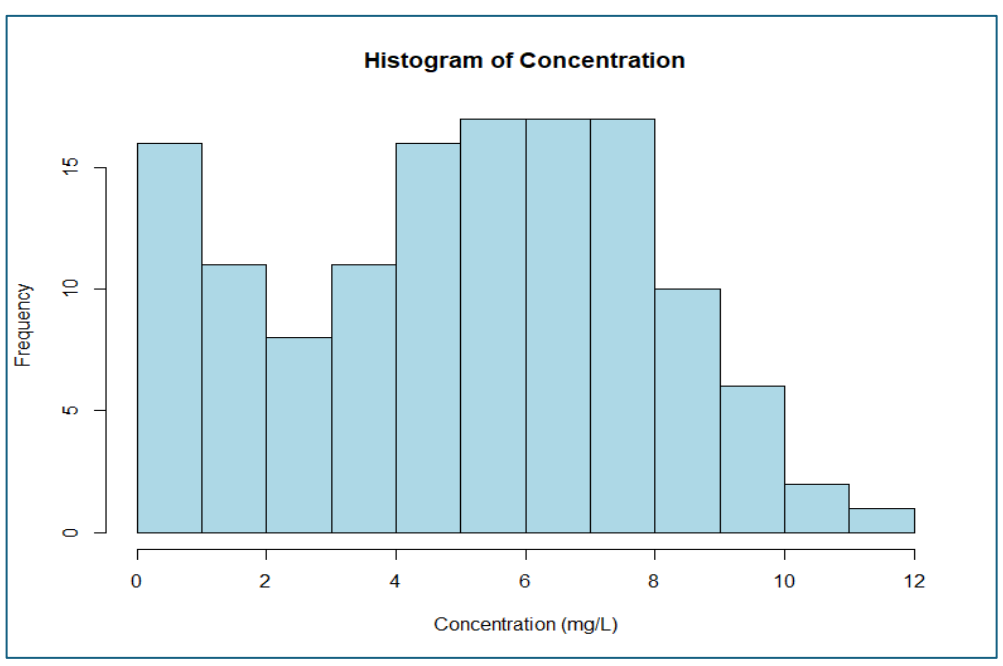
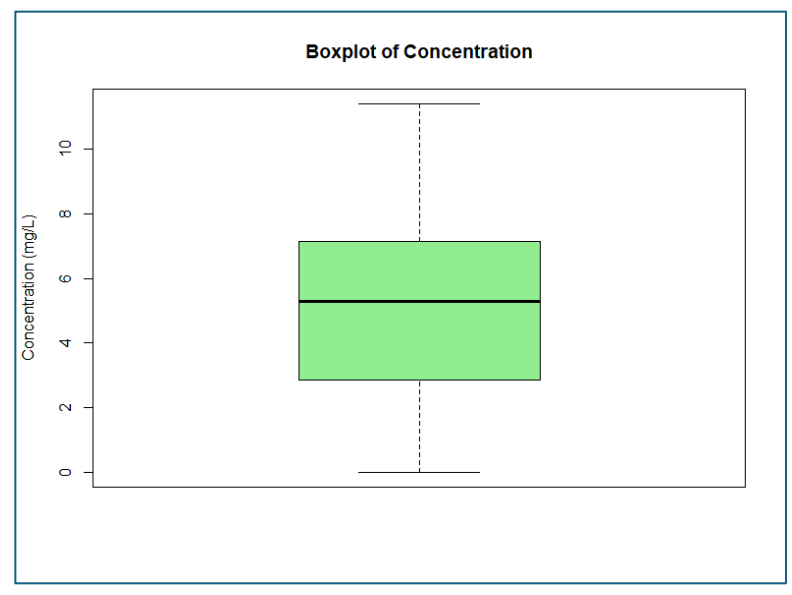
Theoph

1. Data Overview
Structure of the Dataset:
The **Theoph** dataset contains the following variables:
•**Time**: Time since the start of the experiment (in hours)
•**conc**: The concentration of theophylline (in mg/L)
•**Subject**: The subject number
•**Dose**: The dose of theophylline administered (in mg)
•**Wt**: The weight of the subject (in kg)
Dimensions:
•**Number of Observations**: 132
•**Number of Variables**: 5
Observations:
Each observation corresponds to a measurement of theophylline concentration for a subject at a particular time point.

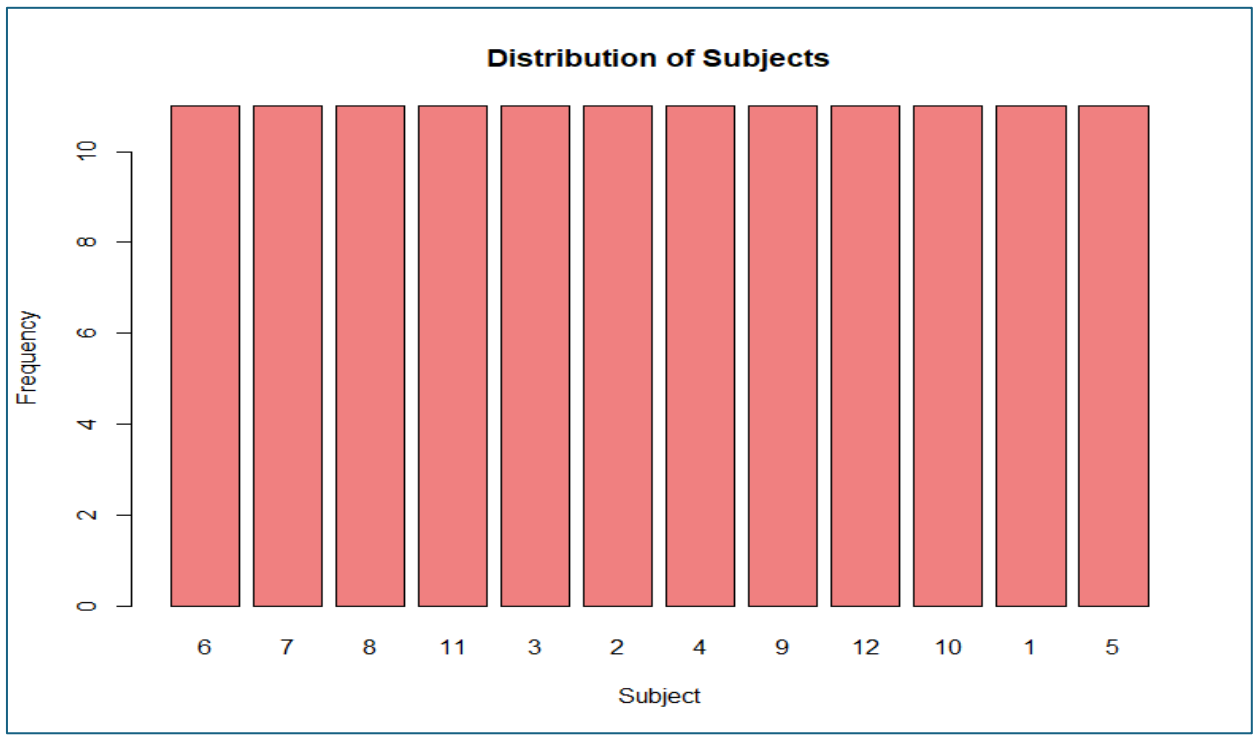
```
Summary Statistics for Concentration (conc):
> cat("Mean of conc:", mean_conc, "\n")
Mean of conc: 4.960455
> cat("Median of conc:", median_conc, "\n")
Median of conc: 5.275
> cat("Standard Deviation of conc:", sd_conc, "\n")
Standard Deviation of conc: 2.867319
> cat("Min of conc:", min_conc, "\n")
Min of conc: 0
> cat("Max of conc:", max_conc, "\n")
Max of conc: 11.4
>
```

2. Summary Statistics
Interpretation:
The mean concentration of theophylline is **4.96 mg/L**, which is close to the **median** value of **5.28 mg/L**, indicating a fairly symmetric distribution. However, the **standard deviation** of **2.87 mg/L** suggests a moderate spread around the mean. The minimum concentration is **0 mg/L**, which may represent either undetectable levels or data errors, while the maximum concentration reaches **11.4 mg/L**, indicating some variability in the concentration levels across subjects. The wide range between minimum and maximum values suggests the presence of some extreme observations, which could be further investigated.





3. Distribution Visualization
Histogram:
The histogram of the concentration variable (**conc**) shows that the distribution is right-skewed. Most of the concentrations are clustered around lower values, with a few higher concentrations.
Boxplot:
The boxplot for **conc** shows the presence of a few outliers, particularly at the higher end of the concentration range. These outliers are potentially extreme values that could be subject to further investigation.



4. Categorical Variable Analysis
•**Categorical Variable**: Subject
•**Bar Plot Insights**:

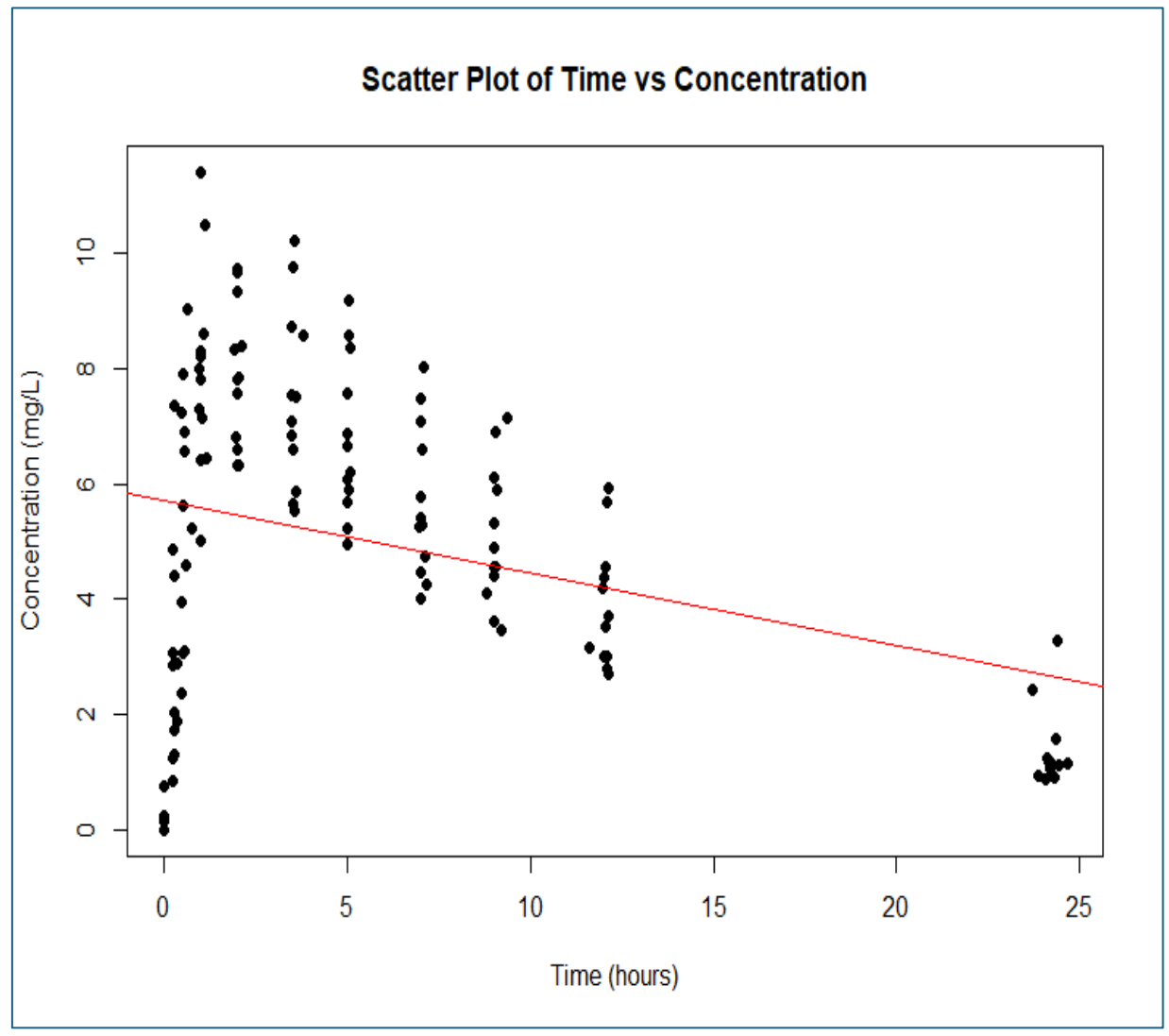
- The distribution of subjects is fairly balanced.
- The counts for each subject are relatively equal, suggesting no subject is overrepresented in the dataset.

This indicates a uniform distribution across subjects.

5. Correlation Analysis
•**Pearson Correlation**: -0.3036

- The negative correlation indicates a moderate inverse relationship between **Time** and **Concentration**. As time increases, concentration tends to decrease, but the relationship is not very strong.

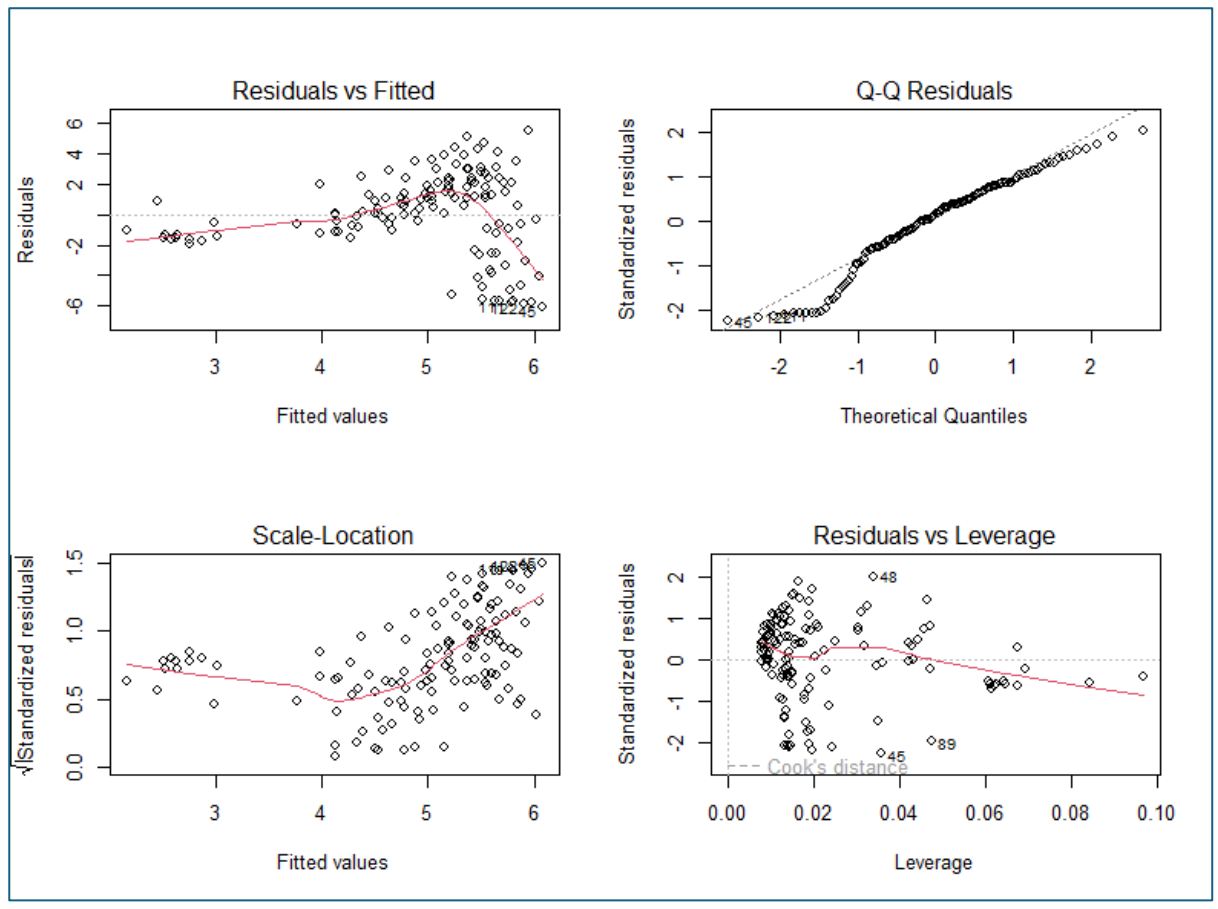
6. Scatter Plot Visualization
•**Scatter Plot**: The scatter plot shows a **moderate negative correlation** between **Time** and **Concentration**. The red trend line indicates a slight downward slope, consistent with the Pearson correlation of -0.3036. As time progresses, concentration tends to decrease, but the relationship appears somewhat scattered, suggesting other factors may influence concentration.




```
Pearson Correlation between Time and Concentration: -0.3036008
> |
```

```
> summary(lm_model_theoph)

Call:
lm(formula = conc ~ Time + Dose, data = Theoph)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0805 -1.4053  0.4145  1.9704  5.4452

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.28008    1.57612   2.716 0.007523 **
Time        -0.12569    0.03462  -3.631 0.000406 ***
Dose         0.30724    0.33388   0.920 0.359179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.744 on 129 degrees of freedom
Multiple R-squared:  0.09809,   Adjusted R-squared:  0.08411
F-statistic: 7.015 on 2 and 129 DF,  p-value: 0.001282
```

7. Multiple Regression
•**Intercept**: 4.28 (p = 0.0075)
•**Time**: -0.126 (p = 0.0004) – Significant negative effect on concentration.
•**Dose**: 0.307 (p = 0.359) – Insignificant predictor.
•**R-squared**: 0.0981 – Model explains 9.8% of variation.
•**F-statistic**: 7.015 (p = 0.0013) – Overall model is significant.
Conclusion:
Time significantly affects concentration, while Dose does not. The model has limited explanatory power.



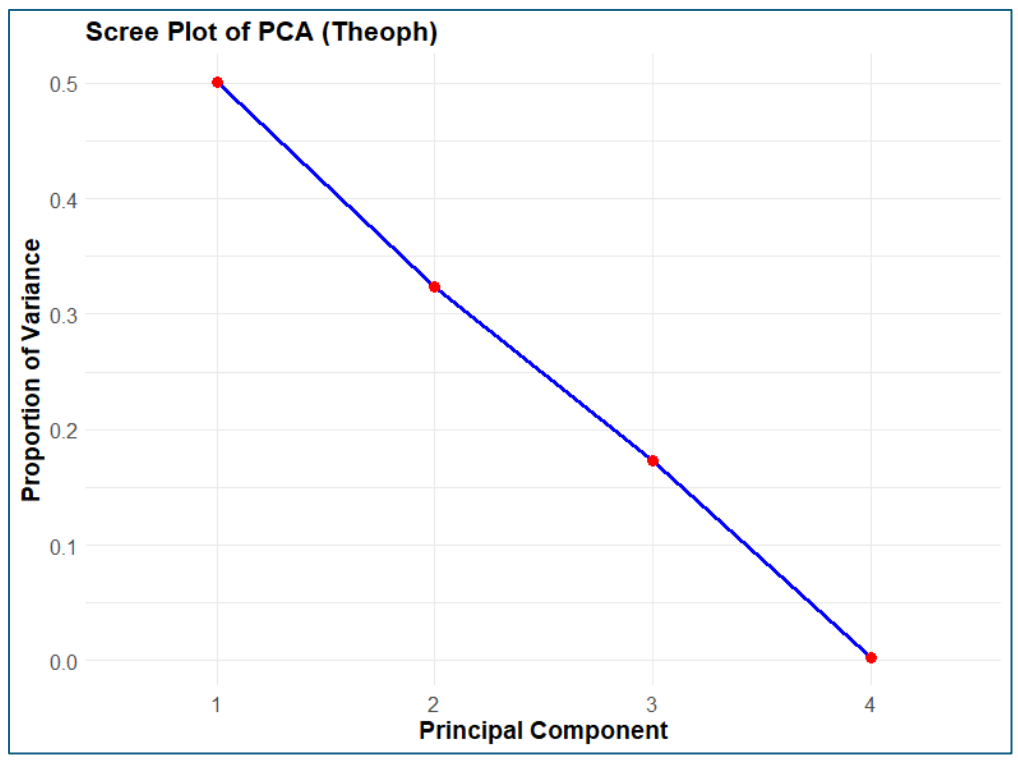
8. Residual Analysis for the Iris Dataset:

•**Residual vs. Fitted Plot**: The plot reveals a relatively random scatter around zero with no clear patterns, which indicates **homoscedasticity** and no obvious issues with the variance of residuals.

•**Normal Q-Q Plot**: The residuals appear to mostly follow a straight line, suggesting that they are approximately normally distributed.

•**Scale-Location Plot**: The residuals appear to be evenly distributed across the fitted values, further confirming that the variance of residuals is constant (homoscedasticity).

•**Residuals vs. Leverage Plot**: There are no points that stand out as high leverage points or influential outliers, meaning the regression model is not unduly influenced by any single data point.



9. Principal Component Analysis (PCA)

PCA Summary:

We performed PCA on the numerical variables (**Time**, **conc**, **Dose**, and **Wt**). The scree plot suggests that the first two principal components explain a significant portion of the total variance in the dataset.

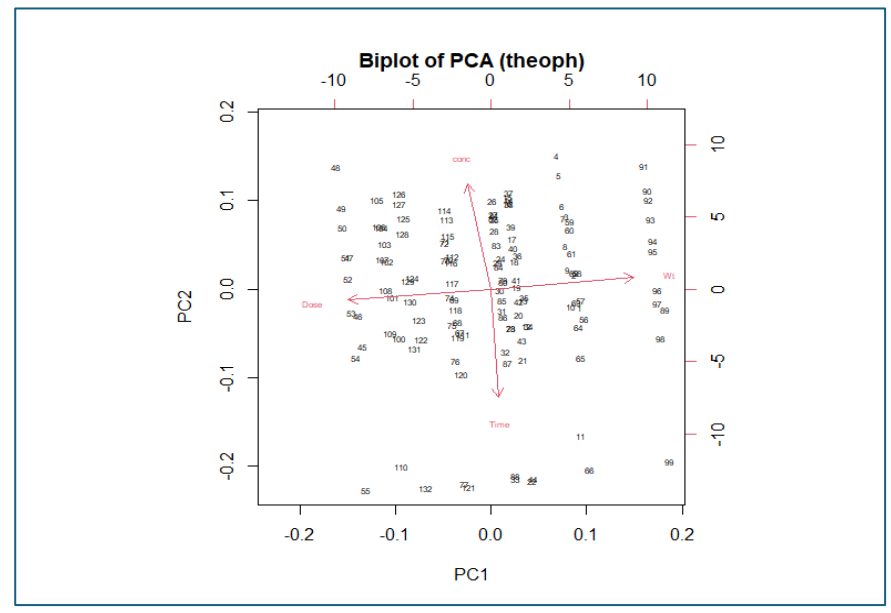
•**PC1**: Accounts for 50.3% of the variance.

•**PC2**: Accounts for 32% of the variance.

Insights:

The first two components together explain 84% of the total variance, indicating that these components largely drive the data.

```
> summary(pca_theoph)
Importance of components:
                          PC1     PC2     PC3      PC4
Standard deviation     1.4149 1.1383 0.8320 0.10180
Proportion of Variance 0.5004 0.3239 0.1730 0.00259
Cumulative Proportion  0.5004 0.8244 0.9974 1.00000
```



Biplot:

The biplot shows how the variables load onto the first two principal components. We observe that **weight** and **Dose** load strongly on **PC1**, while **time & conc** contributes to **PC2**.

Conclusion

Key Findings:

•The dataset shows that **Time** and **Dose** are important factors in predicting **conc**, with **Time** being the stronger predictor.

•The correlation between **Time** and **conc** is moderate, indicating a negative relationship between time and the concentration of theophylline.

•Principal Component Analysis (PCA) revealed that the first two principal components explain most of the variance in the data, highlighting the importance of **Weight** and **Dose** in understanding the variability in the dataset.

•The regression model diagnostics suggest a good fit with no major violations of assumptions.

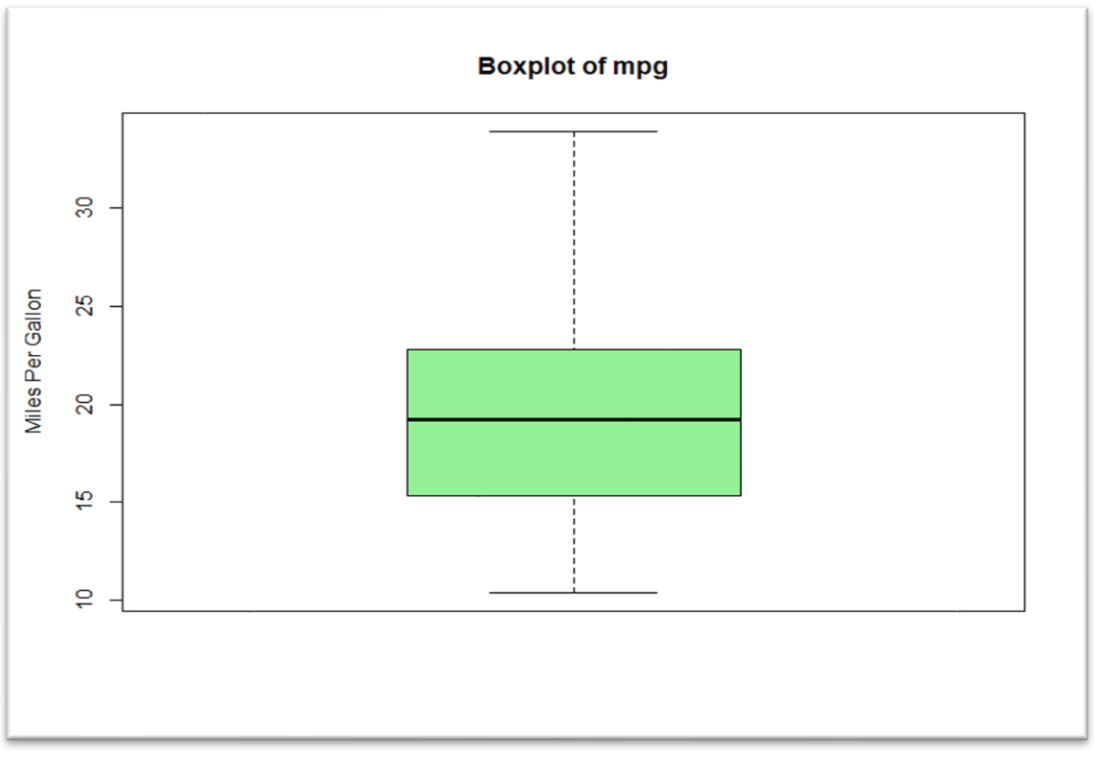
This analysis provides useful insights into the pharmacokinetics of theophylline and highlights the factors that most influence its concentration in the bloodstream over time. Further analysis could explore the effect of additional variables, such as subject-specific characteristics, or refine the model using nonlinear approaches.

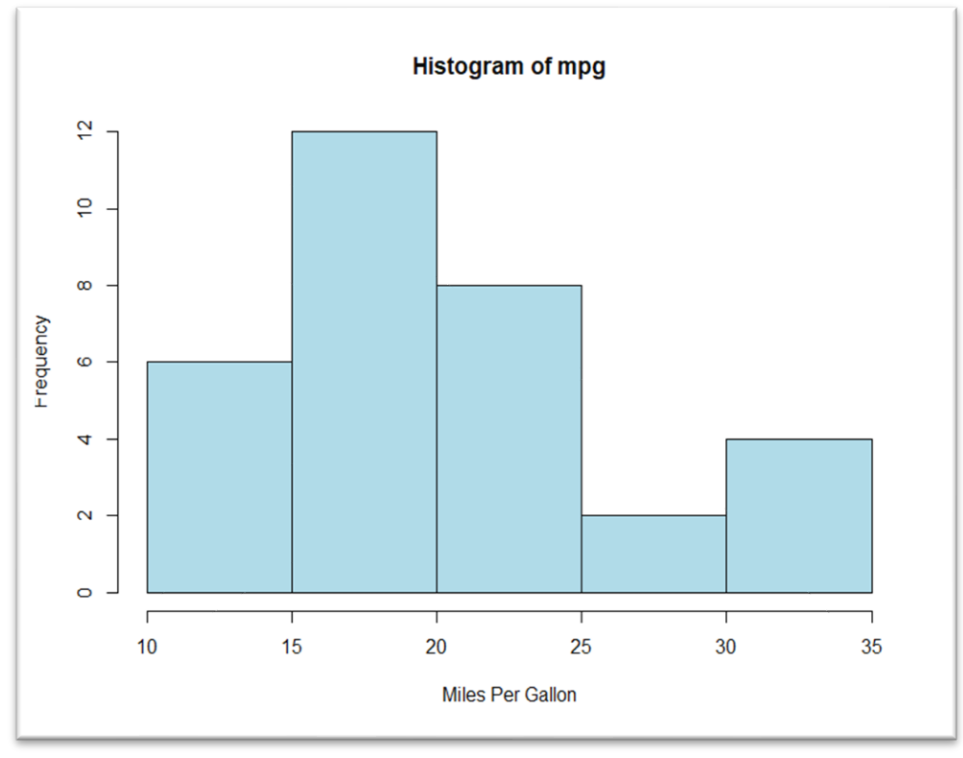
mtcars

1. Data Overview
The **mtcars** dataset consists of **32 observations** and **11 variables**:
•**mpg**: Miles per gallon (continuous variable)
•**cyl**: Number of cylinders (categorical variable)
•**disp**: Displacement (continuous variable)
•**hp**: Horsepower (continuous variable)
•**drat**: Rear axle ratio (continuous variable)
•**wt**: Weight (continuous variable)
•**qsec**: Quarter-mile time (continuous variable)
•**vs**: Engine type (binary variable: V/S)
•**am**: Transmission type (binary variable: Automatic/Manual)
•**gear**: Number of forward gears (categorical variable)
•**carb**: Number of carburetors (continuous variable)

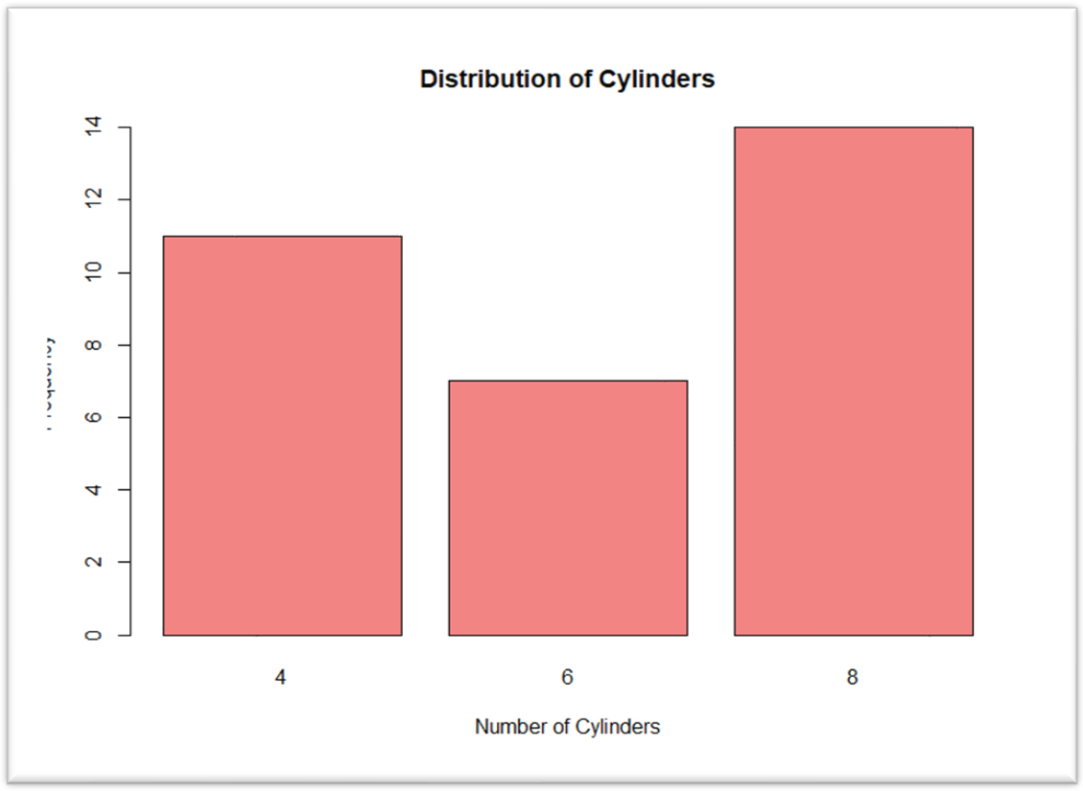
```
> # Display the results
> cat("Mean of mpg:", mean_mpg, "\n")
Mean of mpg: 20.09062
> cat("Median of mpg:", median_mpg, "\n")
Median of mpg: 19.2
> cat("Standard Deviation of mpg:", sd_mpg, "\n")
Standard Deviation of mpg: 6.026948
> cat("Min of mpg:", min_mpg, "\n")
Min of mpg: 10.4
> cat("Max of mpg:", max_mpg, "\n")
Max of mpg: 33.9
```

2. **Summary Statistics**
The **mean mpg** of **20.09** represents the average fuel efficiency of the cars in the dataset.
•The **median mpg** of **19.20** is lower than the mean, indicating a slight **right-skewness** in the distribution (i.e., more cars have lower fuel efficiency).
•The **standard deviation** of **6.03** shows moderate variation in the mpg values, which is expected in a dataset involving various types of vehicles.
•The **minimum** and **maximum** mpg values of **10.40** and **33.90**, respectively, highlight the range of fuel efficiencies in the dataset.





3. Distribution Visualization
•**Histogram**: The histogram of **mpg** shows that the distribution is somewhat **right-skewed**, with a peak near **20 mpg** and a few cars showing much higher mpg values (indicating good fuel efficiency).
•**Boxplot**: The boxplot confirms the right-skewness of the **mpg** distribution, with a few outliers on the higher end of the range. These outliers represent cars with exceptionally high fuel efficiency.



4. Categorical Variable Analysis

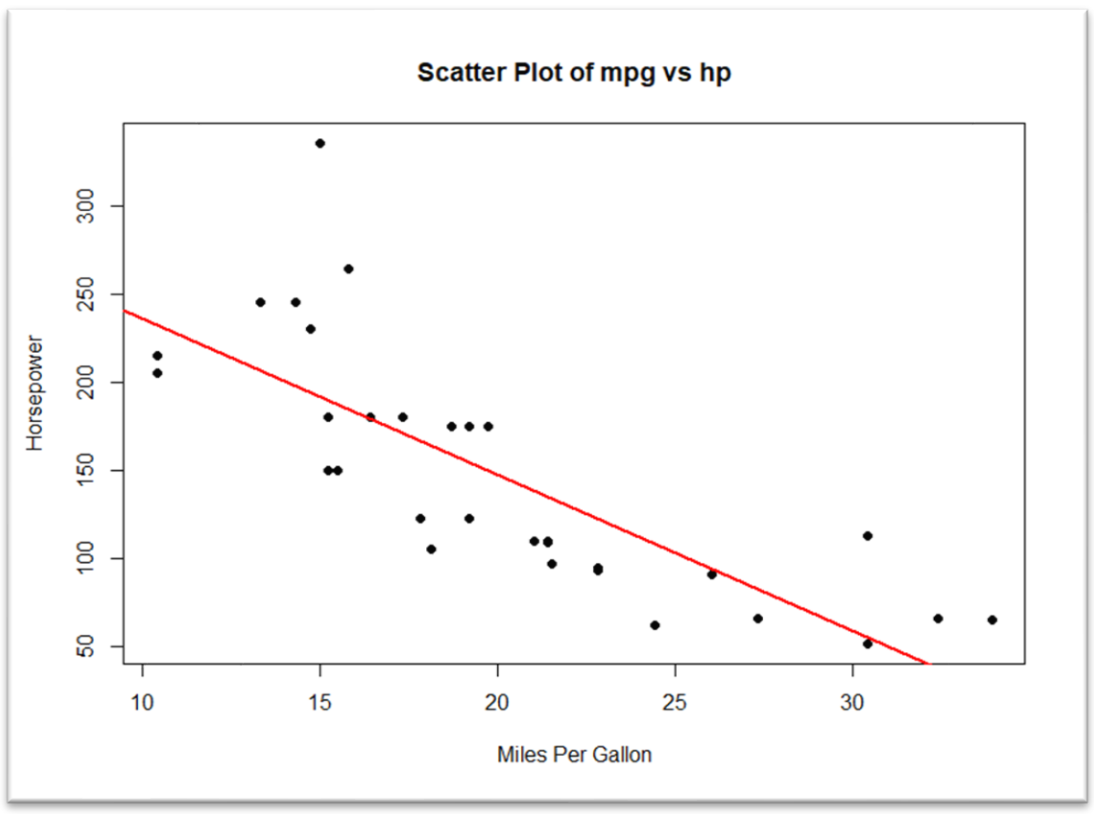
We selected **'cyl'** (the number of cylinders in the car) as the categorical variable for analysis.

The bar plot for the number of cylinders shows that most cars have 8 **cylinders.**

```
> cat("Pearson Correlation between mpg and hp:", cor_mpg_hp, "\n")
Pearson Correlation between mpg and hp: -0.7761684
```

5. Correlation Analysis

•A **negative correlation** of **-0.776** indicates a strong inverse relationship between **horsepower (hp)** and **miles per gallon (mpg)**: as horsepower increases, miles per gallon tends to decrease. This makes sense since higher horsepower engines typically consume more fuel.



6. Scatter Plot Visualization

•The scatter plot confirms the **inverse relationship** between **mpg** and **hp**, with cars having high horsepower showing a marked decrease in miles per gallon.

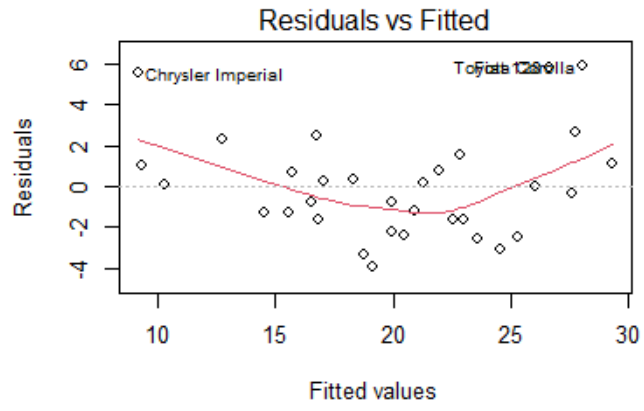
```
> summary(lm_model)

Call:
lm(formula = mpg ~ hp + wt, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

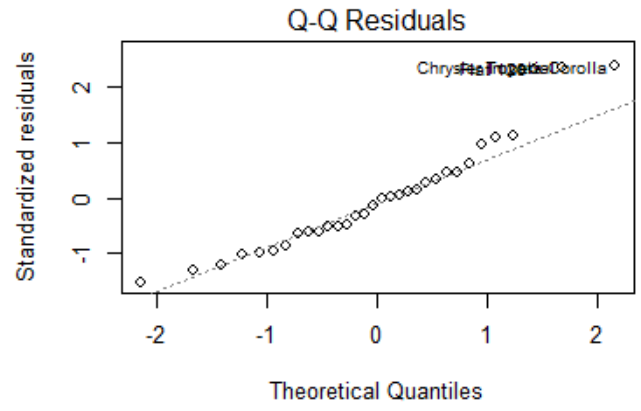
Interpretation of Coefficients:

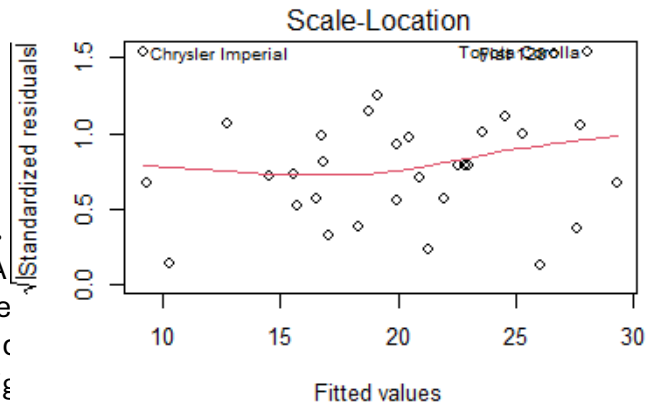
•**Intercept (37.23)**: If **hp** and **wt** were both zero (which is not practically meaningful in this context), the **mpg** would be **37.23**. This value represents the baseline miles per gallon.

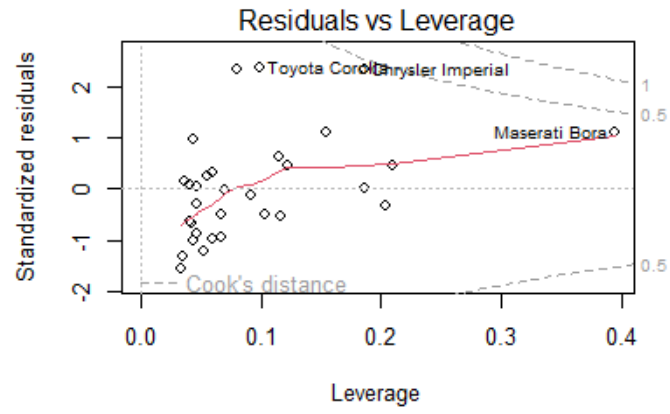
•**Coefficient for hp (-0.03177)**: For each additional unit of horsepower, the **mpg** decreases by **0.03177** units. This result confirms the inverse relationship between horsepower and fuel efficiency, indicating that higher horsepower engines tend to be less fuel-efficient.

•**Coefficient for wt (-3.87783)**: For each additional unit of weight, the **mpg** decreases by **3.88** units. This significant negative effect of weight on fuel efficiency aligns with the expectation that heavier cars typically consume more fuel.









5.
•A
be
inc
hig

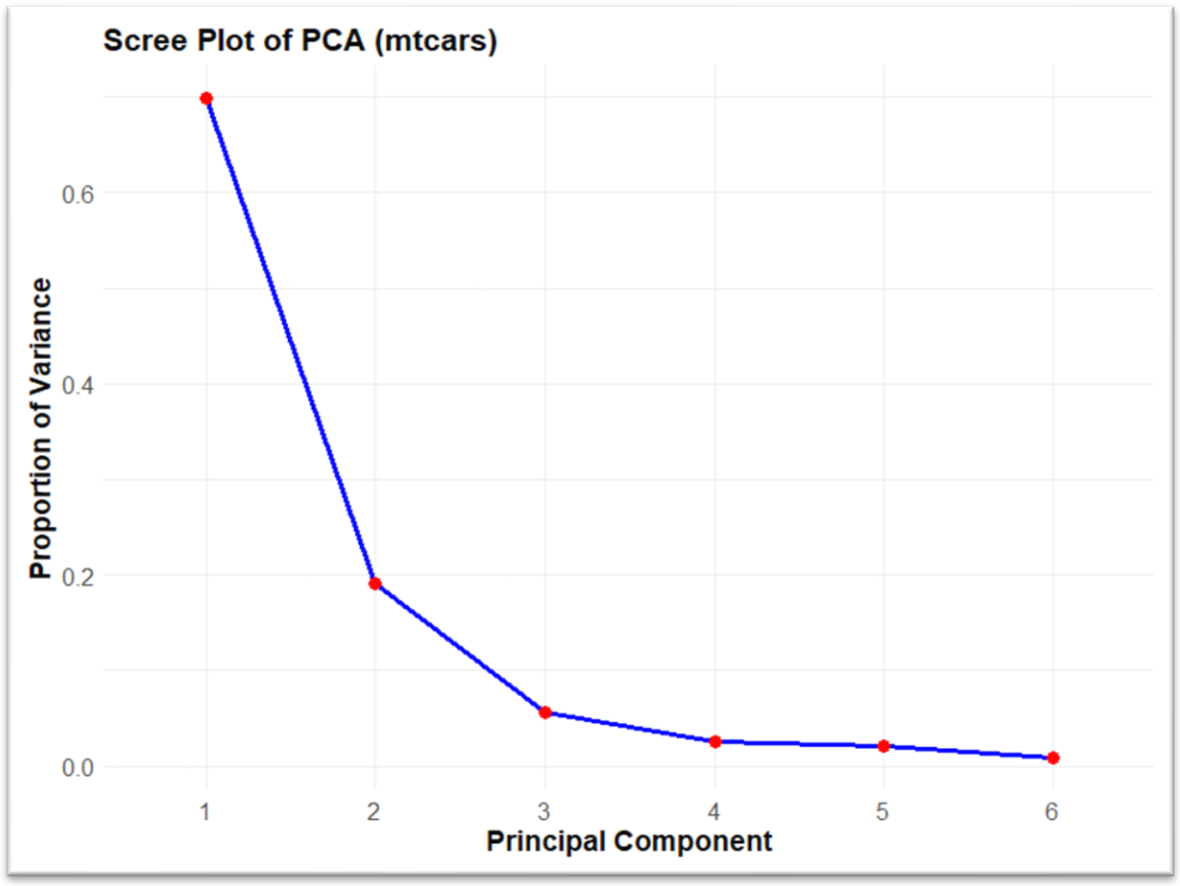
8. Model Diagnostics

Residual plots are created to assess the **homoscedasticity** and **normality** of the regression residuals.

•**Homoscedasticity**: The residual plot indicates **constant variance** across fitted values, supporting the assumption of homoscedasticity.

•**Normality of Residuals**: The Q-Q plot shows that the residuals are reasonably close to a normal distribution, confirming that the regression model fits the data well.

•The residual diagnostics suggest that the model assumptions are largely met, indicating a good fit.



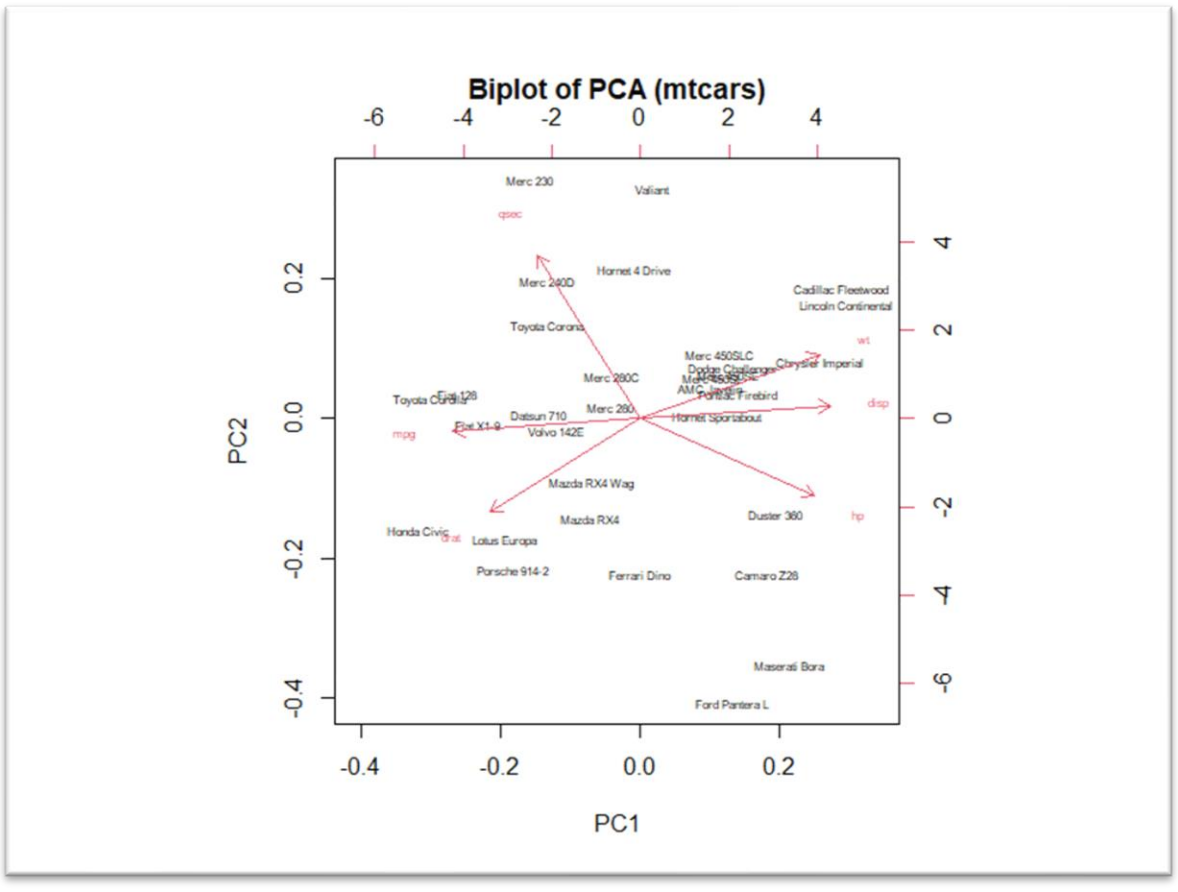
9. Principal Component Analysis (PCA)

•PCA has been done for five of the variables: HP, Weight, MPG, drat, qsec,disp. There would be 5 principal components. PC1 explains 66% of the variability in the data. PC1 and PC2 combined account for 85% of variability.

```
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5707 1.6280 0.79196 0.51923 0.47271 0.46000 0.3678
Proportion of Variance 0.6008 0.2409 0.05702 0.02451 0.02031 0.01924 0.0123
Cumulative Proportion  0.6008 0.8417 0.89873 0.92324 0.94356 0.96279 0.9751
                           PC8     PC9    PC10   PC11
Standard deviation     0.35057 0.2776 0.22811 0.1485
Proportion of Variance 0.01117 0.0070 0.00473 0.0020
Cumulative Proportion  0.98626 0.9933 0.99800 1.0000
```

10. PCA Interpretation

Wt, disp, hp has positive dependence on PC 1
mpg & drat has negative dependence on PC1
mpg,drat, disp,disp,hp seems kess effected by pc2
Qsec has positive dependence of PC2



Conclusion

•**Univariate Analysis**: The mpg distribution is slightly right-skewed, with a relatively high variability in fuel efficiency across the vehicles. No extreme outliers were detected.

•**Multivariate Analysis**: A strong negative correlation exists between mpg and hp, indicating that higher horsepower generally leads to lower fuel efficiency. The multiple regression model shows that both horsepower and weight are significant predictors of fuel efficiency, with weight having a more substantial impact.

•**PCA**: PCA revealed that the first two components explain the majority of the variance in the dataset, with the first component capturing engine power and the second capturing weight and acceleration characteristics.

Overall, the analysis highlights key patterns in vehicle performance, and the use of PCA effectively reduced the dimensionality of the dataset while preserving critical information.

iris

1. Data Overview

The **Iris** dataset contains measurements for 150 iris flowers across 5 variables:

- **Species** (categorical): the species of the iris flower (Setosa, Versicolor, Virginica).
- **Sepal.Length** (continuous): length of the sepal (in cm).
- **Sepal.Width** (continuous): width of the sepal (in cm).
- **Petal.Length** (continuous): length of the petal (in cm).
- **Petal.Width** (continuous): width of the petal (in cm).

Dimensions:

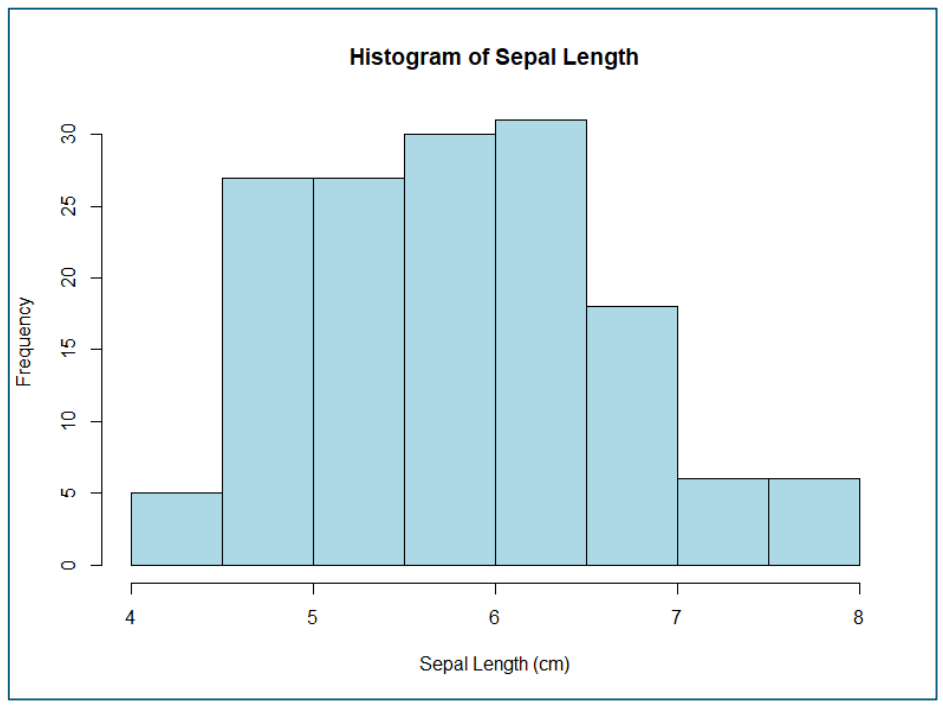
- **Observations:** 150
- **Variables:** 5

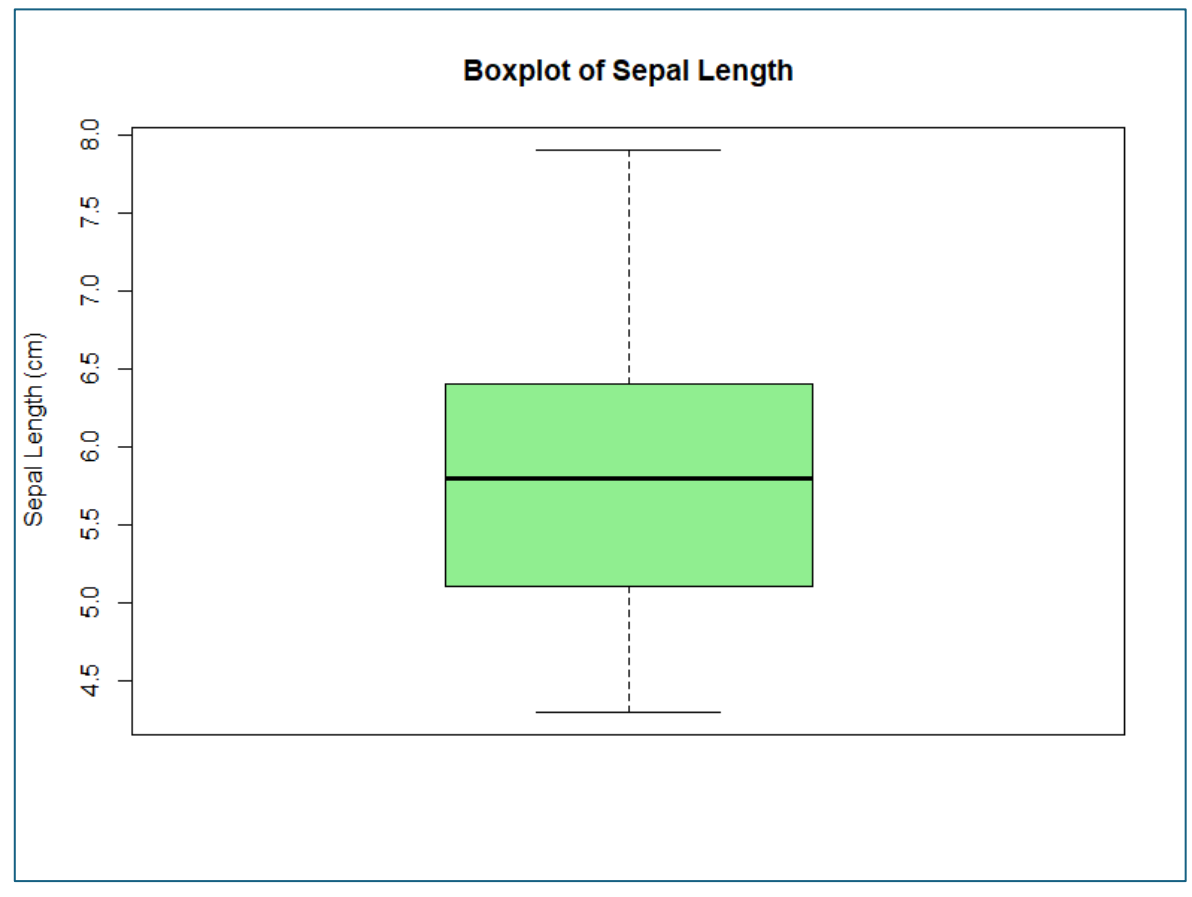
```
ummary statistics for Sepal Length:
 cat("Mean of Sepal Length:", mean_sepal_length, "\n")
ean of Sepal Length: 5.843333
 cat("Median of Sepal Length:", median_sepal_length, "\n")
edian of Sepal Length: 5.8
 cat("Standard Deviation of Sepal Length:", sd_sepal_length, "\n")
tandard Deviation of Sepal Length: 0.8280661
 cat("Min of Sepal Length:", min_sepal_length, "\n")
in of Sepal Length: 4.3
 cat("Max of Sepal Length:", max_sepal_length, "\n")
ax of Sepal Length: 7.9
```

2. Summary Statistics

Mean: 5.843

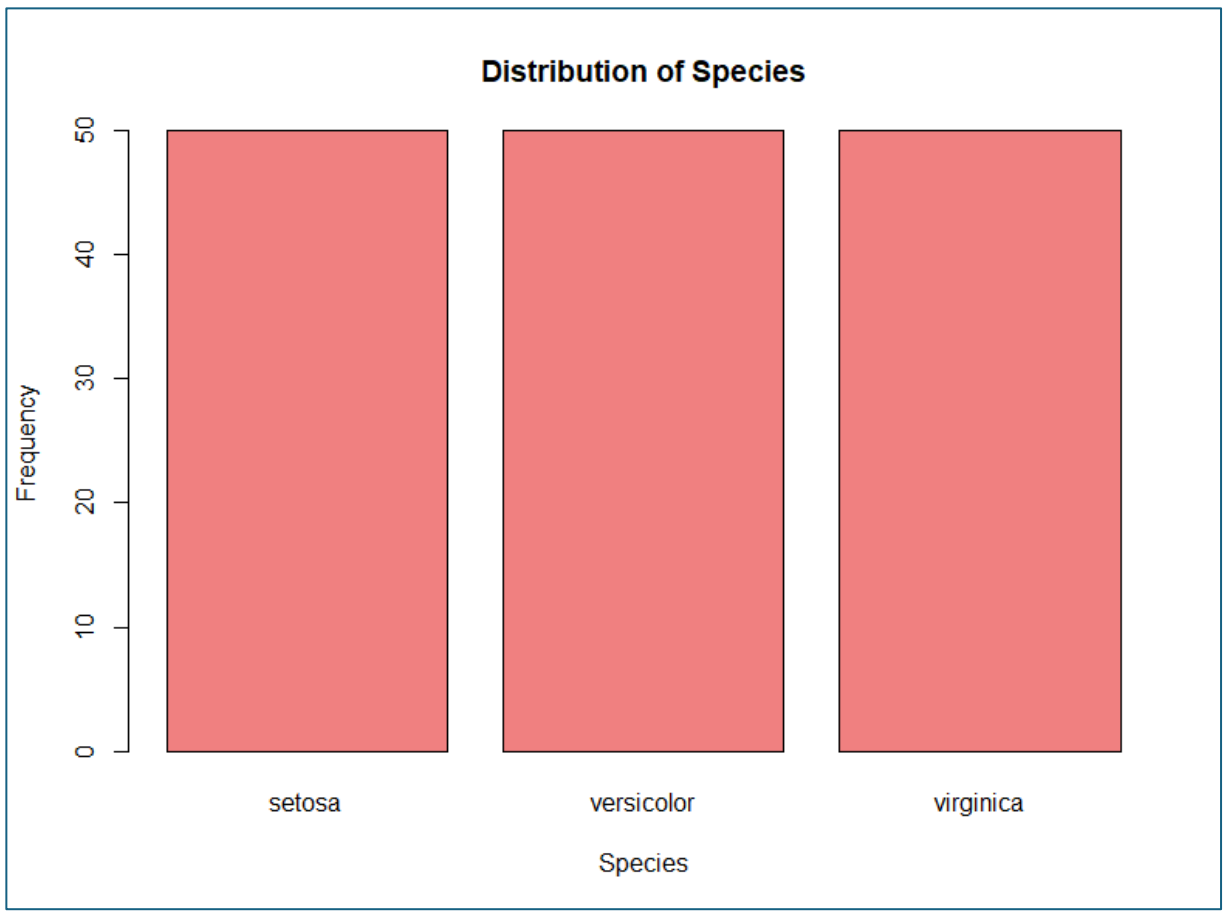
The **Sepal Length** variable has a mean of **5.843** cm, indicating that the average sepal length is about 5.8 cm. The data has a moderate spread (SD = **0.828**) with a minimum of **4.30** cm and a maximum of **7.90** cm.





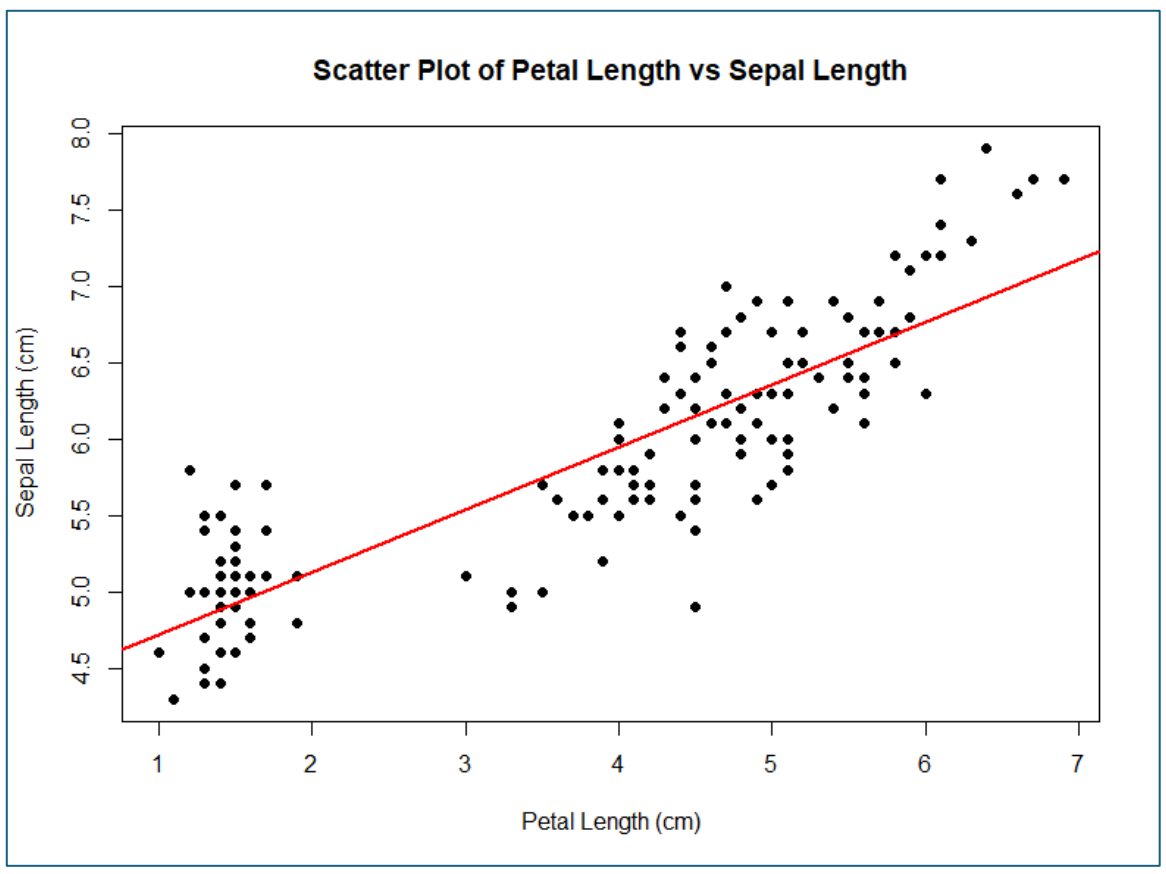
3. Distribution Visualization

- **Histogram and Boxplot:** The histogram shows a roughly normal distribution for **Sepal Length**, with a slight right skew. The boxplot indicates the presence of potential outliers with a few data points above **7.50** cm.



4. Categorical Variable Analysis (for Species)
The **Species** variable is categorical with three levels: **Setosa**, **Versicolor**, and **Virginica**. The bar plot shows even distribution of the three species, with 50 each.

```
Pearson Correlation between Sepal Length and Petal Length: 0.8717538
```

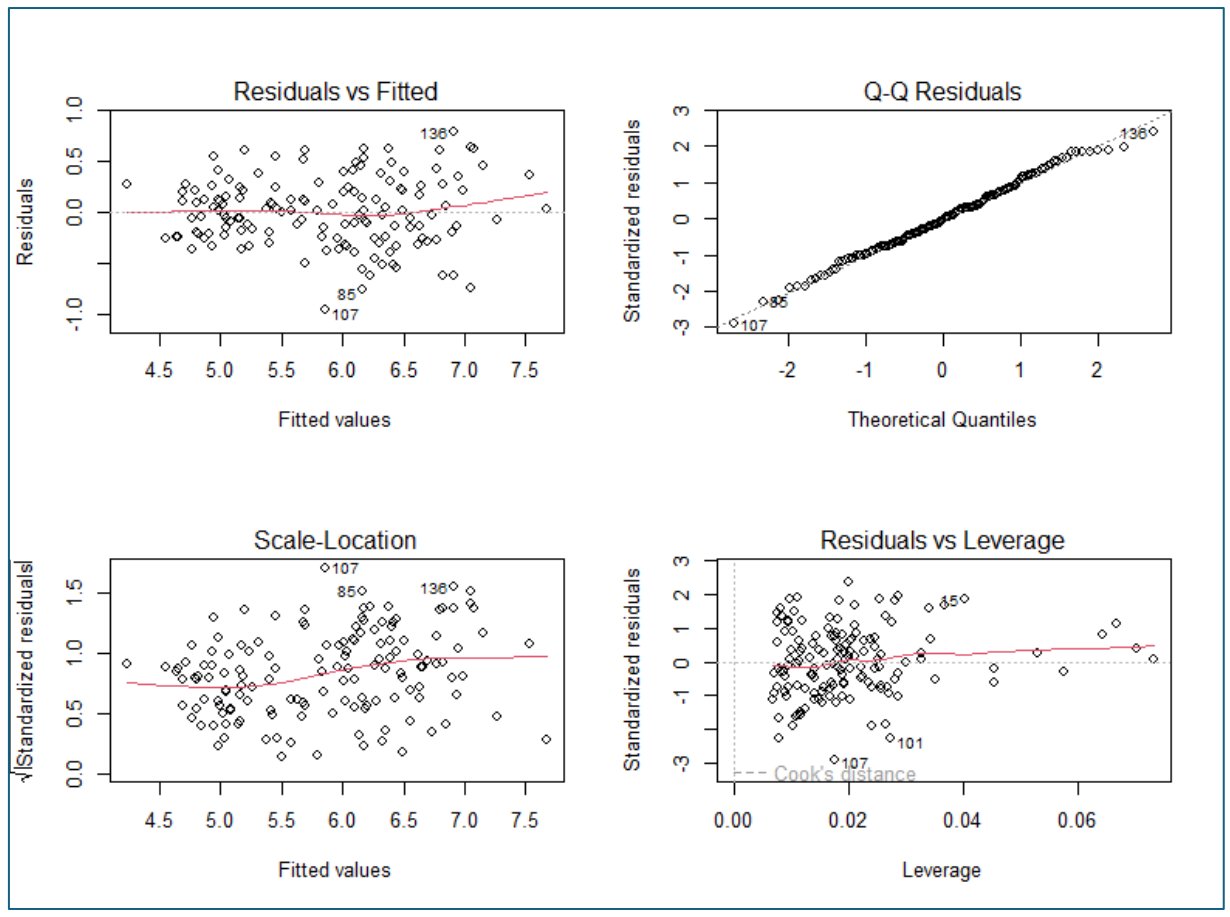


5. Correlation Analysis
The **Pearson Correlation** coefficient between **Sepal Length** and **Petal Length** is **0.8718**, indicating a strong positive relationship between the two variables. As **Sepal Length** increases, **Petal Length** also increases, with the two variables showing a high degree of linear association.

6. Scatter Plot Visualization
The scatter plot shows a clear positive linear relationship between **Sepal Length** and **Petal Length**, with a high concentration of data points forming an upward trend. The **red trend line** confirms the strong correlation.

```
> summary(lm_model_iris)

Call:
lm(formula = Sepal.Length ~ Petal.Length + Sepal.Width, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96159 -0.23489  0.00077  0.21453  0.78557

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.24914    0.24797    9.07 7.04e-16 ***
Petal.Length  0.47192    0.01712   27.57  < 2e-16 ***
Sepal.Width   0.59552    0.06933    8.59 1.16e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3333 on 147 degrees of freedom
Multiple R-squared:  0.8402,    Adjusted R-squared:  0.838
F-statistic: 386.4 on 2 and 147 DF,  p-value: < 2.2e-16
```

7. Multiple Regression Analysis
The multiple linear regression model was fitted to predict **Sepal Length** using **Petal Length** and **Sepal Width** as predictors. The summary of the model is as follows:
•**Intercept**: The estimated intercept is **2.2491**, indicating that when both **Petal Length** and **Sepal Width** are zero, the predicted **Sepal Length** is approximately 2.25 cm.
•**Petal Length**: The coefficient for **Petal Length** is **0.4719**, suggesting that for each unit increase in **Petal Length**, the **Sepal Length** is expected to increase by approximately 0.47 cm, holding **Sepal Width** constant. This variable is highly significant with a **p-value** less than 0.001.
•**Sepal Width**: The coefficient for **Sepal Width** is **0.5955**, meaning that for each unit increase in **Sepal Width**, the **Sepal Length** increases by approximately 0.60 cm, while holding **Petal Length** constant. This variable is also highly significant with a **p-value** less than 0.001.



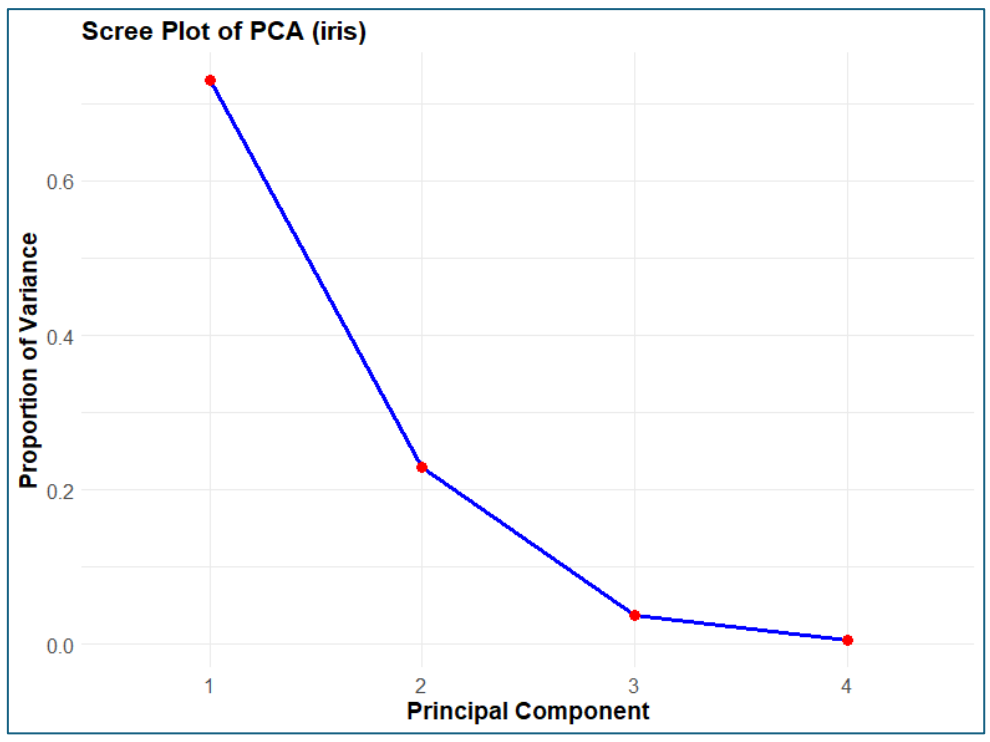
8. Residual Analysis for the Iris Dataset:

•**Residual vs. Fitted Plot**: The plot reveals a relatively random scatter around zero with no clear patterns, which indicates **homoscedasticity** and no obvious issues with the variance of residuals.

•**Normal Q-Q Plot**: The residuals appear to mostly follow a straight line, suggesting that they are approximately normally distributed.

•**Scale-Location Plot**: The residuals appear to be evenly distributed across the fitted values, further confirming that the variance of residuals is constant (homoscedasticity).

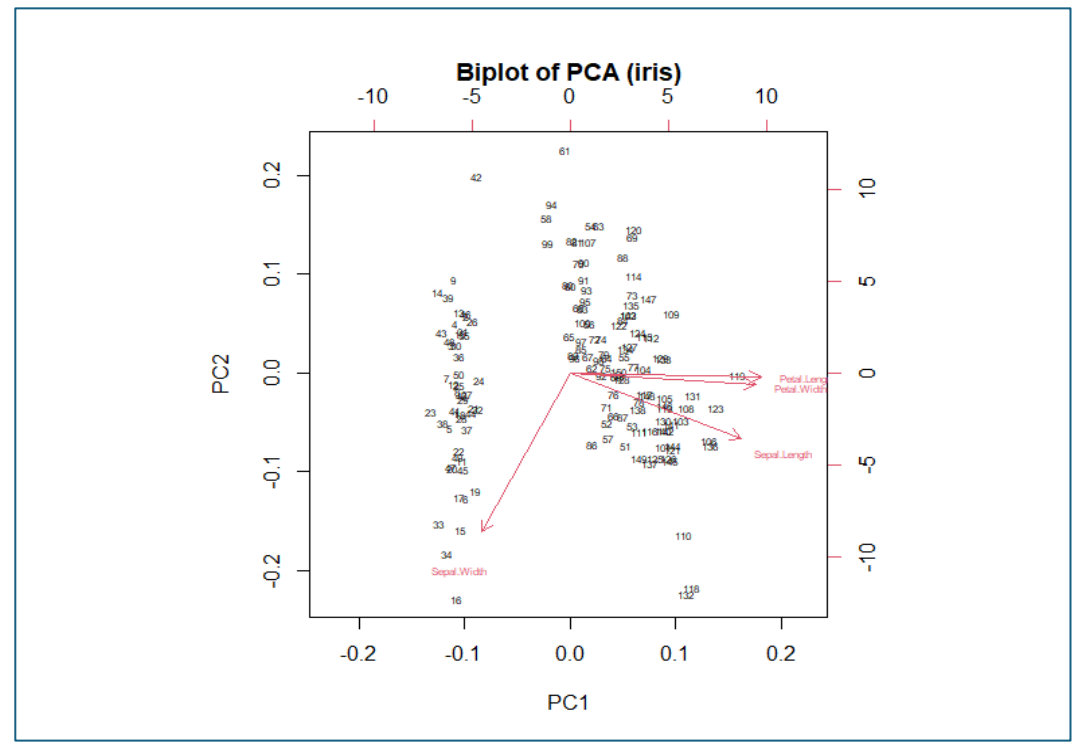
•**Residuals vs. Leverage Plot**: No points stand out as high leverage points or influential outliers, meaning the regression model is not unduly influenced by any single data point.



9. Principal Component Analysis (PCA)

•The **scree plot** indicates that the first two components explain about **92%** of the variance in the data. These two components are sufficient to capture most of the variance, which means dimensionality reduction could be effective with minimal loss of information.

```
> summary(pca_iris)
Importance of components:
                          PC1     PC2      PC3      PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
>
```



10. PCA Interpretation (Biplot)

Petal length, Petal width, sepal length are positively dependent on PC1

Sepal width has negative dependence on PC2

Conclusion:

•The **Iris dataset** is well-suited for analysis, with strong relationships between several variables.

•**Univariate analysis** suggests that **Sepal Length** is roughly normally distributed with slight skewness.

•**Multivariate analysis** reveals that both **Petal Length** and **Sepal Width** are significant predictors of **Sepal Length**, with a strong correlation between **Sepal Length** and **Petal Length**.

•**PCA** highlights that the first two components can explain a large portion of the dataset's variance, and the biplot aids in visualizing the separation between species.

airquality

1. Data Overview

The airquality dataset contains daily air quality measurements from May to September 1973. The dataset has 153 observations across 6 variables:

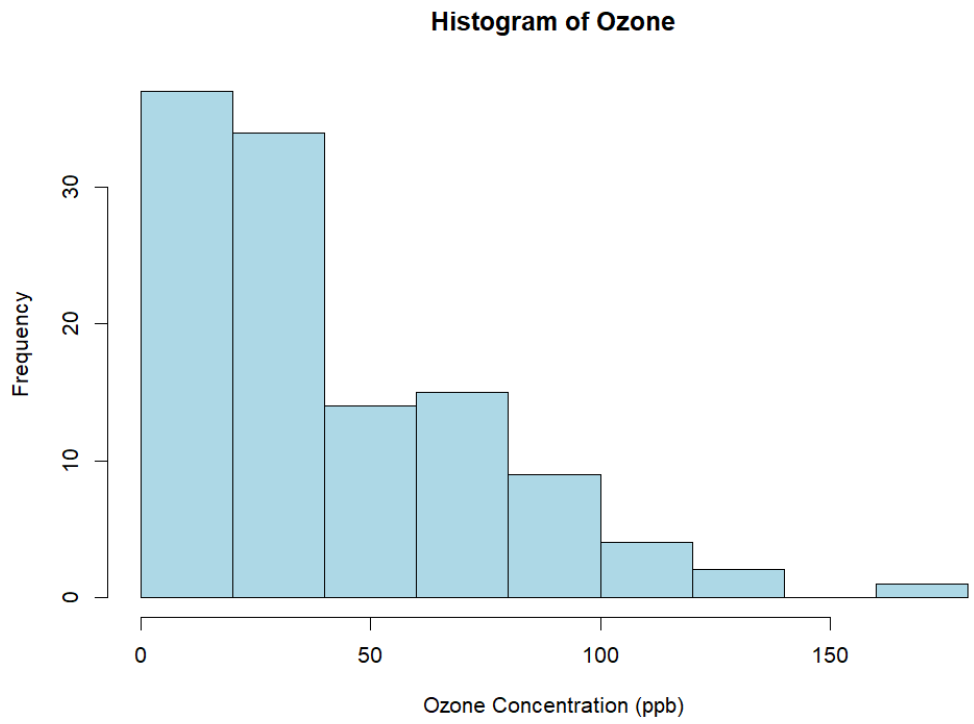
- **Ozone**: Ozone concentration (numeric, ppb)
- **Solar.R**: Solar radiation (numeric, Langley)
- **Wind**: Wind speed (numeric, mph)
- **Temp**: Temperature (numeric, Fahrenheit)
- **Month**: Month of the year (factor, 5-9)
- **Day**: Day of the month (factor, 1-31)

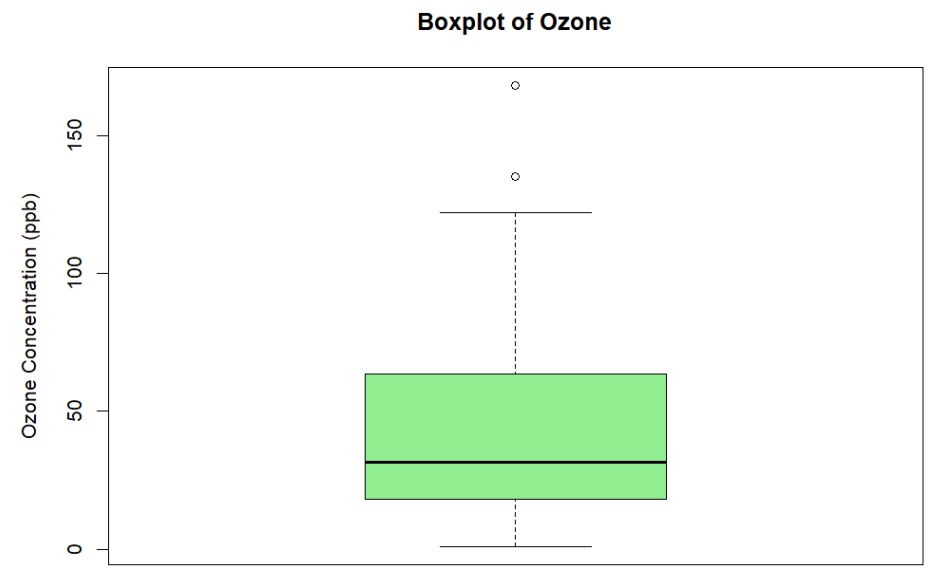
```
> cat("Mean of Ozone:", mean_ozone, "\n")
Mean of Ozone: 42.12931
> cat("Median of Ozone:", median_ozone, "\n")
Median of Ozone: 31.5
> cat("Standard Deviation of Ozone:", sd_ozone, "\n")
Standard Deviation of Ozone: 32.98788
> cat("Min of Ozone:", min_ozone, "\n")
Min of Ozone: 1
> cat("Max of Ozone:", max_ozone, "\n")
Max of Ozone: 168
>
```

2. Summary Statistics

- **Mean**: 42.13 ppb
- **Median**: 31.5 ppb
- **Standard Deviation**: 32.99 ppb
- **Minimum**: 1 ppb
- **Maximum**: 168 ppb

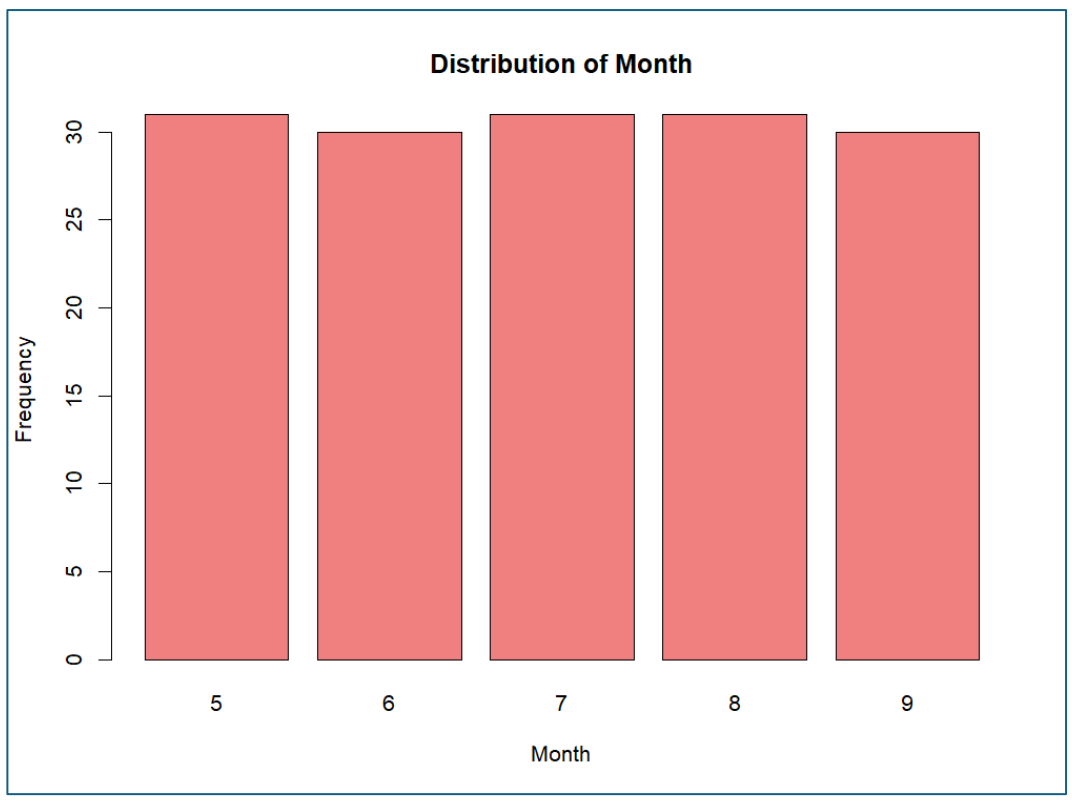
The **mean** is higher than the **median**, indicating a slight **positive skew** in the data. The **standard deviation** suggests significant variability in Ozone levels. The **range** (from 1 to 168) and presence of **outliers** (observations near the upper limit) indicate substantial variability across observations.





3. Distribution Visualization

The **histogram** and **boxplot** for Ozone show a right-skewed distribution with several outliers (as seen from the spread in the boxplot).



4. Categorical Variable Analysis

For the **Month** variable, which represents the month of observation, the distribution of months is as follows:

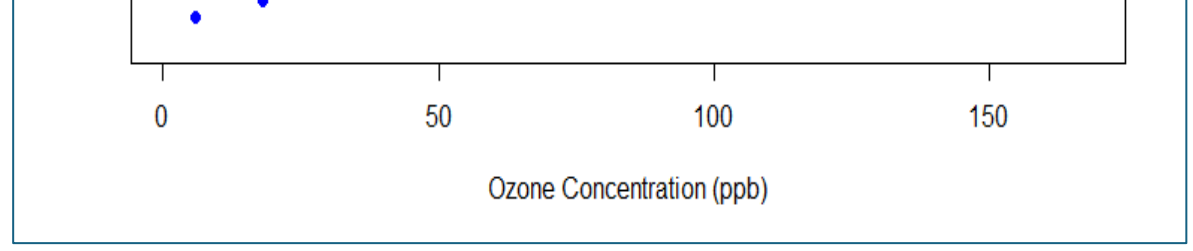
- **May (5)**: 31 observations
- **June (6)**: 30 observations
- **July (7)**: 31 observations
- **August (8)**: 31 observations
- **September (9)**: 30 observations

The data is evenly distributed across the five months, indicating balanced sampling from each month.

```
Pearson Correlation between Ozone and Temp: 0.6983603
>
```

5. Correlation Analysis

- A moderate positive correlation exists between Ozone levels and Temperature, suggesting that higher temperatures are associated with higher ozone concentrations.



6. Scatter Plot Visualization

- **Scatter Plot**: The scatter plot shows a **moderate positive correlation** between **Temp** and ozone **Concentration**..

```
> summary(lm_model)

Call:
lm(formula = Ozone ~ Temp + Wind, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-41.251 -13.695  -2.856  11.390 100.367

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -71.0332    23.5780  -3.013   0.0032 **
Temp          1.8402     0.2500   7.362 3.15e-11 ***
Wind         -3.0555     0.6633  -4.607 1.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.85 on 113 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.5687,    Adjusted R-squared:  0.5611
F-statistic:  74.5 on 2 and 113 DF,  p-value: < 2.2e-16

>
```

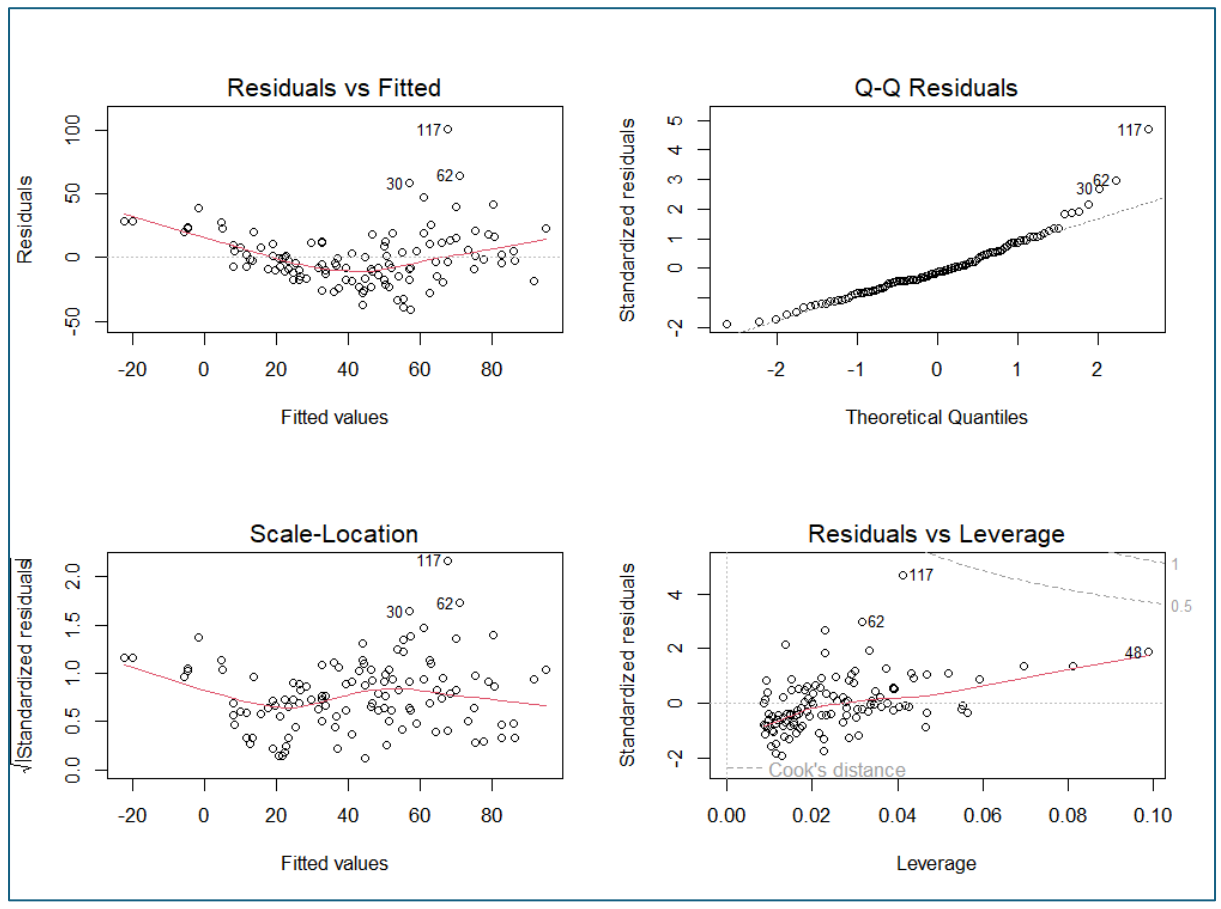
7. Multiple Regression

- **Intercept**: -71.03 (p-value = 0.0032) — Statistically significant.
- **Temp**: 1.84 (p-value < 0.001) — For each 1°F increase in Temperature, Ozone increases by 1.84 ppb.
- **Wind**: -3.06 (p-value = 1.08e-05) — For each unit increase in Wind, Ozone decreases by 3.06 ppb.

Model Statistics

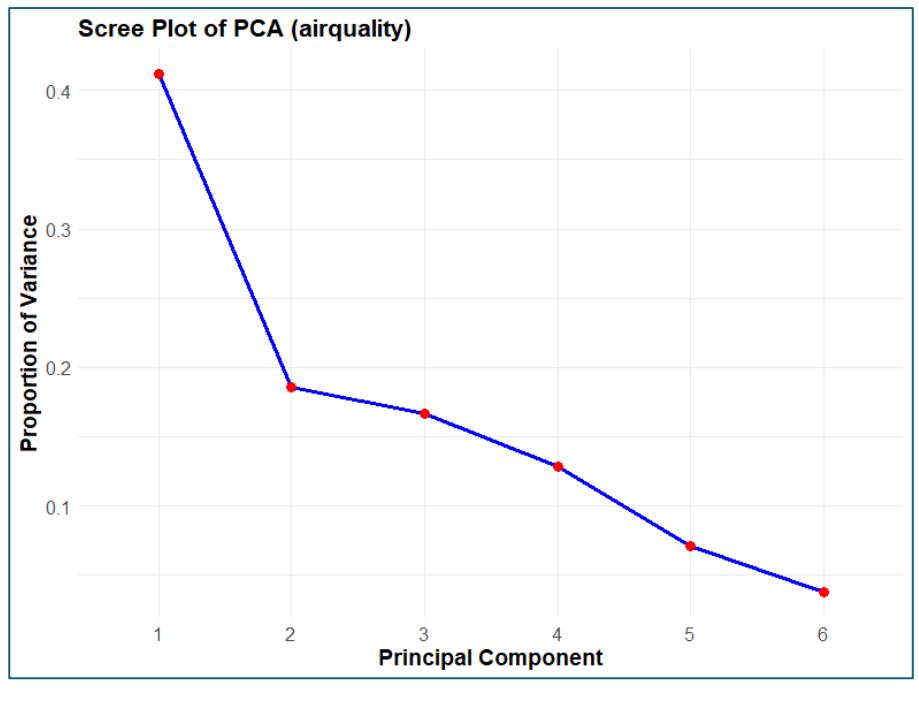
- **Residual Standard Error**: 21.85
- **Multiple R-squared**: 0.5687 — The model explains 56.87% of the variance in Ozone.
- **Adjusted R-squared**: 0.5611
- **F-statistic**: 74.5 (p-value < 2.2e-16) — Overall model is highly significant.

Conclusion: Temp and Wind are significant predictors of Ozone, with Temp positively and Wind negatively impacting Ozone levels. The model explains about 57% of the variance.



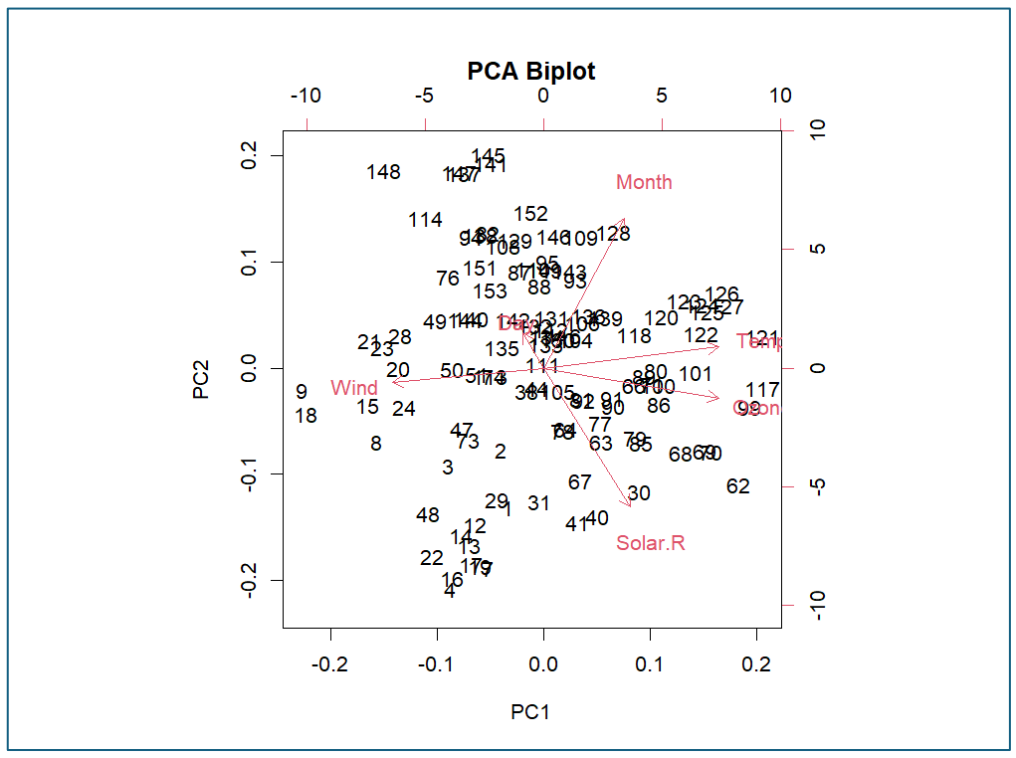
8. Residual Analysis for the Iris Dataset:
The residual plots show some heteroscedasticity and potential outliers in the regression model. The assumption of constant variance might be violated, indicating that further model refinement may be needed.

.



9. Principal Component Analysis (PCA)
PCA revealed that the first two principal components explain a significant proportion of the variance in the data.
The **Scree Plot** (line graph) suggests that the first two components are sufficient to capture most of the variation in the dataset.

```
> summary(pca_airquality)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.5713 1.0550 0.9992 0.8765 0.65169 0.47612
Proportion of Variance 0.4115 0.1855 0.1664 0.1280 0.07078 0.03778
Cumulative Proportion  0.4115 0.5970 0.7634 0.8914 0.96222 1.00000
```



The **biplot** for the first two principal components indicates that temperature and Ozone concentrations have a strong load on the first principal component, ozone has negative dependence od PC1. while month and Solar.R are more strongly associated with the second component.

Conclusion

•**Univariate Analysis**: The ozone data is right-skewed with some extreme values. The mean and median are quite different, indicating the presence of outliers.
•**Multivariate Analysis**: There is a positive correlation between ozone and solar radiation. The multiple regression model identifies solar radiation and wind speed as significant predictors of ozone concentration. Wind speed is negatively correlated with ozone levels, while solar radiation has a positive effect.
•**PCA**: PCA suggests that the first two components explain most of the variance, and the loadings indicate that ozone, solar radiation, and temperature are the most influential variables in the dataset.
This analysis provides a comprehensive understanding of the relationships between the air quality variables, with useful insights into factors affecting ozone concentrations.