**Data Sceince Interview Questions**

1. When to use mean, median and mode

These are all measures of central tendency.

Mean:

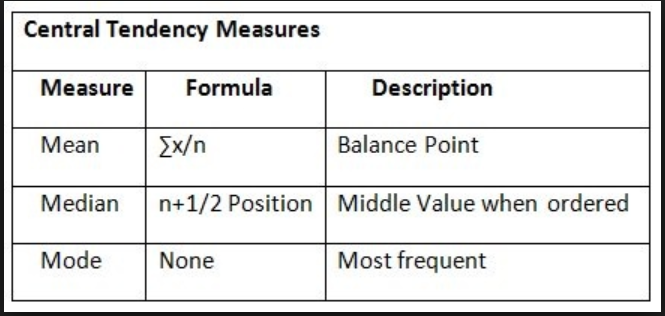
* For symetrical data.
* Gives equal weitage to each value.

Median:

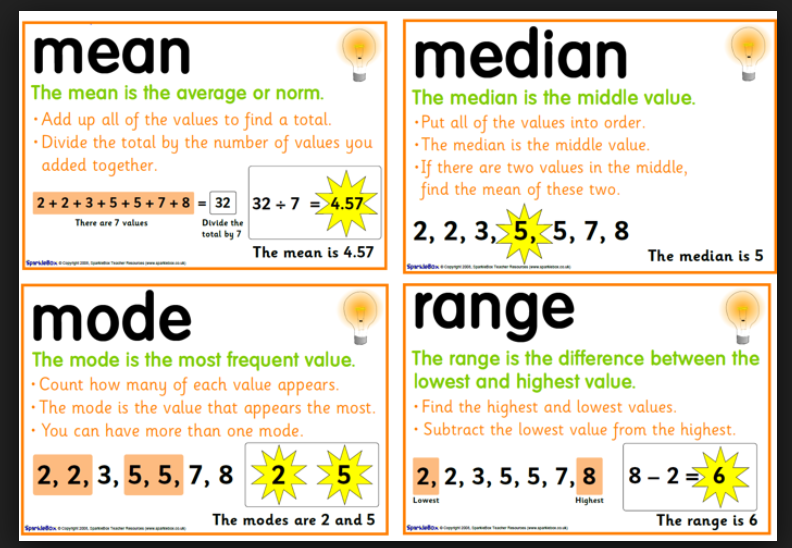
* For asymetircal data
* Gives less weitage to outliers

Mode:

* To find out the most common/frequent value



For Example:

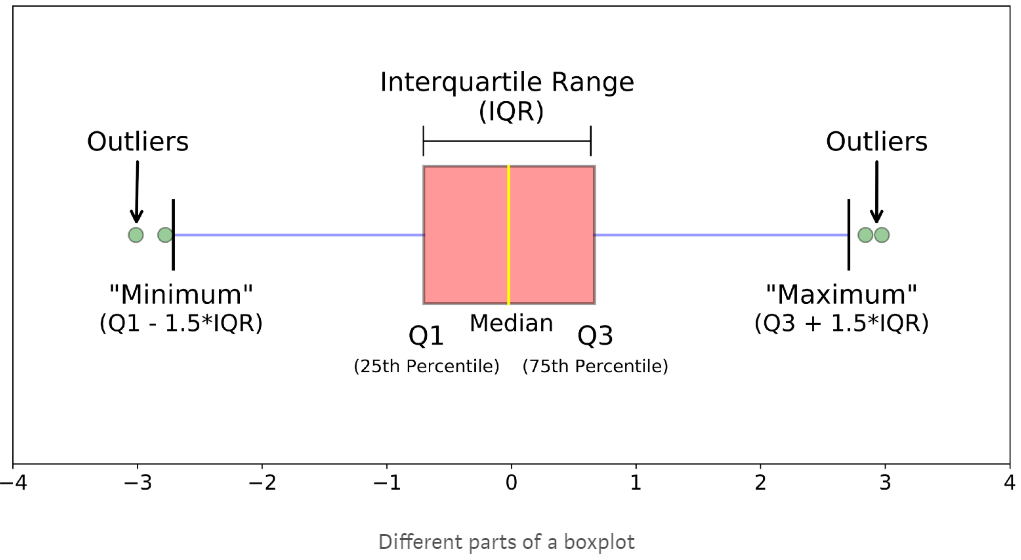


1. Difference between percentage and percentile

* A percentage is simply a representation of a proportion out of 100. To say three out of every six is the same as to say fifty out of every hundred, in other words fifty percent (from the Latin per centum).
* A percentile is a statistical measure of distribution. For a given set of data, it is the level below which a certain percentage of the data falls. In graphic terms, it represents the area under the curve of a distribution.
* For a more human example, if you score in the 72nd percentile on an exam, it means you scored higher than 72 percent of all the people who took the same exam— regardless of what your actual score was. If your baby is in the 28th percentile for length, it means she is longer than 28 percent of all comparable babies (e.g. babies at a certain age).
* Related terms include decile, quintile, quartile, tertile, if you divide up your population by 10, 5, 4, or 3 instead of 100.

1. Understand and explain boxplot

* A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.



Boxplots are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

**median (Q2/50th Percentile)**: the middle value of the dataset.

**first quartile (Q1/25th Percentile)**: the middle number between the smallest number (not the “minimum”) and the median of the dataset.

**third quartile (Q3/75th Percentile)**: the middle value between the median and the highest value (not the “maximum”) of the dataset.

**interquartile range (IQR)**: 25th to the 75th percentile.

**whiskers (shown in blue)**

**outliers (shown as green circles)**

**“maximum”**: Q3 + 1.5\*IQR

**“minimum”**: Q1 -1.5\*IQR

1. Find out what is five number summary

* A five number summary consists of these five statistics: the minimum value, the first quartile, the median, the third quartile, and the maximum value of a set of numbers (data).

1. Learn about central limit theorem and central tendency

* Central limit theorem: The mean of the sample population will be approximately equal to the mean of the whole population.
* Central tendency: The center or the middle location of the distribution is the average value of a probability distribution.
* Difference: Central limit theorem calculates the mean or average while the central tendency gives the central value.

1. Learn about variance and standard deviation.

Standard Deviation:

* The Standard Deviation is a measure of how spread out numbers are.
* Its symbol is σ (the greek letter sigma)
* The formula is easy: it is the square root of the Variance. So now you ask, "What is the Variance?"

Variance:

The average of the squared differences from the Mean.

To calculate the variance follow these steps:

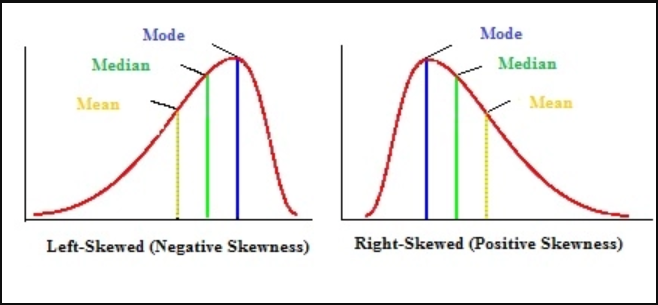
* Work out the [Mean](https://www.mathsisfun.com/mean.html) (the simple average of the numbers)
* Then for each number: subtract the Mean and square the result (the squared difference).
* Then work out the average of those squared differences. ([Why Square?](https://www.mathsisfun.com/data/standard-deviation.html#WhySquare) – to avoid positive num ers cancelling the negatives)

1. Understand what is sample and population data.

your sample is the group of individuals who participate in your study, and your population is the broader group of people to whom your results will apply. As an analogy, you can think of your sample as an aquarium and your population as the ocean. Your sample is small portion of a vaster ocean that you are attempting to understand.

1. Different types of distribution like uniform, normal, positive skewed and negative skewed data.

* In [probability theory](https://en.wikipedia.org/wiki/Probability_theory) and [statistics](https://en.wikipedia.org/wiki/Statistics), the continuous uniform distribution or rectangular distribution is a family of [symmetric](https://en.wikipedia.org/wiki/Symmetric_distribution) [probability distributions](https://en.wikipedia.org/wiki/Probability_distributions) such that for each member of the family, all [intervals](https://en.wikipedia.org/wiki/Interval_(mathematics)) of the same length on the distribution's support are equally probable.
* The normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
* If one tail is longer than another, the distribution is skewed. These distributions are sometimes called asymmetric or asymmetrical distributions as they don’t show any kind of symmetry. Symmetry means that one half of the distribution is a mirror image of the other half. For example, the [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) is a [symmetric distribution](https://www.statisticshowto.datasciencecentral.com/symmetric-distribution-2/)with no skew. The tails are exactly the same.
* [](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/normal-distribution-probability.jpg)
* *A normal curve.*
* A [left-skewed distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/#SkewLeft) has a long left tail. Left-skewed distributions are also called *negatively-skewed* distributions. That’s because there is a [long tail](https://www.statisticshowto.datasciencecentral.com/long-tail-distribution/) in the negative direction on the number line. The mean is also to the left of the peak.
* A [right-skewed distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/#SkewRight) has a long right tail. Right-skewed distributions are also called positive-skew distributions. That’s because there is a long tail in the positive direction on the number line. The [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)is also to the right of the peak.



1. What is the difference between reporting and analysis

Reporting is “the process of organizing data into informational summaries in order to monitor how different areas of a business are performing.”

Analytics is “the process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.”

1. Find the difference between dimension and measure.

Measures are numerical values that mathematical functions work on them. For example, a sales revenue column is a measure because you can find out a total or average the data.

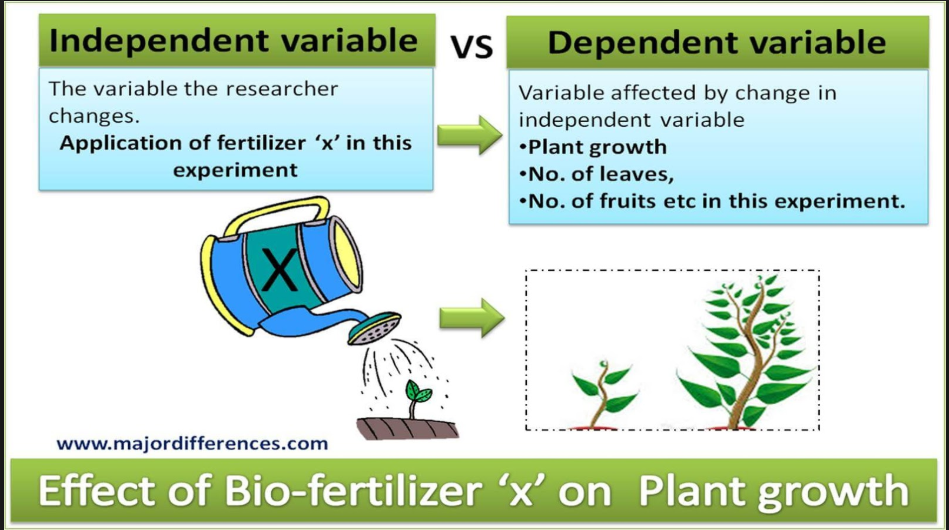
Dimensions are qualitative and do not total a sum. For example, sales region, employee, location, or date are dimensions.

1. Learn about categorical variables and continuous variables.

* Categorical variables contain a finite number of categories or distinct groups. Categorical data might not have a logical order. For example – color , types of car
* Continuous variables are numeric variables that have an infinite number of values between any two values. A continuous variable can be numeric or date/time.

1. Independent variables and dependent variables.

The two main variables in an experiment are the independent and dependent variable. An independent variable is the variable that is changed or controlled in a scientific experiment to test the effects on the dependent variable. A dependent variable is the variable being tested and measured in a scientific experiment.



* Generally the “independent” variable goes on the x-axis (the bottom, horizontal one) and the “dependent” variable goes on the y-axis (the left side, vertical one).
* Also x variables are called predictors / features and y variable is called response.

1. Learn the difference between absolute measure and relative measure. How this makes sense.

* An absolute measure of forecasting error is a measurement that uses numerical variations in investment forecasting to determine the degree of error. For example, if an analyst forecasts that a company's stock will be worth $90 per share by the end of the fiscal year, and the actual stock price at the end of the year is $85, the absolute measure of forecasting error if $5 per share. If another analyst forecast a value of $80, this forecast has the same $5 absolute measure of error; absolute measures take the form of positive numbers, regardless of whether they represent high or low estimations.
* Relative measures of forecasting error are the major alternative to absolute measures. They use statistical variations based on percentages to determine how far from reality a forecast is. For example, a $5 absolute measure of error represents only a 5 percent relative measure of error for an investment with an actual value of $100. However, the same $5 absolute measure of error represents a 25 percent relative error if the investment product is only worth $25.

1. What is EDA ?

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations / data visualizations .