**Types of Data:**

1. Qualitative or Categorical: A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called qualittative or categorical variable. The data collected on such a variable are called qualitative data
   1. Nominal: The values are not ordered. Example: Nationality, Gender, etc
   2. Ordinal: The values are ordered or ranked. Example: Satisfaction score, spiciness of food
2. Quantitative: A variable that can be measured numerically is called quantitative variable. The data collected on a quantitative variable is called quantitative data
   1. Discrete: A variable whose values are countable is called a discrete variable. In other words, a discrete variable can assume only certain values with no intermediate values. Example: Number of heads in 10 tosses
   2. Continuous: A variable that can assume any numberical value over a certain interval or intervals is called a continuous variable. Example: Height of a person

**Types of Statistics:**

* Descriptive: Consists of methods for oorganizing , displaying, and describing data by using tables, graphs and summary measures
  + Measures of central tendency:
    - Mean
    - Median
    - Mode
  + Measures of dispersion
    - Range
    - Standard deviation
  + Frequency distributions
  + Histograms
* Inferential: Inferential statistics consissts of methods that use sample results to help make decisions or predictions about a population.
  + Tools:
    - ANOVA
    - Linear Regression
    - Decision Trees
    - Neural Networks

**Frequency Distribution:**

* A frequency distribution for qualitative data lists all categories and the number of elements that belong to each of the categories. It is a type of descriptive statistics.
* A frequency distribution for quantitative data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution iw called group data.

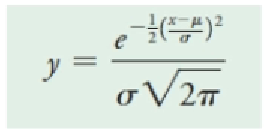
A **bar graph** is a graph made of bars whose heights represent the frequencies of respective categories.

A **histogram** is a graph consisting of bars of equal width drawn adjacent to each other. The **horizontal scale** represents the quantitative data and the **vertical scale** represents frequencies.

Histogram shapes:

* Symmetric
* Skewed(not symmetric)
* Uniform or rectangular

**Normal Distribution:** If a continuous random variable with a graph that is symmetric and bell-shaped, and it can be described by the equation given below, we can say that it has normal distribution.



Properties of normal distribution:

* The total area under the curve is 1.0
* The curve is symmetric about the mean
* The two tails of the curve extend indefinitely

**Comparision of centres:**

* For a symmetric curve with one peak, the values of the mean, median and more are identical
* For a distribution curve skewed to the right, the value of mean is the largest, that of the mode is the smallest, and the value of the median lies between these two
* Mean is not always the best measure of central tendency because it is heavily influenced by outliers. Median is preferred over mean for this reason
* One advantage of the mode is that it can be calculated for both kinds of data - qualitative and qualtitative whereas mean and median can be calculated only for quantitative data

If we have exceptionally large or small values(outliers) -> Either remove the Outliers and then find the mean or prefer median over mean

**Measures of Dispersion:**

The measures that help us learn about the spread of a data set are called measures of dispersion

* Range: The difference between the largest value and the smallest value in a data set. It is influenced by outliers, therefore may not be very useful.
* Standard Deviation: is the square root of variance
* Variance: the average of the squared distances from the mean

Varaince = Standard Deviation ^ 2

**Data Dictionary:** A comprehensive Data Dictionary should include

* Definition of predictors
* Unique identifier for each table(or primary keys)
* Foreign keys or matching keys between tables
* Edxplaination of values n case of categorical values

**Univariate Analysis:** Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships(unlike regression)and it’s major purpose is to describe ; it takes data, summarizes that data and finds patterns in that data.

Ways to describe pattersn found in univariate data:

* Central tendency:
  + mean
  + Median
  + Mode
* Dispersion:
  + Range
  + Varaince
  + Maximun, Minimum
  + Quartiles
  + Standard Deviation
* Count/Null count

**Outlier Treatment:** Outlier is a commonly used terminology by analysts and data scientists, Outlier is an observation that appears far away and diverges from overall pattern in a sample

* Reasons:
* Data Entry Errors
* Measurement error
  + smapling error
* Impact:
  + Increases the error varaince and reduces the power of statistical debt
* Solution:
  + Detect outliers using EDD and visualization methods such as scatter plot, histogram or box plots
  + Impute outliers

Methods of Outlier Treatment:

* Capping and flooring
* Exponential smoothing
* Sigma approach

**Missing value imputation:** Real-world data often has missing values. Data can have missing values for a number of reasons such as observations that were not recorded or data corruption

* Impact:
  + Hadling missing data is important as many machine learning algoriths donot support data with missing values
* Soulution:
  + Remove rows with missing data from your dataset
  + IMpute missing values with mean/median values in your dataset
* Note:
  + Use busingess knowledge to take separate approach for each variable
  + It is advisable to impute instead of remove in case of small smple size or large proportion of observations with missing values
* Methods:
  + Impute with zero
  + Impute with mean/median/mode
  + Segment based imputation

**Seasonality:** is th presence of variaitons that occur at a specific regular interval less than a year, such as weekly, monthly or quarterly

* Reasons:
  + Weather
  + Vacation
  + Holidays
* Examples:
  + Ice cream sales
  + Christmas sales
* Solution:
  + calculate multiplication factor for each month as 
  + multiply each observation with its multiplication factor

**Bivariate analysis:** it is the simultaneous analysis of two variables(attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences

* Scatter plot:
  + indicates the type of (liner or non-linear) and the strength of the relationship between two variables
* Correlation:
  + quatifies the strength of a linear relationship between two numerical variables
  + when there is no correlation between two variables, there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity
  + it is used to drop non usable variables

**Variable Transformation:** transform your existing variable to extract more information out of them.

* Identify:
  + using your business knwoledge and bivariate analysis to modify variable
* Methods:
  + Use mean/median of variables conveying similar type of information
  + Create ratio variable which are more relevant to business
  + transform varaible by taking log,exponential, roots, etc

**Non Usable Variables:** Identify the non usable variables to reduce the dimensions of the dataset

* Variables with a single unique value
* Variables with low fill rate
* Variables with low regularity issue
* Variable with no business sense

**Dummy Variable:** a dummy variable of indicator variable is an artificial variable created to represent an attribute with two or more distinct categories or variables

* Why:
  + regression analysis treats all independent (X) variables in the analysis as numerical
  + Nominal variables, or variables that describe a characteristic using two or more categories, are commonplace in regression research, but are not always isable in their categorical form.
  + dummy coding is a way of incorporating nominal variables into regression analysis
* How:
  + we can make a separate column, or variable for each category
  + This new variable can take value 0 or 1 depending on the value of the categorical variable
* The number of dummy variables necessary to represent a single attribute variable is equal to the number of levels(categories) in that variable minus one.

**Correlation:** it is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable icreases as the other decreases

* Definition:
  + a correlation coeeficient is a way to put a value to the relationship
  + correlation coeeficients have a value of between -1 and 1
  + a 0 means there is no relationship between the variables at all
  + while -1 or 1 means that there is a perfect negative or positive correlation

**Causation:** The relation between something that happens and the thing that causes it . The first thing that happens is the cause and the second thing is the effect

**Correlation Matrix:**

* Definition:
  + it is a table showing correlation coeeficients between variables
  + each cell in the table shows the correlation between two variables
  + A correlation matrix is used as a way to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses
* Application:
  + to summarize a large amount of data where the goal is to see patterns
  + to identify collinearity in the data

**Linear Regression:** it is a linear approach to modelling the relationship between a dependent variable and one or more independent variables

**Simple Linear Regression:**it is an approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y.