ourdiu.com | Shared by Abdullah Tawsif (CSE-232)

Daffodil International University
Faculty of Science & Information Technology
Department of Computer Science and Engineering
Final Examination, Spring 2024
Course Code: CSE315, Course Title: Introduction to Data Science
Level 3 Term 1 Batch: 61 Marks: 40

Time: 2:00 Hrs

Answer ALL Questions

*[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]*

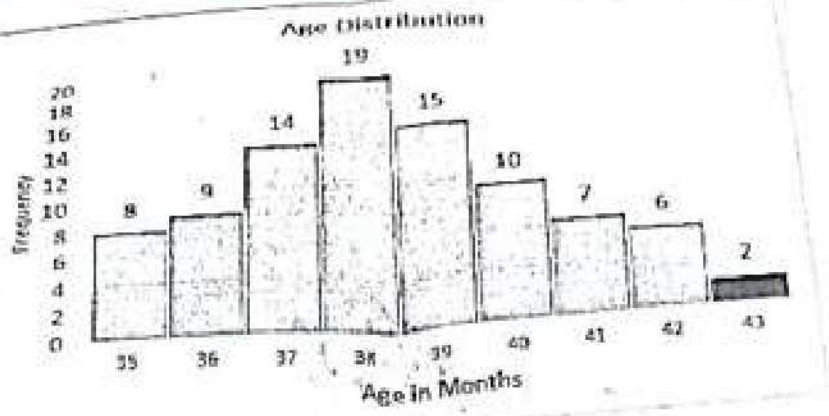
| | | | | |
|---|---|---|---|---|
| 1. | a) | On a multiple-choice test, where each question has 4 possible options but only 1 correct answer, let's explore how likely it is for a student who guesses the answers to get a specific number of questions correct out of a total of 10. First, describe the essential properties of the binomial distribution that govern such scenarios. Then, determine the probability that a student, purely guessing, will answer exactly 3 questions correctly out of the 10. Understanding this helps us measure the chances of guessing accurately on a series of questions, a common situation in various assessments. | [5] | CO2 |
| | b) | In a certain population, adult male heights follow a normal distribution, with an average of 70 inches and a standard deviation of 3 inches. What is the probability of randomly selecting an adult male whose height falls between 65 and 75 inches? Also, briefly explain the difference between discrete and continuous distributions. Understanding this distinction is key to grasping probability concepts in real-world contexts. | [5] | |
| 2. | | Suppose you conducted a study to investigate the relationship between the number of hours spent studying (x) and the scores obtained in a standardized test (y). After collecting data from 8 students, you have to perform linear regression analysis. | [10] | CO2 |

| Hours of Study (x) | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| Test Scores (y) | 76 | 80 | 85 | 88 | 92 | 94 | 97 | 99 |

a) Find the regression model for the given data. Plot the scatter plot of the data points, where the X-axis represents hours of study and the Y-axis represents test scores.
b) Plot the fitted regression line on the same graph.
c) Calculate the predicted test score for a student who studies for 7 hours.
d) Interpret the intercept and slope coefficients in the context of this study.

| | | | | |
|---|---|---|---|---|
| 3. | | Below is a histogram representing the distribution of ages of participants in a survey conducted on a social media platform. Analyze the graph and answer the following questions: | [10] | |

ourdiu.com | Shared by Abdullah Tawsif (CSE-232)




CO3

a) Describe the shape of the histogram.
b) Identify the central tendency measure that best represents the data and explain your choice.
c) Discuss any outliers or unusual features observed in the histogram and their potential implications.
d) Identify the cumulative frequency for age 38.

4. A dataset that has the following features: age, gender, BMI, children, smoker, region, and charges. Write the Python code for the following questions. [10]

| age | gender | BMI | children | smoker | region | charges |
|---|---|---|---|---|---|---|
| 19 | female | 27.9 | 0 | yes | NaN | 16884.92 |
| 18 | male | 33.77 | 1 | no | southeast | 1725.552 |
| 28 | male | 33 | 3 | no | NaN | 4449.462 |
| 33 | male | NaN | 0 | no | northwest | 21984.47 |
| 37 | female | 27.74 | 3 | no | NaN | 7281.506 |
| 31 | female | 25.74 | 0 | no | southeast | 3756.622 |
| 46 | female | NaN | 1 | no | NaN | 8240.59 |
| 37 | female | 27.74 | 3 | no | NaN | 7281.506 |
| 37 | male | 29.83 | 2 | no | northeast | 6406.411 |

CO3

a) Define the types of data in each column in the dataset. Find the average age of individuals.
b) There is some missing data present in the dataset. Explain your course of action to handle the missing information for BMI and region attributes.
c) Encode the gender attribute using customized encoding techniques such as female=1 and male=2 and encode the smoker attribute using dummy attributes.
d) What is the distribution of charges among smokers and non-smokers?
e) How to handle duplicate values in this dataset and change the data permanently in the dataset.