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

i. **Attempt all the questions**
ii. **Use Excel and R where applicable**
iii. **Paste the output of each question on your answer sheet**
iv. **Make sure you submit the right document**

Question 1 (10mks)

a) 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

i) Find the correlation coefficient and interpret your result **(2mks)**
ii) Predict the test score for a student with 7 absences **(2mks)**

b) Discuss the atomic classes in R **(2mks)**

c) 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)
```

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

Question 2 (8mks)

e) Using "Bank Churners" Dataset attached, Conduct Exploratory data analysis on the dataset and comment on few interesting observations **(3mks)**

f) 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)**

g) 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)**

Question 3 (12mks)

a) Discuss process analysis workflow**(3mks)**

b) 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. Use facet_wrap() to arrange the plots based on **Marital status. (3mks)**

a) 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

i) How many students scored between 15 and 22? (**2mks)**
ii) How many students scored above 23? **(2mks)**
iii) How many students scored less 19? **(2mks)**

Formulas

1. $$X - \overline{X} = r \frac{\sigma_x}{\sigma_y} (Y - \overline{Y})$$

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

2.

$$\hat{Y} = a + bx$$

3.

$$Q_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F}{f_{Q_1}} \right) i \qquad Q_3 = L_{Q_3} + \left(\frac{\frac{3n}{4} - F}{f_{Q_3}} \right) i$$

4.

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

.

5.

$$IQR = Q_3 - Q_1$$

7.

Population Variance: $$\sigma^2 = \frac{\sum fx^2 - \frac{(\sum fx)^2}{N}}{N}$$

Variance for sample data: $$s^2 = \frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}$$

Standard Deviation:

Population: $$\sigma^2 = \sqrt{\sigma^2}$$

Sample: $$s^2 = \sqrt{s^2}$$

7.

Finding the y-intercept $b = \frac{\sum x}{n} - m\frac{\sum y}{n}$:

8. Regression equation of x on Y