**Visualizing data**

We can see that in each attribute there are instances with a value of 0 when it’s not possible, especially, there’s a lot of potential missing data in skinfold.

**Identifying missing data**

There is a lot of missing data when it comes to the skin fold thickness attribute. And less than 5% of the population when it comes to bmi and blood pressure.

**Marking missing data**

BMI: Mark as NaN and remove those rows as they are less than 5% of the population

SkinFold: Mark as NaN and impute with the mean of skinfold attribute

BloodP: Mark as NaN and remove rows because less than 5% of population

**Identifying outliers**

BloodP: We have 30 outliers.

Skinfold: There are 34 outliers.

BMI: There are 24 outliers.

**Marking outliers**

BloodP: We have 8 values outside of 3 STDs which would okay to remove

Skinfold: There are 3 values outside of 3 STDs which can definitely be removed especially the instance that is near 100.

BMI: Despite being outside of 3STDs, I would keep these instances but I don’t know why.

**Top 4 Attributes**

Finding top 4 attributes using Linear regression and recursive feature elimination. The top 4 attributes are number of times pregnant, BMI, diabetes pedigree function and plasma glucose concentration.

**Normalization**

Use normalization because the top 4 attributes selected do not follow a bell curve distribution.