DEPARTMENT OF ARTIFICIAL INTELLIGECE AND DATA SCIENCE
QUESTION BANK

Year/Sem/Sec: II/IV/A, B, C

Course Code/ Course Name: U23AIT42/DATA SCIENCE

UNIT 1

1. Define Data Science.

2. What are the benefits of Data Science?

3. List the facets of data.

4. What is meant by the Data Science process?

5. What is data retrieval?

6. Define data cleansing.

7. What is data integration?

8. What is Exploratory Data Analysis (EDA)**?**

9. What is meant by building a model in Data Science?

10, What are data-driven applications?

PART B

1. (a) Define Data Science. Explain the need, benefits, and uses of Data Science in modern industries.
(b) Describe the facets of data and explain structured, semi-structured, and unstructured data with examples.
2**.** Explain in detail the Data Science Process with a neat diagram:

- Setting the research goal
- Data retrieval
- Data cleansing
- Data integration
- Data transformation
- Exploratory Data Analysis
- Model building
- Presentation and application deployment

3. (a) What is Data Cleaning? Explain different techniques used for handling missing data and noisy data.
(b) Discuss data integration and transformation methods with suitable examples.

4. Explain Exploratory Data Analysis (EDA) in detail. Discuss various EDA techniques such as summary statistics, visualization tools, and correlation analysis with examples.

UNIT 2

1. Define frequency distribution.

2. What are outliers?

3. What is meant by data variability?

4. Define Interquartile Range (IQR).

5. What is normal distribution?

6. Define Z-score.

7. What is correlation?

8. What is a scatter plot?

9. Define regression line.

10. What is the standard error of estimate

PART B

1. Explain in detail about Describing Data with Tables and Graphs and solve the following with the data

The IQ scores for a group of 35 high school dropouts:

91	85	84	79	80	87	96	75	86
104	95	71	105	90	77	123	80	100
93	108	98	69	99	95	90	110	109
94	100	103	112	90	90	98	89	

i) Construct a frequency distribution for grouped data

ii) Relative Frequency distribution

iii) Cumulative Frequency distribution

2. (a)Discuss about following measures and calculate them with given "residence changes" data.

1, 3, 4, 1, 0, 2, 5, 8, 0, 2, 3, 4, 7, 11, 0, 2, 3, 4

(i)Range (ii) Variance (iii) Standard Deviation (iv) Inter Quartile Range

(b)In a national entrance exam, the mean score is 70 and the standard deviation is 10. The exam board classifies student performance into four subdivisions:

(i)Scores above 85;
(ii)Scores between 70 and 85
(iii)Scores between 55 and 70;
(iv)Scores below 55

3.(a)Explain in detail about correlations and the types of relationships in correlation.

The wind speed X in miles per hour and wave height Y in feet were measured under various conditions on an enclosed deep water sea, with the results shown in the table.

X	0	2	7	9	13	22
Y	0	5	10	14	22	31

Create a scatter plot and predict the type of correlation. (6)

(b)Explain the concept of least squares regression

Assume that an r of –.80 describes the strong negative relationship between years of heavy smoking *(X)* and life expectancy *(Y)*. Assume, furthermore, that the distributions of heavy smoking and life expectancy each have the following means and sums of squares:

$\overline{X} = 5 \quad \overline{Y} = 60$

$SS_x = 35 \quad SS_y = 70$

Determine the least squares regression equation for predicting life expectancy from years of heavy smoking. (7)

4. Explain in detail about standard error of estimate and multiple regressions with example

UNIT 3

1. Define population and sample.
2. What is random sampling?
3. Define sampling distribution**.**
4. What is standard error of the mean?
5. Define hypothesis testing.
6. What is a Z-test?
7. What is a t-test?
8. What is meant by one-tailed test?
9. Define confidence interval**.**
10. What is level of confidence?

PART B

1. Explain in Detail about Hypothesis testing and Z-test and solve the following

According to the American Psychological Association, members with a doctorate and a full-time teaching appointment earn, on the average, $82,500 per year, with a standard deviation of $6,000. An investigator wishes to determine whether $82,500 is also the mean salary for all

female members with a doctorate and a full-time teaching appointment. Salaries are obtained for a random sample of 100 women from this population, and the mean salary equals $80,100.

i. Someone claims that the observed difference between $80,100 and $82,500 is large enough by itself to support the conclusion that female members earn less than male members. Explain why it is important to conduct a hypothesis test.

i. The investigator wishes to conduct a hypothesis test for what population?

iii. What is the null hypothesis, H0?

iv. What is the alternative hypothesis, H1?

v. Specify the decision rule, using the .05 level of significance.

vi. Calculate the value of z. (Remember to convert the standard deviation to a standard error.)

vii. What is your decision about H0?

viii. Using words, interpret this decision in terms of the original problem

4. (i)Explain one tailed and two tailed test and solve the following

Reading achievement scores are obtained for a group of fourth graders. A score of 4.0 indicates a level of achievement appropriate for fourth grades, a score below 4.0 indicates under achievement. and a score above 4.0 indicates over achievement. Assume that the population standard deviation equals 0.4. A random sample of 64 fourth graders reveals a mean achievement score of 3.82. Construct a 95% confidence interval for the unknown population mean. (Remember to convert the standard deviation to a standard error). Interpret this confidence interval; that is, do you find any consistent evidence either of overachievement or of underachievement?

UNIT 4

1. Define t-test.

2. What is statistical significance?

3. What is two independent sample t-test?

4. Define paired sample t-test**.**

5. What is an F-test?

6. Define ANOVA**.**

7. What is two-factor ANOVA?

8. What is meant by factorial experiment**?**

9. Define chi-square test**.**

10. What is the purpose of ANOVA?

PART B

1. (i) Explain about t-test with formula. A library system lends books for the periods of 21 days. This policy is being reevaluated in view of a possible new loan period that could be either longer or shorter than 21 days. To aid in making this decision, books-lending records were consulted to determine the loan period actually used by the patrons. A random sample of 8 records revealed the following loan periods in days: 21,15,12,24,20,21,13 and 16. Test the null hypothesis with t-test, using the .05 level of significance

2. A random sample of 90 college students indicates whether they most desire love, wealth, power, health, fame, or family happiness.

i. Using the .05 level of significance and the following results, test the null hypothesis that, in the underlying population, the various desires are equally popular.

ii. Specify the approximate p-value for this test result. (APRIL/MAY 2023)

DESIRES OF COLLEGE STUDENTS

FREQUENCY	LOVE	WEALTH	POWER	HEALTH	FAME	FAMILY HAP.	TOTAL
Observed (f_o)	25	10	5	25	10	15	90

3. (a)Blood pressure of 8 patients is before and after is recorded: Before: 180,200,230, 240,170,190,200 and 165 after: 140,145, 150,155,120,130,140 and 130. Find is there any significant difference between BP reading before and after by applying two-sample t-test.

(b) Explain in detail about the chi-square test with an example.

4. (i)Illustrate in detail about one factor and two factors ANOVA with it tables

(ii)Apply one factor ANNOVA for the following data. Estimate the calculation of Sum of Squares (Two-Factor ANOVA)

Teaching Method	Student 1	Student 2	Student 3	Student 4	Student 5
Traditional	85	88	90	86	84
Online	78	74	80	76	82
Hybrid	92	90	94	96	91

UNIT 5

1. Define predictive analytics**.**

2. What is linear regression**?**

3. What is goodness of fit**?**

4. What is logistic regression?

5. Define weighted resampling**.**

6. What are nonlinear relationships**?**

7. What is time series analysis**?**

8. Define moving averages**.**

9. What is autocorrelation**?**

10. What is survival analysis**?**

PART B

1. Solve the following data

X1	X2	Y
2	1	5
4	3	10
6	5	15
8	7	20

(a) Fit a multiple regression model.
(b) Interpret coefficients and R^2.

1) (a) Explain Logistic Regression model with equation and assumptions.
   (b) Differentiate between Linear Regression and Logistic Regression.
2) Given regression output:
   SST = 200, SSE = 50 (a) Calculate R^2. (b) Interpret goodness of fit. (c) Explain adjusted R^2.

4. Given quarterly sales data solve the following:

Quarter	Sales
Q1	200
Q2	250
Q3	300
Q4	350

(a) Compute 2-period moving average.
(b) Comment on trend.

SUBJECT INCHARGE(s) **HOD**