I

SPRING SEMESTER 2023

IST 3015 A: BUSSINESS DATA ANALYTICS INSTRUCTOR: JAPHETH MURSI

DATE: 13TH March 2023

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ASSIGNMENT TWO

Total marks (40)

NJERU OLIVER 663565

Question 1 (8mks)

1. Differentiate between a package and a library in R and show example (2mks)
   1. A package is a collection of R functions, data, and compiled code in a well-defined format, created to add a specific functionality, an example being ggplot2, while a library is the directory where packages are stored, an example being library(tidyverse)
2. How many atomic classes does R programming language have? Discuss them (3mks
   1. Logical which represents Boolean values, i.e, TRUE or FALSE
   2. Real which represents digits that are continuous in nature, i.e, 14.24
   3. Character which represents a sequence of strings that are usually enclosed in either double or single quotes, i.e., “” or ‘’
   4. Integer which represents values that are whole and are denoted by adding an “L” suffix, e.g 65L
   5. Complex which represent values with a real and imaginary part and are usually denoted by adding i to the suffix of the imaginary part, i.e., 5 + 2i
   6. Raw which represents vectors of a specified length
3. Write a R program to convert given dataframe column(s) to a vector (3mks)

|  |  |
| --- | --- |
| Color | Numbers |
| Red | 10 |
| Yellow | 22 |
| Green | 11 |
| Blue | 5 |
| Violet | 2 |

# create the dataframe

df <- data.frame(

Color = c("Red", "Yellow", "Green", "Blue", "Violet"),

Numbers = c(10, 22, 11, 5, 2)

)

# convert the "Color" column to a vector

color\_vector <- df$Color

# convert the "Numbers" column to a vector

numbers\_vector <- df$Numbers

#display color\_vector and numbers\_vector

color\_vector

numbers\_vector

Question 2

Given the shared Excel dataset (Ngara\_district dataset); (16 Mks)

* 1. Write a R program to create a simple histogram for second and third column (2mks)

ngara\_district\_data<-Ngara\_district\_dataset

# Display Histogram for the Second column

hist(ngara\_district\_data$`Total Rural population`)

# Display Histogram for the third column

hist(ngara\_district\_data$`Population\_with\_ Clean Water`)

* 1. Rename the three dataframe columns after “ward” to “A”, “B”, “C” (2mks)

# create a dataframe

ngara\_data\_frame<-data.frame(ward=ngara\_district\_data$Ward,

Total\_Rural\_Population=ngara\_district\_data$`Total Rural population`,

Population\_with\_Clean\_Water=ngara\_district\_data$`Population\_with\_ Clean Water`, Percent\_Population\_Served\_with\_Clean\_Water=ngara\_district\_data$`Percent Population Served with Clean Water`)

# Rename the respective dataframe columns

names(ngara\_data\_frame)[2:4] <- c("A", "B", "C")

# Show the renamed dataframe columns

head(ngara\_data\_frame)

* 1. Write a R program and display a line plot for column B and C as X &Y axis respectively (3mks)
     + # Create a line plot for column B and C
     + line\_plot<-ggplot(ngara\_data\_frame,
     + aes(x=B, y=C)) +
     + geom\_line() +
     + labs(title="Total Rural population vs Population with\_ Clean Water")
     + line\_plot
  2. Label the Line plot as “Total Rural population vs Population with\_ Clean

Water” (2mks)

# Create a line plot for column B and C

line\_plot<-ggplot(ngara\_data\_frame,

aes(x=B, y=C)) +

geom\_line() +

labs(title="Total Rural population vs Population with\_ Clean Water")

line\_plot

* 1. Exclude the variable 'B' and display the top 6 rows of the data frame. (2mks)
     + # Create a new data frame that does not have B
     + df\_new <- ngara\_data\_frame[, c("ward", "A", "C")]
     + head(df\_new, n=6)
  2. Add another column “D” to the dataframe where column“A-B” =D (2mks)
     + # Add a new column 'D' to the dataframe where 'A - B' = 'D'
     + ngara\_data\_frame$D <- (ngara\_data\_frame$A - ngara\_data\_frame$B)
     + head(ngara\_data\_frame)
  3. Write a R program to drop ward “Kabanga” and “Bukiriro” from “ward”

column and show the new total number of wards (3mks)

# Drop rows with 'Kabanga' and 'Bukiriro' in the 'ward' column

df\_updated <- ngara\_data\_frame[!ngara\_data\_frame$ward %in% c("Kabanga", "Bukiriro"), ]

# Count the new total number of wards in the new dataframe

total\_wards <- nrow(df\_updated)

# Print the new dataframe and the total number of wards

print(df\_updated)

cat("\nTotal number of wards: ", total\_wards)

Question 3 (16Mks)

1. Write a R program to call the (built-in) dataset ‘*iris’.* Check whether it is a data frame or not? Display the last rows of the dataset (3mks)
   1. # Load the 'iris' dataset
   2. data(iris)
   3. # Print the first few rows of the dataset
   4. head(iris)
   5. # Check whether it is a data frame
   6. is.data.frame(iris)
   7. tail(iris)
2. Write a R program to create a matrix from a list of given vectors (3mks) Vec1\_Age<- c(34,40,30,23,21,24,44,60,90,94)

Vec2\_Height<-c(1.7,1.65,1.2,1.69,1.51,1.36,2.37,2.08,1.5,1.6)

# Create the vectors

Vec1\_Age <- c(34, 40, 30, 23, 21, 24, 44, 60, 90, 94)

Vec2\_Height <- c(1.7, 1.65, 1.2, 1.69, 1.51, 1.36, 2.37, 2.08, 1.5, 1.6)

# Combine the vectors into a matrix

age\_height <- cbind(Vec1\_Age, Vec2\_Height)

# Print the resulting matrix

age\_height

1. Write a R program to find Sum, Mean and Product the given vectors (4mks)
   1. # Calculate the sum, mean, and product of the vectors
   2. vec1\_sum <- sum(Vec1\_Age)
   3. vec2\_sum <- sum(Vec2\_Height)
   4. vec1\_mean <- mean(Vec1\_Age)
   5. vec2\_mean <- mean(Vec2\_Height)
   6. vec1\_product <- prod(Vec1\_Age)
   7. vec2\_product <- prod(Vec2\_Height)
   8. # Print the results
   9. cat("Results for Vec1\_Age:\n")
   10. cat("Sum =", vec1\_sum, "\n")
   11. cat("Mean =", vec1\_mean, "\n")
   12. cat("Product =", vec1\_product, "\n\n")
   13. cat("Results for Vec2\_Height:\n")
   14. cat("Sum =", vec2\_sum, "\n")
   15. cat("Mean =", vec2\_mean, "\n")
   16. cat("Product =", vec2\_product, "\n")
2. Convert the above vectors in section D to a dataframe and write a R program to create; (6mks)

#q3)d

df <- data.frame(Age = Vec1\_Age, Height = Vec2\_Height)

#print the dataframe

df

* 1. Line chart of Age and Height
     + #q3)d)i
     + library(ggplot2)
     + plot <- ggplot(df, aes(x=Age, y=Height)) + geom\_line()
     + plot
  2. Boxplot of Height
     + #q3)d)ii
     + plot2 <- ggplot(df, aes(y=Height)) +geom\_boxplot()
     + plot2
  3. Histogram of Age
     + #q3)d)iii
     + hist(Vec1\_Age)