MODULE PROJECT - 1

EXECUTIVE SUMMARY REPORT - 1

ALY6000 - Introduction to Data Analytics

BY NIKSHITA RANGANATHAN

FACULTY: PROF. BEHZAD AHMADI

DATE: SEP 25TH, 2022

Introduction

I had been allotted an assignment in Introduction to Analytics course where I had to work on R programming to complete a task using my beginner analytics skills. As I started understanding on the concepts of R, I could learn that R programming is an open-source programming language so I could easily download and install the tool. I explored more on this programming and gained knowledge on the modules and features of it. I also got to know that R programming is widely used as a statistical computing software and data analysis tool which is currently one of the most requested programming languages in the Data Science job market that makes it the trend nowadays.

I referred to the books like McGraw Hill - Bluman Elementary Statistics: A Step-by-Step Approach (10th edition) and R in Action by R. Kabacoff (2nd edition) to understand the step-by step instructions of working practically on the tool. I also went through some of the youtube videos for learning R coding syntax. I have included R console screen shots to support my observations.

Code and Outputs

1. Print your name at the top of the script

```
> print("Nikshita Ranganathan")
[1] "Nikshita Ranganathan"
```

   - Print function is used to display text and values in R

2. Install the *vcd* package

```
> install.packages("vcd")
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.
2/vcd_1.4-10.zip'
Content type 'application/zip' length 1287976 bytes (1.2 M
B)
downloaded 1.2 MB

package 'vcd' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\14086\AppData\Local\Temp\RtmpMDaX62\downlo
aded_packages
```

   - VCD is the abbreviation of **Visualizing Categorical Data.** We use this function to install package for the first time

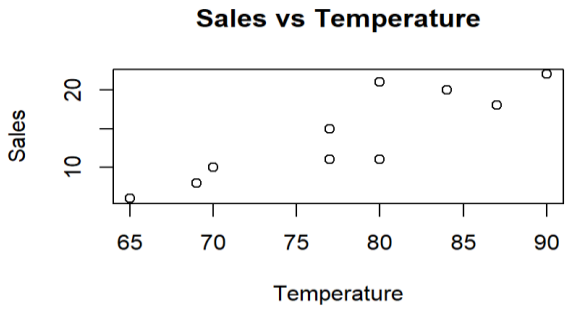
3. Import the *vcd* library

```
> library(vcd)
Loading required package: grid
```

   - library() function is used to load the package

4. Plot a sales ~ temp scatter plot using the data below :
   Sales data: (8,11,15,20,21,11,18,10,6,22)
   Temperature data: (69,80,77,84,80,77,87,70,65,90)

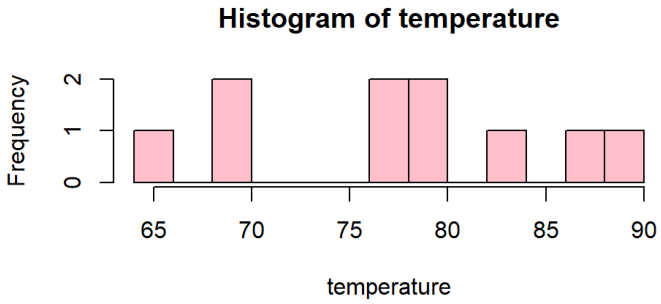
```
> sales<-c(8,11,15,20,21,11,18,10,6,22)
> temperature<-c(69,80,77,84,80,77,87,70,65,90)
> plot(temperature,sales,main="Sales vs Temperature",xlab
="Temperature",ylab="Sales")
```



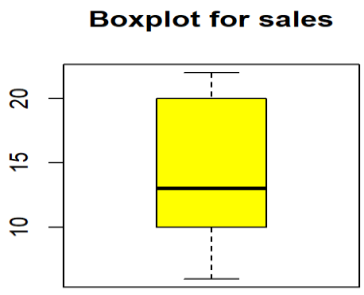
Values	
sales	num [1:10] 8 11 15 20 21 11 18 10 6 22
temperature	num [1:10] 69 80 77 84 80 77 87 70 65 90

- Sales and temp vectors are created, and graph is added using plot function. main is used for the heading of the graph, xlab is for the description x-axis and ylab is for the description of y-axis. I tried some functions like hist(), boxplot, barplot and pie chart with sales and temperature vectors. col() is for adding colours to the graphical representations.

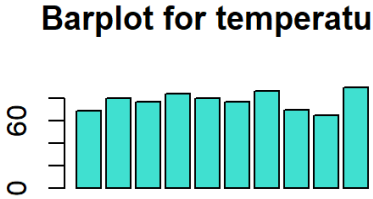
```
> hist(temperature,col="pink",breaks=10)
```




```
> boxplot(sales,main="Boxplot for sales",col="yellow")
```

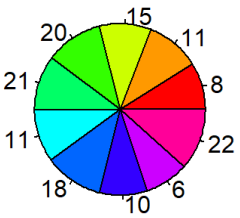



```
> barplot(temperature,main="Barplot for temperature",col="tur
quoise")
```




```
> pie(temperature,sales,main = "Pie chart between sales and t
emp",col=rainbow(length(temperature)))
```

Pie chart between sales and temp



5. Find the mean temperature

```
> mean(temperature)
[1] 77.9
```

- Mean of temperature vector is 77.9. It is calculated by adding all the values and dividing it by the total number of values.

6. Delete the 3rd element from the sales vector

```
> sales<-sales[-3]
> sales
[1]  8 11 20 21 11 18 10  6 22
```

- [-3] is used to remove 3rd element in the vector

7. Insert 16 as the 3rd element into the sales vector

```
> sales<-append(sales,16,2)
> sales
 [1]  8 11 16 20 21 11 18 10  6 22
```

- Append function can be used to easily add an element in any place in the vector. In this vector, 16 is inserted into the 3rd position of the vector.

8. Create a vector <names> with elements Tom, Dick, Harry

```
> names<-c("Tom","Dick","Harry")
> names
[1] "Tom"   "Dick"  "Harry"
```

- This vector has character strings

9. Create a 5 row and 2 column matrix of 10 integers

```
> mymatrix<-matrix(1:10,nrow=5,ncol=2)
> mymatrix
     [,1] [,2]
[1,]    1    6
[2,]    2    7
[3,]    3    8
[4,]    4    9
[5,]    5   10
```

- We can see a matrix with 5 rows and 2 columns. It is filled column wise.

10. Create a data frame <icSales> with sales and temp attributes

```
> icSales<-data.frame(temperature,sales)
> icSales
   temperature sales
1           69     8
2           80    11
3           77    16
4           84    20
5           80    21
6           77    11
7           87    18
8           70    10
9           65     6
10          90    22
```

- Dataframe stores data in form of a table with different datatypes

11. Display the data frame structure of icSales

```
> str(icSales)
'data.frame':   10 obs. of  2 variables:
 $ temperature: num  69 80 77 84 80 77 87 70 65 90
 $ sales      : num  8 11 16 20 21 11 18 10 6 22
```

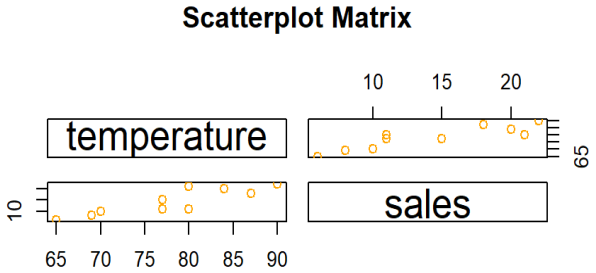
- str() function provides the details of the elements of the vector

12. Display a summary of the icScales data frame

```
> summary(icSales)
  temperature        sales
 Min.   :65.00   Min.   : 6.00
 1st Qu.:71.75   1st Qu.:10.25
 Median :78.50   Median :13.50
 Mean   :77.90   Mean   :14.30
 3rd Qu.:83.00   3rd Qu.:19.50
 Max.   :90.00   Max.   :22.00
```

- Summary function gives the statistical summary of the vectors. This includes mean, median, minimum, maximum etc. For this question, I also used pairs() command for getting a scatterplot matrix.

```
> pairs(icSales,main="Scatterplot Matrix",col="orange")
```



13. Import the dataset Student.csv

```
> getwd()
[1] "C:/Users/14086/Documents"
> studentdata<-read.csv("students.csv")
> studentdata
  StudentID First          Last Math Science Social.Studies
1        11   Bob         Smith   90      80             67
2        12  Jane         Weary   75      NA             80
3        10   Dan Thornton, III   65      75             70
4        40  Mary       O'Leary   90      95             92
```

- Dataset was provided by the faculty. Headings are StudentID, First , Last, Math, Science and Social Studies. getwd() function lets us know the information of current directory. read.csv function helps in importing csv files.

14. Display only the variable names of the Student.csv dataset

```
> colnames(read.csv("students.csv"))
[1] "StudentID"      "First"          "Last"           "Math"
[5] "Science"        "Social.Studies"
```

- colnames function returns the column names of the matrix

Summary

This assignment gave me an opportunity to learn how to download & install R programming packages, import library, understand datasets, calculate mean, create vectors, data frames, arrays, factors and matrices along with plotting the data into meaningful data visualization. I also discovered about Rmarkdown and extracted Rscript into html, pdf and word file. I got some errors while running the code and could solve them using Github and Stackoverflow websites.

I became more confident working with R programming. The online resources and references provided by my professor guided me to gain knowledge of several functionalities in R programming especially importing, computing and executing the code. The books gave me an overview of different types of data, datatypes, variables, measurement levels and sampling.

Bibliography

Kabacoff, Robert.I. (2011). R in Action Data analysis and graphics with R. Manning

Bluman, Allan G. (2017). Elementary statistics A Step by step approach. McGraw Hill

R Programming - DDS. (2021 Feb 15). How to download and install R and RStudio [Video File]. Retrieved from https://www.youtube.com/watch?v=TFGYlKvQEQ4

Edureka!. (2017 May 18). R Programming For Beginners | R Language Tutorial | R Tutorial For Beginners | Edureka [Video File] . Retrieved from https://www.youtube.com/watch?v=fDRa82lxzaU

Eugene O'Loughlin. (2021 Feb 1). How To... Insert Comments in R #03 [Video File]. Retrieved from https://www.youtube.com/watch?v=_vsd5EhzTPw

R Programming 101. (2022 Jun 21). R programming for ABSOLUTE beginners [Video File]. Retrieved from https://www.youtube.com/watch?v=FY8BISK5DpM&t=448s

References (Websites) :

https://stackoverflow.com/

https://resources.github.com/github-and-rstudio/

https://www.geeksforgeeks.org/get-and-set-working-directory-in-r/?ref=gcse

Appendix

```
---
title: "Module Project-1"
author: “Nikshita”
output:  word_document: default
date: "2022-09-23"
---
# 1. Printing name in R
> print("Nikshita Ranganathan")

# 2. Installing a package named vcd in R
> install.packages("vcd")

# 3. Importing vcd library
> library(vcd)

# 4. Plotting a scatter plot between sales and temperature
> sales<-c(8,11,15,20,21,11,18,10,6,22)
# Above command creates sales vector
> temperature<-c(69,80,77,84,80,77,87,70,65,90)
# Above command creates temperature vector
> plot(temperature,sales,main="Sales vs Temperature",xlab="Temperature",ylab="Sales")
# Above command creates a plot between sales and temperature; Temperature in X axis and
sales in Y axis

# 5. Finding the mean in temperature vector
> mean(temperature)

# 6. Deleting the 3rd element from sales vector
> sales<-sales[-3]
> sales

# 7. Inserting 16 as the 3rd element in the sales vector
> sales<-append(sales,16,2)
> sales

# 8. Creating a vector called names with elements Tom, Dick and Harry
> names<-c("Tom","Dick","Harry")
> names

# 9. Creating a matrix with 5 rows and 2 columns
> mymatrix<-matrix(1:10,nrow=5,ncol=2)
```

```
> mymatrix

# 10. Creating a dataframe called icSales with existing sales and temperature vectors
> icSales<-data.frame(temperature,sales)
> icSales

# 11. Displaying the structure of the icSales dataframe created
> str(icSales)

# 12. Displaying the summary of the icSales dataframe created
> summary(icSales)

# 13. Importing a dataset (csv file) called Student into R
> getwd()
> studentdata<-read.csv("students.csv")
> studentdata

# 14. Displaying only the headings of the dataset (csv file) called Student into R
> colnames(read.csv("students.csv"))

# Additional practice
> hist(temperature, col="pink", breaks=10)
> boxplot(sales, main="Boxplot for sales", col="yellow")
> barplot(temperature, main="Barplot for temperature", col="turquoise")
> pie(temperature, sales, main = "Pie chart between sales and temp",
  col=rainbow(length(temperature)))
> pairs(icSales, main="Scatterplot Matrix", col="orange")
```