EDA MID EXAM

TOTAL MARKS: 30 Duration : 2 HOURS

Data Description

`Data File name: StudentsPerformance.csv`

This dataset has the scores of students evaluated on the matters like math, reading and writing.

Also other details about gender, ethnicity, parental level of education, lunch scheme opted and whether or not the student has undergone test preparation course is given. You are required to perform exploratory data analysis on the above file and answer the questions below with appropriate codes.

Section A : 5 Marks

1). List out the set of categorical and numerical variables in the dataset . (1 Mark)

2). Find out the missing values(in %) for all variables in the data set? (1 Mark)

3). Find the Percentage of students who have completed the test preparation course. (1 Mark)

4). Find the levels of parental education. What is the level of education held by majority of parents. (1 Mark)

5). Add a new column to the dataset which is a total score of each student obtained by combining math, reading and writing score. (1 Mark)

Section B: 10 Marks

6.a). Create dummy columns for the categorical variable 'lunch'. (1 mark)

b) Retain the column lunch_standard and remove the other one. (2 marks)

c) Analyze and compare the percentage of female's vs percentage of males who receive standard lunch. (2 Marks)

7.a). Plot a suitable graph for comparing parental level of education vs. the total score of the students? Write your inferences (3 marks)

b) Identify the category which has high scoring students? (2 Marks)

Section C 15 Marks

8.a). Consider the variable 'math score' and check for existence of outliers. If present, then find the list of outliers. How do you treat these outliers (5 Marks)?

b). Find the correlation between reading score and writing score. Visualize the same using a suitable map. And draw your inferences (5 Marks)

c). Which encoding techniques do you recommend to the categorical variables? And apply those encoding techniques to these categorical variables (2 marks)

d). Split the dataset into features(X: independent variables) and target variable (Y: total score). (2 marks)

e) Split the dataset into train and test in 70-30 percent format. (1 Mark)