R Studio

Prepared by: Jasmine A. Tulin

Preparatory work in R

- getwd()
  - Display your working directory
- setwd("<location of your dataset>")
  - change the path where you have stored your dataset

read.xlsx Package

- function provides a high level API for reading data from an Excel worksheet. It calls several low level functions in the process.
- install.packages("readxl")
- library("readxl")

- read.xlsx(**file**, **sheetIndex**, **sheetName** = NULL, **rowIndex** = NULL, **startRow** = NULL, **endRow** = NULL, **colIndex** = NULL)

Parameter	Description
file	the path to the file to read.
sheetIndex	a number representing the sheet index in the workbook.
sheetName	a character string with the sheet name.
rowIndex	a numeric vector indicating the rows you want to extract. If NULL, all rows found will be extracted, unless startRow or endRow are specified.
startRow	a number specifying the index of starting row. For read.xlsx this argument is active only if rowIndex is NULL.
endRow	a number specifying the index of the last row to pull. If NULL, read all the rows in the sheet. For read.xlsx this argument is active only if rowIndex is NULL.
colIndex	a numeric vector indicating the cols you want to extract. If NULL, all columns found will be extracted.

Scatterplots

- show many points plotted in the Cartesian plane. Each point represents the values of two variables. One variable is chosen in the horizontal axis and another in the vertical axis.
- The simple scatterplot is created using the **plot()** function.

Plot Function

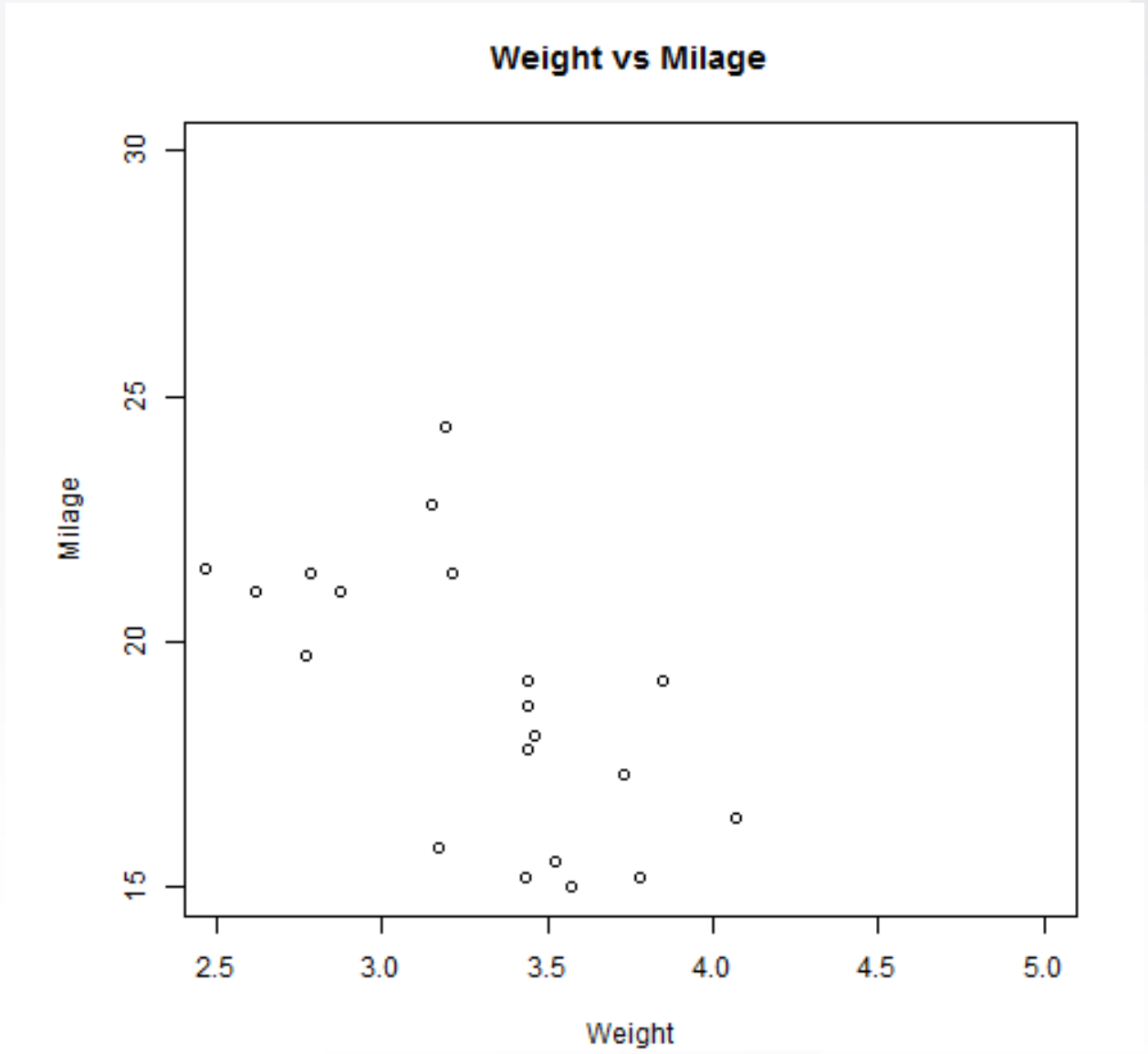
- The basic syntax for creating scatterplot in R is –
  - plot(x, y, main, xlab, ylab, xlim, ylim, axes)
- Following is the description of the parameters used –
  - **x** is the data set whose values are the horizontal coordinates.
  - **y** is the data set whose values are the vertical coordinates.
  - **main** is the tile of the graph.
  - **xlab** is the label in the horizontal axis.
  - **ylab** is the label in the vertical axis.
  - **xlim** is the limits of the values of x used for plotting.
  - **ylim** is the limits of the values of y used for plotting.
  - **axes** indicates whether both axes should be drawn on the plot.

```
# 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()
```


Weight vs Milage
Milage
Weight
30
25
20
15
2.5
3.0
3.5
4.0
4.5
5.0

Scatterplot Matrices

- more than two variables and we want to find the correlation between one variable versus the remaining ones we use scatterplot matrix.
- We use **pairs()** function to create matrices of scatterplots.

Pairs Function

- The basic syntax for creating scatterplot matrices in R is –
  - pairs(formula, data)
- Following is the description of the parameters used –
  - **formula** represents the series of variables used in pairs.
  - **data** represents the data set from which the variables will be taken.

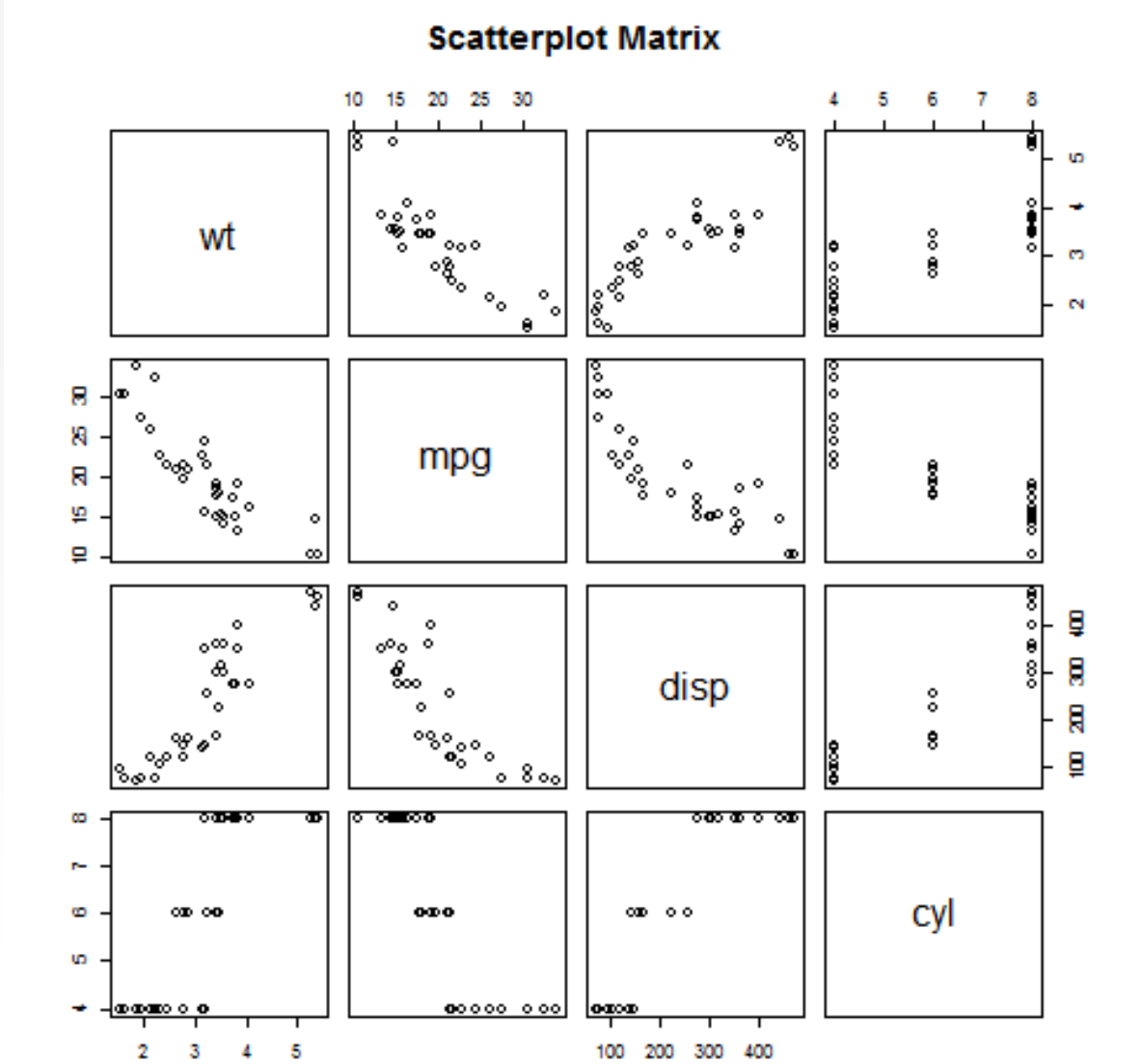
```
# Give the chart file a name.
png(file = "scatterplot_matrices.png")

# 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()
```


Scatterplot Matrix
wt
mpg
disp
cyl

Boxplots

- measure of how well distributed is the data in a data set.
- It divides the data set into three quartiles.
- This graph represents the minimum, maximum, median, first quartile and third quartile in the data set.

Boxplots

- Syntax
  - boxplot(x, data, notch, varwidth, names, main)
    - **x** is a vector or a formula.
    - **data** is the data frame.
    - **notch** is a logical value. Set as TRUE to draw a notch.
    - **varwidth** is a logical value. Set as true to draw width of the box proportionate to the sample size.
    - **names** are the group labels which will be printed under each boxplot.
    - **main** is used to give a title to the graph.

Numerical Measures in R

- mean()
- median()
- quantile()
- quantile(data, c(percentile))
- min() - max()
- IQR()
- var()
- cov(x,y)
- sd()
- cor(x,y)