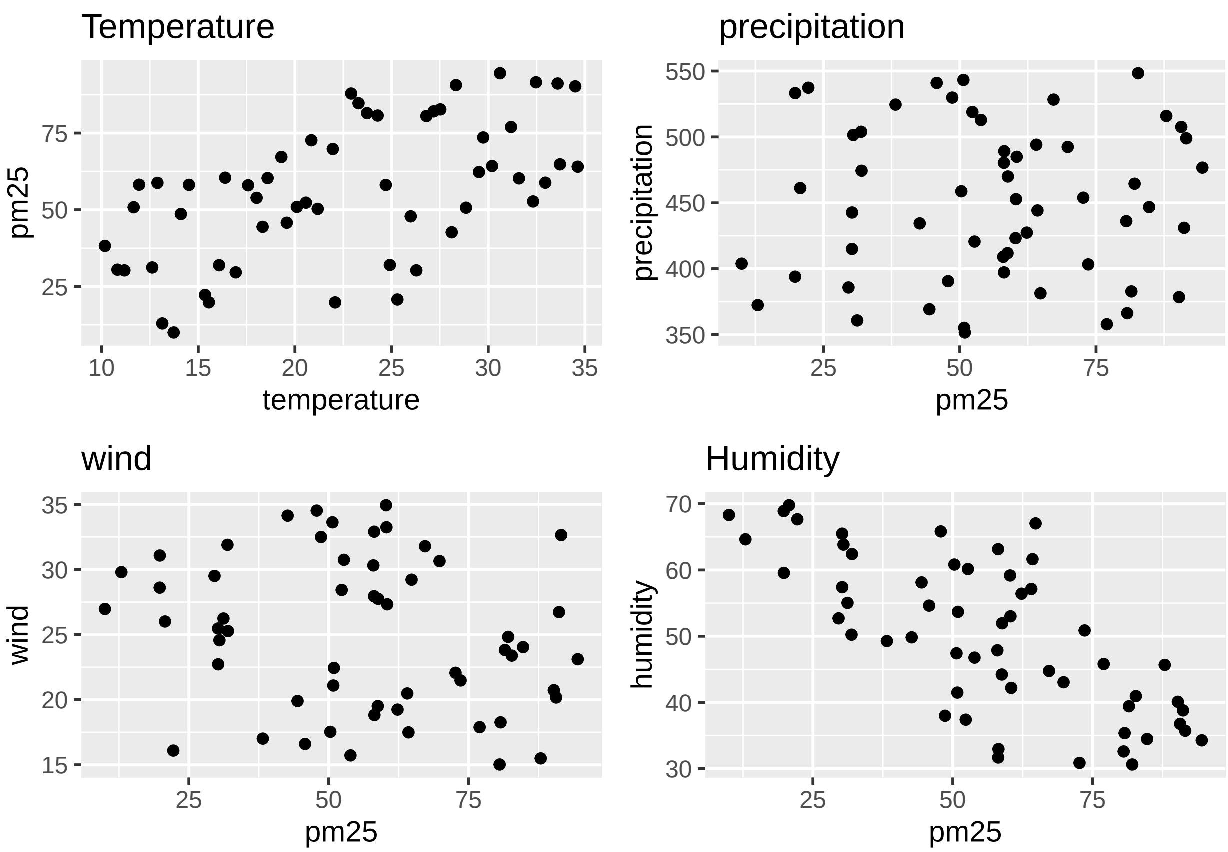
```
  geom_point(mapping=aes(pm25, wind)) +
  labs(title = "wind")



# Create the fourth subplot
p4 <- ggplot(data= data) +
  geom_point(mapping=aes(pm25, humidity)) +
  labs(title = "Humidity")

# Arrange the subplots

grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



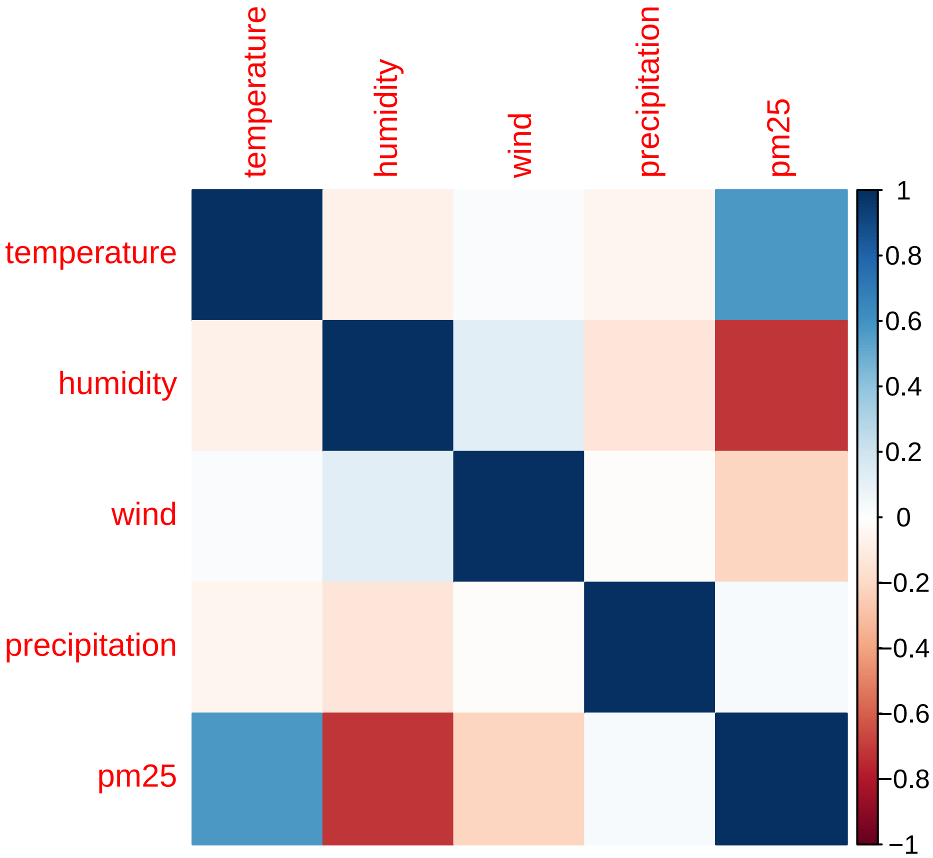
Correlation

```
cor_matrix <- cor(data[, c("temperature", "humidity", "wind", "precipitation", "pm25")])
print(cor_matrix)
```

```
##               temperature    humidity        wind precipitation        pm25
## temperature    1.00000000 -0.07264891  0.02861166   -0.05050014  0.57191961
## humidity      -0.07264891  1.00000000  0.12406351   -0.13550607 -0.71965591
## wind           0.02861166  0.12406351  1.00000000   -0.01525977 -0.21866823
## precipitation -0.05050014 -0.13550607 -0.01525977    1.00000000  0.03759033
## pm25           0.57191961 -0.71965591 -0.21866823    0.03759033  1.00000000
```

Plotting the correlation matrix

```
corrplot(cor_matrix, method = "color")
```



As it can be seen in the correlation matrix that Temperature has the best positive correlation with PM 2.5 which means that when temperature increases the pm 2.5 concentration also increases. If we look at the Humidity correlation with PM 2.5 then we can see that it has a negative correlation which means when humidity increase PM 2.5 concentration tend to decrease and the other two pedictors that is precipitation and wind have close to zero correlation.

Step 2

Performing Multiple Linear Regression Analysis on the dataset

```
model <- lm(pm25 ~ temperature + humidity + wind + precipitation, data = data)
```

Now we will estimate the impact of humidity on PM 2.5 Concentration and will produce a 95 percent confidence interval. For that we need to get the coeficient estimates which basically provides the estimate of the relationship or in other words it will predict the change in the response variable(PM 2.5) which is associated with one unit increase in the predictor variable(humidity). Alongside the coefficient estimate we also need confidence interval which gives us the range in which true population parameter lies.

```
summary(model)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
##     data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.759  -6.804  -1.649   6.857  20.975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   102.72259   14.71953   6.979 5.88e-09 ***
## temperature     1.62142    0.18762   8.642 1.46e-11 ***
## humidity       -1.27742    0.11854 -10.776 9.49e-15 ***
## wind           -0.58016    0.23405  -2.479   0.0165 *
## precipitation  -0.01091    0.02350  -0.464   0.6444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 51 degrees of freedom
## Multiple R-squared:  0.8127, Adjusted R-squared:  0.7981
## F-statistic: 55.34 on 4 and 51 DF,  p-value: < 2.2e-16
```

```r
coefs <- coef(summary(model))
humidity_coefs <- coefs["humidity", ]
humidity_coefs[c("Estimate", "Std. Error", "t value", "Pr(>|t|)", "2.5 %", "97.5 %")]
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)          <NA>
## -1.277423e+00  1.185437e-01 -1.077596e+01  9.490343e-15            NA
##          <NA>
##            NA
```

Where

Estimate: Estimated coefficient of the humidity predictor.

Std.Error: It quantifies coefficient estimate standard deviation.

t value: It tells us the strength of the relationship between predictor and the response variable.

p_value: lower p value suggest significant relationship and higher p value suggest no significant relationship between predictors and response.

2.5% and 97.5% is the lower and upper bounds of the 95% confidence Interval of the estimate.

As it can be seen from the above output that for 1 unit increase in the humidity will result in 1.277423 units decrease in the PM 2.5 concentration.

Step 3

Conducting an F Test for the overall regression.

Mathematical multiple regression model

$$Y = \beta_0 + \beta_1 \cdot \text{Temperature} + \beta_2 \cdot \text{Humidity} + \beta_3 \cdot \text{Wind} + \beta_4 \cdot \text{Precipitation} + \varepsilon$$

$$Y = \text{PM2.5}$$

$$\beta_0 = \text{Intercept}$$

$$\beta_1 = \text{Effect of Temperature}$$

$$\beta_2 = \text{Effect of Humidity}$$

$$\beta_3 = \text{Effect of Wind}$$

$$\beta_4 = \text{Effect of Precipitation}$$

$$\varepsilon = Error$$

Hypothesis for ANOVA test of multiple regression

Null Hypothesis(H0): There is no relationship between the predictors(Temperature, Humidity, Wind, Precipitation) and response(PM2.5 concentration)

Alternative Hypothesis(H1): There is relationship between predictors and the response

Compute Anova Table

Anova specifically compares the amount of variation between groups with the amount of variation within the group.

```
anova_table <- anova(model)

SSReg <- anova_table$"Sum Sq"[1]  # Sum of Squares for Regression
MSReg <- anova_table$"Mean Sq"[1]  # Mean Square for Regression
SSRes <- anova_table$"Sum Sq"[2]  # Sum of Squares for Residuals
n <- nrow(data)  # Sample size
k <- length(coefficients(model)) - 1  # Number of predictors (excluding the intercept)

# Calculate the Mean Square for Residuals
MSRes <- SSRes / (n - k - 1)

# Compute the Total Sum of Squares
SST <- SSReg + SSRes

# Create the ANOVA table
anova_table <- data.frame(Source = c("Regression", "Residual", "Total"),
                          `Sum of Squares` = c(SSReg, SSRes, SST),
                          `Degrees of Freedom` = c(k, n - k - 1, n - 1),
                          `Mean Square` = c(MSReg, MSRes, NA),
                          `F-statistic` = c(MSReg / MSRes, NA, NA))

# Print the ANOVA table
print(anova_table)
```

```
##       Source Sum.of.Squares Degrees.of.Freedom Mean.Square F.statistic
## 1 Regression       9014.394                  4   9014.3941     36.0866
## 2   Residual      12739.744                 51    249.7989          NA
## 3      Total      21754.138                 55          NA          NA
```

F Statistics

As it can be seen in the Anova table that the value of F statistic is 36.0866 which was calculated by the formula Mean Square for Regression divided by Mean Square for Residuals. Below is the code

```
F_statistic <- MSReg / MSRes
print(F_statistic)
```

```
## [1] 36.0866
```

Compute the p-value

```
p_value <- 1 - pf(F_statistic, df1 = k, df2 = n - k - 1)
print(p_value)
```

```
## [1] 2.664535e-14
```

where k is the number of predictors and n is the sample size

Conclusion

```
alpha <- 0.05

# Compare the p-value with the significance level
if (p_value < alpha) {
  conclusion <- print("Reject the null hypothesis. There is a significant relationship between the predi
} else {
  conclusion <- print("Fail to reject the null hypothesis. There is no significant relationship between
}
```

```
## [1] "Reject the null hypothesis. There is a significant relationship between the predictors and the r
```

As the p__value that we got from our F statistic is very small 2.664535e-14 close to almost zero which obviously smaller than the significant level of 0.05 because of which we will reject our null hypothesis and consider the alternative hypothesis which states that there is a significant relationship between the predictors and the response.

Finding R Squared.

```
r_squared <- summary(model)$r.squared
r_squared
```

```
## [1] 0.8127448
```

This value represents how well the predictors collectively account for the variation in PM2.5 concentration.Because the r__squared value is on the higher side which suggests that its a better fit of the model to the data.

Finding the best multiple regression model that explains the data and stating the final fitted regression model.

```
step_model <- step(model, direction = "both", k = 2)
```

```
## Start:  AIC=263.31
## pm25 ~ temperature + humidity + wind + precipitation
##
##                 Df Sum of Sq     RSS    AIC
## - precipitation  1      21.8  5182.4 261.55
## <none>                        5160.6 263.31
## - wind           1     621.7  5782.3 267.68
## - temperature    1    7556.8 12717.5 311.82
## - humidity       1   11750.1 16910.7 327.78
##
## Step:  AIC=261.55
## pm25 ~ temperature + humidity + wind
```

```
##
##                Df Sum of Sq     RSS    AIC
## <none>                        5182.4 261.55
## + precipitation  1      21.8  5160.6 263.31
## - wind           1     622.6  5805.1 265.90
## - temperature    1    7635.2 12817.6 310.26
## - humidity       1   11838.8 17021.2 326.14
```

```
summary(step_model)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.7588  -6.4368  -0.5659   6.4006  20.2813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.3234     8.9561  10.867 5.45e-15 ***
## temperature   1.6267     0.1859   8.753 8.39e-12 ***
## humidity     -1.2698     0.1165 -10.899 4.89e-15 ***
## wind         -0.5806     0.2323  -2.500   0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.983 on 52 degrees of freedom
## Multiple R-squared:  0.812,  Adjusted R-squared:  0.8011
## F-statistic: 74.84 on 3 and 52 DF,  p-value: < 2.2e-16
```

This stepwise model selection selects the best multiple regression model and the final fitted regression model can be seen in the summary above that includes the selected predictors and their corresponding coefficients along with other evaluation metrics.

These both R squared and the adjusted R squared tells us about the goodness of the fit which tells that how well the model describes the changes in the variable of the response.

R-squared(R2) In the context R2 tells us that how much the predictors combined can explain the variations in the PM2.5 concentration. Lets say if the R2 value is .9 which means that 90 percent of the variation in the response(PM2.5) can be explained by the predictors(temperature, humidity, wind, and precipitation). But lets say if we add more predictors into the model, it will increase the R2 which has the potential to explain more variation hence the increase in R2 but adding more predictors can lead to noise because maybe the information that we are adding is irrelevant which makes the value higher. Here adjusted R2 comes.

Adjusted R2 In this the number of the predictors are taken into account and R2 adjusts accordingly and also penalizes if there is addition of unnecessary predictors. Its a more reliable model's performance measure as compare to R2.