Data Analysis

In Data analysis we inspect, transform and model data by locating useful information, informing conclusions, and supporting decision-making. It has multiple ways & diverse techniques with a variety of names and is used in different industries. Data analysis plays an important role in making decisions scientifically.

As a Data analyst whether a beginner or an expert, the basic tool is the same as that is Python programming language.

While Python is meant to be easy to pick up, many of the resources for learning it are vastly overcomplicated and aren’t written using very accessible language. Web frameworks can combine with data manipulation code and web API can construct software which can create with one other language. For example, Ruby is a language for constructing web applications but its EDA libraries are very limited in comparison with Python libraries.

Please find the below-mentioned phases of Data Analysis:-

**Data requirement**

The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis. Data can be collected in various ways like a person or population of people, age and income. And it may be numerical or categorical

**Data collection**

Data is collected from multiple sources. It may also be accumulated from sensors in the environment, traffic cameras, satellites, CCTV, etc. It may also be obtained via interviews, online sources, or reading documentation.

**Data processing**

When initially we received data, firstly we processed or organized the data as per the requirement for our analysis. It may involve organising data into rows and columns known as [structured data](https://en.wikipedia.org/wiki/Data_model).

**Data cleaning**

Once the data is processed and organized successfully, we have to check the data, our data might be incomplete, repeating the same values, contain the missing values or data or any other type of errors. Data cleaning is the process of correcting or fixing errors.

**Exploratory data analysis(EDA)**

Once our data is cleaned, then we have to analyse the data. Analysts can apply multiple techniques, under EDA(Exploratory data analysis).  The Steps under Exploratory Data Analysis are below mentioned:

1. **Description of data:-** Once our data is encoded we can start with describe function i.e. data.describe() in python. Describe will help us in generating descriptive statistics that will summarize the central tendency, dispersion, and shape of a dataset’s distribution, excluding NaN or null values.

Describe output will include count, mean, standard deviation, minimum, maximum and lower and upper percentiles. The value of our 50 percentile is considered the median.

1. **Handling missing data:-** We can find the missing values during data extraction due to multiple reasons. It is mandatory to handle missing values because they reduce the score of any of our performance matrices. It can also lead us to unsatisfactory predictions or classifications. There are numerous options to deal with missing data. However, it depends on the data’s nature.

Below are some of the techniques:

Drop NULL values, fill Missing Values with mean median or mode, handle outliers etc.

**Handling Outliers:-** An outlier is separate or unique from the data or crowd. Outliers can be a result of our mistake at the time of data collection or they may be a sign of variance in our data. Below are some of the methods for detecting and handling outliers:-

Boxplot

[Scatterplot](https://www.edureka.co/blog/python-matplotlib-tutorial/#Scatter)

Z-score

IQR(Inter-Quartile Range)

**Understanding relationships and new insights through plots :**

We can see various relations in our data by visualizing our dataset. Please go through the below-mentioned techniques in order to see the insights:-

Histogram- A histogram is a tool for quickly assessing a probability distribution that is easy to interpret by almost any audience. [Python](https://www.edureka.co/blog/python-programming-language) offers a handful of different options for building and plotting [histograms](https://www.edureka.co/blog/python-matplotlib-tutorial/#Histogram).

Heatmaps- The Heat Map procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient colour scale is used to represent the values of the quantitative variable. The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

**Select, Build, and Test Models**

The next step after EDA (Exploratory Data Analysis) is model selection, building, and testing.

A few considerations will help select one or more appropriate statistical or machine learning models:

* What is the data type i.e. Categorical, ordered, continuous, or mixed?
* Is there any time index to consider?
* Is the response multivariate?
* Are there rules and constraints that need to be incorporated into the model?

After model selection, the next step is we have to select a model building, testing, and tuning. Here models are configured, validated, and tuned with the best parameters to get better accuracy.

For model validation, we will use a popular approach to train the model on one set of data and afterwards we will use the trained or fitted model, to evaluate its predictive ability on separate sets of data. Through the train-validate-test approach, the best performing models and configurations can be selected.

6. **Deploy Models**

Once the model is selected and tuned, the next step is model deployment. The goal of model deployment is to produce outputs that will lead to some decision or action.

In a common scenario, model predictions and other variables are inputs to an optimization problem. The solution to that problem produces raw outputs that are translated and communicated to business experts and decision-makers.

In a data science application, model deployment is often automated while still allowing analyst users to override and influence the model’s recommendations.

7. **Monitor and Validate**

The final or last step in a data analysis process is monitoring and validation. After decisions have been put into play and allowed a short time to work, it's important to go back and check to see if outcomes are as expected.

Monitoring and validating results can take many forms. For example, summary reports and simple charts of actual versus targets or average revenue or sales over time.

The goal is to make sure the results are as expected. Otherwise, review any assumptions, and check for errors in the data feeds or any unexpected changes to data attributes. Look to see if something changed in the market in an unexpected way.

By continually monitoring and going through the above data analysis process steps, problems can be detected early on and corrected before decision-makers find themselves trying to understand non-sensical outputs, or worse, the entire project is branded a disappointing failure. With a good process in place finding and fixing issues will be routine—and with a good complement of software tools, quality and assurance can be built into the system.

## **Data Analysis Tools**

Data analysis tools make it easier to process and manipulate data sets and analysed the relationships and correlations between data sets, and it also helps to identify the patterns & trends for interpretation.

## **Types of Data Analysis: Techniques and Methods**

There are several **types of Data Analysis** techniques which are based on business and technology. However, the important Data Analysis methods are:

* Text Analysis
* Statistical Analysis
* Diagnostic Analysis
* Predictive Analysis
* Prescriptive Analysis

## **Text Analysis**

Text Analysis is also known as Data Mining. It is one of the methods of data analysis to find a pattern in large data sets using a database. It is used to transform the raw data as per the business requirement. Business Intelligence tools are present in the market and are used to make strategic business decisions. However, it offers a way to extract and examine data completely and derive patterns and finally interpretation of the data.

## **Statistical Analysis**

Statistical Analysis helps us to read the data by using past data in the form of dashboards. Statistical Analysis includes collection, analysis, interpretation, presentation, and modelling of data. It analyses a set of data or a sample of data. There are two types of Analysis – Descriptive Analysis and Inferential Analysis.

### Descriptive Analysis

### Descriptive Analysis is an analysis of complete data or a sample of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data.

### Inferential Analysis

### Inferential Analysis is a sample from complete data. In this type of Analysis, you can find so many different conclusions from the same type of data by selecting different samples.

## **Diagnostic Analysis**

Diagnostic Analysis shows “why this happened?” by finding the reasons from the insight found in Statistical Analysis. This Analysis is beneficial to finding the behaviour patterns of data. If any new problem comes in the business process, then data analysts can look into this Analysis part to find out the similar patterns of the problem. And there might be some chances to use the same prescriptions for new problems.

## **Predictive Analysis**

Predictive Analysis shows “what is going to happen” by using previous data. For example if last year I bought two dresses based on my savings and if this year my salary is increasing double then I can buy four dresses. But of course, it’s not easy like this because you have to think about other circumstances like the chances of the prices of clothes being increased this year or maybe instead of dresses you want to buy a new bike, or you need to buy a house!

So here, this Analysis does predictions about future outcomes based on current or past data. Forecasting is just an estimate. Its accuracy is based on how much detailed information you have and how much you dig into it.

## **Prescriptive Analysis**

Prescriptive Analysis combines the insight from all previous Analyses to determine which action to take in a current problem or decision. Most data-driven companies are utilizing Prescriptive Analysis because predictive and descriptive Analyses are not enough to improve data performance. Based on current situations and problems, they analysed the data and make decisions