Date 22/02/2024 Expt. No. 4

Expt. Name Page No. 10

EXPERIMENT - 4

Aim :- Use R data frames to study and analyze real-world datasets, perform basic data manipulations, and generate descriptive statistics using R functions.

Introduction :- The analysis of real-world datasets using R data frames allows for comprehensive exploration and manipulation of data. By leveraging the built-in datasets or importing external data sources, researchers can conduct descriptive statistics, visualize trends, and draw insights, contributing to informed decision-making and deeper understanding of various phenomena.

Objective :-

The primary objectives of this practical are :

1. Import & explore a real-world dataset using R data frames.
2. Perform basic data manipulations, such as subsetting and filteration.
3. Use R functions to generate descriptive statistics for key variables in the dataset.

Materials :-

1. RStudio or R environment installed
2. A real-world dataset for analysis.

Procedure :-

1. Load the Dataset :- Using R's built-in dataset or import an external dataset using functions like `read.csv()`, `read.table`.
2. Explore the Dataset :- Get an overview of the dataset using functions

Teacher's Signature:

Aim:- Use R data frames to study and analyze real-world datasets, perform basic data manipulations and generate descriptive statistics using R functions.

```r
# Load the iris dataset
data(iris)

# View the structure of the dataset
str(iris)

# Display the first few rows of the dataset
head(iris)

# Summary statistics of the dataset
summary(iris)

# Mean of sepal length
mean(iris$Sepal.Length)

# Median of petal width
median(iris$Petal.Width)

# Standard deviation of sepal width
sd(iris$Sepal.Width)

# Create a scatterplot of sepal length vs. sepal width
plot(iris$Sepal.Length, iris$Sepal.Width, main = "Sepal Length vs. Sepal Width", xlab = "Sepal Length", ylab = "Sepal Width", col = iris$Species)

# Add legend to the plot
legend("topright", legend = unique(iris$Species), col = unique(iris$Species), pch = 1)

# Create a boxplot of petal length by species
boxplot(Petal.Length ~ Species, data = iris, main = "Petal Length by Species", xlab = "Species", ylab = "Petal Length")

# Create a histogram of sepal width
hist(iris$Sepal.Width, main = "Histogram of Sepal Width", xlab = "Sepal Width", ylab = "Frequency")
```

like head(), tail(), str() & summary() to understand its structure, variable types and summary statistics.

3. Handle missing values :- Identify and handle missing values using functions like is.na() and na.omit().
Convert variables to appropriate data types using functions like as.numeric() or as.factor().
Remove duplicate rows using duplicated() and unique() functions.
Rename columns if necessary using the names() function.

4. Data Manipulation : Extract subsets of the data based on specific conditions using indexing ([ ]), subset() or dplyr functions like filter() and select().
Merge datasets using functions like merge() or dplyr functions like left_join() and inner_join().
Create new variables based on existing ones using arithmetic operations or functions like mutate() in dplyr.

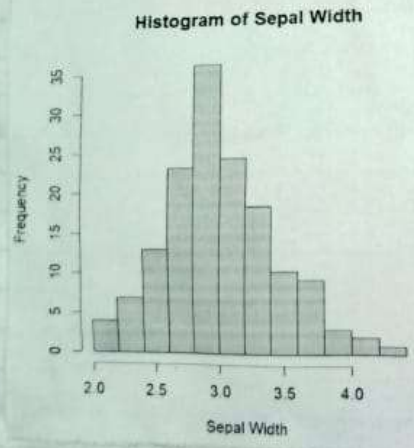
5. Descriptive statistics :- Calculate summary statics functions like mean(), median(), sd(), quantile() etc. to compute summary statistics for numeric variables.
Generate frequency tables for categorical variables using table() or summary().

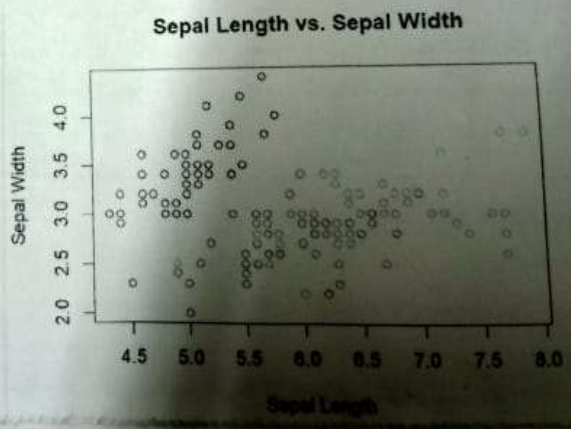
6. Data Visualization :- Create histograms, bar plots, scatter plots etc to visualize the distribution of variables and relationships between them using functions like hist(), barplot(), boxplot(), plot() etc.
Use libraries like ggplot2 for more customized and sophisticated visualizations.

7. Further Analysis :- Conduct hypothesis test, regression analysis or other statistical analyses based on research questions and objectives.
Use appropriate statistical tests like t-test ANOVA, correlation analysis etc, to test hypotheses and explore relationships between variables.

Teacher's Signature: _____

```
> # Load the iris dataset
> data(iris)
> # View the structure of the dataset
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> # Display the first few rows of the dataset
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> # Summary statistics of the dataset
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> # Mean of sepal length
> mean(iris$Sepal.Length)
[1] 5.843333
> # Median of petal width
> median(iris$Petal.Width)
[1] 1.3
> # Standard deviation of sepal width
> sd(iris$Sepal.Width)
[1] 0.4358663
> # Create a scatterplot of sepal length vs. sepal width
> plot(iris$Sepal.Length, iris$Sepal.Width, main = "Sepal Length vs. Sepal Width", xlab = "Sepal Length", ylab = "Sepal Width",
col = iris$Species)
```





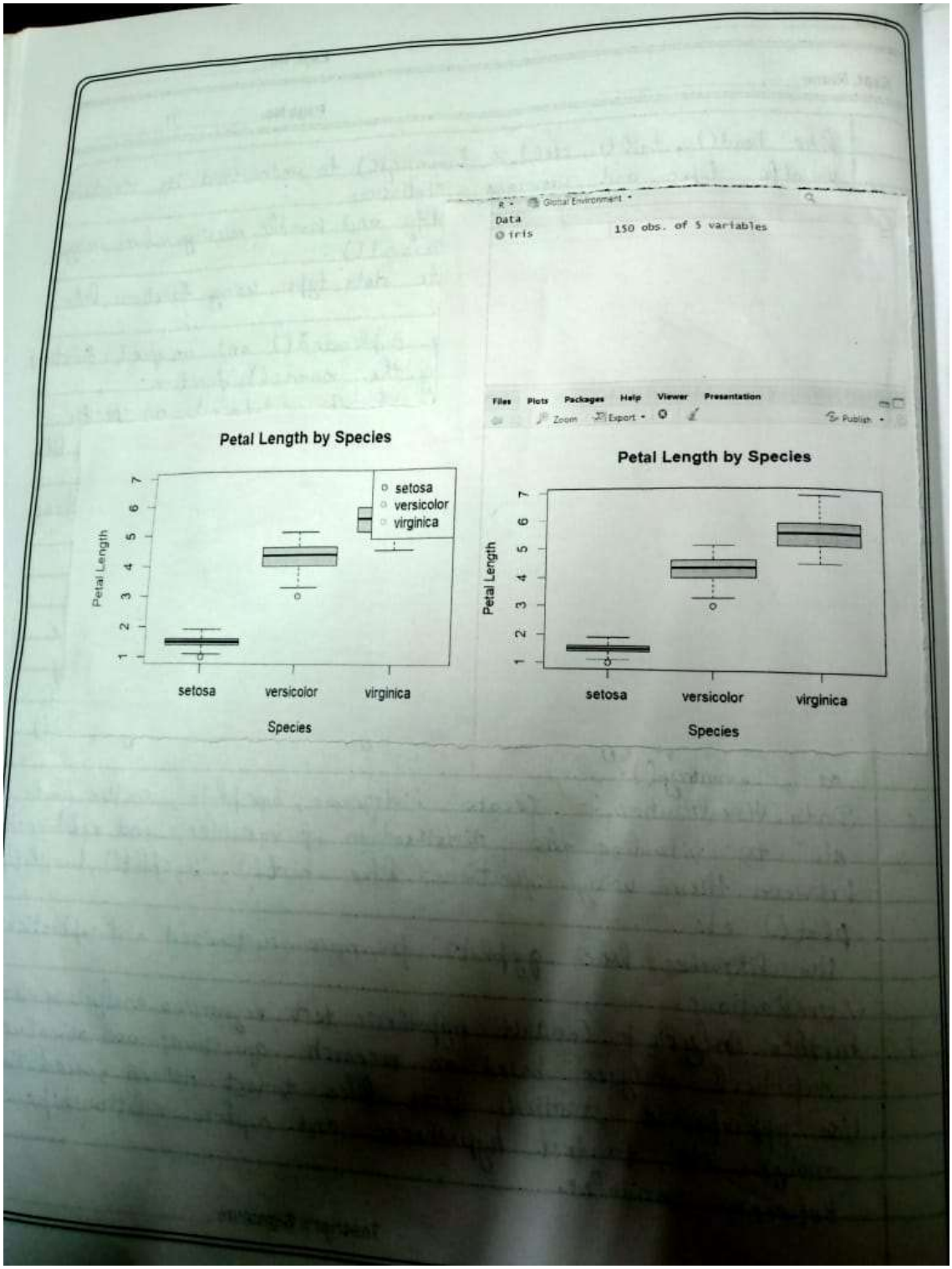
Date _____ Expt. No. _____

Expt. Name _____ Page No. 12

8. Interpretation & Conclusion :- Interpret the result of your analysed in the context of your research questions. Draw conclusions and make recommendations based on the findings from your data analysis

9. Documentation and Reporting :- Document your analysis process including data cleaning steps, manipulations and statistical analyses performed.
Present your findings clearly in reports, presentations or visualizations using Rmarkdown, Shiny apps or other tools.

Conclusion :- Summarize the insights gained from the analysis and discuss the importance of using R data frames for studying and manipulating real-world datasets. Emphasize the practical applicability of the skills acquired in this exercise for future data analysis projects.

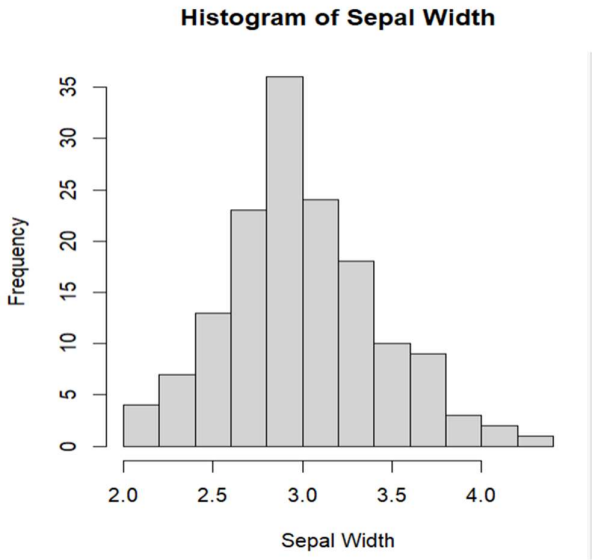


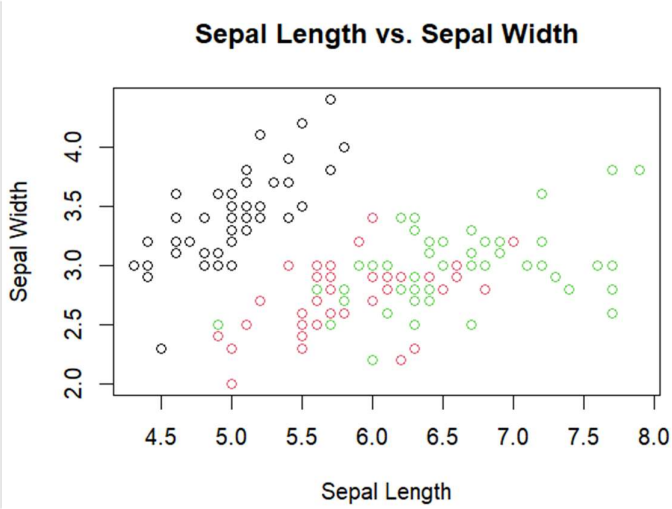
SCREENSHOTS:

Prac 4.R | Untitled3*

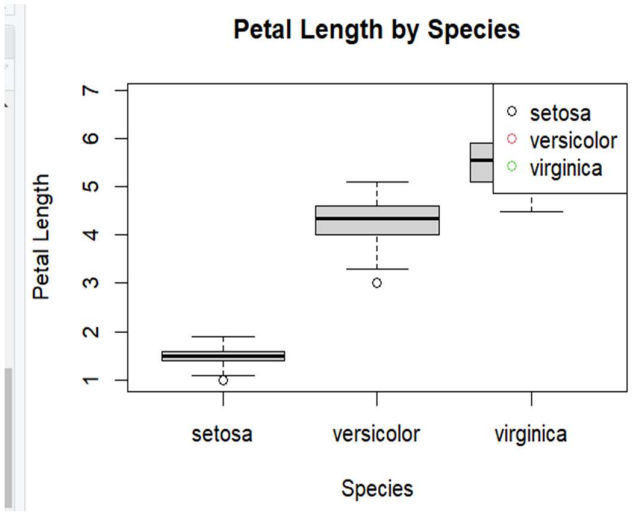
Source on Save | Run | Source

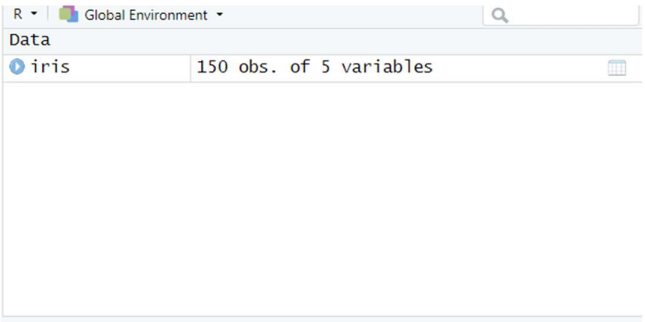
```r
# Load the iris dataset
data(iris)

# View the structure of the dataset
str(iris)

# Display the first few rows of the dataset
head(iris)

# Summary statistics of the dataset
summary(iris)

# Mean of sepal length
mean(iris$Sepal.Length)

# Median of petal width
median(iris$Petal.Width)

# Standard deviation of sepal width
sd(iris$Sepal.Width)

# Create a scatterplot of sepal length vs. sepal width
plot(iris$Sepal.Length, iris$Sepal.Width, main = "Sepal Length vs. Sepal Width", xlab = "Sepal Length", ylab = "Sepal Width", col = iris$Species)

# Add legend to the plot
legend("topright", legend = unique(iris$Species), col = unique(iris$Species), pch = 1)

# Create a boxplot of petal length by species
boxplot(Petal.Length ~ Species, data = iris, main = "Petal Length by Species", xlab = "Species", ylab = "Petal Length")

# Create a histogram of sepal width
hist(iris$Sepal.Width, main = "Histogram of Sepal Width", xlab = "Sepal Width", ylab = "Frequency")

```

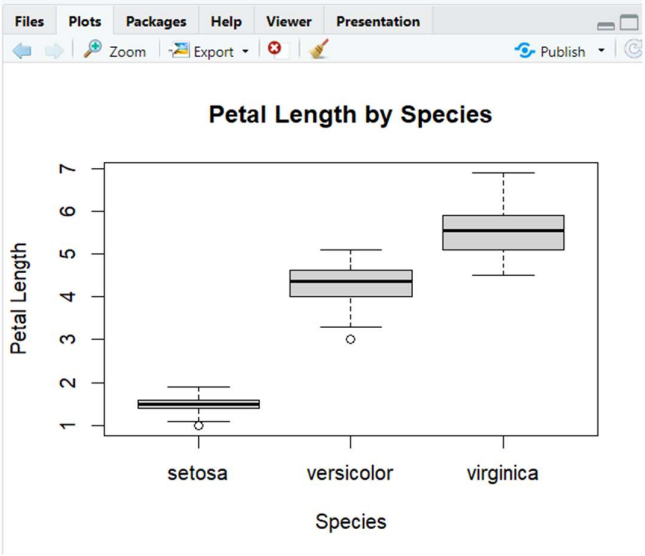





```
> # Load the iris dataset
> data(iris)
> # View the structure of the dataset
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> # Display the first few rows of the dataset
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> # Summary statistics of the dataset
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> # Mean of sepal length
> mean(iris$Sepal.Length)
[1] 5.843333
> # Median of petal width
> median(iris$Petal.Width)
[1] 1.3
> # Standard deviation of sepal width
> sd(iris$Sepal.Width)
[1] 0.4358663
> # Create a scatterplot of sepal length vs. sepal width
> plot(iris$Sepal.Length, iris$Sepal.Width, main = "Sepal Length vs. Sepal Width", xlab = "Sepal Length", ylab = "Sepal Width",
col = iris$Species)
```


Petal Length by Species
Petal Length
1
2
3
4
5
6
7
setosa
versicolor
virginica
Species


R
Global Environment
Data
iris
150 obs. of 5 variables


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Petal Length by Species
Petal Length
1
2
3
4
5
6
7
setosa
versicolor
virginica
Species