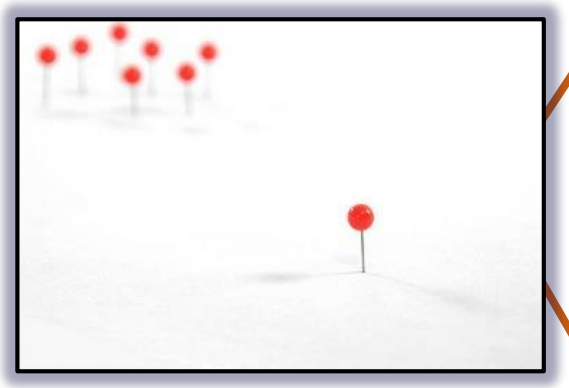
Train In Data

# Outlier Engineering

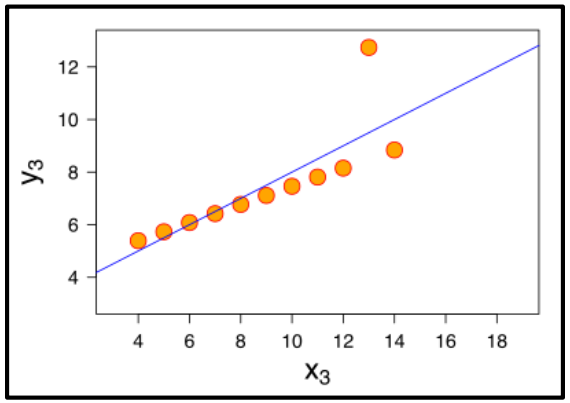
# Outliers

- An outlier is a data point which is significantly different from the remaining data.
- “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” [D. Hawkins. Identification of Outliers, Chapman and Hall , 1980.]

# Algorithms susceptible to outliers



Linear models




Adaboost

Tremendous weights

Bad generalisation

# Handling Outliers

What can we do if we find outliers in our variables?

# Ways to engineer outliers

| | |
|---|---|
| Trimming | • Removing outliers from the data set |
| Missing data | • Treat outliers as missing data and perform missing data imputation<br>• Any technique in section 4 |
| Discretisation | • Put outliers into lower / upper bins<br>• Any technique in section 8 |
| Censoring | • Capping<br>• Top / Bottom coding<br>• Winsorization |

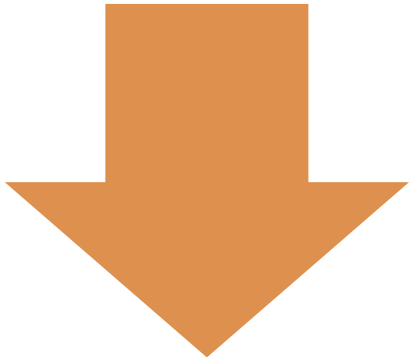
# Ways to engineer outliers

| | |
|---|---|
| Trimming | • Removing outliers from the data set |
| Missing data | • Treat outliers as missing data and perform missing data imputation<br>• **Any technique in section 4** |
| Discretisation | • Put outliers into lower / upper bins<br>• **Any technique in section 8** |
| Censoring | • Capping<br>• Top / Bottom coding<br>• Winsorization |

# Ways to engineer outliers

| | |
|---|---|
| Trimming | • Removing outliers from the data set |
| Missing data | • Treat outliers as missing data and perform missing data imputation<br>• Any technique in section 4 |
| Discretisation | • Put outliers into lower / upper bins<br>• Any technique in section 8 |
| Censoring | • Capping<br>• Top / Bottom coding<br>• Winsorization |

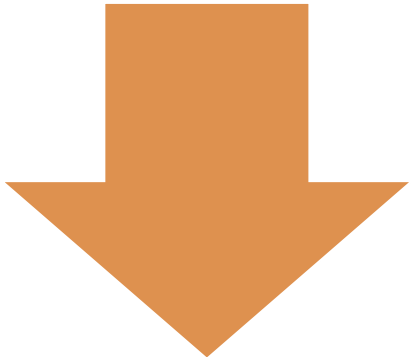
# Trimming: pros and cons



Fast

Can remove big chunk of data

# Capping: pros and cons



No data is removed

Distorts variable distribution

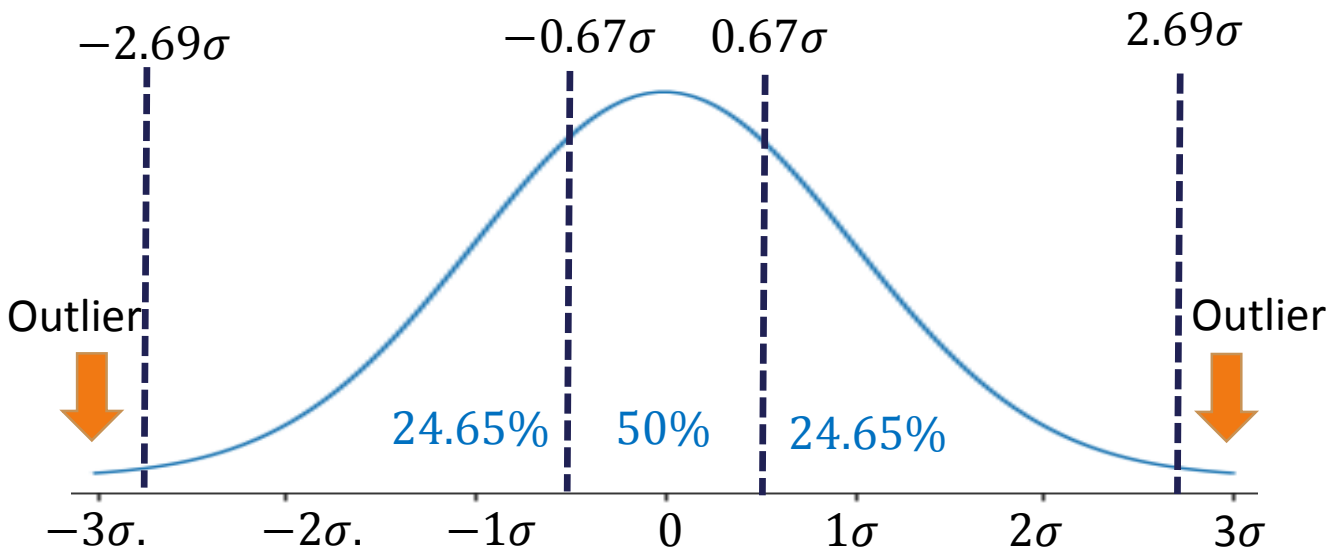
# Detecting Outliers

Extreme Value Analysis

# Detecting outliers

- Gaussian distribution (mean and std)
- Inter-quantal range proximity rule
- Quantiles

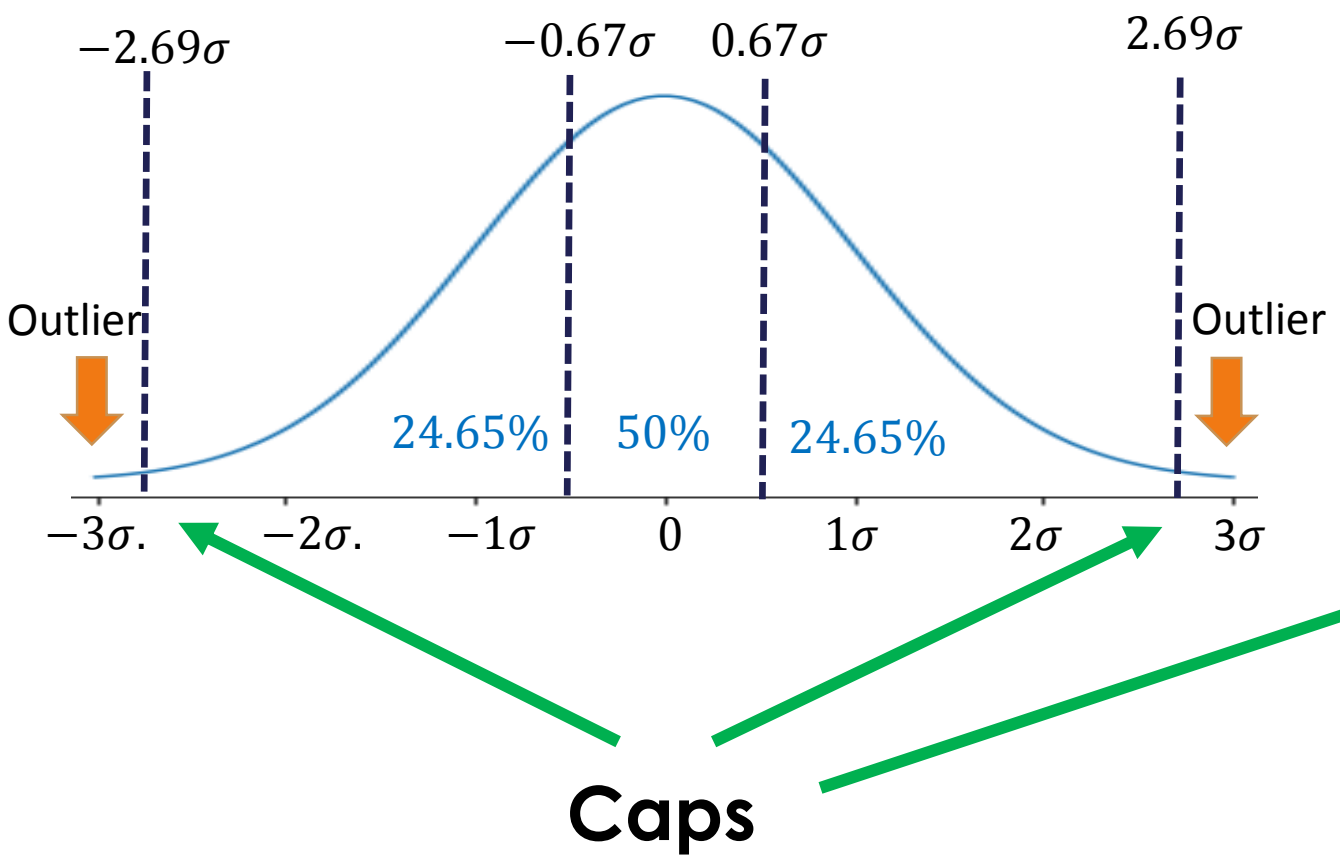
# Normal distribution




- ~99% of the observations of a normally distributed variable lie within the mean ± 3 × standard deviations.

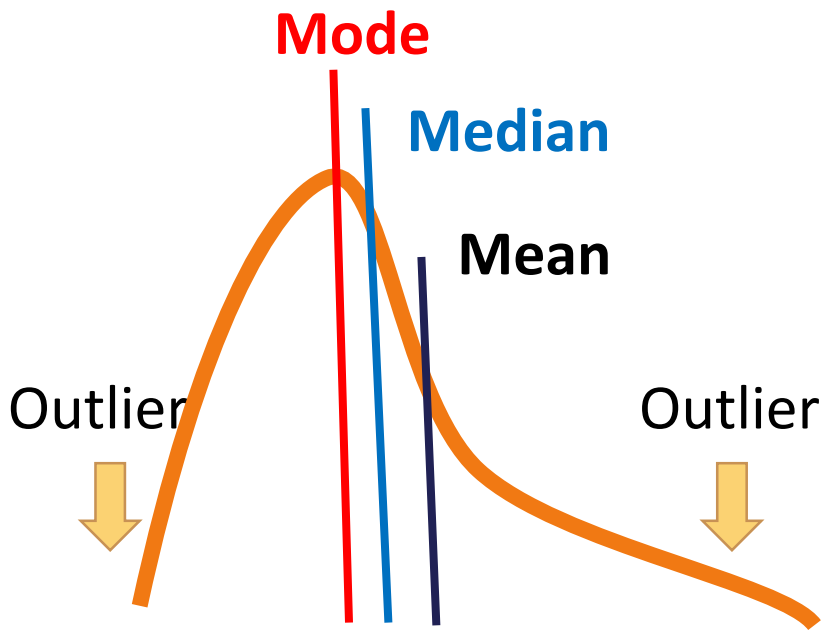
- Values outside **mean ± 3 × standard deviations** are considered outliers

# Normal distribution




- ~99% of the observations of a normally distributed variable lie within the mean ± 3 × standard deviations.
- Values outside **mean ± 3 × standard deviations** are considered outliers

# Skewed distributions

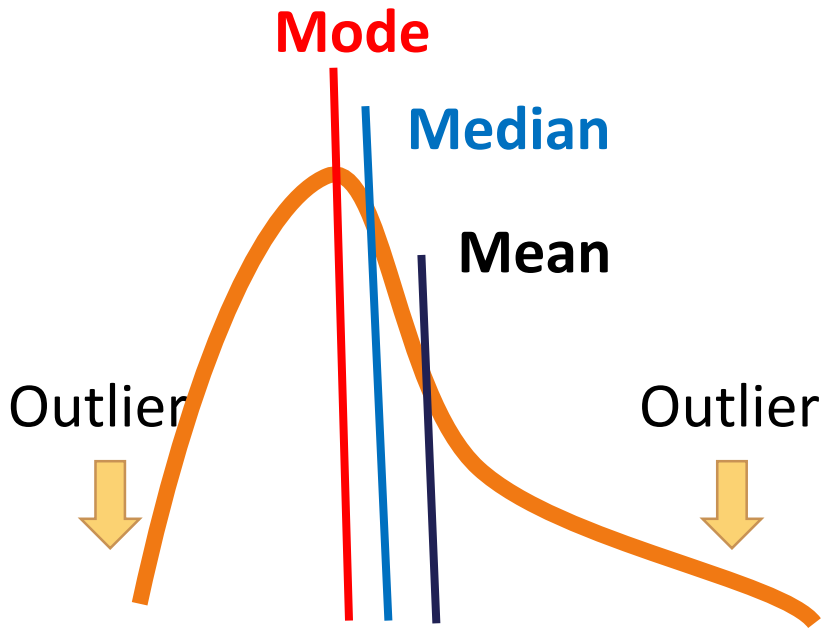



Skewed distribution

- The general approach is to calculate the quantiles, and then the inter-quantile range (IQR), as follows:
- IQR = 75$^{th}$ Quantile – 25$^{th}$ Quantile
- **Upper limit = 75$^{th}$ Quantile + IQR × 1.5**
- **Lower limit = 25$^{th}$ Quantile - IQR × 1.5**

Note, for extreme outliers, multiply the IQR by 3 instead of 1.5

# Skewed distributions



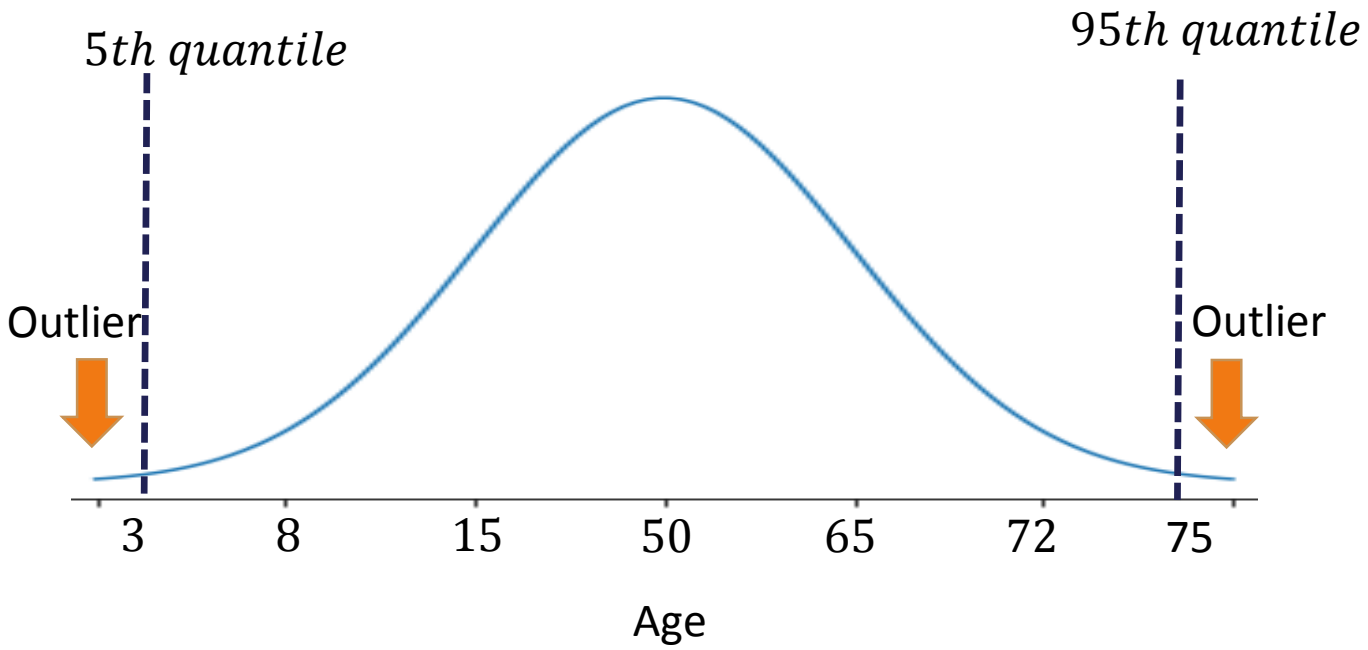

Skewed distribution

- The general approach is to calculate the quantiles, and then the inter-quantile range (IQR), as follows:
- IQR = 75$^{th}$ Quantile – 25$^{th}$ Quantile
- **Upper limit = 75$^{th}$ Quantile + IQR × 1.5**
- **Lower limit = 25$^{th}$ Quantile - IQR × 1.5**

Note, for extreme outliers, multiply the IQR by 3 instead of 1.5

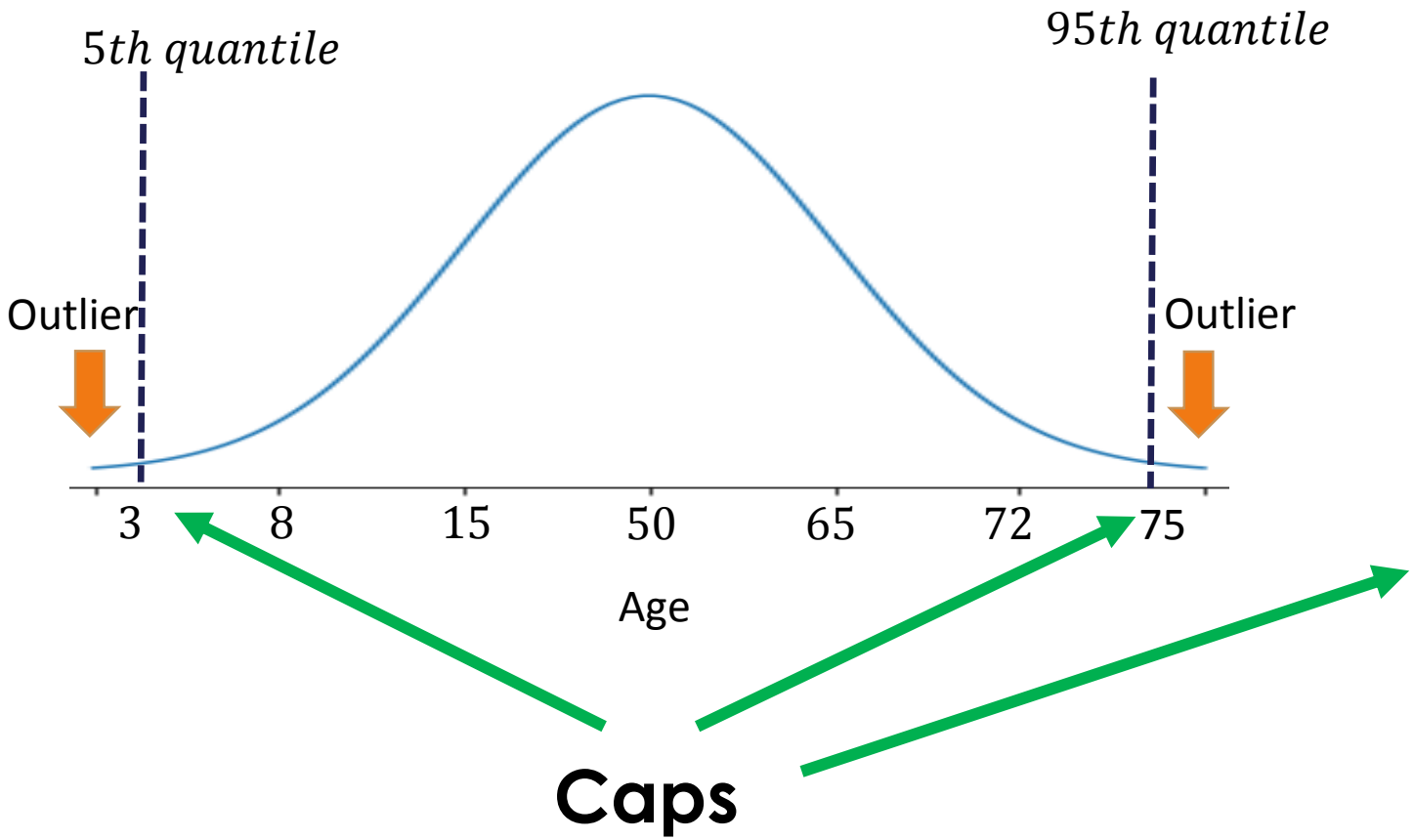
**Caps**

# Normal distribution




- ~95% of the observations above the 5th quantile
- ~95% of the observations below the 95th quantile
- Values **above or below the 95th or 5th quantile** are considered outliers

# Normal distribution




- ~95% of the observations above the 5th quantile
- ~95% of the observations below the 95th quantile
- Values **above or below the 95th or 5th quantile** are considered outliers

# Accompanying Jupyter Notebook


- Accompanying Jupyter Notebook
- How to perform outlier engineering:
  - Gaussian approximation
  - IQR
  - Quantiles
  - Pandas and Feature-engine

Train In Data

# THANK YOU

www.trainindata.com