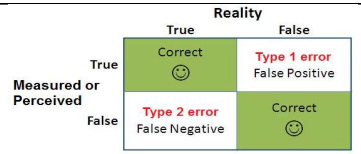
Sl No	Term	Definition	Formula / Comments
1	**Correlation (r)**	Correlation is a statistical technique used to determine the degree to which two quantitative type variables are related without being able to infer causal relationships.	
2	**Quantitative Variables**	Numerical or Continuous variables [eg. Age, Salary, Blood Pressure, Temperature]. Can take any value (infinite) **Discrete:** Variables that can only be represented by whole numbers or integers [eg. No of students in a class] **Continuous / Ratio:** Variables that can take both Integer and Decimal values [eg. Salary, Weight etc.]	Sometimes it depends on the source of data. So ask about the source before answering. [eg. Age - only years (discrete) or Years and Months (continuous) or Age Group (Interval)]
3	**Qualitative Variables**	Categorical or Finite variables [eg. Colous, Qualitative, Yes/No, True/False, Dice, Coin Toss, Cards etc.] **Ordinal:** Can be ranked or arranged [eg. Performance of students in a class, Qualifications] **Interval:** Grouped data [eg. Age range, Salary Range] **Nominal:** Describes the traits or values of a variable [eg. Gender, Hair Color, Departments, City, pin code etc.]	
4	**Scatter Diagram**	Graph between 2 quantitative variables, one dependent and one independent. It helps to identify the degree of correlation between variables visually The relationship can be positive, negative or no relationship	
5	**Independent Variable**	Variable that the experimenter manipulates or changes. They have a direct effect on the dependent variable. Also called Predictor Variable [Eg. House Area, No. of Rooms, Age, Experience etc.]	
6	**Depedent Variable**	Variable that is being measured or tested in an experiment. Also called Target or Predicted Variable [eg. House Price, Salary etc.]	
7	**Correlation Coefficient - (r)** **Pearson's Correlarion** **Product Moment Correlation**	Statistic showing the NATURE and STRENGTH between 2 quantitative variables (Excluding Target Variable) Sign of " r " denotes the Nature of association (" - " means Inversely Correlated, "+" means Directly Correlated) Value of " r " denotes the Strength of association (-1 to +1)	$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$
8	**Standardization**	Process of putting different variables on the same scale	
9	**Inferential Statistics**	Infer some truth / Information about the population based on what we know from sample data. [eg. Exit Polls] Limitations: Bias, Sample Size, Sampling Procedure, Response Rate (Incorrect or No Response)	
10	**Descriptive Statistics**	Organize and Summarize scores from samples. Preprocess data without attempting to draw any inference. There are 3 kinds: **Measure of Central Tendency:** Mean, Median, Mode **Measure of Dispersion:** SD, Variance, Range, IQR **Measure of Symmetricity/Shape:** Skewness and Kurtosis	
11	**Sample**	A smaller (but hopefully representative) collection of units from a population used to determine truth about that population	
12	**Normal Distribution**	It's a Bell Curve where 68% of the data lies between -σ to +σ and 95% of the data lies between -2σ to +2σ. Remaining 5% data are outliers	
13	**Standard Normal Distribution** **Z - Distribution** **Zero - Distribution** **Gaussian - Distribution**	A special normal distribution where the mean is 0 and the standard deviation is 1 Any normal distribution can be standardized by converting its values into z-scores. Z-scores tell you how many standard deviations from the mean each value lies. Bell curve is fatter and Variance is usually High Data Size > 30	μ = Population Mean = 0 σ = Standard Deviation = 1
14	**Z-Score**	Converts Normal Distribution to Standard Normal to make the data unitless and have a standard scale or range for all variables	Z - Score = $(\bar{x}-\mu) / (\sigma / \sqrt{n})$ --> **Z(0.025) = 1.96, Z(0.05) = 1.68** Z - Score = $(X - \mu) / \sigma$
15	**Significance Level (α)**	Threshold value used to judge whether a test statistic is statistically significant, usually 5%. Denoted by Alpha	
16	**Confidence Interval**	The actual data that is contributing to the mean or centre tendency, Usually remaining 95% of the data	CI = 1 - α
17	**Point Estimate (PE)**	Average or Mean of a sample dataset ($\bar{x}$)	$\bar{x}$ = Sample Mean = PE
18	**Margin of Error (MoE)**	The amount of random sampling error in the results of a survey.	MoE = $(Z_{\alpha/2} * \sigma) / \sqrt{n}$
19	**Interval Estimate (IE)**	A range of values where the actual data lies in	IE = PE ± MOE = $\bar{x} \pm (Z_{\alpha/2} * \sigma) / \sqrt{n}$

20	**Sample Size**	30 for most applications 50 for Skewed Samples or for Samples with Outliers 15 for Symmetric Samples that is not Normally Distributed	$n = (Z\alpha/2)^2 \sigma^2 / MoE^2$
21	**T - Distribution**	Bell Curve is Thinner and Variance is usually Low Dataset size is < 30 Depends on Degree of Freedom (n-1)	$IE = PE \pm MOE = \bar{x} \pm (t\alpha/2 * S) / \sqrt{n}$
22	**Degree of Freedom**	Sample Size - 1. If we keep increasing Degree of Freedom to more than 30, T - Distribution will start behaving like Z - Distribution	Degree of Freedom = n - 1
23	**Population Proportion (P)**	A parameter that describes a percentage value associated with a population.	$IE = P \pm Z\alpha/2 * \sqrt{[P(1-P)/n]}$ $Z = P-P_0 / \sigma_P$, Where $\sigma_P = \sqrt{[P_0(1-P_0) / n]}$ --> Hypothesis
24	**Hypothesis**	Hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected	In Hypothesis testing, α value is called P - Value
25	**Null Hypothesis (H_0)**	In Inferential Statistics, the Null Hypothesis is that 2 possibilities are the same	
26	**Alternateive Hypothesis (H_a)**	Opposite of Null Hypothesis	
27	**Type I Error**	When we are rejecting Null Hypothesis but it is true: - Wrong Hypothesis - α < 5% or Confidence Interval > 95%	
28	**Type II Error**	When we are Accepting Null Hypothesis but it is False - α > 10% or Confidence Interval < 90% - α is represented by β	
29	**R^2**	**Coefficient of Determination: The square of Correlation Coefficient. $R^2 = r*r$** The amount of variation in the outcome (dependent variable) that is explained by the predictor (independent variable) Total amount of variation explained by the independent variables which is understood by the target variable Since $-1 < r < 1$ Hence $0 < R^2 < 1$	$R^2 = SSB / (SSB + SSW)$
30	**Adjusted R^2**	- In a model, when the predictor variables are increased, this causes the R^2 value to increase, implying that increasing the count of variables leads to a good model, which is not true - Hence when the number of variables are increased, a better measure to evaluate the model's performance is Adjusted R^2 - In Adjusted R^2, the value of R^2 is adjusted for the no of predictor variables - Now when the variable count increases, the value of Adjusted R^2 will increase only when the added variables really adds to the model's explaination power	
31	**Feature Engineering**	Design features in such a way that the correlated variables do not harm the model. This can be done either by dropping one of the correlated variables or create a new variable that combines the features of 2 or more correlated variables	