INDIAN STATISTICAL INSTITUTE

POST-GRADUATE DIPLOMA IN BUSINESS ANALYTICS (PGDBA): 2024–26

Course: Statistical Structures in Data

Numerical Assignment Report

Submitted by:

Sankalp Davi

Roll No: 24BM6JP48

Submitted to:

Professor Subhajit Dutta

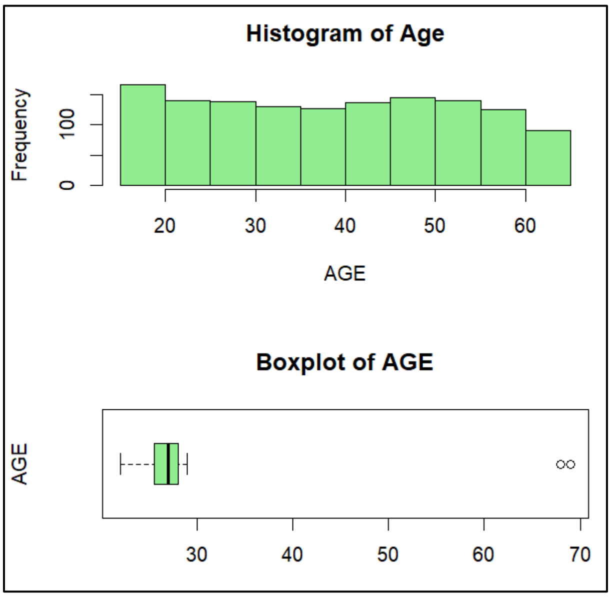
DATASET 1 – Medical Insurance Data

1. Data Overview

- Total Number of observations: **1338**
- Total Number of variables: **7**
- Full Structure of the data is presented in R code file.

2. Summary Statistics of Data

- Mean (Average): **39.20703**
- Median**: 39 -** the median is close to the mean, this suggests the data distribution is relatively symmetric, though not perfectly so.
- Standard Deviation**: 14.04996 -** Most individuals' ages will lie within ±13 years of the mean (approximately 25–51 years).
- Minimum**: 18**
- Maximum**: 64**

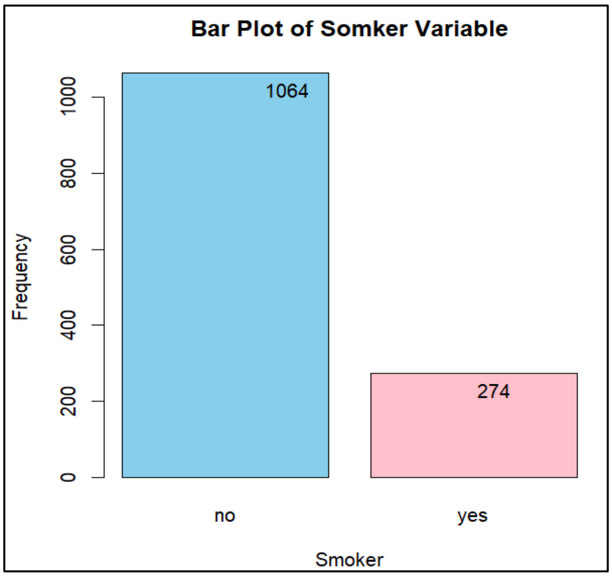


3. Distribution Visualization

- The distribution is more balanced with minor **right** skewness, and there are no extreme values.
- The distribution of age appears **multi-modal**.
- The box plot likely shows a symmetric or slightly skewed distribution with no significant outliers.

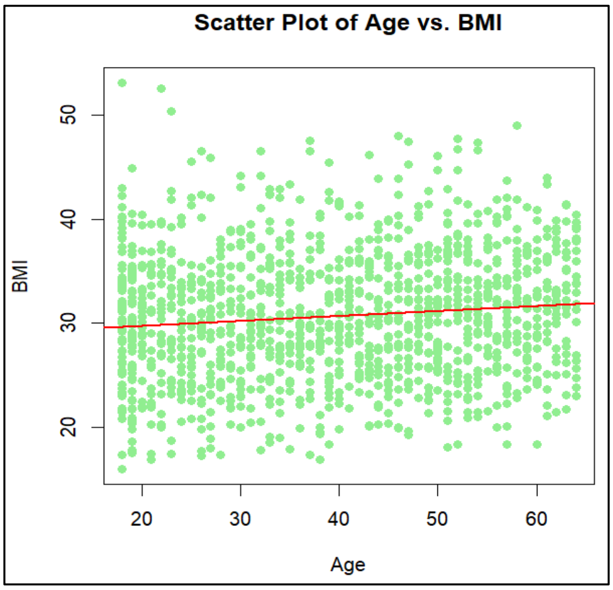
4. Categorical Variable Analysis

- The bar plot shows the distribution of the **Smoker** variable, with two categories: **No** (non-smokers) and **Yes** (smokers).
- The majority of the dataset consists of non-smokers, with a frequency of **1064**, significantly outnumbering smokers while there are **274** smokers in the dataset.



5. Correlation Analysis

- **Pearson correlation coefficient** between Age and BMI is **0.1092719**, which is very close to **0**, which indicates that there is a **weak, positive relationship** between Age and BMI.
- This suggests that age may have a very minimal effect on BMI and other factors might be contributing more to variations in BMI than age alone.



6. Scatter Plot Visualization

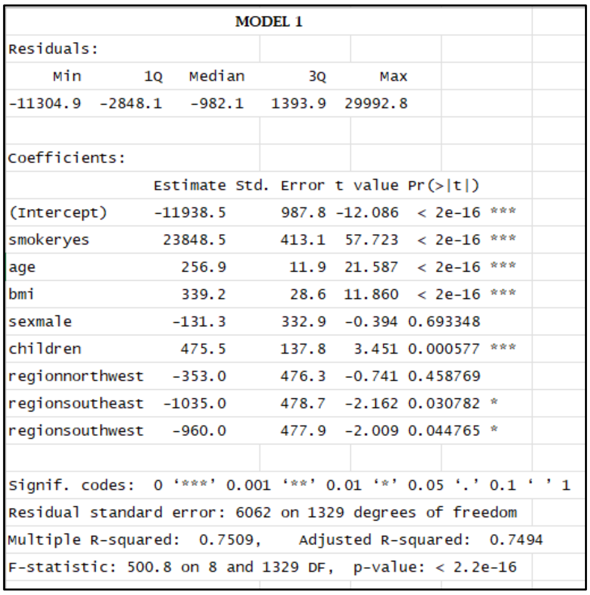
- The points are widely spread, indicating that there isn't a very tight linear relationship between age and BMI.
- The red trend line suggests a slight **positive relationship** between age and BMI, meaning that as age increases, BMI tends to increase slightly as well.

- However, the trend is very subtle, which is in line with the weak correlation.

7. Multiple Regression

Model 1

mlr_model <- lm (charges ~ smoker + age + bmi + sex + children + region, data = dataset1)

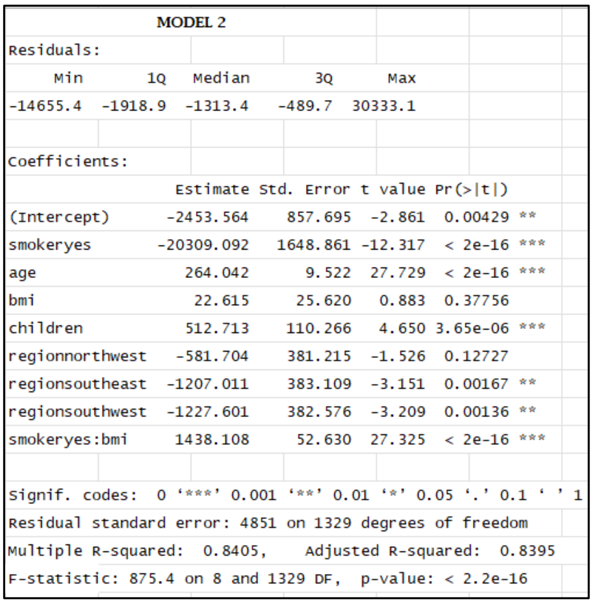


```
MODEL 1
Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -11938.5      987.8 -12.086  < 2e-16 ***
smokeryes         23848.5      413.1  57.723  < 2e-16 ***
age                 256.9       11.9  21.587  < 2e-16 ***
bmi                 339.2       28.6  11.860  < 2e-16 ***
sexmale            -131.3      332.9  -0.394 0.693348
children            475.5      137.8   3.451 0.000577 ***
regionnorthwest    -353.0      476.3  -0.741 0.458769
regionsoutheast   -1035.0      478.7  -2.162 0.030782 *
regionsouthwest    -960.0      477.9  -2.009 0.044765 *

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

- The residuals show the distribution of the errors in the model. The values range from -11304.9 to 29992.8, indicating a widespread in the errors.
- Intercept: The base cost when all predictors are zero is -11938.5.
- Residual standard error: The average error is 6062 on 1329 degrees of freedom.
- R-squared: The model explains 75.09% of the variance in the response variable.
- F-statistic: The overall model is significant with a p-value < 2.2e-16.
- When looking at the coefficients we see that sex is not significant. So, Let’s remove “sex” from the model.

Model 2

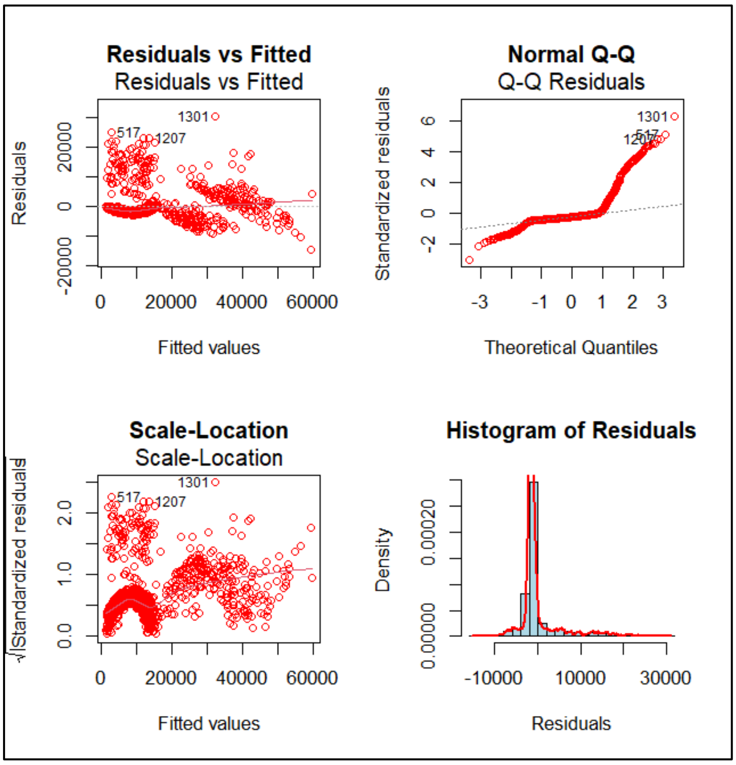
*mlr_model <- lm(charges ~ smoker + age^2 + bmi + children + region + smoker*bmi, data = dataset1)*



```
MODEL 2
Residuals:
     Min       1Q   Median       3Q      Max
-14655.4  -1918.9  -1313.4   -489.7  30333.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2453.564    857.695  -2.861  0.00429 **
smokeryes       -20309.092   1648.861 -12.317  < 2e-16 ***
age                264.042      9.522  27.729  < 2e-16 ***
bmi                 22.615     25.620   0.883  0.37756
children           512.713    110.266   4.650 3.65e-06 ***
regionnorthwest   -581.704    381.215  -1.526  0.12727
regionsoutheast  -1207.011    383.109  -3.151  0.00167 **
regionsouthwest  -1227.601    382.576  -3.209  0.00136 **
smokeryes:bmi     1438.108     52.630  27.325  < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4851 on 1329 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8395
F-statistic: 875.4 on 8 and 1329 DF,  p-value: < 2.2e-16
```

- The residuals show the distribution of errors in the model. The values range from -14655.4 to 30333.1, indicating a widespread in prediction errors.
- Intercept: The estimated intercept is -2453.564, meaning the expected insurance cost without any other factors is -2453.564.
- Residual Standard Error: The RSE is 4851, indicating the average error in predicting insurance costs.
- R-squared: The R-squared value is 0.8405, meaning 84.05% of the variance in insurance costs is explained by the model.
- F-statistic: The F-statistic is 875.4 with a p-value < 2.2e-16, indicating that the model is statistically significant.

8. Model Diagnostics



Homoscedasticity:

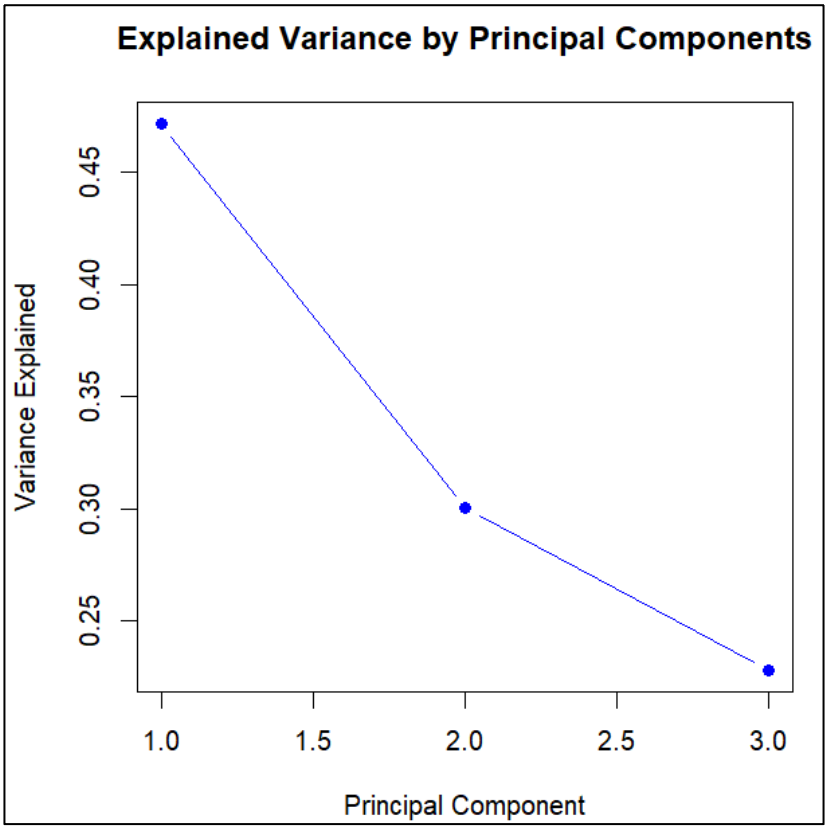
- The Residuals vs. Fitted plot shows a clear pattern (e.g., a fan shape or curvature), indicating heteroscedasticity, meaning the residuals do not have constant variance.
- The Scale-Location plot confirms heteroscedasticity with non-random spread.

Normality of residuals:

- The Normal Q-Q Plot shows significant deviation from the diagonal line, especially at the tails. This suggests the residuals are not normally distributed.
- The Histogram of Residuals shows a skewed or non-normal distribution of residuals, further confirming non-normality.
- The model may not be well-fitted and may require adjustments, such as transforming variables.

9. Principal Component Analysis (PCA)

- Based on the scree plot, I would choose two principal components
- The explained variance shows a sharp drop after the second component.
- This is consistent with the "elbow rule," where the components before the elbow capture the most variance, and subsequent components contribute marginally.



10. PCA Interpretation

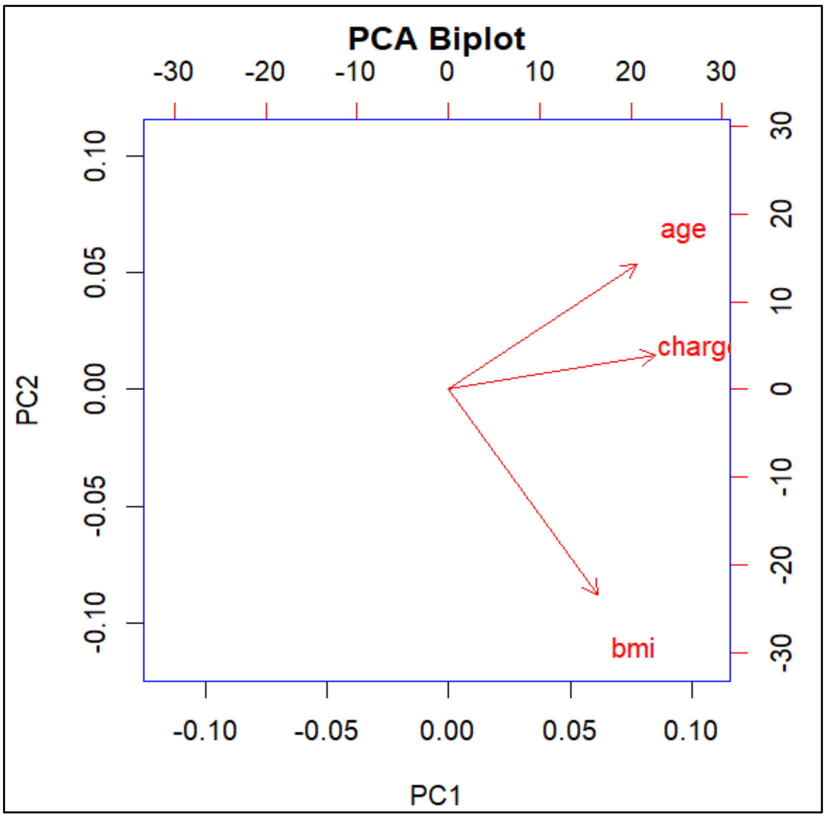
This PCA biplot shows the loadings of the first two principal components (PC1 and PC2) for the variables *age*, *charges*, and *BMI*.

The primary variability in the dataset (PC1) seems to be driven by age and charges, which are strongly aligned with this axis.

- PC1: Variables like charges and age are positively correlated with PC1 (vectors pointing in the same direction and close alignment with the PC1 axis).
- BMI shows a slightly negative correlation with PC1.

The second component (PC2) captures variation mostly attributed to BMI, which has a distinct relationship compared to age and charges.

- PC2: The BMI variable has a stronger loading along the negative direction of PC2.
- Age and charges are weakly associated with PC2, as their vectors are closer to horizontal.



DATASET 2 – Chick Weight Data

1. Data Overview

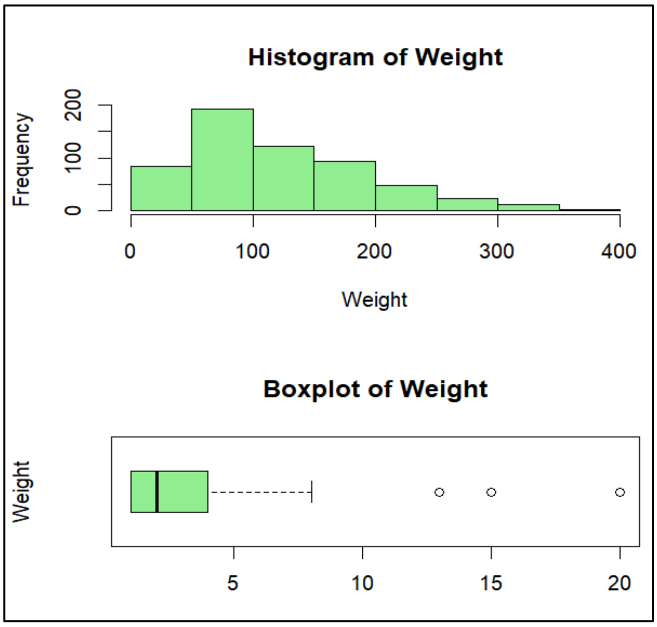
- Total Number of observations: 578
- Total Number of variables: 4
- Structure of the data is presented in R code file.

2. Summary Statistics of Data

- Mean (Average): **121.8183**
- Median**: 103**
- Standard Deviation**: 71.071**
- Minimum**: 35**
- Maximum**: 373**

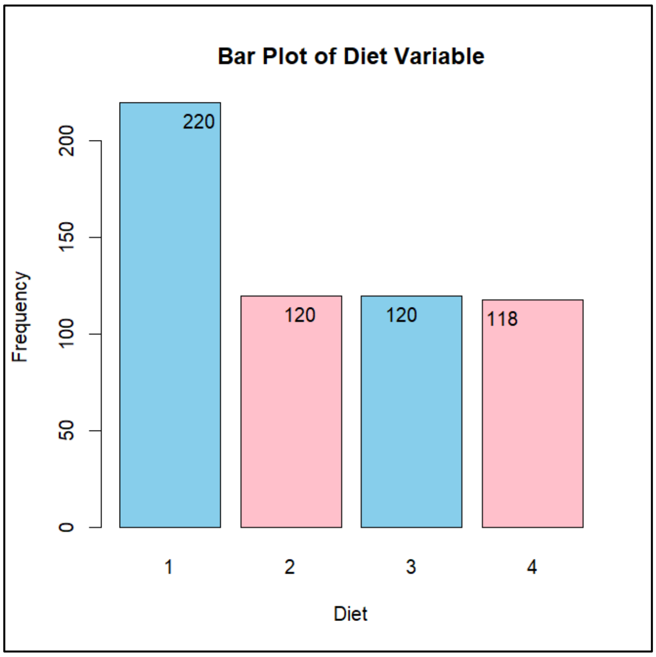
3. Distribution Visualization

- The distribution is more balanced with **right** skewness, and there are no extreme values.
- The distribution of Weight appears **uni-modal**.
- The box plot likely shows a symmetric or slightly skewed distribution with some significant outliers.



4. Categorical Variable Analysis

- The bar plot shows the distribution of the **Diet** variable, with four categories: 1,2,3,4 indicating type of diets for Chicken.
- The majority of the dataset consists of Diet 1, with a frequency of **220**, while the remaining diets having close to equal proportion.

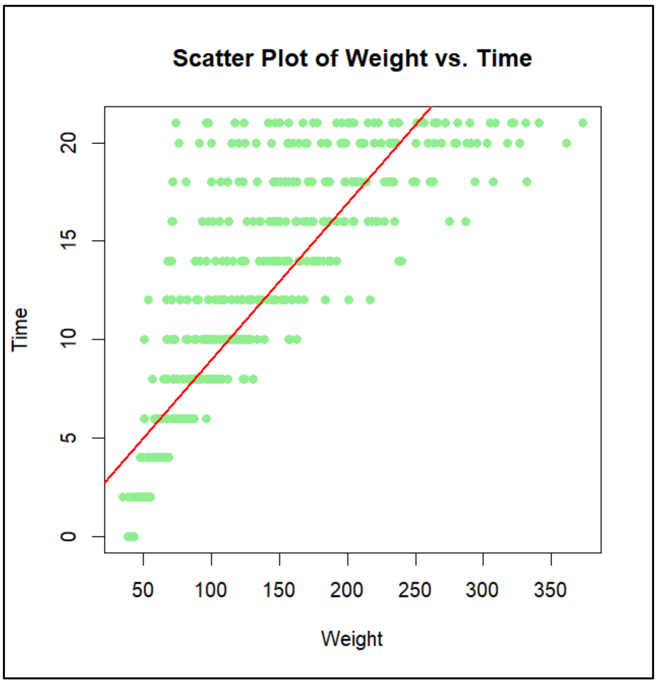


5. Correlation Analysis

- **Pearson correlation coefficient** between Age and BMI is **0.8371017**, which is close to **1**, which indicates that there is a **strong, positive relationship** between Weight and Time.

6. Scatter Plot Visualization

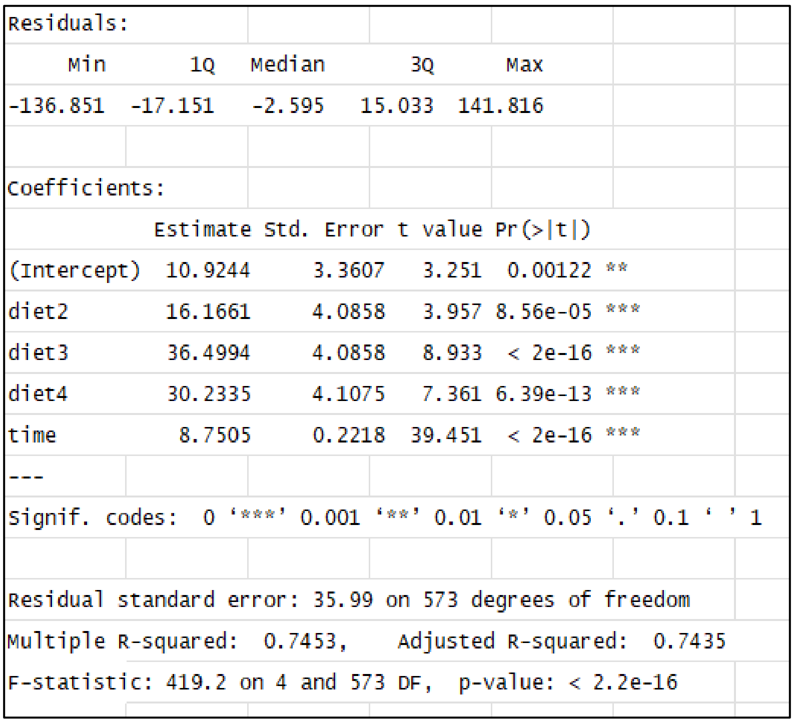
- The points are spread at specific Time interval, indicating that there is very tight linear relationship between Weight and Time.
- The red trend line suggests a **positive relationship** between Weight and Time, meaning that as Weight increases, Time tends to increase as well.
- However, the trend is very clear, which is in line with the strong correlation.



7. Multiple Regression

Summary of the Model is displayed in R code

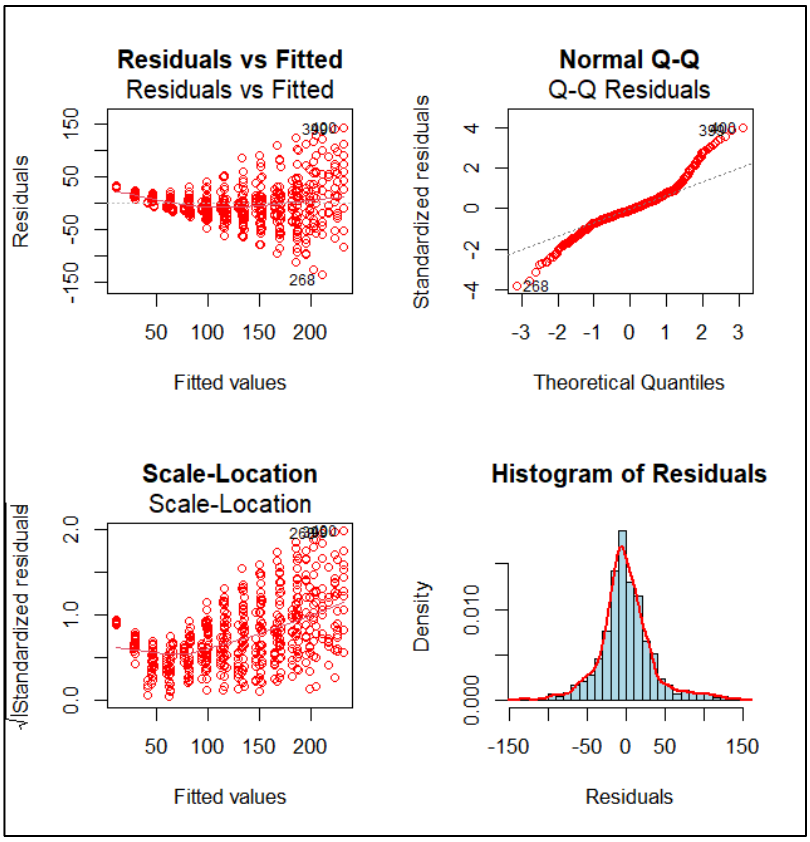
mlr_model <- lm (weight ~ diet + time, data = dataset2)



```
Residuals:
     Min       1Q   Median       3Q      Max 
-136.851  -17.151   -2.595   15.033  141.816 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.9244     3.3607   3.251  0.00122 ** 
diet2        16.1661     4.0858   3.957 8.56e-05 ***
diet3        36.4994     4.0858   8.933  < 2e-16 ***
diet4        30.2335     4.1075   7.361 6.39e-13 ***
time          8.7505     0.2218  39.451  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.99 on 573 degrees of freedom
Multiple R-squared:  0.7453,	Adjusted R-squared:  0.7435 
F-statistic: 419.2 on 4 and 573 DF,  p-value: < 2.2e-16
```

- The residuals show the distribution of errors in the model. The values range from -136.85 to 141.816, indicating a widespread in prediction errors.
- Intercept: When diet, diet2, diet3, diet4, and time are all zero, the predicted value of the response variable is 10.9244. meaning the expected weight without any other factors is 10.9244
- For a one-unit increase in `diet2`, we expect the weight to increase by 16.1661 units on average.
- A one-unit increase in `diet3` is associated with an increase of 36.4994 units in the weight,
- A one-unit increase in `diet4` leads to an increase of 30.2335 units in the weight. holding other variables constant.
- A one-unit increase in `time` is associated with an increase of 8.7505 units in the weight.
- R-squared: The R-squared value is 0.7453, meaning 74.53% of the variance in weight is explained by the model.
- F-statistic: The F-statistic is 419.2 with a p-value < 2.2e-16, indicating that the model is statistically significant.

8. Model Diagnostics



Homoscedasticity

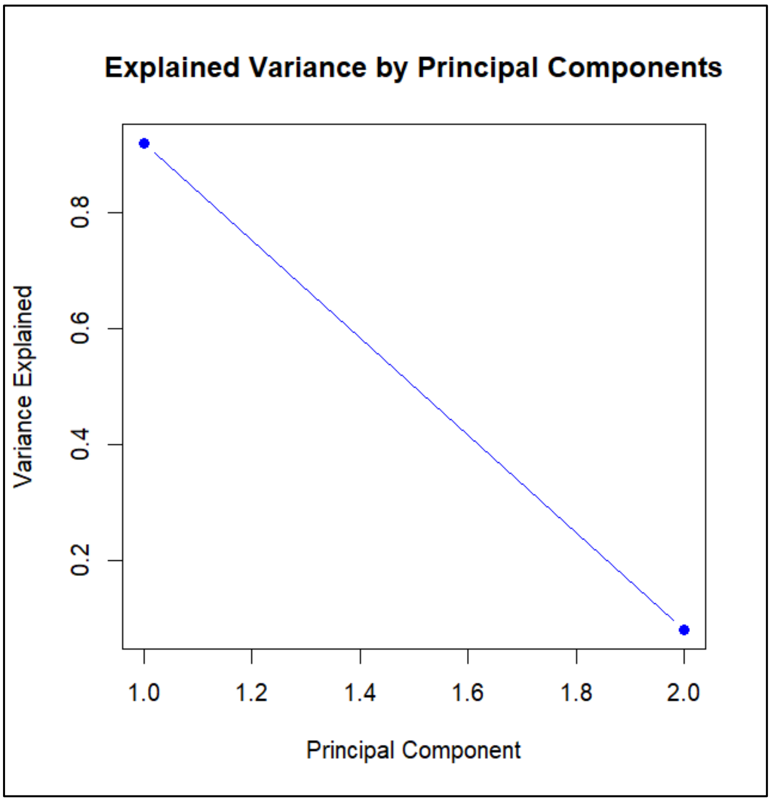
- Residuals vs Fitted: This plot shows some slight curvature, indicating that the variance of the residuals increases with the fitted values. This suggests potential heteroscedasticity.
- Scale-Location: This plot is like the "Residuals vs Fitted" plot, we see a similar pattern as in the "Residuals vs Fitted" plot, suggesting potential heteroscedasticity.

Normality of Residuals

- Normal Q-Q: We see some deviation from the straight line, especially in the tails, indicating that the residuals might not be perfectly normally distributed.
- Histogram of Residuals: The histogram appears slightly skewed, which again suggests that the residuals might not be perfectly normally distributed.
- Applying transformations to the response or predictor variables might help stabilize the variance and improve normality.

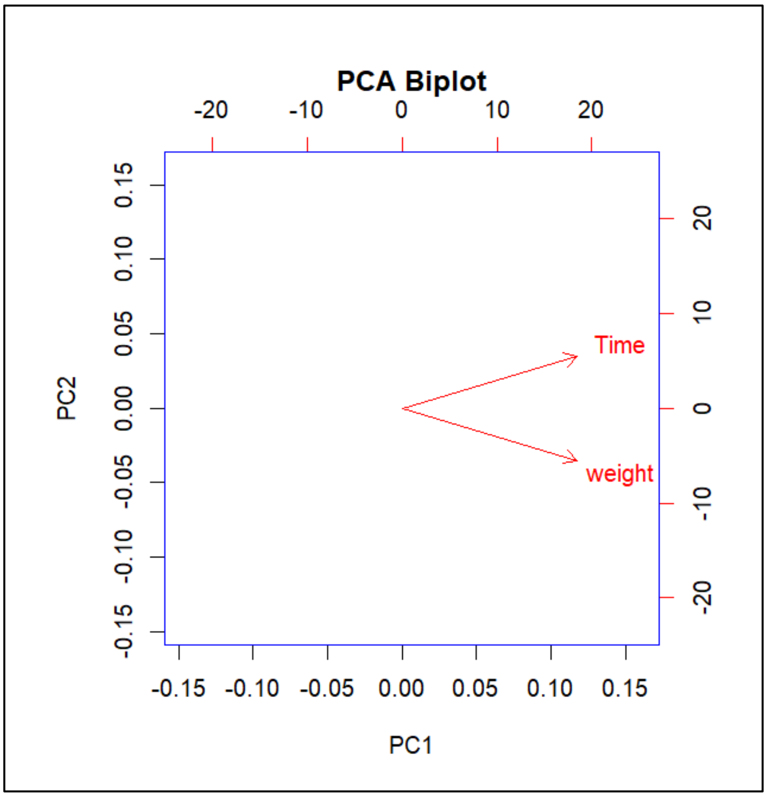
9. PRINCIPAL COMPONENT ANALYSIS (PCA)

- Based on the scree plot, I would choose one principal components.
- In this case, the scree plot has a distinct elbow at the 1st principal component. This suggests that the first two components capture most of the variance in the data.
- Adding more components beyond the first would result in diminishing returns.



10. PCA INTERPRETATION

- The biplot shows the loadings of the first two principal components (PC1 and PC2) for the variables "Time" and "Weight". The direction and length of the arrows indicate the contribution of each variable to the corresponding principal component.
- The arrow for "Time" points towards the positive direction of PC1. This suggests that "Time" has a positive correlation with PC1. As "Time" increases, the scores on PC1 tend to increase as well.
- The arrow for "Weight" points towards the negative direction of PC1. This indicates a negative correlation between "Weight" and PC1. As "Weight" increases, the scores on PC1 tend to decrease.
- The relative positions of the arrows for "Time" and "Weight" suggest that these variables are positively correlated.



DATASET 3 – Mtcars

1. Data Overview

- Total Number of observations: **32**
- Total Number of variables: **11**
- Full Structure of the data is presented in R code file.

2. Summary Statistics of Data

- Mean (Average): **20.09**
- Median: **- 19.2**
- Standard Deviation: **6.026**
- Minimum: **10.4**
- Maximum: **33.9**

3. Distribution Visualization

- The distribution is more balanced with minor **right** skewness, and there are no extreme values.
- The distribution of mpg appears **uni-modal**.
- The boxplot of mpg doesn't show any clear outliers. The boxplot is relatively symmetrical, indicating that the distribution is not heavily skewed.

4. Categorical Variable Analysis

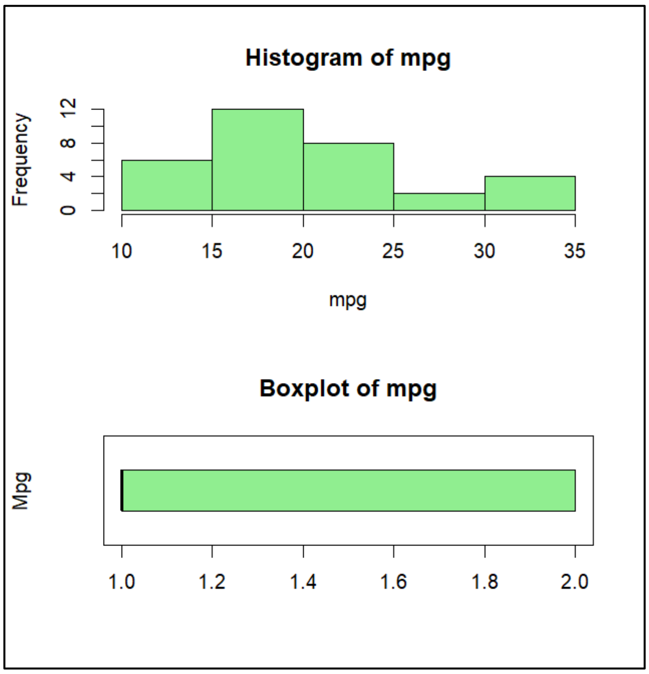
- The bar plot shows the distribution of the **no. of gears** variable, with three categories: 3, 4 & 5.
- The bar plot shows that cars with 3 gears are the most common, followed by 4-gear cars. 5-gear cars are the least common. The distribution is skewed towards lower gear numbers.

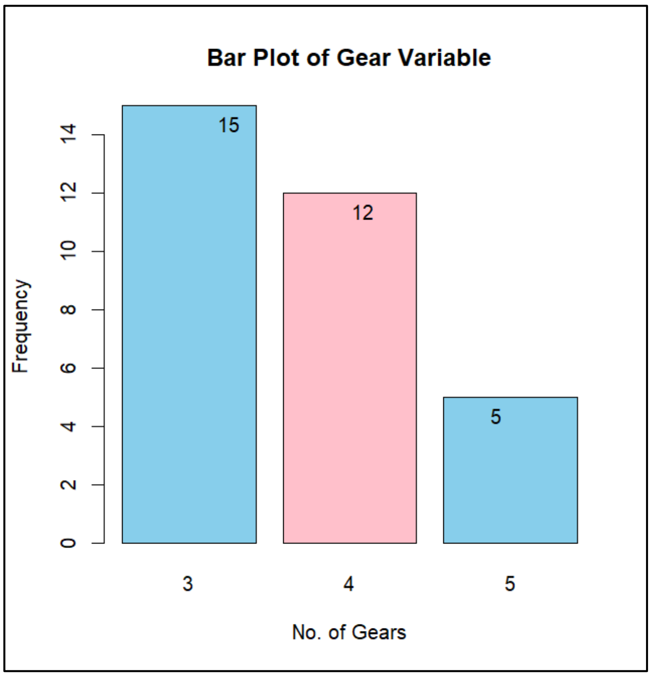
5. Correlation Analysis

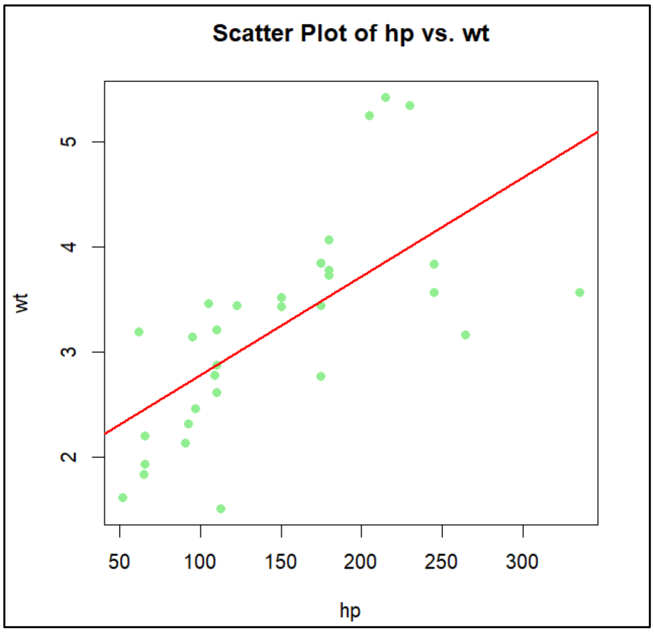
- **Pearson correlation coefficient** between hp and wt is **0.6587479**, which indicates that there is a **positive relationship** between hp and wt.
- This suggests that hp may have an effect on wt.

6. Scatter Plot Visualization

- The upward trend of the data points indicates a positive correlation between hp and wt.
- The data points seem to follow a linear pattern, suggesting that the relationship between hp and wt can be approximated by a straight line.
- However, the data points are not perfectly aligned, indicating some degree of variability in the relationship.

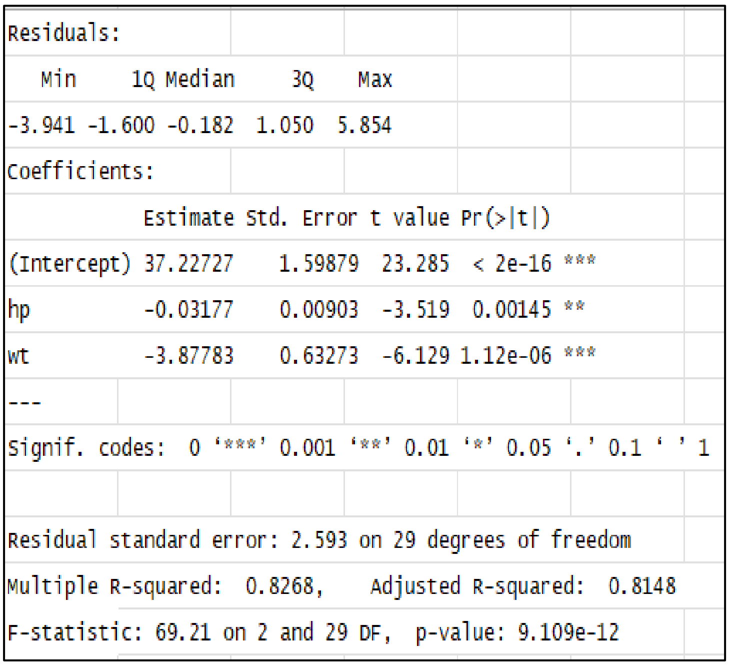






7. Multiple Regression

mlr_model <- lm (mpg ~ hp + wt, data = dataset3)



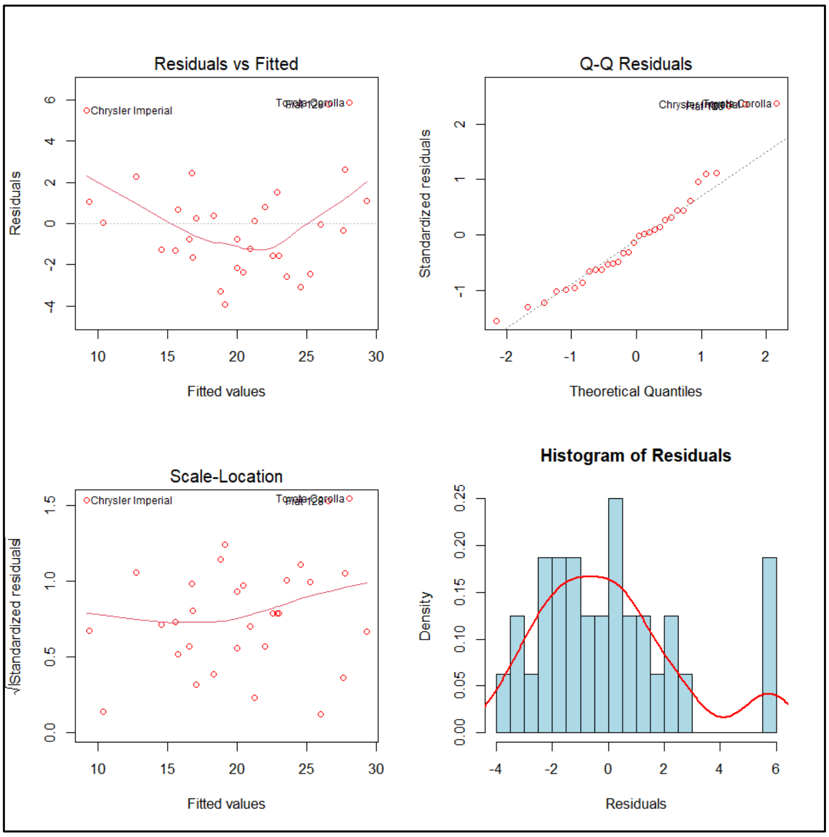
```
Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
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
```

- The residuals show the distribution of the errors in the model. The values range from -3.941 to 5.854, indicating a small spread in the errors.
- Intercept: The base mpg when all predictors are zero is -37.227. In other words, if a car has zero horsepower and zero weight, its predicted mpg is 37.22727.
- hp: The coefficient for hp is -0.03177. This means that for every one-unit increase in horsepower (hp), the predicted mpg decreases by 0.03177 units, holding weight constant.
- wt: The coefficient for wt is -3.87783. This means that for every one-unit increase in weight (wt), the predicted mpg decreases by 3.87783 units, holding horsepower constant.
- Residual standard error: The average error is 2.593 on 29 degrees of freedom.
- R-squared: The model explains 82.68% of the variance in the response variable.
- F-statistic: The overall model is significant with a p-value < 9.109 e-12.

8. Model Diagnostics

Homoscedasticity

- Residuals vs Fitted: We see some slight curvature, indicating that the variance of the residuals might be increasing with the fitted values. This suggests potential heteroscedasticity.
- Scale-Location: We see a similar pattern as in the "Residuals vs Fitted" plot, suggesting potential heteroscedasticity.

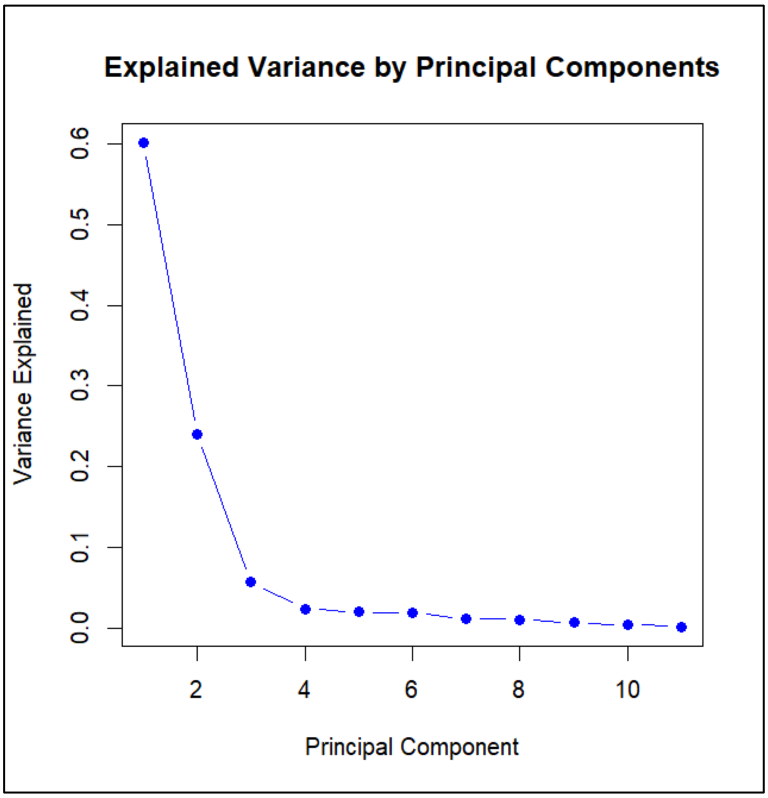


Normality of Residuals

- Normal Q-Q: Here see some deviation from the straight line, especially in the tails, indicating that the residuals might not be perfectly normally distributed.
- Histogram of Residuals: The histogram appears slightly skewed, which again suggests that the residuals might not be perfectly normally distributed.

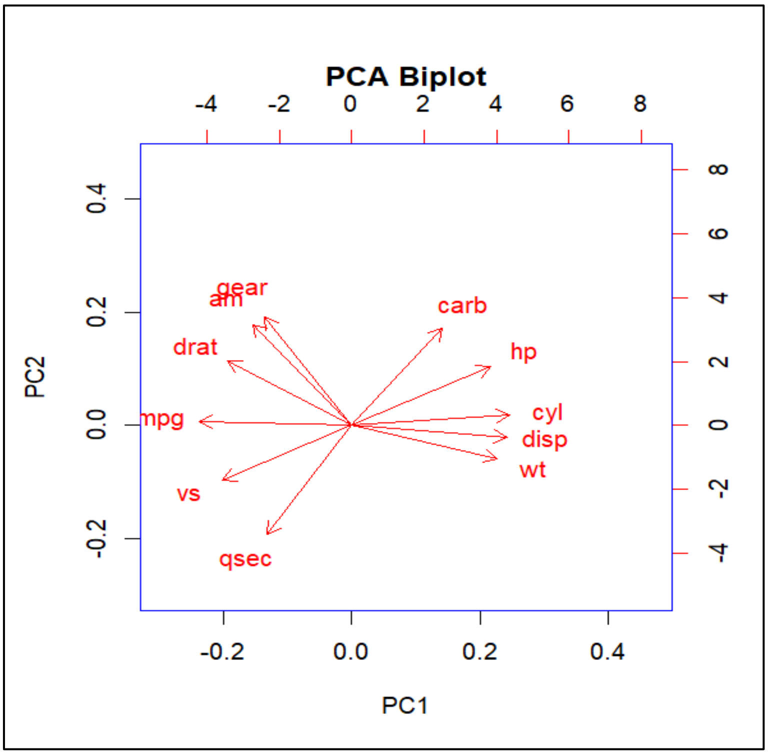
9. Principal Component Analysis (PCA)

- Based on the scree plot, I would choose three principal components
- The explained variance shows a sharp drop after the third component.
- This is consistent with the elbow rule, where the components before the elbow capture the most variance, and subsequent components contribute marginally.
- As here after the third component subsequent components contribute minimally to the explained variance.



10. PCA Interpretation

- The biplot shows the loadings of the first two principal components (PC1 and PC2) for the different variables. The direction and length of the arrows indicate the contribution of each variable to the corresponding principal component.
- PC1: Variables like hp, disp, wt, and cyl have arrows pointing in the positive direction of PC1. This suggests that these variables are positively correlated with PC1. On the other hand, variables like mpg, qsec, vs, and drat have arrows pointing in the negative direction of PC1. This suggests a negative correlation with PC1.
- PC1 Represents a performance dimension, with high values indicating high horsepower, displacement, weight, and cylinder count.
- PC2: Variables like gear, am, and carb have arrows pointing in the positive direction of PC2. This suggests a positive correlation with PC2. As these variables increase, the scores on PC2 tend to increase.
- PC2 Represents a transmission dimension, with high values indicating more gears, automatic transmission, and more carburettors.



DATASET 4 – Iris

1. Data Overview

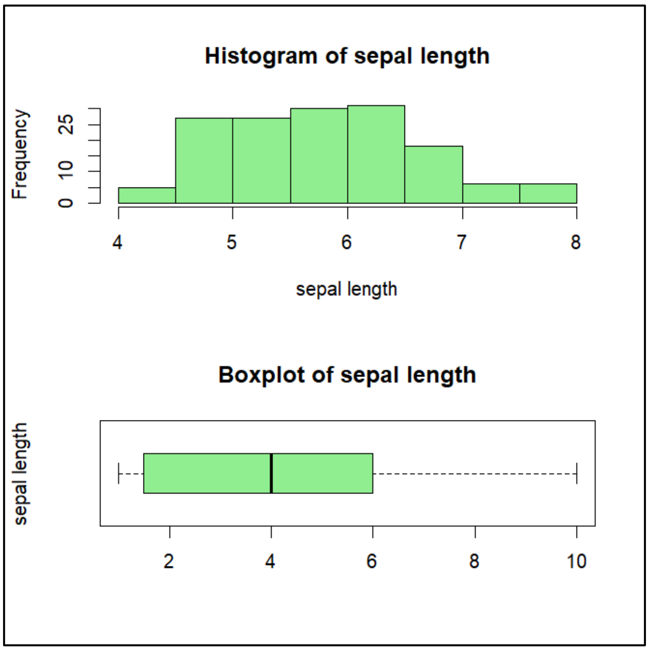
- Total Number of observations: 150
- Total Number of variables: 5
- Full Structure of the data is presented in R code file.

2. Summary Statistics of Data

- Mean (Average): **5.843**
- Median**: - 5.8**
- Standard Deviation**: 0.828**
- Minimum**: 4.3**
- Maximum**: 7.9**

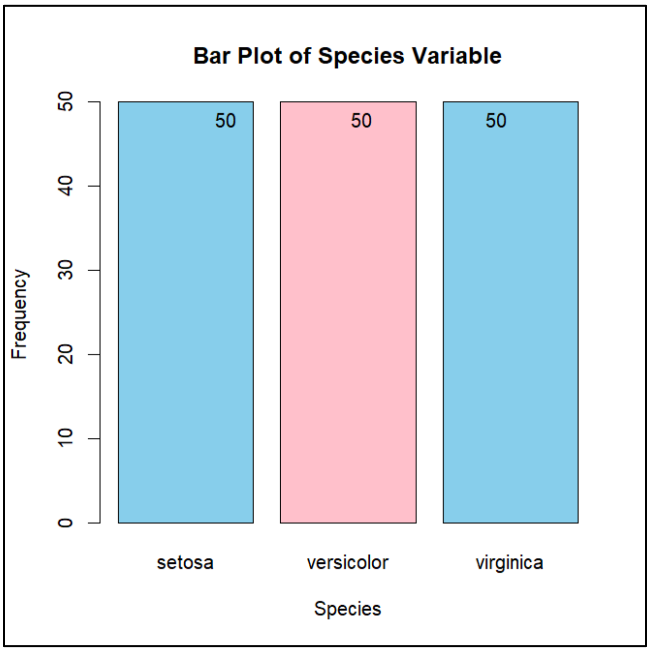
3. Distribution Visualization

- The distribution is slightly right-skewed distributed, and there are not many extreme values.
- The distribution of sepal length appears **uni-modal**.
- The boxplot of sepal length doesn't show any clear outliers. , the long whisker on the right side suggests that there might be some data points with higher sepal length values that are farther away from the majority of the data.



4. Categorical Variable Analysis

- The bar plot shows the distribution of the **Species** variable, with three categories: setosa, versicolor, and virginica.
- Each species has 50 observations, resulting in a balanced dataset.

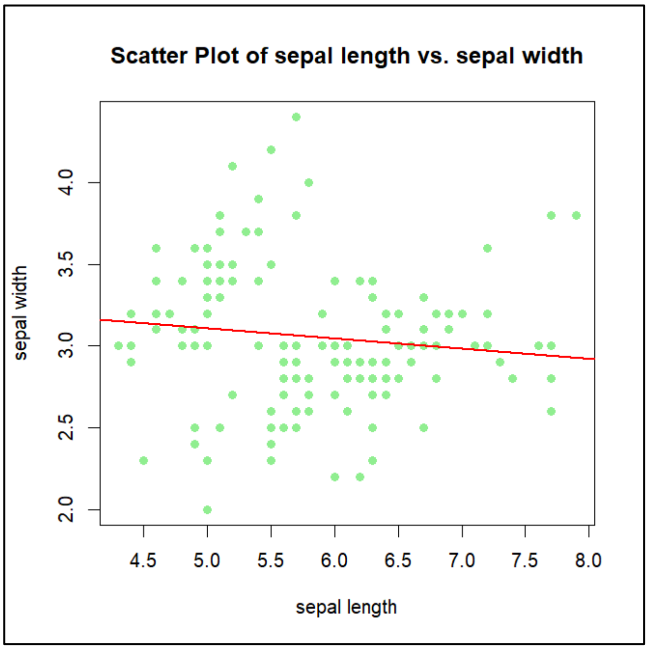


5. Correlation Analysis

- **Pearson correlation coefficient** between Sepal length and Sepal width is **-0.11756**, which is very close to **0**, which indicates that there is a **weak, negative relationship**.
- This suggests that Sepal length may have a very minimal effect on Sepal width.

6. Scatter Plot Visualization

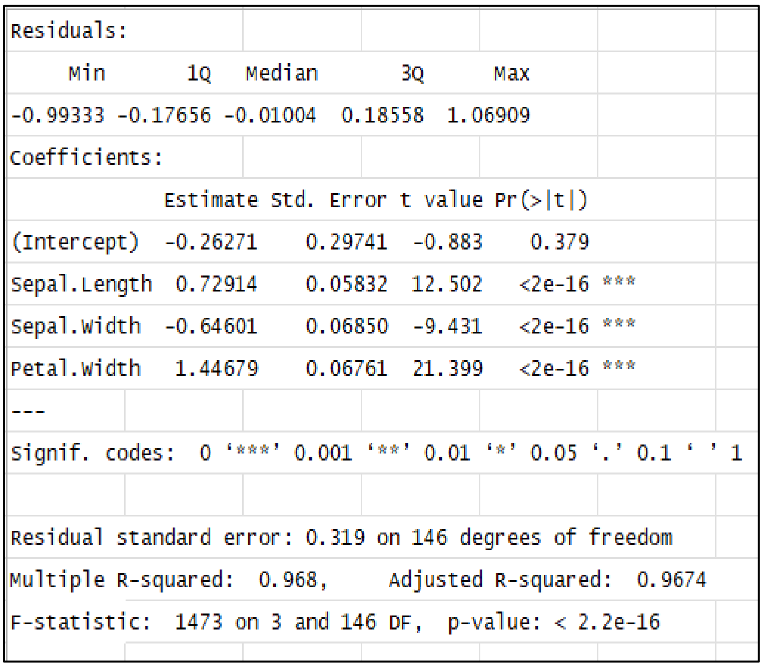
- The slight downward trend of the data points indicates a weak negative correlation between sepal length and sepal width.
- The data points seem to follow a linear pattern, suggesting that the relationship between sepal length and sepal width can be approximated by a straight line.
- The data points are quite scattered, indicating a significant amount of variability in the relationship.



7. Multiple Regression

Summary of the Model is displayed in R code

mlr_model <- lm (Petal.Length ~ Sepal.Length + Sepal.Width + Petal.Width, data = dataset4)

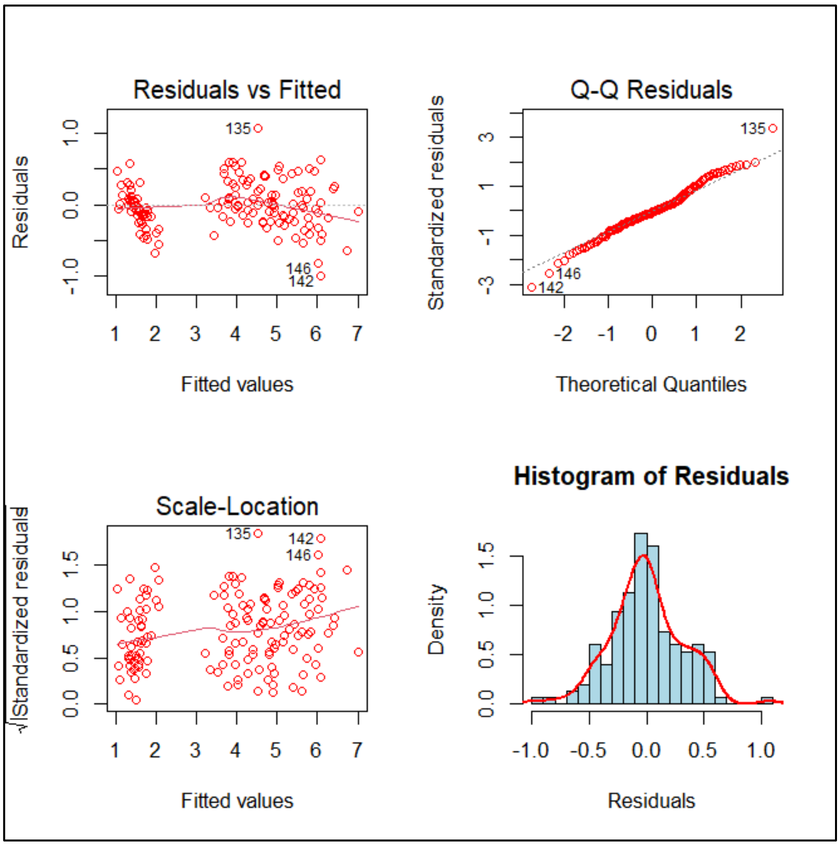


```
Residuals:
     Min       1Q   Median       3Q      Max
-0.99333 -0.17656 -0.01004  0.18558  1.06909
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.26271    0.29741  -0.883    0.379
Sepal.Length  0.72914    0.05832  12.502   <2e-16 ***
Sepal.Width  -0.64601    0.06850  -9.431   <2e-16 ***
Petal.Width   1.44679    0.06761  21.399   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.319 on 146 degrees of freedom
Multiple R-squared:  0.968,     Adjusted R-squared:  0.9674
F-statistic:  1473 on 3 and 146 DF,  p-value: < 2.2e-16
```

The residuals show the distribution of the errors in the model. The values range from -0.993 to 1.069, indicating a very small spread in the errors.

- Intercept: The base cost when all predictors are zero is -0.26271.
- Sepal.Length: The coefficient for sepal length is 0.72914. This means that for every one-unit increase in sepal length, the predicted sepal width increases by 0.72914 units, holding other variables constant.
- Sepal.Width: The coefficient for sepal width is -0.64601. This means that for every one-unit increase in sepal width, the predicted sepal length decreases by 0.64601 units, holding other variables constant.
- Petal.Width: The coefficient for petal width is 1.44679. This means that for every one-unit increase in petal width, the predicted sepal length increases by 1.44679 units, holding other variables constant.
- Residual standard error: The average error is 0.319 on 146 degrees of freedom.
- R-squared: The model explains 96.8% of the variance in the response variable.
- F-statistic: The overall model is significant with a p-value < 2.2e-16.

8. Model Diagnostics



Homoscedasticity

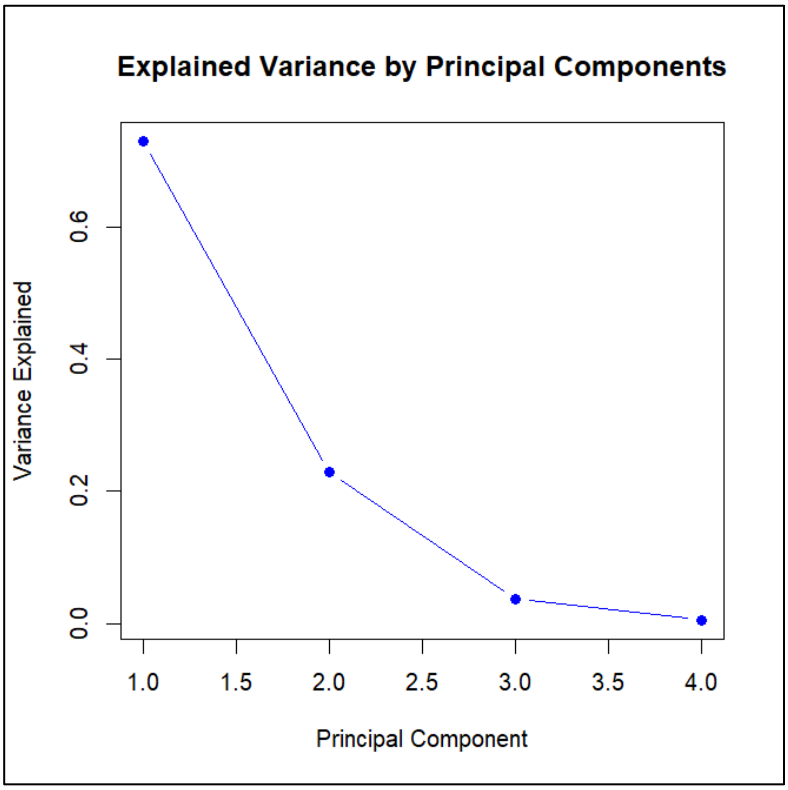
- Residuals vs Fitted: We see some slight curvature, indicating that the variance of the residuals might be increasing with the fitted values. This suggests potential heteroscedasticity.
- Scale-Location: We see a similar pattern as in the "Residuals vs Fitted" plot, suggesting potential heteroscedasticity.

Normality of Residuals

- Normal Q-Q: Here, the points deviate slightly from the straight line, especially in the tails, indicating that the residuals might not be perfectly normally distributed.
- Histogram of Residuals: In this case, the histogram appears slightly skewed, which again suggests that the residuals might not be perfectly normally distributed.
- The model may not be well-fitted and may require adjustments, such as transforming variables.

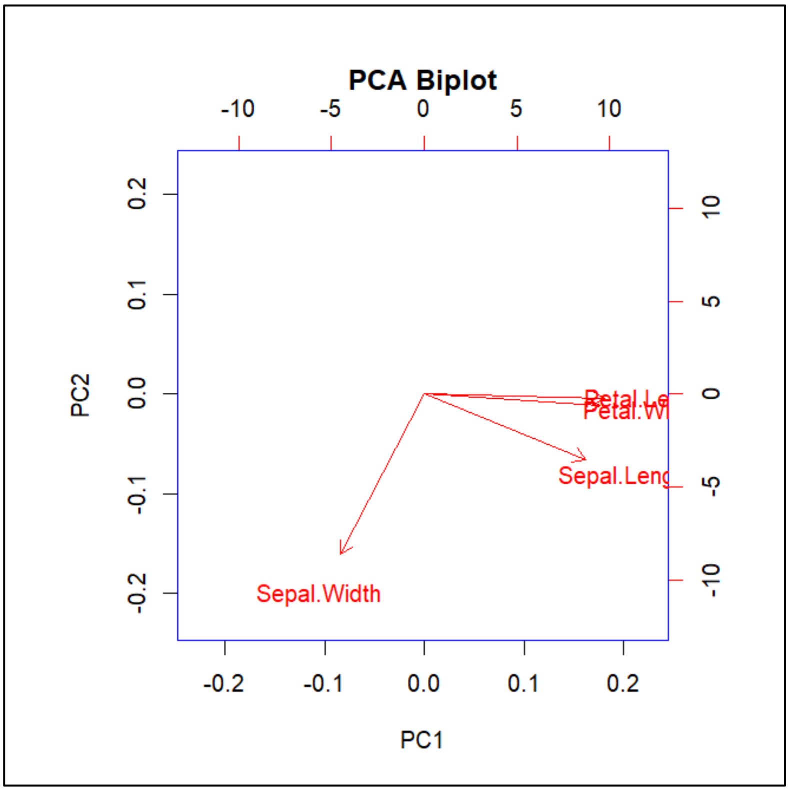
9. Principal Component Analysis (PCA)

- Based on the scree plot, I would choose two principal components
- The explained variance shows a sharp drop after the second component.
- This is consistent with the "elbow rule," where the components before the elbow capture the most variance, and subsequent components contribute marginally.



10. PCA Interpretation

- PC1: Variables like Petal.Length and Petal. Width has arrows pointing in the positive direction of PC1. This suggests that these variables are positively correlated with PC1
- On the other hand, variables like Sepal.Length and Sepal.Width have arrows pointing in the negative direction of PC1. This suggests a negative correlation with PC1.
- PC1 Represents a dimension related to petal size, with high values indicating larger petals.
- PC2: All variables have relatively short arrows along PC2, indicating that PC2 captures less variation in the data compared to PC1.
- PC2 Captures less variation and doesn't show a clear grouping of variables.
- Overall, the biplot suggests that the first two principal components capture two main dimensions of variation in the data:


PCA Biplot
PC2
Petal.Le
Petal.Wi
Sepal.Leng
Sepal.Width
PC1