Everything about Data Preparation

Data preparation is the process of preparing the raw data by assuring the quantity and quality of data. This is achieved by looking at the distribution of data.

A frequency distribution is a count of how often each variable contains each value in a data set. The easiest way to visualize a distribution is to plot a histogram.

Problems during Data Preparation

1) **Data entry errors** : Typographic errors during data input

2) **Outliers** : Values that are much larger or smaller than all others

Outliers can be removed or collect more data to represent that aspect of the world. Can be caused by data entry errors

3) **Minority Values** : Values that only appear infrequently in the data

Can be removed or just collect more data to represent them. Can be caused by data entry errors (e.g Male/Female, M/F)

4) **Flat and Wide Variables** : Variables that all their values are minority values

They should be excluded from the model

Variability

The correct way to draw a straight line through a scatter plot is to find one that minimises the distance between all the points and the line. This distance is the error of the model and is known as MSE - Mean Squared Error

The process of learning is the process of minimizing the MSE