Data Analysis- Introduction

1- Introduction

Data has been the buzzword for ages now. Either the data being generated from large-scale enterprises or the data generated from an individual, each and every aspect of data needs to be analyzed to benefit yourself from it. But how do we do it? Well, that's where the term 'Data Analytics' comes in.

2- Some definitions before starting:

- **Data analysis** is a process of inspecting, cleansing, transforming, and modeling 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.

- **Data mining** is a particular data analysis technique that focuses on statistical modelling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information. In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing hypotheses. Predictive analytics focuses on the application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All of the above are varieties of data analysis.

- **Data integration** is a precursor to data analysis, and data analysis is closely linked to data visualization and data dissemination.

3- Why is Data Analytics important?

Data Analytics has a key role in improving your business as it is used to gather hidden insights, generate reports, perform market analysis, and improve business requirements.

4- What is the role of Data Analytics?

You can refer below:

- **Gather Hidden Insights** – Hidden insights from data are gathered and then analyzed with respect to business requirements.
- **Generate Reports** – Reports are generated from the data and are passed on to the respective teams and individuals to deal with further actions for a high rise in business.
- **Perform Market Analysis** – Market Analysis can be performed to understand the strengths and weaknesses of competitors.
- **Improve Business Requirement** – Analysis of Data allows improving Business to customer requirements and experience.

Now that you know the need for Data Analytics, let me quickly elaborate on what is Data Analytics for you.

5- What is Data Analytics for Beginners?

Data Analytics refers to the techniques used to analyze data to enhance productivity and business gain. Data is extracted from various sources and is cleaned and categorized to analyze various behavioral patterns. The techniques and the tools used vary according to the organization or individual.

So, in short, if you understand your Business Administration and have the capability to perform Exploratory Data Analysis, to gather the required information, then you are good to go with a career in Data Analytics.

So, now that you know what is Data Analytics, let me quickly cover the top tools used in this field.

6- What are the tools used in Data Analytics?

With the increasing demand for Data Analytics in the market, many tools have emerged with various functionalities for this purpose. Either open-source or user-friendly, the top tools in the data analytics market are as follows.

- **R programming** – This tool is the leading analytics tool used for statistics and data modeling. R compiles and runs on various platforms such as UNIX, Windows, and Mac OS. It also provides tools to automatically install all packages as per user-requirement.

- **Python** – Python is an open-source, object-oriented programming language that is easy to read, write, and maintain. It provides various machine learning and visualization libraries such as Scikit-learn, TensorFlow, Matplotlib, Pandas, Keras, etc. It also can be assembled on any platform like SQL server, a MongoDB database or JSON
- **Tableau Public** – This is a free software that connects to any data source such as Excel, corporate Data Warehouse, etc. It then creates visualizations, maps, dashboards etc with real-time updates on the web.
- **QlikView** – This tool offers in-memory data processing with the results delivered to the end-users quickly. It also offers data association and data visualization with data being compressed to almost 10% of its original size.
- **SAS** – A programming language and environment for data manipulation and analytics, this tool is easily accessible and can analyze data from different sources.
- **Microsoft Excel** – This tool is one of the most widely used tools for data analytics. Mostly used for clients' internal data, this tool analyzes the tasks that summarize the data with a preview of pivot tables.
- **RapidMiner** – A powerful, integrated platform that can integrate with any data source types such as Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase etc. This tool is mostly used for predictive analytics, such as data mining, text analytics, machine learning.
- **KNIME** – Konstanz Information Miner (KNIME) is an open-source data analytics platform, which allows you to analyze and model data. With the benefit of visual programming, KNIME provides a platform for reporting and integration through its modular data pipeline concept.
- **OpenRefine** – Also known as GoogleRefine, this data cleaning software will help you clean up data for analysis. It is used for cleaning messy data, the transformation of data and parsing data from websites.
- **Apache Spark** – One of the largest large-scale data processing engine, this tool executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk. This tool is also popular for data pipelines and machine learning model development.

Now that you know all this about Data Analysis, let me tell you what you can become by gaining knowledge about this field.

Well, you can become a well-renowned Data Analyst. Now, if you ask me *Who is a Data Analyst?* then my answer would be that a Data Analyst is a professional who can analyze data by applying various tool and techniques and gathering the required insights.

7- How to Become a Data Analyst?

Data analysts translate numbers into plain English. A Data Analyst delivers value to their companies by **taking information** about specific topics and then **interpreting,**

analyzing, and presenting findings in comprehensive **reports**. So, if you have the capability to collect data from various sources, analyze the data, gather hidden insights, and generate reports, then you can become a Data Analyst. Refer to the image below:

Process of Data Analysis – What is Data Analytics

Apart from the above-mentioned capabilities, a Data Analyst should also possess skills such as Statistics, Data Cleaning, Exploratory Data Analysis, and Data Visualization. Also, if you have a knowledge of Machine Learning, then that would make you stand out from the crowd.

8- **The process of data analysis**

Analysis refers to dividing a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data, and subsequently converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses, or disprove theories.

Statistician John Tukey defined data analysis in 1961, as:

"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases. The CRISP framework, used in data mining, has similar steps.

- **Data requirements**

The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analytics (or customers, who will use the finished product of the analysis). The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

- **Data collection**

Data is collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data; such as, Information Technology personnel within an organization. The data may also be collected from sensors in the environment, including traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

- **Data processing**

The phases of the intelligence cycle used to convert raw information into actionable intelligence or knowledge are conceptually similar to the phases in data analysis.

Data, when initially obtained, must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (known as structured data) for further analysis, often through the use of spreadsheet or statistical software.

- **Data cleaning**

Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that the data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data, deduplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers that are believed to be reliable. Unusual amounts, above or below predetermined thresholds, may also be reviewed. There are several types of data cleaning, that are dependent upon the type of data in the set; this could be phone numbers, email addresses, employers, or other values. Quantitative data methods for outlier detection can be used to get rid of data that appears to have a higher likelihood of being input incorrectly. Textual data spell checkers can be used to lessen the amount of mis-typed words. However, it is harder to tell if the words themselves are correct.

- **Exploratory data analysis**

Once the datasets are cleaned, they can then be analyzed. Analysts may apply a variety of techniques, referred to as exploratory data analysis, to begin understanding the messages contained within the obtained data. The process of data exploration may result in additional data cleaning or additional requests for data; thus, the initialization of the iterative phases mentioned in the lead paragraph of this section. Descriptive statistics, such as, the average or median, can be generated to aid in understanding the data. Data visualization is also a technique used, in which the analyst is able to examine the data in a graphical format in order to obtain additional insights regarding the messages within the data.

- **Modelling and algorithms**

Mathematical formulas or models (known as algorithms), may be applied to the data in order to identify relationships among the variables; for example, using correlation or causation. In general terms, models may be developed to evaluate a specific variable based on other variable(s) contained within the dataset, with some residual error depending on the implemented model's accuracy (e.g., Data = Model + Error).

Inferential statistics, includes utilizing techniques that measure the relationships between particular variables. For example, regression analysis may be used to model whether a change in advertising (independent variable X), provides an explanation for the variation in sales (dependent variable Y). In mathematical terms, Y (sales) is a function of X (advertising).[39] It may be described as (Y = aX + b + error), where the model is designed such that (a) and (b) minimize the error when the model predicts Y for a given range of values of X. Analysts may also attempt to build models that are descriptive of the data, in an aim to simplify analysis and communicate results.

- **Data product**

A data product is a computer application that takes data inputs and generates outputs, feeding them back into the environment. It may be based on a model or algorithm. For instance, an application that analyzes data about customer purchase history, and uses the results to recommend other purchases the customer might enjoy.

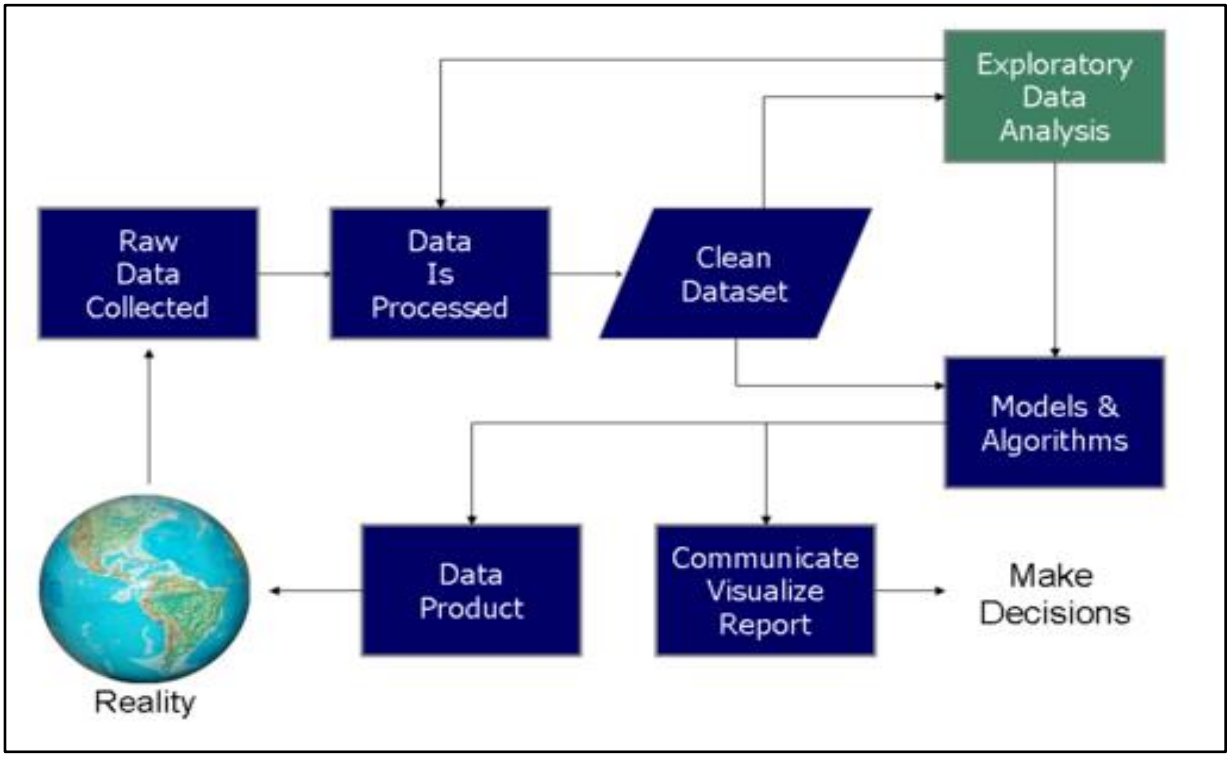
- **Communication**

Data visualization is used to help understand the results after data is analyzed.

Once data is analyzed, it may be reported in many formats to the users of the analysis to support their requirements. The users may have feedback, which results in additional analysis. As such, much of the analytical cycle is iterative.

When determining how to communicate the results, the analyst may consider implementing a variety of data visualization techniques to help communicate the message more clearly and efficiently to the audience. Data visualization uses information

displays (graphics such as, tables and charts) to help communicate key messages contained in the data. Tables are a valuable tool by enabling the ability of a user to query and focus on specific numbers; while charts (e.g., bar charts or line charts), may help explain the quantitative messages contained in the data.



9- Techniques for analyzing quantitative data

recommended a series of best practices for understanding quantitative data. These include:

- Check raw data for anomalies prior to performing an analysis;
- Re-perform important calculations, such as verifying columns of data that are formula driven;
- Confirm main totals are the sum of subtotals;
- Check relationships between numbers that should be related in a predictable way, such as ratios over time;
- Normalize numbers to make comparisons easier, such as analyzing amounts per person or relative to GDP or as an index value relative to a base year;
- Break problems into component parts by analyzing factors that led to the results, such as DuPont analysis of return on equity.

For the variables under examination, analysts typically obtain descriptive statistics for them, such as the mean (average), median, and standard deviation. They may also analyze the distribution of the key variables to see how the individual values cluster around the mean

10-Example:

https://rigorousthemes.com/blog/best-power-bi-dashboard-examples/

https://hevodata.com/learn/top-10-best-power-bi-dashboard-examples-in-2021/

https://www.spec-india.com/blog/power-bi-dashboard-examples

https://learn.microsoft.com/en-us/power-bi/create-reports/sample-datasets

https://www.imensosoftware.com/top-9-best-power-bi-dashboard-examples-for-2022