Data Analytics (IT – 3006)
Practice Questions (Unit 2)

Q1. What do you understand by data exploration? Illustrate the answer with an example.

Q2. Why is data exploration important?

Q3. What is the Difference between univariate, bivariate, and multivariate analysis?

Q4. Name few tools for exploratory data analysis.

Q5. What are the advantages of using exploratory data analysis?

Q6. A survey will be given to 100 students randomly selected from the freshmen class at Lincoln High School. What is the sample and population?

Q7. Fifty bottles of water were randomly selected from a large collection of bottles in a company's warehouse. These fifty bottles are referred to as ___ and the large collection of bottles is referred to as ___?

Q8. A group of librarians is interested in the numbers of books and other media that patrons check out from their library. They examine the checkout records of 150 randomly selected adult patrons. Identify the population and sample in this setting.

Q9. Marco is conducting an experiment on training certain breeds of dogs. He wants to know how long, on average, it would take to teach a Labrador to fetch an object. He gets a group of dogs to conduct his experiment. 5 of the dogs are Labradors and 3 of the dogs are Dalmatians. What is the population and sample in this experiment?

Q10. A school takes a poll to find out what students want to eat at lunch. 70 students are randomly chosen to answer the poll questions. What are the population and the sample?

Q11. The marks obtained out of 25 by 30 students of a class in the examination are: 20, 6, 23, 19, 9, 14, 15, 3, 1, 12, 10, 20, 13, 3, 17, 10, 11, 6, 21, 9, 6, 10, 9, 4, 5, 1, 5, 11, 7, and 24. Draw the frequency distribution table and frequency distribution graph.

Q12. Weekly pocket expenses (in $) of 30 students of class VIII are 37, 41, 39, 34, 71, 26, 56, 61, 58, 79, 83, 72, 64, 39, 75, 39, 37, 59, 57, 37, 53, 38, 49, 45, 70, 82, 44, 37, 79, 76. Construct the frequency distribution table and frequency distribution graph with the class interval of equal width such as 30 - 35. Also, find the range of the weekly pocket expenses.

Q13. Construct a frequency distribution table and frequency distribution graph for the following weights (in gm) of 30 oranges using the equal class intervals, one of them is 40-45. The weights are: 31, 41, 46, 33, 44, 51, 56, 63, 71, 71, 62, 63, 54, 53, 51, 43, 36, 38, 54, 56, 66, 71, 74, 75, 46, 47, 59, 60, 61, and 63. In addition, answer the following:

- How many class intervals are there?
- What is the range of the above weights?
- Which class interval has the lowest frequency?
- Which class interval has the highest frequency?

Q14. Find the mean of the following data.

- 9, 7, 11, 13, 2, 4, 5, 5
- 16, 18, 19, 21, 23, 23, 27, 29, 29, 35
- 2.2, 10.2, 14.7, 5.9, 4.9, 11.1, 10.5

Q15. The mean of 8, 11, 6, 14, x and 13 is 66. Find the value of the observation x.

Q16. The mean of 6, 8, x + 2, 10, 2x - 1, and 2 is 9. Find the value of x and also the value of the observation in the data.

Q17. Find the mean of the following distribution.

(a)

Age in Years	12	10	15	14	8
Number of Boys	5	3	2	6	4

(b)

Marks	25	30	15	20	24
Number of Students	8	12	10	6	4

(c) The daily wages of 50 employees in an organization are given below:

Daily wages in INR	100 - 150	150 - 200	200 - 250	250 – 300
Number of Students	12	13	17	8

Q18. The runs scored in a cricket match by 11 players are: 7, 16, 121, 51, 101, 81, 1, 16, 9, 11, and 16. Find the mean, mode, median of this data.

Q19. The weights in kg of 10 students are: 39, 43, 36, 38, 46, 51, 33, 44, 44, 43. Find the mode of this data. Is there more than 1 mode? If yes, why?

Q20. The marks obtained by 40 students out of 50 in a class are given below. Find the mode of the above data.

Marks	42	36	30	45	50
Number of Students	7	10	13	8	2

Q21. The following observations are arranged in ascending order. The median of the data is 25 find the value of x. The observations: 17, x, 24, x + 7, 35, 36, 46

Q22. The mean of the following distribution is 26. Find the value of p and also the value of the observation.

x_i	0	1	2	3	4	5
f_i	3	3	p	7	p - 1	4

Q23. The number of students in 7 different classes is given below. Represent this data on the bar chart and pie chart.

Class	6th	7th	8th	9th	10th	11th	12th
Number of Students	130	120	135	130	150	80	75

Q24. The weekly sale of pencil boxes in a stationary shop is given in the table below. Using a suitable scale, represent the given information on a bar chart.

Day	Mon	Tues	Wed	Thurs	Fri	Sat
Pencil Boxes Sold	10	25	30	40	50	10

From the bar chart, answer the following:

- On which day were the maximum pencil boxes sold?
- If the shopkeeper decides to close his shop for one more day each week, selection of which days would lead to minimum loss of sale and maximum loss of sale?

- On which day were the least number of pencil boxes sold?
- Which two days was equal number of pencil boxes sold?

Q25. A survey was conducted with 1000 participants out of which 55% are female and 45% are male. From 55% of female, 65% are married. From 45% of male, 75% are married. In addition, from 55% of female, 25 % belongs to place 1, 35 % belongs to place 2 and rest belongs to place 3. From 45% of female, 35 % belongs to place 1, 45 % belongs to place 2 and rest belongs to place 3. Considering the concept of descriptive measures for categorical variables, draw the frequencies in the following table.

Male	Female

	Female	Male
Married		
Single		

		Place 1	Place 2	Place 3
Married	Female			
	Male			
Single	Female			
	Male			

Q26. Considering the concept of descriptive measures for categorical variables, draw the proportions in reference to the problem statement of Q25 in 3 tables (as per Q25).

Q27. A survey was conducted with 1500 participants out of which 45% are female and 55% are male. From 45% of female, 65% are married. From 55% of male, 75% are married. Draw the following tables by capturing:

- Row and column wise frequency marginal
- Row and column wise percentage marginal

Q28. In reference to Q27 and consideration of dummy variables, draw the following tables by capturing:

- Row and column wise frequency marginal
- Row and column wise percentage marginal

Q29. Find the inter quartile range (IQR), Q0, Q1, Q2, Q3 and Q4 for the data set: 23, 45, 32, 29, 37, 47, 21, 36, and 52.

Q30. A mutual fund achieved the following rates of growth over an 11-month period: {3% 2% 7% 8% 2% 4% 3% 7.5% 7.2% 2.7% 2.09%}. Determine 75^{th}, 25^{th}, 85^{th}, 50^{th}, 90^{th}, and 5^{th} percentile.

Q31. Explain the empirical rules for interpreting standard deviation with standard normal distribution diagram.

Q32. Given the following distribution of returns, check the validity of 68-95-99 rule and justify the claim.

{10% 23% 12% 21% 14% 17% 16% 11% 15% 19%}

Q33. What is the skeweness of the normal distribution?

Q34. What is the coefficient of skewness and kurtosis of the following distribution?

Score	**Frequency**
60	3
65	4
70	3
75	4

80	2
85	3
90	2
55	4

Q35. Using the data from dataset (12, 13, 54, 56, 25), determine excess kurtosis and the type of kurtosis present i.e., Mesokurtic/Platykurtic/Leptokurtic distribution.

Q36. Using the data from dataset (42, 20, 38, 78, 54, 26), determine excess kurtosis and the type of kurtosis present i.e., Mesokurtic/Platykurtic/Leptokurtic distribution.

Q37. What are outliers? What are the various ways to detect them?

Q38. Detect the outlier from the list: [20, 24, 22, 19, 29, 18, 4300, 30, 18] with box plot analysis.

Q39. Determine the outliers of the dataset: [35, 75, 20, 25, 15, 30, 30, 15, 45, 40, 110] using z-score.

Q40. For the data set including values 2, 5, 6, 9, 12, determine five-number summary.

Q41. Using the data from dataset (42, 20, 38, 78, 54, 26, 150, 125), determine lower and upper whisker.

Q42. Which imputation is better for numerical data with missing value, mean or median? What is the reason behind them?

Q43. What is the difference between univariate and multivariate imputation of the missing data? Give examples.

Q44. Does missing data have a big impact on analysis?

Q45. What are the types of missing values?

Q46. The following table shows last year's revenue for each location. If the mean of the dataset is 158 thousand rupees, find the revenue for the location D.

Location	**Revenue**
A	Rs. 121 K
B	Rs. 189 K
C	Rs. 147 K
D	

Q47. The total values of food grains (rice and wheat) imported during these years are given below. Draw the scatterplots.

1971 : Rs. 123 crore
1980 : Rs. 80 crore
1981 : Rs. 314 crore
1982 : Rs. 295 crore
1983 : Rs. 587 crore
1984 : Rs. 158 crore

Q48. Is correlation transitive i.e., Suppose that X, Y, and Z are random variables. X and Y are correlated and Y and Z are likewise correlated. Does it follow that X and Z must be correlated?

Q49. Is Pearson correlation coefficient sensitive to outliers?

Q50. Eight tomato plants of the same variety were selected at random in which x grams of fertilizer was dissolved in a fixed quantity of water. This yields y kilograms of tomatoes which were recorded below:

Plant	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
Y	3.9	4.4	5.8	6.6	7.0	7.1	7.3	7.7

Determine correlation coefficient and comment on the relationship.

Q51. Eight tomato plants of the same variety were selected at random in which x grams of fertilizer was dissolved in a fixed quantity of water. This yields y kilograms of tomatoes which were recorded below:

Plant	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	3.9	4.4	5.8	6.6	7.0	7.1	7.3	7.7

Determine correlation coefficient between x and y, and comment on the relationship.

Q52. The dataset shows a verbal reasoning test score x and an English test score y for each of a random sample of 8 children who took both tests.

Child	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
x	112	113	110	113	112	114	109	113
Y	69	65	75	70	70	75	68	76

Determine covariance between x and y, and comment on the relationship.

Q53. Find the relationships of salary between male and female of below sample by illustrating with the box plot.

1	Gender	Salary
2	Male	81600
3	Female	61600
4	Female	64300
5	Female	71900
6	Male	76300
7	Female	68200
8	Male	60900
9	Female	78600
10	Female	81700
11	Male	60200
12	Female	69200
13	Male	59000
14	Male	68600
15	Male	51900

Q54. Table 1 represents the sample of salary drawn by different age of employers working for a retail business. Find the relationships of salary between young, middle and old aged people of below sample by illustrating with the box plot. Young aged are characterised by age range 21 – 44, middle by 45 – 59 and old aged by more than 60.

Table 1: Salary in Thousands

Age	25	45	55	57	65	30	62	61	63	35	42	55	57	32	29	64

Salary	112	113	114	117	115	112	111	108	117	119	121	122	114	105	78	120

Q55. A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with sample of size two and compute the sample mean for each one. Use them to find the probability distribution and plot the sampling distribution.

Q56. A rowing team consists of four rowers who weigh 72, 66, 80, 71, and 84 pounds. Find all possible random samples with sample of size one and compute the sample mean for each one. Use them to find the probability distribution and plot the sampling distribution.

Q57. With the increase of sample size, what would happen to the sampling error?

Q58. The probability of selecting an item in probability sampling, from the population is __?

Q59. Hayley wants to carry out some research on her class. She wants a sample of 12 people out of the 30 in her class. Use a random sampling technique to determine the reference number of the students in the class who should be included in the sample. Do not include duplicated data. List the sample.

Q60. A drinks company produces 1200 bottles of pop every 30 minutes. For quality control purposes, 12 bottles are selected and checked. Each bottle passes through the machine in a single file. Using a systematic sampling technique, determine the bottles that will be selected for the sample.

Q61. The sample with the test scores in data analytics after end semester examination is 55, 65, 80, 95, 90, 90, 95, 75, 75, 85, 90 and 80. Calculate the confidence limit and margin error. The Z-score for 95% confidence level is 1.96.

Q62. Calculate the best point variance estimate from the sample: 15.22, 14.34, 18.12, 12.61, 15.61, 14.22, 19.41, 12.22, 17.12, 14.22, 12.91 and 18.12.

Q63. A sample of 40 packages of rice has a mean weight of 5.7 kg with a standard deviation of 0.4 kg. Find the best estimate of the population mean?

Q64. In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists want to test a new medication to see if it has either a positive or negative effect on intelligence or not effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence? The z-value is 1.96.

Q65. For the following dataset, using Chi-square test for independence, determine whether categorical variables are related to each other or not. We have a list of movie genres; this is the first variable. The second variable is whether or not the patrons of those genres bought snacks at the theater. The idea (or null hypothesis) is that the type of movie and whether or not people bought snacks are unrelated. The owner of the movie theater wants to estimate how many snacks to buy. If movie type and snack purchases are unrelated, estimating will be simpler than if the movie types impact snack sales.

Type of movie	**Snack**	**No snack**
Action	50	75
Comedy	125	175
Family	90	30
Error	45	10

*** The End ***