**Visualizing data**

We can see that both mpg and horsepower have instances with a value of 0 and this isn’t possible. We can also see that the data is positively skewed in pretty much all the attributes analysed.

**Identifying missing data**

There are 4 instances that have an mpg below 0 and 6 instances that had a question mark which have been replaced by -1/

**Marking missing data**

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

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

**Identifying outliers**

mpg: We have 0 outliers.

Cylin: There are 4 outliers.

Displ: There are 0 outliers.

Hpow: There are 4 outliers.

**Marking outliers**

We should remove all outliers as they don’t reach 5% of the population. Ask about this.

**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.