Data Analysis

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

- Summarize the main characteristics often with visual methods
- Convert raw data into useful information for decision making.
- Decision supporting system

Data Analysis is a lifeline of any business. Whether one wants to arrive at some marketing decisions or new product launch strategy. Data Analysis is the key to all the problems.

Data analytics professionals are primarily mathematicians, statisticians, database/data warehouse engineers, data miners and IT professionals with data warehousing skills.

Facts:

- A full 90% of all the data in the world has been generated over the last two years.(2013)
- 80% of our global data is unstructured.
- One flight produces 140 TB of Data.
- Google processes 800 TB of data per day.
- Facebook generates 500 TB of data per data.
- In 2012, Harvard Business review named Data Scientist the "sexiest job of the 21st century".
- More recently, Glassdoor named it the "Best job of the year" for 2016.

Data Explosion every minute:

- Facebook users share nearly 2.5 million pieces of content.
- Twitter users tweet nearly 300,000 times.
- Instagram users post nearly 220,000 new photos.
- YouTube users upload 72 hours of new video content.
- Apple users download nearly 50,000 apps.
- Email users send over 200 million messages.
- Amazon generates over $80,000 in online sales.

Analytical CRM

The role of analytical CRM systems is to analyse customer data collected through multiple sources, and present it so that business managers can make more informed decisions.

Analytical CRM systems use techniques such as **data mining**, correlation, and pattern recognition to analyse the customer data.

These analytics help improve customer service by finding small problems which can be solved, perhaps, by marketing to different parts of a consumer audience differently.

For example, through the analysis of a customer base's buying behaviour, a company might see that this customer base has not been buying a lot of products recently. After scanning through this data, the company might think to market to this subset of consumers differently, in order to best communicate how this company's products might benefit this group specifically.

Sorting data

Sorting data is an integral part of data analysis. You might want to arrange a list of names in alphabetical order, compile a list of product inventory levels from highest to lowest, or order rows by colors or icons. Sorting data helps you quickly visualize and understand your data better, organize and find the data that you want, and ultimately make more effective decisions.

Excel Auto Filter

- The basic Excel filter (also known as the Excel Autofilter) allows you to view specific rows in an Excel spreadsheet, while hiding the other rows.
- When a filter is added to the header row of a spreadsheet, a drop-down menu appears in each cell of the header row. This provides you with a number of filter options that can be used to specify which rows of the spreadsheet are to be displayed.

Pivot table

- A pivot table is a data processing tool used to query, organize and summarize data or information between spreadsheets, tables or databases. Dragging and dropping fields into a pivot table facilitates rotational, or pivotal, structural changes.
- Among other functions, a pivot table can automatically sort, count, total or give the average of the data stored in one table or spreadsheet, displaying the results in a second table showing the summarized data. Pivot tables are also useful for quickly creating unweighted cross tabulations.(No duplicates)

Power Query

- Power Query is known as Get & Transform which Power Query provides data discovery, data transformation and enrichment for the desktop to the cloud. (clean data)
- Power Query enhances self-service business intelligence (BI) for Excel with an intuitive and consistent experience for discovering, combining, and refining data across a wide variety of sources.

Database-Collection of Data, facts, figures, Tables that can be processed to produce information

Database Model

A database model is a type of data model that determines the logical structure of a database and fundamentally determines in which manner data can be stored, organized, and manipulated.

The most popular example of a database model is the relational model, which uses a table-based format.

Data Model: Multiples tables with relationship.

- A data model explicitly determines the structure of data. Typical applications of data models include database models, design of information systems, and enabling exchange of data. Usually data models are specified in a data modeling language.
- Data models define how the logical structure of a database is modeled. Data models define how data is connected to each other and how they are processed and stored inside the system.

RDBMS (Structured Data)

- RDBMS stands for Relational Database Management System.
- RDBMS is the basis for SQL, and for all modern database systems such as MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access.
- The data in RDBMS is stored in database objects called tables.
- A table is a collection of related data entries and it consists of columns and rows

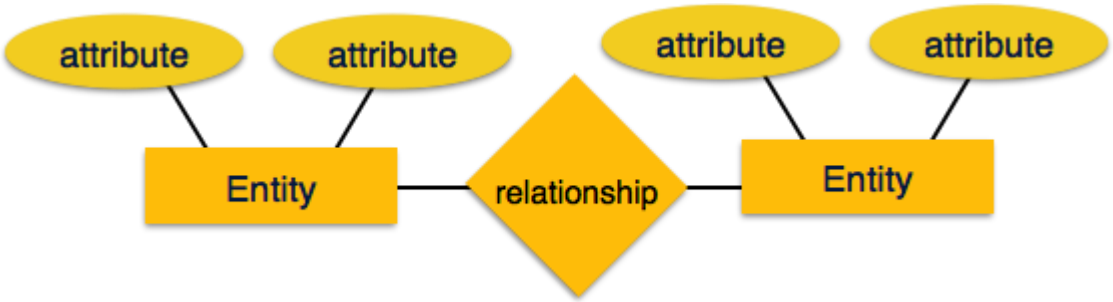
Entity-Relationship Model

Entity-Relationship (ER) Model is based on the notion of real-world entities and relationships among them. While formulating real-world scenario into the database model, the ER Model creates entity set, relationship set, general attributes and constraints.

ER Model is best used for the conceptual design of a database.

ER Model is based on −

- **Entities** and their *attributes.*
- **Relationships** among entities.These concepts are explained below.



- **Entity** −An entity in an ER Model is a real-world entity having properties called **attributes**. Every **attribute** is defined by its set of values called **domain**.
- For example, in a school database, a student is considered as an entity. Student has various attributes like name, age, class, etc. (Rows-Record, / Columns- Field, Attributes)
- **Relationship** –The logical association among entities is called ***relationship***. Relationships are mapped with entities in various ways. Mapping cardinalities define the number of association between two entities. (Table- Relation) (One entire row- Tuple)

Transaction

A transaction can be defined as a group of tasks. A transaction is a very small unit of a program and it may contain several low level tasks.

A transaction in a database system must maintain **A**tomicity, **C**onsistency, **I**solation, and **D**urability − commonly known as ACID properties − in order to ensure accuracy, completeness, and data integrity.

Atomicity − This property states that a transaction must be treated as an atomic unit, that is, either all of its operations are executed or none. There must be no state in a database where a transaction is left partially completed. States should be defined either before the execution of the transaction or after the execution/abortion/failure of the transaction.

Consistency −The database must remain in a consistent state after any transaction. No transaction should have any adverse effect on the data residing in the database. If the database was in a consistent state before the execution of a transaction, it must remain consistent after the execution of the transaction as well.

Durability −The database should be durable enough to hold all its latest updates even if the system fails or restarts. If a transaction updates a chunk of data in a database and commits, then the database will hold the modified data. If a transaction commits but the system fails before the data could be written on to the disk, then that data will be updated once the system springs back into action.

Isolation − In a database system where more than one transaction are being executed simultaneously and in parallel, the property of isolation states that all the transactions will be carried out and executed as if it is the only transaction in the system. No transaction will affect the existence of any other transaction.

Data analysis used in Retail, Healthcare, Weather Forecast and Banking

1) **Retail-** Customer behavior and buying pattern analysis
2) **Healthcare-** Quality of treatment analysis, Maintaining patient records etc
3) **Banking -** Fraud detection

Advantages of Data Analysis

1) Control Data Redundancy- No Duplication, Distinct
2) Data Consistency
3) Data Sharing
4) Data Integration- many tables relationship
5) Data Security
6) Data Constraints
7) Creating forms
8) Data Independence
9) Backup and Recovery

Big Data- (Unstructured Data)

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.

This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

Digital Information such as audios, videos, 3D Images, Medical records, Business reports, Email, Research, and Text files

This data is big data and mainly constitutes the unstructured data. This Big Data offers challenge in term of storage and further analysis in rest of in real time. One can dig gold mine if we are able to make sense out of big data.

6V of Big Data: Velocity, Variety, Volume, veracity, Validity, Volatility

Volume
The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Variety
The type and nature of the data. This helps people who analyze it to effectively use the resulting insight. Structured or Unstructured data.

Velocity
In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

Variability
Inconsistency of the data set can hamper processes to handle and manage it.

Veracity
The quality of captured data can vary greatly, affecting accurate analysis.

The important parts of Data Analysis are:

1) Data Mining 2) Data Warehouse/Data Collection 3) Data Wrangling
4) Data Model 5) Data Visualization / Data Interpretation 6) Predictive Model
7) Data Cleaning / Data Validation 8) Statistical Analysis

Data Mining

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data Mining is about pattern recognition.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Knowledge Discovery Process:-

Data Cleaning − In this step, the noise and inconsistent data is removed.

Data Integration − In this step, multiple data sources are combined.

Data Selection − In this step, data relevant to the analysis task are retrieved from the database.

Data Transformation − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

Data Mining − In this step, intelligent methods are applied in order to extract data patterns.

Pattern Evaluation − In this step, data patterns are evaluated.

Knowledge Presentation − In this step, knowledge is represented

Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data Warehouse

A large store of data accumulated from a wide range of sources within a company and used to guide management decisions. Data warehousing is defined as a process of centralized data management and retrieval.

Data collection

Data is collected from a variety of sources.
The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

Data Wrangling

Data Wrangler is an interactive tool for data cleaning and transformation.
Spend less time formatting and more time analyzing your data.

Data cleaning / Data Validation

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 data is entered and stored. Data cleaning is the process of preventing and correcting these errors.

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Predictive Model: Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends.

Predictive modeling is used widely in information technology (IT). In spam filtering systems, for example, predictive modeling is sometimes used to identify the probability that a given message is spam. Other applications of predictive modeling include customer relationship management (CRM), capacity planning, change management, disaster, recovery, engineering, meteorology, security management and city planning.

1) Connect data to effective actions by drawing reliable conclusion
2) To identify potential risk and opportunities
3) Question of hypothesis

Statistical Analysis: Collection, examination, summarization, manipulation, and interpretation of quantitative data to discover its underlying causes, patterns, relationships, and trends.

Hypothesis testing: A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.

1) Hypothesis testing: Distribution, Sampling, Probability

i) Null Hypothesis (Ho) ii) Alternative hypothesis (H1)

2) Linear Regression- Sales, price, Ads, Brand how 3 factors it influence sales

3) Logistic Regression- Predict if an event occurs, doesn't estimate value. Bank credit check

4) Cluster Analysis- Used to find items similar to other, organizing data into groups - Recommendation

5) Factor Analysis- Groups columns similar to other, large no. of variables

SQL- Structured Query Language

- SQL is an ANSI (American National Standards Institute) standard

SQL PRIMARY KEY Constraint

- The PRIMARY KEY constraint uniquely identifies each record in a database table.
- Primary keys must contain UNIQUE values.
- A primary key column cannot contain NULL values.
- Most tables should have a primary key, and each table can have only ONE primary key.

SQL FOREIGN Key:

- A foreign key is a key used to link two tables together. This is sometimes called a referencing key.
- Foreign Key is a column or a combination of columns whose values match a Primary Key in a different table
- The relationship between 2 tables matches the Primary Key in one of the tables with a Foreign Key in the second table

SQL Commands:

The standard SQL commands to interact with relational databases are CREATE, SELECT, INSERT, UPDATE, DELETE and DROP. These commands can be classified into groups based on their nature:

DDL - Data Definition Language:

Command	Description
CREATE	Creates a new table, a view of a table, or other object in database
ALTER	Modifies an existing database object, such as a table.
DROP	Deletes an entire table, a view of a table or other object in the database.

DML - Data Manipulation Language:

Command	Description
SELECT	Retrieves certain records from one or more tables
INSERT	Creates a record
UPDATE	Modifies records
DELETE	Deletes records

DCL - Data Control Language:

Command	Description
GRANT	Gives a privilege to user
REVOKE	Takes back privileges granted from user

SQL has many built-in functions for performing processing on string or numeric data. Following is the list of all useful SQL built-in functions:

1) **SQL COUNT Function** - The SQL COUNT aggregate function is used to count the number of rows in a database table.

2) **SQL MAX Function** - The SQL MAX aggregate function allows us to select the highest (maximum) value for a certain column.

3) **SQL MIN Function** - The SQL MIN aggregate function allows us to select the lowest (minimum) value for a certain column.

4) **SQL AVG Function** - The SQL AVG aggregate function selects the average value for certain table column.

5) **SQL SUM Function** - The SQL SUM aggregate function allows selecting the total for a numeric column.

6) **SQL SQRT Functions** - This is used to generate a square root of a given number.

7) **SQL RAND Function** - This is used to generate a random number using SQL command.

8) **SQL CONCAT Function** - This is used to concatenate any string inside any SQL command.

9) **SQL Numeric Functions** - Complete list of SQL functions required to manipulate numbers in SQL.

SQL String Functions - Complete list of SQL functions required to manipulate strings in SQL.

What Can SQL do?

- SQL can execute queries against a database
- SQL can retrieve data from a database
- SQL can insert records in a database
- SQL can update records in a database
- SQL can delete records from a database
- SQL can create new databases
- SQL can create new tables in a database
- SQL can create stored procedures in a database
- SQL can create views in a database
- SQL can set permissions on tables, procedures, and views

- CREATE TABLE CUSTOMERS (
- ID INT NOT NULL,
- NAME VARCHAR (20) NOT NULL,
- AGE INT NOT NULL,
- ADDRESS CHAR (25) ,
- SALARY DECIMAL (18, 2),

PRIMARY KEY (ID)

Select ID, Name, DOB, Age
From table_Name
Where conditions
Order by - Asc/Des

Logical Operator- And, Or, Not ,True, false
Arithmetic- +, -, *, /
Comparison- >, <, in like, between
Group by and **Having clause** works with **Aggregate function**

First Name- Char – Single quotes required ‘’
Age – Int Single quotes not required
First name Start with S – **like ‘s%’ ‘_a%’**
Each record should have a unique identifier

Data Frame- Tabular data 1) www.rstudio.com 2) www.kaggle.com 3) SAS 4) IBM

Text- Left
Number.- Right

Time – No Space 7am
Date- mm/dd/yyyy

1) Visual C++
2) MySQL for Excel 1.3.6
3) MySQL Workbench 6.3.6
4) MySQL for Visual Studio 1.2.6
5) MySQL Fabric 1.5.6
6) MySQL Connector Python 3.4

Amazon- Oracle RDBMS
Facebook- MySQL
Twitter – Bootstrap

Hadoop

Apache Hadoop is an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.
The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce.

Marketo

Marketo is a web service that provides tools for managing marketing campaigns. It is provided by Marketo, SaaS revenue performance management company specializing in marketing automation and sales effectiveness software for mid-size and enterprise business-to-business (B2B) companies.

Hubspot

HubSpot provides tools for social media marketing, content management, web analytics, landing pages and search engine optimization.
The HubSpot suite of online tools has three primary applications:
1) Content management tools for creating or managing blogs, templates, forms and landing pages;
2) Exposure optimization applications that help the content be found, such as through search engine optimization;
3) Lead tracking and intelligence tools, which track and manage e-mail marketing, customer interactions, qualified prospects, reports and analysis.

1 Bit = Binary Digit
· 8 Bits = 1 Byte
· 1024 Bytes = 1 Kilobyte
· 1024 Kilobytes = 1 Megabyte
· 1024 Megabytes = 1 Gigabyte
· 1024 Gigabytes = 1 Terabyte
· 1024 Terabytes = 1 Petabyte
1024 Exabytes = 1 Zettabyte
1024 Zettabytes = 1 Yottabyte
1024 Yottabytes = 1 Brontobyte
1024 Brontobytes = 1 Geopbyte
1024 Petabytes = 1 Exabyte

Excel Keyboard commands

1) Esc Key:

i. Closes Backstage View (like Print Preview)
ii. Closes most dialog boxes.
iii. If you are in Edit mode in a Cell, Esc will revert back to what you had in the cell before you put the Cell in

2) F2 Key = Puts formula in Edit Mode and shows the rainbow colored Range Finder.

3) **SUM** Function: Alt + =

4) **Ctrl + Shift + Arrow** = **Highlight column** (Current Region).

5) **Ctrl + Backspace** = Jumps back to Active Cell

6) **Ctrl + Z** = Undo.

7) **Ctrl + Y** = Undo the Undo.

8) **Ctrl + C** = Copy.

9) **Ctrl + X** = Cut.

10) **Ctrl + V** = Paste.

11) **Ctrl + PageDown** =expose next sheet to right.

12) **Ctrl + PageUp** =expose next sheet to left.

13) **Ctrl + 1** = Format Cells dialog box, or in a chart it opens Format Chart Element Task Pane.

14) **Ctrl + Arrow**: jumps to the bottom of the "Current Region", which means it jumps to the last cell that has data, right before the first empty cell.

15) **Ctrl + Home** = Go to Cell A1.

16) **Ctrl + End** = Go to last cell used.

17) **Alt** keyboards are keys that you hit in succession. Alt keyboards are keyboards you can teach yourself by hitting the Alt key and looking at the screen tips.

i. Create PivotTable dialog box: **Alt, N, V**
ii. Page Setup dialog box: **Alt, P, S, P**
iii. Keyboard to open Sort dialog box: **Alt, D, S**

18) **ENTER** = When you are in Edit Mode in a Cell, it will put thing in cell and move selected cell DOWN.

19) **CTRL + ENTER** = When you are in Edit Mode in a Cell, it will put thing in cell and keep cell selected.

20) **TAB** = When you are in Edit Mode in a Cell, it will put thing in cell and move selected cell RIGHT.

21) **SHIFT + ENTER** = When you are in Edit Mode in a Cell, it will put thing in cell and move selected cell UP.

22) **SHIFT + TAB** = When you are in Edit Mode in a Cell, it will put thing in cell and move selected cell LEFT.

23) **Ctrl + T** = Create Excel Table (with dynamic ranges) from a Proper Data Set.

i. Keyboard to name Excel Table: **Alt, J, T, A**
ii. Tab = Enter Raw Data into an Excel Table.

24) **Ctrl + Shift + ~ (`)** = General Number Formatting Keyboard.

25) **Ctrl + ;** = Keyboard for hardcoding today's date.

26) **Ctrl + Shift + ;** = Keyboard for hardcoding current time.

27) **Arrow Key** = If you are making a formula, Arrow key will "hunt" for Cell Reference.

28) **Ctrl + B** = Bold the Font

29) **Ctrl + *** (on Number Pad) or **Ctrl + Shift + 8** = Highlight Current Table.

30) **Alt + Enter** = Add Manual Line Break (Word Wrap)

31) **Ctrl + P** = Print dialog Backstage View and Print Preview Edit mode.

32) **F4 Key** = If you are in Edit mode while making a formula AND your cursor is touching a particular Cell Reference, F4 key will toggle through the different Cell References:

i. **A1** = Relative
ii. **A1** = Absolute or "Locked"
iii. **A$1** = Mixed with Row Locked (Relative as you copy across the columns AND Locked as you copy down the rows)
iv. **$A1** = Mixed with Column Locked (Relative as you copy down the rows AND Locked as you across the columns)

33) **Ctrl + Shift + 4** = Apply Currency Number Formatting

34) **Tab key** = When you are selecting a Function from the Function Drop-down list, you can select the function that is highlighted in blue by using the Tab key.

35) **F9 Key** = To evaluate just a single part of formula while you are in edit mode, highlight part of formula and hit the F9 key.

i. If you are creating an Array Constant in your formula: Hit F9.
ii. If you are evaluating the formula element just to see what that part of the formula looks like,

REMEMBER: to **Undo** with **Ctrl + Z.**

36) **Alt, E, A, A** = Clear All (Content and Formatting)

37) Evaluate Formula One Step at a Time Keyboard: **Alt, M, V**