



# Outliers

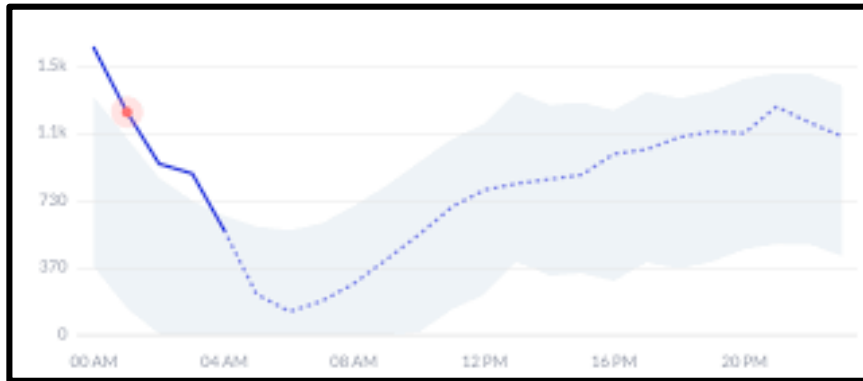


# Outliers

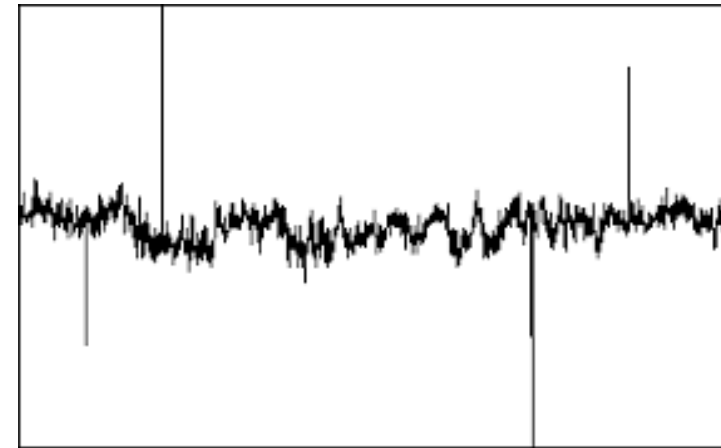
- An outlier is a data point which is significantly different from the remaining data.
- “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” [D. Hawkins. Identification of Outliers, Chapman and Hall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

# Should outliers be removed?

Revenue forecasting



Credit card transactions

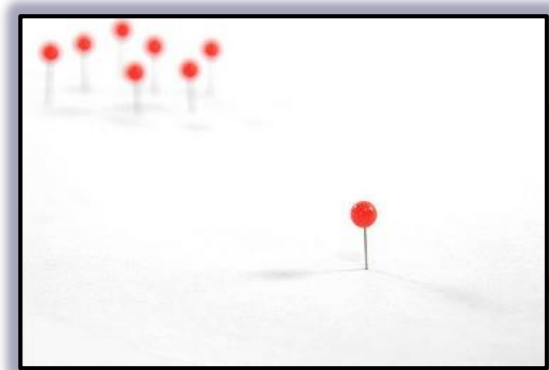


Depending on the context, outliers either deserve special attention or should be completely ignored.

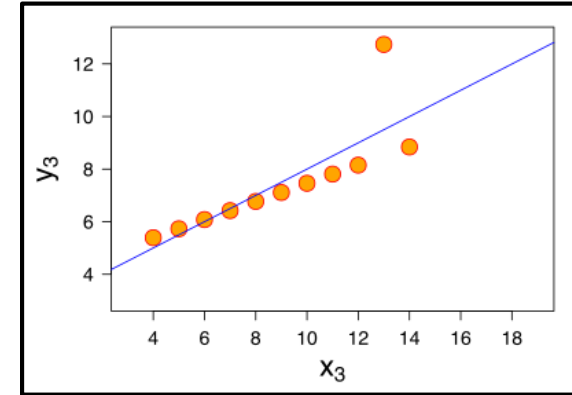
# Approach to outliers in this course

- 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

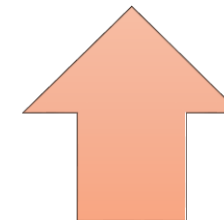
# Algorithms susceptible to outliers



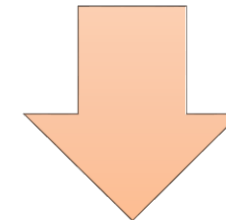
Linear  
models



Adaboost



Tremendous  
weights



Bad  
generalisation

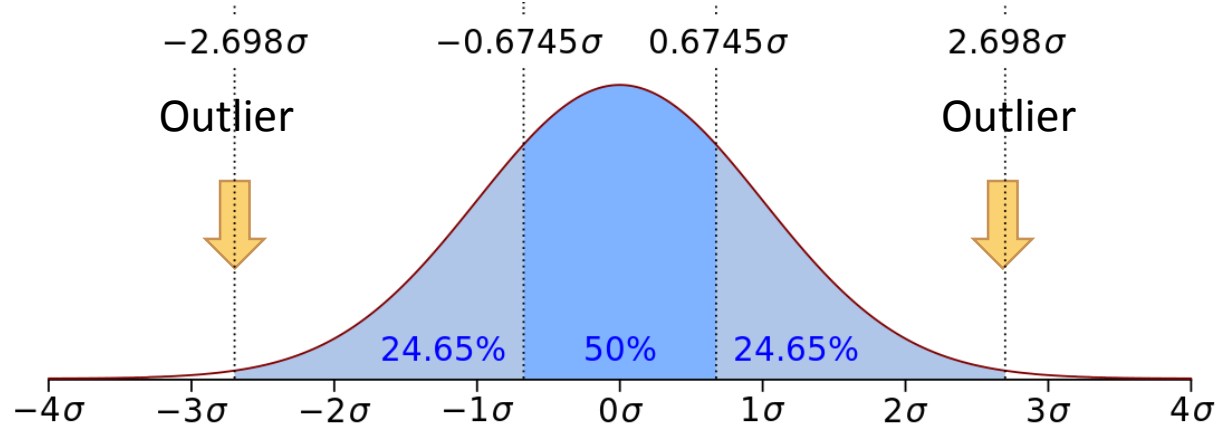


# Detecting Outliers

Extreme Value Analysis

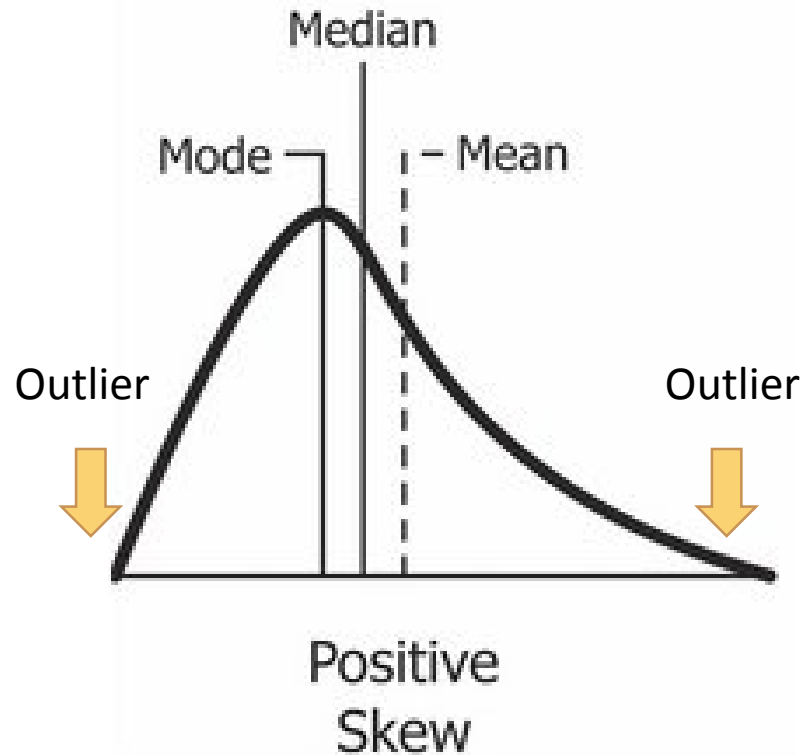


# Normal distribution



- ~99% of the observations of a normally distributed variable lie within the mean  $\pm 3 \times$  standard deviations.
- Values outside mean  $\pm 3 \times$  standard deviations are considered outliers

# Skewed distributions



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

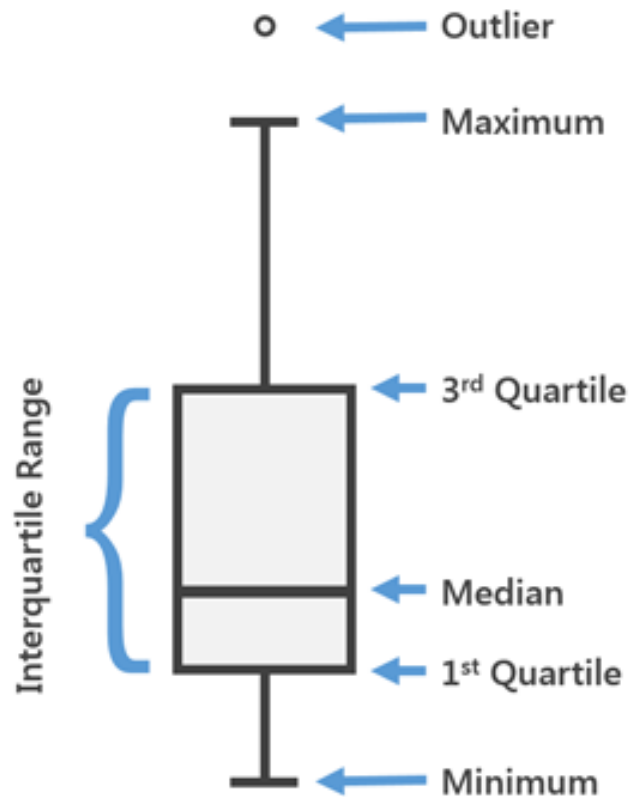
Note, for extreme outliers, multiply the IQR by 3 instead of 1.5



# Notes on quantiles

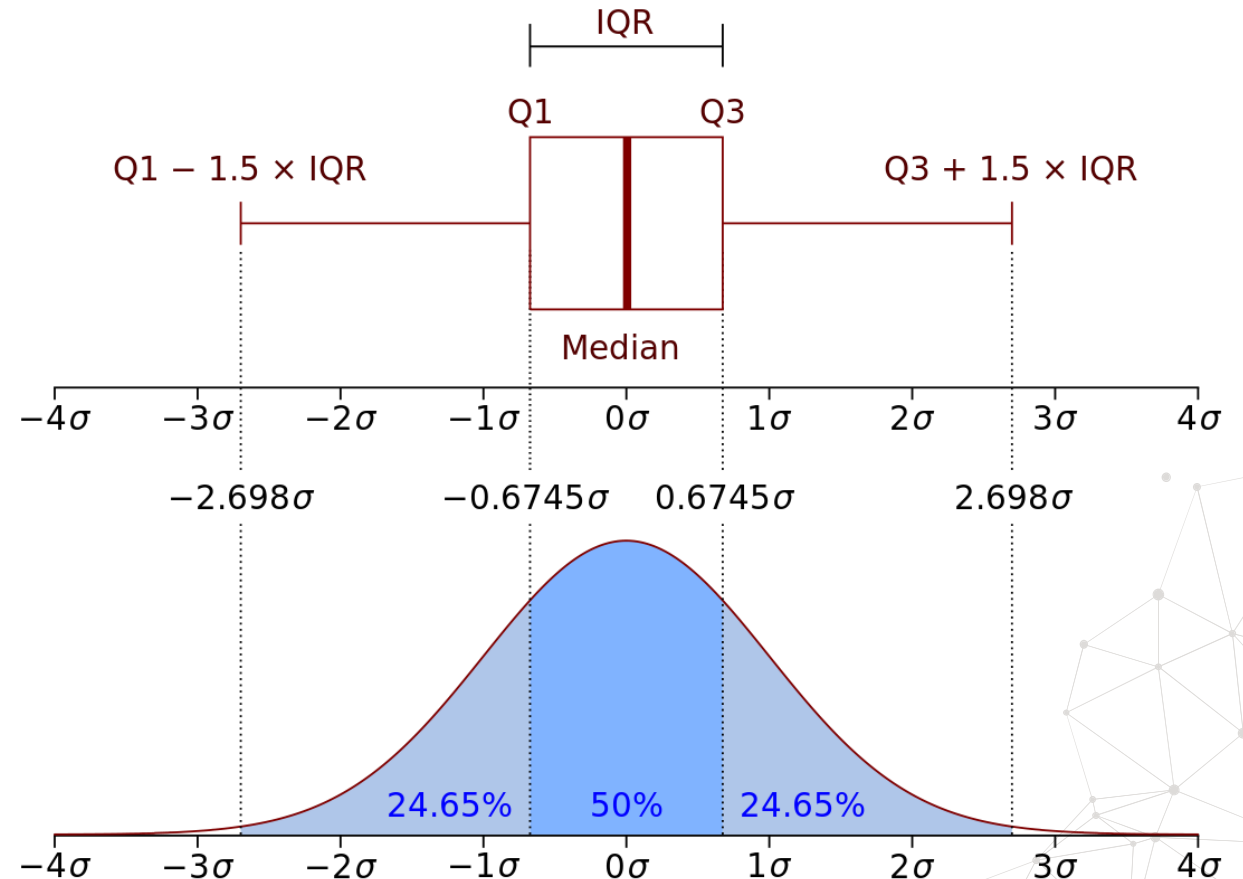
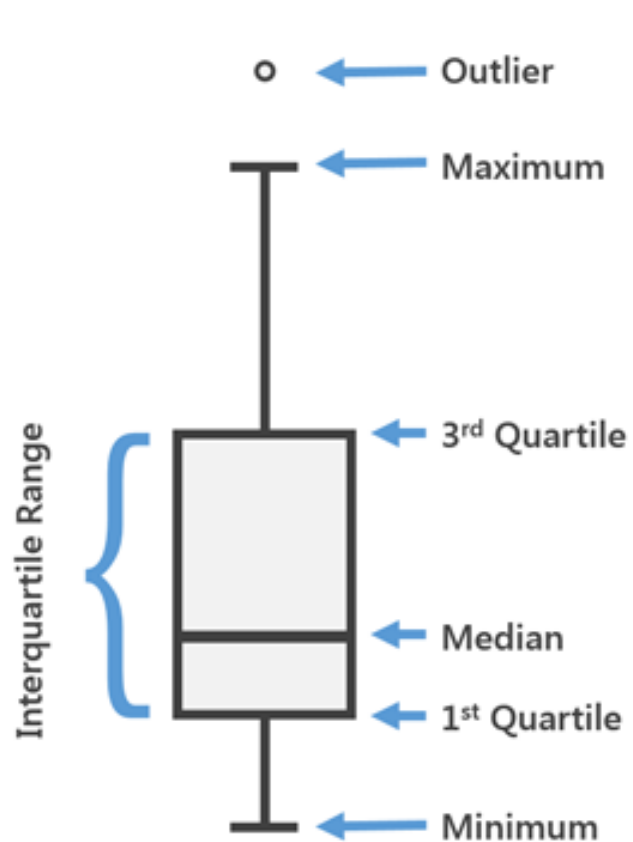
- Quartiles = dividing the distribution in 4
- Quantiles = dividing the distribution into 100
- 1<sup>st</sup> Quartile = 25<sup>th</sup> Quantile
- 3<sup>rd</sup> Quartile = 75<sup>th</sup> Quantile
- 2<sup>nd</sup> Quartile = 50<sup>th</sup> Quantile = Median
- $\text{IQR} = 75^{\text{th}} \text{ Quantile} - 25^{\text{th}} \text{ Quantile} = 3^{\text{rd}} \text{ Quartile} - 1^{\text{st}} \text{ Quartile}$

# Visualising outliers - Boxplots



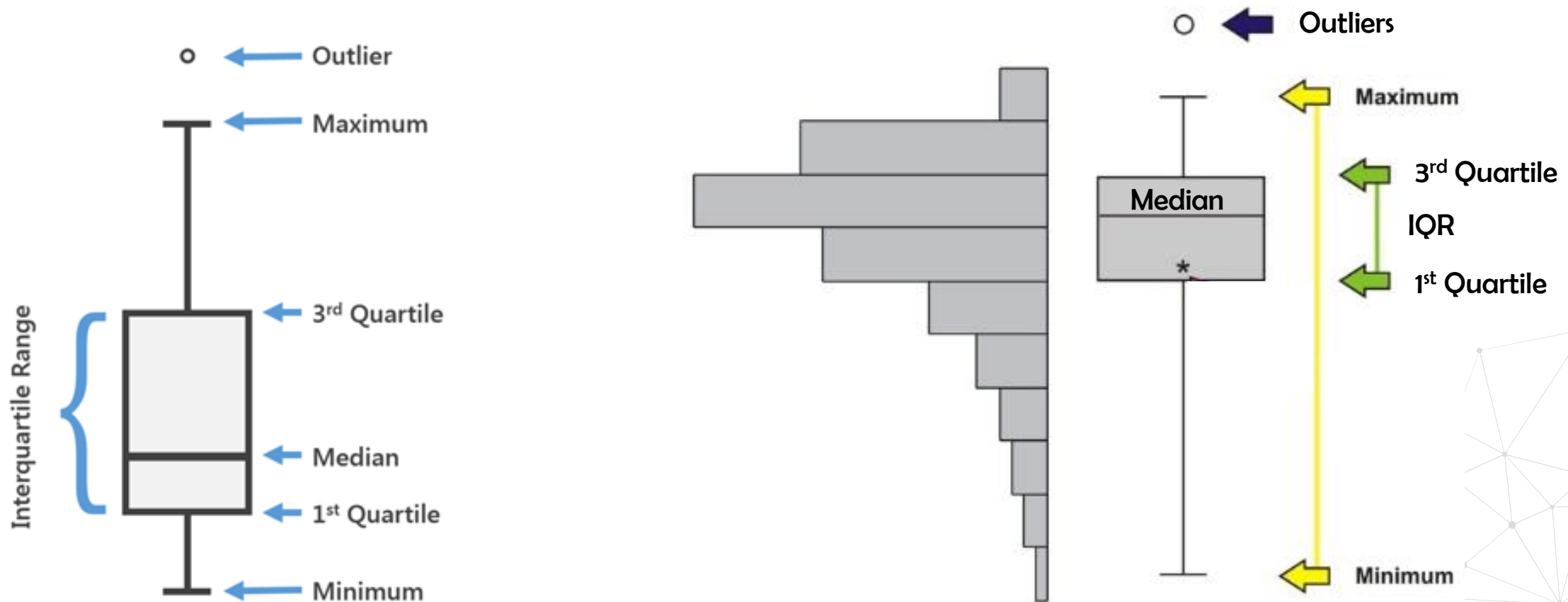
Images taken from [pro.arcgis.com](https://pro.arcgis.com) and [wiki.commons](https://wiki.commons)

# Visualising outliers - Boxplots



Images taken from [pro.arcgis.com](https://pro.arcgis.com) and [wiki.common](https://wiki.common)

# Visualising outliers - Boxplots



Images taken from [pro.arcgis.com](https://pro.arcgis.com) and [wiki.common](https://wiki.common)

# 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](http://www.trainindata.com)