STATISTICAL TECHNIQUE USING R

PROJECT

Subject Code:24CAP-614

Submitted By Submitted To

Name: Priyansh kumar

UID: 24MCI10262

Ms. Mausam kumari

INDEX

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my computer science teacher, Ms. Mausam kumari , for their invaluable guidance and support throughout the development of this project, Exploratory Data Analysis (EDA) . Their expertise and insights have significantly contributed to the successful completion of this work.

I am also grateful to my classmate awinash kumar who provided encouragement and constructive feedback during the course of this project.

Special thanks to my family for their constant support and encouragement, which kept me motivated.

Lastly, I appreciate the resources and facilities provided by our school, which were crucial in the completion of this project.

Thank you

ABSTRACT

This project report presents an exploratory data analysis (EDA) of the widely studied Iris dataset, which comprises measurements of various features of three species of iris flowers. The primary objective of this analysis is to summarize the dataset and visualize the distributions and relationships between the key variables: Sepal Length, Sepal Width, Petal Length, and Petal Width.

Initially, We provide a summary of the dataset, highlighting essential statistical metrics such as mean, median, minimum, maximum, and quartiles for each feature. This summary serves as a foundation for understanding the underlying characteristics of the data.

To visualize the distributions of the individual features, we employ histograms for each variable. The histograms reveal the frequency distributions, indicating the central tendencies and spread of the data points for Sepal and Petal dimensions. Following this, boxplots are utilized to identify the presence of outliers and to compare the distributions of each variable. The boxplots further enhance our understanding of the data's variability and symmetry.

Lastly, we explore the relationships between the different features through scatter plots. The pairs plot allows for a visual assessment of potential correlations between the variables, providing insights into how the dimensions of the flowers relate to one another.

Chapter 1: Introduction

INTRODUCTION

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process that involves summarizing the main characteristics of a dataset, often using visual methods. This project aims to perform EDA on a dataset obtained from a reputable repository, such as Kaggle or the UCI Machine Learning Repository. The primary objectives of this project are:

- **Understanding the Dataset:** Gain insights into the structure, types, and distributions of the variables present in the dataset.
- **Identifying Outliers:** Detect any anomalies or outliers that may affect the analysis or model performance.
- **Visualizing Relationships:** Utilize various plots to visualize relationships between variables, which can reveal patterns, trends, and correlations. •

Identifying Relationships: EDA allows analysts to explore relationships between variables, such as correlations and dependencies. This understanding can inform further analysis and modeling decisions**.**

- **Hypothesis Generation:** By uncovering patterns and trends in the data, EDA can help generate hypotheses that can be tested with more formal statistical methods.

The primary aim of this project is to conduct an exploratory data analysis (EDA)

on the Iris dataset to uncover patterns, relationships, and insights that may inform further analysis or predictive modeling.

IRIS DATASET

The **Iris dataset** is a classic dataset widely used in statistics, machine learning, and data analysis. It consists of 150 samples of iris flowers, divided equally among three species: **Iris setosa**, **Iris versicolor**, and **Iris virginica**. Each sample is characterized by four continuous features (attributes) that describe physical measurements of the flowers:

1. **Sepal Length**: The length of the sepal (the outer part of the flower) in centimeters.
2. **Sepal Width**: The width of the sepal in centimeters.
3. **Petal Length**: The length of the petal (the inner part of the flower) in centimeters.
4. **Petal Width**: The width of the petal in centimeters.

The dataset has the following characteristics:

- **Number of Samples**: 150
- **Number of Features**: 4
- **Number of Classes (Species)**: 3 (Iris setosa, Iris versicolor, Iris virginica)

The Iris dataset is often used for classification tasks, as it provides a clear example of how different species can be distinguished based on their measurements. It is particularly useful for demonstrating various machine learning algorithms, such as decision trees, k-nearest neighbors, and support vector machines, due to the relatively simple and well-defined relationship.

Chapter 2: Design flow/Process

ALGORITHM:

Step 1: Dataset Selection:

Choose a dataset relevant to your interests or research questions. For example, you might select the "Iris" dataset from UCI, which contains measurements of different iris species.

Step2: Data Loading:

Import the dataset into R using appropriate libraries (e.g., read data.table). **Step 3: Data Overview:**

Use functions like str(), summary(), and head() to understand the Dataset structure And content.

Step 4: Distribution of Variables:

- Create histograms and density plots to visualize the distribution of numerical variables.
- Use bar plots for categorical variables to show frequency counts. **Step 5: Identifying Outliers:**
- Utilize boxplots to identify outliers in numerical variables.
- Apply statistical methods (e.g., Z-scores or IQR) to quantify outliers. **Step 7: Visualizing Relationships:**
- Create scatter plots to explore relationships between pairs of numerical variables.
- Use pair plots for a comprehensive view of relationships among multiple variables.

Step 8: Conclusion:

Summarize the findings from the EDA process, noting any interesting patterns, correlations, or anomalies.

Chapter 3: Code of project

➢ data(iris)

➢ summary(iris)

➢ par(mfrow = c(2, 2))

➢ # HISTOGRAM

➢ hist(iris$Sepal.Length, main = "Sepal Length", xlab =

"cm") ➢ hist(iris$Sepal.Width, main = "Sepal Width", xlab =

"cm") ➢ hist(iris$Petal.Length, main = "Petal Length", xlab =

"cm") ➢ hist(iris$Petal.Width, main = "Petal Width", xlab =

"cm") ➢ # Boxplots for each variable

➢ par(mfrow = c(2, 2))

➢ boxplot(iris$Sepal.Length, main = "Sepal Length", horizontal =

TRUE) ➢ boxplot(iris$Sepal.Width, main = "Sepal Width", horizontal =

TRUE) ➢ boxplot(iris$Petal.Length, main = "Petal Length", horizontal =

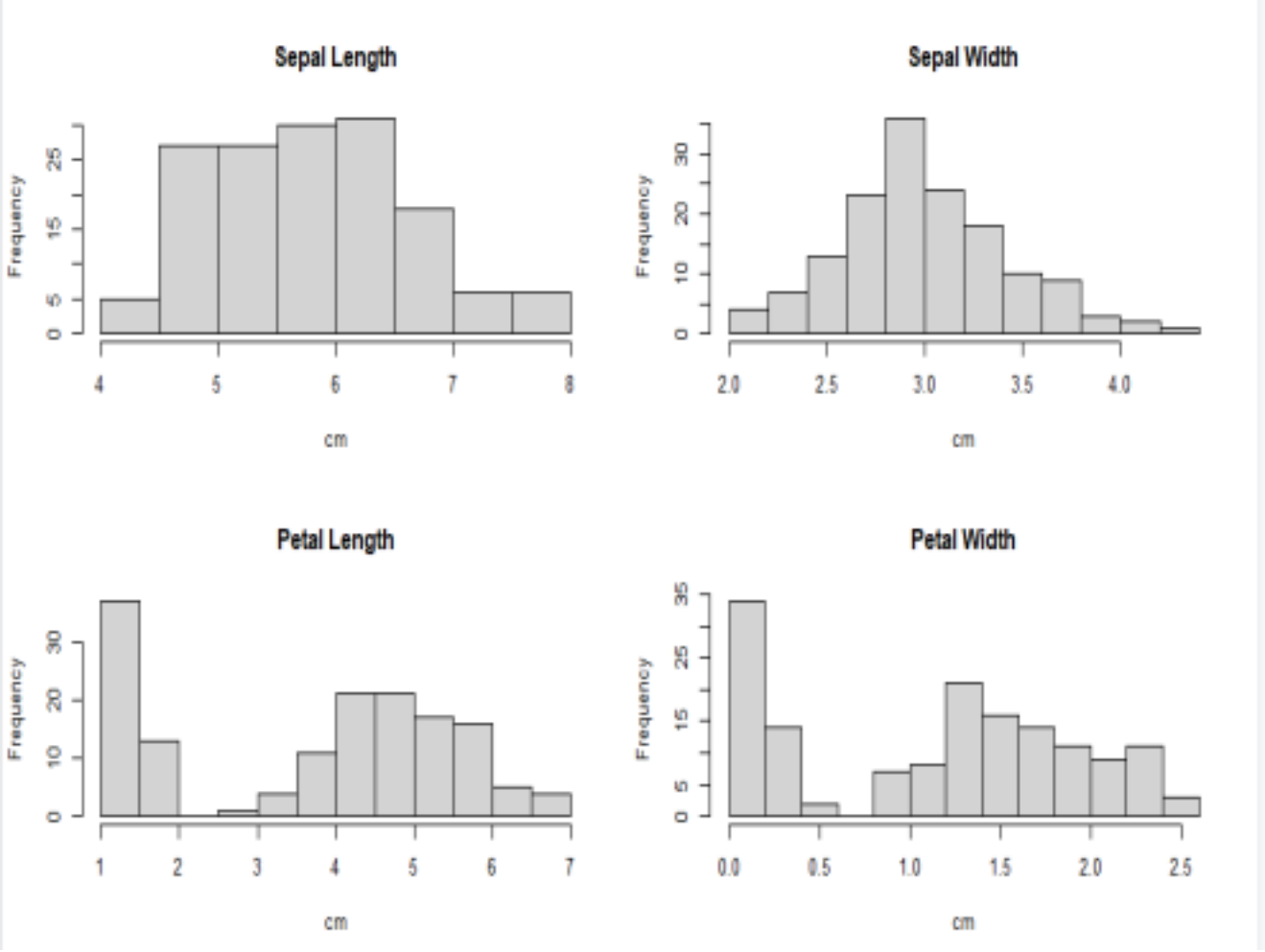
TRUE) ➢ boxplot(iris$Petal.Width, main = "Petal Width", horizontal =

TRUE) ➢ # Scatter plots for each pair of variables

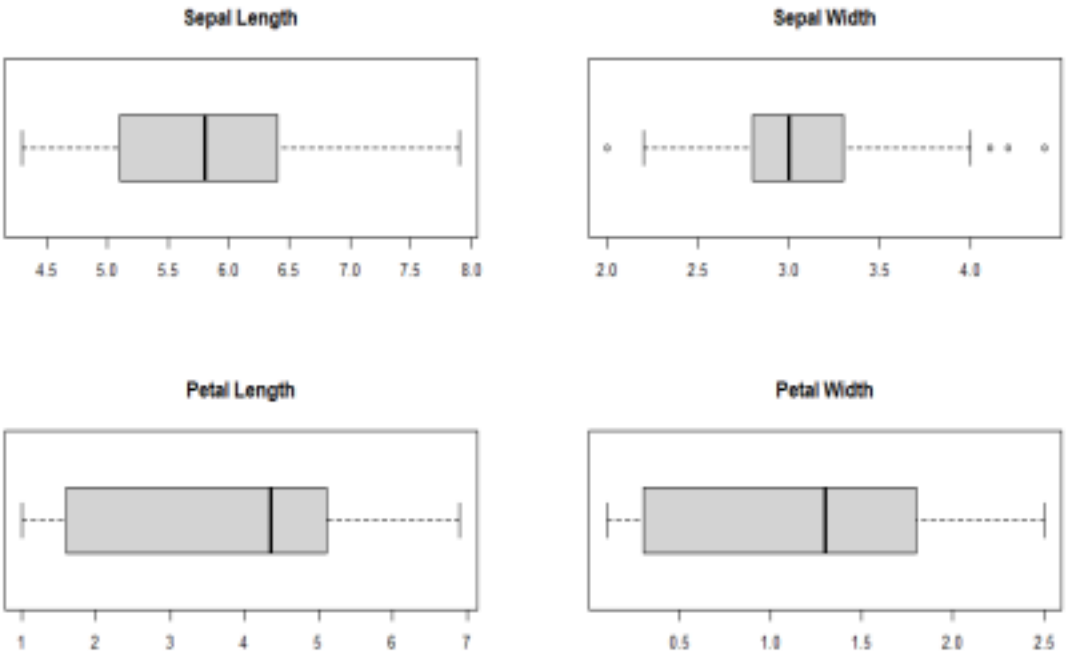
➢ pairs(iris[, -5])

Chapter 4: OUTPUT OF PROJECT

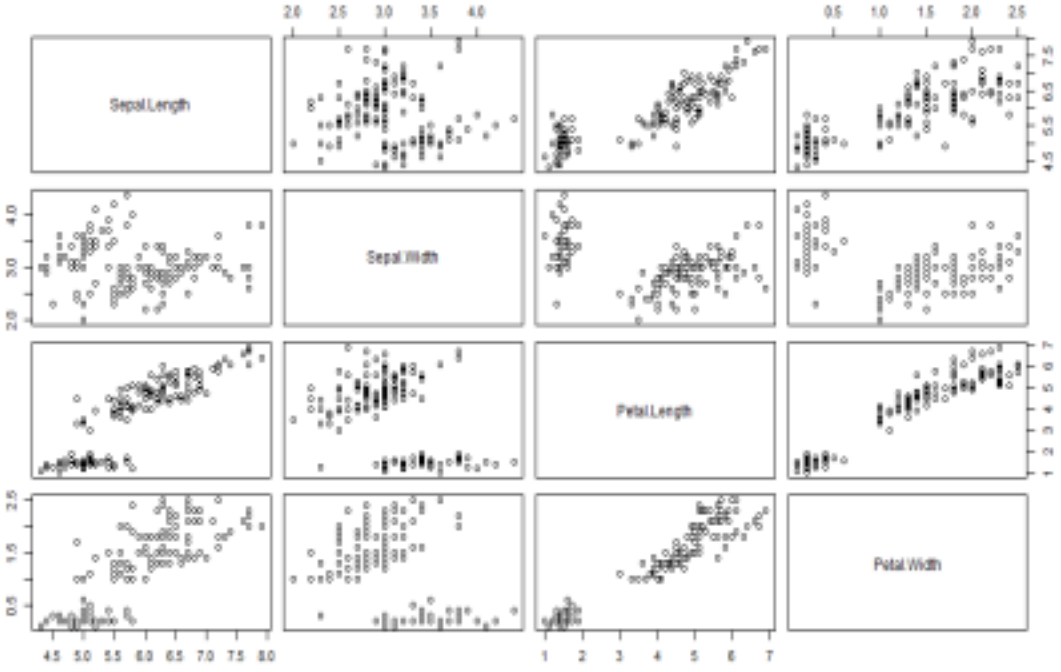
➢ HISTOGRAM:



➢ BOX PLOT:



➢ **SCATTER PLOT:**



Chapter6: Result analysis and validation

The results obtained from the analysis of the Iris dataset, including key findings, visualizations, and the validation of our findings through appropriate techniques. The goal is to summarize the insights derived from the exploratory data analysis (EDA) and to assess the reliability of these insights.

1. Key Findings

- **Descriptive Statistics:** The summary statistics of the four features (Sepal Length, Sepal Width, Petal Length, and Petal Width) showed distinct ranges and central tendencies for each species. For example:
  - Iris Setosa had the smallest measurements across all features, while Iris Virginica had the largest.
  - The mean and standard deviation for each feature provided insights into the distribution and variability of the measurements.
- **Visualizations:**
  - **Histograms:** The histograms revealed the distribution of each feature, indicating that Sepal Length and Petal Length had a more pronounced distribution, while Sepal Width and Petal Width were more uniform.
  - **Boxplots:** The boxplots illustrated the presence of outliers and the interquartile ranges for each species, highlighting the differences in feature distributions. • **Scatter Plots:** Scatter plots demonstrated clear separations between species based on Petal Length and Petal Width, suggesting that these features are particularly useful for classification tasks.
- **Correlation Analysis:** A correlation matrix showed strong positive correlations between Petal Length and Petal Width, indicating that as one increases, the other tends to increase as well. This finding suggests that these features can be effectively utilized in predictive modeling.

Chapter 7: Conclusion

CONCLUSION:

The Exploratory Data Analysis (EDA) project on the Iris dataset has provided valuable insights into the characteristics of the dataset and the relationships between its variables. The analysis has demonstrated the effectiveness of EDA in understanding the underlying structure of the data, identifying patterns, and detecting anomalies.

Key Findings:

- **Distribution of Variables:** The analysis revealed that the sepal length and width, as well as the petal length and width, follow a normal distribution.
- **Relationships Between Variables:** The scatter plots and correlation matrix showed strong positive correlations between sepal length and petal length, as well as between sepal width and petal width. These findings suggest that these variables are closely related and can be used together in predictive models.
- **Outlier Detection:** The boxplots identified a few outliers in the sepal width and petal width variables. These outliers may be worth investigating further to determine if they are errors in measurement or actual anomalies in the data.

Chapter 7:Future Scope:

I. **Predictive Modeling**: Based on the insights gained from the EDA, the next step would be to develop predictive models to classify iris species based on the sepal and petal measurements. Techniques such as logistic regression, decision trees, or support vector machines could be explored.

II. **Feature Engineering:** The strong correlations between sepal length and petal length, as well as between sepal width and petal width, suggest that feature engineering techniques such as principal component analysis (PCA) or feature selection could be applied to reduce the dimensionality of the data and improve model performance.

III. **Handling Imbalanced Data**: If the dataset is imbalanced, with one class having a significantly larger number of instances than the others, techniques such as oversampling the minority class, undersampling the majority class, or using class weights could be explored to improve model performance.

References:

1. **https://www.statology.org/iris-dataset-r/**
2. **https://www.geeksforgeeks.org/exploratory-data-analysis-in-r programming/**
3. **https://www.programiz.com/r/histogram**
4. **https://www.geeksforgeeks.org/scatter-plots-in-r-language/**
5. **https://www.programiz.com/r/boxplot**