# Statistics Assignment 3

1. Write the Gaussian Distribution empirical formula.
    - If any random variable X, follows normal distribution or gaussian distribution then
    - 68% of its data lie within range of one standard deviation region (mean-1*SD, mean+1*SD)
    - 95% of its data lie within range of two standard deviation region (mean-2*SD, mean+2*SD)
    - 99.7% of data lie within range of three standard deviation region (Mean-3*SD, mean+3*SD)

2. What is the Z-score, and why is it important?
    - Z-score is difference between data point x and mean of dataset divided by standard deviation of that dataset.
    - Z-score used in standardization of data. Most of dataset have many features and values of features have different scales like one feature has unit in Km, one feature has unit in Kg.
    - So, to draw all features under one scale z-score formula is used.
    - Z-score is standardized value which help to compare between different datasets.
3. What is an outlier, exactly?
    - Outlier is nothing but abnormal value in dataset which is too far away from mean value.
    - Due to presence of outlier our mean shift from actual value and gives wrong information about middle part of dataset.

4. What are our options for dealing with outliers in our dataset?
    - There are three ways to deal with outliers in dataset:
        - **Trimming:**

          In trimming, if number of outliers is less then we can directly eliminate that datapoint from dataset.

But it is preferred to use trim only when number of outliers is less and we have larger dataset and elimination of data point do not affect actual distribution.

- **Imputation with median and mode**
  In this case, we just find out outliers and replace outliers with either median or mode so that our distribution is not affected. So, with this no need to delete data.
- **Boxplot**
  In boxplot, just replace outliers which are above mean by max_value=Q3+1.5*IQR
  And replace outliers below mean by min_value=Q1-1.5*IQR

5. Write the sample and population variances equations and explain Bessel Correction.
   - Population variance, σ^2= ∑ (x-μ) ^2/N
   - Sample variance, s^2=∑ (x-$\hat{x}$) ^2/(n-1)
   - x= random data point, μ=Population mean, N= no of population data point, $\hat{x}$ = sample population mean, n=no of sample data points
   - In sample variance equation we usually divide by (n-1). If we divide summation by n, value of sample variance is much far away from population means.
   - This result in biasness. We underestimate true value of population parameter.
   - If we divide summation (n-1) we get value of sample variance close to population variance and biasness will be eliminated.
   - This is Bessel correction