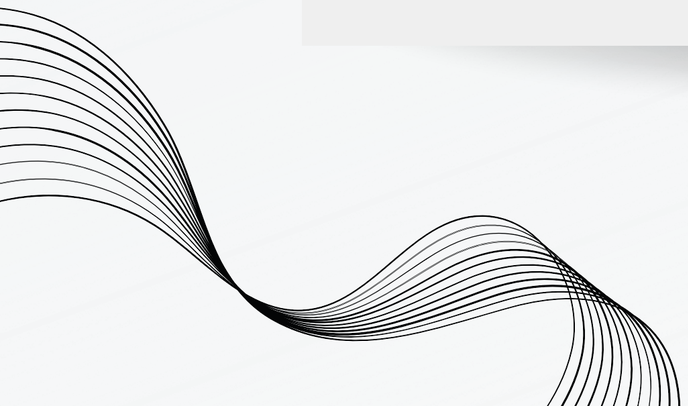
DAY 2

NAYAN PANDEY

SCALING

Different features have different scales of data. While some features can have values in the millions, other features can be quite small.

To make sure that the features with small values also contribute when using them in our models, we need to scale them properly.

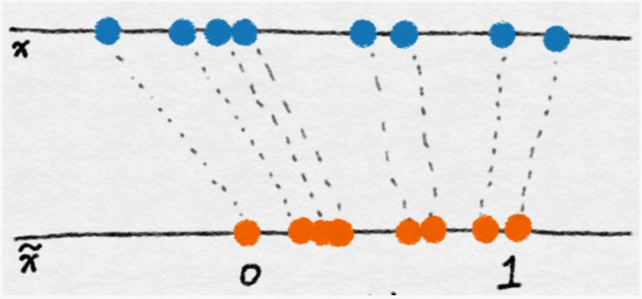


MIN MAX SCALING

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

- In this technique, we take the maximum and minimum value of the feature and scale all the other data points in the [0, 1] range using those min and max values.

- It transforms the data linearly preserving the relative distance between the data points.
- However, if outliers are present, then this can impact the range of the remainder of data.

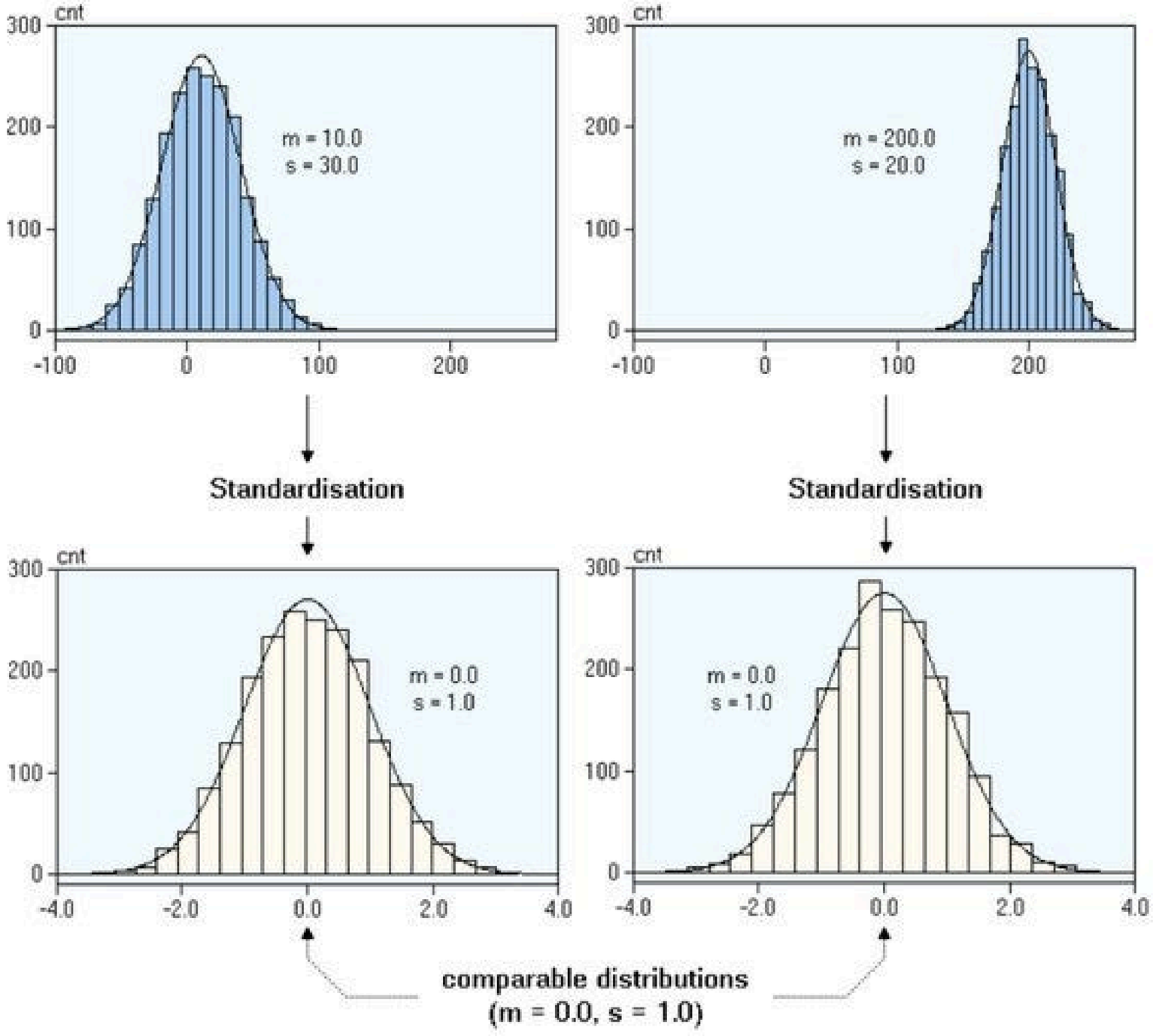


STANDARDIZATION

$$z = \frac{x_i - \mu}{\sigma}$$

- In this technique, we calculate the mean and standard deviation of the feature and scale the feature using those two values.
- The resulting distribution has a mean of 0 and a variance of 1.

- The outliers does not affect the final distribution of the data and they are also retained.
- Its range is not bounded like min-max scaling


cnt
m = 10.0
s = 30.0
cnt
m = 200.0
s = 20.0
Standardisation
Standardisation
cnt
m = 0.0
s = 1.0
cnt
m = 0.0
s = 1.0
comparable distributions
(m = 0.0, s = 1.0)

OUTLIERS

Some data points can lie extremely far from the rest of the observations. Those data points are outliers.

Problem 1

Outliers can affect the statistical measures like mean, median, variance giving us a false impression of the dataset

Problem 2

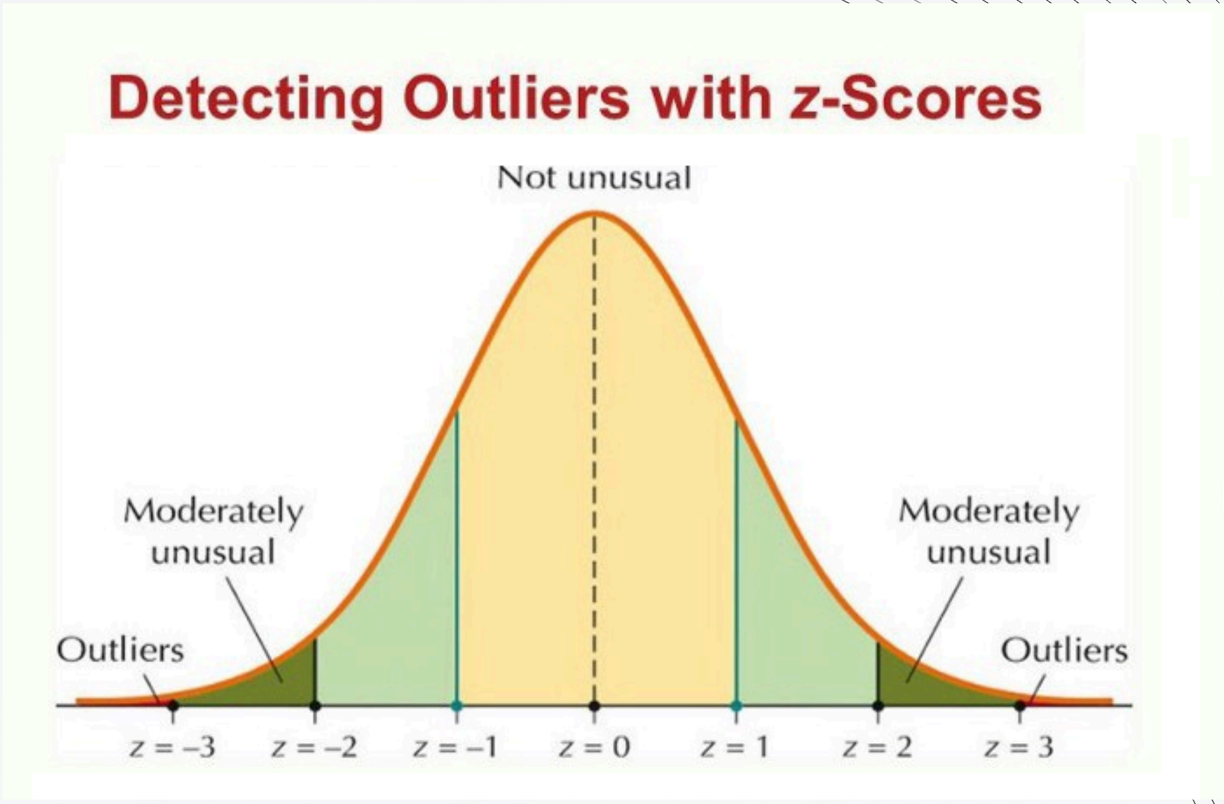
They can affect the model performance and hinder our predictive ability.

Problem 3

It also complicates interpretability making it harder to find patterns and relationships among features in the data.

DETECT OUTLIERS

- We can detect outliers using the box-plots and scatter plots.
  - We can also use methods like IQR, Z-score to find outliers.



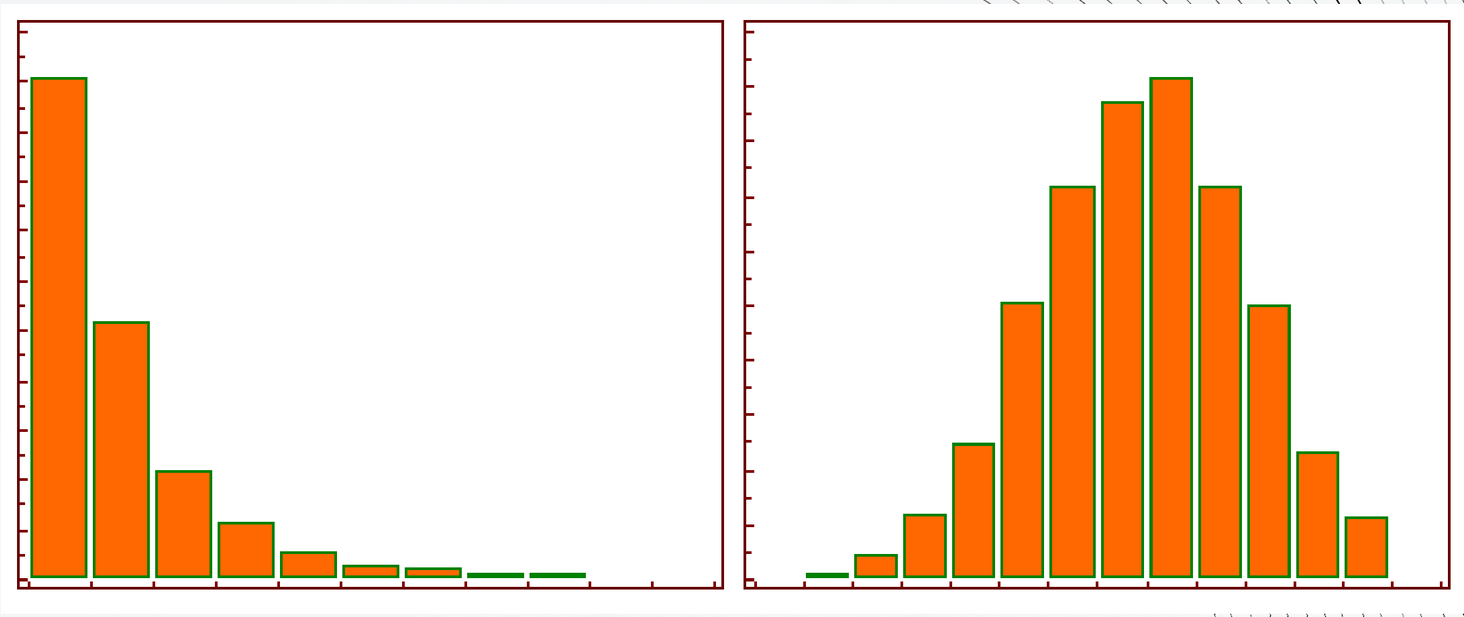
CUSTOMERS

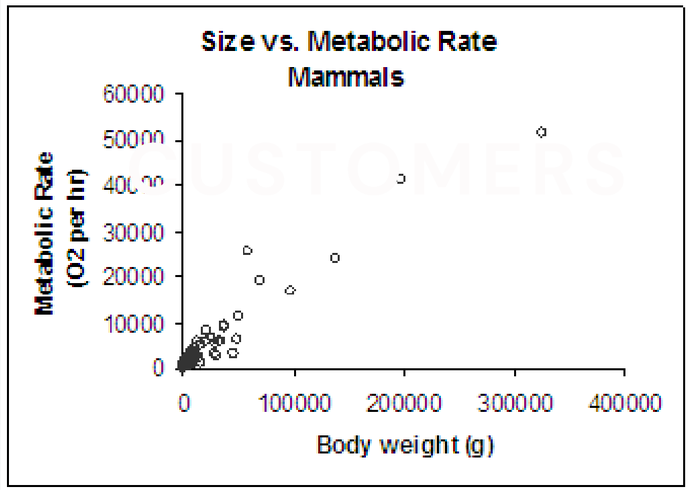
HANDLE OUTLIERS

1. Removal of outliers

2. Transformation: log, square root etc.

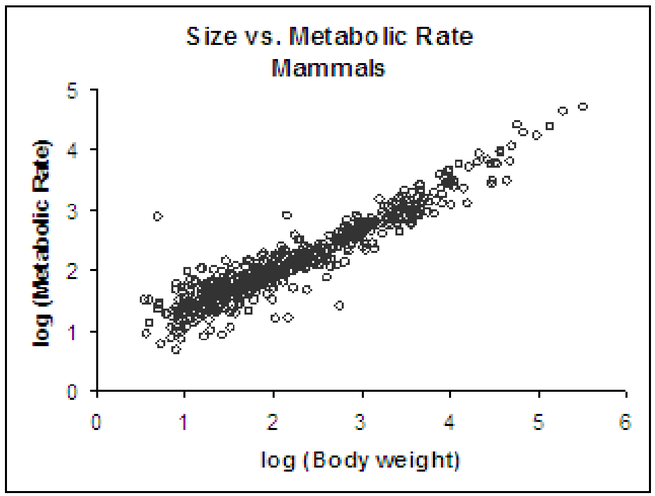
3. Set a maximum and minimum threshold.





Log

Trans-form



THANK YOU