**SEGMENT WISE ANALYSIS OF BANK TELEMARKETING DATASET**

**The whole data set problem statement solution is divided into 5 segment step by step step I described all the segment and the file of solution is also added**

## Session- 2, Data Cleaning

### Segment- 2, Data Types

There are multiple types of data types available in the data set. some of them are numerical type and some of categorical type. You are required to get the idea about the data types after reading the data frame.

Following are the some of the types of variables:

* **Numeric data type**: banking dataset: salary, balance, duration and age.
* **Categorical data type**: banking dataset: education, job, marital, poutcome and month etc.
* **Ordinal data type**: banking dataset: Age group.
* **Time and date type**
* **Coordinates type of data**: latitude and longitude type.

### Segment- 3, Fixing the Rows and Columns

Checklist for fixing rows:

* **Delete summary rows**: Total and Subtotal rows
* **Delete incorrect rows**: Header row and footer row
* **Delete extra rows**: Column number, indicators, Blank rows, Page No.

Checklist for fixing columns:

* **Merge columns for creating unique identifiers**, if needed, for example, merge the columns State and City into the column Full address.
* **Split columns to get more data**: Split the Address column to get State and City columns to analyse each separately.
* **Add column names**: Add column names if missing.
* **Rename columns consistently**: Abbreviations, encoded columns.
* **Delete columns**: Delete unnecessary columns.
* **Align misaligned columns**: The data set may have shifted columns, which you need to align correctly.

### Segment- 4, Impute/Remove missing values

Take aways from the lecture on missing values:

* **Set values as missing values**: Identify values that indicate missing data, for example, treat blank strings, "NA", "XX", "999", etc., as missing.
* **Adding is good, exaggerating is bad**: You should try to get information from reliable external sources as much as possible, but if you can’t, then it is better to retain missing values rather than exaggerating the existing rows/columns.
* **Delete rows and columns**: Rows can be deleted if the number of missing values is insignificant, as this would not impact the overall analysis results. Columns can be removed if the missing values are quite significant in number.
* **Fill partial missing values using business judgement**: Such values include missing time zone, century, etc. These values can be identified easily.

Types of missing values:

* **MCAR**: It stands for Missing completely at random (the reason behind the missing value is not dependent on any other feature).
* **MAR**: It stands for Missing at random (the reason behind the missing value may be associated with some other features).
* **MNAR**: It stands for Missing not at random (there is a specific reason behind the missing value).

### Segment- 5, Handling Outliers

Major approaches to the treat outliers:

* **Imputation**
* **Deletion of outliers**
* **Binning of values**
* **Cap the outlier**

### Segment- 3, Categorical ordered univariate analysis

Ordered variables have some kind of ordering. Some examples of bank marketing dataset are:

* Age group= <30, 30-40, 40-50 and so on.
* Month = Jan-Feb-Mar etc.
* Education = primary, secondary and so on.

## Session- 4, Bivariate and Multivariate Analysis

### Segment-2, Numeric- numeric analysis

There are three ways to analyse the numeric- numeric data types simultaneously.

* **Scatter plot**: describes the pattern that how one variable is varying with other variable.
* **Correlation matrix**: to describe the linearity of two numeric variables.
* **Pair plot**: group of scatter plots of all numeric variables in the data frame.

### Segment- 5, Categorical categorical variable