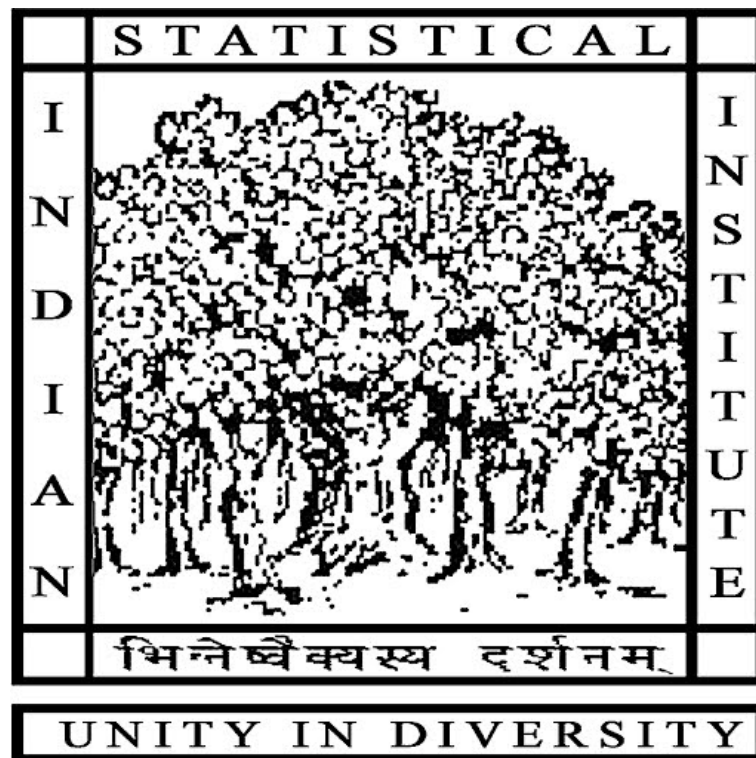


DECEMBER 8, 2024



NUMERICAL ASSIGNMENT STATISTICAL STRUCTURES IN DATA

INSTRUCTOR: DR SUBHAJIT DUTTA

HEMRAJ CHAKRAVARTI

24BM6JP22

PGDBA -Batch 10

Univariate Analysis Dataset 1

1. Data Overview

The mtcars dataset contains data about car models and their attributes. It comprises 32 observations and 11 variables.:

- Numerical: mpg (miles per gallon), hp (horsepower), wt (weight), etc.
- Categorical: cyl (number of cylinders), gear (number of gears).

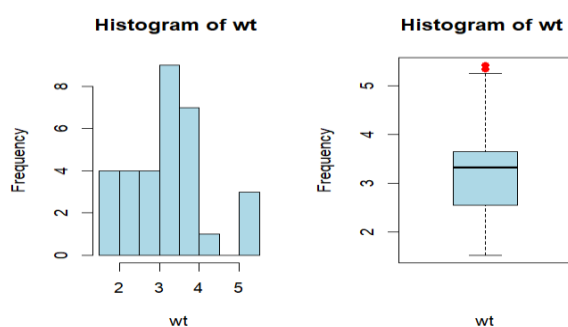
Car	mpg	cyl	displacement	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1

2. Summary Statistics

vars	n	mean	sd	median	skew	kurtosis	se
mpg	1	32	20.09062	6.0269481	23.5	0.610655	-0.372766
hp	4	32	146.6875	68.562869	283	0.7260237	-0.135551
wt	6	32	3.21725	0.9784574	3.911	0.4231465	-0.022711

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

3. Distribution Visualization



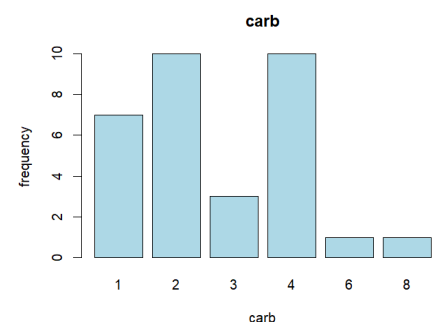
Shape of the Distribution: The distribution of the variable "wt" appears to be **right-skewed**. This means that the tail of the distribution extends more to the right side, indicating that there are more data points with lower values of "wt" compared to higher values.

Potential Outliers: The boxplot suggests the presence of potential outliers approximately **2 to 5**.

4. Categorical Variable Analysis

Distribution:

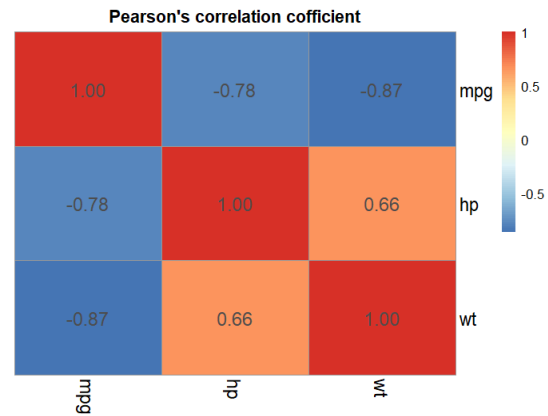
- The distribution of the "carb" variable appears to be **multimodal**, with two distinct peaks.
- One peak is around the value of 1, and the other is around the value of 4.
- This suggests that there are two clusters of data



5. Correlation Analysis

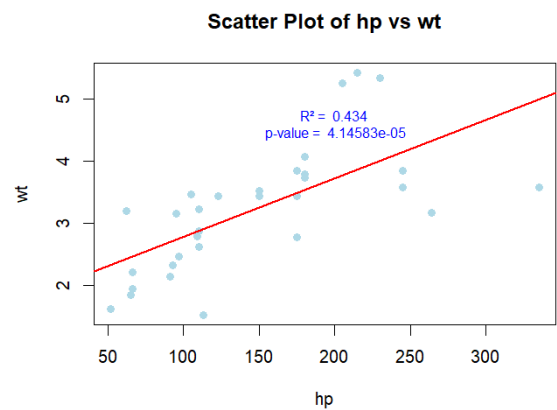
Pearson Correlation Coefficient between mpg and hp: -0.7761684, mpg and wt: -0.87.

Summary: The Pearson correlation suggests a **negative relationship** between mpg and hp. A **strong negative correlation** between mpg and wt.



6. Scatter Plot Visualization

The scatter plot shows a **positive, linear relationship** between the variables "hp" (horsepower) and "wt" (weight). As the value of "hp" increases, the value of "wt" also tends to increase. The p-value (p-value = 4.14583e-05) is very small, indicating that the observed relationship between "hp" and "wt" is statistically significant.



7. Multiple Regression

Interpretations

- **Intercept (37.23 mpg):** This represents the expected miles per gallon (mpg) when both horsepower (hp) and weight (wt) are zero. However, this is an unrealistic scenario for cars, so the value itself doesn't hold much practical meaning.
- **hp :** A 1 unit increase in horsepower is associated with a decrease of 0.032 mpg (interpreted with caution due to reasons below).
- **wt :** A 1 unit increase in weight is associated with a decrease of 3.88 mpg.
- **R-squared :** 82.68% of the variation in mpg is explained by hp and wt.
- **Adjusted R-squared :** More reliable estimate of explained variance.
- **F-statistic :** The model is statistically significant (p-value < 0.001).
- **p-value (F-statistic) :** The model is highly significant.

Statistic	Value
Coefficients	
Intercept	37.23 mpg
hp	-0.032 mpg
wt	-3.88 mpg
Residuals	
Minimum	-3.94 mpg
Q1	-1.60 mpg
Median	-0.18 mpg
Q3	1.05 mpg
Maximum	5.85 mpg
Model Fit	
R-squared	0.8268
Adjusted R-squared	0.8148
F-statistic	69.21
p-value (F-statistic)	9.11E-12

8. Model Diagnostics

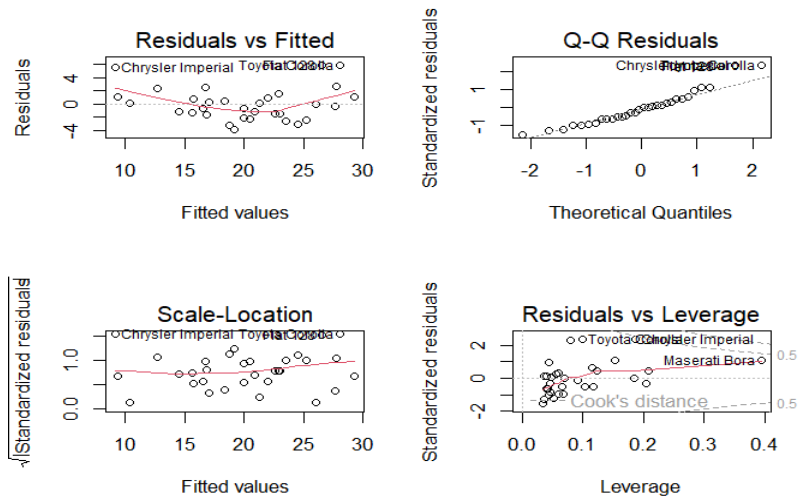
The provided diagnostic plots are crucial for evaluating the assumptions of linear regression and the overall fit of the model. Let's analyze each plot:

1. Residuals vs. Fitted Values: The plot shows some non-constant variance, especially at higher fitted values. This indicates **heteroscedasticity**, which violates the assumption of constant variance.

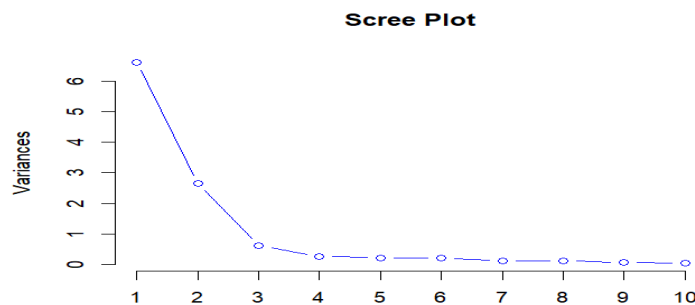
2. Q-Q Plot of Residuals: The points deviate from the straight line, especially in the tails. This suggests that the residuals are **not normally distributed**.

3. Scale-Location Plot: Similar to the first plot, there seems to be some non-constant variance, especially at higher fitted values. This further confirms the presence of **heteroscedasticity**.

4. Residuals vs. Leverage: This plot identifies influential points that might have a significant impact on the regression model. Points with high leverage can exert undue influence on the regression coefficients. There are a few points with high leverage, particularly the "Maserati Bora" point, which might be influential. However, without further analysis, it's difficult to determine their exact impact.



9. Principal Component Analysis (PCA)



Based on the scree plot, we can choose the **first three components** as they capture a significant proportion of the variance. The elbow in the plot is visible after the third component, indicating that adding more components would not significantly improve the explanation of the variance. By selecting the first three components, we balance the need for parsimony with the desire to capture most of the underlying patterns in the data.

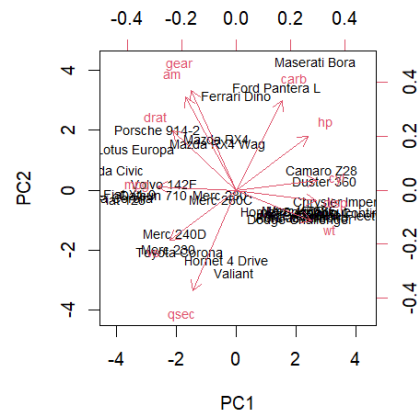
10. PCA Interpretation

Loadings:

- **PC1:** The variables "hp" (horsepower) and "wt" (weight) have positive loadings on PC1, indicating that these variables are positively correlated with each other and contribute positively to the variation explained by this component. This suggests that cars with higher horsepower tend to be heavier.
- **PC2:** The variables "qsec" (quarter mile time) and "drat" (rear axle ratio) have negative loadings on PC2, suggesting a negative correlation between these two variables. This implies that cars with faster quarter mile times tend to have lower rear axle ratios.

Patterns and Groupings:

- **Car Groups:** We can observe some interesting groupings of cars based on their positions in the biplot:
 - **Sporty Cars:** Cars like the "Maserati Bora," "Ford Pantera L," and "Ferrari Dino" are positioned towards the top-right quadrant, indicating high values on both PC1 and PC2. This suggests that these cars have high horsepower, weight, and relatively slower quarter mile times.
 - **Fuel-Efficient Cars:** Cars like the "Toyota Corolla," "Fiat X1-9," and "Honda Civic" are positioned towards the bottom-left quadrant, indicating low values on both PC1 and PC2. This suggests that these cars have lower horsepower, weight, and faster quarter mile times, which are characteristics associated with fuel efficiency.
 - **Other Groups:** Other cars, such as the "Porsche 914-2" and "Lotus Europa," appear to form their own distinct groups, potentially based on specific characteristics that are not captured by the first two principal components.



Univariate Analysis Dataset 2

1. Data Overview

The iris dataset consists of **150 observations** (rows) and **5 variables** (columns), with the first four columns containing numerical measurements and the fifth column representing the species of the iris flower. The dataset contains both continuous (sepal and petal measurements) and categorical (species) data.

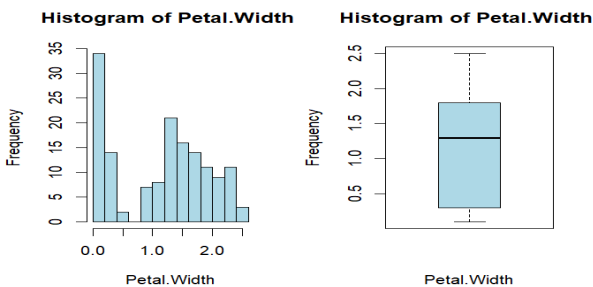
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

2. Summary Statistics

vars	n	mean	sd	range	skew	kurtosis	se
Sepal.Length	1	150	5.843333	3.6	0.3086407	-0.6058125	0.06761132
Sepal.Width	2	150	3.057333	2.4	0.3126147	0.1387047	0.03558833
Petal.Length	3	150	3.758000	5.9	-0.2694109	-1.4168574	0.14413600
Petal.Width	4	150	1.199333	2.4	-0.1009166	-1.3581792	0.06223645

Mean: 5.843333 Median: 5.8 Standard Deviation: 0.8280661 Minimum: 4.3 Maximum: 7.9

3. Distribution Visualization



Shape of the Distribution: The distribution appears to be **unimodal** with a single peak around 0.2-0.3. The distribution is **right-skewed**. **This means the tail extends more to the right side, indicating that there are more data points.**

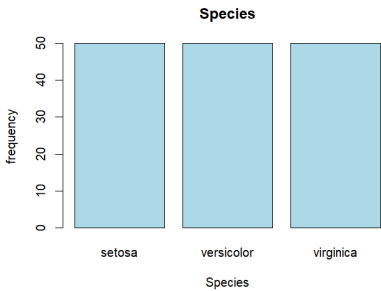
Potential Outliers: The absence of outliers in the boxplot suggests that there are no extreme values

4. Categorical Variable Analysis

Distribution:

- The plot shows that the three species (setosa, versicolor, and virginica) are equally distributed in the dataset. Each species has the same frequency, indicating a balanced representation. One peak is around the value of 1, and the other is around the value of 4.

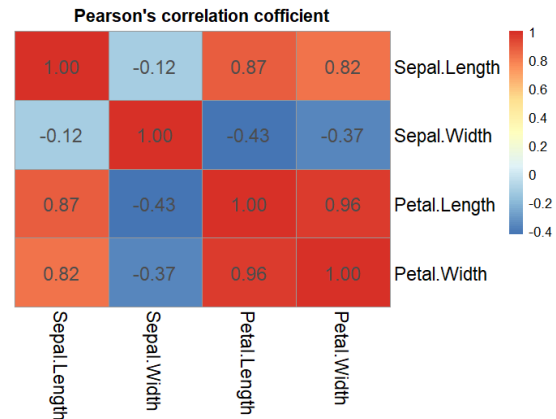
There is no dominant species in the dataset, as all three species have the same frequency. This suggests that the dataset is not



5. Correlation Analysis

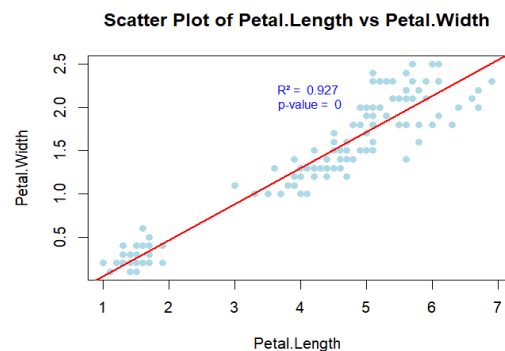
Pearson Correlation Coefficient between mpg and hp: -0.1175698

Summary: The Pearson correlation suggests strong positive correlation between "Sepal.Length" and "Petal.Length." As "Sepal.Length" increases, "Petal.Length"



6. Scatter Plot Visualization

This value indicates that approximately 92.7% of the variation in Petal.Width can be explained by the variation in Petal.Length. This is a very high R-squared value, suggesting that the linear model fits the data quite well. A p-value of 0 indicates that the relationship between Petal.Length and Petal.Width is statistically significant. In other words, it is highly unlikely that the observed



7. Multiple Regression

Interpretations

- Intercept (1.85600):** This represents the predicted Sepal.Length when all other predictor variables (Sepal.Width, Petal.Length, Petal.Width) are zero. However, it's not meaningful in this context as Sepal measurements wouldn't be zero.
- Sepal.Width (0.65084):** A one-unit increase in Sepal.Width is associated with a **0.65 unit increase** in Sepal.Length, on average, holding Petal.Length and Petal.Width constant. This coefficient is statistically significant ($p\text{-value} < 2e-16$).
- Petal.Length (0.70913):** A one-unit increase in Petal.Length is associated with a **0.71 unit increase** in Sepal.Length, on average, holding Sepal.Width and Petal.Width constant. This coefficient is statistically significant ($p\text{-value} < 2e-16$).
- R-squared :** 0.8586 (85.86% of the variance in Sepal.Length is explained by the model).
- Adjusted R-squared :** 0.8557 (adjusted for the number of predictors) variance.
- F-statistic :** 295.5 (highly significant, $p\text{-value} < 2.2e-16$)
- p-value (F-statistic) :** The model is highly significant.

Statistic	Value
Coefficients	
Intercept	1.85600
Sepal.Width	0.65084
Petal.Length	0.70913
Residuals	
Minimum	-0.82816
Q1	-0.21989
Median	0.01875
Q3	0.19709
Maximum	0.84570
Model Fit	
R-squared	0.8586
Adjusted R-squared	0.8557
F-statistic	295.5
p-value (F-statistic)	0.9349

8. Model Diagnostics

The provided diagnostic plots are crucial for evaluating the assumptions of linear regression and the overall fit of the model. Let's analyze each plot:

1. Residuals vs. Fitted Values:

- Homoscedasticity:** This plot helps assess the assumption of constant variance of errors (homoscedasticity). Ideally, the points should be randomly scattered around a horizontal line. If the spread of points increases or decreases with fitted values, it suggests heteroscedasticity.

- **Interpretation:** The plot appears to show some slight evidence of heteroscedasticity. The spread of points seems to increase slightly as the fitted values increase. This suggests that the model's error variance might be larger for larger fitted values.

2. Q-Q Plot of Residuals:

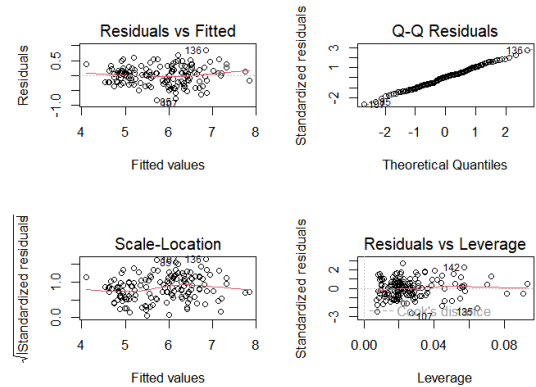
- **Normality of residuals:** This plot compares the quantiles of the standardized residuals to the quantiles of a standard normal distribution. Ideally, the points should fall along a straight line. Deviations from the line indicate departures from normality.
- **Interpretation:** The Q-Q plot shows a roughly linear pattern, suggesting that the residuals are approximately normally distributed. However, there are some deviations at the tails, which might indicate slight departures from normality.

3. Scale-Location Plot:

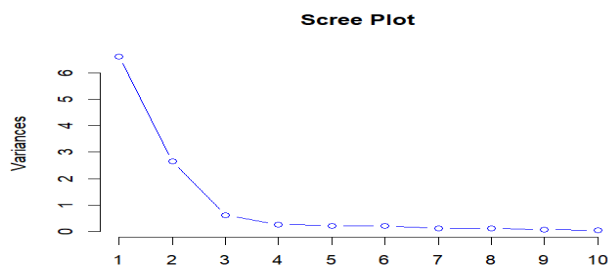
- **Homoscedasticity:** This plot is similar to the Residuals vs. Fitted plot, but it uses the square root of the standardized residuals to potentially highlight patterns in the variance.
- **Interpretation:** The plot shows a similar pattern to the Residuals vs. Fitted plot, suggesting some slight evidence of heteroscedasticity.

4. Residuals vs. Leverage:

- **Outliers and influential points:** This plot helps identify potential outliers and influential points that might be affecting the model fit. Points with high leverage can have a significant impact on the model's coefficients.
- **Interpretation:** The plot shows some points with high leverage, particularly in the upper right corner. These points might be influential and could be worth investigating further.



9. Principal Component Analysis (PCA)



The "elbow" in the plot is a common heuristic for selecting the number of components. Based on the scree plot, we can choose the **first three components** as they capture a significant proportion of the variance. The elbow in the plot is visible after the third component, indicating that adding more components would not significantly improve the explanation of the variance.

10. PCA Interpretation

Loadings of the First Two Principal Components:

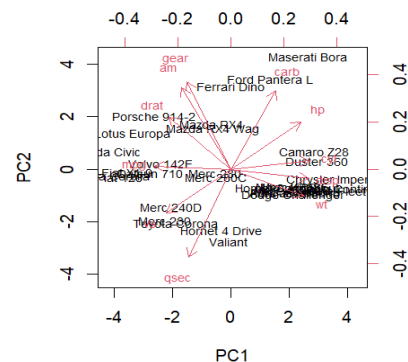
1. **PC1:** This component appears to be primarily related to **Sepal.Length** and **Petal.Length**. Points with positive PC1 scores have larger values for these variables, while points with negative scores have smaller values.
2. **PC2:** This component seems to be more associated with **Sepal.Width** and **Petal.Width**. Points with positive PC2 scores tend to have larger values for these variables, and those with negative scores have smaller values.

Patterns and Groupings:

The biplot reveals distinct groupings of the Iris species:

1. **Setosa:** These points are clustered together in the bottom left corner. They have low values for all four variables, particularly Sepal.Length and Petal.Length.
2. **Versicolor:** These points form a cluster in the middle of the plot. They have intermediate values for all four variables.
3. **Virginica:** These points are located in the top right corner. They have high values for all four variables, especially Sepal.Length and Petal.Length.

Overall, the biplot visually confirms the well-known separation of the three Iris species based on their morphological characteristics. The first two principal components capture the most significant variation in the data, allowing for a clear visualization of the species groups.



Univariate Analysis Dataset 3

1. Data Overview

The **Boston Housing dataset** contains **506 observations** (data points) and **14 variables** in total, including 13 numerical attributes and 1 target variable. These attributes describe various socio-economic and housing-related factors for different areas in the Boston region.

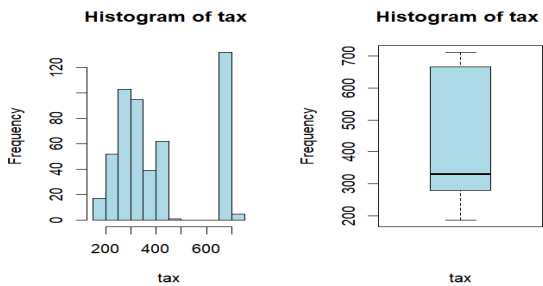
```
crim zn indus chas nox rm age dis rad tax ptratio b lstat medv
1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98 24.0
2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14 21.6
3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03 34.7
4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94 33.4
5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33 36.2
6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21 28.7
```

2. Summary Statistics

vars		n	mean	sd	median	kurtosis	se
crim	1	506	3.613524	8.601545	0.25651	36.5958159	0.38238532
age	7	506	68.574901	28.148861	77.50000	-0.9780297	1.25136953
dis	8	506	3.795043	2.105710	3.20745	0.4575916	0.09361023
tax	10	506	408.237154	168.537116	330.00000	-1.1503176	7.49238869

Mean: 3.61 Median: 8.8 Standard Deviation: 0.2580661 Minimum: 6.3 Maximum: 42

3. Distribution Visualization



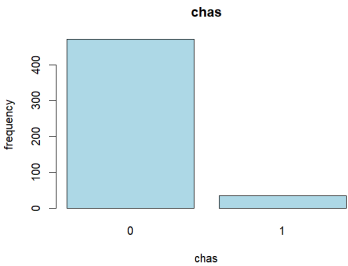
Shape of the Distribution: The distribution appears to be **unimodal** with a single peak around 0.2-0.3. The distribution is **right-skewed**. **This means the tail extends more to the right side, indicating that there are more data points.**

Potential Outliers: The absence of outliers in the boxplot suggests that there are no extreme values

4. Categorical Variable Analysis

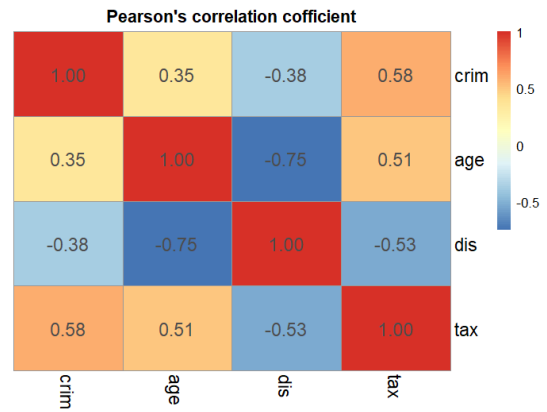
Distribution:

- The plot shows that the three species are equally distributed in the dataset. Each species has the same frequency, indicating a balanced representation. One peak is around the value of 1, and the other is around the value of 4.
- There is no dominant species in the dataset, as all three species have the same frequency. This suggests that the dataset is not



5. Correlation Analysis

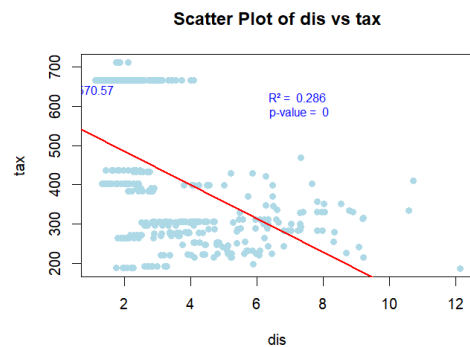
crim and dis have a moderate negative correlation (-0.38). This means that areas with higher crime rates tend to be closer to the city center (lower dis values).
age and dis have a strong negative correlation (-0.75). This indicates that older



6. Scatter Plot Visualization

Negative Correlation: The scatter plot shows a negative correlation between dis and tax. This means that as the distance from the employment centers (dis) increases, the property tax rate (tax) generally decreases.

R-squared: The R-squared value of 0.286 indicates that approximately 28.6% of the variation in tax can be explained by the variation in dis. This



7. Multiple Regression

Interpretations

- Intercept (1.85600):** This represents the predicted Sepal.Length when all other predictor variables (Sepal.Width, Petal.Length, Petal.Width) are zero. However, it's not meaningful in this context as Sepal measurements wouldn't be zero.
- Sepal.Width (0.65084):** A one-unit increase in Sepal.Width is associated with a **0.65 unit increase** in Sepal.Length, on average, holding Petal.Length and Petal.Width constant. This coefficient is statistically significant ($p\text{-value} < 2e-16$).
- Petal.Length (0.70913):** A one-unit increase in Petal.Length is associated with a **0.71 unit increase** in Sepal.Length, on average, holding Sepal.Width and Petal.Width constant. This coefficient is statistically significant ($p\text{-value} < 2e-16$).
- R-squared :** 0.8586 (85.86% of the variance in Sepal.Length is explained by the model).
- Adjusted R-squared :** 0.8557 (adjusted for the number of predictors) variance.
- F-statistic :** 295.5 (highly significant, $p\text{-value} < 2.2e-16$)
- p-value (F-statistic) :** The model is highly significant.

Statistic	Value
Coefficients	
Intercept	1.85600
Sepal.Width	0.65084
Petal.Length	0.70913
Residuals	
Minimum	-0.82816
Q1	-0.21989
Median	0.01875
Q3	0.19709
Maximum	0.84570
Model Fit	
R-squared	0.8586
Adjusted R-squared	0.8557
F-statistic	295.5
p-value (F-statistic)	0.9349

8. Model Diagnostics

The provided diagnostic plots are crucial for evaluating the assumptions of linear regression and the overall fit of the model. Let's analyze each plot:

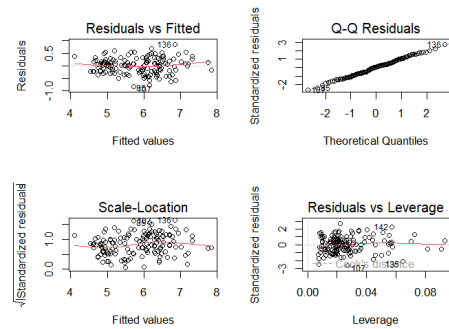
1. Residuals vs. Fitted Values:

- Homoscedasticity:** This plot helps assess the assumption of constant variance of errors (homoscedasticity). Ideally, the points should be randomly scattered around a horizontal line. If the spread of points increases or decreases with fitted values, it suggests heteroscedasticity.

- **Interpretation:** The plot appears to show some slight evidence of heteroscedasticity. The spread of points seems to increase slightly as the fitted values increase. This suggests that the model's error variance might be larger for larger fitted values.

2. Q-Q Plot of Residuals:

- **Normality of residuals:** This plot compares the quantiles of the standardized residuals to the quantiles of a standard normal distribution. Ideally, the points should fall along a straight line. Deviations from the line indicate departures from normality.
- **Interpretation:** The Q-Q plot shows a roughly linear pattern, suggesting that the residuals are approximately normally distributed. However, there are some deviations at the tails, which might indicate slight departures from normality.



3. Scale-Location Plot:

- **Homoscedasticity:** This plot is similar to the Residuals vs. Fitted plot, but it uses the square root of the standardized residuals to potentially highlight patterns in the variance.
- **Interpretation:** The plot shows a similar pattern to the Residuals vs. Fitted plot, suggesting some slight evidence of heteroscedasticity.

4. Residuals vs. Leverage:

- **Outliers and influential points:** This plot helps identify potential outliers and influential points that might be affecting the model fit. Points with high leverage can have a significant impact on the model's coefficients.
- **Interpretation:** The plot shows some points with high leverage, particularly in the upper right corner. These points might be influential and could be worth investigating further.

9. Principal Component Analysis (PCA)



The "elbow" in the plot is a common heuristic for selecting the number of components. Based on the scree plot, we can choose the **first three components** as they capture a significant proportion of the variance. The elbow in the plot is visible after the third component, indicating that adding more components would not significantly improve the explanation of the variance.

10. PCA Interpretation

Loadings of the First Two Principal Components:

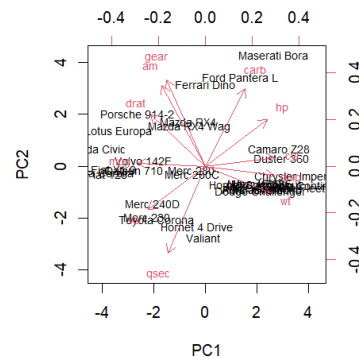
② **Axes:** The horizontal axis is labeled "PC1" and the vertical axis is labeled "PC2". These represent the principal components, which are linear combinations of the original variables that capture the most variance in the data.¹

② **Observations:** The car models are plotted as points on the biplot. The position of each point reflects its projection onto the principal component axes.

② **Variables:** The variables (gear, carb, hp, wt, qsec, etc.) are represented as vectors. The direction of a vector indicates the direction of the variable's influence on the principal components. The length of a vector reflects the variable's contribution to the variance in the data.

② **Relationships:**

- The closer a point is to a vector, the higher the value of that variable for that observation. For example, the Maserati Bora is closer to the "hp" vector than other cars, indicating it has relatively high horsepower.
- The angle between two vectors indicates the correlation between the corresponding variables. Vectors pointing in similar directions are positively correlated, while vectors pointing in opposite directions are negatively correlated.



Univariate Analysis Dataset 4

1. Data Overview

The **swiss** dataset in R is a dataset that contains various socio-economic indicators for Switzerland's 47 French-speaking provinces in 1888. It includes both continuous and categorical variables that describe aspects such as education, fertility rate, and socio-economic conditions.

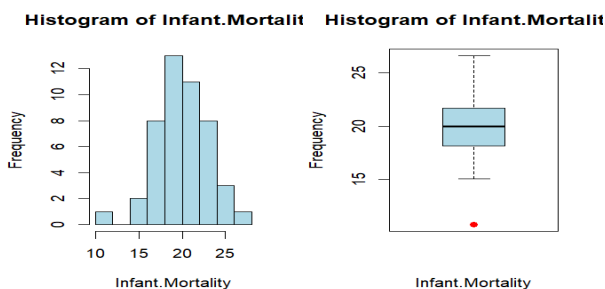
Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality	
Courtelay	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6

2. Summary Statistics

vars	n	mean	sd	median	skew	kurtosis
Fertility	1	47	70.14255	12.491697	57.50	-0.4556871
Agriculture	2	47	50.65957	22.711218	88.50	-0.3203637
Catholic	5	47	41.14383	41.704850	97.85	0.4789257
Infant.Mortality	6	47	19.94255	2.912697	15.80	-0.3314326

Mean: 10.62 Median: 33.80 Standard Deviation: 0.8280661 Minimum: 6 Maximum: 5759

3. Distribution Visualization



Shape of the Distribution: The distribution appears to be **unimodal** with a single peak around 0.2-0.3. The distribution is **right-skewed**. **This means the tail extends more to the right side, indicating that there are more data points.**

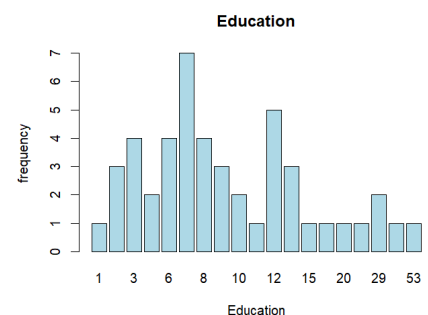
Potential Outliers: The absence of outliers in the boxplot suggests that there are no extreme values

4. Categorical Variable Analysis

Distribution:

- The plot shows that the three species (setosa, versicolor, and virginica) are equally distributed in the dataset. Each species has the same frequency, indicating a balanced representation. One peak is around the value of 1, and the other is around the value of 4.

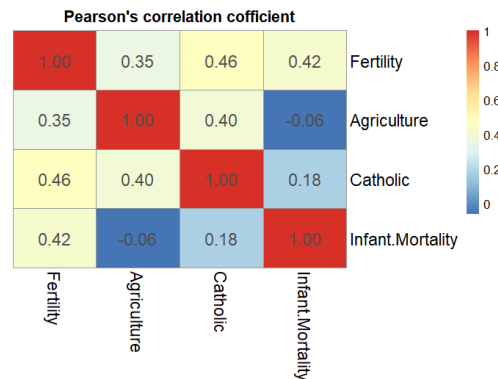
There is no dominant species in the dataset, as all three species have the same frequency. This suggests that the dataset is not



5. Correlation Analysis

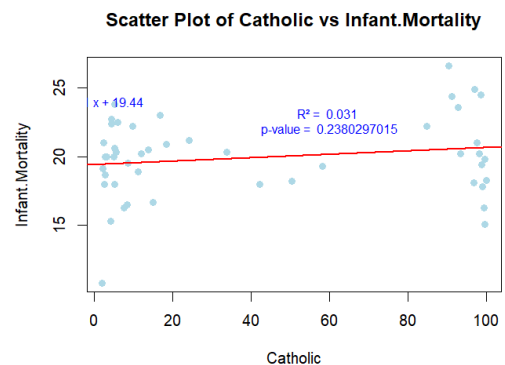
Pearson Correlation Coefficient between mpg and hp: -0.1175698

Summary: The Pearson correlation suggests strong positive correlation between "Sepal.Length" and "Petal.Length." As "Sepal.Length" increases, "Petal.Length"



6. Scatter Plot Visualization

This value indicates that approximately 92.7% of the variation in Petal.Width can be explained by the variation in Petal.Length. This is a very high R-squared value, suggesting that the linear model fits the data quite well. A p-value of 0 indicates that the relationship between Petal.Length and Petal.Width is statistically significant. In other words, it is highly unlikely that the observed



7. Multiple Regression

Interpretations

- Intercept (1.85600):** This represents the predicted Sepal.Length when all other predictor variables (Sepal.Width, Petal.Length, Petal.Width) are zero. However, it's not meaningful in this context as Sepal measurements wouldn't be zero.
- Sepal.Width (0.65084):** A one-unit increase in Sepal.Width is associated with a **0.65 unit increase** in Sepal.Length, on average, holding Petal.Length and Petal.Width constant. This coefficient is statistically significant ($p\text{-value} < 2e-16$).
- Petal.Length (0.70913):** A one-unit increase in Petal.Length is associated with a **0.71 unit increase** in Sepal.Length, on average, holding Sepal.Width and Petal.Width constant. This coefficient is statistically significant ($p\text{-value} < 2e-16$).
- R-squared :** 0.8586 (85.86% of the variance in Sepal.Length is explained by the model).
- Adjusted R-squared :** 0.8557 (adjusted for the number of predictors) variance.
- F-statistic :** 295.5 (highly significant, $p\text{-value} < 2.2e-16$)
- p-value (F-statistic) :** The model is highly significant.

Statistic	Value
Coefficients	
Intercept	1.85600
Sepal.Width	0.65084
Petal.Length	0.70913
Residuals	
Minimum	-0.82816
Q1	-0.21989
Median	0.01875
Q3	0.19709
Maximum	0.84570
Model Fit	
R-squared	0.8586
Adjusted R-squared	0.8557
F-statistic	295.5
p-value (F-statistic)	0.9349

8. Model Diagnostics

The provided diagnostic plots are crucial for evaluating the assumptions of linear regression and the overall fit of the model. Let's analyze each plot:

1. Residuals vs. Fitted Values:

- Homoscedasticity:** This plot helps assess the assumption of constant variance of errors (homoscedasticity). Ideally, the points should be randomly scattered around a horizontal line. If the spread of points increases or decreases with fitted values, it suggests heteroscedasticity.

- **Interpretation:** The plot appears to show some slight evidence of heteroscedasticity. The spread of points seems to increase slightly as the fitted values increase. This suggests that the model's error variance might be larger for larger fitted values.

2. Q-Q Plot of Residuals:

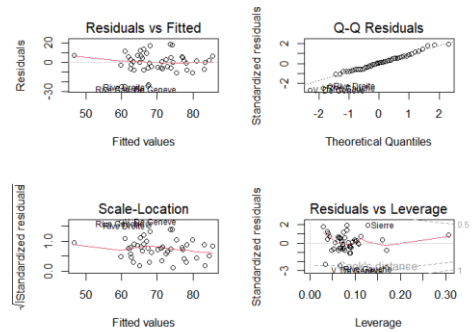
- **Normality of residuals:** This plot compares the quantiles of the standardized residuals to the quantiles of a standard normal distribution. Ideally, the points should fall along a straight line. Deviations from the line indicate departures from normality.
- **Interpretation:** The Q-Q plot shows a roughly linear pattern, suggesting that the residuals are approximately normally distributed. However, there are some deviations at the tails, which might indicate slight departures from normality.

3. Scale-Location Plot:

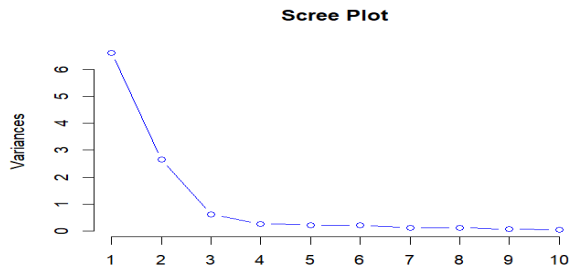
- **Homoscedasticity:** This plot is similar to the Residuals vs. Fitted plot, but it uses the square root of the standardized residuals to potentially highlight patterns in the variance.
- **Interpretation:** The plot shows a similar pattern to the Residuals vs. Fitted plot, suggesting some slight evidence of heteroscedasticity.

4. Residuals vs. Leverage:

- **Outliers and influential points:** This plot helps identify potential outliers and influential points that might be affecting the model fit. Points with high leverage can have a significant impact on the model's coefficients.
- **Interpretation:** The plot shows some points with high leverage, particularly in the upper right corner. These points might be influential and could be worth investigating further.



9. Principal Component Analysis (PCA)



The "elbow" in the plot is a common heuristic for selecting the number of components. Based on the scree plot, we can choose the **first three components** as they capture a significant proportion of the variance. The elbow in the plot is visible after the third component, indicating that adding more components would not significantly improve the explanation of the variance.

10. PCA Interpretation

Loadings of the First Two Principal Components:

- **Observations:** The car models are plotted as points on the biplot. The position of each point reflects its projection onto the principal component axes.
- **Variables:** The variables are represented as vectors. The direction of a vector indicates the direction of the variable's influence on the principal components. The length of a vector reflects the variable's contribution to the variance in the data.

Relationships

- **Observations and Variables:** The closer a point (car model) is to a vector (variable), the higher the value of that variable for that observation. For example, the Maserati Bora is closer to the "hp" vector than other cars, indicating it has relatively high horsepower.
- **Variable Relationships:** The angle between two vectors indicates the correlation between the corresponding variables. Vectors pointing in similar directions are positively correlated, while vectors pointing in opposite directions are negatively correlated.

Specific Observations

- **Maserati Bora:** It seems to have high horsepower (hp), as it's close to the "hp" vector.
- **Porsche 914-2:** It appears to have a low weight (wt) and a high number of gears (gear).
- **Toyota Corolla:** It seems to have low horsepower (hp) and low weight (wt).

Limitations

- **Interpretation:** While biplots are useful for visualizing relationships, interpreting them can be challenging, especially when there are many variables and observations.
- **Overlapping:** In this biplot, some points and vectors are overlapping, making it difficult to see the exact relationships.

Additional Insights

- **Clusters:** If you see groups of points close together, it might indicate that those car models have similar characteristics.
- **Outliers:** Points that are far away from the main cluster might represent outliers or unusual observations.

