Data Science Interview Question

1. What is data science?

Data science is a field that uses **scientific methods, algorithms, and systems to extract insights and knowledge** from data in various forms. It combines aspects of statistics, computer science, and domain expertise to **understand and analyze complex phenomena**.

2. Difference between Data Analytics and Data Science?

- Data analytics focuses on analyzing data to **derive insights and make informed decisions**, often using **statistical and visualization** techniques.

- Data science is a broader field that encompasses data analytics but also involves **building predictive models,** machine learning algorithms, and extracting deeper insights from data using **various tools and techniques**.

3.Name a few libraries in python used for data science?

Here are a few popular libraries in Python for data science:

- Pandas
- NumPy
- Matplotlib
- Scikit-learn

4.What are the important skills to have in python with regard to data analysis?

- **Programming Fundamentals:** Basic understanding of Python **syntax, data types, loops, and functions**.
- **Data Manipulation:** Proficiency in libraries like Pandas for **data manipulation and cleaning**.
- **Data Visualization:** Knowledge of libraries like **Matplotlib and Seaborn** for creating visualizations.
- **Statistical Analysis:** Understanding of statistical concepts and libraries like **NumPy and SciPy** for numerical computations.
- **Machine Learning:** Familiarity with libraries like **Scikit-learn for implementing** machine learning algorithms.
- **Problem-Solving:** Ability to **tackle real-world data analysis problems** with creative solutions.

5. What is the data type for qualitative and quantitative?

For qualitative data, the data type is **categorical**. For quantitative data, the data type is **numerical**.

6.What is the continuous and discrete value?

Continuous values are those that can take on any value within a certain range, like **height or weight**, where there are **infinite possibilities** between any two points. Discrete values, on the other hand, are specific, separate values with **no "in-between" possibilities,** like the number of students in a class or the outcome of flipping a coin.

7. What is Nominal & Ordinal?

Nominal refers to categories **without any inherent order, like colors or types of animals**. Ordinal implies a **ranking or order**, like **grades in school or levels of satisfaction**.

8. What is the difference between data mining, statistics, machine learning and AI?

- Data mining is about **finding patterns** in large datasets.
- Statistics is the study of **collecting, analyzing, and interpreting data** to make inferences and predictions.
- Machine learning is a **subset of AI** where systems learn from data to **improve their performance** on a task.
- AI (Artificial Intelligence) is a broad field focused on **creating systems** that can perform tasks that would typically require **human intelligence**.

9. What is the difference between descriptive statistics and inferential statistics?

Descriptive statistics describe and summarize data, giving you a **snapshot** of what happened. Inferential statistics help you **draw conclusions or make predictions** about a population based on a sample of data.

10.What is EDA?

EDA stands for **Exploratory Data Analysis**. It's a process in data science where you **visually explore and summarize** data sets to understand their main characteristics.

11. What are the key elements of an EDA report and how do they contribute to understanding a dataset?

An EDA (Exploratory Data Analysis) report typically includes:

- **Summary Statistics:** Basic stats like **mean, median, mode, and standard deviation** give an overview of the data's **central tendency and variability**.
- **Data Visualization:** Graphs like **histograms, box plots, and scatter plots** help visualize distributions, trends, and relationships within the data.
- **Data Quality Assessment:** Identification and handling of **missing values, outliers, and inconsistencies** ensure data integrity.
- **Feature Analysis:** Examination of individual features' distributions and relationships with the **target variable provides insights into predictive power**.
- **Correlation Analysis:** Assessing pairwise relationships between **variables** helps identify **potential multicollinearity** and informs feature selection.

These elements collectively enable analysts to gain insights, detect patterns, and make informed decisions about the dataset.

12.What is the normal distribution?

The normal distribution is a **bell-shaped curve** that describes the distribution of many naturally occurring phenomena. It's characterized by its **symmetric shape**, with the **mean, median, and mode** all being equal and located at the **center** of the curve.

13. What is the difference between the Bernoulli and binomial distribution?

The Bernoulli distribution describes the **probability of success or failure in a single trial,** like flipping a coin. Binomial distribution, on the other hand, deals **with the number of successes in a fixed number of independent** Bernoulli trials, like the number of heads in 10 coin flips.

14. What is eigenvalue and eigenvector?

Eigenvalues and eigenvectors are concepts from **linear algebra**. Eigenvalues represent the scalars that scale eigenvectors but **leave their direction unchanged when a linear transformation** is applied to a vector space. Eigenvectors are the vectors that, when operated on by a transformation, **only change in scale, not direction**.

15.What is parametric & non parametric?

Parametric methods make assumptions about the **underlying data distribution**, while non-parametric methods **do not make such assumptions and instead rely on the data itself** to determine the model's structure.

16.What is the central limit theorem?

The central limit theorem states that when you have a **large enough sample size**, the distribution of sample means will be approximately normally distributed, regardless of the **original distribution of the population,** as long as certain conditions are met.

17. How do you transform a skewed distribution into a normal distribution?

To transform a skewed distribution into a normal distribution, you can use techniques like **logarithmic transformations, square root transformations, or Box-Cox transformations**. These methods help to make the data **more symmetric and closer** to a normal distribution.

18. What is a probability distribution?

A probability distribution tells us how likely **different outcomes are in a given situation**. It shows the probabilities of **all possible outcomes** of a random variable.

19. What is the difference between Binomial and geometric distribution?

The binomial distribution counts the **number of successes in a fixed number of trials with a constant probability** of success, while the geometric distribution models the **number of trials needed to achieve the first success in a series of independent trials with the same probability** of success.

20. What is the cumulative distribution function?

The cumulative distribution function (CDF) gives the probability that a random variable takes on a **value less than or equal to a given number**.

21. Difference between univariate, bivariate and multivariate analysis?

- **Univariate analysis:** Looks at **one variable at a time** to understand its **characteristics or behavior**.
- **Bivariate analysis:** Examines the relationship between **two variables** to see how they **interact or influence** each other.
- **Multivariate analysis:** Considers **multiple variables** simultaneously to **explore complex** relationships among them.

22. How to perform univariate analysis for numerical and categorical variables?

-

For numerical variables:

- **Mean, Median, Mode:** Calculate these measures to understand **central tendency**.
- **Range and Interquartile Range (IQR):** Determine the **spread** of data.
- **Histograms and Boxplots:** Visualize the **distribution and identify outliers**.

For categorical variables:

- **Frequency Tables:** **Count** occurrences of each category.
- **Bar Charts:** Display the **frequency** of each category.
- **Pie Charts:** Show the **proportion** of each category in the whole.

23. What is the difference between probability mass function and density probability function

The main difference is the type of random variable they apply to. Probability mass function (PMF) is for **discrete random variables,** while probability density function (PDF) is for **continuous random variables**. PMF gives the **probability of each possible outcome**, while PDF gives the relative **likelihood of different values** occurring **within a range**.

24. what is overfitting and underfitting?how will you identify and tackle this in your model

- Overfitting happens when a **model** learns the training data **too well, including noise,** so it **performs poorly** on new data.
- Underfitting occurs when a **model is too simple** to capture the underlying structure of the data, resulting in **poor performance** even on the training data.
- To identify overfitting and underfitting, we can use techniques like **cross-validation**, where we split the data into **training and validation** sets.
- If the model performs **well on the training** set but **poorly on the validation set,** it might be overfitting. If it **performs poorly on both**, it might be underfitting.
- To tackle overfitting, we can use methods like **regularization (e.g., L1, L2), reducing** model complexity, or increasing training data.
- For underfitting, we might try **increasing** model complexity, **adding more features**, or using a more complex model.

25. What are some of the techniques to avoid overfitting and underfitting?

To avoid overfitting:

- **Cross-validation:** Split data into **training and validation** sets.
- **Regularization:** Penalize **complex models with extra terms** in the cost function.
- **Dropout:** Randomly **deactivate neurons** during training to prevent reliance on specific ones.

To avoid underfitting:

- **Increase model complexity:** Add more layers or neurons to capture **underlying patterns**.
- **Feature engineering:** **Select or create** more relevant features.
- **Decrease regularization:** **Relax constraints** on the model to allow for more complexity.

26. What is the difference between feature selection Vs feature engineering

Feature selection involves **choosing a subset** of relevant features from the **original dataset.** It's like picking the best ingredients for a recipe. Feature engineering, on the other hand, involves **transforming or creating** new features to **improve model performance.** It's akin to modifying the ingredients or even inventing new ones to enhance the dish's flavor.

27. Difference between feature selection and feature extraction?

Feature selection is about choosing a subset of the original features from the dataset, while feature extraction involves **transforming the original features into a new set of features**.

28. what is preprocessing? what are preprocessing steps do you know?

Preprocessing is the initial stage of data preparation where raw data is **transformed into a clean,** organized format suitable for **analysis or modeling**. Preprocessing steps include:

- **Data Cleaning:** Removing or correcting errors, **inconsistencies, and missing values** in the data.
- **Data Transformation:** Standardizing, normalizing, or encoding data to make it **suitable for analysis**.
- **Data Reduction:** Reducing the dimensionality of the data through techniques like **feature selection or extraction**.
- **Data Integration:** Combining data from **different sources into a unified dataset**.
- **Data Discretization:** **Converting continuous data** into discrete bins or categories.
- **Data Sampling:** Selecting a **representative subset** of the data for analysis.
- **Data Scaling:** Rescaling data to ensure all features have a **similar scale for accurate analysis**.

These steps ensure that the data is of high quality and ready for further analysis or modeling.

29. Explain Training Dataset, testset and validation dataset

- **Training Dataset:** It's like a **teacher's lesson plan**. It's the data used to teach a machine learning model how to make **predictions or classifications**.

- **Test Dataset:** This is like a **quiz for the model**. It's data that the model **hasn't seen before,** used to evaluate how well it learned **during training**.

- **Validation Dataset:** Think of this as **practice before the big test**. It's another set of data used to **fine-tune the model's performance** during training, ensuring it generalizes well to new, unseen data.

30. What is Bayes' theorem and how is it used in data science?

- Bayes' theorem is a **statistical principle** that calculates the probability of an event based on **prior knowledge or conditions** related to that event.

- In data science, it's used to update beliefs about the likelihood of an event occurring when **new evidence** is obtained. It's particularly handy for tasks like **classification**, where it helps refine predictions based on observed **data and prior probabilities**.

31. What are the measures of central tendency?

The measures of central tendency are ways to find the average or typical value in a set of data. They include **mean (average), median (middle value), and mode (most common value).**

32. What will be the case in which the Mean, Median, and Mode will be the same for the dataset?

When every value in the dataset is **exactly the same, the mean, median, and mode will all be the same.** When the mean, median, and mode are all the same, we call it a **"symmetric distribution" or "symmetric data."**

33. What is the difference between k-means and k-medians and when would you use one over another?

- K-means and k-medians are both **clustering algorithms**, but they differ in how they calculate **cluster centers and assign data points**.

- K-means calculates cluster centers by **minimizing the sum of squared** distances between **data points and cluster centroids,** while k-medians calculates cluster centers by **minimizing the sum of absolute deviations**.

- Use k-means when your data is normally distributed and you want **clusters with spherical shapes**. Use k-medians when your data contains **outliers or has a skewed distribution,** as it's more robust to such scenarios.

34. When do we use mean?

We use the mean when we want to **find the average value** of a set of numbers.

35. When do we use median?

We use the median when we want to find the **middle value** in a dataset. It's handy when there are extreme values that could **skew the average (mean)**.

36. What is percentile?

A percentile is a **statistical measure** indicating the percentage of data points that **fall below a certain value** in a dataset. For example, if you scored in the **80th percentile** on a test, it means you scored better than **80%** of the people who took the test.

37. What is IQR?

IQR stands for **Interquartile Range**. It's a measure of statistical dispersion, showing the range within which the **middle 50%** of data points fall in a dataset. It's calculated as the difference between the **third quartile (Q3) and the first quartile (Q1)**.

38. When would you use the interquartile range?

The interquartile range is used to understand the spread of data by **focusing on the middle 50%.** You'd use it when you want to analyze the **variability** within a dataset while **ignoring outliers, like in assessing income disparities or test scores**

39.Does Standard deviation get influenced by Outliers?

Yes, outliers can influence the standard deviation.

40. How to handle outliers?

Handling outliers involves several steps:

- **Identify outliers:** Use statistical methods like **box plots, Z-score, or interquartile range (IQR)** to detect outliers in the dataset.
- **Understand the cause:** **Investigate the reason behind** outliers. They could be **genuine data points or errors** in measurement.
- **Choose an approach:** Depending on the situation, decide whether to **remove, transform, or keep the outliers**. Sometimes, outliers contain v**aluable information and shouldn't be discarded** outright.

Handling outliers:

- **Remove outliers:** Exclude them from analysis if they're due to **measurement errors or if they significantly skew** the results.
- **Transform data:** Apply transformations like log transformation to **reduce the impact of outliers**.
- **Treat separately:** Analyze outliers separately if they represent a **distinct subgroup** in the data.
- **Evaluate impact:** Assess how handling outliers affects the **overall analysis and results**. Be transparent about the methods used and their implications.

41. What are the different ways in which we can find outliers in the data?

There are mainly two ways to find outliers in data:

- **Statistical methods:** Using mathematical techniques like **z-score, interquartile range (IQR), or standard deviation** to identify values that significantly differ from the rest.

- **Machine learning approaches:** Employing algorithms such as **isolation forest or k-nearest neighbors (KNN)** to detect anomalies based on patterns in the data.

42. What is frequency? Why it is used? how we can use frequency

- Frequency refers to the rate at which something occurs or **repeats over a specific period**, typically measured in **hertz (Hz)**. It is used to describe how often a phenomenon happens within a **given timeframe**.

- In various fields such as **physics, engineering, and signal processing,** frequency helps understand and analyze **patterns, cycles, and vibrations**.

- We can use frequency to **tune musical instruments, transmit radio signals, analyze sound waves, and determine the behavior of waves** in physics, among other applications.

43. What is the difference between a box plot and a histogram?

A box plot summarizes **data distribution, showing median, quartiles, and outliers,** while a histogram displays the **frequency or count of data** within intervals.

44. How can you define Skewness?

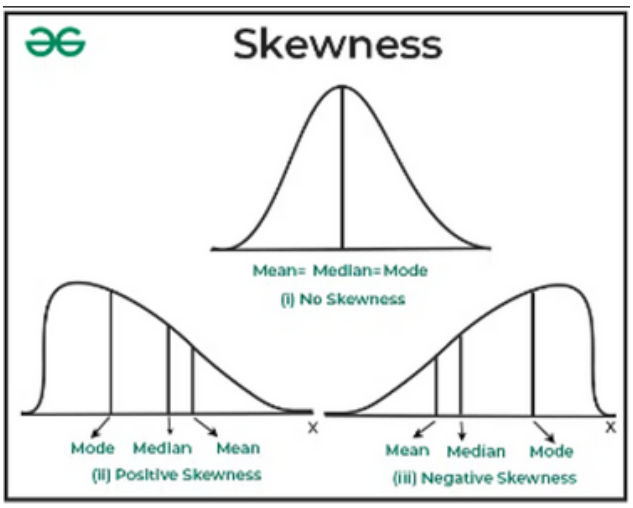
Skewness measures the **asymmetry of a probability distribution**. If the distribution leans towards the **left, it's negatively skewed**; if it leans towards the **right, it's positively skewed.**

45. Type of skewness , explain mean,median,mode in graph?

Skewness refers to the asymmetry in a distribution.

- If the distribution is **negatively skewed,** the **mean < median < mode**. This means the tail is **longer on the left side** of the distribution.

- If the distribution is **positively skewed**, the **mean > median > mode**. This means the tail is **longer on the right side** of the distribution.



46. What do you understand about Kurtosis?

Kurtosis measures how much a distribution's tails **differ from a normal distribution**. High kurtosis means heavy tails, while **low kurtosis indicates light tails.**

47. What does Right Skewed mean?

Right skewed means that the **tail of the distribution in a graph** leans towards the **right side**.

48. Left (or) Negative Skewed data mean?

Left (or Negative) skewed data means that the tail of the distribution **stretches towards the lower values**, indicating that the majority of the data points are concentrated on the **higher side of the distribution with fewer values on the lower side**.

49. Difference between kurtosis and skewness?

Kurtosis measures how **peaked or flat** a distribution is, while skewness measures the **asymmetry** of the distribution.

50. What is variance?

Variance measures how much data points **deviate from the mean value** in a dataset. It quantifies the **spread or dispersion** of the data points around the mean.

51. What is standard deviation

Standard deviation is a measure of how **spread out or dispersed the values in a dataset are from the mean (average)**. It tells us how much individual data points deviate from the average of the dataset.

52. Compare Standard Deviation and Variance?

Standard deviation and variance are both **measures of dispersion** in a dataset. While variance gives the **average of the squared differences from the mean, standard deviation is the square root of variance.** So, standard deviation is in the same units as the original data, making it **easier to interpret**, while variance is in squared units.

53. What are the disadvantages of Variance?

The main disadvantages of variance are:

- It's **sensitive to outliers,** meaning extreme values can **heavily influence** it.
- It's **not in the same units** as the original data, making **interpretation difficult**.
- Squaring the differences can **amplify the effect of errors or discrepancies** in the data.

54. Can the measure of dispersion 'Standard deviation' be negative?

No, the standard deviation **cannot be** negative.

55. Difference between pdf and cdf?

PDF stands for **Probability Density Function**, while CDF stands for **Cumulative Distribution Function**.

The main difference between them lies in their function and purpose.

- **PDF:** It gives the **probability density** at a specific point in a **probability distribution.** It represents the likelihood of a random variable taking on a particular value. **PDF doesn't** provide **cumulative probabilities**.

- **CDF:** It gives the **cumulative probability** that a random variable takes on a **value less than or equal** to a specific point. In other words, it gives the probability of a random variable being less than or equal to a certain value. CDF **accumulates probabilities** as it progresses through the distribution.

56. What is the standard normal distribution and why is it used?

The standard normal distribution is a **symmetric bell-shaped curve with a mean of 0 and a standard deviation of 1**. It's used because it simplifies statistical calculations and helps analyze data by providing a common baseline for comparison.

57. What is z-score?

A z-score is a measure that tells you how many **standard deviations a data point is from the mean** of a dataset.

58. Difference between covariance and correlation?

Covariance measures how **two variables change together**, while correlation **measures the strength and direction** of their relationship on a standardized scale **(-1 to 1).**

59. Can you explain the concept of correlation and covariance?

Correlation measures how **two variables change together**, either **positively** (both increase or decrease) or **negatively** (one increases **while the other decreases**). Covariance is a measure of how **two variables vary together**; it's similar to correlation but **doesn't have standardized units** like correlation does.

60. What is Multicollinearity?

Multicollinearity is when **two or more independent variables** in a regression model are **highly correlated** with each other, which can cause issues in interpreting the **coefficients and reliability** of the model's predictions.

61. Define Range. How to evaluate it?

Range is the difference between the **highest and lowest values** in a dataset. To evaluate it, **subtract** the lowest value from the highest value in the dataset.

62. What is vif why it is used and How do you calculate it?

VIF stands for **Variance Inflation Factor**. It's used to detect multicollinearity in regression analysis, where predictor variables are **highly correlated**. It measures how much the variance of an estimated **regression coefficient is increased** because of **collinearity**.

To calculate VIF:

- **Fit a regression model** for each predictor variable against all other predictor variables.
- Calculate the **R-squared value** for each regression.
- For each predictor variable, calculate the VIF using the formula:

  VIF = 1 / (1 - R-squared)
- **Higher VIF** values indicate **stronger** multicollinearity. Typically, a VIF above **5 or 10** indicates a problematic level of multicollinearity.

63.What is t-test?

A t-test in data science is a statistical method used to determine if there is a **significant difference between the means of two groups**. It assesses whether the means are statistically different from each other or if the differences observed are due to **random chance.**

64. How does an ANOVA test work and its types?

ANOVA, or Analysis of Variance, tests if there are significant differences between means of **three or more groups.** It works by comparing variation within groups to variation between groups. There are three types: **one-way ANOVA for one independent variable, two-way ANOVA for two independent variables, and MANOVA for multiple dependent variables.**

65. What is the significance of p-value?

The p-value tells us the probability of observing results as extreme as the **ones in our data,** assuming the **null hypothesis is true**. It helps us decide whether the results are statistically significant or just due to chance.

66. Why is hypothesis testing useful for a data scientist?

Hypothesis testing helps data scientists make **informed decisions** by allowing them to assess if there's a significant relationship between **variables in their data**, helping **validate or reject assumptions** about the population based on sample data.

67. State null and alternate hypothesis

- **Null hypothesis:** There is **no** significant difference or effect.
- **Alternate hypothesis:** There **is a significant** difference or effect.

68. How would you test hypotheses using Bayes' Rule?

To test hypotheses using Bayes' Rule, you'd start with **prior beliefs** about the hypotheses, **gather new data**, update those beliefs using Bayes' Rule, and calculate the **posterior probabilities** for each hypothesis based on the **new evidence**.

69. What is the purpose of the chi-square test in statistics?

The purpose of the chi-square test in statistics is to determine if there's a significant difference between **observed and expected frequencies in categorical data**.

70. Mention some techniques used for sampling.

Here are some sampling techniques:

- Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling
- Convenience Sampling

71. What is systematic sampling?

Systematic sampling is a method where every **nth item** in a population is selected after **randomly choosing a starting point**, ensuring an even distribution and representing the whole population fairly.

72. What is cluster sampling?

Cluster sampling is a method where the population is divided into **groups or clusters**, and then a random sample of these clusters is chosen for **analysis**. Each cluster ideally represents the **diversity** of the entire population, making it **efficient for large and diverse** populations.

73. Why is resampling done?

Resampling is done to **improve the reliability and accuracy** of statistical estimates by generating **new samples or redistributing existing data**.

74. What is variance in Data Science?

Variance in data science refers to how much the values in a dataset **deviate or spread out** from the **mean (average)**. It measures the **variability or dispersion** of data points around the average.

75. What is variance error?

Variance error is a measure of how much **predictions of a model differ from the actual values,** indicating the model's sensitivity to fluctuations in the training data.

76. What do you do if you have a high bias problem?

If you have a high bias problem, you need to **address it** by **increasing the complexity** of your model or **gathering more diverse data to better** represent the underlying patterns in the data.

77. What is the difference between variance and bias in Statistics?

Variance is the **variability of model predictions** from one another, **reflecting sensitivity to changes** in the training data. Bias is the **error** due to overly simplistic assumptions in the learning algorithm, leading to **inaccurate predictions**.

78. In statistics , what is the difference between bias and error?

Bias refers to the **systematic error** in a measurement process that causes it to **consistently deviate from the true value**. Error, on the other hand, encompasses all deviations of measurements from the true value, including both **systematic biases and random fluctuations**.

79. Define bias-variance trade-off

The bias-variance trade-off is a fundamental concept in machine learning. It refers to the balance between **errors caused by underfitting (high bias) and overfitting (high variance)** when building predictive models. Essentially, **reducing bias** typically **increases variance and vice versa,** so finding the optimal trade-off is crucial for model performance.

80.Difference between expected and mean value

The mean value is the **average of a set of values**, while the expected value is what we anticipate to happen based on **probabilities or predictions.**

81. How do you decide whether a model is suffering from high bias or high variance?

High bias and high variance in a model can be identified through performance evaluation:

High Bias:

- **Symptom:** **Poor performance** on both training and testing data.
- **Indication:** The model is **too simple** to capture the underlying patterns in the data.
- **Solution:** **Increase** model **complexity**, add more features, or use a **more sophisticated algorithm**.

High Variance:

- **Symptom:** **Excellent performance** on **training data** but **poor performance** on **testing data**.
- **Indication:** The model is **overly sensitive to small fluctuations** in the training data and fails to generalize.
- **Solution:** Simplify the model, **reduce** the number of features, **increase training data,** or use **regularization** techniques.

By analyzing performance metrics like training and testing error, bias and variance issues can be diagnosed, guiding adjustments to improve model performance.

82. What are some techniques for balancing bias and variance in a model?

To balance bias and variance in a model, you can:

- **Regularization:** Penalize **complex** models to **reduce** variance.
- **Cross-validation:** Assess model performance on **different data splits** to understand **bias-variance** trade-off.
- **Feature selection:** Choose relevant features to **reduce model complexity and variance**.
- **Ensemble methods:** Combine multiple models to **average out biases and reduce variance**.
- **Bias-variance decomposition:** Analyze sources of **error to understand** where adjustments are needed.

83. What is Expected value?

Expected value is the **average outcome** we anticipate from a probability distribution, calculated by **multiplying each possible outcome** by its probability and summing them up. It helps **predict long-term results** in uncertain situations.

84.What is the difference between Type I and Type II errors?

Type I error is when you **incorrectly reject** a **true** null hypothesis (false positive), while Type II error is when you **fail to reject** a **false** null hypothesis (false negative).

85. What is confidence interval and confidence level?

A confidence interval is a **range of values** that we are **fairly certain includes the true value** of a population parameter. The confidence level is the **probability** that the **interval actually does contain** the population parameter.

86. How can less training data give higher accuracy?

In data science, less training data can sometimes lead to higher accuracy through techniques like **transfer learning,** where knowledge gained from one task is applied to another, or by employing **more sophisticated algorithms** that can **effectively generalize from limited data**.

87. What are the differences between supervised and unsupervised learning?

Supervised learning uses **labeled** data to train algorithms to make **predictions or decisions**, while unsupervised learning deals with **unlabeled** data to discover **patterns or structures** within the data without explicit guidance.

88.What is the difference between upsampling and downsampling?

Upsampling **increases** the resolution or size of an image or signal, while downsampling **decreases** it.

89. What is data leakage and how can we identify it?

- Data leakage refers to the unauthorized or unintended **transmission of sensitive data outside** of a secure environment.
- It can happen through various means like **human error, malicious intent, or technical vulnerabilities**.
- To identify data leakage, organizations can use tools like **data loss prevention (DLP) software, encryption methods, access controls, and regular monitoring of network traffic and data access logs**.

90. What is the difference between Manhattan Distance and Euclidean distance?

Manhattan distance measures the distance between two points in a **grid** based on the **sum of the absolute** differences of their **coordinates**. Euclidean distance calculates the **straight-line** distance between two points **in space**.

91. What is the use of the violin plot?
A violin plot is used to visualize the **distribution of data and its probability density**.

92.What are the five statistical measures represented in a box plot?

The five statistical measures represented in a box plot are:

- Minimum
- First Quartile (Q1)
- Median (Second Quartile, Q2)
- Third Quartile (Q3)
- Maximum

93. In data science how will you analyse categorical column

- To analyze a categorical column in data science, you typically use techniques like **frequency counts, distribution plots, or statistical tests such as chi-square test**.

- These methods help understand the distribution and relationships within categorical data, which is **crucial** for making informed decisions in d**ata analysis and modeling**.

94. What are the graphs used to analyse numerical column?

The graphs used to analyze numerical columns are typically **histograms and box plots**.

95. What is collinearity and multicollinearity?

- Collinearity and multicollinearity are both about how variables in a statistical model relate to each other.

- Collinearity is when **two or more predictors** in a regression model are **highly correlated**.

- Multicollinearity is a more serious form of collinearity, where **three or more variables** are highly correlated, making it **hard to separate** their individual effects on the dependent variable.

96.How would you screen for outliers and what should you do if you find one?

To screen for outliers, I'd use statistical methods like **box plots, z-scores, or IQR.** If I find one, I'd first **double-check data accuracy**. Then, decide whether to **remove, transform, or analyze separately**, depending on its impact and context in the dataset.

97. What is A/B testing?

A/B testing is a method used to compare two versions of something **(like a webpage or app)** to see which one performs **better**. You divide your audience into **two groups** and show each group a **different version**, then analyze which version gets better results, like **more clicks or purchases**.

98. Which library would you prefer for plotting in python language : seaborn or matplot or bokeh

For **simplicity**, I'd choose **Matplotlib**.

99. Suppose you found that your model is suffering from low bias and high variance? How you tackle it?

To tackle low bias and high variance in a model, you can:

Reduce Variance:

- Gather **more data** if possible.
- Simplify the model by **reducing** its complexity.
- Regularize the model by **adding penalties** to prevent overfitting.

Increase Bias:

- Use **more features** or improve feature engineering.
- Choose a **more complex model** if underfitting persists.

By finding a balance between bias and variance, you can improve your model's performance.

100. What is the difference between "long" and "wide" format data?

In **long format data**, each row represents a single observation with **multiple variables**, while in **wide format data,** each row represents a single variable with **multiple observations**.

101. How does data cleaning plays a vital role in analysis?

Data cleaning is crucial in analysis because it ensures that the data is **accurate, reliable, and consistent**. Without clean data, analysis results can be **skewed or inaccurate, leading to flawed** insights and decisions.

102.What is the difference between the standard error of the mean and standard deviation?

The standard deviation measures the **spread of individual data points from the mean,** while the standard error of the mean tells us how much the sample mean is likely to **vary from the true population mean**.

103.How do you decide the size of your validation and test sets?

When deciding the size of validation and test sets, I typically aim for a balance between having enough data to accurately **evaluate performance and ensuring** that there's still **sufficient data** for training. It's important to avoid overfitting by not allocating **too much data to validation and test sets,** while also ensuring they're representative of the overall dataset.

104. What are various steps involved in an analytics project?

- **Define objectives:** **Clearly outline** what the project aims to achieve.
- **Data collection:** Gather relevant data from **various sources**.
- **Data preparation:** **Clean, organize, and preprocess** the data for analysis.
- **Data analysis:** Apply **statistical methods and algorithms** to extract insights.
- **Interpretation:** Analyze the **results** and draw meaningful **conclusions**.
- **Visualization:** Present findings through **charts, graphs, or other visual aids**.
- **Implementation:** **Incorporate insights** into decision-making or action plans.
- **Monitoring:** Continuously assess **performance and adjust strategies** as needed.

105. What is vector?

A vector is a **quantity** that has both **magnitude and direction**.

106. What is the difference between normalization and scaling?

Normalization and scaling are both techniques used in **data preprocessing**, but they serve different purposes:

Normalization:

- Normalization rescales the features to a range between **0 and 1, or sometimes -1 and 1**.
- It's useful when the features have **varying scales** and need to be on a **similar scale for algorithms** that rely on distance measures.
- Common normalization techniques include **Min-Max scaling and Z-score normalization**.

Scaling:

- Scaling, on the other hand, adjusts the range of the features **without changing their distribution**.
- It's used to **standardize** the range of independent variables in the dataset.
- Scaling doesn't necessarily bound the data to a specific range; it just ensures that the **mean** of the data is **0** and the **standard deviation** is **1**.
- Common scaling techniques include **StandardScaler and RobustScaler**.

In essence, normalization adjusts the values of features to a target range, while scaling standardizes the range of features without necessarily bounding them to a specific range.

107. Difference between normalisation and standardisation

Normalization and standardization are both techniques used in **data preprocessing,** but they serve different purposes:

- **Normalization** rescales the values of a feature to a range between **0 and 1,** making them comparable, which is useful for algorithms that rely on measures of **distance or similarity**.

- **Standardization**, on the other hand, rescales the distribution of values to have a **mean of 0 and a standard deviation of 1**. This makes the features have comparable scales, which is beneficial for algorithms that assume **Gaussian distribution** or require features to be **centered at 0**.

In simpler terms, normalization **adjusts values** to fit within a specific range, while standardization **adjusts the distribution** of values to a common scale.

108.Describe the importance of data preprocessing in AI.

- Data preprocessing is crucial in AI because it **cleans, organizes, and transforms** raw data into a format that machine learning algorithms can **understand and work with effectively**.

- It helps in improving the **quality of data, removing noise, handling missing values,** and standardizing features, thus enhancing the accuracy and reliability of AI models.

109. How to handle nan and null values

To handle NaN (Not a Number) and null values, you can:

- **Identify:** Recognize where **NaN or null values exist** in your dataset.
- **Remove or Replace:** Decide whether to remove those **entries** or replace them with **appropriate values**.
- **Imputation:** If replacing, use methods like **mean, median, mode,** or predictive modeling to **fill missing values**.
- **Data Types:** Ensure your data types are **consistent and appropriate** for handling missing values.
- **Documentation:** Document your approach thoroughly for **transparency and reproducibility**.

Data Science Interview Questions

1. What is data science?
2. Difference between Data Analytics and Data Science?
3.Name a few libraries in python used for data science?
4.What are the important skills to have in python with regard to data analysis?
5. What is the data type for qualitative and quantitative?
6.What is the continuous and discrete value?
7. What is Nominal & Ordinal?
8. What is the difference between data mining, statistics, machine learning and AI?
9. What is the difference between descriptive statistics and inferential statistics?
10.What is EDA?
11. What are the key elements of an EDA report and how do they contribute to understanding a dataset?
12.What is the normal distribution?
13. What is the difference between the Bernoulli and binomial distribution?
14. What is eigenvalue and eigenvector?
15.What is parametric & non parametric?
16.What is the central limit theorem?
17. How do you transform a skewed distribution into a normal distribution
18. What is a probability distribution?
19. What is the difference between Binomial and geometric distribution
20. What is the cumulative distribution function?
21. Difference between univariate, bivariate and multivariate analysis?
22. How to perform univariate analysis for numerical and categorical variables?

23. What is the difference between probability mass function and density probability function?
24. what is overfitting and underfitting?How will you identify and tackle this in your model?
25. What are some of the techniques to avoid overfitting and underfitting?
26. what is the difference between feature selection Vs feature engineering?
27. Difference between feature selection and feature extraction?
28. what is preprocessing? what are preprocessing steps do you know?
29. Explain Training Dataset, test set and validation dataset
30. What is Bayes' theorem and how is it used in data science?
31. What are the measures of central tendency?
32. What will be the case in which the Mean, Median, and Mode will be the same for the dataset?
33. What is the difference between k-means and k-medians and when would you use one over another?
34. When do we use mean?
35. When do we use median?
36. What is percentile?
37. What is IQR?
38. When would you use the interquartile range
39.Does Standard deviation get influenced by Outliers?
40. How to handle outliers?
41. What are the different ways in which we can find outliers in the data?
42. What is frequency? Why it is used? how we can use frequency
43. What is the difference between a box plot and a histogram?
44. How can you define Skewness?
45. Type of skewness , explain mean,median,mode in graph?

46. What do you understand about Kurtosis?
47. What does Right Skewed mean?
48. Left (or) Negative Skewed data mean?
49. Difference between kurtosis and skewness?
50. What is variance?
51. What is standard deviation
52. Compare Standard Deviation and Variance?
53. What are the disadvantages of Variance?
54. Can the measure of dispersion 'Standard deviation' be negative?
55. Difference between pdf and cdf?
56. What is the standard normal distribution and why is it used?
57. What is z-score?
58. Difference between covariance and correlation?
59. Can you explain the concept of correlation and covariance?
60. What is Multicollinearity?
61. Define Range. How to evaluate it?
62. What is vif why it is used and How do you calculate it?
63.What is t-test?
64. How does an ANOVA test work and its types?
65. What is the significance of p-value?
66. Why is hypothesis testing useful for a data scientist?
67. State null and alternate hypothesis
68. How would you test hypotheses using Bayes' Rule?
69. What is the purpose of the chi-square test in statistics?
70. Mention some techniques used for sampling.
71. What is systematic sampling?
72. What is cluster sampling?
73. Why is resampling done?
74. What is variance in Data Science?
75. What is variance error?
76. What do you do if you have a high bias problem?

77. What is the difference between Variance and Bias in Statistics?
78. In statistics , what is the difference between bias and error?
79. Define bias-variance trade-off
80.Difference between expected and mean value
81. How do you decide whether a model is suffering from high bias or high variance?
82. What are some techniques for balancing bias and variance in a model?
83. What is Expected value?
84.What is the difference between Type I and Type II errors?
85. What is confidence interval and confidence level
86. How can less training data give higher accuracy?
87. What are the differences between supervised and unsupervised learning?
88.What is the difference between upsampling and downsampling?
89. What is data leakage and how can we identify it?
90. What is the difference between Manhattan Distance and Euclidean distance?
91. What is the use of the violin plot?
92.What are the five statistical measures represented in a box plot?
93. In data science how will you analyse categorical column
94. What are the graphs used to analyse numerical column?
95. What is collinearity and multicollinearity?
96.How would you screen for outliers and what should you do if you find one?
97. What is A/B testing?
98. Which library would you prefer for plotting in python language? 99. Suppose you found that your model is suffering from low bias and high variance? How you tackle it?

100. What is the difference between “long” and “wide” format data
101. How does data cleaning plays a vital role in analysis?
102.What is the difference between the standard error of the mean and standard deviation?
103.How do you decide the size of your validation and test sets?
104. What are various steps involved in an analytics project?
105. What is vector?
106. What is the difference between normalization and scaling?
107. Difference between normalisation and standardisation
108.Describe the importance of data preprocessing in AI.
109. How to handle nan and null values?