Data Analysis

Data Analysis Definition ?

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

Why is data analysis is Required ?

Every organization makes attempts to gather data, for instance, by monitoring its competitors’ performance, sales figures, and buying trends etc. in an effort to be more competitive. However, nobody can understand customers’ behaviours and competitors’ performance without the skills to analyse all that data. Data analysis, therefore, is a necessity for making well-informed and efficient decisions. Data analysis is what helps organizations determine their positions in the market relative to competitors. It is what helps us identify the potential risks that need to be avoided and the opportunities that must be grabbed in order to grow. It is, in fact, data analysis that enables us to gauge the satisfaction level of the customers and their needs in order to come up with new products and services that provide greater satisfaction to them. Therefore, it is an understatement to say that data analysis is important for the success of businesses.

Data Analysis Process?

Data Analysis is a process of collecting, transforming, cleaning, and modeling data with the goal of discovering the required information. The results so obtained are communicated, suggesting conclusions, and supporting decision-making. Data visualization is at times used to portray the data for the ease of discovering the useful patterns in the data. The terms Data Modelling and Data Analysis mean the same.

Data Analysis Process consists of the following phases that are iterative in nature −

1. **Data Requirements Specification**

First of all we have to need data selection on which we have to work to find decision and prediction on data of particular subject .The data required for analysis is based on a question or an experiment. Based on the requirements of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data may be numerical or categorical.

1. **Data Collection**

Data Collection is the process of gathering information on targeted variables identified as data requirements. The emphasis is on ensuring accurate and honest collection of data. Data Collection ensures that data gathered is accurate such that the related decisions are valid. Data Collection provides both a baseline to measure and a target to improve.

Data is collected from various sources ranging from organizational databases to the information in web pages. The data thus obtained, may not be structured and may contain irrelevant information. Hence, the collected data is required to be subjected to Data Processing and Data Cleaning.

1. **Data Processing**

Data Collection is the process of gathering information on targeted variables identified as data requirements. The emphasis is on ensuring accurate and honest collection of data. Data Collection ensures that data gathered is accurate such that the related decisions are valid. Data Collection provides both a baseline to measure and a target to improve.

The data that is collected must be processed or organized for analysis. This includes structuring the data as required for the relevant Analysis Tools. For example, the data might have to be placed into rows and columns in a table within a Spreadsheet or Statistical Application. A Data Model might have to be created.

For Processing we are using Jupiter note book and use below code and run .

# import below Library in Jupiter note book

import pandas as pd

import numpy as np

1. df=pd.read\_csv("dataset .csv)

df

1. df.head()
2. df.columns
3. df.shape
4. df.info()
5. **Data Cleaning**

The processed and organized data may be incomplete, contain duplicates, or contain errors. Data Cleaning is the process of preventing and correcting these errors. There are several types of Data Cleaning that depend on the type of data. For example, while cleaning the financial data, certain totals might be compared against reliable published numbers or defined thresholds. Likewise, quantitative data methods can be used for outlier detection that would be subsequently excluded in analysis.

# For checking Null value …

1. df.isnull().values.any()

# For checking Null value in columns

1. df.isnull().sum()

# Removing Null value with Replace and fillna from below code

1. df['XYZ']=df['XYZ'].fillna(df['XYZ'].mean())

df['XYZ']=df['XYZ'].fillna(df['XYZ'].median())

df['XYZ']=df['XYZ'].fillna(df['XYZ'].mode())

df['XYZ']=df['XYZ'].replace(0,np.NaN)

# Import Below Library for doing numerical variable for better visualization and correlation in Statistics

from sklearn.preprocessing import LabelEncoder

1. le=LabelEncoder()

df['XYZ']=le.fit\_transform(df['XYZ’])

df

1. **Data Analysis**

Data that is processed, organized and cleaned would be ready for the analysis. Various data analysis techniques are available to understand, interpret, and derive conclusions based on the requirements. Data Visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

Statistical Data Models such as Correlation, Regression Analysis can be used to identify the relations among the data variables. These models that are descriptive of the data are helpful in simplifying analysis and communicate results.

The process might require additional Data Cleaning or additional Data Collection, and hence these activities are iterative in nature.

# **Summary Statistics**

1. df1.describe()

( for this summary we are finding count , mean , std., min , 25% ,50%, 75% and Max of all Variables ) and for this summary we can understand No any null value are remains for remove , its show any outliers are present in dataset or not . what is min value , max value, standard value and Mean value of this dataset and any particular columns .

# # To check the Correlation

1. df1.corr()

(for this summary we can understand the correlation of any columns with target value , means how much any columns are giving impact of target variable )

# # Data Visulization

importing below Library for statically data drawn.

import matplotlib.pyplot as plt

import seaborn as sns

1. Univariate analysis
2. Distplot
3. Hist Plot
4. Bivariate analysis
5. Numerical vs Numerical
6. Scatterplot
7. Line plot
8. Heatmap
9. Categorical vs Numerical
10. Bar Chart
11. Violin plot
12. Categorical box plot
13. Categorical vs categorical
14. Bar Chart
15. Grouped bar chart
16. Point plot

**# Check outliers and Remove Outliers**

1. **df['XYZ'].plot.box()**

**( Box Plot is the best visualization for checking outliers of any columns of dataset )**

**# Removing outliers**

**from scipy.stats import zscore**

z=np.abs(zscore(df))

threshold=3

print(np.where(z>3))

**z**

df\_new=df[(z<3).all(axis=1)]

df\_new

# **Check Skewness and Remove Skewness**

1. **df.skew()**
2. **for col in df.columns:**

**if df[col].skew()>0.55:**

**df[col]=np.log1p(df[col])**

**Now Your Dataset are ready for model training and testing .**