DATA EXPLORATION

- Here, we gain insights of data.
    - 7 steps in data exploration :
        - Reading the data
        - Variable Identification
        - Univariate Analysis
        - Bi-Varaite Analysis
        - Missing value treatment
        - Outlier detection
        - Variable Transformation

- **Reading the data** : Here we just read the data from a given dataset using pandas..
    - We can read HDF5,Local Clipboard,Excel,CSV etc
    - Can't read MP4 files using pandas...
- **Variable Identification :** Here,get to know what are
    - Independent and dependent variables
    - Continuous and categorical variables..

  Why ?
    - Techniques like supervised learning require identification of dependent variable
    - Different data processing techniques for categorical and continuous variables

  How to identify ?
    - We can predict dependent and independent from the problem statement itself.
    - Pandas store categorical variables as object data type,continuous as int,float data types.
- **Univariate Analysis :**

- ➢ First Explore one variable at a time and summarize the variable
- ➢ Discover insights from that variable detect Anomaly` in data(using box plot we can detect)

How do we perform ?

- ➢ On Continuous variables :
  - ■ Tabular method : for analysing mean,median,SD..
  - ■ Graphical method : For distribution of variables and presence of outliers.
- ➢ On Categorical variables :
  - ■ Count : absolute frequency of each category
  - ■ Count% : Proportions of different categories.
  - ■ We can visualize this in plots

- ❖ **Bivariate Analysis :**
  - ➢ It is for when you want to see 2 variables associated with each other or not
  - ➢ It is also used to find relation b/w target & Independent variables as well as relation b/w 2 independent variables
  - ➢ If 2 var's are associated one is used to to infer other
  - ➢ It also helps in prediction ,detecting anomalies..

How do we perform ?

- ➢ On Continuous-Continuous :
  - ■ Analyze the relation graphically by "scatterplot"
  - ■ Perform analysis test(i.e correlation)
- ➢ On Categorical-Continuous :
  - ■ Here we use barplot to visualize
  - ■ Analysis test : 2 sample t test
- ➢ On Categorical-Categorical :
  - ■ We visualize through 2-way table
  - ■ Analysis test : Chi-Square

- **Missing Value Treatment** :Can occur by
    - Non-Response(Eg: Salary,they won’t respond)
    - Error in data collection
    - Error in reading data
    - 3 types of missing :
        - MCAR(Missing Completely At Random)
        - MAR(Missing At Random)
        - MNAR(Missing Not At Random)

    How to identify ?

    - describe() - for continuous variables
    - isnull() - for all variables

    How do we deal ?

    - continuous variable
        - can be imputed with mean,median,Regression model
    - Categorical variable
        - we can impute with mode,classification model
    - Numeric Data
        - We can impute with mean,median,mode..
- **Outlier detection :** Reasons for outliers
    - Data Entry Errors
    - Measurement errors
    - Preprocessing error
    - Types of outliers:
        - Univariate(We can analyze only variable to get outliers)
        - Bivariate(we analyze 2 variables to detect outliers)

    How to identify ?

    - Univariate : Boxplot
    - Bivariate : Scatter Plot
    - Formula method :
        - <Q1-1.5*IQR (or) >Q3+1.5*IQR are treated as outliers
        - Q1= First quartile.IQR= Q3-Q1

- **How do we treat it ?**
  - Deleting observations
  - Transforming and binning values
  - Imputing outliers like missing values
  - Treat the as separately

- **Transforming the variables :**
  - It is a process of which we replace a var with some function of that avr (i.e replacing x with its algorithm)
  - Transforming non-linear to linear relationship
  - Create symmetric distribution from skewed distribution
  - **Methods :**
    - Log transformations : reduces right skewness of variables.
    - Square root : used for reducing skewness with positive values only.
    - Cube root : can be used to reduce with any values
    - Binning : Used for converting continuous to categorical variables..

Prepared By
Sudheer Kumar Puppala