**INDIAN STATISTICAL INSTITUTE, KOLKATA**

****

**Statistical Structures in Data**

**Prof. Subhajit Dutta**

**Numerical Assignment Report:**

**Exploration of Univariate and Multivariate Data Using R**

**By**

**Sushrut Joshi**

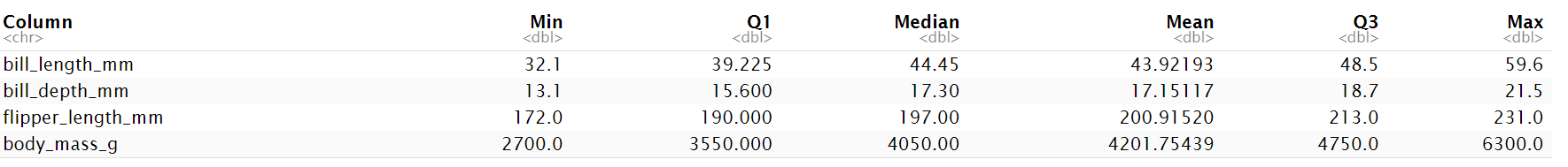
**24BM6JP55**

**1) About the data set:**

PENGUINS Data-set

The penguin dataset comprises 344 entries across three species: Adelie, Gentoo, and Chinstrap, distributed over three islands: Biscoe, Dream, and Torgersen. It includes measurements such as bill length, bill depth, flipper length, and body mass, with some missing values in these fields and the sex category.

**2) Summary Statistics:**

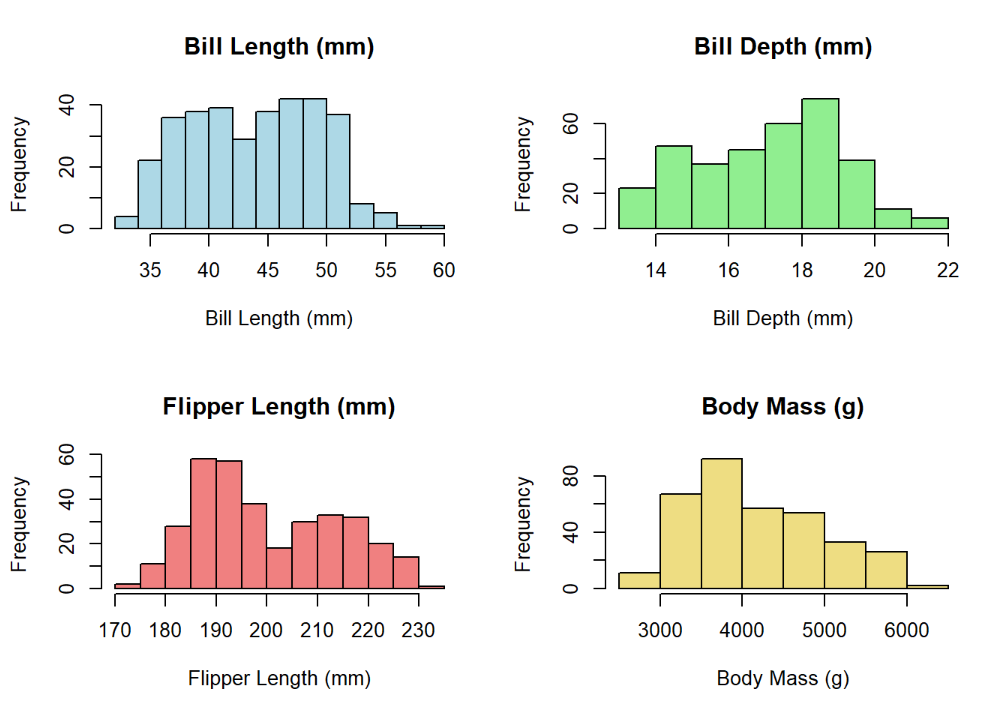
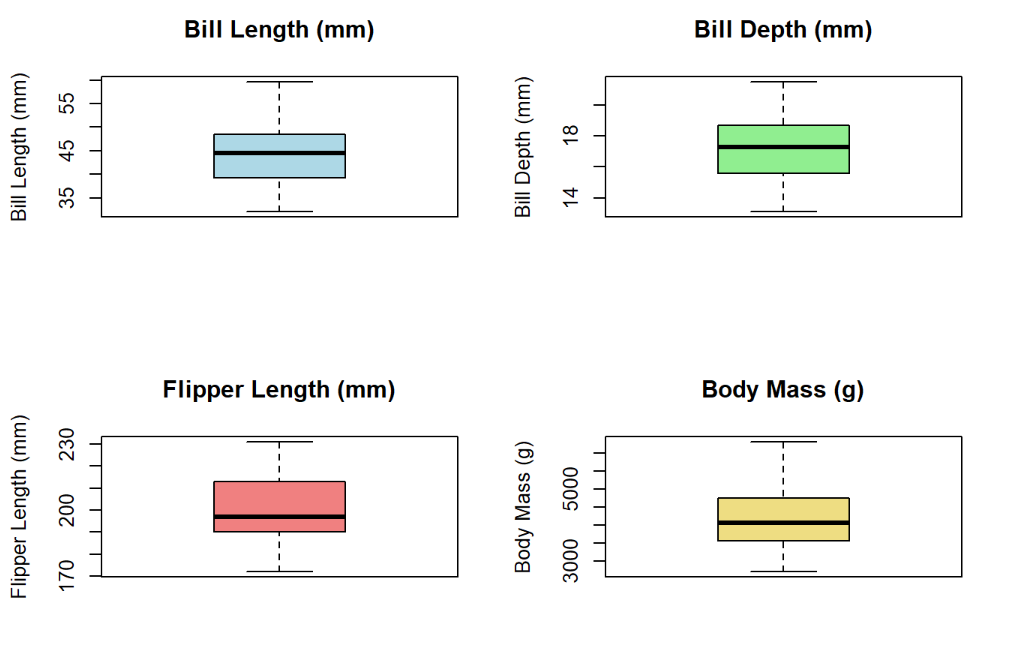


**Measurement Range:** Body mass shows the largest range, indicating high variability, while bill length and depth have smaller ranges.

**Mean vs Median:** Body mass has a noticeable difference between mean and median, suggesting skewness, unlike the other measurements which are more symmetrical.

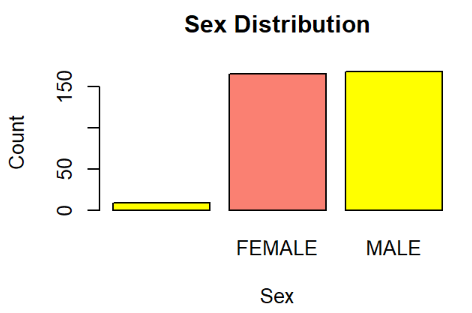
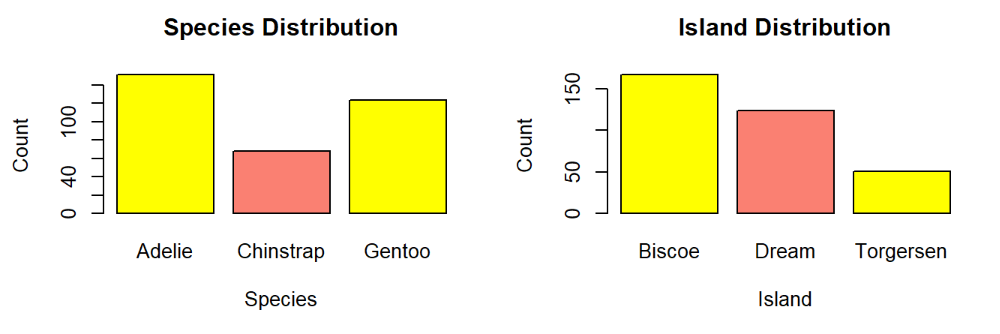
**Distribution:** Body mass data is more spread out compared to other categories, which display more consistent distributions.

**3) Distribution Visualization:**

** **

**Spread**: Bill length and body mass show moderate spread; flipper length is uniform.  
**Potential** **Outliers**: Bill depth has potential outliers beyond 14-22 mm; body mass below 3000 g or above 5000 g.  
**Distribution**: Bill length and body mass are slightly skewed; others are more uniform.

**4) Categorical Variable Analysis:**

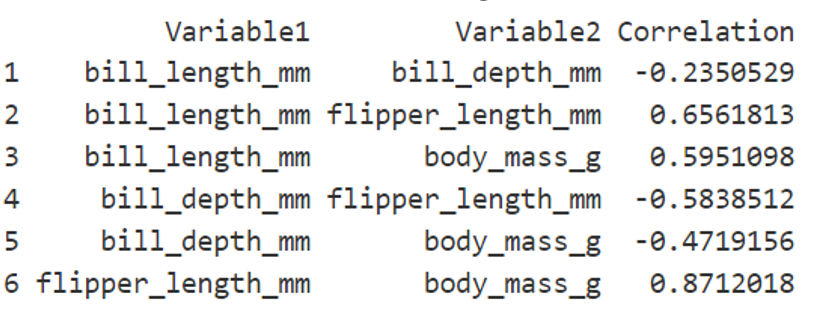
****

There are 3 categories for penguin species: Adelie, Chinstrap and Gentoo. Adelie are more common penguins followed by Gentoo. Whereas Chinstrap are least in numbers. Penguins are seen in 3 different islands. Biscoe has most number of penguins, dream has less number of penguin compared to Dream and Torgersen has least number of penguins among all 3 islands. Based on sex, there aren’t much difference between number of female and number of male penguins.

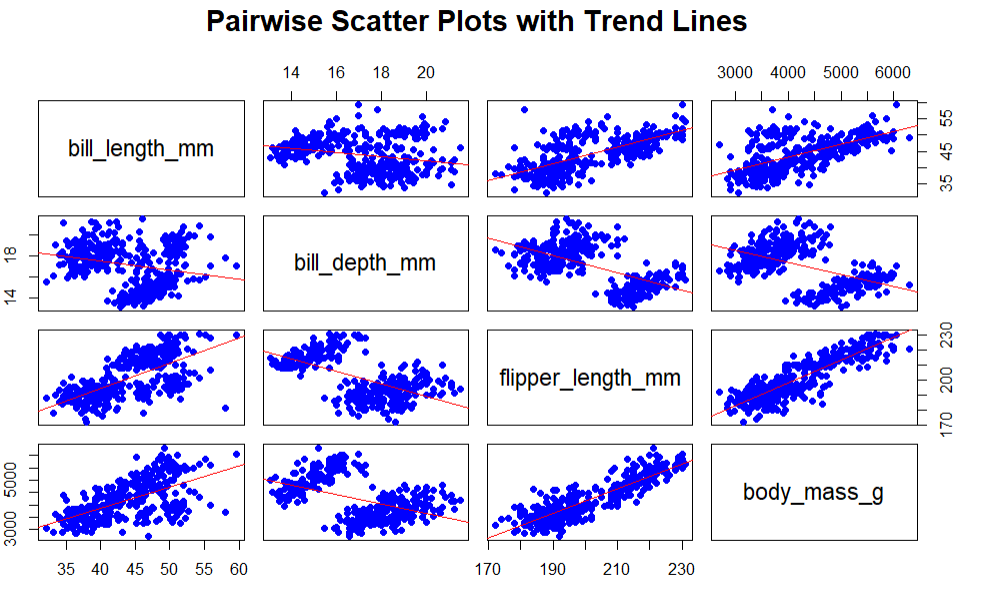
**Multivariate Analysis**

**5) Correlation Analysis**

* As bill\_length\_mm increases, flipper\_length\_mm and body\_mass\_g increases. And as bill\_length\_mm increases bill\_depth\_mm decreases.
* Flipper\_length\_mm has very high positive correlation body\_mass\_g.
* As bill\_depth\_mm increases both flipper\_length\_mm and body\_mass\_g decreases.

****

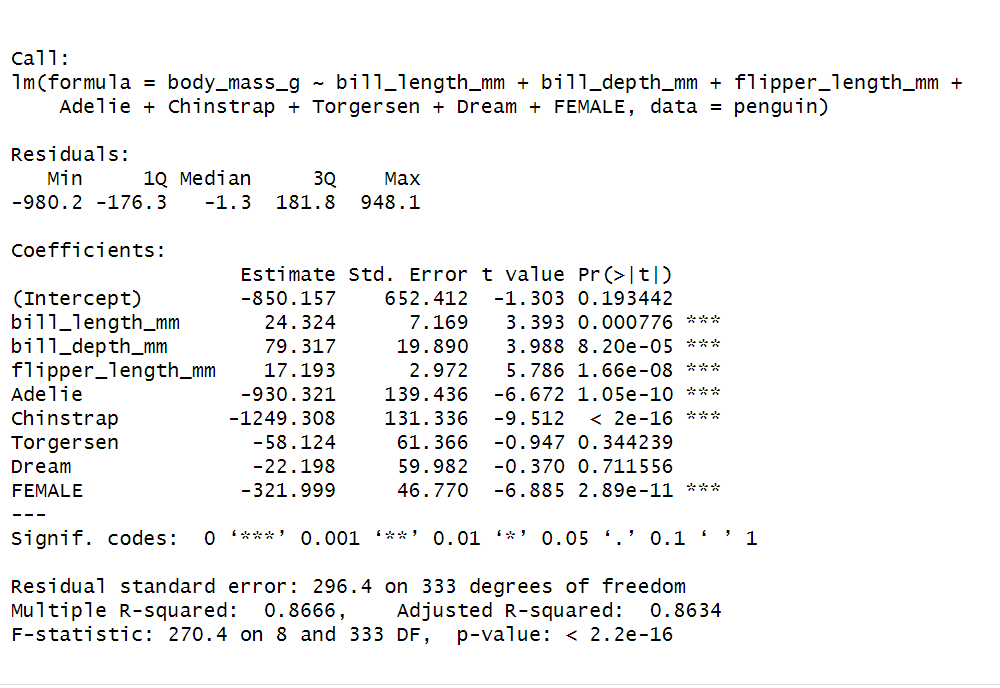
**6) Scatter Plot Visualization**

****

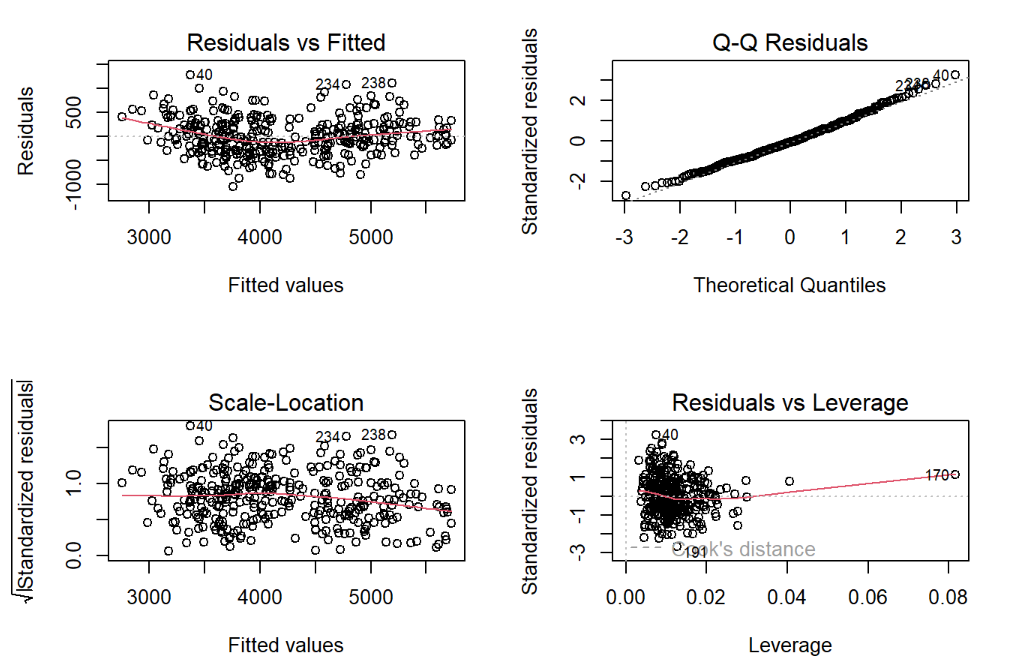
* The scatter plot matrix shows pairwise relationships between four variables: bill length, bill depth, flipper length, and body mass.
* Each plot includes a trend line indicating correlation.
* Bill length and flipper length have a positive relationship with body mass, while bill depth shows weaker correlations with other variables.

**7) Multiple Regression**

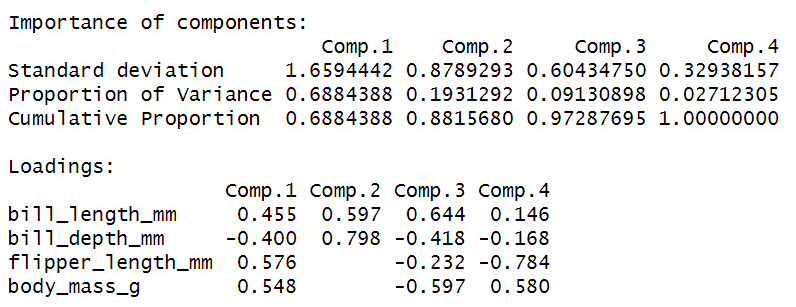
* The linear model predicts *body mass* using *bill length*, *bill depth*, *flipper length*, and penguin species, with a high R-squared of 0.8666.
* Significant predictors include bill length, bill depth, and flipper length (p < 0.001).
* The model's F-statistic is 270.4, indicating overall significance (p < 2.2e-16)
* The coefficient for FEMALE is -321.999, showing a significant impact on body mass (p < 0.001).
* Residuals range from -980.2 to 948.1, with a standard error of 296.4.

****

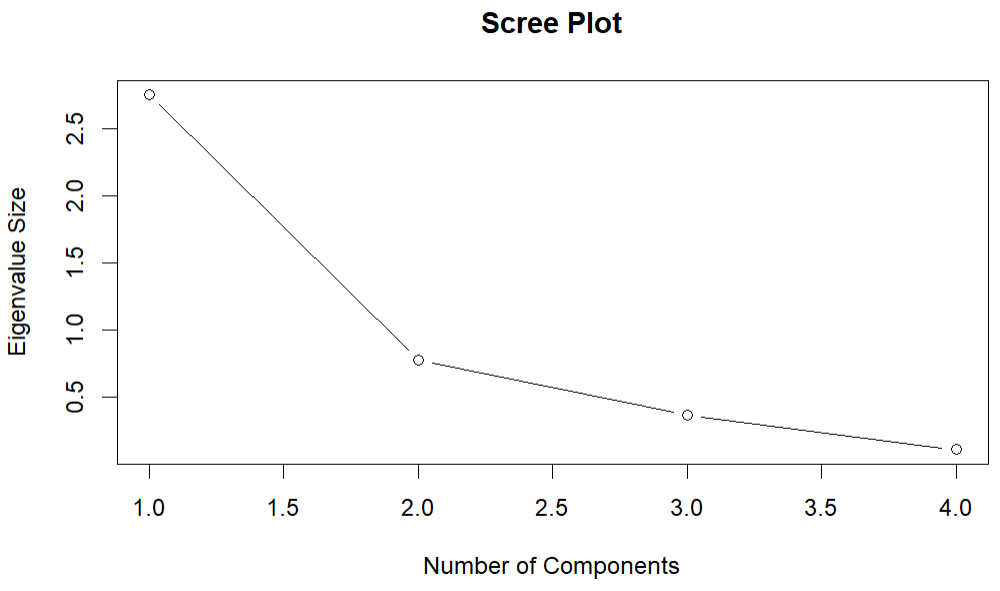
**8) Model Diagnostics**

****

* Diagnostic plots show residuals are approximately normally distributed, though minor heteroscedasticity is observed, as indicated by a slight spread in residual patterns.
* Overall, the model fits well, accurately capturing the relationships between predictors and body mass, but the slight deviation from assumptions.

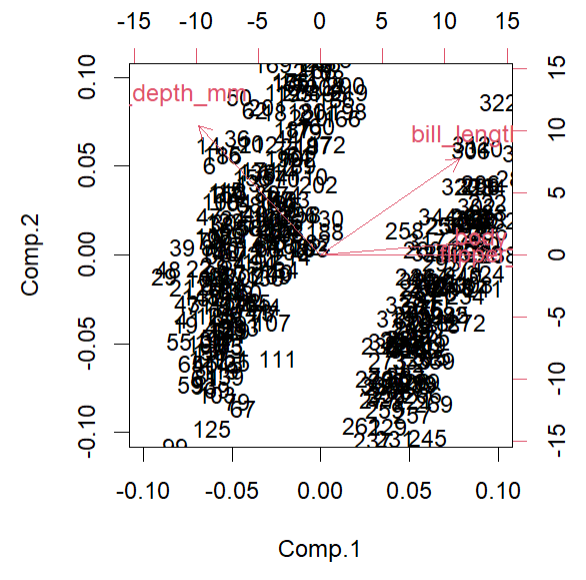
**9) PCA  
**

* The PCA highlights the dominant patterns in the penguin dataset while reducing its complexity.
* Two principal components explain the majority of the variance, with the first capturing size-related features like body mass, bill length, and flipper length, and the second emphasizing bill depth.
* The scree plot confirms the appropriateness of focusing on the first two components, as additional components contribute marginally to the variance.
* Choosing of 2 PCs: The scree plot shows a sharp decline after the second component, indicating that these two capture most variance. Subsequent components contribute minimally, justifying the choice of two PCs.

****

**10) PCA Interpretation:**

* Clustering of Larger Penguins: Heavier and larger penguins tend to cluster together in the biplot, with flipper length closely correlated to body mass, showing a strong relationship between these traits.
* Divergence of Bill Depth: Bill depth shows a somewhat different influence compared to other traits, indicating its distinct role in the morphological variation of penguins.
* Species Separation: The PCA biplot suggests that species differences are reflected in clusters, highlighting variations in body size and shape, which simplifies understanding the morphological diversity within the penguin dataset.

****

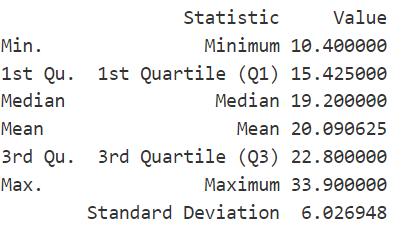
MTCars Data-set

**1) About the data set:**

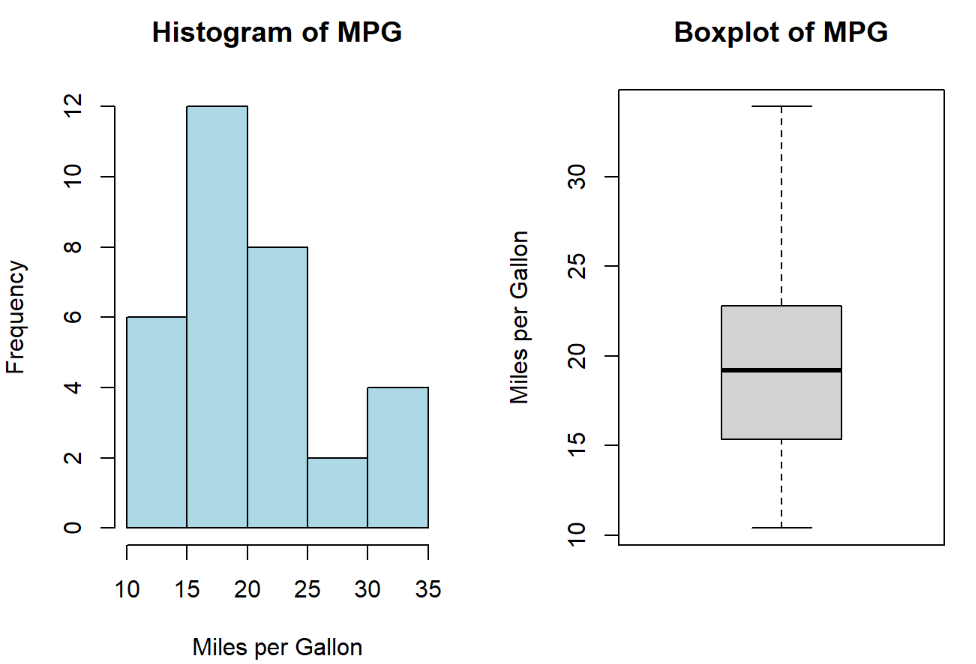
The mtcars dataset contains 32 entries representing various car models with numeric attributes such as miles per gallon (mpg), horsepower (hp), weight (wt), and more. It includes performance, design, and efficiency metrics like cylinder count and transmission type. Widely used in regression modeling, it contains no missing values.

**2) Summary Statistics: for mpg**

* **Central Tendency**: Median mpg is 19.2; mean is 20.09, indicating a slight skew towards higher values.
* **Distribution**: Mpg ranges from 10.4 to 33.9 with a standard deviation of 6.03.
* **Quartiles**: IQR is 7.37, with Q1 at 15.43 and Q3 at 22.8.

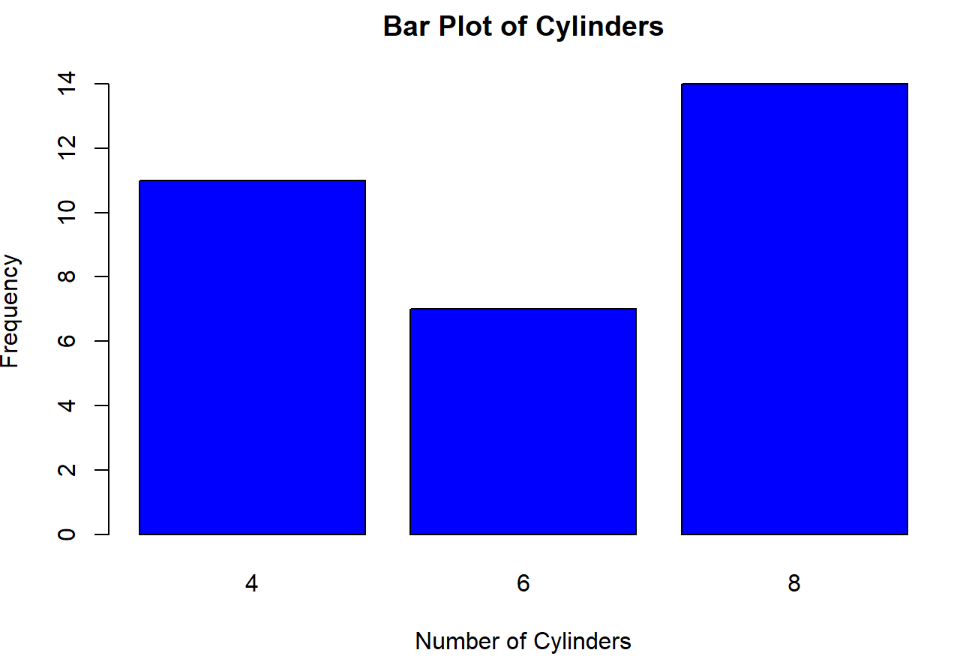


**3) Distribution Visualization:**



* **Histogram**: Shows mpg distribution; most cars cluster between 15 and 25 mpg, with fewer cars at extremes.
* **Boxplot**: Median mpg is 19.2; IQR is 7.37, showing middle 50% of values between 15.43 and 22.8.
* **Statistics Summary**: Mpg ranges from 10.4 to 33.9; mean is 20.09, standard deviation is 6.03, indicating moderate variability.

**4) Categorical Variable Analysis:**

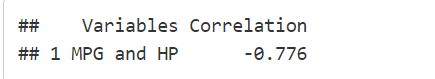


* The bar plot shows the frequency of vehicles with different numbers of cylinders: 4, 6, and 8.
* Vehicles with 8 cylinders are the most frequent, with a count of 14.
* Vehicles with 6 cylinders are the least frequent, with a count of 7. Vehicles with 4 cylinders have a frequency of 11.

**Multivariate Analysis**

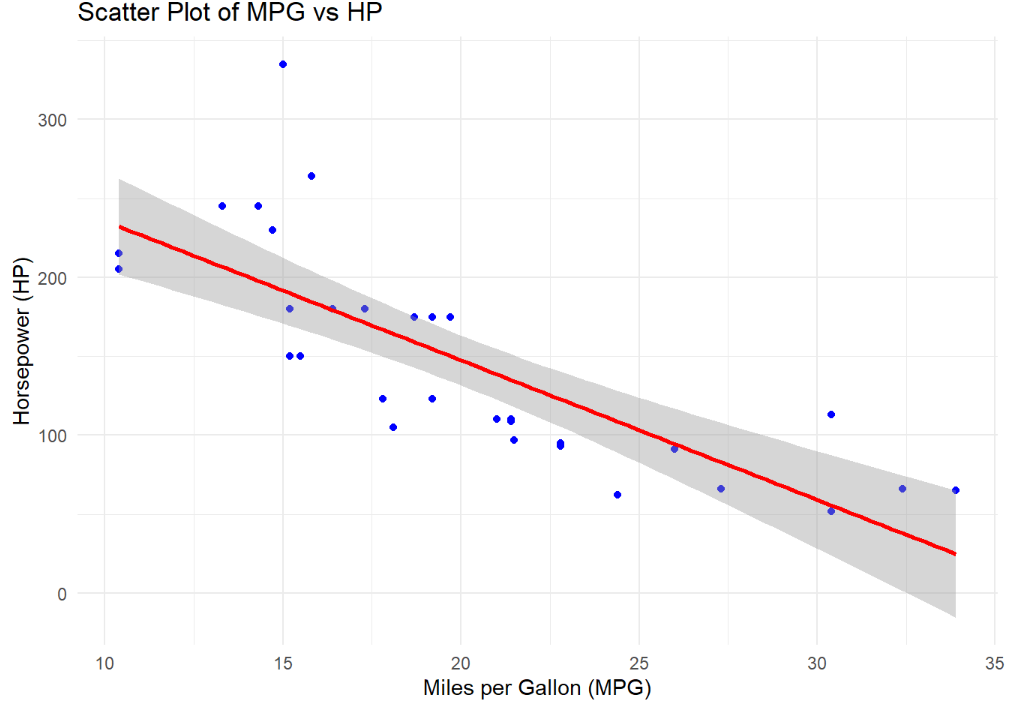
**5) Correlation Analysis**

* **Correlation Analysis:** There is a strong negative correlation of -0.776 between miles per gallon (MPG) and horsepower (HP), indicating that as horsepower increases, fuel efficiency tends to decrease.



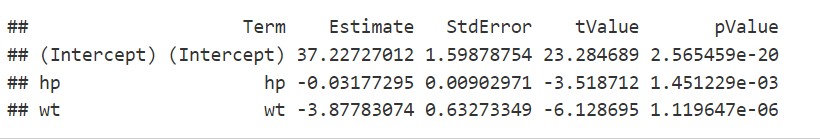
**6) Scatter Plot Visualization:**

* **Negative Correlation**: The scatter plot shows a negative correlation between miles per gallon (MPG) and horsepower (HP), indicating that as MPG increases, HP tends to decrease.
* **Linear Trend:** A red trend line is fitted to the data, highlighting the linear relationship between the two variables, with a shaded area representing the confidence interval.
* **Data Distribution:** The data points are spread across the plot, with most points concentrated between 10-30 MPG and 50-250 HP.

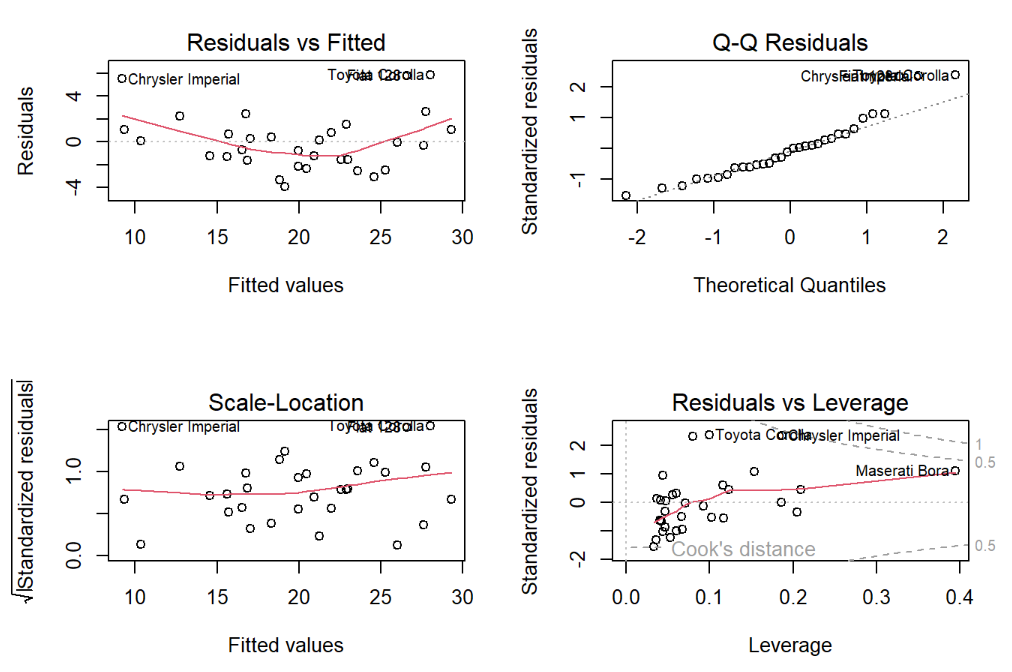
****

**7) Multiple Regression**

* **Intercept and Coefficients:** Baseline is 37.23; hp and wt decrease the dependent variable with coefficients -0.0318 and -3.877.
* **Statistical Significance:** Low p-values for hp (0.0015) and wt (1.12×10−61.12×10−6) indicate significance.
* **Effect Size**: wt has a larger impact than hp, shown by higher absolute coefficient and t-value.

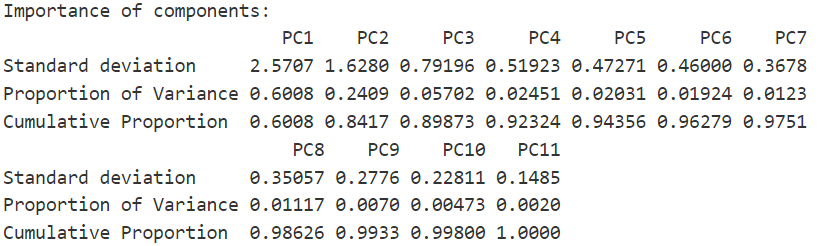
****

**8) Model Diagnostics**

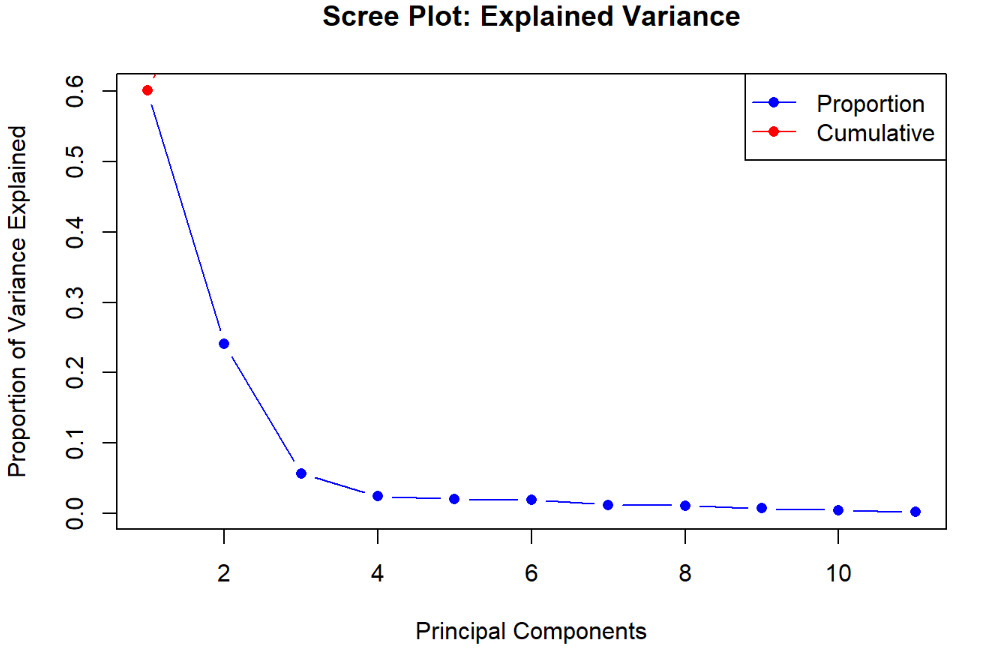
****

* **Residuals vs Fitted Plot**: Shows a non-linear pattern, indicating model fit issues.
* **Q-Q Plot**: Residuals deviate from normal distribution, affecting statistical validity.
* **Residuals vs Leverage Plot**: Identifies influential points like "Toyota Corolla" and "Chrysler Imperial."

**9) PCA**

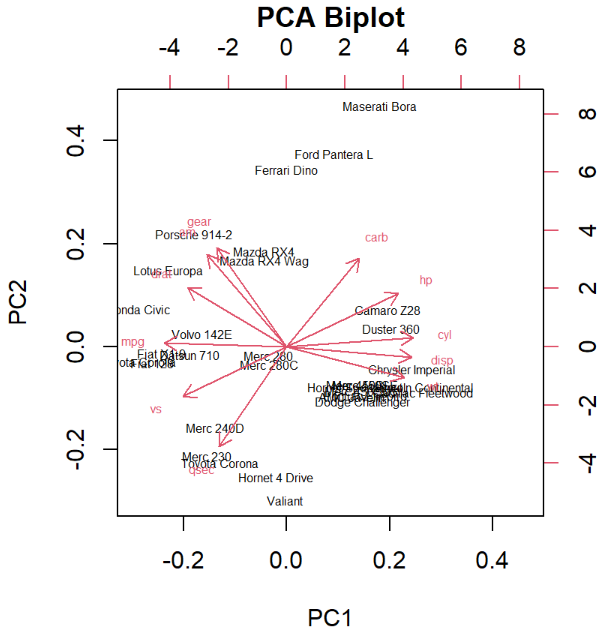
****

* **Variance Explanation:** PC1 explains 60.08% of the variance, and PC2 adds 24.09%, totaling 84.17%.
* **Diminishing Returns:** Variance contribution drops significantly after PC2.
* **Cumulative Variance:** Over 93% of the variance is captured by the first four components.
* **Scree Plot Insight:** The steep decline after PC2 indicates that only a few components are needed for effective dimensionality reduction.

****

**10) PCA Interpretation:**

* **Axes**: The plot shows two principal components, PC1 and PC2, which are linear combinations of the original variables.
* **Points**: Each point represents an observation (e.g., different car models).
* **Arrows**: The red arrows indicate the direction and magnitude of the original variables' contribution to the principal components.
* **Interpretation**: The position of the points and arrows helps in understanding the relationships between observations and variables. Observations close to each other are similar, and the direction of arrows shows the influence of variables on these components.



Titanic Data-set

**1) About the data set:**

The Titanic dataset contains 891 entries and 12 attributes, including survival status, age, gender, class, and fare. Notable features are missing values in 'Age' and 'Cabin'. It is widely used in machine learning for survival prediction and classification tasks.

**2) Summary Statistics: for mpg**

**Age**

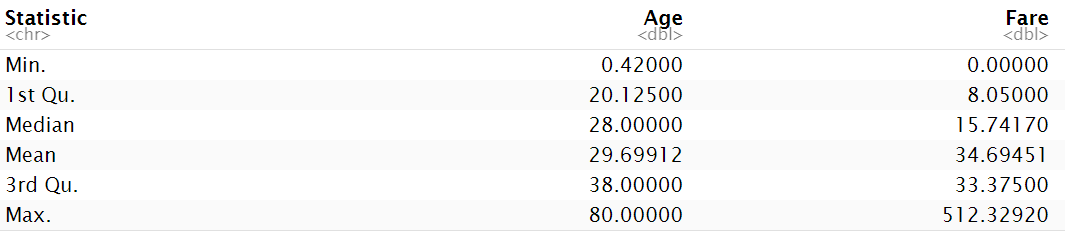
**Range:** Minimum is 0.42 years, and maximum is 80 years.

**Central Tendency:** Median is 28 years, mean is approximately 29.7 years.

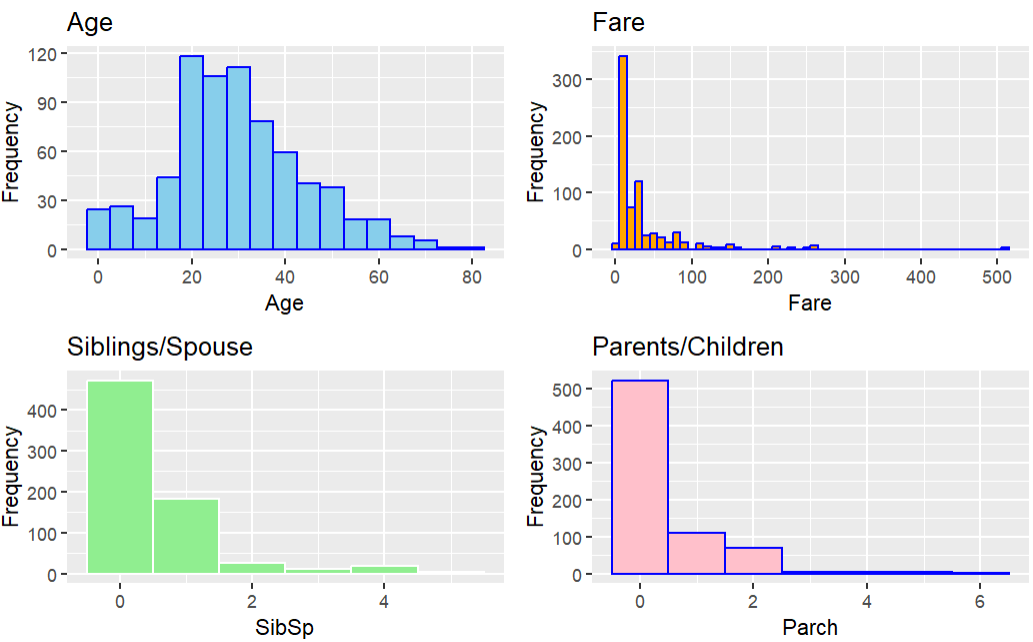
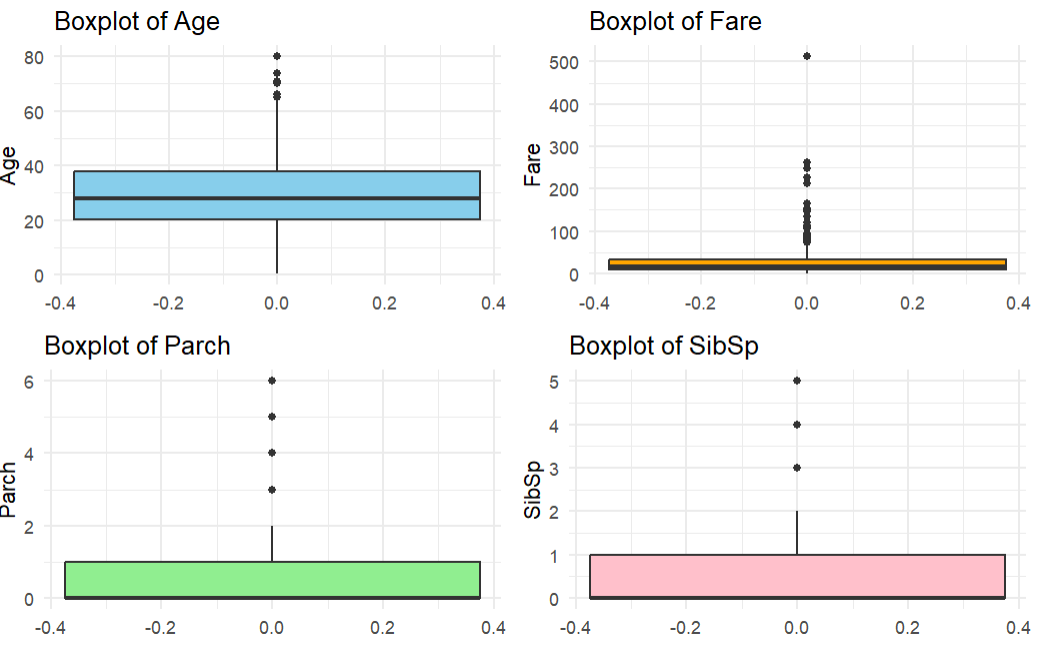
**Fare**

**Range:** Minimum is $0.00, and maximum is $512.33.

**Central Tendency:** Median is $15.74, mean is approximately $34.69

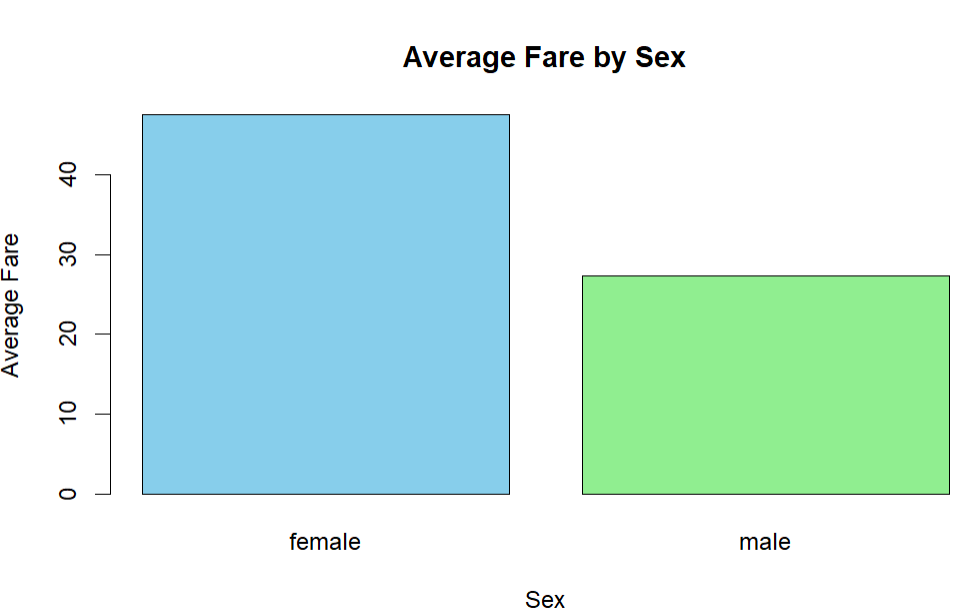


**3) Distribution Visualization:**

* **Age and Fare Distribution**: The age distribution is approximately normal, with most passengers aged between 20 and 40. The fare distribution is highly skewed, with most fares below 100 and a few extreme outliers above 500.
* **Family Variables (SibSp and Parch)**: Most passengers traveled with no siblings/spouses (SibSp = 0) or parents/children (Parch = 0). A small number had up to 4 family members in these categories.
* **Outliers**: Boxplots reveal outliers in all variables, particularly in "Fare," which has significant extreme values. "Age," "SibSp," and "Parch" also show some outliers, though less prominent.

**4) Categorical Variable Analysis:**

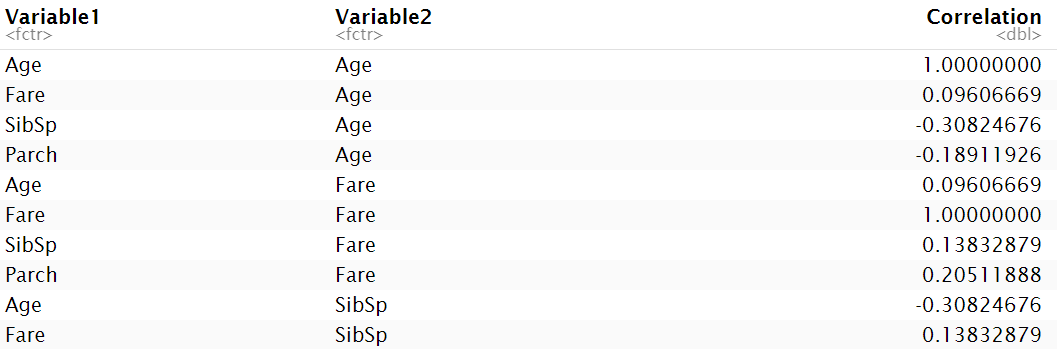


* **Higher Average Fare for Females**: Females paid higher average fares (~45) compared to males (~20), as shown by the taller bar.
* **Gender-Based Disparity**: This disparity may reflect differences in travel class, group dynamics, or other socioeconomic factors.

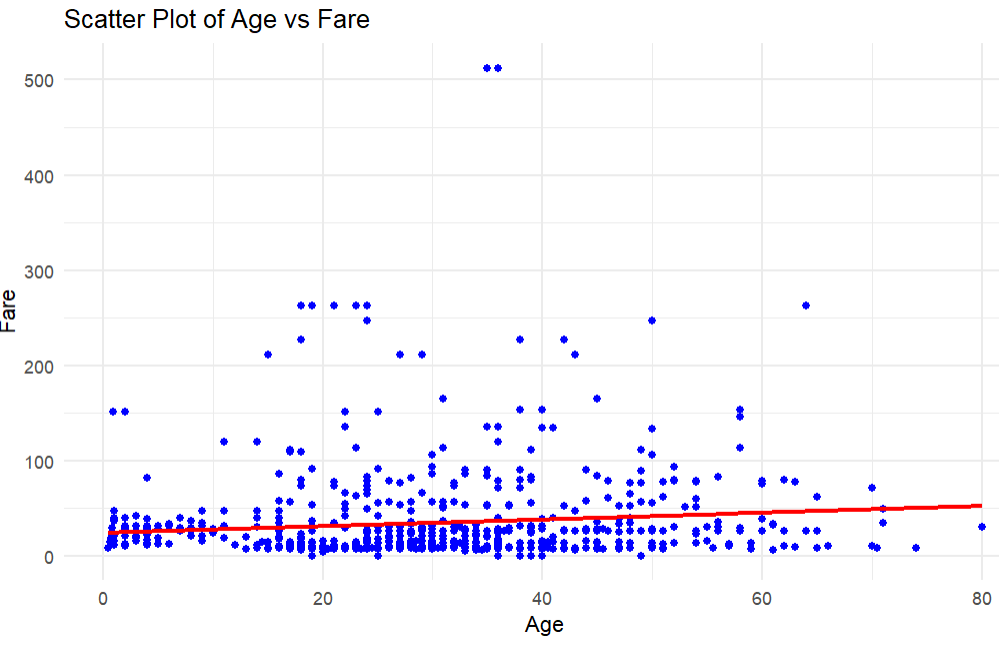
**Multivariate Analysis**

* **Strongest correlation**: Fare and Parch show the highest positive correlation (0.21), indicating a moderate relationship.
* **Negative correlations:** SibSp and Age (-0.31) exhibit the strongest negative correlation, followed by Parch and Age (-0.19).
* **Weakest correlation**: Fare and Age have the weakest positive correlation (0.10), suggesting minimal association.

**5) Correlation Analysis**



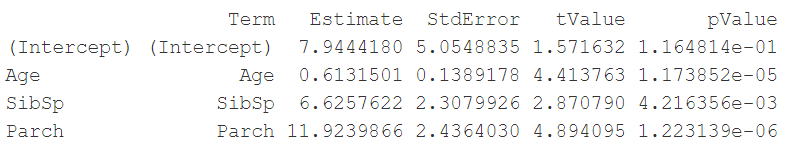
**6) Scatter Plot Visualization:**

****

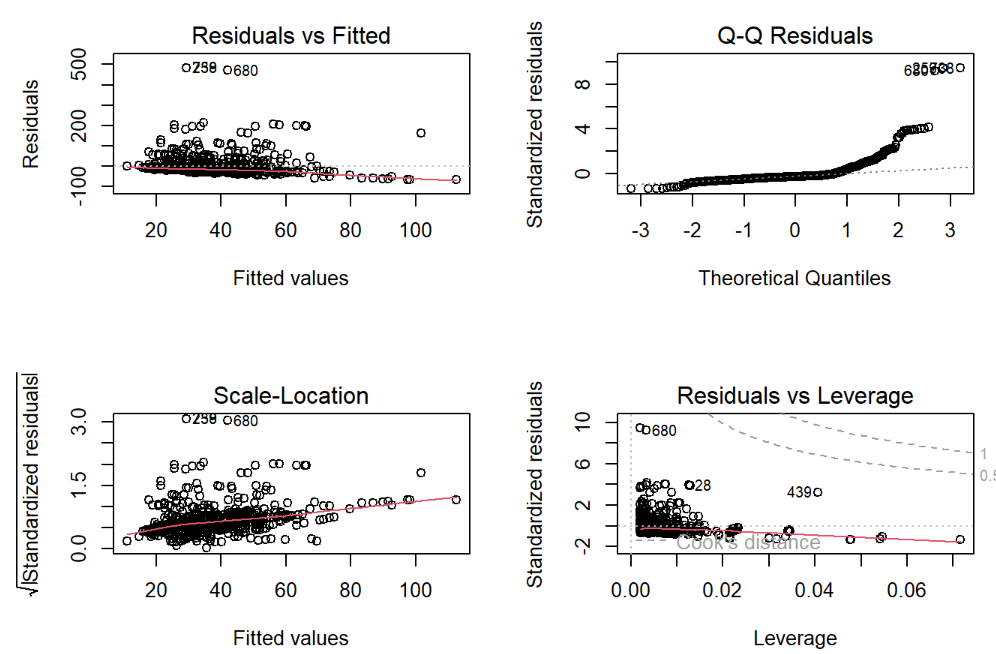
The scatter plot shows a weak positive correlation between age and fare, as indicated by the red trend line. Most fares cluster below 100, with a few outliers above.

* The predictors **Age**, **SibSp**, and **Parch** are statistically significant as their p-values are less than 0.05.
* The intercept is not statistically significant (p-value = 0.1648).
* The predictor with the highest effect size is **Parch** (Estimate = 11.92), while **Age** has the lowest effect size (Estimate = 0.61).
* **Age** is the most precise predictor (lowest StdError = 0.1389), while the intercept is the least precise (highest StdError = 5.0548).
* Based on t-values, **Parch** is the strongest predictor (t-value = 4.894), while the intercept is the weakest (t-value = 1.572).

**7) Multiple Regression**

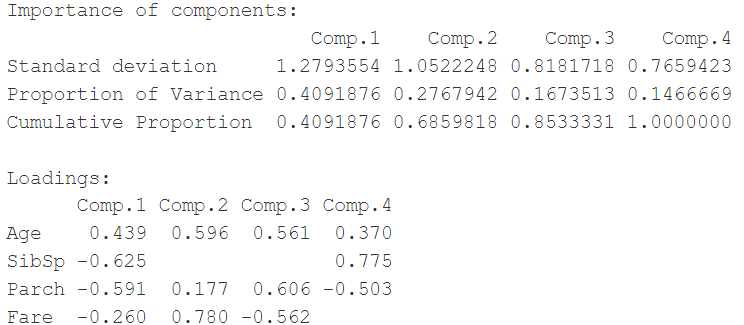
****

**8) Model Diagnostics**

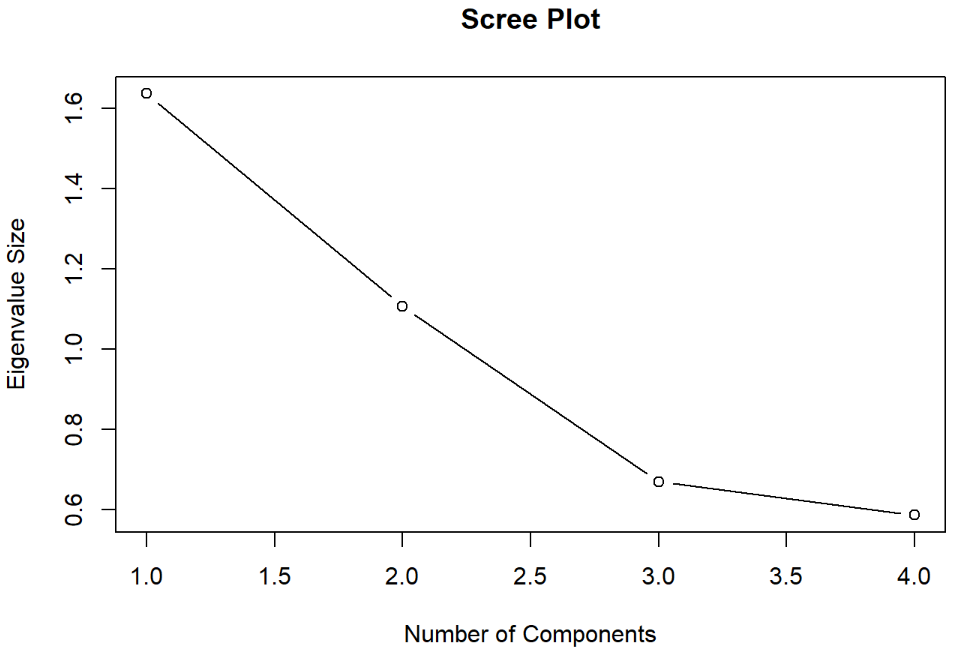


* **Residuals vs Fitted Plot**: Shows a non-linear pattern, indicating model fit issues.
* **Q-Q Plot**: Residuals deviate from normal distribution, affecting statistical validity.
* **Residuals vs Leverage Plot**: Identifies influential points like "Toyota Corolla" and "Chrysler Imperial."

**9) PCA**

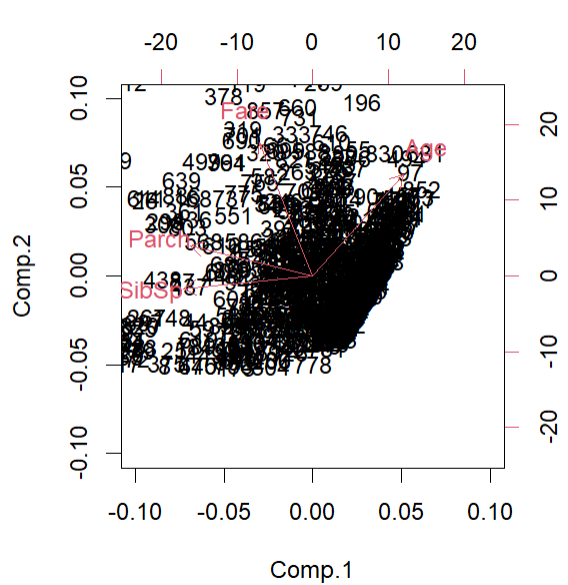
****

* **Variance Explained:** Comp.1 and Comp.2 explain 68.59% of the variance, making them the most significant components.
* **Key Loadings:** Comp.1 is influenced by "SibSp" (-0.625) and "Parch" (-0.591), while Comp.2 is driven by "Fare" (0.780) and "Age" (0.596).
* **Scree Plot:** A steep drop after the second component suggests retaining only the first two for analysis.
* **Dimensionality Reduction:** The first two components are sufficient for simplifying the dataset effectively.

****

**10) PCA Interpretation:**

* **Variance Explained:** The first two components (Comp.1 and Comp.2) explain 68.59% of the variance, making them the most significant for analysis.
* **Key Loadings:** Comp.1 is strongly influenced by "SibSp" (-0.625) and "Parch" (-0.591), while Comp.2 is dominated by "Fare" (0.780) and "Age" (0.596).
* **Scree Plot Insight:** The scree plot shows a steep decline after the second component, suggesting that only the first two components are necessary for dimensionality reduction.
* **Biplot Interpretation:** The biplot indicates that "Fare" and "Age" align closely with Comp.2, while "SibSp" and "Parch" align more with Comp.1, reflecting distinct variable contributions to these components.

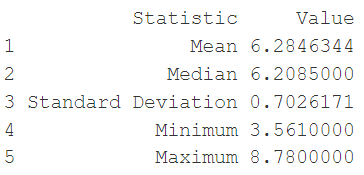


Boston Housing Data-set

**1) About the data set:** The `Boston` dataset contains 506 observations and 14 variables related to housing prices and socioeconomic factors in Boston. It is widely used for regression and statistical modeling tasks.

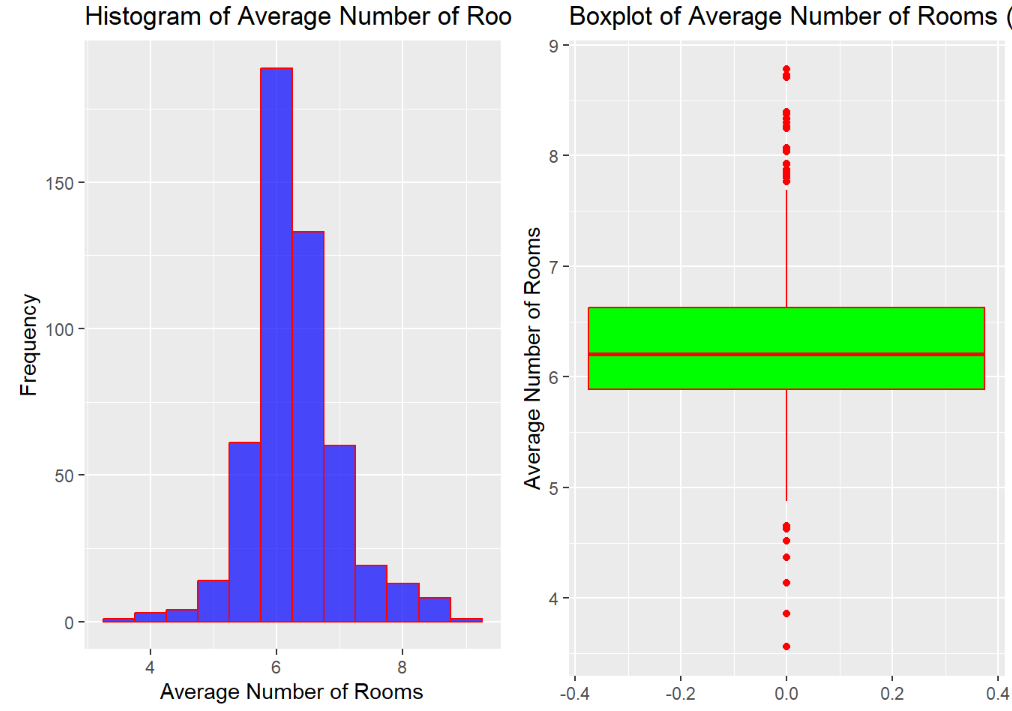
**2) Summary Statistics: for mpg**

The average number of rooms per dwelling (rm) is approximately 6.28, with a median of 6.21, indicating a fairly symmetric distribution. The standard deviation of 0.70 suggests moderate variability, while the number of rooms ranges from a minimum of 3.56 to a maximum of 8.78

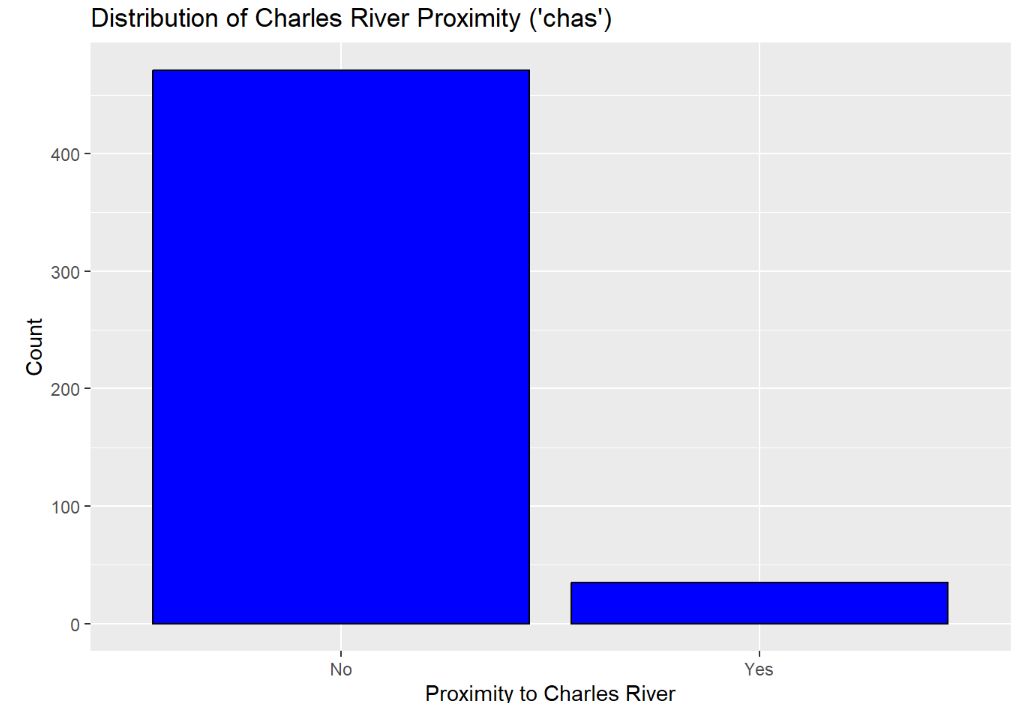


**3) Distribution Visualization:**

* The spread of the data, as shown in the histogram, is relatively narrow, with most values concentrated around the mean (approximately 6 rooms). The distribution appears symmetric and bell-shaped, indicating a normal distribution.
* The boxplot highlights the presence of outliers, with several data points above the upper whisker. These outliers represent properties with significantly more rooms than the majority.
* The interquartile range (IQR) is well-defined, with the majority of data falling between approximately 5 and 7 rooms. The green box in the boxplot represents this central spread of data.



**4) Categorical Variable Analysis:**

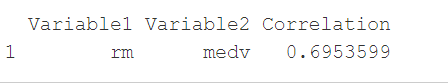


* Proximity to Charles River (Bar Chart): The majority of properties are not located near the Charles River, as indicated by the overwhelming count in the "No" category compared to the much smaller "Yes" category.

**Multivariate Analysis**

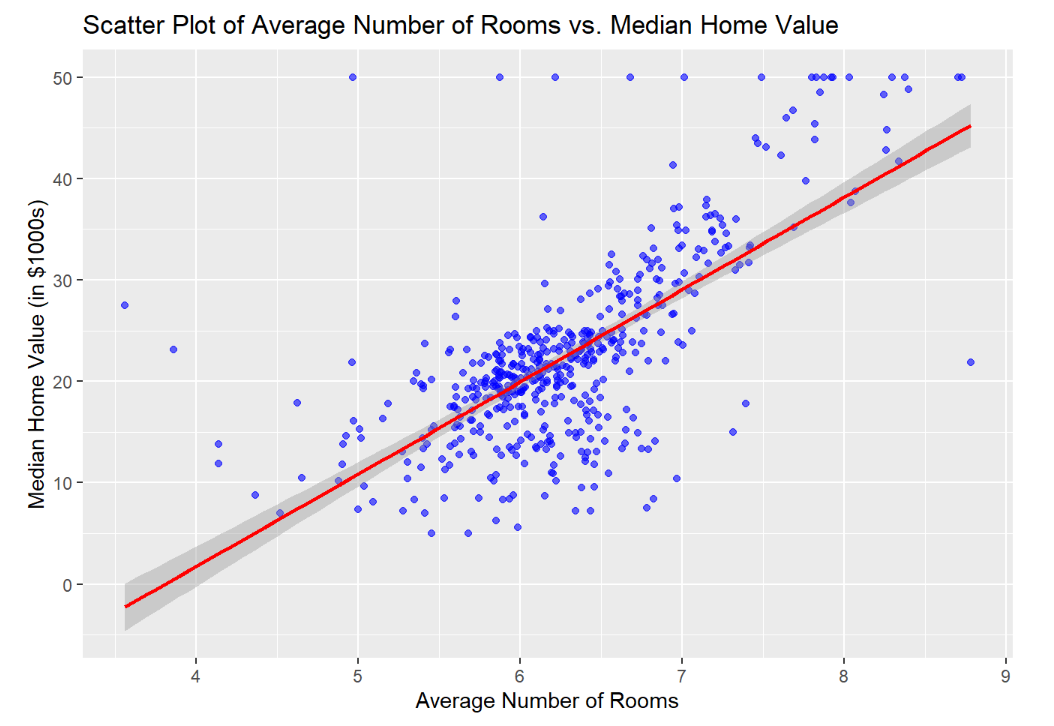
**5) Correlation Analysis**

* For rm (average number of rooms) and medv (median home value), a positive correlation (e.g., 0.7) would suggest that as the number of rooms increases, the home value tends to increase as well.



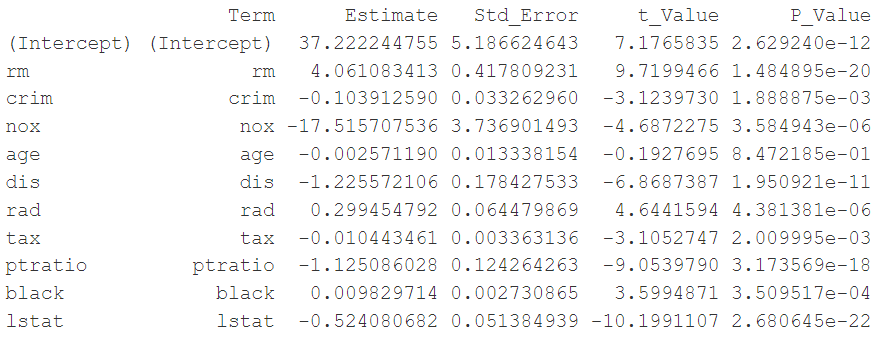
**6) Scatter Plot Visualization:**

* he scatter plot shows a **positive linear relationship** between the average number of rooms in a home and the median home value. As the number of rooms increases, the median home value tends to rise.
* The red regression line indicates the trend, with a **narrow confidence interval (shaded area)** suggesting a relatively strong correlation.
* There is some **variability in home values** for homes with 5–7 rooms, but the overall trend remains consistent, with fewer outliers as room numbers increase.

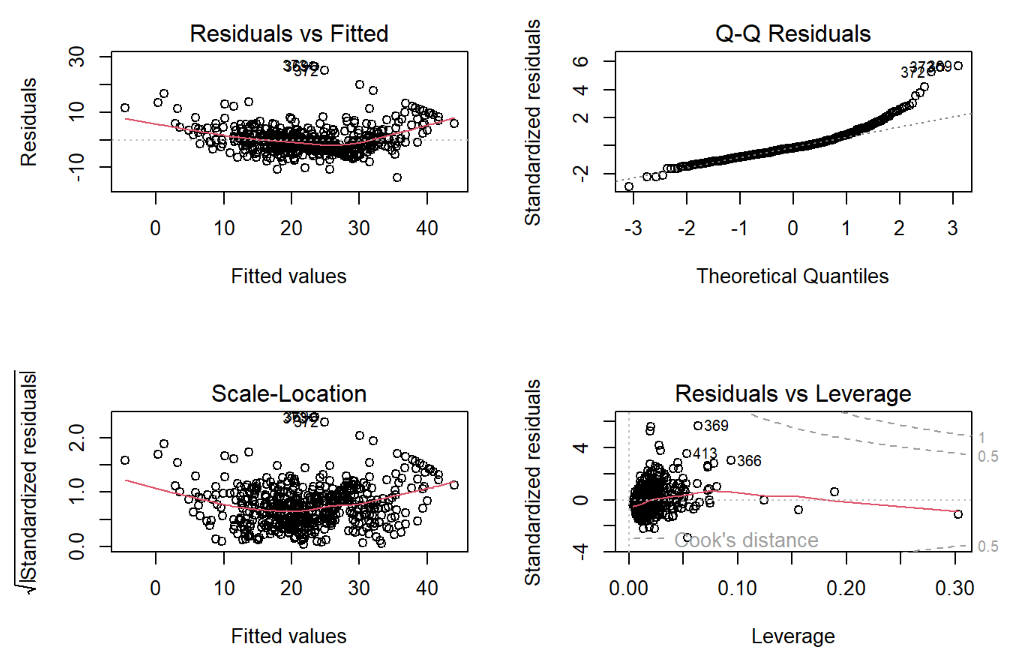
****

**7) Multiple Regression**

* **Significant Predictors:** The variable rm (number of rooms) has a positive coefficient (0.061) and a very low p-value, indicating it is a significant positive predictor. Similarly, lstat (lower status population percentage) has a negative coefficient (-0.524) and is also highly significant.
* **Insignificant Variables:** Variables like age and dis have high p-values (> 0.05), suggesting they are not statistically significant predictors in this model.
* **Other Notable Findings:** Variables such as nox, rad, and ptratio are statistically significant but have varying effects, as indicated by their coefficients and low p-values.
* Based on t-values, **Parch** is the strongest predictor (t-value = 4.894), while the intercept is the weakest (t-value = 1.572).

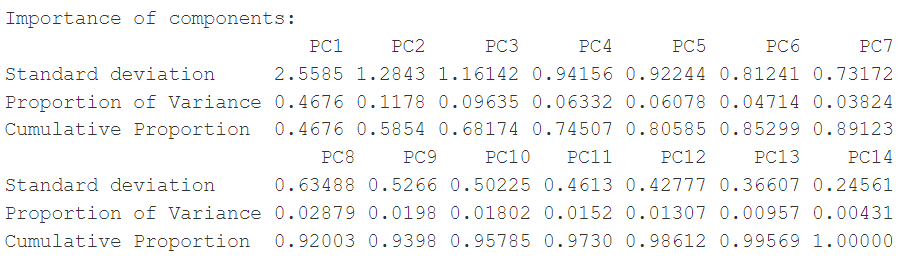
****

**8) Model Diagnostics**

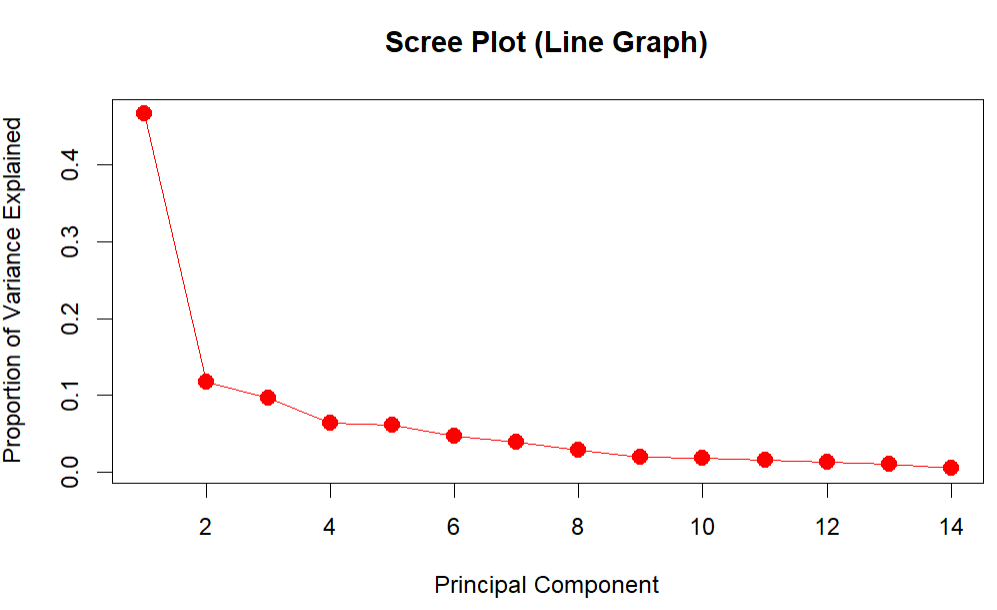
****

1. **Non-Linearity in Residuals:** The "Residuals vs Fitted" plot shows a curved pattern, indicating the model may not fully capture the relationship.
2. **Non-Normal Residuals:** The "Q-Q Residuals" plot displays deviations from the straight line, suggesting non-normality of residuals.
3. **Outliers and High Leverage Points:** The "Residuals vs Leverage" plot highlights influential points (Cook's distance), which may affect the model.

**9) PCA**



1. **PC1 explains the largest variance**: It accounts for **46.76% of the total variance**, making it the most significant component.
2. **Five components explain over 80% of the variance**: The cumulative proportion reaches **80.59%** by including PC1 to PC5, sufficient for dimensionality reduction.
3. **Elbow point at PC5 or PC6**: The scree plot shows diminishing returns in variance explained beyond PC5, marking it as the "elbow point."

****

* **Principal Components:**PC1 (46.76%) and PC2 (11.76%) explain 58.52% of variance**.**
* **Red Arrows:**Represent variable contributions; direction shows correlation, length indicates **strength.**
* **Data Points:**Observations, colored by chas (Charles River proximity).
* **Patterns:**Similar observations cluster; chas = 1 concentrated in specific regions**.**

**10) PCA Interpretation:**

