Modern Data Ecosystem and the Role of Data Analytics

**Key players in the Data Ecosystem**:

**Data engineering** converts raw data into usable data. **Data analytics** uses this data to generate insights. **Data scientists** use data analytics and data engineering to predict the future using data from the past, **business analysts** and business intelligence analysts use these insights and predictions to drive decisions that benefit and grow their business.

Data Scientist:

Data scientists analyze data for actionable insights and build machine learning or deep learning models that train on past data to create predictive models. Data scientists are people who answer questions such as, how many new social media followers am I likely to get next month, or what percentage of my customers am I likely to lose to competition in the next quarter, or is this financial transaction unusual for this customer? Data scientists require knowledge of mathematics, statistics, and a fair understanding of programming languages, databases, and building data models. They also need to have domain knowledge.

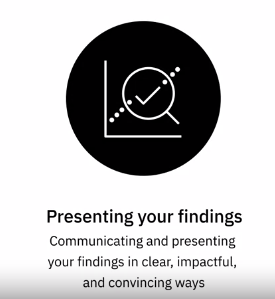
Data Analysis:

Data analysis is the process of gathering, cleaning, analyzing and mining data, interpreting results, and reporting the findings. With data analysis we find patterns within data and correlations between different data points. And it is through these patterns and correlations that insights are generated, and conclusions are drawn. Data analysis helps businesses understand their past performance and informs their decision-making for future actions. Using data analysis, businesses can validate a course of action before committing to it. Saving valuable time and resources and also ensuring greater success.

The Data Analysis Process:







Data Analytics:

Data analytics as a process or a phenomenon of taking information gathered from a relevant population, maybe our customers or our social audience, and breaking that information down into subsets, and using that data to make decisions about products or services that we want to offer, or in cases of the digital environment that we're in, making decisions about certain pieces of content that we want to publish so that it appeals to our target audience.

**Descriptive Analytics,** that helps decode “What happened.”

**Diagnostic Analytics**, that helps us understand “Why it happened.”

**Predictive Analytics**, that analyzes historical data and trends to suggest “What will happen next.”

**Prescriptive Analytics**, that prescribes “What should be done next.”

Responsibilities of a Data Analyst:

Acquiring data from primary and secondary data sources, Creating queries to extract required data from databases and other data collection systems, Filtering, cleaning, standardizing, and reorganizing data in preparation for data analysis, Using statistical tools to interpret data sets, Using statistical techniques to identify patterns and correlations in data, Analyzing patterns in complex data sets and interpreting trends, Preparing reports and charts that effectively communicate trends and patterns, Creating appropriate documentation to define and demonstrate the steps of the data analysis process.

Skills that a Data Analyst need:

**For technical skills**:

Expertise in using spreadsheets such as **Microsoft Excel or Google Sheets**

Proficiency in statistical analysis and visualization tools and software

**IBM Cognos, IBM SPSS, Oracle Visual Analyzer, Microsoft Power BI, SAS, and Tableau**

Proficiency in programming languages

**R, Python, and in some cases C++, Java, and MATLAB**

Databases

Good knowledge of **SQL**, and ability to work with **data in relational and NoSQL** databases

The ability to access and extract data from data repositories such as **data marts, data warehouses, data lakes, and data pipelines**

Familiarity with Big Data processing tools such as

**Hadoop, Hive, and Spark.**

**Functional Skills:**

Proficiency in Statistics

Analyze data, validate the analysis, identify fallacies and logical errors

Analytical skills

Research and interpret data, theorize, make forecasts

Problem-solving skills

Come up with possible solutions for a given problem

Probing skills

Identify and define the problem statement and desired outcome

Data Visualization skills

Create clear and compelling visualizations to present the analysis

Project management skills

Manage the process, people, dependencies and timelines

**Soft Skills:**

**Ability to**:

* Work collaboratively with business and cross-functional teams
* Communicate effectively to report and present findings
* Tell a compelling and convincing story
* Gather support and buy-in for work

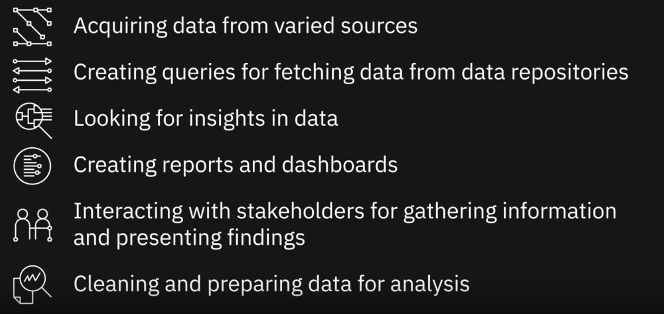
**Curiosity**

Allowing new questions to surface and challenging own assumptions and hypotheses

**Intuition**

Having a sense of the future based on pattern recognition and past experiences

A day in the life of a Data Analyst:



What are some of the applications of Data Analytics in todays’ world?

* Use of sentiment analysis and tweets and stories to inform investment decisions
* Use of satellite imagery data to track the development of industrial activities
* Use of geolocation data to track store traffic and predict sales volume

**The Data Ecosystem**:

We will learn about the different types of data structures, file formats, sources of data, and the languages data professionals use in their day-to-day tasks. We will gain an understanding of various types of data repositories such as Databases, Data Warehouses, Data Marts, Data Lakes, and Data Pipelines. In addition, we will learn about the Extract, Transform, and Load (ETL) Process, which is used to extract, transform, and load data into data repositories. We will gain a basic understanding of Big Data and Big Data processing tools such as Hadoop, Hadoop Distributed File System (HDFS), Hive, and Spark.

Learning Objectives

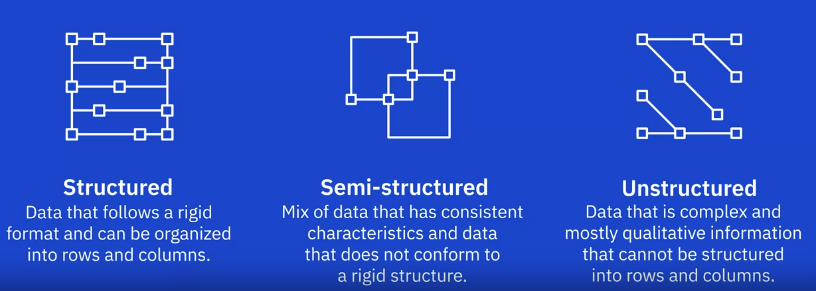
* Describe and differentiate between relational and non-relational database management systems.
* Classify data structures, file formats, and sources of data by their different types.
* Explain the features and use of the different languages used by data professionals.
* Describe how Data Warehouses, Data Marts, Data Lakes, and Data Pipelines work.
* Explain how the Extract, Transform, and Load process works to make raw data ready for analysis.
* Explain what Big Data is.
* Summarize the features and use of some of the Big Data processing tools.

Overview of the Data Analyst Ecosystem

A data analyst’s ecosystem includes the infrastructure, software, tools, frameworks, and processes used to

* Gather data
* Clean data
* Mine data
* Visualize data

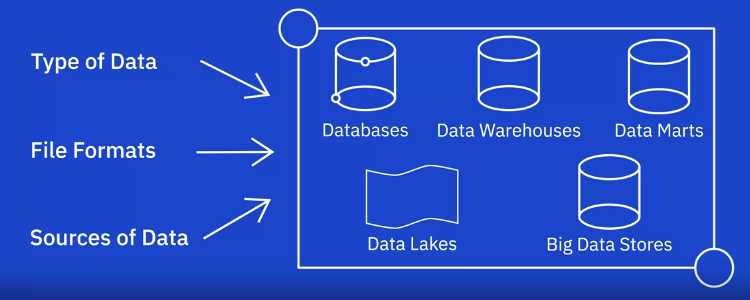
**3 types of Data:**



Data can come in a variety of file formats, such as

* Relational database
* Non-relational database
* APIs
* Web services
* Data streams
* Social platforms
* Sensor devices

**Data Repositories:**



**Languages:**

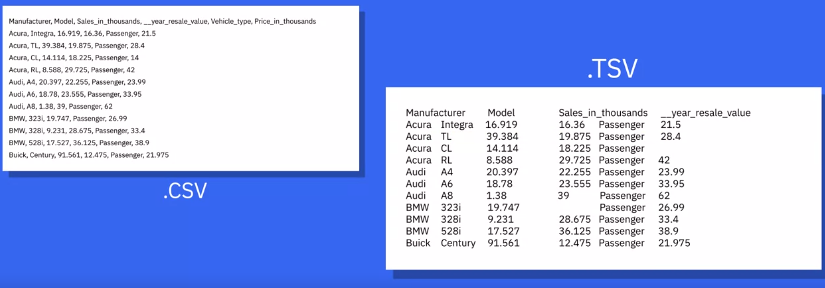
* Query languages such as SQL for querying and manipulating data
* Programming languages such as Python for developing data applications
* Shell and Scripting languages for repetitive operational tasks

Understanding Different types of file formats:

Standard file formats:

1. Delimited text file formats or .CSV
2. Microsoft Excel Open .XML Spreadsheet or .XLSX
3. Extensible Markup Language or .XML
4. Portable Document Format or .PDF
5. JavaScript Object Notation or. JSON

**Delimiter** – a sequence of one or more characters for specifying the boundary between independent entities or values. Comma, Tab, Colon, Vertical Bar, Space



**JSON** – is a text-based open standard designed for transmitting structured data over the web.

**Sources of Data:**

* Relational database (MSSQL Server, ORACLE, MySQL, IBM DB2)
* Flat files and XML Datasets (Spreadsheet, XML)
* APIs and Web services
* Web scraping (Beautiful Soup, Scrapy, Pandas, Selenium)
* Data streams and feeds (IoT devices, GPS data from cars, computer programs website and social media posts)

**APIs and Web Services**:

APIs and Web Services typically listen for incoming requests, which can be in the form of web requests from users or network requests from applications and return data in plain text, XML, HTML, JSON, or media files.

**Popular examples of APIs** such as Twitter and Facebook APIs for customer sentiment analysis

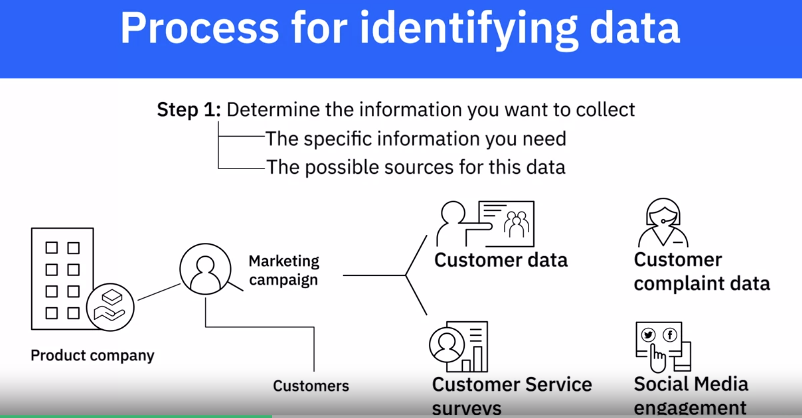
**Stock Market APIs** for trading and analysis

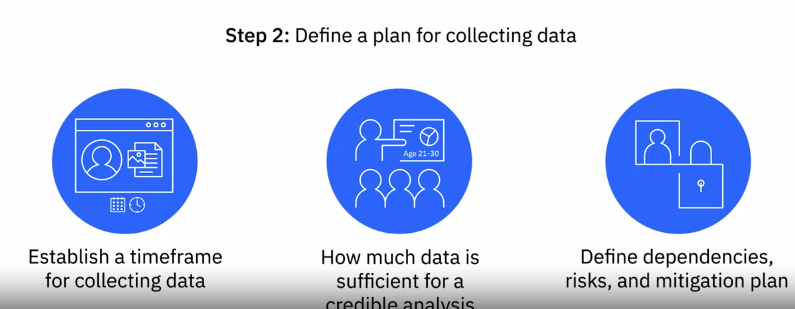
**Data Lookup and Validation APIs** for cleaning and co-relating data

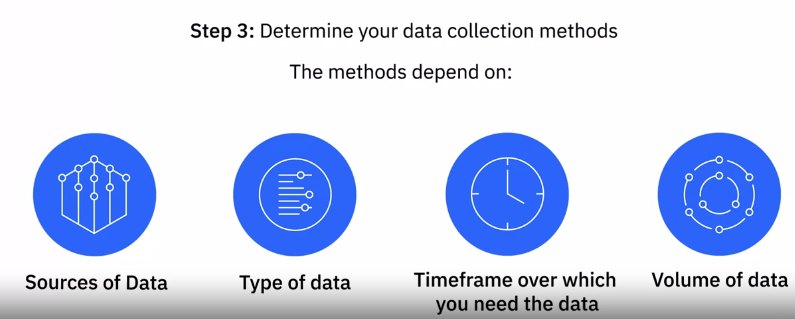
**Gathering and Wrangling Data**

We will learn about the process and steps involved in identifying, gathering, and importing data from disparate sources. We will learn about the tasks involved in wrangling and cleaning data in order to make it ready for analysis. In addition, we will gain an understanding of the different tools that can be used for gathering, importing, wrangling, and cleaning data, along with some of their characteristics, strengths, limitations, and applications.

**Identifying Data Analysis**:







**Summary:**

Having identified the data, your next step is to identify the sources from which you will extract the required data and define a plan for data collection. Decisions regarding the timeframe over which you need your data set, and how much data would suffice for arriving at a credible analysis also weigh in at this stage.

Data Sources can be internal or external to the organization, and they can be primary, secondary, or third-party, depending on whether you are obtaining the data directly from the original source, retrieving it from externally available data sources, or purchasing it from data aggregators.

Some of the data sources from which you could be gathering data include databases, the web, social media, interactive platforms, sensor devices, data exchanges, surveys and observation studies.

Data that has been identified and gathered from the various data sources is combined using a variety of tools and methods to provide a single interface using which data can be queried and manipulated.

The data you identify, the source of that data, and the practices you employ for gathering the data have implications for quality, security, and privacy, which need to be considered at this stage.

**How to Gather and Import Data**:

**Specific data repositories are optimized for certain types of data.**

**Structured data**

* Relational databases store structured data with a well-defined schema
* Sources include data from OLTP system, spreadsheets, online forms, sensors, network and web logs
* Can also be stored in NoSQL databases

**Semi-structured data**

* Sources include emails, XML, zipped files, binary executables, and TCP/IP protocols
* Can be stored in NoSQL clusters
* XML and JSON are commonly used for storing and exchanging semi-structured data

**Unstructured data:**

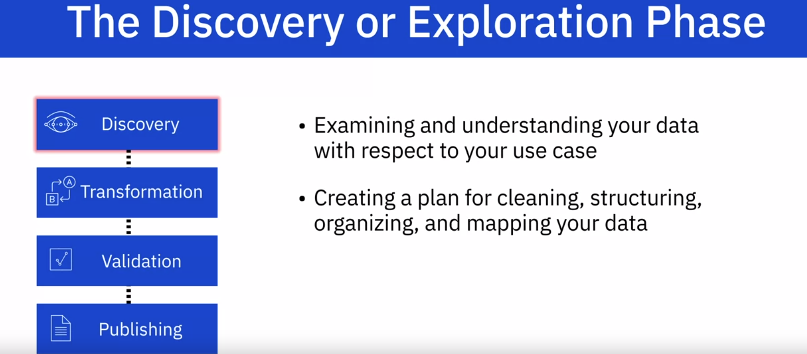
Unstructured data is data that does not have a structure and cannot be organized into a schema, such as data from web pages, social media feeds, images, videos, documents, media logs, and surveys. NoSQL databases and Data Lakes provide a good option to store and manipulate large volumes of unstructured data. Data lakes can accommodate all data types and schema. ETL tools and data pipelines provide automated functions that facilitate the process of importing data. Tools such as Talend and Informatica, and programming languages such as Python and R, and their libraries, are widely used for importing data.

**Wrangling Data**:

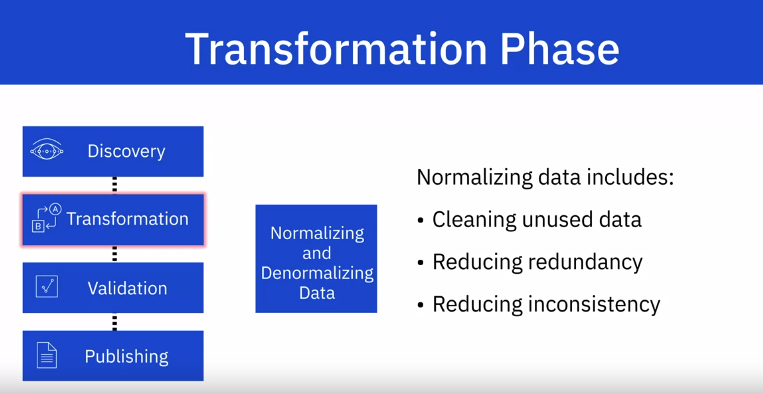
**Data wrangling**:

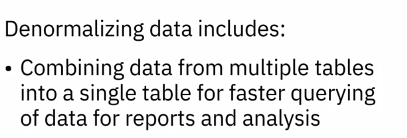
Data wrangling, also known as data munging, is an iterative process that involves data exploration, transformation, validation, and making it available for a credible and meaningful analysis. It includes a range of tasks involved in preparing raw data for a clearly defined purpose, where raw data at this stage is data that has been collated through various data sources in a data repository. Data wrangling captures a range of tasks involved in preparing data for analysis. Typically, it is a 4-step process that involves—Discovery, Transformation, Validation, and Publishing.

**Discovery or Exploration Phase**:



The next phase, which is the Transformation phase, forms the bulk of the data wrangling process. It involves the tasks you undertake to transform the data, such as structuring, normalizing, de-normalizing, cleaning, and enriching the data.





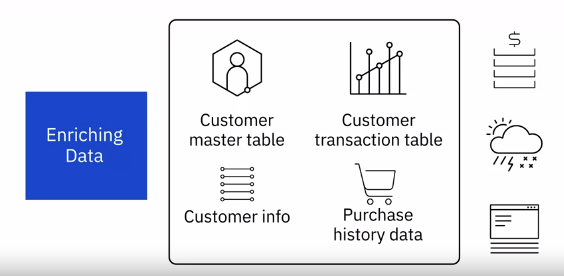
## **Cleaning Data**

* Fixing irregularities in data in order to product a credible and accurate analysis



Enriching data:

* Adding data points that make our analysis more meaningful



After transformation, the next phase in Data Wrangling is Validation.

Validation:

This is where we check the quality of the data post structuring, normalizing, cleaning, and enriching. Validation rules refer to repetitive programming steps used to verify the consistency, quality, and security of the data we have. This brings us to Publishing—the fourth phase of the data wrangling process. Publishing involves delivering the output of the wrangled data for downstream project needs. What is published is the transformed and validated version of the input dataset along with the metadata about the dataset. Lastly, it is important to note the criticality of documenting the steps and considerations we have taken to convert the raw data to analysis-ready data. All phases of data wrangling are iterative in nature. In order to replicate the steps and to revisit your considerations for performing these steps, it is vital that we document all considerations and actions.

**Tools for Data Wrangling**:

Some of the popularly used data wrangling software and tools, include:

* Excel Power Query / Spreadsheets
* Open Refine
* google Data Prep
* Watson studio refinery
* Trifecta wrangler
* Python
* R

**Spreadsheets**:

Spreadsheets such as Microsoft Excel and Google Sheets have a host of features and in-built formulae that can help you identify issues, clean, and transform data. Add-ins are available that allow you to import data from several different types of sources and clean and transform data as needed—such as Microsoft Power Query for Excel and Google Sheets Query function for Google Sheets.

**OpenRefine:**

OpenRefine is an open-source tool that allows you to import and export data in a wide variety of formats, such as TSV, CSV, XLS, XML, and JSON. Using OpenRefine, you can clean data, transform it from one format to another, and extend data with web services and external data. OpenRefine is easy to learn and easy to use. It offers menu-based operations, which means you don’t need to memorize commands or syntax.

**Python:**

Python has a huge library and set of packages that offer powerful data manipulation capabilities. Let’s look at a few of these libraries and packages.

**Jupyter Notebook** is an open-source web application widely used for data cleaning and transformation, statistical modeling, also data visualization.

**Numpy**, or Numerical Python, is the most basic package that Python offers. It is fast, versatile, interoperable, and easy to use. It provides support for large, multi-dimensional arrays and matrices, and high-level mathematical functions to operate on these arrays. **Pandas** is designed for fast and easy data analysis operations. It allows complex operations such as merging, joining, and transforming huge chunks of data, performed using simple, single-line commands. Using Pandas, we can prevent common errors that result from misaligned data coming in from different sources.

**R:**

R, also offers a series of libraries and packages that are explicitly created for wrangling messy data—such as Dplyr, Data.table, and Jsonlite. Using these libraries, you can investigate, manipulate, and analyze data.

**Dplyr** is a powerful library for data wrangling. It has a precise and straightforward syntax. **Data.table** helps to aggregate large data sets quickly.

**Jsonlite** is a robust JSON parsing tool, great for interacting with web APIs.

**Data Cleaning**:

Data Cleaning forms a very significant and integral part of the Transformation phase in a data wrangling workflow. A typical data cleaning workflow includes: Inspection, Cleaning, and Verification.

The first step in the data cleaning workflow is to detect the different types of issues and errors that your dataset may have. We can use scripts and tools that allow us to define specific rules and constraints and validate our data against these rules and constraints. We can also use data profiling and data visualization tools for inspection.

Data profiling helps us to inspect the source data to understand the structure, content, and interrelationships in our data. It uncovers anomalies and data quality issues. For example, blank or null values, duplicate data, or whether the value of a field falls within the expected range. Visualizing the data using statistical methods can help us to spot outliers.

For example, plotting the average income in a demographic dataset can help us spot outliers. That brings us to the actual cleaning of the data. The techniques you apply for cleaning your dataset will depend on our use case and the type of issues you encounter.

Let’s look at some of the more common data issues. Let’s start with missing values. Missing values are very important to deal with as they can cause unexpected or biased results. We can choose to filter out the records with missing values or find a way to source that information in case it is intrinsic to your use case.

For example, missing age data from a demographics study. A third option is a method known as imputation, which calculates the missing value based on statistical values. Our decision on the course of action we choose needs to be anchored in what’s best for our use case.

We may also come across duplicate data, data points that are repeated in your dataset. These need to be removed. Another type of issue we may encounter is that of irrelevant data. Data that does not fit within the context of your use case can be considered irrelevant data. For example, if you are analyzing data about the general health of a segment of the population, their contact numbers may not be relevant for us.

Cleaning can involve data type conversion as well. This is needed to ensure that values in a field are stored as the data type of that field—for example, numbers stored as numerical data type or date stored as a date data type.

We may also need to clean your data in order to standardize it. For example, for strings, you may want all values to be in lower case. Similarly, date formats and units of measurement need to be standardized.

Then there are syntax errors. For example, white spaces, or extra spaces at the beginning or end of a string is a syntax error that needs to be rectified. This can also include fixing typos or format, for example, the state name being entered as a full form such as New York versus an abbreviated form such as NY in some records.

Data can also have outliers, or values that are vastly different from other observations in the dataset. Outliers may, or may not, be incorrect. For example, when an age field in a voters database has the value 5, we know it is incorrect data and needs to be corrected. Now let’s consider a group of people where the annual income is in the range of one hundred thousand to two hundred thousand dollars—except for that one person who earns a million dollars a year. While this data point is not incorrect, it is an outlier, and needs to be looked at.

Depending on your use case, we may need to decide if including this data will skew the results in a way that does not serve your use case.

This brings us to the next step in the data cleaning workflow—Verification. In this step, we inspect the results to establish effectiveness and accuracy achieved as a result of the data cleaning operation. We need to re-inspect the data to make sure the rules and constraints applicable on the data still hold after the corrections we made.

And in the end, it is important to note that all changes undertaken as part of the data cleaning operation need to be documented. Not just the changes, but also the reasons behind making those changes, and the quality of the currently stored data. Reporting how healthy the data is, is a very crucial step.

Analyzing and Mining Data:

**Statistics:**

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data.

It’s all around us in our day to day lives. Whether we’re talking about average income, average age, or highest-paid professions—it’s all statistics. Today, statistics is being applied across industries for decision-making based on data. For example, researchers using statistics to analyze data from the production of vaccines to ensure safety and efficacy, or companies using statistics to reduce customer churn by gaining greater insight into customer requirements.

**Statistical Analysis**:

Statistical Analysis is the application of statistical methods to a sample of data in order to develop an understanding of what that data represents. It includes collecting and scrutinizing every data sample in a set of items from which samples can be drawn. A sample, in Statistics, is a representative selection drawn from a total population, where population is a discrete group of people or things that can be identified by at least one common characteristic for purposes of data collection and analysis.

For example, in a certain use case, population may be all people in a state that have a driving license, and a sample of this population that is a part, or subset, of the population could be men drivers over the age of 50. Statistical methods are mainly useful to ensure that data is interpreted correctly, and apparent relationships are meaningful and not just happening by chance.

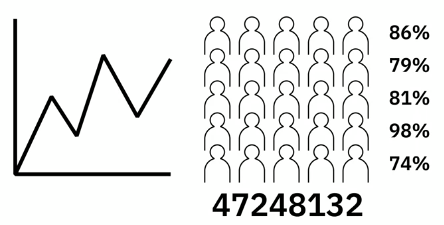
Two different types of statistics:

1. Descriptive Statistics [To summarize information about the sample]
2. Inferential Statistics [To make inferences or generalizations about broader population]

Descriptive Statistics:

Descriptive Statistics enables you to present data in a meaningful way allowing simpler interpretation of the data. Data is described using summary charts, tables, and graphs without any attempts to draw conclusions about the population from which the sample is taken.

The objective is to make it easier to understand and visualize raw data without making conclusions regarding any hypotheses that were made. For example, we want to describe the English test scores in a specific class of 25 students. We record the test scores of all students, calculate the summary statistics, and produce a graph.



Common measure of Descriptive Statistical Analysis:

* Central Tendency
* Dispersion
* Skewness

**Central Tendency**:

Central Tendency, or locating the center of a data sample. Some of the common measures of central tendency include **mean, median, and mode**. These measures tell you where most values in your dataset fall.

**Dispersion:**

Dispersion is the measure of variability in a dataset. Common measures of statistical dispersion are **Variance, Standard Deviation, and Range**.

**Variance**:

Variance defines how far away the data points fall from the center, that is, the distribution of values. When a distribution has lower variability, the values in a dataset are more consistent. However, when the variability is higher, the data points are more dissimilar, and extreme values become more likely. Understanding variability can help us grasp the likelihood of an event happening.

**Standard deviation**:

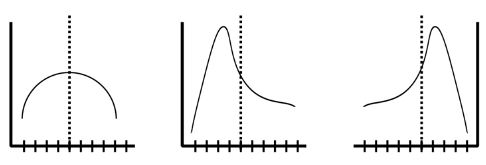
Standard deviation tells us how tightly our data is clustered around the mean.

**Range**:

Range gives you the distance between the smallest and largest values in your datasets.

**Skewness:**

Skewness is the measure of whether the distribution of values is symmetrical around a central value or skewed left or right. Skewed data can affect which types of analyses are valid to perform. These are some of the basic and most commonly used descriptive statistics tools, but there are other tools as well, for example, using **correlation** and **scatterplots** to assess the relationships of paired data.



**Inferential Statistics**:

Inferential statistics takes data from a sample to make inferences about the larger population from which the sample was drawn. Using methods of inferential statistics, we can draw generalizations that apply the results of the sample to the population as a whole. Some common methodologies of Inferential Statistics include:

* Hypothesis Testing
* Confidence Intervals
* Regression Analysis

**Hypothesis Testing**—For example, for studying the effectiveness of a vaccine by comparing outcomes in a control group, hypothesis tests can tell us whether the efficacy of a vaccine observed in a control group is likely to exist in the population as well.

**Confidence Intervals** incorporate the uncertainty and sample error to create a range of values the actual population value is like to fall within.

**Regression Analysis** incorporates hypothesis tests that help determine whether the relationships observed in the sample data actually exist in the population rather than just the sample.

Data Mining:

Data mining or the process of extracting knowledge from data, is the heart of the data analysis process. It is an interdisciplinary field that involves the use of pattern recognition technologies, statistical analysis and mathematical techniques. Its goal is to identify correlations in data, find patterns and variations. Understand trends and predict probabilities.

Applications of Data Mining:

Data mining has applications across industries and disciplines. For example, profiling customer behaviors needs and disposable income in order to offer targeted campaigns, financial institutions, tracking customer transactions for unusual behaviors, and flagging fraudulent transactions using data mining models. The use of statistical models to predict a patient’s likelihood for specific health conditions and prioritizing treatment.

Data Mining Techniques:

There are several techniques we can use to detect patterns and build accurate models for discovery, be it descriptive, diagnostic, predictive modeling.

**Classification**:

Classification is a technique that classifies attributes into target categories, for example, classifying customers into low, medium, or high spenders based on how much they earn.

**Clustering**:

Clustering is similar to classification, but involves grouping data into clusters so they can be treated as groups. For example, clustering customers based on geographic regions anomaly or outlier detection is a technique that helps find patterns and data that are not normal or unexpected. For example, spikes in the usage of a credit card that can flag possible misuse.

**Decision Tree**:

Decision trees help build classification models in the form of a tree structure with multiple branches, where each branch represents a probable occurrence. This technique helps to build a clear understanding of the relationship between input and output.

**Regression**:

Regression is a technique that helps identify the nature of the relationship between two variables, which could be causal or correlational. For example, based on factors such as location and covered area, a regression model could be used to predict the value of a house.

**Data Mining Tools**:

The commonly used software and tools for data mining, such as: Spreadsheets, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio; and SAS.

Spreadsheets:

Spreadsheets, such as Microsoft Excel and Google Sheets, are commonly used for performing basic data mining tasks. Spreadsheets can be used to host data that has been exported from other systems in an easily accessible and easy-to-read format. We can **pivot tables** to showcase specific aspects of our data, which is vital when we have huge amounts of data to sort through and analyze. They also make it relatively easier to make comparisons between different sets of data.

**Add-ins** available for Excel, such as the Data Mining Client for Excel, XLMiner, and KnowledgeMiner for Excel, allow us to perform common mining tasks, such as classification, regression, association rules, clustering, and model building. **Google Sheets** also has an array of add-ons that can be used for analysis and mining, such as Text Analysis, Text Mining, Google Analytics.

R:

R is one of the most widely used languages for performing statistical modeling and computations by statisticians and data miners. R is packaged with hundreds of libraries explicitly built for data mining operations such as regression, classification, data clustering, association rule mining, text mining, outlier detection, and social network analysis.

Some of the popular R packages include **tm** and **twitteR**. tm, a framework for text mining applications within R, provides functions for text mining. twitteR provides a framework for mining tweets. R Studio is a popularly used open-source Integrated Development Environment (or IDE) for working with the R programming language.

Python:

Python libraries like **Pandas** and **NumPy** are commonly used for Data Mining. Pandas is an open-source module for working with data structures and analysis. It is possibly one of the most popular libraries for data analysis in Python.

It allows you to upload data in any format and provides a simple platform to organize, sort, and manipulate that data. Using Pandas, we can: perform basic numerical computations such as mean, median, mode, and range; calculate statistics and answer questions regarding correlation between data and distribution of data; explore data visually and quantitatively; visualize data with help from other Python libraries.

NumPy is a tool for mathematical computing and data preparation in Python. NumPy offers a host of built-in functions and capabilities for data mining. Jupyter Notebooks have become the tool of choice for Data Scientists and Data Analysts when working with Python to perform data mining and statistical analysis.