1. Modern email servers and anti-spam filters attempt to identify spam emails and direct them to a junk folder. There are various ways to detect spam, and research still continues. In this regard, an information security officer tries to confirm that the chance for an email to be spam depends on whether it contains images or not. The following data were collected on n = 1000 random email messages.

Spam status	**Image containing status**		
	With image	No image	Total
With spam	160	240	400
No spam	140	460	600
Total	300	700	1000

   Assess whether being spam and containing images are independent factors at 1% level of significance.

2. The following data related to the number of children classified according to the type of feed and the nature of teeth

	Nature of teeth	
Type of feed	Normal	Defective
Breast	18	12
Bottle	2	13

   Do the information provide sufficient evidence to conclude that type of feeding and nature of teeth are dependent? Use chi square test at 5% level of significance.

3. Social media users use a variety of derives to access social networking, mobile phones are increasingly popular . However, is there a difference in the various age groups in the proportion of social media users who use their mobile phone to access social networking? A study showed the following results for the different age groups

		Age	
Use mobile phones to access social networking	18 – 34	35 – 64	65+
Yes	60	37	14
No	40	63	86

   At the 0.05 level of significance, is there evidence of a different among the age groups with respect to use of mobile phone for accessing social networking?

4. A random sample of 200 married men, all retired, were classified according to education and number of children.

Education	**Number of children**		
	0-1	**2-3**	**Over 3**
Elementary	14	37	32
Secondary	19	42	17

College	12	17	10

Test the hypothesis, at the 1% level of significance, that the number of children is independent of the level of education attained by the father.

5. A psychologist wishes to verify that a certain drug increases the reaction time to given stimulus. The following reaction times (in tenth of seconds) were recorded before and after injection of the drug for each of four subjects

	Subject	1	2	3	4
Reaction time	Before	7	2	12	12
	After	13	3	18	13

   Test at 5% level of significance to determine whether the drug significantly increases reaction time. Use non parametric test
6. What do you mean by non parametric test? Write down advantages of non parametric tests over the parametric tests
7. Bank of Nepal recorded the sex of first 30 customers who appeared last Monday with notation MMFMMFMFFMMMFFMFFMFFMFFMFMMMFF . At the 0.05 level of significance , test the randomness of this sequence
8. Define level of significance. Describe run test with some relevant examples.
9. What do you mean by run? Marks secured by a sample of 15 students in Final exam of Statistics II are found to be 27, 34, 48, 21, 7, 56, 44, 32, 25, 42, 33, 28, 41, 5, 49, Are marks in random order? Use 5% level of significance.
10. What is median test? Following data represents marks secured by students of section A and section B of a college in mid-term exam of statistics II

Section A	30	27	19	22	28	25	9	13	20
Section B	24	28	16	22	19	29	7	11	

    Is there any significant difference in marks of section A and section B? Use median test at 5% level of significance.
11. Two computer makers, A and B, compete for a certain market. Their users rank the quality of computers on a 4-point scale as "Not satisfied", "Satisfied", "Good quality", and "Excellent quality, will recommend to others. The following counts were observed:

Computer maker	Not satisfied	Satisfied	Good quality	Excellent quality
A	20	40	70	20
B	10	30	40	20

    Is there a significant difference in customer satisfaction of the computers produced by A and by B using Mann-Whitney U test at 5% level of significance.

12. A chemist uses three catalysts for distilling alcohol and layout were tabulated below

Catalyst	Alcohol(in cc)				
C_1	380	430	410		
C_2	290	350	270	250	270
C_3	400	380	450		

Are there any significant differences between catalyst? Test at 5% level of significance. Use Kruskal Walli's H test.

13. There are three brand of computers Dell, Lenovo and HP . The following are life time of 15 computers in years

Serial number	Computer brand	Life time in years
1	Dell	15
2	Lenovo	10
3	HP	9
4	Dell	12
5	Lenovo	6
6	HP	7
7	Dell	4
8	Lenovo	8
9	HP	13
10	Dell	11
11	HP	5
12	Lenovo	7
13	Dell	3
14	HP	5
15	Lenovo	4

Apply appropriate statistical test to identify whether the average life time in years is significantly different across three brand of computers at 5% level of significance. You can again tabulate data initially in the required format for statistical analysis

14. Marks secured by students in three chapter tests in a subject are as follows

Student	A	B	C	D	E	F	G	H
Chapter test I	13	11	16	19	6	14	18	5
Chapter test II	14	10	18	11	12	9	18	7

Chapter test III	15	19	13	10	11	5	17	4

Is there any significant difference in marks in three chapter tests? Use Friedman's two way ANOVA test at 10% level of significance

15. It was reported somewhere that children whenever plays the game in computer, they used the computer very roughly which may reduce the lifetime of a computer. The random access memory (RAM) of a computer also plays a crucial role on the lifetime of a computer. A researcher wanted to examine how the lifetime of a personal computer which is used by children is affected by the time (in hours) spends by the children per day to play games and the available random access memory (RAM) measured in megabytes (MB) of a used computer. The data is provided in following table.

Lifetime (years)	5	1	7	2	3	4	6
Play time (hours)/day	2	8	1	5	6	3	2
RAM in MB)	8	2	6	3	2	4	7

Identify which one is dependent variable? Solve this problem using multiple linear regression model and provide problem specific interpretations based on the regression model developed.

16. What are required conditions for error variable in multiple regression analysis? The Internal Revenue Service is trying to estimate the monthly amount of unpaied taxes discovered by its auditing division. The Internal Revenue Service estimated this figure on the basis of field auditing labour hours and numbers of hours of its computers are used. The table given below presents these data for the last ten months

Month	(x1) Field audit labour hours in 100	(x2) Computer hours in 100	(y)Annual unpaid taxes discovered million of dollars
Jan	45	16	29
Feb	42	14	24
Mar	44	15	27
Apr	45	13	25
May	43	13	26
Jun	46	14	28
Jul	44	16	30
Aug	45	16	28
Sep	44	15	28
Oct	43	15	27

Given $\sum yx1 = 12005$, $\sum yx2 = 4013$, $\sum x1x2 = 6485$, $\sum y\text{^}2 = 7428$, $\sum x1\text{^}2 = 19461$, $\sum x2\text{^}2 = 2173$

(i) Develop the estimating equation best describing these data

(ii) Interpret the value of regression coefficients

(iii) Estimate the actual unpaid tax for field audit labour hour is 4200 and computer hours is 1600 hours

17. A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data and how many tables are used to arrange each data set. Efficiency will be measured the number of processed requests per hour. Applying the program to data set of different sizes and number of tables are used , she gets the following results.

Processed requests, Y	16	26	17	41	50	55	40
Data size (Giga bites)X1	15	10	10	8	7	7	6
Number of tables X2	1	2	10	10	20	20	4

The regression equation obtained is Y = 52.7- 2.87x1 + 0.85 x2

Total sum of square = 1452

Sum of square due to regression = 1143.3

a) Interpret the values of regression coefficients b1 and b2
b) Test the significance of the regression model at 0.05 level of significance
c) Is there significant relationship between processed request and number of tables at 0.05 level of significance? Given standard error of b2 = 0.55
d) What percentage of variation of processed requests is explained by data size and number of tables?
e) Compute standard error of estimate.
f) Estimate the number of processed requests if data size is 9 Giga bites and number of tables used are 8

18. A computer manager is keenly interested to know how efficiency of her new computer program depends on the size of incoming data and data structure. Efficiency will be measured by the number of processed requests per hour. Data structure may be measured on how many tables were used to arrange each data set. All the information was put together as follows.

Data size(gigabytes)	6	7	7	8	10	10	15
Number of tables	4	20	20	10	10	2	1
Processed requests	40	55	50	41	17	26	16

Identify which one is dependent variable? Fit the appropriate multiple regression model and provide problem specific interpretations of the fitted regression coefficients.

19. What is multiple Linear Regression(MLR)? From following information of variables x_1 , x_2 and y.
$\Sigma x_1 = 272$, $\Sigma x_2 = 441$, $\Sigma y = 147$, $\Sigma x_1^2 = 7428$, $\Sigma x_2^2 = 19461$,
$\Sigma y^2 = 2137$, $\Sigma x_1 y = 4013$, $\Sigma x_1 x_2 = 12005$, $\Sigma x_2 y = 6485$, n = 10 . Fit a regression equation of y on x_1 and x_2

20. Suppose we are given following information with n = 7, multiple regression model is
$\hat{y}$ = 8.15 + 0.56x_1 + 0.54x_2
Here, Total sum of square = 1493
Sum of square due to error = 91
Find (i) R^2 and interpret it (ii) Test the overall significance of model

21. Define multiple correlation. In a trivariate distribution X1, X2 and X3, the simple correlation coefficients are given as r12 = 0.5, r23 = 0.6 r13 = 0.7 find
(i) Partial correlation coefficient between X1 and X2 keeping X3 constant
(ii) Multiple correlation coefficient assuming X1 is dependent variab

22. The following ANOVA summary table was obtained from a multiple regression model with two independent variable

SV	SS	df	MS	F ratio
Regression	12.62	2	?	?
Error	0.78	12	?	
Total	13,4	14		

(i) Determine the mean sum of square due to regression, the mean sum of square due to error and F value

(ii) Test the significance of the overall regression model at 5% level of significance

(iii) Compute coefficient of determination and interpret its value

(iv) Find standard error of estimate

23. Write short notes on

i. Partial and multiple correlation coefficient

ii. Required assumptions for linear regression model

iii. Rationale of using non parametric statistical test