

**SPRING SEMESTER 2023**

**IST 3015 (A): BUSSINESS DATA ANALYTICS INSTRUCTOR: JAPHETH MURSI**

**DATE: 23rd March 2023, Venue: Online**

**QUIZZ 2**

**Duration: 1hr 45 Mins Total marks (20)**

**NJERU OLIVER - 663565**

**Instruction**

**1. Show the output below each question**

Question 1 (8mks)

1. How would you apply plotly function to plots in R package? **(2mks)**
   1. **Plotly function is used in R to make an already plotted graph interactive through methods such as hovering on the graph’s plot to see key metrics being displayed that are dynamic and update as you move the cursor pointer over the graph’s plot. To apply the plotly function, assuming we have already installed it, we load it up using library(plotly). Say we have a dataset called Survey Dataset stored in variable df1 and it has a gender and salary column, we can plot it as follows**

**plot1<-ggplot(data = df1,**

**aes(x = sex,**

**y = weight,**

**col=species\_id,**

**shape=sex)) +**

**geom\_jitter()**

**plot1**

**we can then apply plotly as follows**

**ggplotly(plot1)**

1. Elucidate four data structures in R? **(2mks)**
   1. **Lists which are a collection of data of different types**
   2. **Factors which store data used to represent categories, e.g M for male, F for Female**
   3. **Vectors which are one-dimensional arrays that store data of the same type**
   4. **Matrices which are two-dimensional arrays that store data of the same type**
2. Create a vector with 8 of your friend’s names. Display the 2nd and 3rd friends **(2mk)**

**# create friends vector**

**friends<-c("Robert",**

**"Ochiel",**

**"Onchienku",**

**"Michelle",**

**"Mursi",**

**"Japeth",**

**"Rambo",**

**"Lorraine")**

**# display 2nd and 3rd friend**

**friends[2:3]**

1. Sort your friends by names using two sorting methods **(2mks)**

**To sort, we can use two functions, order() and sort() functions**

**# sort() function**

**friends\_sorted <- sort(friends)**

**print(friends\_sorted)**

**# order() function**

**friends\_ordered <- friends[order(friends)]**

**print(friends\_ordered)**

# Question 2 (12mks)

In the CRB\_Score \_dataset attached, Subset the dataset and create two data frames for JobID 1 and 2.

To do this, we would need to load the dplyr library so that we can manipulate the data. The code is as follows

library(dplyr)

# JobID 1 subset

JobID\_1 <- CRB\_score\_dataset %>%filter(JobID == 1)

JobID\_1

# JobID 2 subset

JobID\_2 <- CRB\_score\_dataset %>%filter(JobID == 2)

JobID\_2

* 1. Merge the two data frames and sort by “Housing”. **(2mks)**

**To merge we will use the rbind() function and pass in the two dataframes. To sort by housing, we will use the arrange function and pass in the merged dataframe and the column we want to use to sort the rest of the data.**

**# Merge the data frames**

**merged\_df <- rbind(JobID\_1, JobID\_2)**

**merged\_df**

**# Sort by "Housing"**

**sorted\_merged\_df <- arrange(merged\_df, Housing)**

**sorted\_merged\_df**

* 1. Using the merged dataframe in (a) above, create a boxplot for *Credit Amount and Duration*. Facet the plot by the column *'Jobid'.* **(2mks)**

**To do this, we need to load up ggplot2 library**

**library(ggplot2)**

**plot1<-ggplot(merged\_df,**

**mapping=aes(x=`Credit amount`,**

**y=`Duration`,**

**color=`Housing`)) +**

**geom\_boxplot()+**

**facet\_wrap(merged\_df$JobID)**

**plot1**

* 1. Test the hypothesis whether the customers default score is independent of their credit amount at .05 significance level. **(2mks)**
     + **To do this we need to create a contingency table as shown**
     + **contigency\_tbl<- table(merged\_df$Defaulter, cut(merged\_df$`Credit amount`, breaks = 5))**
     + **chisq.test(contigency\_tbl)**
     + **the answer is 0.1769 which is higher than significance level meaning it is independent.**
  2. Using the whole dataset, create a scatterplot for *Age* and Duration column. Use *color and shape functions* in your plot. Label the plot as “Duration distribution”. Connect the plot points using geomline. **(4mks)**

**plot2<-ggplot(data = CRB\_score\_dataset,**

**aes(x = Age,**

**y = Duration,**

**color = Housing,**

**shape = Sex)) +**

**geom\_point() +**

**geom\_line() +**

**labs(title = "Duration distribution",**

**x = "Age",**

**y = "Duration")**

**plot2**

* 1. Is there a relationship between the age of the customer and the Duration of payment? **(2mks)**

To do this, we will use the correlation function and pass in the variables that correspond to age and duration. The code is as shown

# check for correlation

cust\_age<-CRB\_score\_dataset$Age

duration<-CRB\_score\_dataset$Duration

cor(cust\_age, duration)

The correlation is -0.337 meaning that there is a moderate negative correlation between age and duration meaning that if the value of age increases, duration decreases and vice versa.