Train In Data

Outliers

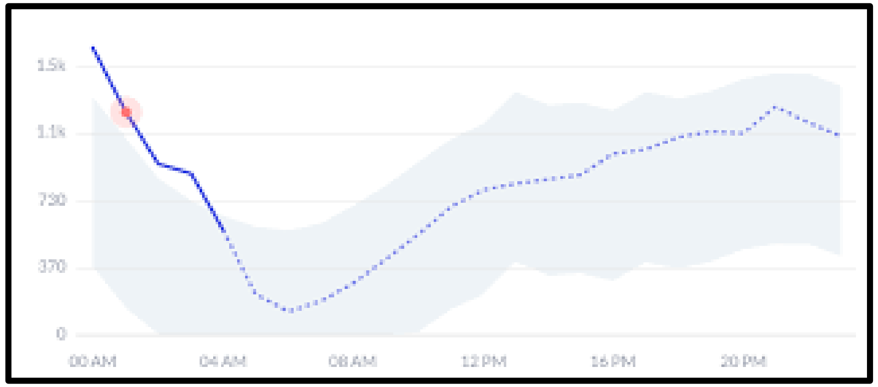
Outliers

- An outlier is a data point which is significantly different from the remaining data.

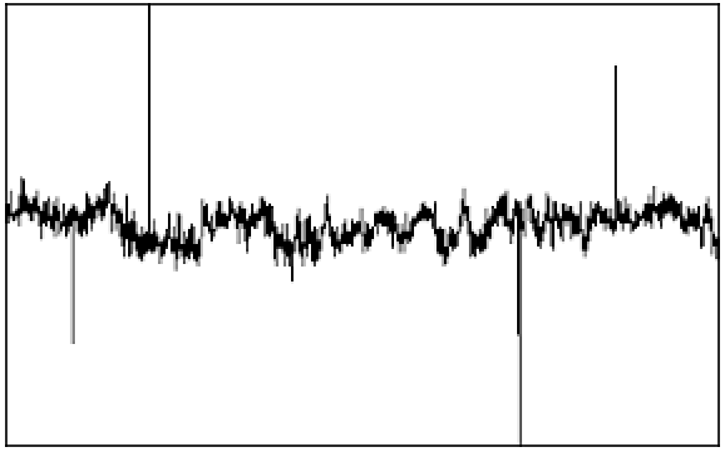
- “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” [D. Hawkins. Identification of Outliers, Chapman and Hall , 1980.]

Should outliers be removed?

Revenue forecasting



Credit card transactions

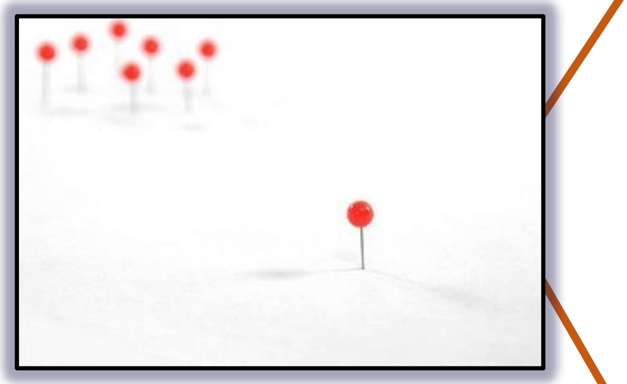


Depending on the context, outliers either deserve special attention or should be completely ignored.

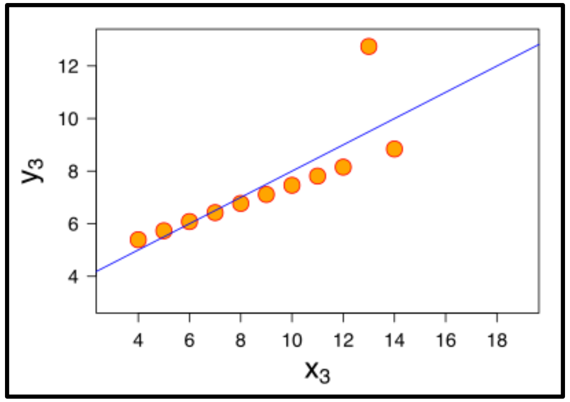
Approach to outliers in this course

- Handle outliers in cases where they may affect model performance
- The course is tailored to improve model performance
- Out of scope: outlier detection
  - A massive field with lots of techniques

Algorithms susceptible to outliers



Linear models



Adaboost

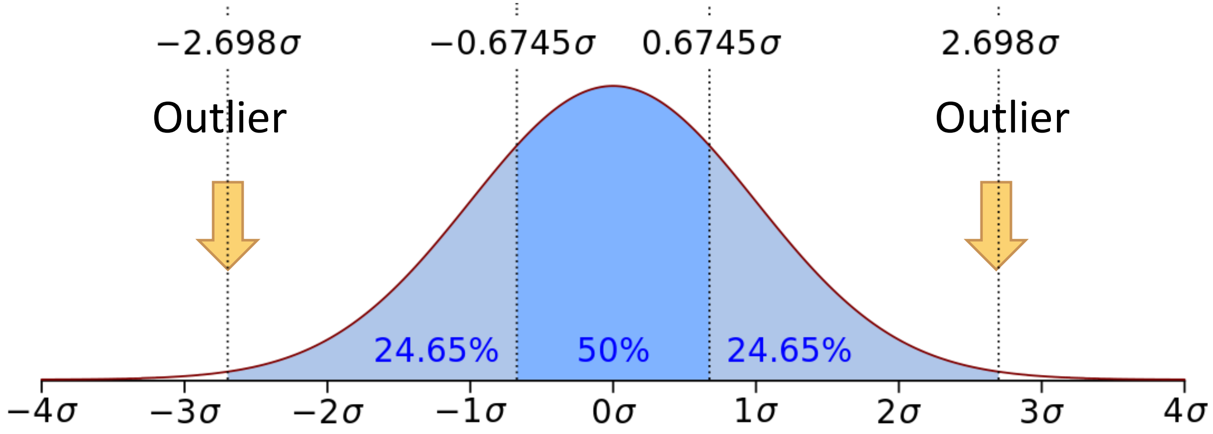
Tremendous weights

Bad generalisation

Detecting Outliers

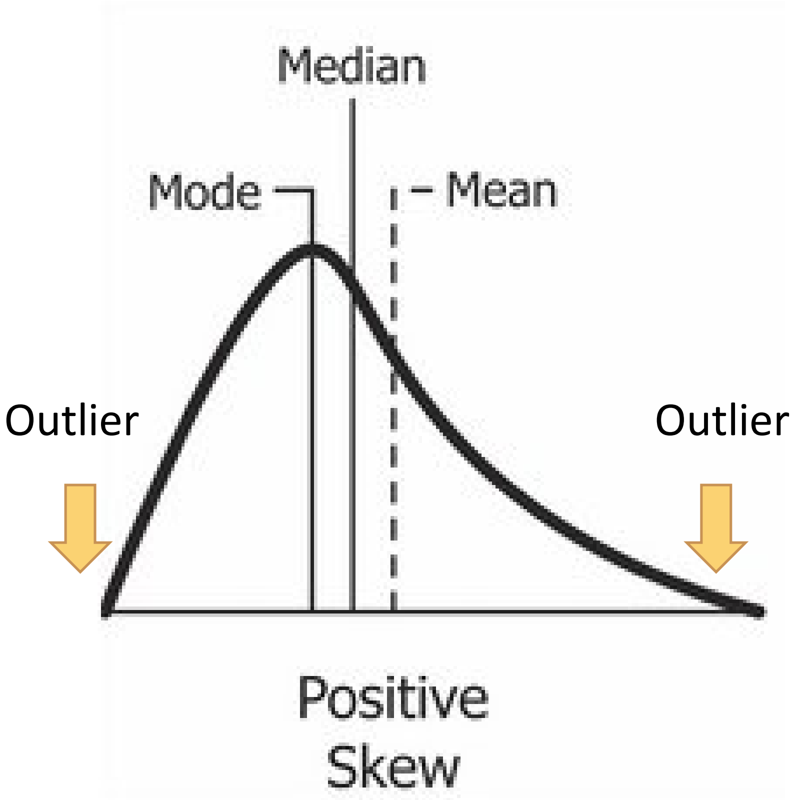
Extreme Value Analysis

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



- The general approach is to calculate the quantiles, and then the inter-quantile range (IQR), as follows:

- IQR = 75th Quantile – 25th Quantile

- Upper limit = 75th Quantile + IQR × 1.5

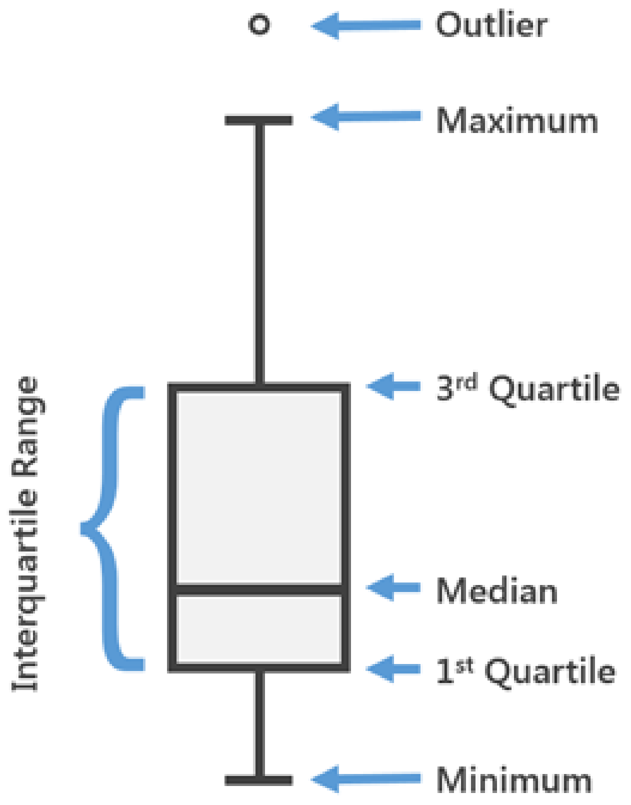
- Lower limit = 25th Quantile - IQR × 1.5

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

Notes on quantiles

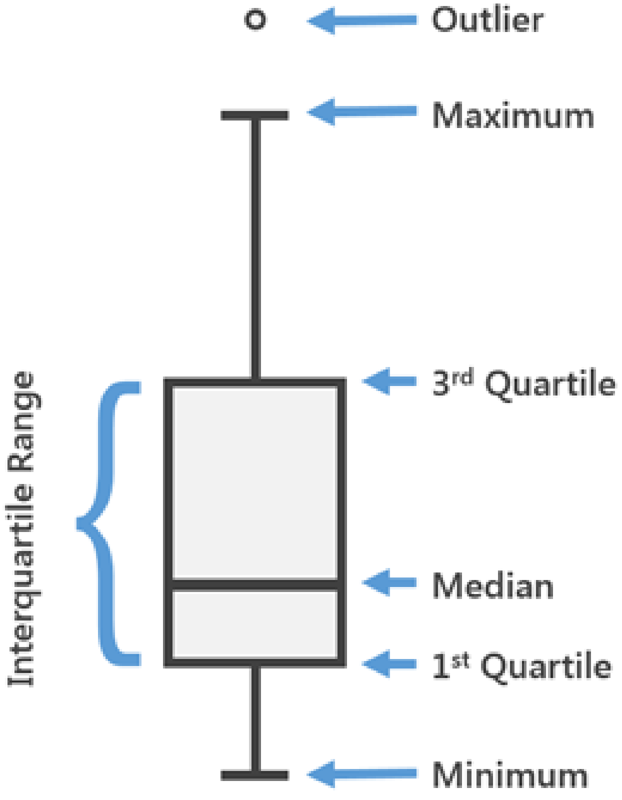
- Quartiles = dividing the distribution in 4
- Quantiles = dividing the distribution into 100
- 1st Quartile = 25th Quantile
- 3rd Quartile = 75th Quantile
- 2nd Quartile = 50th Quantile = Median
- IQR = 75th Quantile – 25th Quantile = 3rd Quartile – 1st Quartile

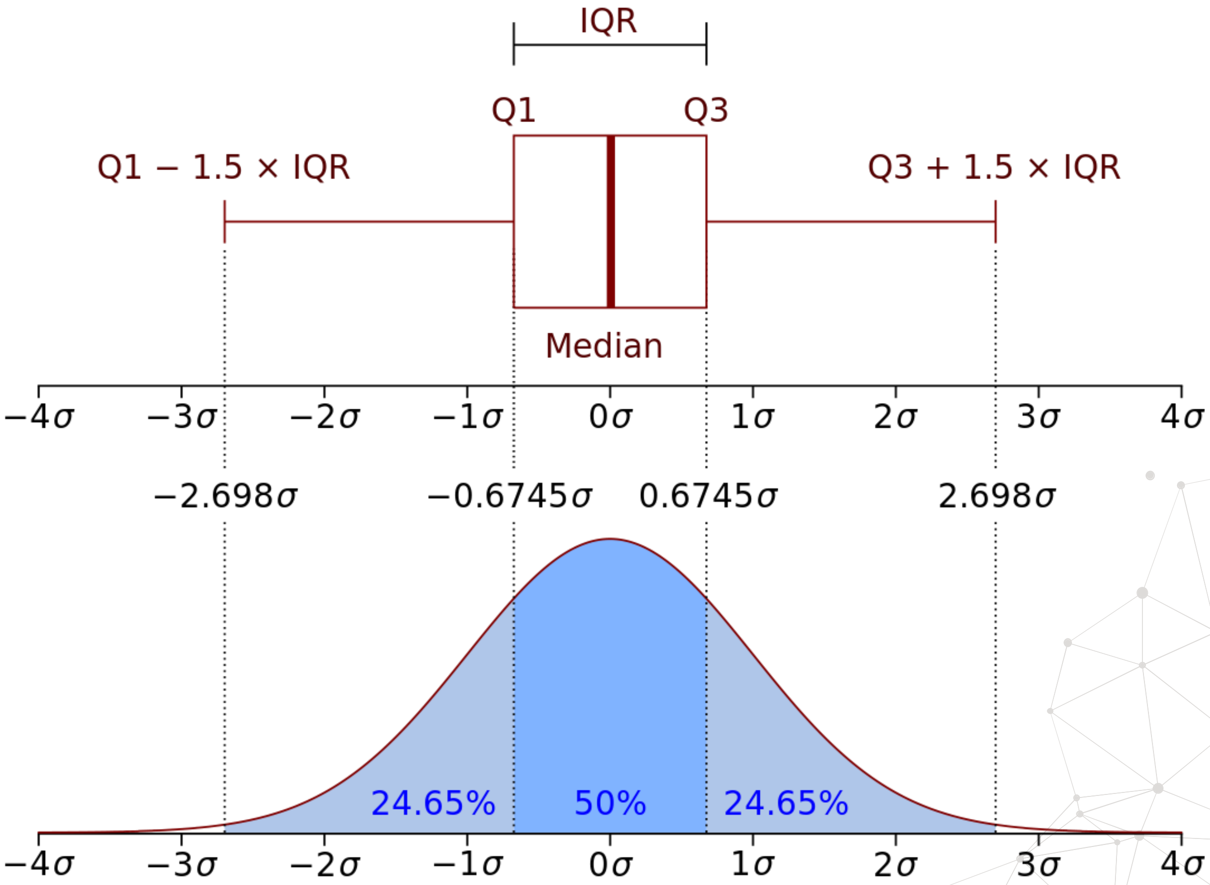
Visualising outliers - Boxplots



Images taken from pro.arcgis.com and wiki.commons

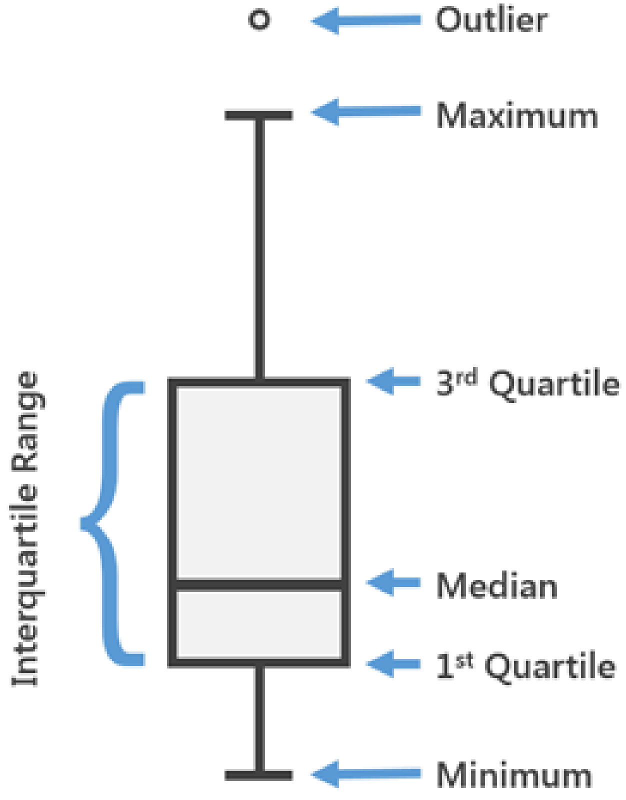
Visualising outliers - Boxplots

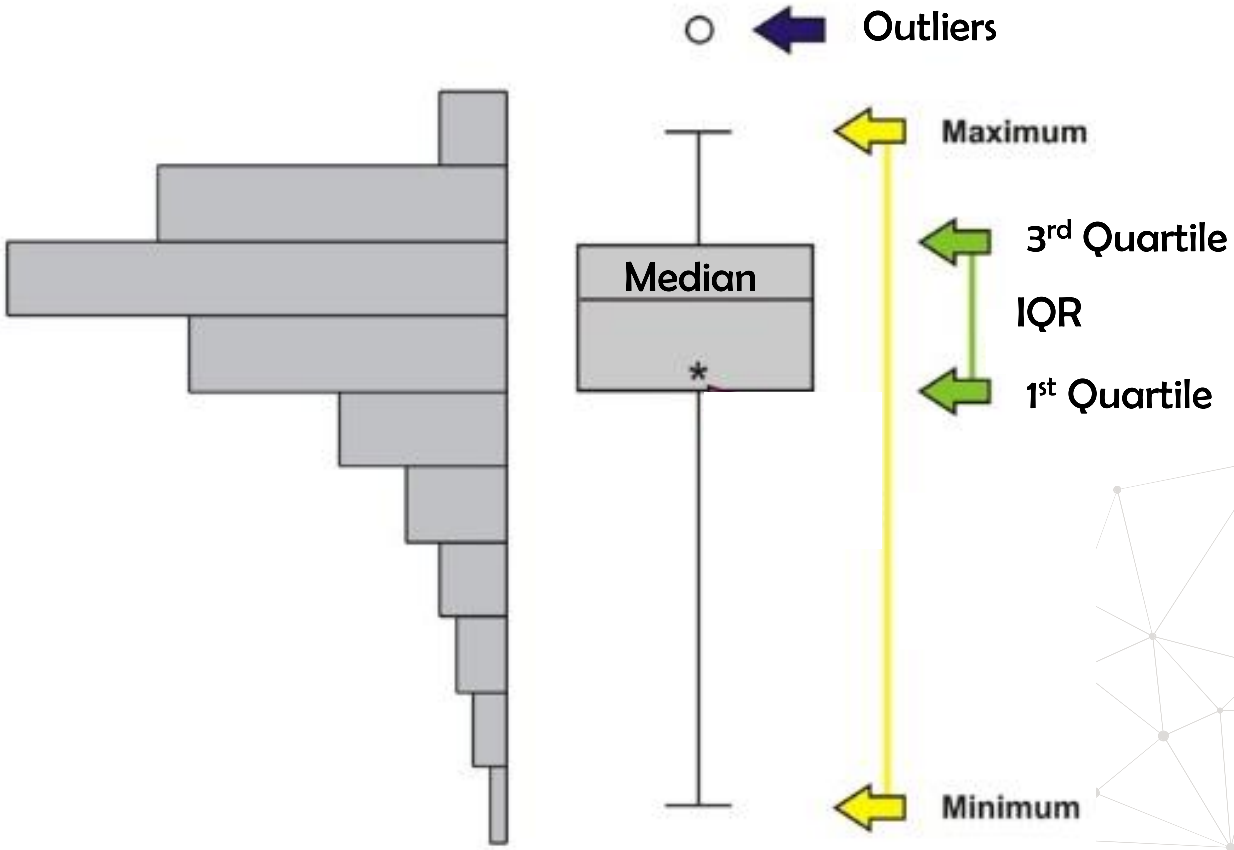




Images taken from pro.arcgis.com and wiki.commons

Visualising outliers - Boxplots





Images taken from pro.arcgis.com and wiki.commons

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook
- Extreme Value Analysis to detect outliers in normal and skewed variables in 2 different datasets

THANK YOU

www.trainindata.com