

## SPRING SEMESTER 2023

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

**DATE: 11th APRIL 2023, Venue: LAB 3**

**END OF SEMESTER EXAMS**

**Duration: 1hr 45 Mins Total marks (30)**

**Instruction**

1. **Attempt all the questions**
2. **Use Excel and R where applicable**
3. **Paste the output of each question on your answer sheet**
4. **Make sure you submit the right document**

**NJERU OLIVER NJIRU - 663565**

**Question 1 (10mks)**

1. The table below shows the number of absences X, in an IST 3015 course and the final exam grade Y, for seven students

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 9 | 2 | 6 | 4 | 3 | 3 |
| Y | 75 | 60 | 70 | 45 | 70 | 80 | 85 |

* 1. Find the correlation coefficient and interpret your result

## (2mks)

## Using cor() function in R and storing the values of X and Y as below: correlation is -0.6

## Graphical user interface, application, table, Excel Description automatically generated

## The correlation coefficient value is -0.6 meaning that there is a negatively moderate correlation between the number of absences X and the final exam grade Y. This also means that in the case of an increase in the value of X, there will be a decrease in the value of Y due the negative correlation value between the two variables.

* 1. Predict the test score for a student with 7 absences **(2mks)**

**55.7**

Graphical user interface, application

Description automatically generated

1. Discuss the atomic classes in R **(2mks)**

* **Numerics – This includes both whole numbers in the positive and negative scale and also decimals.[2, 15.5]**
* **Integers – This includes only whole numbers. [2L]**
* **Characters – this includes strings. [“oliver”]**
* **Logical – this includes only Boolean logic. [TRUE, FALSE]**
* **Complex – this consists of complex numbers with real and imaginary parts. [1+4i]**
* **Raw – this includes data types that is in raw bytes form.**

# Using below vectors, create a dataframe “Dataframe1”. Create a new column “Gross\_profit” and calculate Gross\_profit (Sales -COS) **(2mks)**

Month<-c(July, August, September, October) sales <- c(65000,80000,123000,75000,45000) C\_O\_S <- c(15000,20000,34000,32000,36000)

# Create the vectors

Month <- c("July", "August", "September", "October", "November")

sales <- c(65000, 80000, 123000, 75000, 45000)

c\_o\_s <- c(15000, 20000, 34000, 32000, 36000)

# Create the data frame

Dataframe1 <- data.frame(Month, sales, c\_o\_s)

# Add Gross\_profit Column

Dataframe1$Gross\_profit <- Dataframe1$sales - Dataframe1$c\_o\_s

# Print the data frame

Dataframe1

1. Create a row that will be the total sum of the numeric columns (Sales, COS, GP), name it “Total”. **(2mks)**

# Add Total row using sum function

Dataframe1[nrow(Dataframe1) + 1,] <- c("Total", sum(Dataframe1$sales), sum(Dataframe1$c\_o\_s), sum(Dataframe1$Gross\_profit))

# Print the updated data frame

Dataframe1

## Question 2 (8mks)

1. Using “Bank Churners” Dataset attached, Conduct Exploratory data analysis on the dataset and comment on few interesting observations **(3mks)**

**We have 18 features. The dataset does not contain missing data so we do have to perform data preprocessing. Since we have categorical data such as Education\_Level, Marital\_Status, Card\_Category… etc, we will have to perform one hot encoding to covert this categorical data into numerical data that can be used to train and test machine learning models so as to get more accurate predictions. We can also see from the plotted graphs that there is a close relationship between the age of the customers between 40-50 and the number of bank churners available suggesting the existence of some correlation.**

1. Display the top (6) categories of Churners whose “Card Category” was **Gold & Platinum** and whose educational level was ***Postgraduate*** and **Uneducated**. What is their average “**Months\_on\_book**” **(3mks)**

**# load the dataset**

**bnk\_crs <- BankChurners\_1\_**

**# Filter Gold or Platinum and an educational level of Postgraduate or Uneducated**

**subset\_data <- subset(bnk\_crs, Card\_Category %in% c("Gold", "Platinum") & Education\_Level %in% c("Post-Graduate", "Uneducated"))**

**# Display top 6 categories**

**top\_categories <- head(sort(table(subset\_data$Attrition\_Flag)), 6)**

**print(top\_categories)**

**# average months\_on\_book**

**# Calculate the average "Months\_on\_book" for the subset of churners**

**avg\_months\_on\_book <- mean(subset\_data$Months\_on\_book)**

**# Display the average "Months\_on\_book"**

**print(avg\_months\_on\_book)**

1. Create a new data frame that contains Churners with **Credit*\_Limit*** between “2400 to 5200”. Display the top and last 6 rows of the data frame**(2mks)**

**# Select only churners with Credit\_Limit between 2400 and 5200**

**credit\_limit\_df <- subset(bnk\_crs, Credit\_Limit >= 2400 & Credit\_Limit <= 5200)**

**# Display the new data frame**

**print(credit\_limit\_df)**

## Question 3 (12mks)

1. Discuss process analysis workflow**(3mks)**

**This is examining an organization’s workflow usually in an attempt to understand it and optimize it to improve it for the future. How do we do this? From our analysis of say a certain process, we get to see redundancies in the process and bottlenecks among other things that come up. With this we now try to fit our intended outcome of the workflow with the actual result of that workflow.**

1. Using “Bank Churners Dataset create a scatter plot using ggplot2, where each plot shows the relationship between **“Months\_on\_book”** and **“Credit\_Limit”** and show the different **education levels** in your plot. . **(3mks)**

**# Load libraries**

**library(ggplot2)**

**library(dplyr)**

**# Convert Education\_Level column to factor**

**bnk\_crs$Education\_Level <- factor(bnk\_crs$Education\_Level)**

**# Create the scatter plot with education levels**

**ggplot(data = bnk\_crs, aes(x = Months\_on\_book, y = Credit\_Limit, color = Education\_Level)) +**

**geom\_point() +**

**labs(title = "Relationship between Months\_on\_book and Credit\_Limit by Education\_Level",**

**x = "Months\_on\_book", y = "Credit\_Limit")**

**# implemented facet wrap**

**ggplot(data = bnk\_crs, aes(x = Months\_on\_book, y = Credit\_Limit, color = Education\_Level)) +**

**geom\_point() +**

**facet\_wrap(~ Marital\_Status) +**

**labs(title = "Relationship between Months\_on\_book and Credit\_Limit by Education\_Level and Marital\_Status",**

**x = "Months\_on\_book", y = "Credit\_Limit")**

1. In a sample of 75 students, the mean of test 1 is 20 and standard deviation is 4.5. Assuming the distribution to be normal, find
   1. How many students scored between 15 and 22? (**2mks)**

**40**

**Graphical user interface, application, table, Excel

Description automatically generated**

* 1. How many students scored above 23? **(2mks)**

**56**

**Graphical user interface, application, table, Excel

Description automatically generated**

* 1. How many students scored less 19? **(2mks)**

**31**

**Graphical user interface, application, table, Excel

Description automatically generated**

**Formulas**

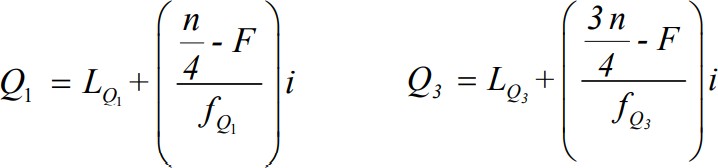
1.



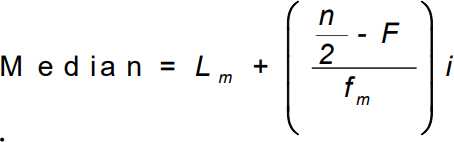
2.



3.



4.

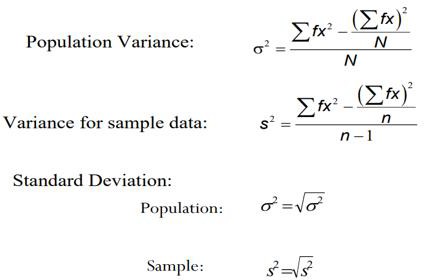


5.

IQR = Q3 – Q1

7.

6



7.



8. Regression equation of x on Y

