FERRERO

Business Case

Welcome to the business case study. Below are a set of questions that you would need to attempt and present the results in powerpoint on the day of the interview. There are two datasets that are available to you for this business case. Please be ready to share the underlying codes and outputs and explain the reasoning of your answers.

Questions:

1. Read 'california_housing_train.csv' and print the first 15 rows
2. Check the following characteristics of the dataset:
   - Number of rows and columns
   - Data type
   - Print the names of columns
3. Does the dataset contain any null values
4. Create a dataframe with the total number of rooms (aggregated) and total number of bedrooms (aggregated) if the housing median age is less than 25
5. Display a scatterplot that compares median income with median house value
6. Split california_housing_train into training and validation set
   Set seed: 42 and validation size 10%
   What does the "seed" part do here?
7. Build a Linear Regression model on the training set to predict: 'median_house_value'.
   a) Evaluate the performance both on training and validation set. Chart actual vs predicted.
   b) Any enhancements you would suggest in this exercise?
8. List some differences between supervised and unsupervised learning and give an example of each type
9. You are provided with two tables : "Order table" and "Customer table"
   Order table has fields : OrderID, CustID, OrderNumber, Amount
   Customer table has fields : ID, FirstName,LastName,City
   Write a SQL query that lists down all the orders with customer information
10. How would you deal with unbalanced binary classification?