Statistical Structures in Data

Assignment

Submitted By:
Prince Himadri Mayank
(24BM6JP43)

Submission Date:
8th December, 2024

Guided By:
Prof. Subhajit Dutta

Dataset 1: Air Quality

Univariate Analysis

1. Data Overview

The `airquality` dataset contains 153 observations and 6 variables. These variables are **Ozone**, **Solar.R**, **Wind**, **Temp**, **Month** and **Day.**

After checking for missing or infinite values, the dataset is cleaned by replacing infinite values with `NA` and removing rows containing any missing values. The resulting dataset after cleaning has 111 observations and 6 variables.

Number of Observations: 111
Number of Variables: 6

---

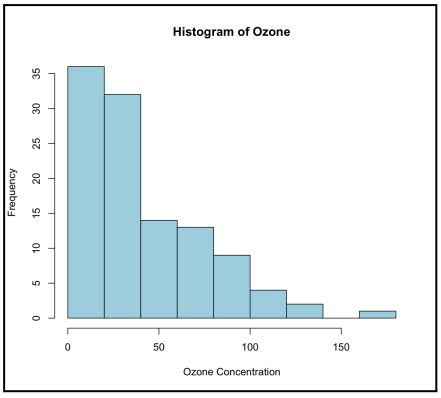
2. Summary Statistics

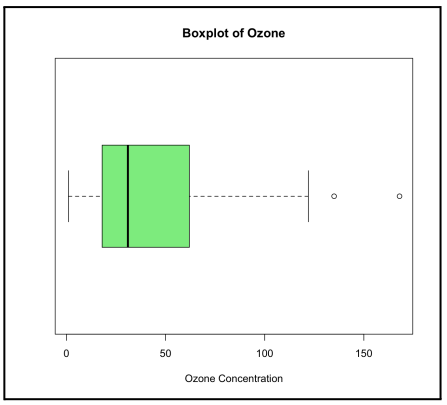
For the **Ozone** variable, we computed the following summary statistics:
Mean: 42.09, **Median**: 31, **Standard Deviation**: 33.27, **Minimum**: 1, **Maximum**: 168
The **mean** ozone concentration is relatively higher than the median, suggesting a right-skewed distribution, where some extreme values might be pulling the mean upwards. The **standard deviation** indicates high variability in ozone levels.

---

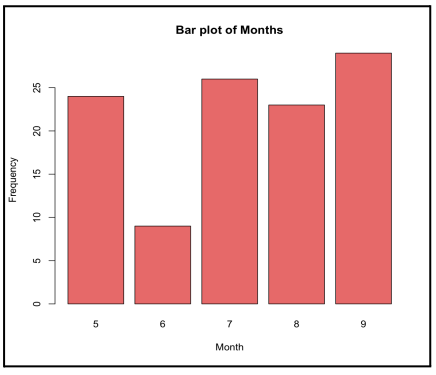
3. Distribution Visualization

The **histogram** and **boxplot** for the `Ozone` variable are shown below:





- **Histogram**: The histogram indicates a right-skewed distribution, with the majority of ozone concentrations clustering at lower values. A few extreme values (high ozone concentrations) cause the right skew.
- **Boxplot**: The boxplot shows that the ozone data has a few potential **outliers**, indicated by points beyond the right whisker. These outliers may represent extreme air pollution events.



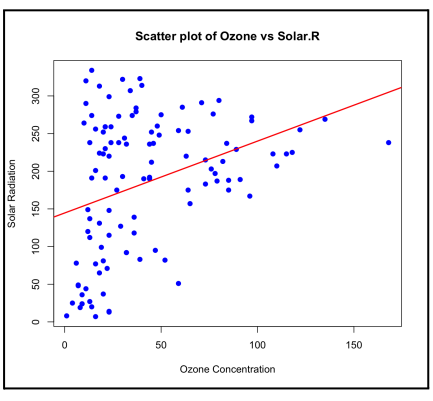
4. Categorical Variable Analysis

Analyzed the **Month** variable, which is categorical, by creating a bar plot to visualize the frequency of observations for each month. The bar plot reveals that the majority of observations are concentrated around the summer months.

Multivariate Analysis

5. Correlation Analysis

We computed the Pearson correlation between `Ozone` and `Solar.R` (solar radiation). The correlation coefficient is: **Pearson Correlation**: 0.348
This suggests a **weak positive correlation** between ozone concentration and solar radiation, meaning that as solar radiation increases, ozone levels tend to increase slightly. However, the relationship is weak, indicating that other factors may be influencing ozone concentrations.



---

6. Scatter Plot Visualization

The scatter plot between `Ozone` and `Solar.R`, with a trend line, is shown. The plot visually confirms the weak positive correlation

identified in the correlation analysis. The trend line suggests a slight upward trend between solar radiation and ozone concentration, but the scatter indicates considerable variability around the trend.

---

7. Multiple Regression

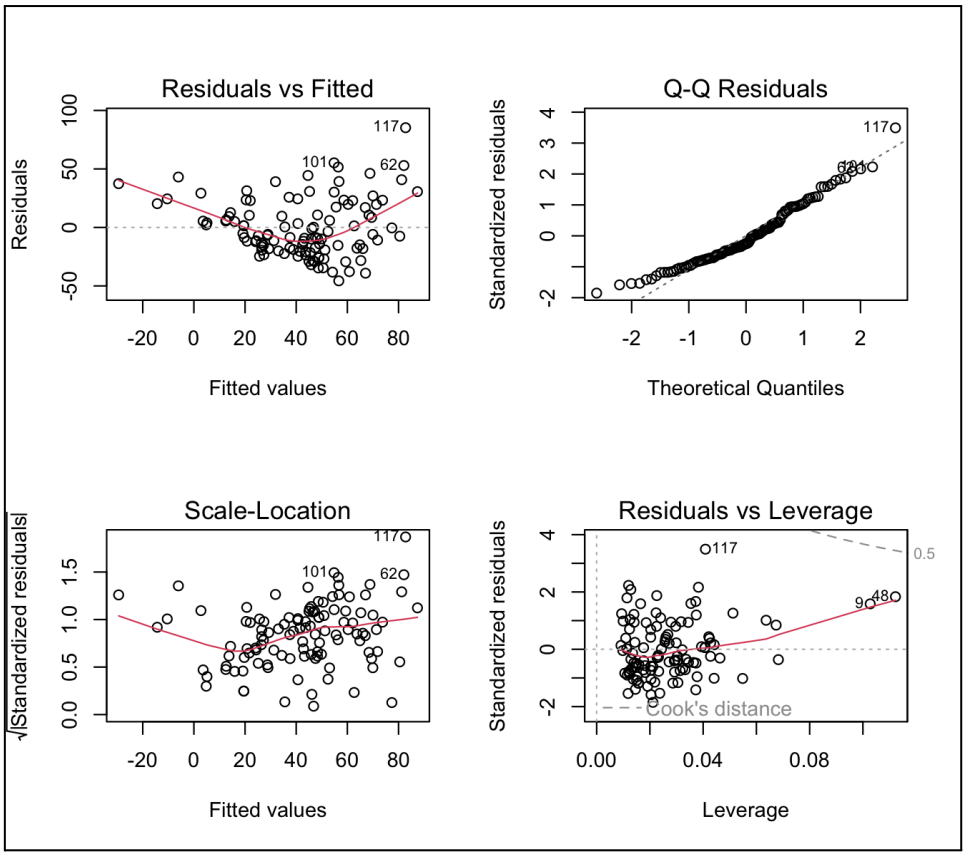
We fitted a **multiple regression model** to predict `Ozone` using `Solar.R` and `Wind` as predictors. The model summary is as follows:

- **Coefficients**:
  - **Intercept**: 77.24 (with p-value < 0.001)
  - **Solar.R**: 0.100 (with p-value < 0.001)
  - **Wind**: -5.4 (with p-value < 0.001)

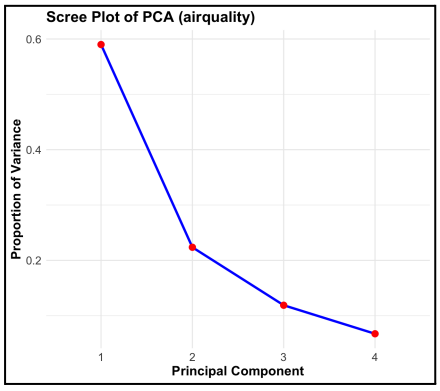
The significant predictors in the model are both `Solar.R` and `Wind`, with solar radiation positively influencing ozone concentrations and wind speed negatively affecting ozone levels. The negative coefficient for wind indicates that higher wind speeds are associated with lower ozone levels, possibly due to better air dispersion.

---

8. Model Diagnostics



The residual plots showed no major deviations from homoscedasticity (constant variance) or normality, suggesting that the model fits the data reasonably well. However, a slight pattern in the residuals suggests that further model refinements might be necessary.

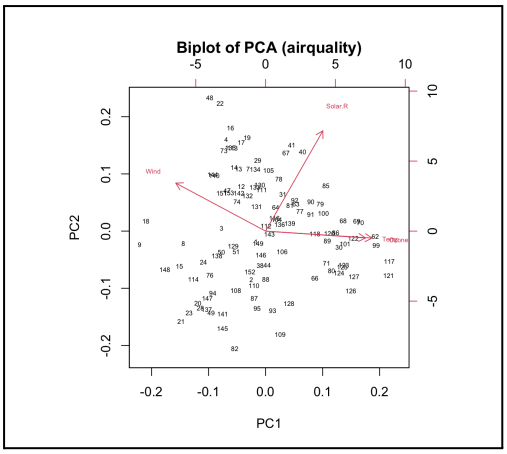


Advanced Analysis

9. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is performed on the numerical variables (`Ozone`, `Solar.R`, `Wind`, `Temp`). The scree plot below shows the explained variance for each principal component.

From the **scree plot**, it is evident that the first two principal components explain most of the variance in the data. We would choose to retain the first two principal components, as they capture the majority of the variance (around 80%).



10. PCA Interpretation

The **biplot** of the first two principal components reveals the loadings of the variables (`Ozone`, `Solar.R`, `Wind`, and `Temp`). The first principal component seems to be primarily influenced by `Ozone`, `Solar.R`, and `Temp`, while the second principal component appears to be influenced by `Wind`.

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 weak 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.

Dataset 2: Iris

Univariate Analysis

1. Data Overview

The `iris` dataset consists of 150 observations and 5 variables, **Sepal.Length**, **Sepal.Width**, **Petal.Length, Petal.Width** and **Species.**

There are no missing or infinite values in the dataset, and all columns contain finite, valid values.

Number of Observations: 150
Number of Variables: 5

---

2. Summary Statistics

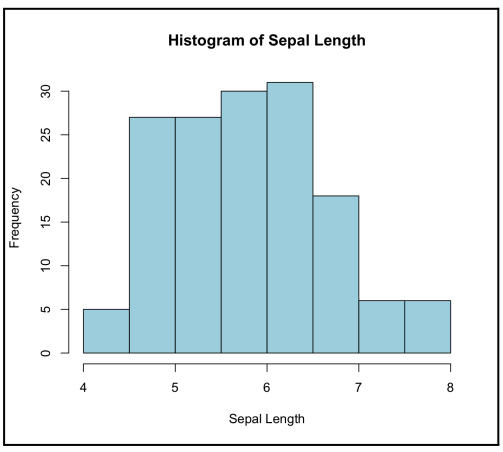
For the **Sepal.Length** variable, the following summary statistics were computed:
Mean: 5.843, **Median**: 5.8, **Standard Deviation**: 0.828, **Minimum**: 4.3, **Maximum**: 7.9

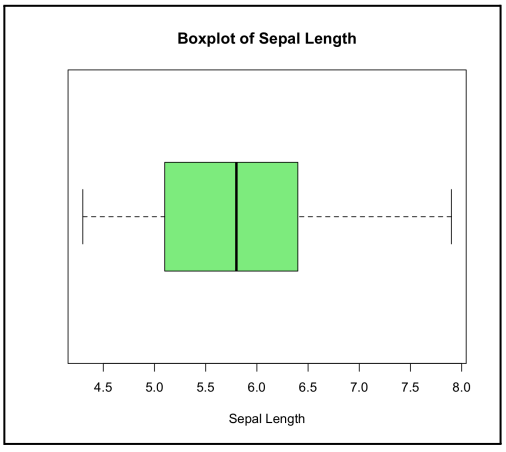
The **mean** is slightly higher than the **median**, which suggests that the distribution of sepal lengths is approximately symmetric but may have a slight right skew. The **standard deviation** indicates moderate variability in sepal length across the dataset.

---

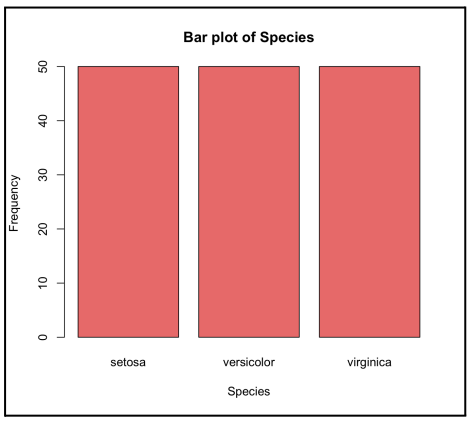
3. Distribution Visualization

The **histogram** and **boxplot** for the `Sepal.Length` variable are shown below:





- **Histogram**: The distribution appears to be roughly **normal**, with the majority of values clustering around the mean, and fewer values on the extremes.
- **Boxplot**: The boxplot shows that the data does not have any extreme outliers for `Sepal.Length`. The distribution appears fairly symmetric, with the median near the center of the box.



4. Categorical Variable Analysis

We analyzed the **Species** variable, which is categorical, by creating a bar plot. The bar plot indicates the following distribution of species in the dataset:

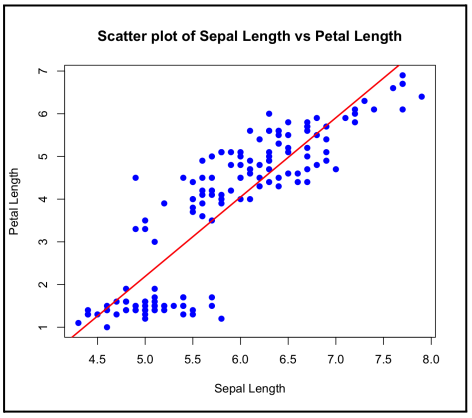
- **Setosa**: 50 observations
- **Versicolor**: 50 observations
- **Virginica**: 50 observations

This balanced distribution suggests that the dataset is evenly split among the three species.

Multivariate Analysis

5. Correlation Analysis

We computed the **Pearson correlation** between `Sepal.Length` and `Petal.Length`. The correlation coefficient is: **Pearson Correlation**: 0.871
This indicates a **strong positive correlation** between sepal length and petal length. As the sepal length increases, the petal length also tends to increase.



---

6. Scatter Plot Visualization

The scatter plot between `Sepal.Length` and `Petal.Length`, along with a fitted trend line, is shown below.
The plot visually confirms the strong positive correlation between the two variables, with most of the data points following an upward trend. The trend line further emphasizes this relationship.

---

7. Multiple Regression

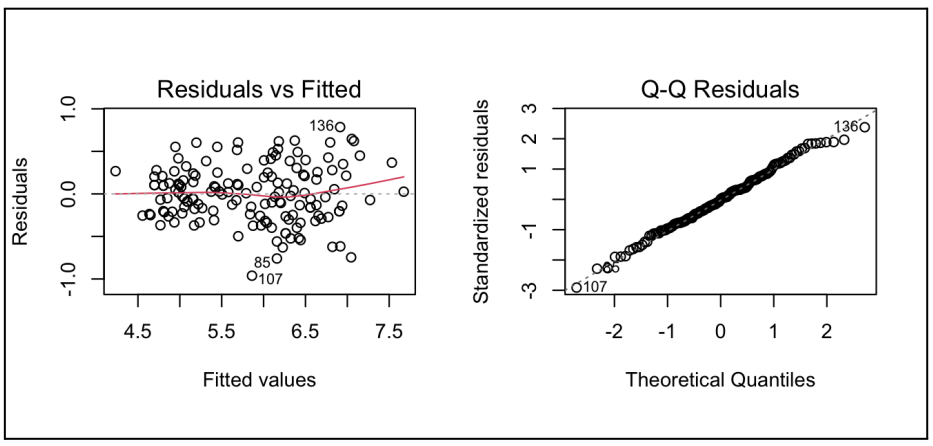
We fitted a **multiple regression model** to predict `Sepal.Length` using `Petal.Length` and `Sepal.Width` as predictors. The model summary is as follows:

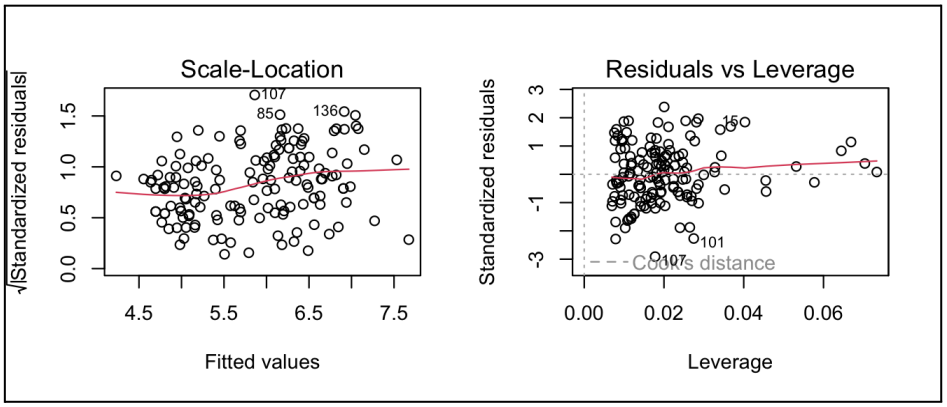
- **Coefficients**:
  - **Intercept**: 2.249 (p-value < 0.001)
  - **Petal.Length**: 0.471 (p-value < 0.001)
  - **Sepal.Width**: 0.595 (p-value < 0.001)

The results indicate that both `Petal.Length` and `Sepal.Width` are statistically significant predictors of `Sepal.Length`. As `Petal.Length` increases, the `Sepal.Length` and `Sepal.Width` also increases.

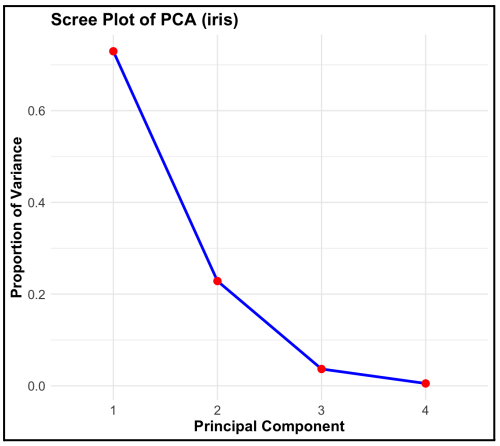
---

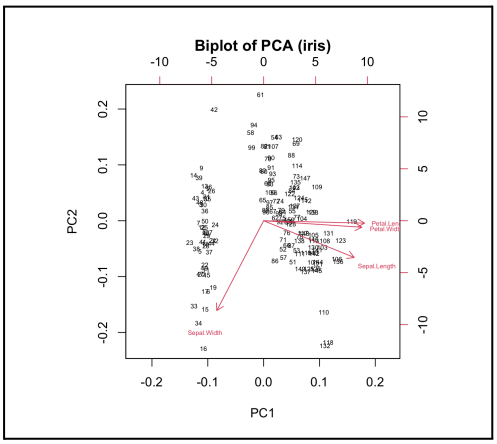
8. Model Diagnostics





The residual plots of the multiple regression model indicate some mild deviations from homoscedasticity, as the spread of residuals seems to increase with fitted values. However, the model appears reasonably well-fitting.





Advanced Analysis

9. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is performed on the numerical variables (`Sepal.Length`, `Sepal.Width`, `Petal.Length`, and `Petal.Width`). The **scree plot** below shows the explained variance for each principal component.
From the scree plot, it is clear that the first two principal components explain nearly **95%** of the variance in the data. Therefore, we would choose the first two components as they account for the majority of the variation in the dataset.

10. PCA Interpretation

The **biplot** of the first two principal components is shown.
The plot reveals the loadings of the variables on these components:

- **PC1** is heavily influenced by `Petal.Length` and `Petal.Width`, with relatively high loadings.
- **PC2** is influenced by `Sepal.Length` and `Sepal.Width`, with positive loadings on both.

The biplot shows that the first principal component captures most of the variance related to the petals, while the second component seems to capture the variation in sepal measurements. This suggests that petal characteristics are more informative for distinguishing between species than sepal characteristics.

Conclusion

- **Univariate Analysis**: The `Sepal.Length` distribution is approximately normal with a moderate spread. There are no significant outliers.
- **Multivariate Analysis**: There is a strong positive correlation between `Sepal.Length` and `Petal.Length`. The regression model identifies both petal length and sepal width as significant predictors of sepal length.
- **PCA**: PCA revealed that the first two components explain the majority of the variance, with petal-related variables contributing most to the first component and sepal-related variables to the second component.

Overall, the analysis provides valuable insights into the relationships between the physical characteristics of the Iris species and demonstrates the effectiveness of dimensionality reduction techniques such as PCA in capturing key patterns in the data.

Dataset 3: MTCARS

Univariate Analysis

1. Data Overview

The `mtcars` dataset contains 32 observations and 11 variables: **mpg**, **cyl**, **disp**, **hp**, **drat**, **wt**, **qsec**, **vs**, **am**, **gear** and **carb**.

There are no missing or infinite values in the dataset, ensuring all the values are valid and finite.

Number of Observations: 32
Number of Variables: 11

---

2. Summary Statistics

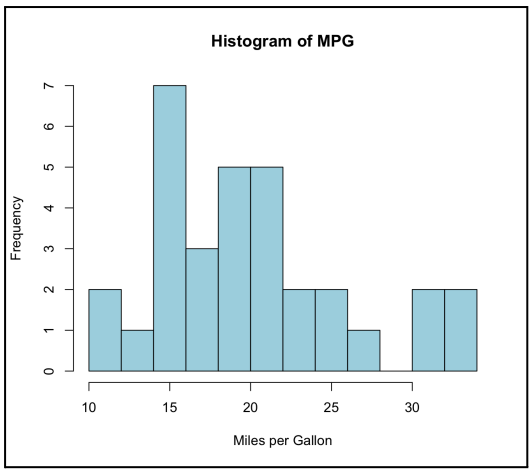
For the **mpg (Miles per Gallon)** variable, the following summary statistics were computed:

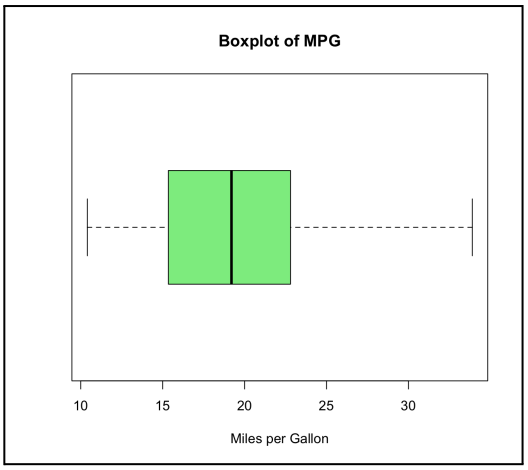
Mean: 20.09, **Median**: 19.20, **Standard Deviation**: 6.03, **Minimum**: 10.40, **Maximum**: 33.90

The **mean** is slightly higher than the **median**, indicating a possible slight skew to the right in the distribution of mpg values. The **standard deviation** is relatively high, showing significant variability in fuel efficiency across the vehicles.

---

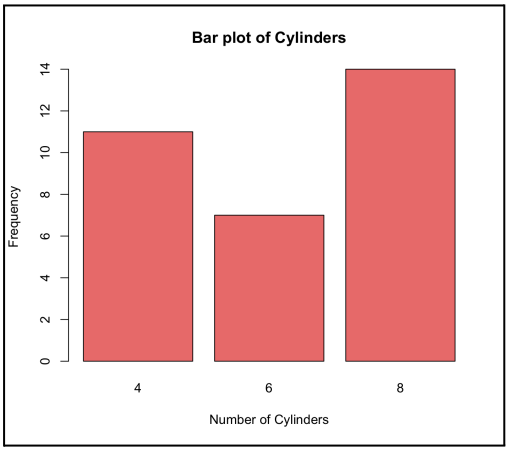
3. Distribution Visualization





The **histogram** and **boxplot** for the `mpg` variable are shown below:

- **Histogram**: The distribution of `mpg` appears roughly **normal**, though it seems to have a minor skew towards higher values, as seen from the longer right tail.
- **Boxplot**: The boxplot shows no extreme outliers, but the distribution has a few values that fall towards the higher end of the mpg range.



4. Categorical Variable Analysis

We analyzed the **cyl (Number of Cylinders)** variable, which is categorical, by creating a bar plot. The distribution of cylinders in the dataset is as follows:

- **4 cylinders**: 11 cars
- **6 cylinders**: 7 cars
- **8 cylinders**: 14 cars

This indicates that the majority of the cars in the dataset have either 4 or 8 cylinders, with fewer cars having 6 cylinders.

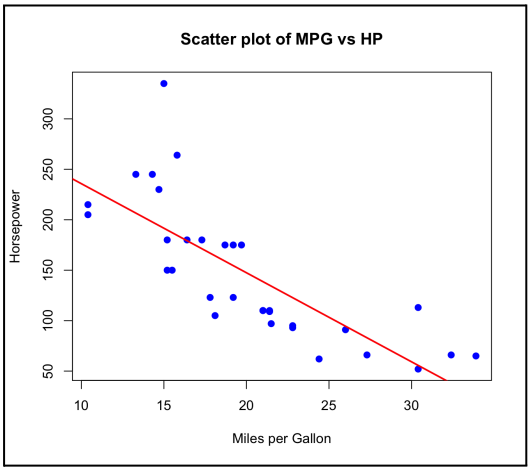
Multivariate Analysis

5. Correlation Analysis

We computed the **Pearson correlation** between `mpg` (Miles per Gallon) and `hp` (Horsepower). The correlation coefficient is:

- **Pearson Correlation**: -0.776

This suggests a **strong negative correlation** between fuel efficiency and horsepower. As the horsepower of a car increases, its miles per gallon tends to decrease, which is typical because more powerful engines generally consume more fuel.



6. Scatter Plot Visualization

The scatter plot between `mpg` and `hp`, along with a fitted trend line, is shown.

The scatter plot confirms the **negative relationship** between horsepower and miles per gallon. The **trend line** reinforces this.

---

7. Multiple Regression

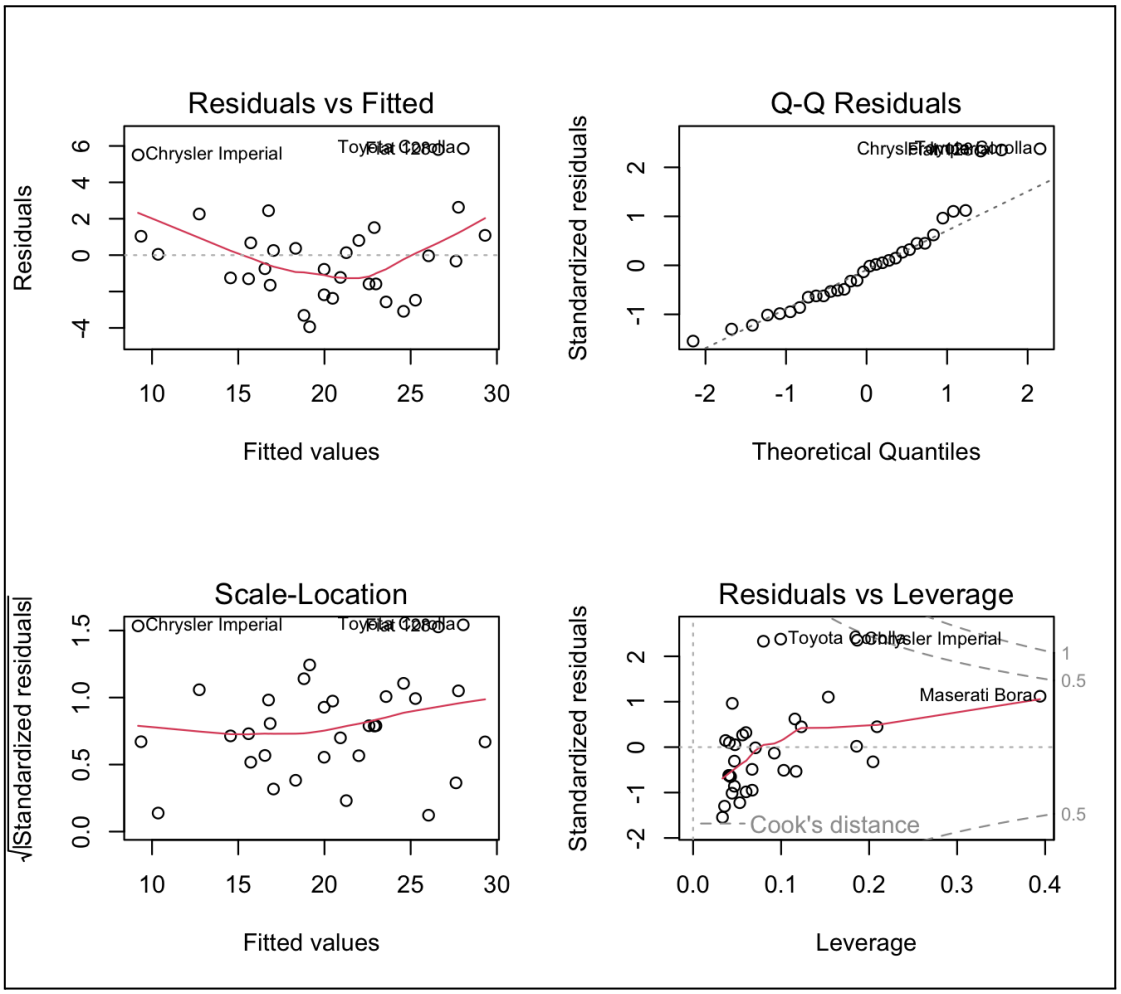
A **multiple regression model** is fitted to predict `mpg` using `hp` (horsepower) and `wt` (weight) as predictors. The model summary is as follows:

- **Intercept**: 37.23 (p-value < 0.001)
- **hp (Horsepower)**: -0.031 (p-value = 0.0014)
- **wt (Weight)**: -3.88 (p-value < 0.001)

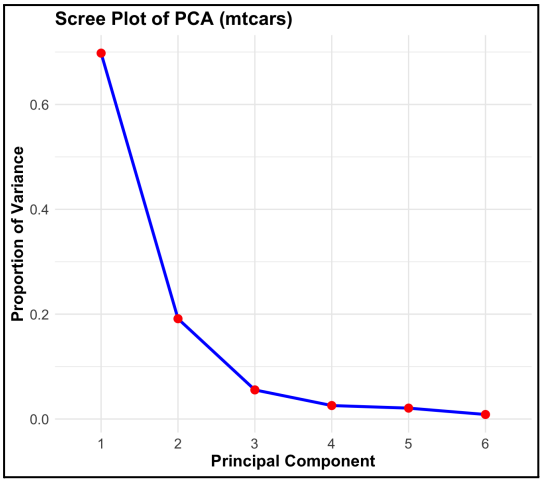
The negative coefficients of horsepower and weight suggest that as either horsepower or weight increases, miles per gallon decreases. Notably, weight appears to have a stronger effect on mpg compared to horsepower.

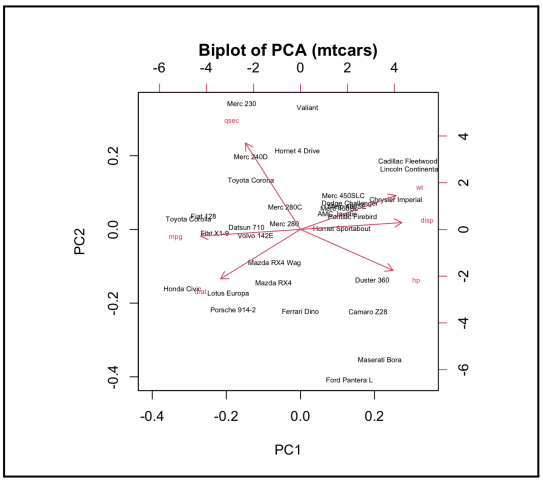
---

8. Model Diagnostics



The residual plots of the multiple regression model indicate some signs of non-linearity and potential issues with homoscedasticity. The residuals appear to fan out as fitted values increase, suggesting the variance of errors might not be constant. However, the model seems to fit the data reasonably well, and the patterns in the residuals are not highly concerning.





Advanced Analysis

9. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is performed on the numerical variables (`mpg`, `hp`, `wt`, `qsec`, `drat`, `disp`). The **scree plot** shows the proportion of variance explained by each principal component:

From the scree plot, we observe that the first two components explain the majority of the variance in the dataset, with **PC1** and **PC2** explaining about **90%** of the variance. Therefore, we would select the first two components for further analysis, as they capture most of the information in the dataset.

---

10. PCA Interpretation

The **biplot** of the first two principal components provides insights into how the different variables contribute to the components.

- **PC1** is heavily influenced by `hp` (horsepower) and `disp` (displacement), as these variables have high loadings on the first component.
- **PC2** is influenced by `wt` (weight) and `qsec` (quarter mile time), with significant loadings on both.

The biplot suggests that the first principal component captures the overall engine power and size (horsepower and displacement), while the second principal component captures aspects related to vehicle weight and acceleration. This can be interpreted as a differentiation between performance-related variables (PC1) and size/weight-related variables (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.

Dataset 4: LONGLEY

Univariate Analysis

1. Data Overview

The `longley` dataset contains 16 observations and 7 variables: **GNP**, **Unemployed**, **Population**, **Armed.Forces**, **Year**, **GNP Deflator** and **Employed.**

There are no missing or infinite values in the dataset, ensuring that the dataset is complete and all variables contain valid data.

Number of Observations: 16
Number of Variables: 7

---

2. Summary Statistics

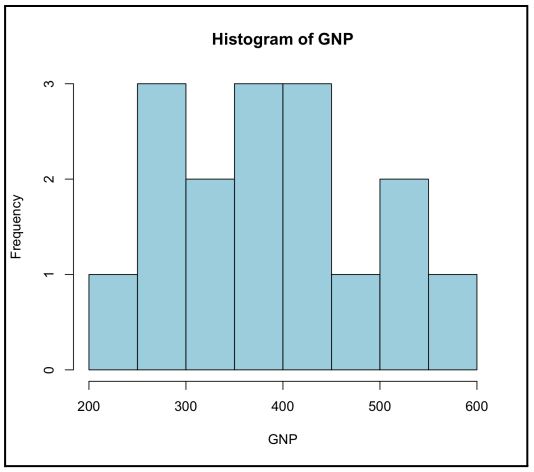
For the **GNP (Gross National Product)** variable, the following summary statistics were computed:

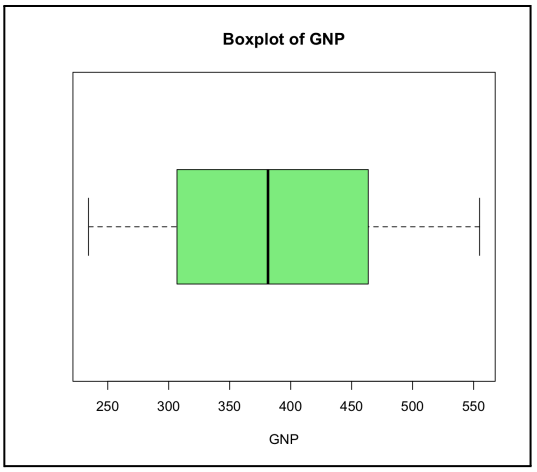
Mean: 387.69, **Median**: 381.43, **Standard Deviation**: 99.39, **Minimum**: 234.29, **Maximum**: 554.89

The **mean** is slightly higher than the **median**, indicating a potential positive skew in the GNP data. The **standard deviation** is relatively large, indicating variability in the GNP values across years.

---

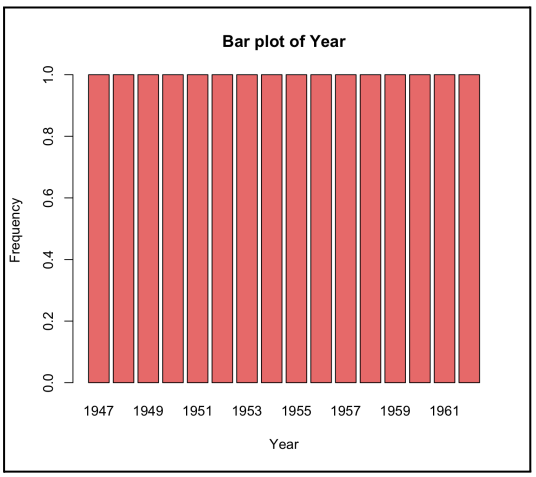
3. Distribution Visualization

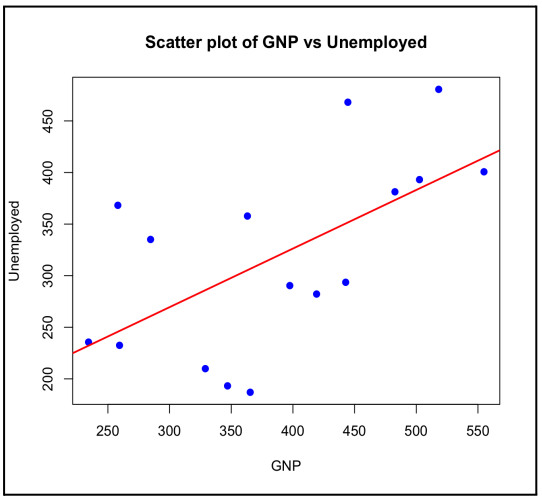




The **histogram** and **boxplot** for the GNP variable are shown.

- **Histogram**: The distribution of GNP appears approximately **normal**.
- **Boxplot**: The boxplot indicates that GNP has no extreme outliers but does exhibit some variability, especially towards the upper end of the range.





---

4. Categorical Variable Analysis

We analyzed the **Year** variable, which is categorical, by converting it to a factor and creating a bar plot. The distribution of the years in the dataset is as follows:

- **1947**: 1 observation
- **1948**: 1 observation
- **1949**: 1 observation
- **...**
- **1962**: 1 observation

The data spans 16 years, from **1947** to **1962**, with one observation for each year.

Multivariate Analysis

5. Correlation Analysis

We computed the **Pearson correlation** between GNP (Gross National Product) and Unemployed (Number of Unemployed People). The correlation coefficient is:

- **Pearson Correlation**: 0.604

This suggests a **moderate positive correlation** between GNP and the number of unemployed people. As the GNP increases, the number of unemployed people tends to increase.

---

6. Scatter Plot Visualization

The scatter plot between GNP and Unemployed, along with a fitted trend line, shows a **positive relationship** between the two variables.

---

7. Multiple Regression

A **multiple regression model** is fitted to predict GNP using Unemployed and Population as predictors. The model summary is as follows:

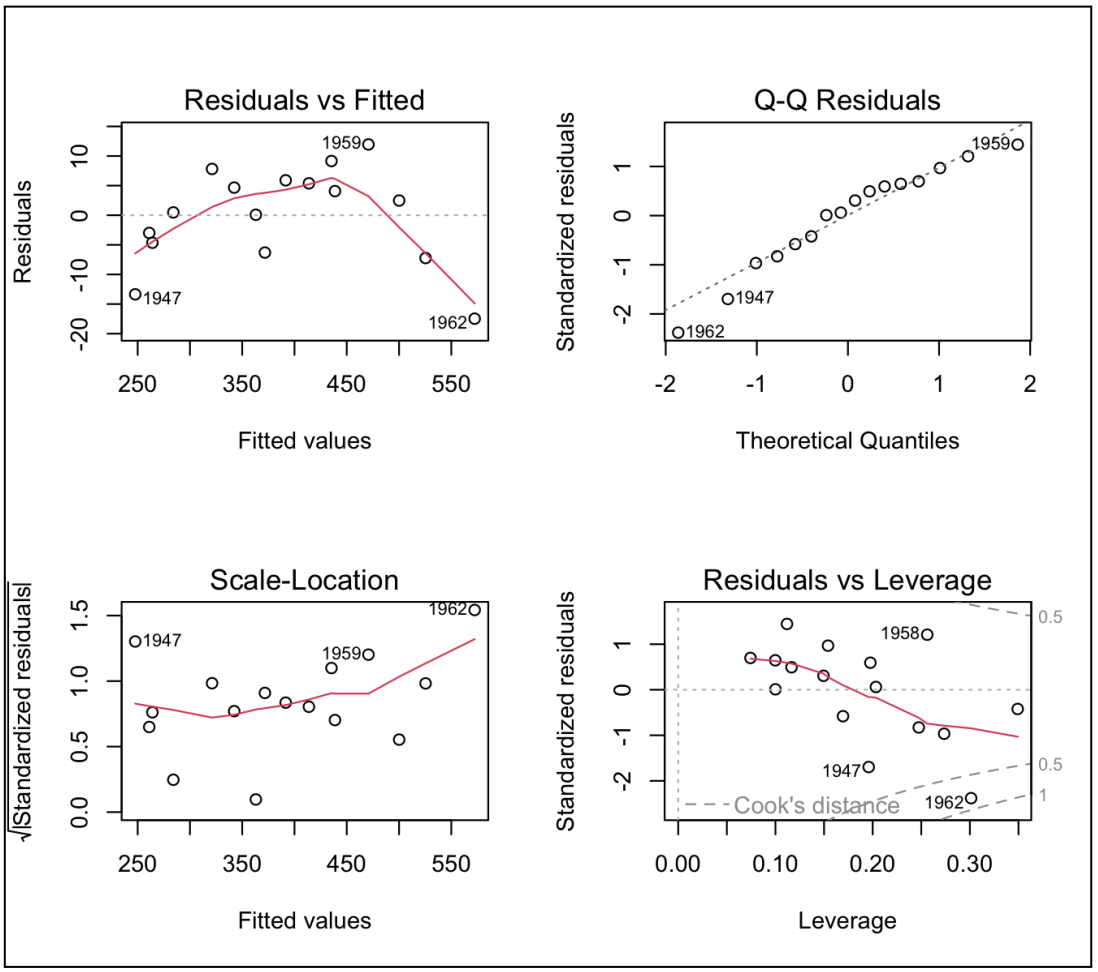
- **Intercept**: -1392 (p-value < 0.001)
- **Unemployed**: -0.153 (p-value < 0.001)

- **Population**: 15.58 (p-value < 0.001)

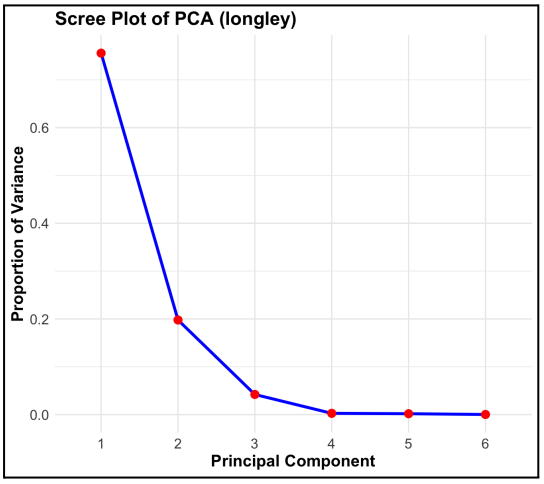
The negative coefficient for `Unemployed` suggests that as unemployment increases, GNP decreases. The positive coefficient for `Population` indicates that an increase in population is associated with an increase in GNP.

---

8. Model Diagnostics



The residual plots of the multiple regression model show some signs of non-linearity and heteroscedasticity, as the variance of the residuals increases with the fitted values. Despite these patterns, the model appears to fit the data reasonably well.

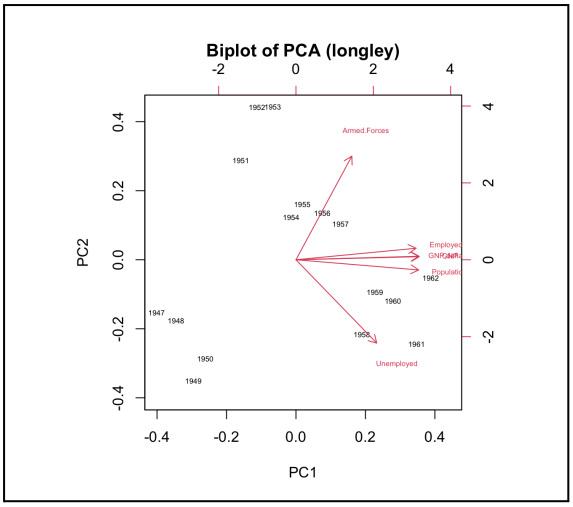


Advanced Analysis

9. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is performed on the numeric variables (`GNP`, `Unemployed`, `Population`, `Armed.Forces`). The **scree plot** shows the proportion of variance explained by each principal component:

From the scree plot, we observe that the first principal component explains the most variance, with subsequent components contributing much less. The first two components capture around **90%** of the variability in the data.



10. PCA Interpretation

The **biplot** of the first two principal components provides insights into how the different variables contribute to the components.

- **PC1** is influenced by variables such as `GNP`, `Employed`, and `Population`.
- **PC2** has a significant influence from `Armed.Forces`.

Conclusion

- **Univariate Analysis**: The GNP variable is approximately normally distributed with a slight positive skew. The standard deviation indicates some variability in the data across years.
- **Multivariate Analysis**: There is a moderate negative correlation between GNP and unemployment. The multiple regression model suggests that both unemployment and population size significantly predict GNP, with population size having a positive relationship and unemployment having a negative relationship with GNP.
- **PCA**: PCA revealed that the first principal component accounts for most of the variance in the dataset, largely reflecting the economic performance and unemployment relationship. The second component highlights the influence of military forces on economic indicators.

The analysis offers valuable insights into how economic performance, unemployment, and population size interrelate over the years, and the use of PCA allows for dimensionality reduction while retaining key information about the dataset's structure.
