**Analytics** is the application of logical and computational reasoning to the component parts obtained in an analysis.

**Business Intelligence** is the process of analyzing and reporting historical business data. It aims to explain past events using business data. It fits into Data Science as it is the preliminary step of predictive analytics.

**Machine Learning** is the ability of machines to predict outcomes without being explicitly programmed to do so. It is about creating and implementing algorithms that let machines receive data and use this date to *make* *predictions, analyze patterns and give recommendations* on their own. For example: creating real-time dashboards.

**Artificial Intelligence** is the simulation of human knowledge and decision making with computers. AI is possible due to machine learning.

**Data Collection Methods**: Surveys, Web Cookies,

* **Pre-processing:**

*Class Labeling: Numerical/Categorical, identifies which of the collected data can be manipulated or not.*

*Data Cleansing: Misspelt data etc.*

*Missing Values: Fields are left missing, hence an average can be used*

* **Balancing Techniques:** Example if 80% of survey respondents are from a certain demographic the results will be representative of that demographic, we have to balance the data before using it.
* **Data Shuffling:** Prevents unwanted patterns and helps avoid misleading results.
* **Text Data Mining:** The process of deriving valuable, unstructured data from a text
* **Data Masking:** Conceals the original data with random and false data

**Data Analysis for BI (Past):** Observations -> Quantifications -> Measures -> Metrics -> KPIs (Dashboards, Visuals)

* **Observations:** Keeping track of key variables. Manipulations cannot be applied until the data is quantified
* **Quantifications:** The process of representing observations as numbers.
* **Measures:** The accumulation of observations to show some information
* **Metric:** Way to gauge business performance/progress KPI = Measure + Business Meaning. Useful for comparisons.
* **KPIs:** Not smart to track entire load of business data, only useful to pay attention to key indicators
* **Example:** Metric provides info on company website traffic, KPI provides traffic only generated by users who clicked on a web link in a marketing campaign.
* **Real Life Use:** Price Optimization, Inventory Management etc.

**Predictive Analysis (Future):** Analyzing potential future scenarios by using advanced statistical methods.

* **Regression:** A model for quantifying casual relationships among the different variables included in your analysis. Linear Models: y = bx (where y is the vertical axis, x is the horizontal and b is the coefficient).
* **Logistic Regression:** the values on the vertical line will be 1s and 0s only.
* **Cluster Analysis:** Grouping observations that exhibit similar behaviours. Location affecting house prices.
* **Factor Analysis:** Grouping explanatory variables that measure a similar attitude into one factor.
* **Time Series:** Development of certain values over time. Plotting them against time on the x-axis (indep.).
* **Real Life Use:** Clustering: UX experience across various continents, Time Series: Sales Forecasting.

**Machine Learning:** Creating an algorithm which a computer uses to find a model that best fits the data and makes very accurate predictions based on that.

* **ML Algorithm:** Trial and Error process where each consecutive trial is at least as good as the previous one. Data -> Model -> Objective Function -> Optimization Algorithm
* **Example:** Data (Arrow Types) -> Model (Usage of Bow) -> Objective Function (Distance from Target on every shot, measure inaccuracy) -> Optimization Algorithm (Mechanics that will improve the model’s performance e.g. posture, pull, etc.)
* **Training the model:** No rules on firing the bow. only to hit the target. It stops when the model is trained, the objective function is minimized, and the optimization function has done its job in leaning the most accurate technique.
* **Benefits + Uses:** Machines can learn to act more effectively than humans + Improves complex computational models.
* **Supervised Learning + Labelled Data:** Associating or labelling a target to a type of arrow. Meaning we know the target prior to the shot, and we can associate that shot with a target.
* **Unsupervised Learning + Unlabelled Data:** Looks for a model that divides the arrows in a certain way. There are no predefined targets. It shoots the arrows at random and gathers the results of the shots per type of arrow. In practice, this saves times and resources that are required to predefine outcomes.
* **Uses:** One may start with unsupervised learning to create an arrow categorization model and then apply supervised learning to optimize a shooting model for using the bow.
* **Reinforcement Learning:** A reward system, the model is aimed at maximizing rewards as opposed to minimizing errors, positive reinforcement.
* **Real Life Use:** Fraud Prevention at Banks by analysing legitimate and illegitimate techniques. Client Retention using reward programs that are personalized to customer shopping styles or needs.

**Programming Languages:** Allow us to design programs that can execute specific operations and we can reuse these programs whenever we need to execute the same action.

* **Python + R:** Commonly used across all disciplines.
* **Excel:** Useful with traditional Data and Visualizations
* **SQL:** Useful for big data and machine queries.

**Descriptive Statistical Analysis:** Definition of key terms.

* **Population (N):** Collection of all items of interest. The numbers we obtain when working with a population is called parameters. Hard to define and observe in real life.
* **Sample (n):** A subset of the population. The numbers we obtain when working with a sample is called statistics. Easier to focus on. For a sample to be valid, it must meet the following conditions:
  1. **Randomness:** A random sample is a collected when each member of the sample is chosen from the population by chance.
  2. **Representativeness:** A representative sample is a subset of the population that accurately represents the member of the population

**Different Types of Variables:** Classifying Data based on Types and Measurement Levels

* **Categorical Data:** Classifies categories or groups e.g. Car Brands, Yes or No Answers
* **Numerical Data:** Data that is quantified or recorded in numbers.
  1. **Discrete:** Can be counted in a finite number.E.g. number of children, SAT scores
  2. **Continuous:** Infinite and impossible to count. E.g. Body Weight, Height, Time

**Levels of Measurement:**

* **Qualitative Data:** These are non-numerical and can either be Nominal or Ordinal
  1. **Nominal:** Similar to categorical, they cannot be ordered. E.g. Car brands, states etc.
  2. **Ordinal:** Groups or categories that can be placed in a particular order. E.g. Disgusting to Delicious
* **Quantitative Data:** These are represented by numbers and can be Interval or Ratio.
  1. **Interval:** These do not have areal 0 and cannot be represent in ratios. E.g. Temperature in C or F
  2. **Ratios:** These are common and have a true 0 E.g. number of objects, Temperature in Kelvin.

**Types of Graphs (One Variable)** “Visualizing data is the most intuitive way to visualize it”.

* **Representation of Categorical Data** Frequency Distribution Tables, Bar Charts, Pie Charts, Pareto Diagrams (similar to a bar graph but with a curve showing the cumulative frequency). etc.
  1. **Frequency** is the number of occurrences of an item.
  2. **Pareto Principle:** 80% of the effect come from 20% of the causes: 80-20 Rule. The pareto diagram shows how subtotals change with each additional category and provide us with a better understanding of our data.

* **Representation of Numerical Data:**
  1. **Frequency Distribution Tables** representing data by the frequency of occurrence. We can use equal width intervals to reduce the size of the collected data. A histogram is a common way to visualize this data. Also, intervals may be of varying widths.

**Types of Graphs (Multiple Variables)** Used to represent the relationship between two variables.

* **Cross tables:** Useful for categorical data. Can be visualized with side-by-side bar chart.
* **Scatter Plot:** Useful for analysing numerical data. Used to represent lots of observations. Interest is usually in the major distribution of the data. Outliers go against the main logic of the whole data set.

**Measures of Central Tendency:** Mean, Median and Mode. These are to be used together as one is not entirely representative of the data set.

* **Mean:** The simple average of a data set. , . Calculated by adding up all the components and dividing by the number of components. It is easily affected by outliers. Thus, the mean is not enough to make definite conclusions.
* **Median:** The middle number in an ordered data set. This is the number in position (n+1z)/2 in the ordered list, where n is the number of observations. Although it is not affected by outliers, it does not get the full picture.
* **Mode:** Can be used for both categorical and numerical data. It is the value that occurs the most often in the data set.

**Measures of Asymmetry:** This is the link between measures of central tendency and probability theory

* **Skewness:** Indicates whether the data is concentrated on one side.
  1. **Positive Skew:** Mean > Median, that is outliers are to the right of the graph. Right skewness means that the tail is leaning to the right.
  2. **Zero Skew:** Mean = Median = Mode. The frequency of occurrence is symmetrical.
  3. **Negative Skew:** Mean < Median, that is outliers are to the left of the graph. Left skewness means that the tail is leaning to the left.

**Measures of Variability:**

* **Variance:** Measures the dispersion of a set of data points around their mean.
  1. **Population Variance:**

Numerator: The sum of differences between the observed variables and the mean, squared. Lower differences mean observed variables are closer to the mean and vice versa. The squaring of the result is to get non-negative results as dispersion (distance) is perceived to be non-negative and the sum of positive and negative results will cancel out. Also, squaring amplifies the effect of large differences.

* 1. **Sample Variance:**

With a sample there is more uncertainty as the entire population is unknown. Thus, the sample variance formula yields a slightly larger result (n-1) in order to reflect the higher potential variability.

* **Standard Deviation:** This is the square root of the obtained variance.
* **Coefficient of Variation:** This is the relative standard deviation. That is the standard deviation divided by the mean or relative to the mean.

**Measuring the relationship between two variables**

* **Covariance:** Sample**:**  , Population: if the covariance is positive > 0, the two variables move together. < 0, the two variables move in opposite directions. = 0, the variables are independent.
* **Correlation:** Adjusts covariance, so that the relationship between the two variables becomes easy and intuitive to interpret. Correlation coefficient is the , for either sample or population.
* **Correlation Coefficient:** It is a value between 1 and -1. A value of 1 denotes the entire variability of one variability is explained by the other. 0 means the variables are independent. Could be perfect negative coefficient of -1 or an imperfect negative correlation between -1,0. Common practice is to disregard value below 0.2.
* **Causality:** Although correlation between x,y is the same as y,x, causality may not be true in both directions. It is important to understand the direction of causal relationships. For example, the sizes of houses affect or cause the price, but prices do not cause the size. Therefore, correlation does not imply causality.

**Inferential Statistics:** Methods that rely on probability theory or distributions to predict population values based on sample data.

**Distribution:** This is a function that shows the possible values for a variable and the probability of their occurrence. It usually means a probability distribution, it can be normal, binomial or uniform in nature.

* **Discrete Uniform Distribution:** This means all possible outcomes have the same probability of occurrence.
* **Normal/Gaussian Distributions:** Often referred to as the bell curve, its mean, median and mode are equal. Thus, there is no skew. The spread of the graph is determined by the standard deviation. Controlling for the standard deviation, the graph keeps its shape but moves either to the right (lower mean) or to the left (higher mean). Controlling for the mean changes the shape of the graph, higher middle and flatter tails (lower std. deviation) or flatter middle and fatter tails (higher std. deviation)
* **Standardization:** ­-> :

Is the process of transforming the variable to a mean of 0 and standard deviation of 1, using Z = .

Z is the same as the Z-score, z is the critical value from the z table

**Central Limit Theorem:** No matter the underlying distribution, the sampling distribution approximates a normal. The mean is the same for the population and sample, but the sample variance is the population variance divided by the sample size. Thus, the more samples we draw, the lower the variance and the closer we are to approximating the population. The more samples we draw, the more accurate we get. For CLT t apply we need a sample of n>30.

* **Standard error**: = : This is the standard deviation of the samples we draw. It is denoted by the root of the sample variance, which is the population variance divided by the number of samples drawn. It is the variability of the means of the samples drawn. Standard error decreases when sample size increases.
* **Estimators:** Point estimates vs confidence intervals. A point estimate is a single number given by an estimator. Example: Dataset with a mean of 10, in this case 10 is a point estimate. (The estimator in this case is a point estimator and is the formula for the mean). The point estimate is the mid-point of the confidence interval. Efficiency: The most efficient estimator means the unbiased estimator with the smallest variance.

**Confidence levels and Confidence Intervals:** A confidence level is the prediction that an estimate is within the suggested confidence interval. It is denoted by , where alpha is . Thus, a confidence interval of 95% requires an alpha of 5%. Basically, the confidence interval is the mean the Margin of Error. A higher statistic increases the ME. A higher standard deviation increases the ME. A higher sample size decreases the ME.

* **Confidence intervals with KNOWN Population Variance:** Assume a large enough sample to get normal distribution. The confidence interval is denoted by , where The Z of alpha is obtained using the z-table (1 – alpha/2 and then we find the closest value within the table and take the corresponding x,y axis values). Thus, higher confidence levels result in a broader confidence interval.
* **Student’s T distribution:** The Student’s T distribution approximates the Normal distribution but has fatter tails. This means the probability of values being far away from the mean is bigger. For big enough samples, the Student’s T distribution coincides with the Normal distribution.
* **Confidence intervals with UNKNOWN Population Variance:**  In this the confidence interval is denoted by where we are using t-statistics and s, sample std. deviation. Given the same confidence level, we get a narrower confidence interval with a known population variance compared to an unknown scenario.

**Dependent vs. Independent Samples:** This depends on the relatability or cause and effect each sample has on the other.

* **Dependent:** where d is the relating to the difference before and after.
  1. Before and After situation (same sample being used/tested)
  2. Cause and Effect
* **Independent Scenarios**
  1. Population Variance Known
  2. Population variance unknown but assumed equal
  3. Population variance unknown but assumed to be different.

1. **Population Variance Known**

* Confidence Interval **z**:
* Variance of the difference is: **,** n is sample size.

1. **Population variance unknown but assumed equal**

* Confidence interval **t**:
* T: degrees of freedom = total sample size minus the number of variables:
* Use a **Pooled Variance formula** to estimate the variance for both samples
* Pooled Variance Formula: , based on sample sizes and standard devs.
* Variance of the Difference:

1. **Population variance unknown but assumed to be different.**

* Confidence interval **t**:
* T: degrees of freedom, v = *complex and of no value*
* Variance Difference:  **,**

**Hypothesis Testing: Data Drive Decision Making**

**Steps:**

1. Formulate a Hypothesis
2. Find the Right Test
3. Execute the Test
4. Review and make decision

A hypothesis is an idea or statement that can be tested.

**Null hypothesis (Ho):** The idea to be tested = $X

**Alternative (Hi):** Everything outside of the idea to be tested $X

Outcomes of tests refer to the population parameter rather than the sample statistic. Often times we are trying to reject the Null hypothesis (present state of affairs) and establish the alternative (our idea/statement).

**Rejection Region and Significance Level**

**Significance Level:** = The probability of rejecting the null hypothesis, if it is true. 0.01, 0.05\*, 0.1. If the test value falls into the rejection region, you will reject the null hypothesis.

**Errors in Hypothesis Testing**

**Type I error (False Positive):** When true null hypothesis is rejected. The probability of making this error lies in , the level of Significance. The tester is responsible for choosing alpha.

**Type II error (False Negative):** When a false hypothesis is accepted. The probability of making this error lies in , which depends on sample size (n) and magnitude of effect (variance). The probability of reject a false null hypothesis is = , the power of the test.

**Variance Known: z**

**Note that a -X < a negative z = +X > a positive z:**

**Thus, Reject if: absolute value of the Z-score > positive critical value (z).**

**Level of Significant after which we cannott reject the null hypothesis.: P-Value.**

* P value for one-sided = 1 – number from the table, two-sided = 2\*(1 – number from table)
* Reject the null hypothesis, if the P-Value <

**Variance Unknown: t**

**Note that a -X < a negative z = +X > a positive z:**

**Thus, Reject if: absolute value of the t-score > positive critical value (t).**

* P value for one-sided = 1 – number from the table, two-sided = 2\*(1 – number from table)
* Reject the null hypothesis, if the P-Value <
* **If we cannot reject a test at 0.95 significance level, we cannot reject at smaller levels either**

**Z - statistic: Big Samples, Known population variances**

**t – statistic: Small Samples. Unknown population variances**

True value tends to be much lower than hypothesized value, thus is z/t-score is negative in an absolute test, the actual figure may be lower.

**As a rule of thumb, researchers believe 2 to be sufficiently significant for a T-test**

**SQL: Structured Query Language**

Used to create, manipulate and share Data from Relational Database Management Systems.

**Query:** A piece of code that induces the computer into an action to produce a desired output.

**Record:** Row of data

**Field**: A column in a table containing specific information about every record in the data

**Relational Algebra:** Allows us to use mathematical logic and create a relational connection between a few tables in a way that allows us to retrieve data efficiently, a few related tables form a relational database

**Entity:** The smallest unit that can contain a meaningful set of data. Rows/columns = horizontal/vertical entity

**Relational Database Essentials:**

**Database goal**: Organize huge amounts of data that can be quickly retrieved. Picture databases as circles which form a Venn diagram with an overlap of field similar to both tables (relationships).

* Compact
* Well-structured
* Efficient (speed of extraction)

**Relations:** Smallest units in the entire system that can carry integral logical meaning.

**Database vs. Spreadsheets**

**Spreadsheets:** An electronic ledger or an electronic version of worksheets or tabular data.

**Database:** the field types are pre-set and cannot allow a different type of data; it is raw and does not allow formatting. Formulas cannot be applied until after the data is retrieved from the database. It is aimed to improve data integrity.

* Data Consistency
* Data Integrity
* Speed of Retrieval and Update

**Database Terminology:**

**Database Design:** Plot the entire database system on a canvas using a visualization tool.

* Entity-Relationship (ER) diagram: Different figures represent different data entities and the specific relationships there are between entities.
* Relational Schema: An existing idea of how the database must be organized

**Database Creation:** Physical or actual setup of the design in a database management tool, SQL

**Data Manipulation:** Allows the use of database in extracting meaningful business insights

**Database Administration:** Daily maintenance of the database

**Relational Schema – Primary Key**

**Primary Key:** A column (or set of columns) whose values exists and is unique for every record in a table.

* Each data table must have only one primary key, e.g. Purchase No.
* In terms of a set of columns, more than one column may be combined to forma unique pair e.g. Purchase No. + Date of purchase
* The unique items or pairs cannot occur more than once.
* **Cannot contain null values, there must always be a value.**
* Not all tables will necessarily have a primary key

A screenshot of a cell phone

Description automatically generated**A screenshot of a cell phone

Description automatically generated**

A close up of a map

Description automatically generated

**Relational Schema – Foreign Key (FK):** Identifies the relationship between tables, not the tables themselves

**A picture containing screenshot

Description automatically generatedA picture containing screenshot

Description automatically generatedA screenshot of a cell phone

Description automatically generated**

* A foreign key which is a primary key in another table does not need to meet the primary key requirements in the foreign table. However, the foreign field will only accept values that are listed or stored in the primary field. Example: Primary - Customer ID: 1 to 4, Foreign-Customer ID cannot have ID numbers above 4 (error).

**Relational Schema – Unique Key and Null Values**

**Unique Key:** Used whenever you would like to specify that duplicates data cannot be given in a field e.g. phone numbers. As no two companies have the same phone number.

* They **differ from Primary Key as these values can be null.**
* There can be multiple unique key fields.
* Can be combined in pairs of multiple fields

**Relationships:** Tell how much of the data from a foreign key can be seen in the primary key column of the table the data is related to and vice versa.

**One-to-many (Many-to-One): One** value from the *customer\_id* column under “Customer” table can be found **many** times in the *customer\_id* column in the “Sales” table.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Flowing in this direction Sales Customers: A customer has to make at least one purchase to appear in the sales data.

* The angular symbol: means Many
* The Line symbol:means One

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Flowing in this direction Sales Customers: Regardless of how many purchases made, a customer’s info can only be listed once in the customer table.

* **Using the symbols closest to the rectangles**, we can establish the relationships
* Sales Customers: One-to-Many & Sales Customers: Many-to-One,
* The symbols used to express relationships are called cardinality constraints
* **Other relationship types:** One-to-One, Many-to-Many

**Relationship schemas:** Represents the concept database administrators must implement

* Depict how a database is organised
* Blueprints or a plan for a database
* Help in writing queries

**Clean Code:** code that is *focused* and *understandable*, which means it must be readable logical and changeable.

* Good code is not the code computers understand but code that humans can understand.
* Good coding practice implies choosing a version that will be easiest to read and understand
* As simple as possible, perfectly organized and maintaining a good logical flow

**Naming:** when assigning names to variables or SQL object choose names that are **short, meaningful and convey specific information**. It should be **pronounceable**, where one word per concept had been picked.

* No blank space between names, either Xxx\_Yyy or XxxYyy.

**Best Practices:**

* Use ad-hoc software that re-organizes code and colours different words consistently.
* Use the relevant analogical tool provided in workbench (brush symbol or Ctrl B)
* Intervene manually and adjust code as you like. E.g. Indentation

**Comments:** Lines of text that workbench will not run as code; they convey a message to anyone reading the code.

* Use /\* … \*/ for large comments.
* Use #, ##, -- for one-line comments

**Running Code:** Semi-colons**;** are used to separate blocks of code in workbench.

* Use the lightning sign (Ctrl+Shift+Enter) to run the block of code associated with the cursor placement.
* Use the lightning/cursor sign (Ctrl+Enter) to execute the line or statement at the cursor placement.

**Aggregate Functions:** They are applied on *multiple rows* of a *single column* and ***return***an output of **a single value**