**Q1.What is the difference between data analysis and machine learning?**

Data analysis requires strong knowledge of coding and basic knowledge of statistics Machine learning, on the other hand, requires basic knowledge of coding and strong knowledge of statistics and business.

**Q2.What is big data?**

Big data has 3 major components – volume (size of data), velocity (inflow of data) and variety (types of data)Big data causes “overloads”

**Q3.What are the four main things we should know before studying data analysis?**

Descriptive statistics

Inferential statistics

Distributions (normal distribution / sampling distribution)

Hypothesis testing

Q4.**What is the difference between inferential statistics and descriptive statistics?**

Descriptive statistics – provides exact and accurate information.

Inferential statistics – provides information of a sample and we need to inferential statistics to reach to a conclusion about the population.

**Q5.What is the difference between population and sample in inferential statistics?**

From the population we take a sample. We cannot work on the population either due to computational costs or due to availability of all data points for the population.

From the sample we calculate the statistics

From the sample statistics we conclude about the population

Q6. **What are descriptive statistics?**

Descriptive statistic is used to describe the data (data properties)

5-number summary is the most commonly used descriptive statistics

Q7. **Most common characteristics used in descriptive statistics?**

* Center – middle of the data. Mean / Median / Mode are the most commonly used as measures.
  + Mean – average of all the numbers
  + Median – the number in the middle
  + Mode – the number that occurs the most. The disadvantage of using Mode is that there may be more than one mode.
* Spread – How the data is dispersed. Range / IQR / Standard Deviation / Variance are the most commonly used as measures.
  + Range = Max – Min
  + Inter Quartile Range (IQR) = Q3 – Q1
  + Standard Deviation (σ) = √(∑(x-µ)2 / n)
  + Variance = σ2
* Shape – the shape of the data can be symmetric or skewed
  + Symmetric – the part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median
  + Left skewed – the left tail is longer than the right side
  + Right skewed – the right tail is longer than the left side
* Outlier – An outlier is an abnormal value
  + Keep the outlier based on judgement
  + Remove the outlier based on judgement