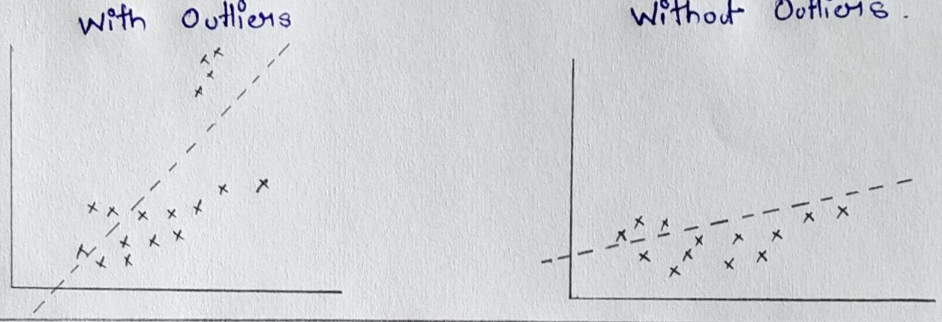
TREATMENT OF OUTLIERS

Q) Should we drop or not to drop Outliers?

Ans → Dropping data is always a harsh step and should be taken when we are sure that outlier is a measurement error, which we generally don't know.

- When we drop the data, we loose information in terms of variability in data. Suppose, we are doing segmentation, variability of data becomes major role because if there is no variation then we cannot divide/ segment the data into groups.
- So when we have too many observations and outliers are very very few, we can think of dropping the observations.

Example → Regression Line



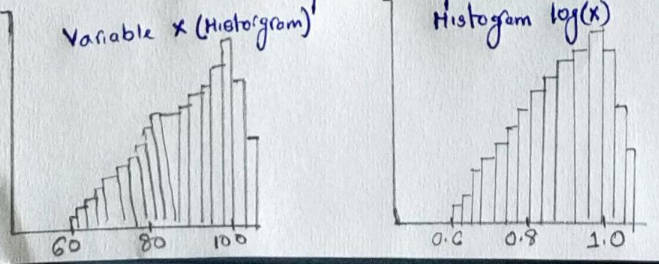
What are the steps to remove/deal with Outliers?

Ans - I) Capping   II) Log-Scale Transformation.   III) Binning

I) Capping → This methods involve setting the extreme values of an attribute to some specified value.

For example, bottom 5% of value are set equal to minimum value in 5th percentile, while the upper 5% of values are set equal to the maximum value in 95th percentile.

II) Log scale transformation → This method is often used to reduce the variability of data including outlying observations. Here value is change with log i.e, y → log(y). It's often preferred when the response variable follow exponential distribution/right skewed.



→ This is controversial steps as it does not reduce variance, only thing is it change the scale
60, 80, 100 → 0.6, 0.8, 1.0

III) Binning → This refers to dividing a list of continuous variable into groups. We do this to discover set of patterns in continuous variable, which is difficult to analyze otherwise. But it also lead to loss of information and loss of power because suppose if we group them, then even if they are/ have variability they are one group.

Q) Which are the algorithims affected by outliers?

Ans - ML models like linear & logistic regression are easily affected by the outliers.

Random forest /decision tree are not affected by outliers.