Statistics Assignment: Exploratory Data Analysis (EDA)

Perform Exploratory Data Analysis (EDA) on the data-set given below.

Download the dataset from this link. CLICK HERE TO DOWNLOAD DATASET

About Dataset:

Abstract:

The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y).

Data Set Information:

The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Attribute Information:

Bank client data:

- Age (numeric)
- Job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- Marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
- Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- Default: has credit in default? (categorical: 'no', 'yes', 'unknown')
- Housing: has a housing loan? (categorical: 'no', 'yes', 'unknown')
- Loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Related with the last contact of the current campaign:

- Contact: contact communication type (categorical: 'cellular','telephone')
- Month: last contact month of year (categorical: 'jan', 'feb', 'mar', …, 'nov', 'dec')
- Day*of*week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- Duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

- Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- Previous: number of contacts performed before this campaign and for this client (numeric)
- Poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Challenges:

- Data Cleaning : Some of the columns are merged together e.g education and occupation has been merged together your task would also include to find such columns and find a way to clean or separate them.
- Data Manipulation : Try to play around with the variables to find out whether there is a scope for selecting more important variables or deriving some columns from the existing ones.

If you are facing any difficulty in performing EDA, follow the steps mentioned below:

Step - 1 - Introduction -> Give a detailed data description and objective
Step - 2 - Import the data and display the head, shape and description of the data.
Step - 3 - Univariate Analysis -> PDF, Histograms, Boxplots, Count plots, etc..

- Find the outliers in each numerical column
- Understand the probability and frequency distribution of each numerical column
- Understand the frequency distribution of each categorical Variable/Column
- Mention observations after each plot.

Step - 4 - Bivariate Analysis

- Discover the relationships between numerical columns using Scatter plots, hexbin plots, pair plots, etc..
- Identify the patterns between categorical and numerical columns using swarmplot, boxplot, barplot, etc..
- Mention observations after each plot.

Step - 5 - Conclusion

NOTE: Mention observations after each plot.

Statistics Manipulations:

1. Find the **correlation** between the columns and draw the observations from it.
2. What is the mean age and duration time of the customers with respect to every column?
3. Find the **mean** and **median** of every column **response** wise and draw the observations.
4. Find the probabilities with respect to the job role and education with customer responses.
5. Find the Best features using **correlation** and **Chi-square** test.
6. Find the relation of **salary** and **age** column using **statistical** tests and draw the observations from it.
7. Using **statistical analysis, find** whether the **age** column is impacting the **duration** column or not.
8. Show that the columns are following the **Normal Distribution** or not, if not following try to convert it non-normal to normal. (use transformation techniques)
9. Let's check if we have any statistical patterns in the Data frame (using plots or analysis).