plot(x, y, main, xlab, ylab, xlim, ylim, axes)

The parameters of the plot functions are as follows:

- parameter **x** is the data values for x-axis.
- parameter **y** is the data values for y-axis
- parameter **main** is used for title of the grpah
- parameters **xlab** and **ylab** are used to specify the Labels for x-axis and y-axis respectively.
- Parameters **xlim** and **ylim** used define the limits of values of x and y respectively.
- **axes** specifies if the plot should include both axes.

```
input <- mtcars[,c('wt','mpg')]
print(head(input))

# Get the input values.
input <- mtcars[,c('wt','mpg')]

# Give the chart file a name.
png(file = "scatterplot.png")

# Plot the chart for cars with weight between 2.5 to 5 and mileage between 15 and 30.
plot(x = input$wt,y = input$mpg,
     xlab = "Weight",
     ylab = "Milage",
     xlim = c(2.5,5),
     ylim = c(15,30),
     main = "Weight vs Milage"
)

# Save the file.
dev.off()
```

Figure 14.28: Plot function to draw Scatter Plot

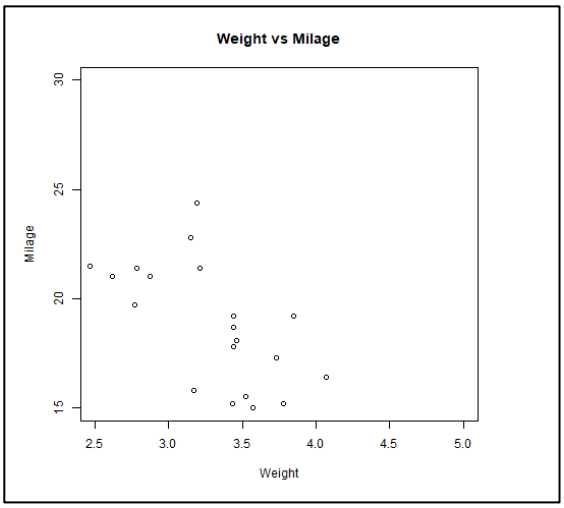


Figure 14.29: Scatter plot for the data of Figure 14.28

Scatterplot Matrices

The scatterplot matrix is used when there are more than two variables and you want to identify the correlation between one variable and the others. To make scatterplot matrices, we use **pairs()** function.

Syntax:

pairs(formula, data)

where,

- The **formula** represents a set of variables that are utilised in pairs.
- The data set from which the variables will be derived is referred to as **data**.

```
# Give the chart file a name.
png(file = "scatterplot_matrices.png")

# Plot the matrices between 4 variables giving 12 plots.

# One variable with 3 others and total 4 variables.

pairs(~wt+mpg+disp+cyl,data = mtcars,
      main = "Scatterplot Matrix")

# Save the file.
dev.off()
```

Figure 14.30: Function for Scatterplot matrix

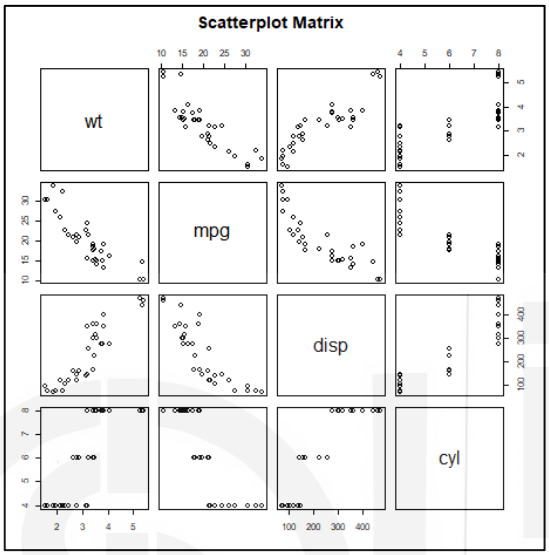


Figure 14.31: A Scatterplot matrix

Check your Progress 2

1. What is scatter plot?
2. When you will use histogram and when you will use bar chart in R?
3. What type of chart you consider when trying to demonstrate "relationship" between variables/parameters?

14.6 Summary

In this unit you have gone though various file types that can be processed for data analysis in R and further discussed their interfaces. R can read and write a variety of file types outside the R environment, including CSV, Excel, binary, XML and JSON. Further, R can readily connect to various relational databases, such as MySQL, Oracle, and SQL Server, and retrieve records as a data frame that can be modified and analysed with all of R's sophisticated packages and functions. The data can also be programmatically extracted from websites using R applications. "RCurl," "XML," and "stringr" are some R packages that are used to scrape data from the web. The unit also explains the concept of data cleaning and pre-processing which is the process of identifying, correcting and removing incorrect raw data, familiarization with the dataset, checking data for structural errors and data irregularities and deciding on how to deal with missing values are the steps involved in cleaning and preparing data which is mainly considered among the best practices. The unit finally explores the concept of

visualisations in R. There are various types of graphs and charts including- bar charts, box plots, histograms, line graphs and scatterplots that can be used to visualise the data effectively. The unit explained the usage and syntax for each of the illustration with graphics.

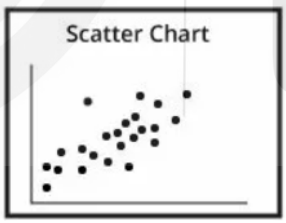
14.7 Answers

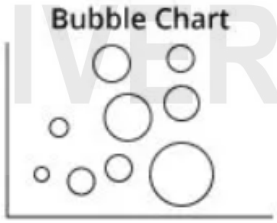
Check your progress 1

1. Install.packages("**rjson"**)
   library(**rjson**)
2. rb mode opens the file in the binary format for reading and wb mode opens the file in the binary format for writing.
3. The checklist points used for cleaning/ preparing data:
   Check for data irregularities: You may check for the invalid values and outliers.
   Decide on how to deal missing values: Either delete the observations if they are not providing any meaningful insights to our data or imputing the data with some logical values like mean or median based on the observations.

Check your progress 2

1. A scatter plot is a chart used to plot a correlation between two or more variables at the same time
2. We use a histogram to plot the distribution of a continuous variable, while we can use a bar chart to plot the distribution of a categorical variable.
3. When you are trying to show "relationship" between two variables, you will use a scatter plot or chart. When you are trying to show "relationship" between three variables, you will have to use a bubble chart.





14.8 REFERENCES AND FURTHER READINGS

1. De Vries, A., & Meys, J. (2015). *R for Dummies*. John Wiley & Sons.
2. Peng, R. D. (2016). *R programming for data science* (pp. 86-181). Victoria, BC, Canada: Leanpub.
3. Schmuller, J. (2017). *Statistical Analysis with R For Dummies*. John Wiley & Sons.
4. Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.
5. Lander, J. P. (2014). *R for everyone: Advanced analytics and graphics*. Pearson Education.
6. Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.
7. Heumann, C., & Schomaker, M. (2016). *Introduction to statistics and data analysis*. Springer International Publishing Switzerland.
8. Davies, T. M. (2016). *The book of R: a first course in programming and statistics*. No Starch Press.
9. https://www.tutorialspoint.com/r/index.html