**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**

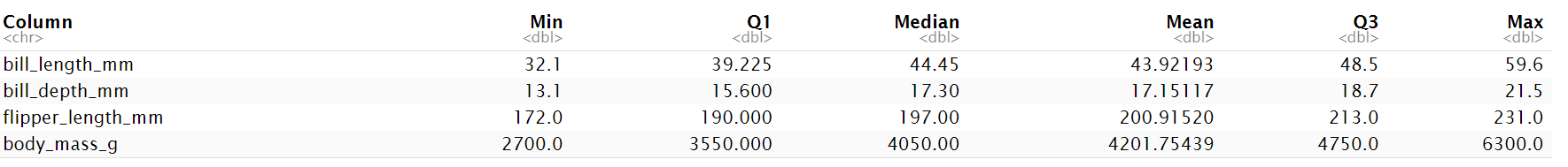
**24BM6JP55**

**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:**

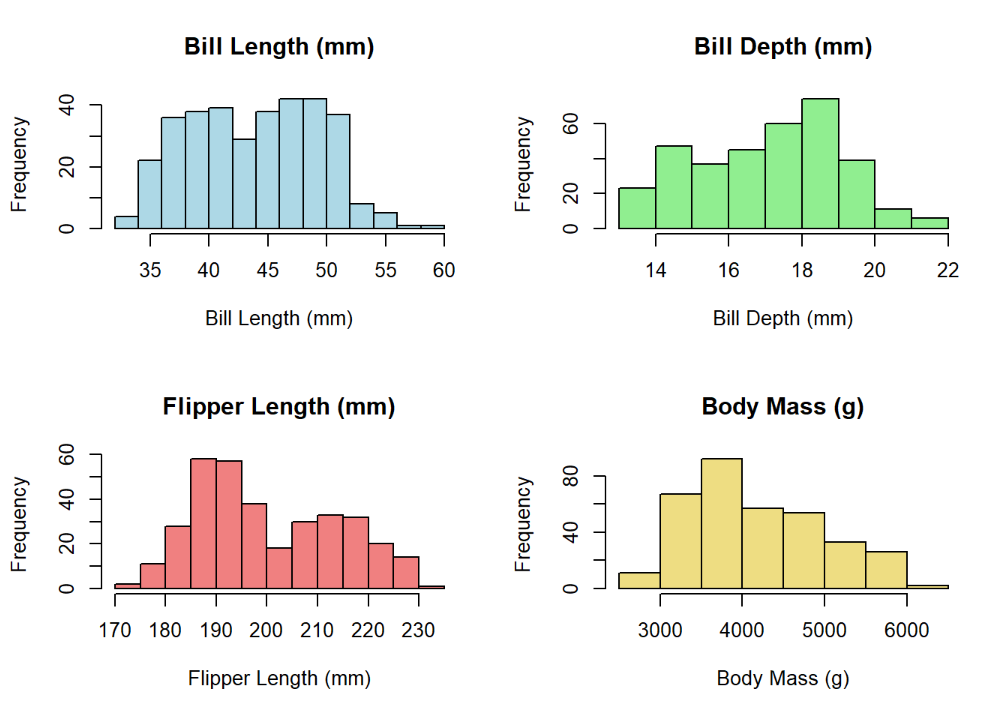
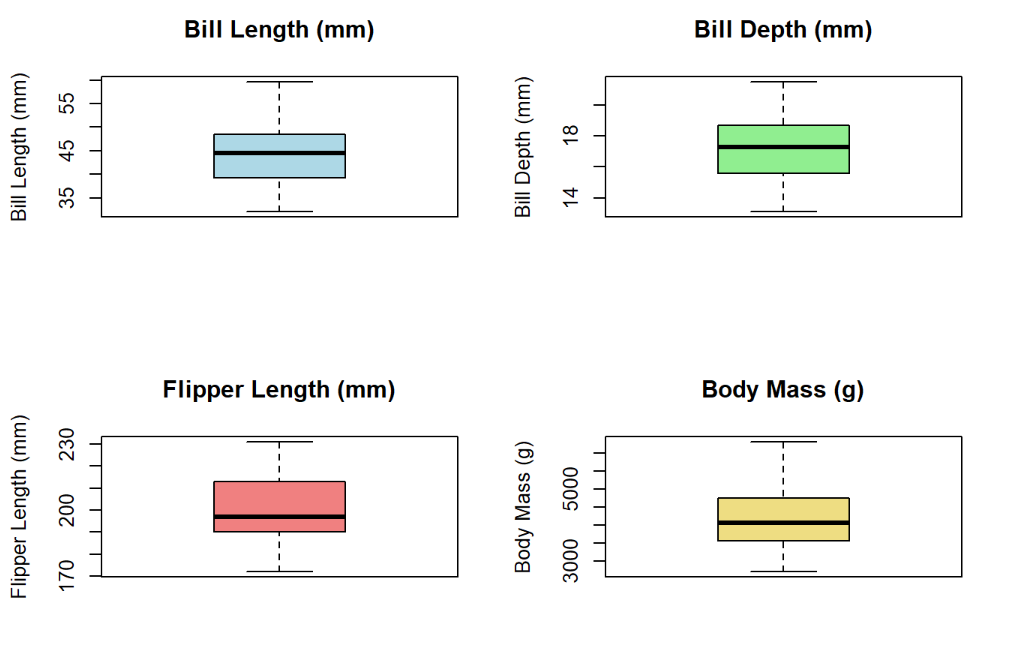


**Measurement Range:** Body mass shows the largest range, indicating high variability, while bill length and depth have smaller ranges.

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

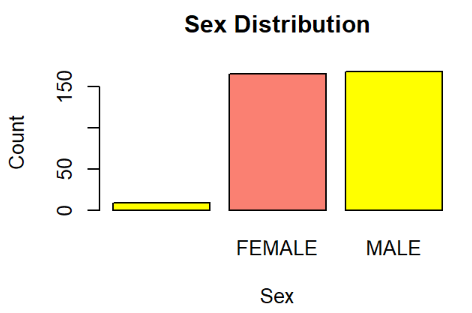
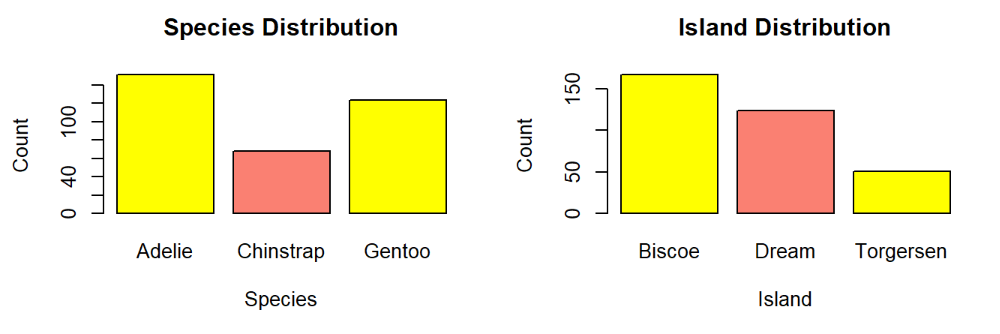
**Distribution:** Body mass data is more spread out compared to other categories, which display more consistent distributions.

**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:**

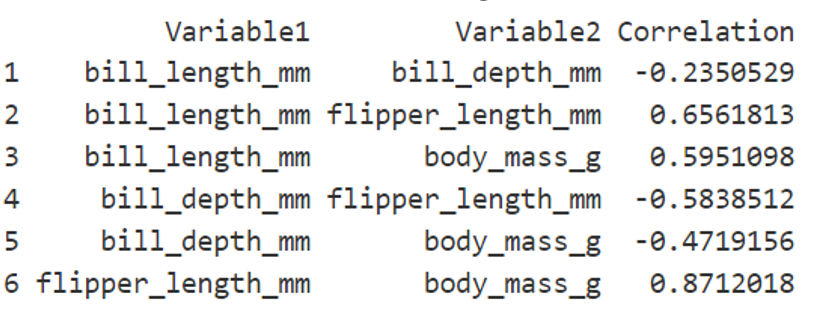
****

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.

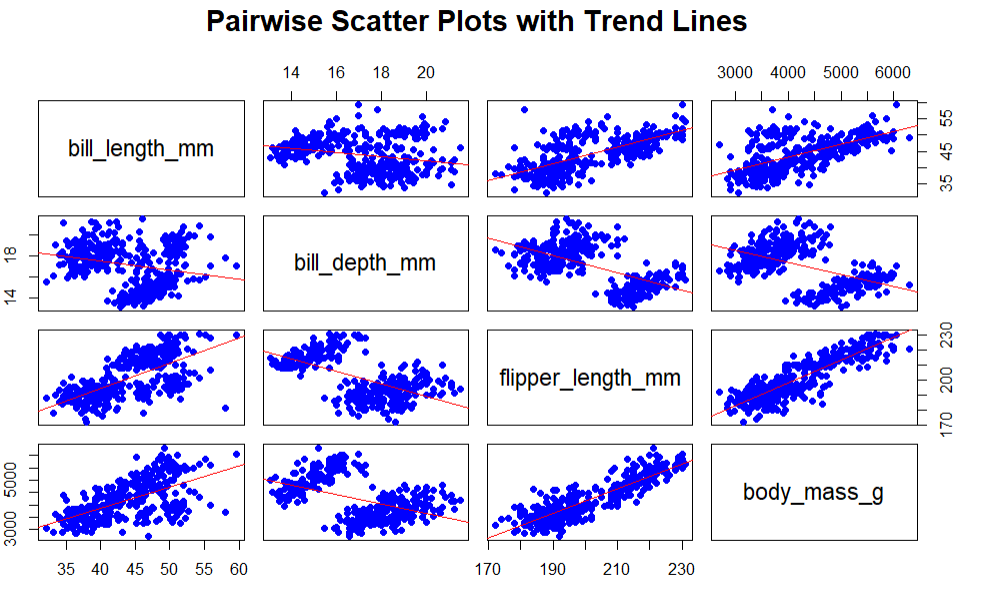
**Multivariate Analysis**

**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.

****

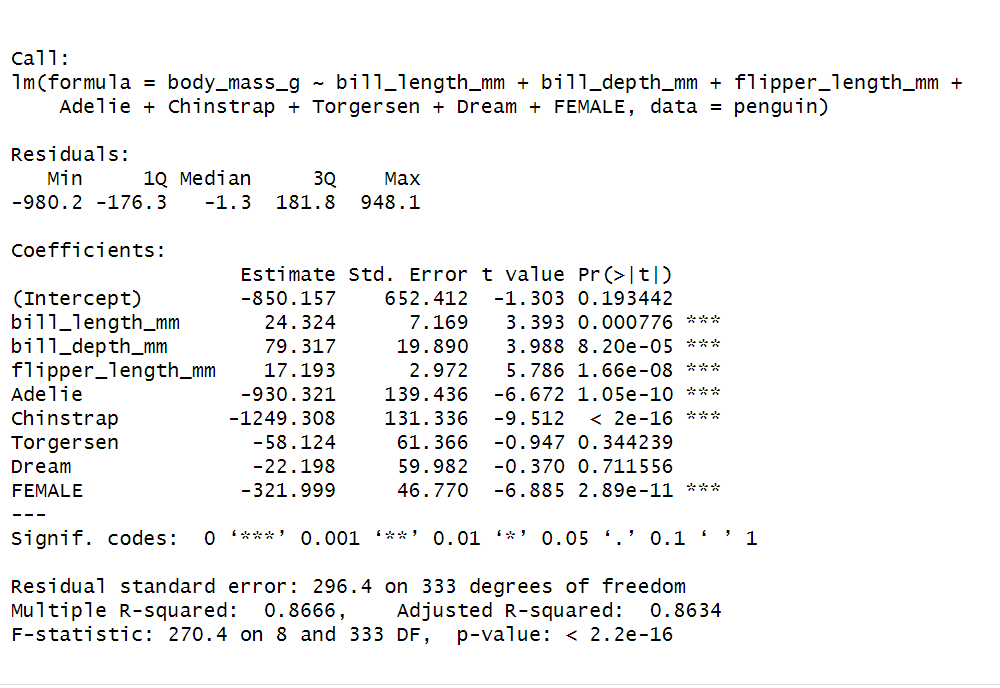
**6) Scatter Plot Visualization**

****

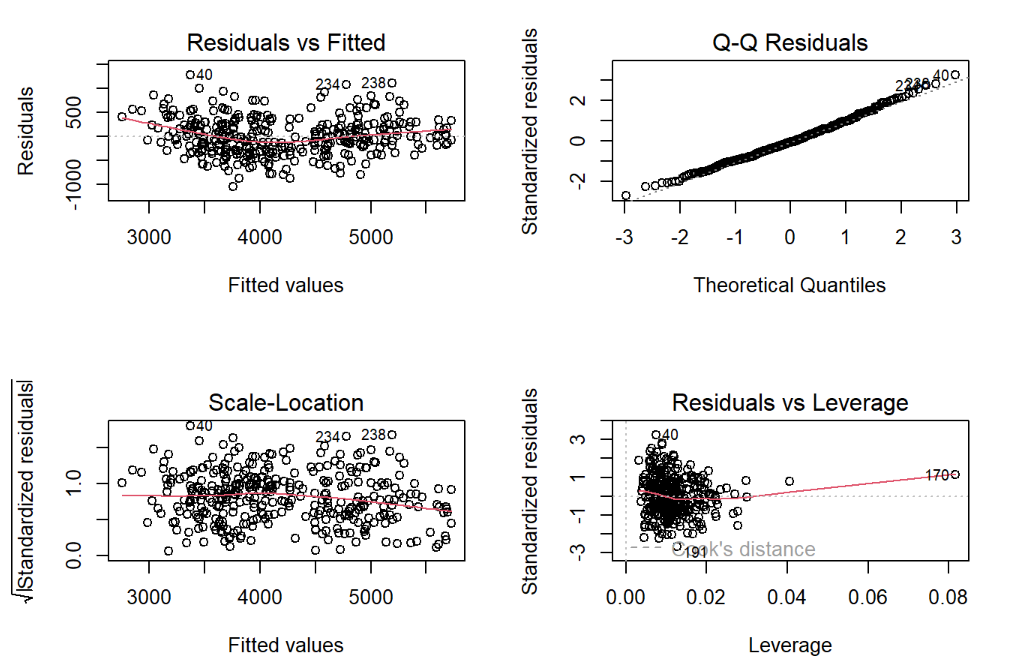
* The scatter plot matrix shows pairwise relationships between four variables: bill length, bill depth, flipper length, and body mass.
* Each plot includes a trend line indicating correlation.
* Bill length and flipper length have a positive relationship with body mass, while bill depth shows weaker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

**7) Multiple Regression**

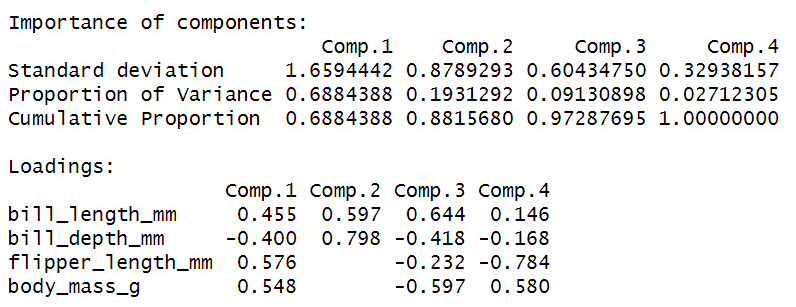
* The linear model predicts *body mass* using *bill length*, *bill depth*, *flipper length*, and penguin species, with a high R-squared of 0.8666.
* Significant predictors include bill length, bill depth, and flipper length (p < 0.001).
* The model's F-statistic is 270.4, indicating overall significance (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.

****

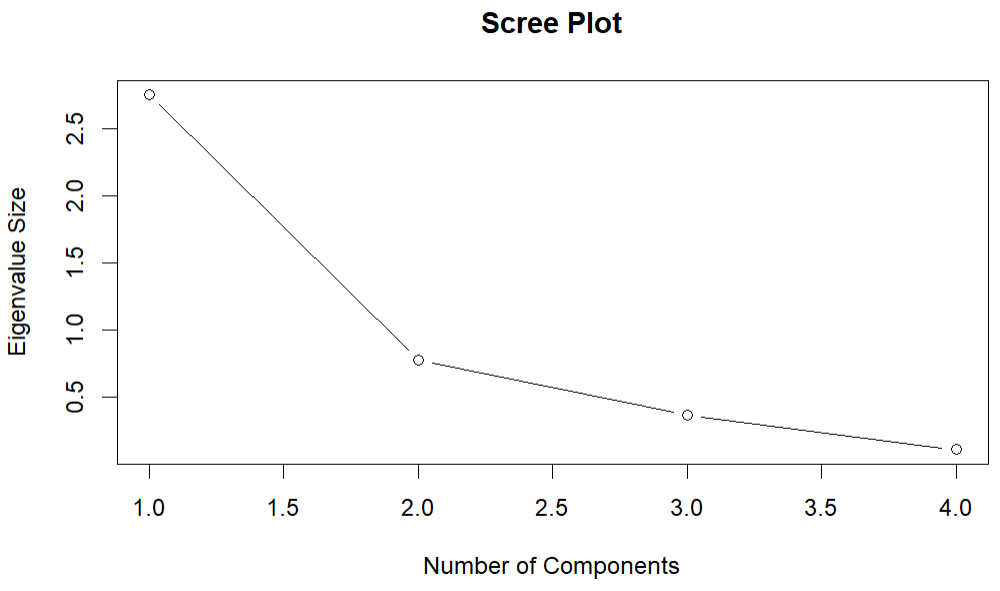
**8) Model Diagnostics**

****

* 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.

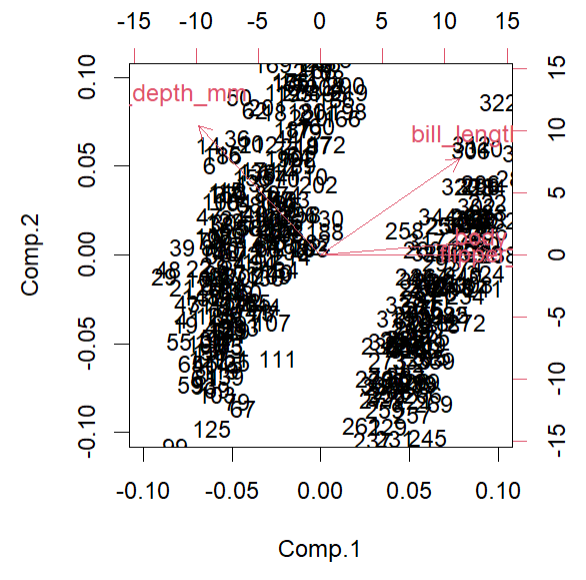
**9) PCA  
**

* 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.

****

**Conclusion:** The multiple regression analysis of penguin body mass identifies several key factors influencing their weight. Significant predictors include bill dimensions (length and depth) and flipper length, all of which correlate strongly with body mass. Gentoo penguins are the heaviest species, while Adelie and Chinstrap species show lower body masses. Additionally, female penguins are generally lighter than males, reflecting sexual dimorphism. Island location has minimal impact on body mass, as Torgersen and Dream islands do not significantly affect the results. The model explains nearly 87% of the variance in body mass, indicating strong performance.Model diagnostics reveal that residuals are approximately normally distributed, although some minor heteroscedasticity is present. Overall, the model fits well, effectively capturing the relationships between predictors and body mass.Principal Component Analysis (PCA) further simplifies the dataset by highlighting dominant patterns. The first two principal components account for most of the variance, with the first component focusing on size-related features and the second emphasizing bill depth. The scree plot confirms that additional components contribute little to variance.The PCA biplot illustrates clear relationships among variables, showing clusters of heavier and larger penguins, with flipper length closely linked to body mass. This analysis also suggests potential species separations based on size and shape differences among penguins. Overall, PCA provides a valuable visualization of interrelationships between traits, aiding in understanding the complex morphological variations within the dataset

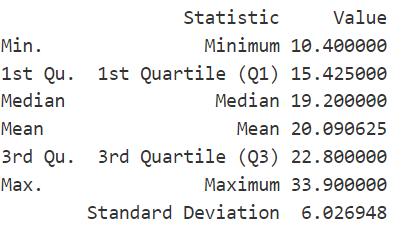
MTCars Data-set

**1) About the data set:**

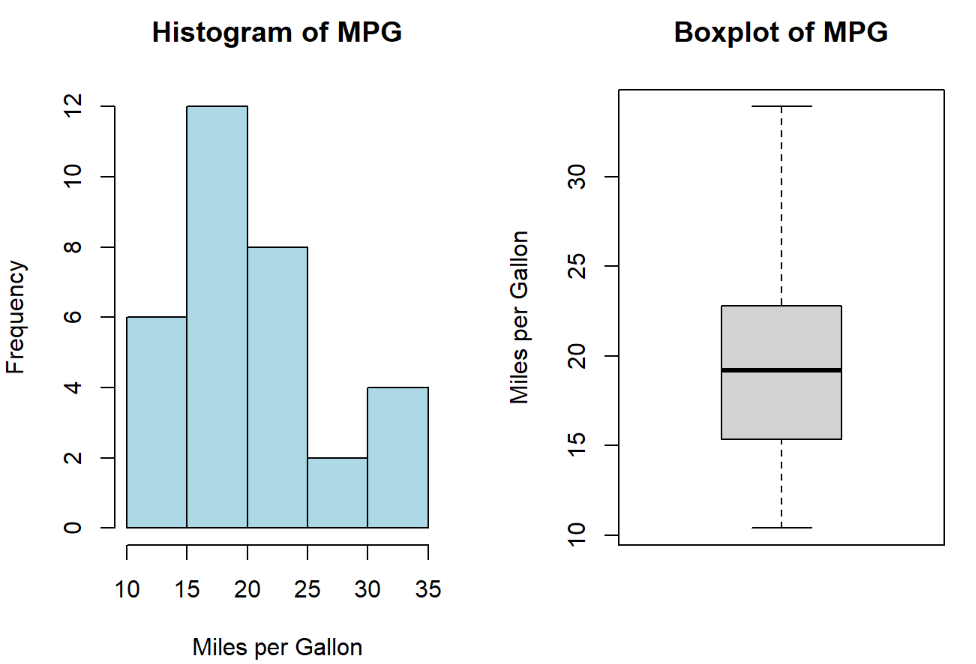
The mtcars dataset contains 32 entries representing various car models with numeric attributes such as miles per gallon (mpg), horsepower (hp), weight (wt), and more. It includes performance, design, and efficiency metrics like cylinder count and transmission type. Widely used in regression modeling, 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.

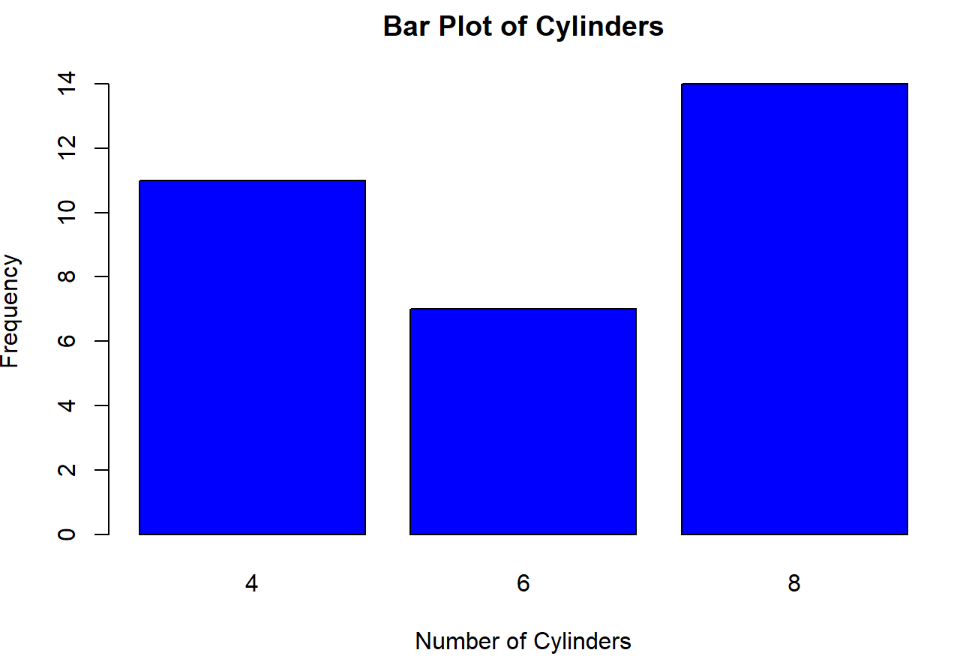


**3) Distribution Visualization:**



* **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:**

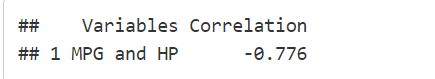


* The bar plot shows the frequency of vehicles with different numbers of cylinders: 4, 6, and 8.
* Vehicles with 8 cylinders are the most frequent, with a count of 14.
* Vehicles with 6 cylinders are the least frequent, with a count of 7. Vehicles with 4 cylinders have a frequency of 11.

**Multivariate Analysis**

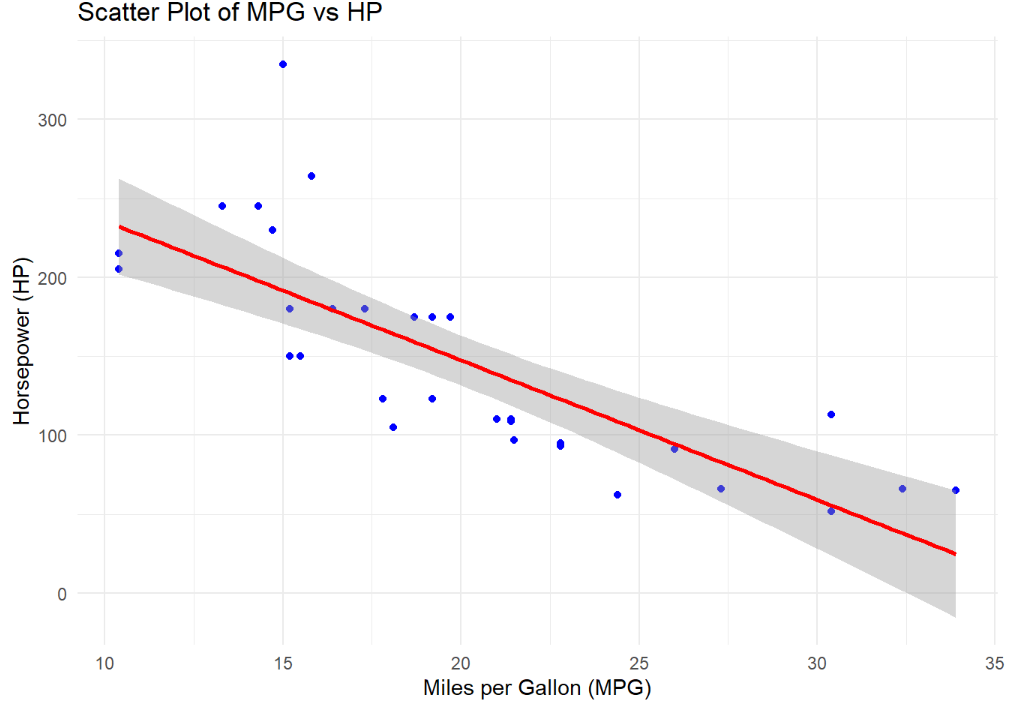
**5) Correlation Analysis**

* **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.



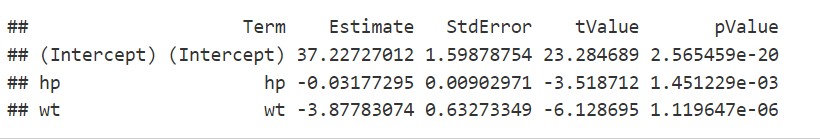
**6) Scatter Plot Visualization:**

* **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.

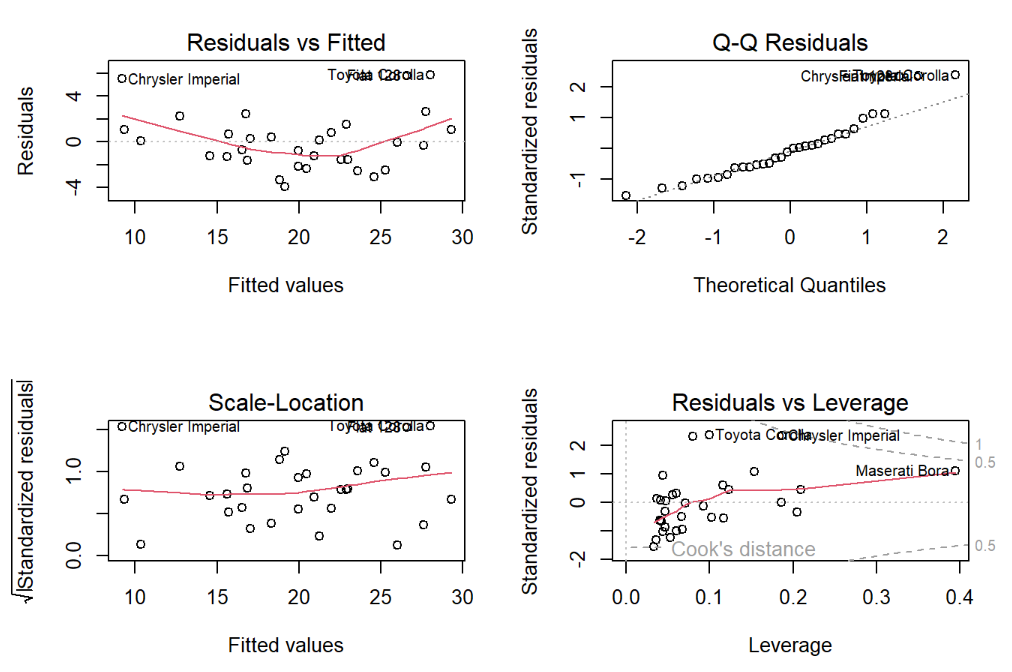
****

**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.

****

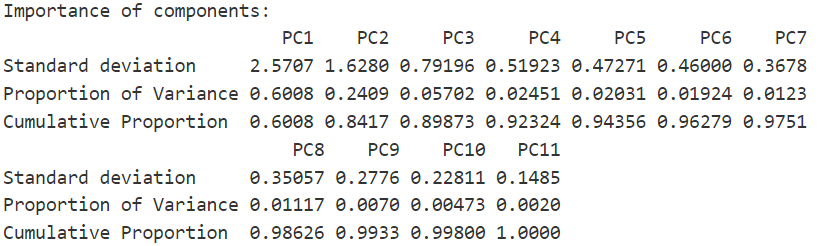
**8) Model Diagnostics**

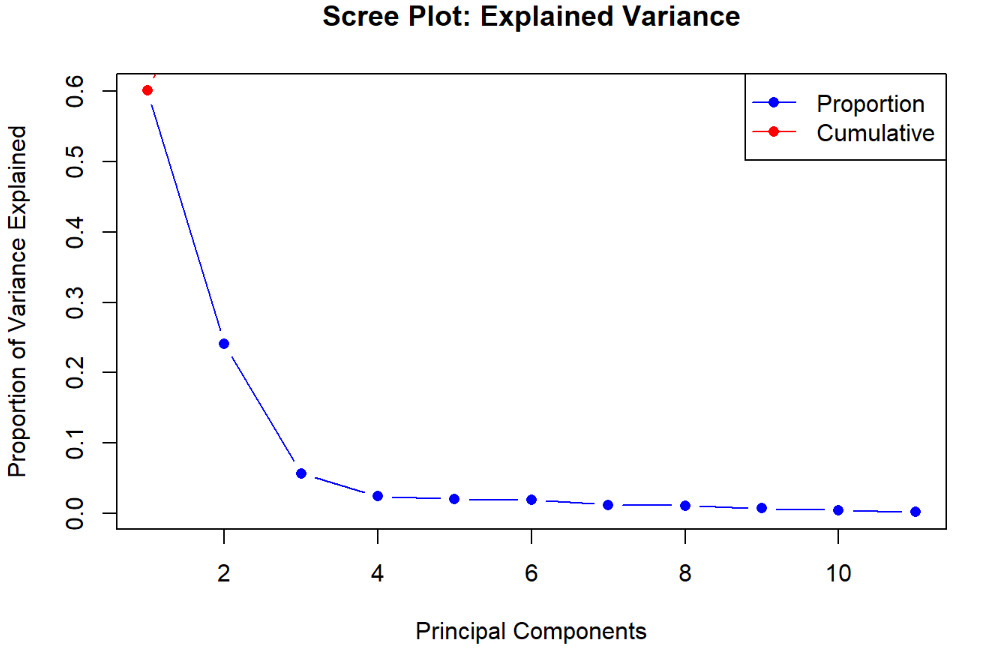
****

* **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**

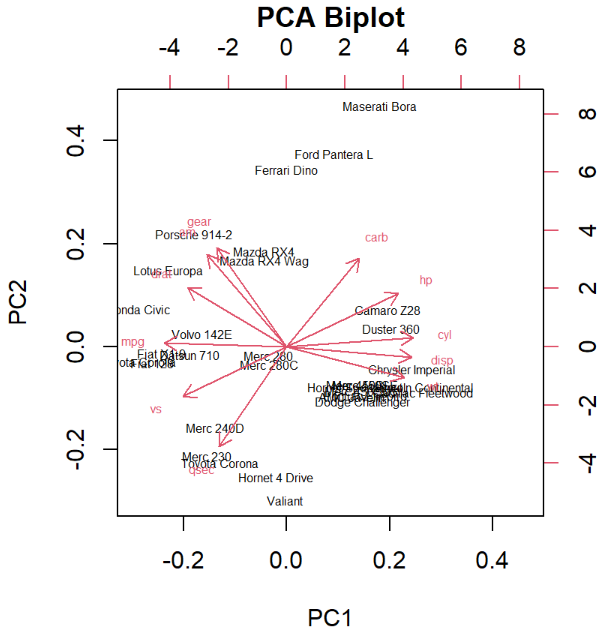
* **Variance Explanation:** PC1 explains 60.08% of the variance, and PC2 adds 24.09%, totaling 84.17%.
* **Diminishing Returns:** Variance contribution drops significantly after PC2.
* **Cumulative Variance:** Over 93% of the variance is captured by the first four components.
* **Scree Plot Insight:**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

****

****

* **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).
* **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.

**10) PCA Interpretation:**



**Conclusion:** The mtcars dataset analysis reveals a fascinating diversity of car attributes, showcasing everything from fuel-efficient models to high-performance vehicles. In the univariate analysis, we identify outliers in variables like miles per gallon (mpg), which represent unique or extreme car models. The multivariate analysis highlights an interesting trade-off between fuel efficiency and factors such as horsepower (hp) and weight (wt), indicating that more powerful cars often sacrifice fuel efficiency.

Additionally, there is a strong correlation between horsepower and weight, which jointly influences a car's mileage. Clear segments emerge based on the number of cylinders and gears, with 4-cylinder cars favoring efficiency while 8-cylinder cars focus on performance.Principal Component Analysis (PCA) simplifies the dataset by capturing most of the variability with just two principal components. The PCA biplots reveal natural clusters in car attributes that align with concepts of efficiency and performance. Key insights from this analysis emphasize the importance of understanding trade-offs for targeting specific market demands. The distinct groupings based on car features can guide effective market positioning, while dimensional reduction highlights the core attributes driving variability, ultimately simplifying analysis and decision-making. Overall, this analysis provides valuable insights into the relationships among car attributes, aiding manufacturers and marketers in navigating the automotive landscape effectively.

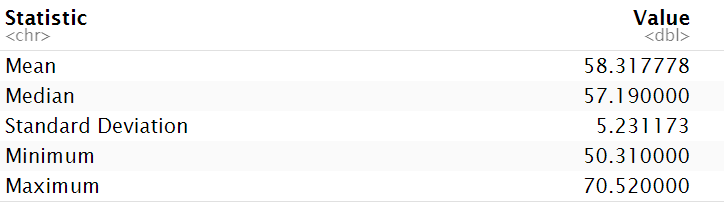
Decathelon2 Data-set

**1) About the data set:**

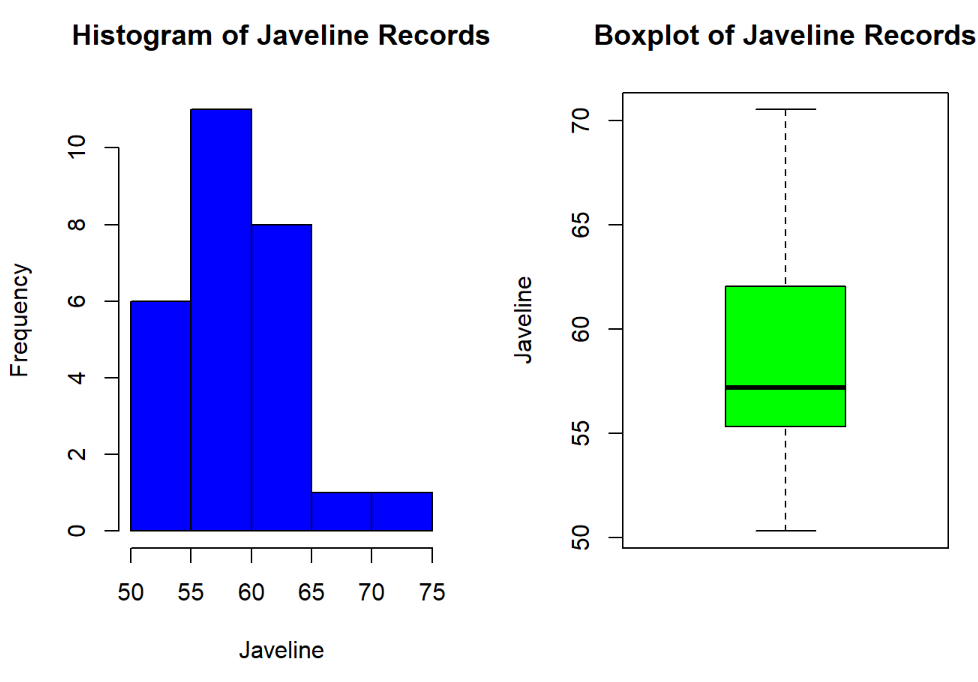
The decathlon2 dataset contains decathlon performance data with 23 athletes and 10 variables (events). It includes results for events like 100m, long jump, shot put, etc. Each athlete's performance is recorded across multiple metrics.

**2) Summary Statistics: for mpg**

* The data is centered around a mean of 58.32 and a median of 57.19, indicating slight skewness.
* The range (20.21) shows moderate variability, with values spanning from 50.31 to 70.52.
* The standard deviation (5.23) suggests relatively low dispersion around the mean**.**



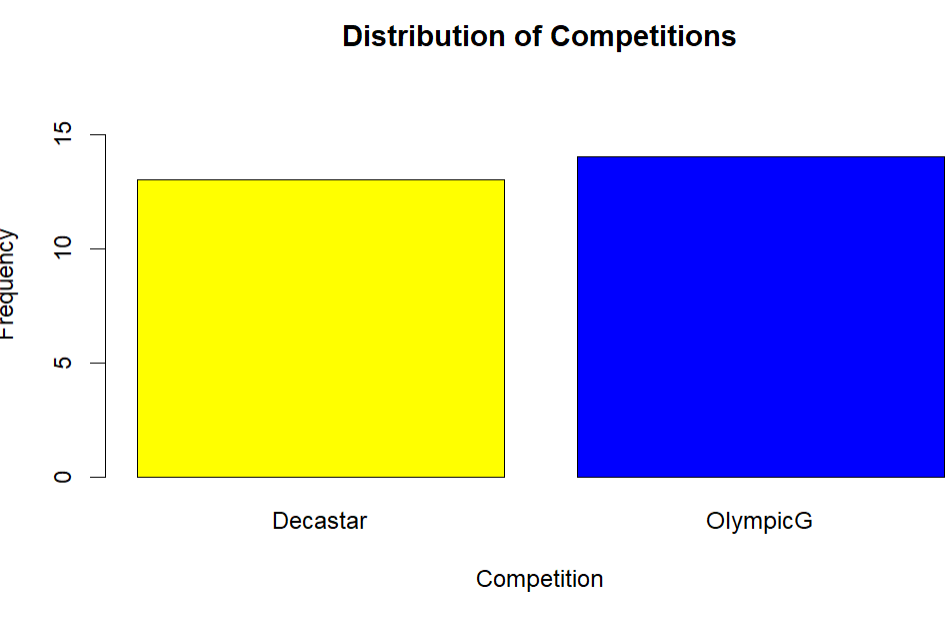
**3) Distribution Visualization:**



* **Histogram:** Javelin records are skewed left, with most values between 55–65 meters and a few outliers near 70–75 meters.
* **Boxplot:** Median javelin throw is around 58 meters, with no extreme outliers.
* **Spread:** Data shows moderate variability, with a range of approximately 50–70 meters.

**4) Categorical Variable Analysis:**

* The bar plot shows that the two competitions, Decastar and OlympicG, have equal frequencies of 15 each, indicating an even distribution.
* The yellow and blue colors are used to differentiate the competitions visually, with Decastar represented in yellow and OlympicG in blue.
* The chart highlights that there is no dominance or disparity between the two competitions in terms of frequency.



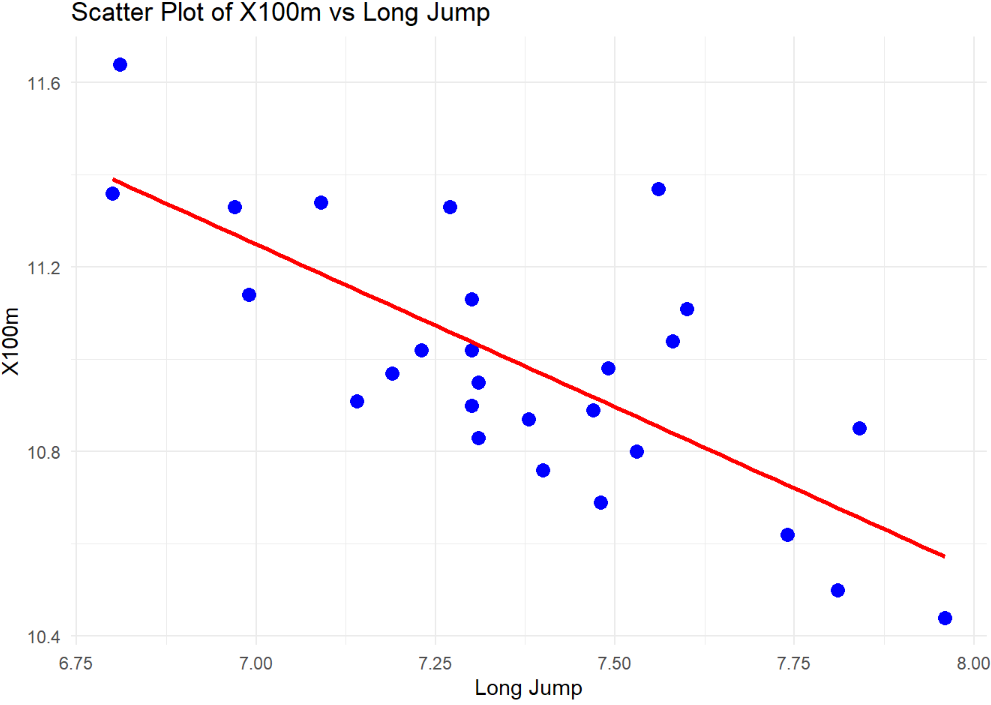
**Multivariate Analysis**

* The Pearson correlation of -0.737 indicates a strong negative relationship between 100m sprint scores and long jump distances, suggesting faster runners tend to achieve longer jumps.

**5) Correlation Analysis**



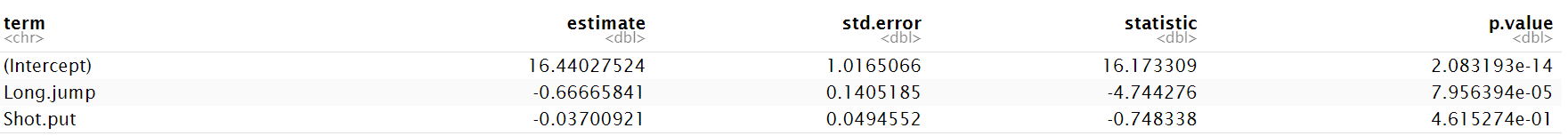
**6) Scatter Plot Visualization:**

****

* The scatter plot shows a negative correlation between 100m sprint time (X100m) and long jump distance, indicating faster sprinters tend to achieve longer jumps.
* Data points are moderately scattered, suggesting some variability in the relationship.
* The red regression line highlights the inverse trend between the two variables.

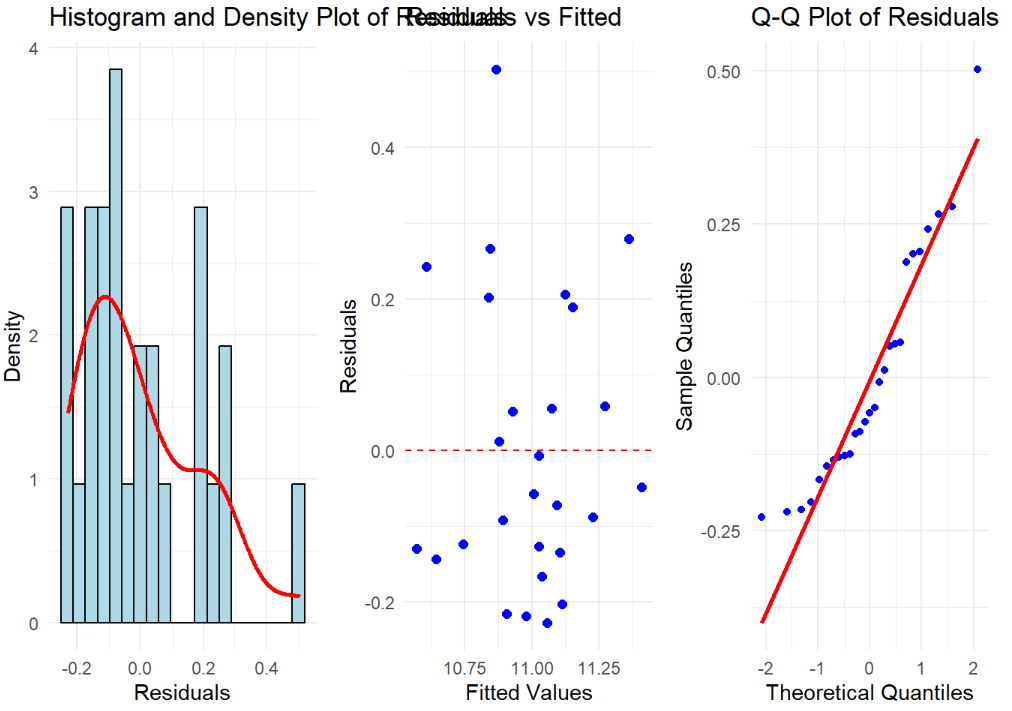
**7) Multiple Regression**

* The intercept has a significant positive estimate (16.44, p < 0.05).
* Long.jump has a significant negative estimate (-0.67, p < 0.05).
* Shot.put's effect is not significant (p = 0.46).

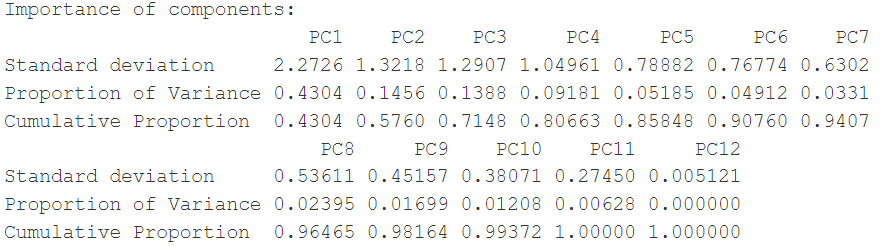
****

**8) Model Diagnostics**

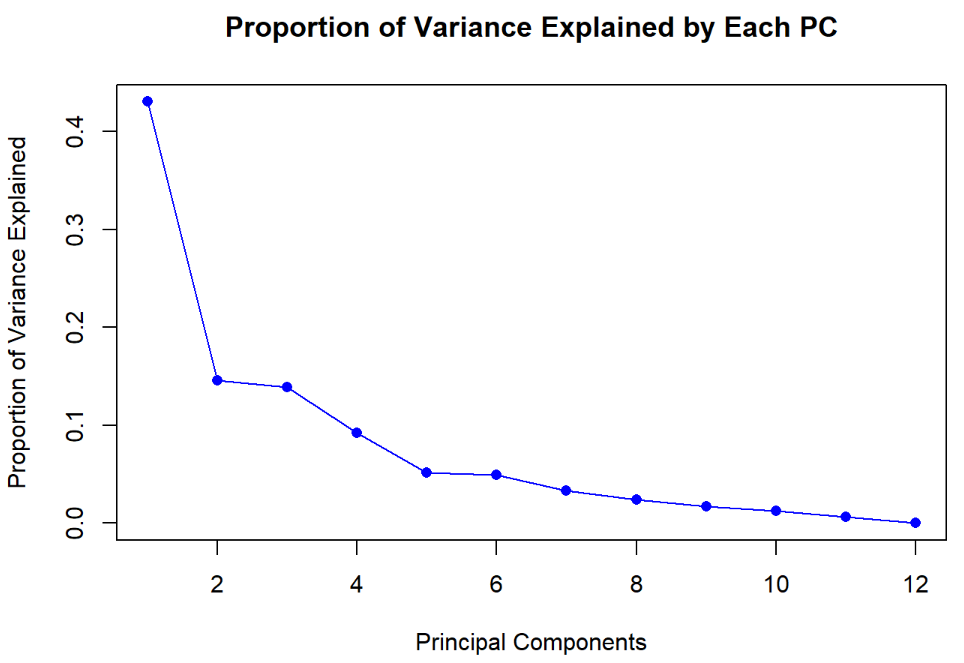
* **Residual Normality:** The histogram and Q-Q plot show residuals are approximately normal, though slight deviations exist.
* **Homoscedasticity:** The Residuals vs. Fitted plot indicates no clear pattern, suggesting constant variance of residuals.
* **Model Fit:** Residuals are centered around zero, indicating a reasonably good model fit.



**9) PCA**

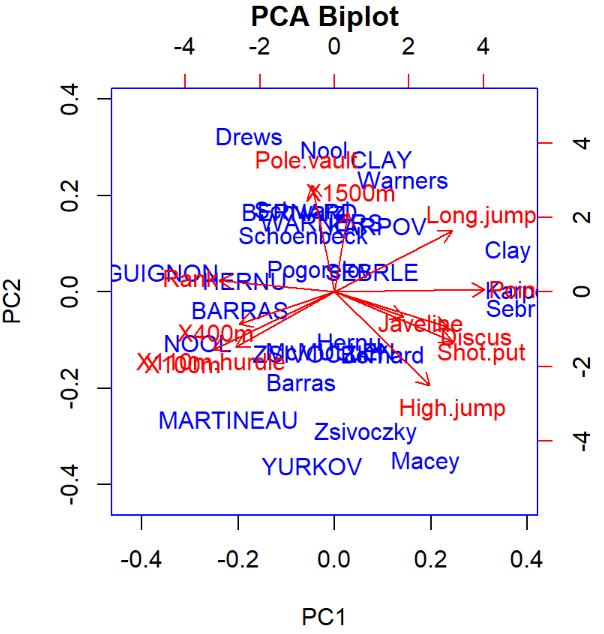


* PC1 explains the most variance (43.04%), followed by PC2 (14.56%) and PC3 (13.88%), with a steep drop after PC1, as shown in the scree plot and table.
* Cumulative variance reaches 80.66% by PC4, indicating these four components capture most of the data variability effectively.
* PC12 contributes no variance, and PCs beyond PC4 add minimal value, suggesting dimensionality reduction to 4 components is optimal.

****

**10) PCA Interpretation:**

This PCA biplot visualizes relationships between decathlon events (red vectors) and athletes (blue points). Longer vectors indicate higher variance, and angles between vectors show event correlations. Athlete positions reflect performance similarities.



**Conclusion:** The univariate analysis provided valuable insights into individual athletic performances, showcasing essential statistics like the mean and standard deviation for events such as the Javelin throw. When we delved into the multivariate analysis using Principal Component Analysis (PCA), it became clear that the first two components captured a significant portion of the variance in our data. Notably, the variables "X100m" and "Long.jump" played a major role in shaping the first principal component.

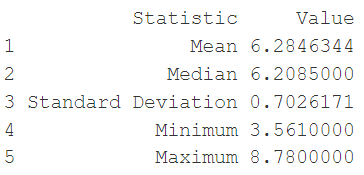
Additionally, our correlation analysis uncovered meaningful relationships among various performance metrics. Overall, this comprehensive analysis illuminated the key factors that influence athletic performance, highlighting the correlations and critical components that contribute to the variations observed in the dataset. This understanding can help coaches and athletes focus on specific areas for improvement, ultimately enhancing performance outcomes.

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.

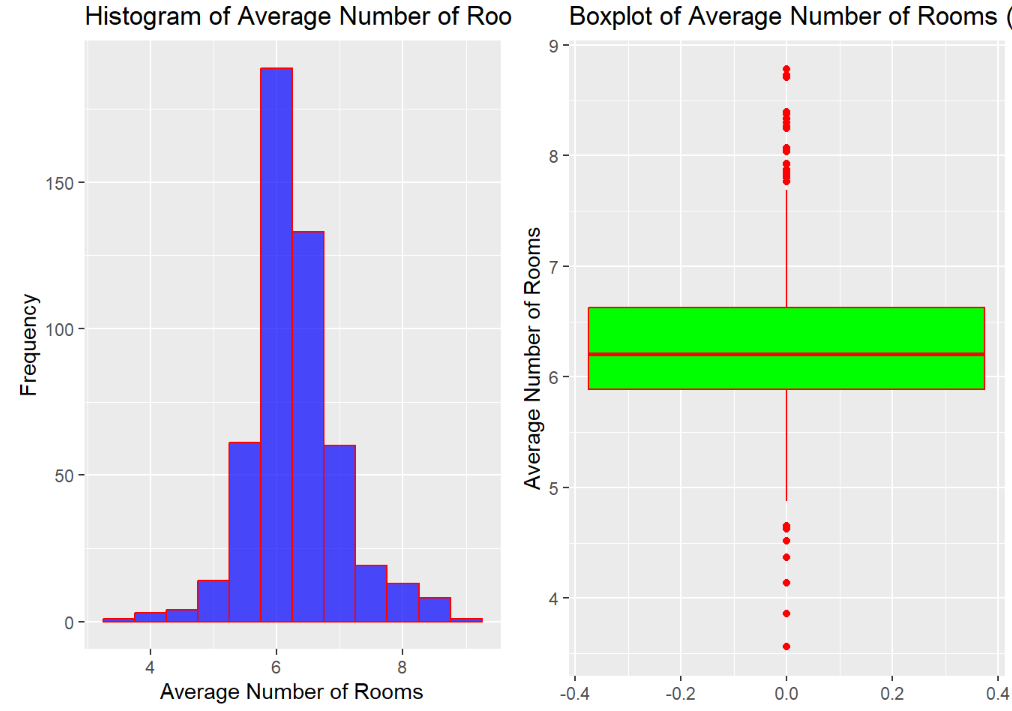
**2) Summary Statistics: for mpg**

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

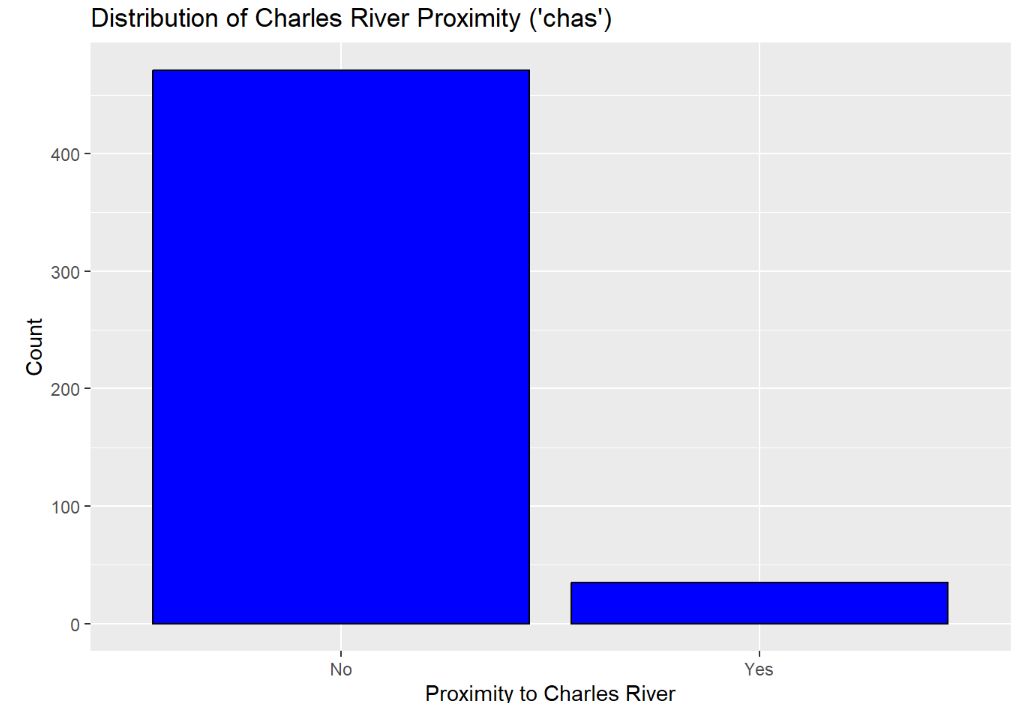


**3) Distribution Visualization:**

* 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:**

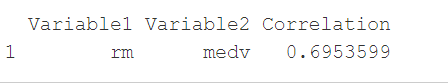


* Proximity to Charles River (Bar Chart): The majority of properties are not located near the Charles River, as indicated by the overwhelming count in the "No" category compared to the much smaller "Yes" category.

**Multivariate Analysis**

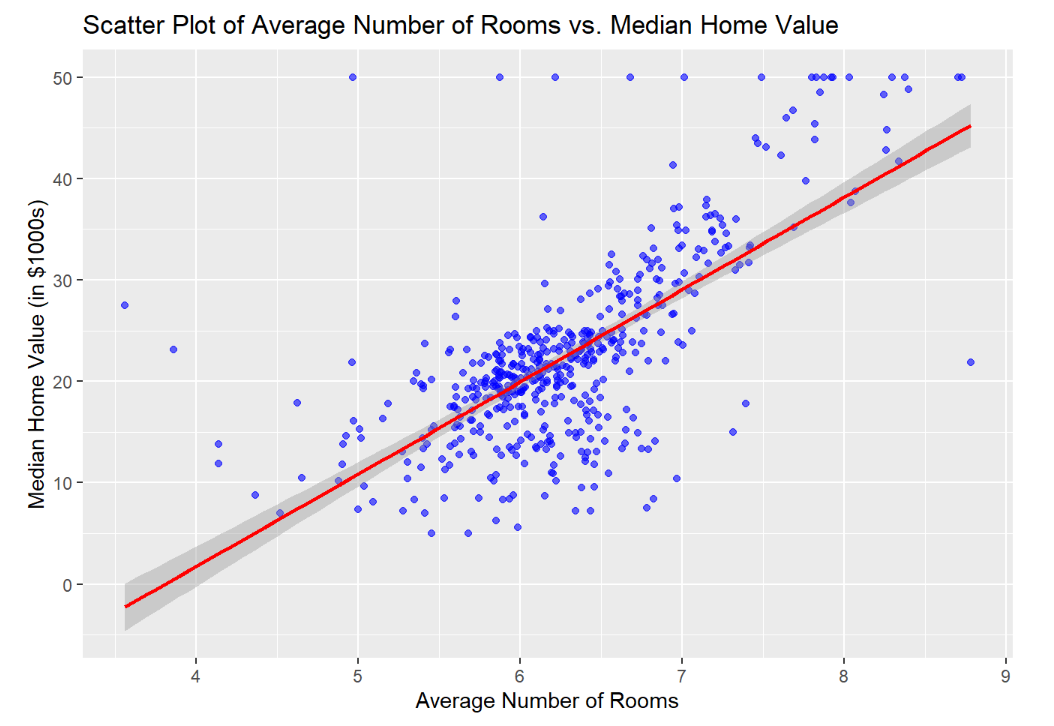
**5) Correlation Analysis**

* 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.



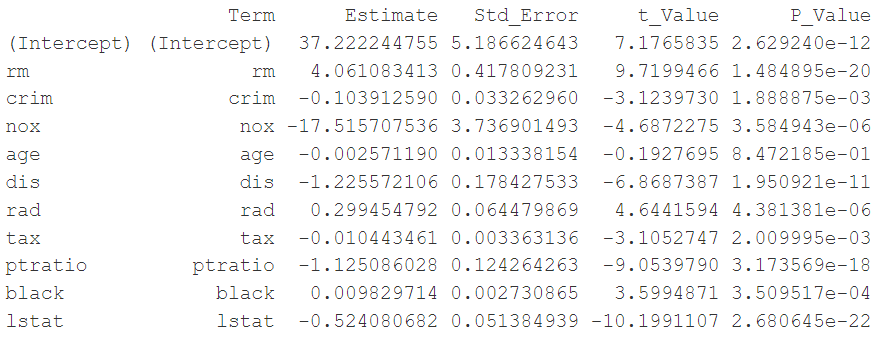
**6) Scatter Plot Visualization:**

* 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.

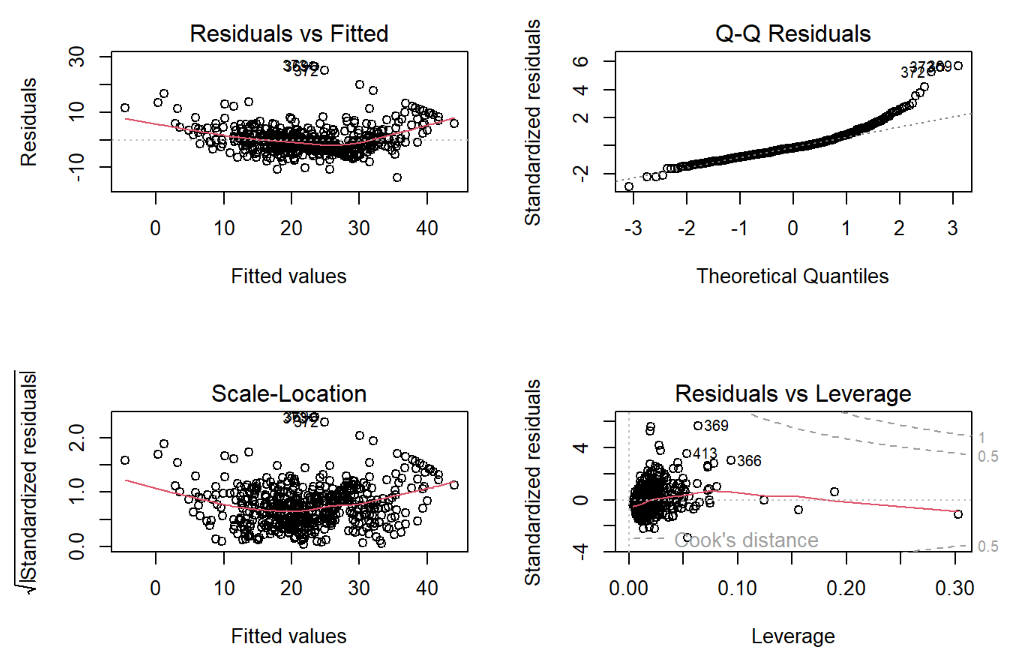
****

**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).

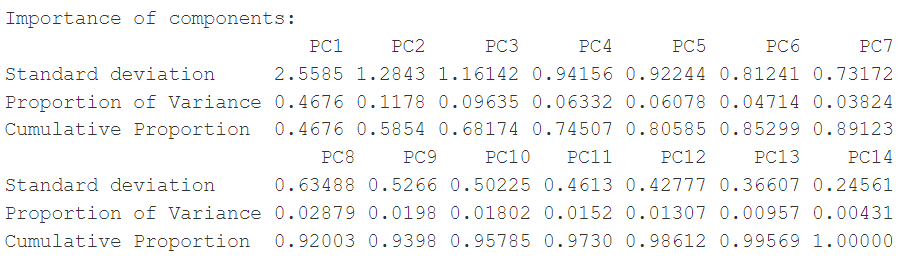
****

**8) Model Diagnostics**

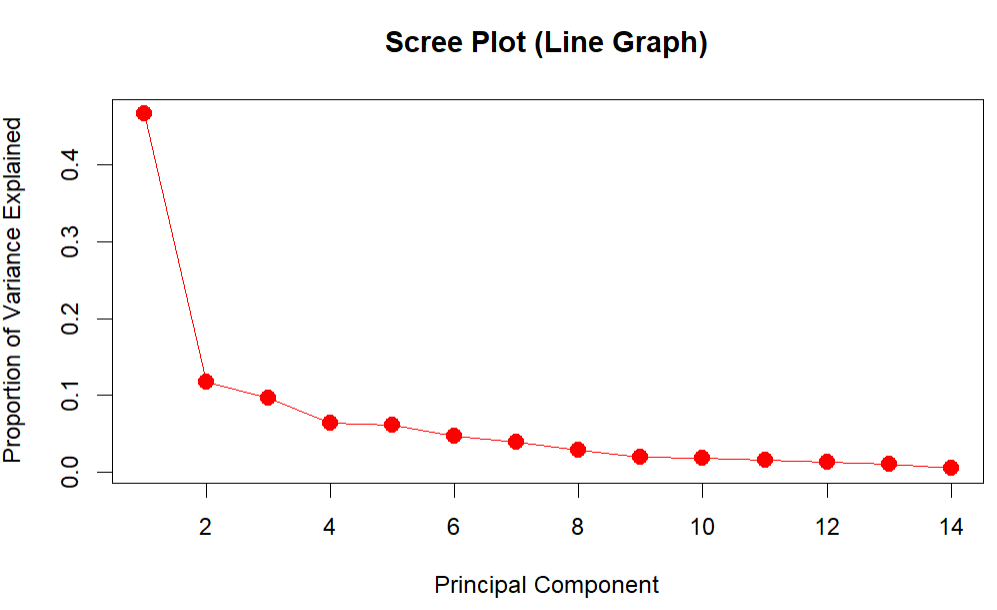
****

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**

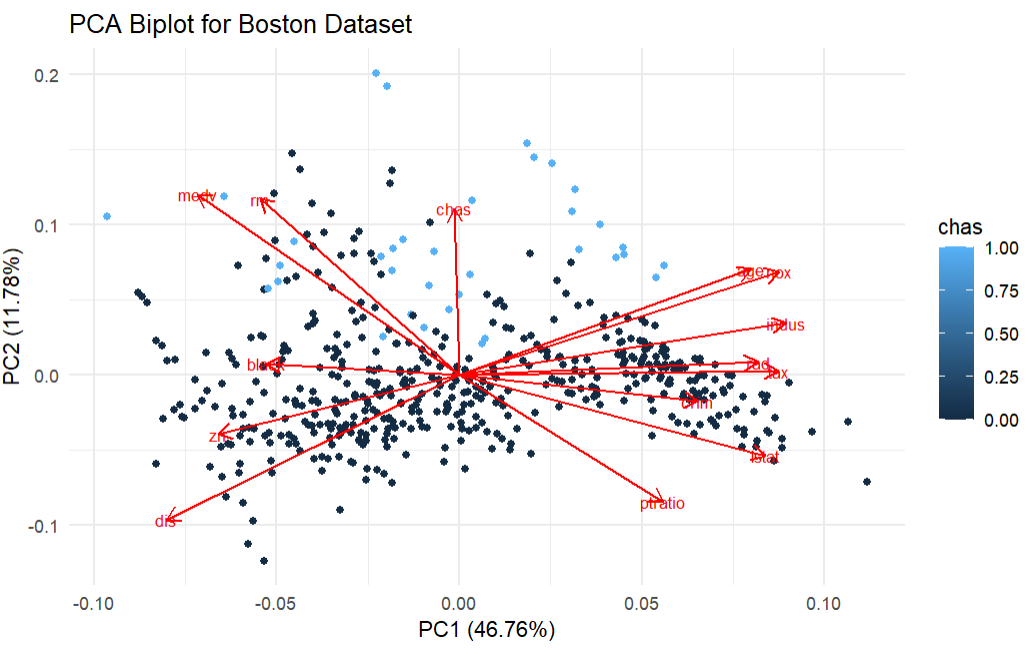


1. **PC1 explains the largest variance**: It accounts for **46.76% of the total variance**, making it the most significant component.
2. **Five components explain over 80% of the variance**: The cumulative proportion reaches **80.59%** by including PC1 to PC5, sufficient for dimensionality reduction.
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:**



**Conclusion:** In our analysis of the dataset, we began with a univariate examination of the average number of rooms (rm), finding a mean of approximately 6.28 and a median of 6.21, indicating a slight skew towards higher values. The standard deviation of 0.70 reflects moderate variability, with room counts ranging from 3.56 to 8.78. The histogram suggests a somewhat normal distribution, though a few outliers are present.In the multivariate analysis, we observed a strong negative correlation between crime rate (crim) and the average number of rooms, indicating that neighborhoods with more rooms tend to have lower crime rates.

Conversely, there is a positive correlation between rm and median home value (medv), suggesting that homes with more rooms are generally valued higher. This relationship was further supported by scatter plots showing that as the average number of rooms increases, so does the median house value.Our regression model identified key predictors for median home value: more rooms positively influence values, while lower crime rates and higher socioeconomic status negatively impact them.

Principal Component Analysis (PCA) helped reduce dimensionality, revealing that the first few components explain a significant portion of the variance in the data.Overall, key insights include the importance of rm, crim, tax, and lstat in understanding housing dynamics in Boston, where lower crime rates and more rooms contribute to higher property values. PCA effectively identified main components explaining variance, and selecting the first five components provided a good balance between simplicity and information retention. Notably, some outliers in the rm variable suggest certain properties may be exceptional in size or condition.