Deleting and Capping Outliers

Handling Outliers

Example:

- The age variable in our dataset which predicts loan approval.
- The normal range of age can be 15-60 years but there may be some people with age greater than 60.
- So in this case, where people having age greater than 60 applied for the loan are not outliers, we can not simply delete them.

Deleting Outliers

1. We are sure that the outliers are due to an entry error or due to measurement error.
2. If the outliers create a significant relationship between two independent variables which is against the assumption of many of our machine learning algorithms.

Capping Outliers

- Capping refers to replacing the outliers to a near value so that we can keep the point in our analysis and it also does not skew the data.
- **Note:** Other than Deleting and Capping, there are two more ways of handling outliers.
    1. Imputing.
    2. Binning.

Ninja Tip

- ***Imputing:*** If an outlier seems to be due to some mistake and we recognise the mistake.
- ***Binning:*** It is the process of transforming numerical variables into categorical type.
    - ***Example:*** we can bin the age variable into categories such as 20-40, 40-60, 60-90 and above.