What is the Central Limit Theorem and why is it important?

- "Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible.
- While we can't obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample.

The Central Limit Theorem addresses this question exactly.

- Formally, it states that if we sample from a population using a sufficiently large sample size, the mean of the samples (also known as the sample population) will be normally distributed (assuming true random sampling).
- What's especially important is that this will be true regardless of the distribution of the original population.
- In the study of probability theory, the central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution (also known as a "bell curve"), as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population distribution shape.

Raise two kind of questions –
1. What is sampling and technique of sampling
2. Give me an example of CLT – Measuring the average height among all the people in the state

What is sampling? How many sampling methods do you know?

"Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined."

Two types of Sampling can be done:
1. Probability Sampling
2. Non-Probability Sampling

Probability sampling – Randomization. Equal chance to be picked up.
Type –
1. **Simple Random sampling:** Equal chance of getting selected
2. **Stratified Sampling:** Divide elements of population into smaller group based on similarity (age/sex/location) and select one.
3. **Cluster Sampling:** Entire population divided into cluster. Then
    a. **Single stage clustering sample** – Random selection of clusters.
    b. **Two stage clustering sampling** – First we randomly select clusters and then from selected clusters we randomly select for sampling
4. **Systematic sampling:** Selection of sample are systematic not random. Suppose we will select only those elements who are odd numbered position. Example 1,3,5,7,9, …
5. **Multi-stage sampling:** Combination of one or more above methods. For example – Country can be divided based into states, cities, urban or rural and based on similarity merge or form strata.

Non probability sampling – No randomization. May have bias result.
Type – Convenience sampling, Purposive sampling, Quota sampling, Referral or snowball sampling.
1. **Convenience sampling**: Selected based on availability. Rare and costly
2. **Purposive sampling**: Select only those which suits best for study.
3. **Quota sampling**: Elements are selected until exact proportions of certain type of data is obtained.
4. **Referral or Snowball sampling**: When population is unknown. There we select first element and ask for recommendation to the same element who fit the description. Example – Corona virus testing.

What is selection bias?

- Ideally, you should randomly select every participant in a survey. But, sometimes biases creep in, whether intentional or unintentional. Selection bias takes away from the "randomness" you are hoping to achieve. It's usually a result of not using the correct procedures to choose your participants.
- **For Example** - A classic example of undercoverage is the 2013 Indian Election voter survey, which predicted that Congress would beat BJP in the 2013 Parliament Election. The survey sample suffered from undercoverage of Kerala voters, who tended to be Congress voters. Undercoverage is often a problem with **convenience samples**.