



**SRM VALLIAMMAI ENGINEERING COLLEGE**

**(An Autonomous Institution)**

SRM Nagar, Kattankulathur – 603 203



**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**QUESTION BANK**



**IV SEMESTER**

**AD3464 FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS**

**Regulation – 2023**

**Academic Year 2024 – 2025(EVEN SEMESTER)**

*Prepared by*

**R. Vaishnavi, Assistant Professor (O.G)**

---



# SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM Nagar, Kattankulathur – 603 203



## DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

### QUESTION BANK

**SUBJECT: FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS**

**SEM / YEAR: IV SEMESTER/ SECOND YEAR**

<b>UNIT 1 INTRODUCTION TO DATA SCIENCE</b>			
Need for data science - benefits and uses - facets of data - data science process - setting the research goal - retrieving data - cleansing, integrating, and transforming data - exploratory data analysis - build the models - presenting and building applications.			
<b>PART – A</b>			
<b>Q.No</b>	<b>Question</b>	<b>Level</b>	<b>Competence</b>
1	Define Data Science.	BTL 1	Remember
2	What is Bigdata?	BTL 2	Understand
3	What is machineLearning?	BTL 1	Remember
4	Define DataMining?	BTL 1	Remember
5	List the characteristics of bigdata.	BTL 1	Remember
6	Mention the categories of data.	BTL 2	Understand
7	List some of the application domains of datascience.	BTL 1	Remember
8	What is structured data? Give some examples.	BTL 1	Remember
9	Define unstructured data. Give examples.	BTL 1	Remember
10	What is machine generated data	BTL 2	Understand
11	Why the data is to be cleaned.	BTL 2	Understand
12	List the phases involve in the data science process.	BTL 1	Remember
13	What is meant by data cleaning?	BTL 2	Understand
14	What is project charter?	BTL 1	Remember
15	Identify the important contents of a project charter.	BTL 1	Remember

16	List some of the visualization techniques	BTL 2	Understand
17	Name some problems associated with real world data.	BTL 2	Understand
18	Define data warehouse, datamart and datalake.	BTL 2	Understand
19	List some of the factors involved in selecting the modeling technique.	BTL 2	Understand
20	What is a dummy variable?	BTL 1	Remember
21	What do you meant by exploratory data analysis?	BTL 1	Remember
22	List out the methods for combining data from different table.	BTL 1	Remember
23	Why we need to build a model?	BTL 2	Understand
24	On what factors the modeling technique is being selected.	BTL 2	Understand

### PART – B

Q.No	Question	Level	Competence
1	Discuss the applications of data science and bigdata with suitable examples.	BTL 6	Create
2	Illustrate the overview of the data science process.	BTL 4	Analyze
3	Elaborate any five application domains of datascience.	BTL 5	Evaluate
4	Describe the categories of data for data mining.	BTL 3	Apply
5	Discuss the significance of setting the research goal for the data science project.	BTL 4	Analyze
6	Discuss the categories involved in retrieving relevant data from different sources of data.	BTL 5	Evaluate
7	Explain the different stages of data preparation phase.	BTL 6	Create
8	Elucidate the techniques involved in data cleansing.	BTL 4	Analyze
9	Illustrate the steps involved in combining data from different data sources.	BTL 6	Create
10	Explain the impact of variable reduction on data science project highlighting its pros and cons.	BTL 6	Create
11	Elaborate on the steps involve in model building with suitable diagrams.	BTL 3	Apply
12	Discuss briefly about facets of data.		
13	Justify Exploratory Data Analysis.	BTL 4	Analyze
14	Explain briefly about Data science and its life cycle.		
15	Compare and contrast Data science and Big Data.	BTL 4	Analyze
16	Compare and contrast Cloud Computing and Big Data.	BTL 4	Analyze
17	Explain the impact of Big Data technologies on the field of Data Science. How do these technologies enhance the capacity of Data Scientists to solve complex problems?	BTL 3	Apply

## UNIT 2 DESCRIPTIVE ANALYTICS

Frequency distributions - Outliers - interpreting distributions – graphs - averages – describing variability - interquartile range - variability for qualitative and ranked data - Normal distributions - z scores – correlation - scatter plots – regression - regression line - least squares regression line - standard error of estimate - interpretation of  $r^2$  - multiple regression equations - regression toward the mean.

### PART – A

Q.No	Question	Level	Competence
1	What is meant by frequency distribution?	BTL 1	Remember
2	What is meant by qualitative data? Give examples.	BTL 2	Understand
3	What is meant by quantitative data? Give examples.	BTL 1	Remember
4	Differentiate qualitative and quantitative data	BTL 1	Remember
5	Compare discrete and continuous variables.	BTL 1	Remember
6	State the difference between nominal and ordinal data	BTL 1	Remember
7	Mention the types of frequency distribution?	BTL 1	Remember
8	Define an outlier?	BTL 2	Understand
9	What is percentile rank?	BTL 1	Remember
10	Provide the equation for percentile rank.	BTL 2	Understand
11	State the differences between a histogram and bar graph.	BTL 2	Understand
12	Give the measures of central tendency	BTL 2	Understand
13	Define mode	BTL 1	Remember
14	Define median.	BTL 1	Remember
15	What is the interpretation of $r^2$ ?	BTL 1	Remember
16	What is the standard error of estimate?	BTL 2	Understand
17	Define standard deviation.	BTL 1	Remember
18	What is normal curve?	BTL 2	Understand
19	Define z-score.	BTL 2	Understand
20	Give the equation for z-score.	BTL 1	Remember
21	How will convert the z-score to the original score.	BTL 1	Remember
22	Define Correlation.	BTL 1	Remember
23	Mention the types of correlation.	BTL 2	Understand
24	Define Scatterplot	BTL 2	Understand

## PART – B

Q.No	Question	Level	Competence																						
1	Explain the different types of frequency distribution with suitable examples and diagrams.	BTL 4	Analyze																						
2	Elaborate the different ways to describe or represent data using tables with suitable examples.	BTL 5	Evaluate																						
3	Explain the various ways by which data can be represents or describes using graphs with suitable examples.	BTL 4	Analyze																						
4	Compute the mean, median and mode for the following datasets .I)9,10,12,13,13,13,15,15,16,16,18,22,23,24,24,25	BTL 3	Apply																						
5	The following data are the shoe sizes of 50 male students. The sizes are discrete data since shoe size is measured in whole and half units only. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars. 9;9;9.5;9.5;10;10;10;10;10;10;10.5;10.5;10.5;10.5;10.5;10.5;10.5;11;11;11;11;11;11;11;11;11;11;11;11;11;11.5;11.5;11.5;11.5;11.5;11.5;11.5;11.5;12;12;12;12;12;12;12;12;12.5;12.5;12.5;12.5;14	BTL 3	Apply																						
6	What are scatterplots? Illustrate on the various types with suitable examples.	BTL 5	Evaluate																						
7	Elaborate on the correlation coefficient. Compare the various correlation coefficients																								
8	Explain the characteristics of a normal distribution. Discuss why the normal distribution is widely used in statistics and how it relates to other probability distributions. How can you check if a dataset approximates a normal distribution?	BTL 6	Create																						
9	What is a z-score, and how is it used to standardize data in statistical analysis? Discuss its role in comparing data points from different distributions and how it helps in identifying outliers.	BTL 4	Analyze																						
10	Explain the meaning and interpretation of $r^2$ in the context of regression analysis. How does $r^2$ help assess the goodness of fit of a model, and what does it reveal about the relationship between independent and dependent variables?	BTL 5	Evaluate																						
11	Find Karl Pearson's Correlation Coefficient for the following paired data. <table border="1" style="margin-left: 20px;"> <tbody> <tr> <td>X</td> <td>38</td> <td>45</td> <td>46</td> <td>38</td> <td>35</td> <td>38</td> <td>46</td> <td>32</td> <td>36</td> <td>38</td> </tr> <tr> <td>Y</td> <td>28</td> <td>34</td> <td>38</td> <td>34</td> <td>36</td> <td>36</td> <td>28</td> <td>29</td> <td>25</td> <td>26</td> </tr> </tbody> </table>	X	38	45	46	38	35	38	46	32	36	38	Y	28	34	38	34	36	36	28	29	25	26	BTL 6	Create
X	38	45	46	38	35	38	46	32	36	38															
Y	28	34	38	34	36	36	28	29	25	26															
12	Discuss Multiple Regression Equations.	BTL 5	Evaluate																						
13	A random sample of 5 college students is selected and their grades in operating system and software engineering are found to be? <table border="1" style="margin-left: 20px;"> <tbody> <tr> <td>Subject</td> <td>1</td> <td>2</td> <td>3</td> <td>4</td> <td>5</td> </tr> <tr> <td>Operating System</td> <td>85</td> <td>60</td> <td>73</td> <td>40</td> <td>90</td> </tr> <tr> <td>Software Engineering</td> <td>93</td> <td>75</td> <td>65</td> <td>50</td> <td>80</td> </tr> </tbody> </table> Calculate Pearson's rank correlation coefficient?	Subject	1	2	3	4	5	Operating System	85	60	73	40	90	Software Engineering	93	75	65	50	80	BTL 4	Analyze				
Subject	1	2	3	4	5																				
Operating System	85	60	73	40	90																				
Software Engineering	93	75	65	50	80																				
14	Define the interquartile range (IQR) and explain its significance in understanding data spread. How can the IQR be used to identify outliers, and what does it reveal about the central tendency and	BTL 4	Analyze																						

	distribution of the data?		
15	Explain how measures of variability (such as range, variance, and standard deviation) describe the spread of data. Provide examples of when each measure is most appropriate to use.	BTL 5	Evaluate
16	Discuss the phenomenon of regression toward the mean and how it can influence statistical interpretation. Discuss how this concept is related to the correlation between variables and its implications in predictive modeling.	BTL 6	Create
17	Describe the process of simple linear regression analysis and explain the significance of the regression line.	BTL 4	Analyze

### UNIT 3 - INFERENTIAL STATISTICS

Populations – samples - random sampling - Sampling distribution - standard error of the mean - Hypothesis testing - z-test - z-test procedure - decision rule – calculations - decisions - interpretations - one-tailed and two-tailed tests – Estimation - point estimate - confidence interval - level of confidence - effect of sample size.

#### PART – A

Q.No	Question	Level	Competence
1.	Define population? Give an example.	BTL 1	Remember
2	What is real population?	BTL 2	Understand
3	List the different types of population.	BTL 1	Remember
4	What is hypothetical population?	BTL 1	Remember
5	Define Samples.	BTL 1	Remember
6	List the categories of sample.	BTL 2	Understand
7	What is random sampling?	BTL 1	Remember
8	Mention the types of random sampling.	BTL 1	Remember
9	Differentiate population and sample.	BTL 1	Remember
10	List the types of non-probability sampling.	BTL 2	Understand
11	Define snow ball sampling.	BTL 2	Understand
12	Differentiate non-probability and probability sampling.	BTL 1	Remember
13	Give the optimal sample size.	BTL 2	Understand
14	What is systematic sampling?	BTL 1	Remember
15	Define cluster sampling.	BTL 1	Remember
16	Mention the advantages of random sampling.	BTL 2	Understand
17	Define consecutive sampling.	BTL 1	Remember
18	Provide the standard error of the mean	BTL 1	Remember
19	Give the level of confidence.	BTL 2	Understand
20	Compare two tailed and one tailed test.	BTL 1	Remember

21	Define estimation.	BTL 1	Remember
22	What are the possible decisions you can make after performing a hypothesis test?	BTL 1	Remember
23	Describe the basic steps involved in conducting a Z-test.	BTL 2	Understand
24	State the importance of random sampling in statistical analysis.	BTL 2	Understand
<b>PART – B</b>			
Q.No	Question	Level	Competence
1	Discuss on population and samples with suitable examples.	BTL 4	Analyze
2	Discuss the different types of random sampling techniques.	BTL 6	Create
3	Elaborate on the different types of non-probability based sampling techniques.	BTL 5	Evaluate
4	Illustrate the hypothesis testing with an example.	BTL 6	Create
5	Explain the procedure of z-test with an example.	BTL 5	Evaluate
6	Explain in detail about Estimation and the significance of point estimates.	BTL 5	Evaluate
7	Elaborate on Confidence interval and level of confidence.	BTL 6	Create
8	Discuss z-Test Problem.	BTL 4	Analyze
9	Illustrate Decision Rule.	BTL 5	Evaluate
10	What is data interpretation? Discuss Qualitative and Quantitative Data Interpretation.	BTL 3	Apply
11	Discuss the effect of sample size.	BTL 4	Analyze
12	Find the standard error of mean of given observations, $x=10,20,30,40,50$	BTL 3	Apply
13	Compare and Contrast one-tailed test and a two-tailed test in hypothesis testing.	BTL 4	Analyze
14	What is the sampling distribution of the sample mean? Discuss the role of the Central Limit Theorem in determining the shape of the sampling distribution as the sample size increases.	BTL 4	Analyze
15	Explain how sample size influences the width of a confidence interval. Why does increasing the sample size lead to a more precise estimate?	BTL 5	Evaluate
16	Describe the procedure for conducting a z-test. Outline the steps involved in performing a z-test, from formulating hypotheses to making a decision.	BTL 6	Create
17	Elaborate the steps to test a hypothesis.	BTL 4	Analyze

### UNIT 4 - ANALYSIS OF VARIANCE

t-test for one sample - sampling distribution of t - t-test procedure - t-test for two independent samples - p-value - statistical significance - t-test for two related samples. F - test – ANOVA – Two-factor experiments - three f-tests - two-factor ANOVA - Introduction to chi-square tests.

#### PART – A

Q.No	Question	Level	Competence
1	Define categorical variable. Give example.	BTL 1	Remember
2	Mention the types of categorical variable	BTL 2	Understand
3	Give the difference between one way and two way anova.	BTL 1	Remember
4	What is t -test?	BTL 1	Remember
5	Give the measures of the t-test	BTL 2	Understand
6	When to use the t-test?	BTL 2	Understand
7	Provide the difference between a one-sample t-test and a paired t-test.	BTL 1	Remember
8	Can the t-test is used to measure the difference among several groups.	BTL 2	Understand
9	Define chi-square test and write its formulae.	BTL 1	Remember
10	Specify the purpose of chi-square test.	BTL 2	Understand
11	How the chi-square test is interpreted.	BTL 2	Understand
12	What is an acceptable value in chi-square method	BTL 2	Understand
13	Define f-test.	BTL 1	Remember
14	Write the decision criteria for a right tailed F-test.	BTL 1	Remember
15	Give the critical value for the F-test.	BTL 1	Remember
16	Why does Anova uses F-test?	BTL 2	Understand
17	Is it possible for a negative F-statistic in a F-test.	BTL 1	Remember
18	How F-test is differentiated from T.	BTL 2	Understand
19	Differentiate one way Anova from two way Anova.	BTL 2	Understand
20	How Anova's statistical significance is determined.	BTL 1	Remember
21	What is factorial anova?	BTL 1	Remember
22	Where does the chi-square test is used?	BTL 1	Remember
23	What is meant by P-Value?	BTL 2	Understand
24	How is P-Value Calculated?	BTL 2	Understand

#### PART – B

Q.No	Question	Level	Competence
1	Elaborate T-test Problem and theory.	BTL 5	Evaluate



2	Discuss F-test Problem and theory.		BTL 4	Analyze			
3	Explain Chi-Square test Problem and theory.		BTL 5	Evaluate			
4	Discuss ANOVA Problem and theory.		BTL 4	Analyze			
5	Explain briefly about Sampling Distribution of T.		BTL 5	Evaluate			
6	a) Illustrate in detail about one factor ANOVA with example.	(8)	BTL 6	Create			
	b) A random sample of 90 college students indicates whether they most desire love, wealth power, health, fame or family happiness . Using the .05 level of significance and the following results, test the null hypothesis that in the null underlying Population, the various desires are equally popular using chi-square test.	(8)					
Desires of College Students							
Frequency	Love	Wealth			Power	Health	Fame
Observed ( $f_o$ )	25	10	5	25	10	15	90
8	A manufacturer of a gas additive claims that it improves gas mileage. A random sample of 30 drivers tests this claim by determining their gas mileage for a full tank of gas that contains the additive ( $X_1$ ) and for a full tank of gas that does not contain the additive ( $X_2$ ). The sample mean difference, $D$ , equals 2.12 miles (in favor of the additive), and the estimated standard error equals 1.50 miles. (i) Using t, test the null hypothesis at the .05 level of significance,(6) (ii) Specify the p-value for this result. (5) (iii) Are there any special precautions that should be taken with the present experimental design? (5)		BTL 4	Analyze			
9	Compare and Contrast One-Way ANOVA and Two-Way ANOVA.		BTL 4	Analyze			
10	a) A research team wants to study the effects of a new drug on insomnia. 8 tests were conducted with a variance of 600 initially. After 7 months 6 tests were conducted with a variance of 400. At a significance level of 0.05 was there any improvement in the results after 7 months? Evaluate by using f-test.	(8)	BTL 5	Evaluate			
	b) Elaborate the difference between F - Test and T – Test.	(8)					
11	a) A library system lends book for period 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, book-lending records were consulted to determine the loan periods actually used by the patrons. A random sample of eight 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.	(8)	BTL 4	Analyze			
	b) Discuss effect size estimation.	(8)					
12	Elaborate F Test in Statistics. Importance of F-Test.		BTL 5	Evaluate			
13	Explain the assumptions of the chi-square test and how violations can affect the results.		BTL 4	Analyze			

14	Explain the concept of the F-distribution and how it is used to determine statistical significance in ANOVA.	BTL 5	Evaluate
15	Derive the formula for the paired t-test statistic.	BTL 4	Analyze
16	Discuss the relationship between the p-value and Type I and Type II errors.	BTL 5	Evaluate
17	Discuss the shape and characteristics of the t-distribution. How does the t-distribution differ from the normal distribution?	BTL 4	Analyze

### UNIT 5 PREDICTIVE ANALYTICS

Linear least squares - implementation - goodness of fit - testing a linear model - weighted resampling. Regression using Stats Models - multiple regression - nonlinear relationships - logistic regression - estimating parameters - Time series analysis - moving averages - missing values - serial correlation - autocorrelation. Introduction to survival analysis.

### PART – A

Q.No	Question	Level	Competence
1	How do you calculate least squares	BTL 1	Remember
2	List the methods the available to calculate least square	BTL 2	Understand
3	Define the principle of least square.	BTL 1	Remember
4	Defne least square.	BTL 1	Remember
5	What is least square curve fitting?	BTL 1	Remember
6	Why do we need Time series Analysis?	BTL 2	Understand
7	Give some examples for time series analysis.	BTL 1	Remember
8	Mention the types of Time series Analysis	BTL 1	Remember
9	Mention the applications of Time Series Analysis	BTL 1	Remember
10	Give the limitations of Timeseries Analysis.	BTL 2	Understand
11	List the Datatypes of Time series.	BTL 2	Understand
12	What does Goodness of fit mean?	BTL 1	Remember
13	Why is Goodness of fit is important?	BTL 2	Understand
14	Provide the most common goodness of fit tests.	BTL 1	Remember
15	Why do we test goodness of fit.	BTL 2	Understand
16	Define multiple linear regression.	BTL 2	Understand
17	How the error is calculated in linear regression model.	BTL 1	Remember
18	What Is Predictive Analytics?	BTL 1	Remember
19	What are the applications of predictive models?	BTL 2	Understand
20	Define Credit.	BTL 1	Remember
21	What is meant by Forecasting?	BTL 1	Remember
22	Define Underwriting	BTL 1	Remember
23	Compare Predictive Analytics vs. Machine Learning	BTL 2	Understand
24	Define Regression.	BTL 2	Understand

## PART – B

Q.No	Question	Level	Competence																														
1	Explain Multiple regression.	BTL 5	Evaluate																														
2	Explain the concept of survival curve.	BTL 4	Analyze																														
3	Linear least square problem and theory.	BTL 4	Analyze																														
4	Explain in detail about logistic regression.	BTL 5	Evaluate																														
5	Explain Time series analysis.	BTL 5	Evaluate																														
6	Write in detail about goodness of fit.	BTL 3	Apply																														
7	a) Compare and Contrast between multiple regression and logistic regression with examples.	(8)	BTL 6 Create																														
	b) A company manufactures an electronic device to be used in a very wide temperature range. The company knows that increased temperature shortens the life time of the device, and a study is therefore performed in which the life time is determined as a function of temperature. The following data is found:	(8)																															
	<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr> <td>Temperature in Celcius (t)</td> <td>10</td> <td>20</td> <td>30</td> <td>40</td> <td>50</td> <td>60</td> <td>7</td> <td>8</td> <td>9</td> </tr> <tr> <td>Life time in hours (y)</td> <td>42</td> <td>36</td> <td>285</td> <td>22</td> <td>176</td> <td>117</td> <td>6</td> <td>3</td> <td>5</td> </tr> <tr> <td></td> <td>0</td> <td>5</td> <td></td> <td>0</td> <td></td> <td></td> <td>9</td> <td>4</td> <td></td> </tr> </table>	Temperature in Celcius (t)		10	20	30	40	50	60	7	8	9	Life time in hours (y)	42	36	285	22	176	117	6	3	5		0	5		0			9	4		
	Temperature in Celcius (t)	10		20	30	40	50	60	7	8	9																						
Life time in hours (y)	42	36	285	22	176	117	6	3	5																								
	0	5		0			9	4																									
Find the linear regression equation. Also find the estimated life time when temperature is 55.																																	
8	Illustrate in depth about time series forecasting, its components, moving averages and its various methods with examples.	BTL 5	Evaluate																														
9	Explain in brief about the various steps of Data Analysis.	BTL 3	Apply																														
10	Describe in detail Introduction to survival analysis.	BTL 4	Analyze																														
11	Explain in detail serial correlation and autocorrelation.	BTL 4	Analyze																														
12	Describe Regression using Stats Models.	BTL 4	Analyze																														
13	Illustrate nonlinear relationships.	(8)	BTL 5 Evaluate																														
	How to estimate the coefficients using maximum likelihood estimation (MLE) and interpret the estimated parameters.	(8)																															
14	How would you handle missing values and account for them in time series analysis?	BTL 3	Apply																														
15	Define heteroscedasticity and explain why it is a problem in linear regression.	BTL 3	Apply																														
16	How to assess the goodness of fit of the model using R-squared and Adjusted R-squared.	BTL 3	Apply																														
17	Justify the results of the fitted model, including the significance of the coefficients.	BTL 5	Evaluate																														