1. Write the Gaussian Distribution empirical formula.

P [μ – 2σ <= X <= μ + 2σ] ≈ 95 % So the second formula basically says that the probability of a variable that falls within the range of μ – 2σ and μ + 2σ is 95 %. which means 95% of the data points belonging to the random variable X fall within the range of the second standard deviation.

2. What is the Z-score, and why is it important?

The value of the z-score tells you how many standard deviations you are away from the mean. If a z-score is equal to 0, it is on the mean. A positive z-score indicates the raw score is higher than the mean average. For example, if a z-score is equal to +1, it is 1 standard deviation above the mean.

3. What is an outlier, exactly?

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

4. What are our options for dealing with outliers in our dataset?

3 different methods for dealing with outliers: the univariate method, the multivariate method and the Minkowski error. These methods are complementary and, if our data set has many and difficult outliers, we might need to try them all.

5. Write the sample and population variances equations and explain Bessel Correction.

Bessel correction refers to the n-1 part used as the denominator in the formula of sample variance or sample distribution.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$