Register Number:

DATE:

ST. JOSEPH'S COLLEGE (AUTONOMOUS), BENGALURU-27

M.Sc. STATISTICS - II SEMESTER

SEMESTER EXAMINATION - JULY 2022

ST 8521 – INTRODUCTION TO DATA SCIENCE

Time: 1½ Hours **Max Marks: 35**

This question paper has **THREE** printed pages and **TWO** sections

Note: Scientific calculators are allowed.

SECTION – A

I Answer any THREE of the following: 3x 3 = 9

1. Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. A data frame with 272 observations on 2 variables. First being eruptions time in minutes and second being Waiting time to next eruption in minutes. Descriptive summary of the data is given below.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.600   2.163   4.000   3.488   4.454   5.100
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   43.0    58.0    76.0    70.9    82.0    96.0
> cor(x,y)
[1] 0.9008112
```

   a) Observing the descriptive statistics given, comment on it.
   b) Give your impression the relation between variables using pearson correlation coefficients provided.
   c) List the analysis that can be carried out on the data set.

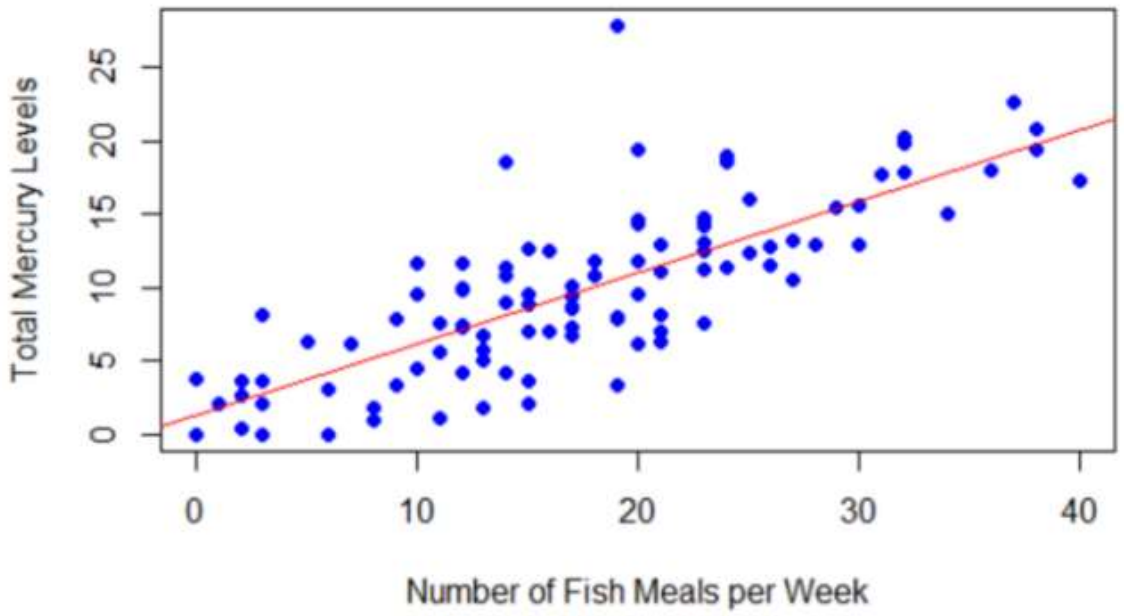
2. Mention at least three different methods/algorithms to sort a dataset involving both variables and categorical data.
3. Create a 4x4 matrix and illustrate the code to find its inverse.
4. What is EDA and why should one perform it.
5. Define any three types of objects in R?

SECTION – B

II Answer any TWO of the following: 13 x2 = 26

6. A study is conducted to see if there was a significant linear relationship between the number of fish meals consumed per week and the total mercury levels found amongst fishermen. Following are the outputs obtained on analyzing the data.

Fishermen Fish Meal Consumption and Mercury Levels



Where:

$\hat{y}$ is the expected/average predicted value for a given x

$\widehat{\beta_0}$ and $\widehat{\beta_1}$ are least-square estimates for β_0 (y-intercept) and β_1 (slope)

The estimates for β_0 and β_1 are given by the following equations:

$$\boldsymbol{\beta_1 = r\frac{s_y}{s_x}} \text{ and } \boldsymbol{\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}}$$

$s_y = 5.771316$
$s_x = 9.298088$
$\bar{x} = 17.51$
$\bar{y} = 9.81$

$$\beta_1 = 0.78\frac{5.771316}{9.298088} = 0.4841454 \approx 0.4841$$

$$\beta_0 = \bar{y} - \widehat{\beta_1}\bar{x} = 9.81 - (0.4841 * 17.51) = 1.333914 \approx 1.3339$$

Therefore, the equation for the least-squares regression line is:

$$\hat{y} = 1.3339 + 0.4841x$$

	SS (Sum of Squares)	Df (Degrees of Freedom)	MS (Mean Square)	F-Statistic	p-value
Regression	1995.8	1	1995.84	150.26	2.2e-16
Residual	1301.7	98	13.28		
Total	3297.5				

Table 2: Estimates, standard error, t-statistic and p-value for intercept and number of fish meals

	Estimate	SE	t-statistic	p-value
Intercept	1.35583	0.78014	1.738	0.0854
Fish Meals	0.48289	0.03939	12.258	<2e-16

a) Draw the valid conclusions using above outputs to meet the objective mentioned above.

b) List out the steps involved in validating the above statistical methods/models used.

c) Define P-value and its significance.

(5+4+4)

7. Explain the steps involved in cleaning the following data. (13)

Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission
Ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual
sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual
Ciaz	2017	7.25	9.85	6900	Petrol	Owner	Manual
wagon r	2011		4.15	5200		Dealer	Manual
Swift	NA	4.6	6.87	42450	Diesel	Dealer	Manual
Brezza	2018	9.25	9.83	NA	Diesel	Owner	Automatic
Ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual
s cross	2015	6.5	8.61	33429	Diesel	Dealer	Automatic
Ciaz	2016	8.75	8.89	20273	Diesel	Owner	Manual
Ciaz	2015	7.45	8.92	42367	Diesel	Dealer	
alto 800	2017		3.6	2135	Petrol	Dealer	Manual
Ciaz	2015	6.85	100.38	51000	Diesel	Dealer	Manual
Ciaz	2015	7.5	9.94	15000	Petrol	Owner	Automatic
Ertiga	2015	6.1	7.71	26000	Petrol	Dealer	Manual
Dzire	2009	2.25	7.21	77427	Petrol	Dealer	Manual

8. a) What are some of the outlier detection techniques? Illustrate a code for the same.
   b) Briefly explain the various graphs one plots after fitting a model.
   What is the code for the same?
   c) Why does one transform data? Illustrate a code used for transforming a variable.
   d) What is the use of set.seed () function.
   e) Differentiate between support vector machine and Decision Trees.

(3+5+2+1+2)