UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Predictive modelling

Components of Predictive modelling

Predictive Model:

$$y = f(X_1, X_2, \dots, X_p)$$

- $X_1, X_2, \dots, X_p$: features (explanatory/independent variables)
  p: number of predictors
- y: target (label, response, or dependent variable)

In terms of dataset/observation, we denote vectors in boldface

- $\boldsymbol{X}$: an $n \times p$ matrix with observations as rows, $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})^{\mathrm{T}}$
  the i-th observation of the j-th predictor x_{ij}, $i = 1 \dots n, j = 1 \dots p$
- $\boldsymbol{y}$: an $n \times 1$ vector

The function f is what we want to learn.
Use available dataset $\{\boldsymbol{X}, \boldsymbol{y}\}$ to learn the relationship f then use f to predict $\hat{y}_i$ given $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip})$.

Edinburgh visitor information			
Index	**Age**	**Nationality**	**Visited?**
1	24	Scottish	1
2	56	English	0
3	42	Scottish	1
4	43	Welsh	1
5	77	English	0

?

Index	**Age**	**Nationality**	**Visited**
1	78	Scottish	?
2	24	English	?

What f might be?

- Rule 1: English & >50 ➔ not visit
- Rule 2: English || > 50 ➔ not visit

Example: Predicting visitor count

Quiz: Predictive modelling

Q1. What component is not part of the predictive analytics process?

- A. Formulating the business questions.
- B. Building the predictive model.
- C. Interpreting the result.
- D. They are all part of the process.

Q2. What is the ideal outcome of a predictive process?

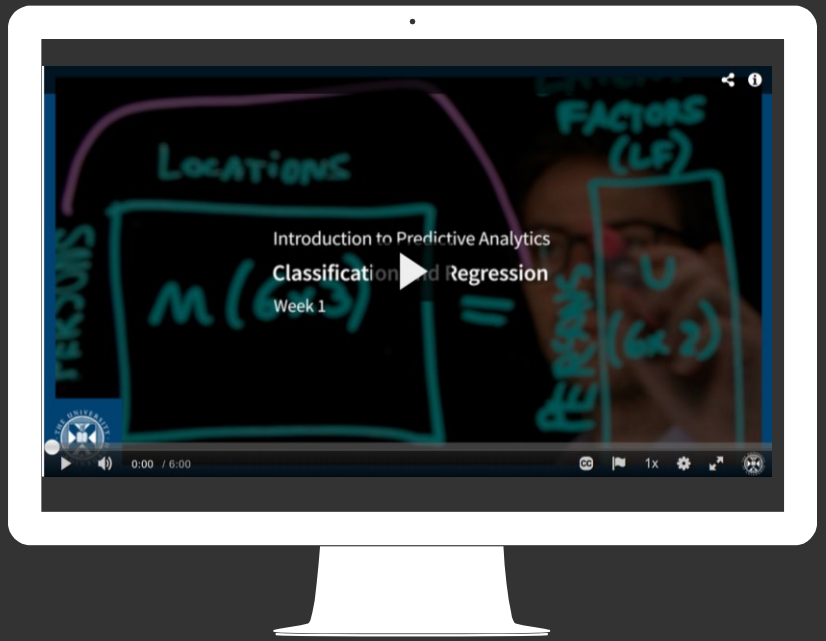
- A. A predictive model.
- B. Decision rules.
- C. An interpretation the results.
- D. A predictive model and its decision rules
- E. None of the above.

Q2. Decision rules outputted by a predictive model need to be...

- A. Understandable or accurate
- B. At least understandable
- C. At least accurate
- D. None of the above.

Classification and Regression

Classification and Regression



- Please watch the following video via link https://media.ed.ac.uk/media/Classification+and+Regression/1_m7frvdk9/117185481

Share your experiences on discussion board

Classification and regression: discussion

- Find the three most prevalent applications for regression, classification, and time series analysis by doing some research online.

Defining Your Problem and Variables

The different types of VAR

The weather condition and air pressure of the last 3 hours captured in the following table:

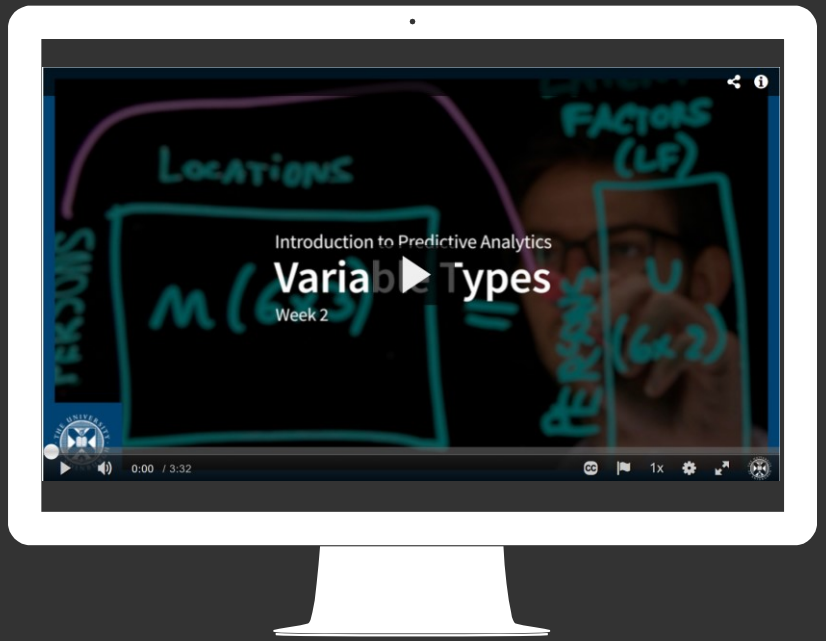
Nominal or Categorical

Weather condition	Air pressure last 3 hours(hPa)
Sunny	1,007
Cloudy	1,002
Heavy rain	960
Sunny	1,020

Ordinal (natural order)

- Continuous variable: Any value (in an interval), e.g. real numbers $\mathcal{R}$
  continuous y ➔ Regression
- Discrete variable: Finite number of possible outcomes, natural (counting) numbers $\mathcal{N}$
  discrete y ➔ Classification

Variable types



- Please watch the video through this link

https://media.ed.ac.uk/media/Variable+Types/1_q68bphal/117185481

Quiz: Defining Your Problem and Variables

Q1. What is the most appropriate technique for identifying the level of product stuck?

- A. Binary classification
- B. Regression
- C. Time series analysis
- D. Multiclass classification

Q2. What is the most appropriate technique for identifying someone's province/county of living?

- A. Binary classification
- B. Regression
- C. Time series analysis
- D. Multiclass classification

Q3. What is the most appropriate technique for identifying the GDP of a country?

- A. Binary classification
- B. Regression
- C. Time series analysis
- D. Multiclass classification

Q4. What is the most appropriate technique for identifying the attendance of an event?

- A. Binary classification
- B. Regression
- C. Time series analysis
- D. Multiclass classification

Summary Statistics

Summary statistics:

Sample **Mean (average):**

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\overline{x}$ measures the central tendency

Sample **Median** $\tilde{x}_{0.5}$:

$$\tilde{x}_{0.5} = \begin{cases} x_{\frac{n+1}{2}}, n\ is\ odd \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), n\ is\ even \end{cases}$$

And the 20%, 75% quantiles $\tilde{x}_{0.25}$, $\tilde{x}_{0.75}$

Sample **Variance**:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

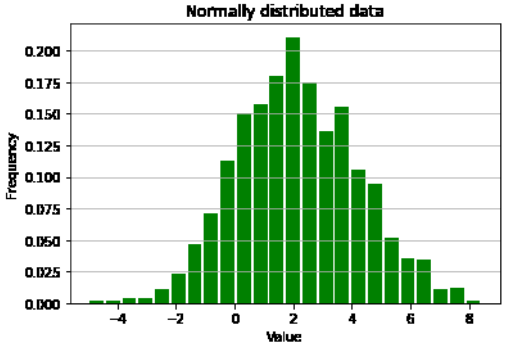
Sample Standard Deviation: s

s^2 measures the spread of variable values around the mean.

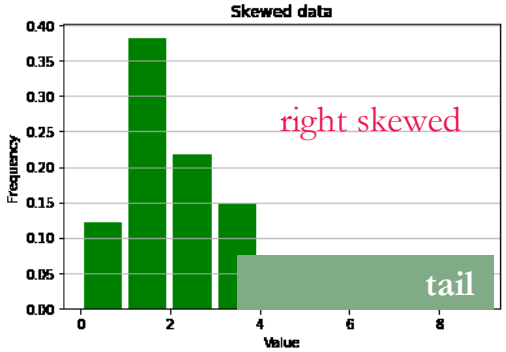
- Min $x^{\min}$ and Max$x^{\max}$
- Range: $x^{\max} - x^{\min}$
- Interquartile range: $\tilde{x}_{0.75} - \tilde{x}_{0.25}$

Frequency Table		
Value	Frequency	Proportion
Class A	30	0.3
Class B	50	0.5
…	…	

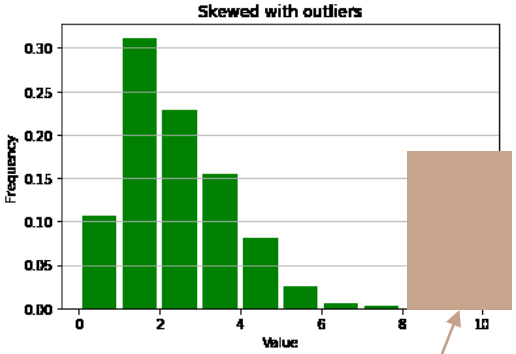
Summary statistics: histogram



Mean: 2.148 Median: 2.0339

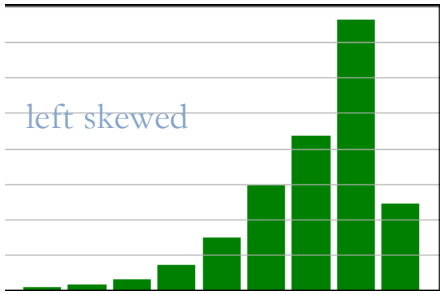


Mean: 1.88 Median: 1.0



Mean: 2.488 Median: 2.0

Possible outliers



Quiz: Summary Statistics

Q1. What summary statistic works best for getting a grasp on the central tendency of skewed distributions?

- A. Median
- B. Mean
- C. Range
- D. Interquartile range

Quiz: Summary Statistics

Q2.

Series 1= [9, 2, 1, 10, 5, 2 ,2, 8, 6, 2]

Series 2 = [8 ,8 ,1, 1, 2, 2, 2, 1, 1, 8]

Based on these two series, which of the following statements is true?

- A. 1 has a higher median, but lower mean than 2.
- B. 1 has a higher median, and higher mean than 2.
- C. 1 has a lower median, but higher mean than 2.
- D. 1 has a lower median, and lower mean than 2.
- E. None of these statements are true.

Quiz: Summary Statistics

Q2.

Series 1= [9, 2, 1, 10, 5, 2 ,2, 8, 6, 2]

Series 2 = [8 ,8 ,1, 1, 2, 2, 2, 1, 1, 8]

Based on these two series, which of the following statements is true?

- A. 1 has a higher standard deviation, but lower interquartile range than 2.
- B. 1 has a higher standard deviation, and higher interquartile range than 2.
- C. 1 has a lower standard deviation, but higher interquartile range than 2.
- D. 1 has a lower standard deviation, and lower interquartile range than 2.
- E. None of these statements are true.

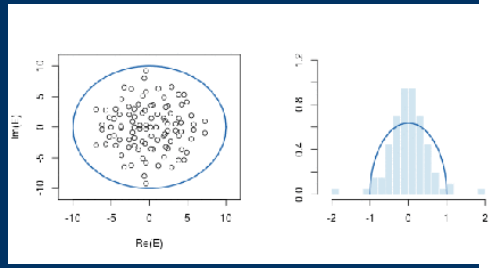
- Study the file
  5 - datasets_and_summary_statistics_reading.ipynb
- Try the following exercise
  6 - calculating_extra_summary_statistics_activity3.ipynb

Activity: Reading datasets and calculating summary statistics

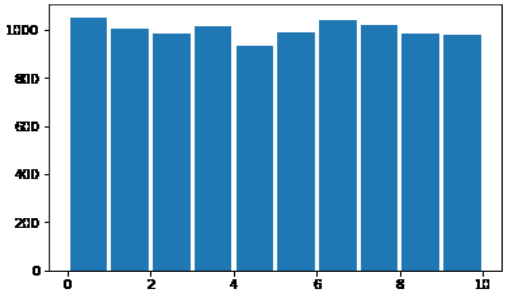
Plot from https://mathematica.stackexchange.com/questions/75711

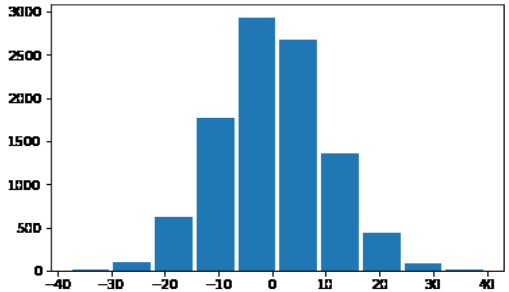
Distributions in datasets

Common distributions



Distributions in datasets





Uniform distribution U(a,b), pdf

$$P(x) = \begin{cases} 0, & x < a \\ \dfrac{1}{b-a}, & a \le x \le b \\ 0, & x > b \end{cases}$$

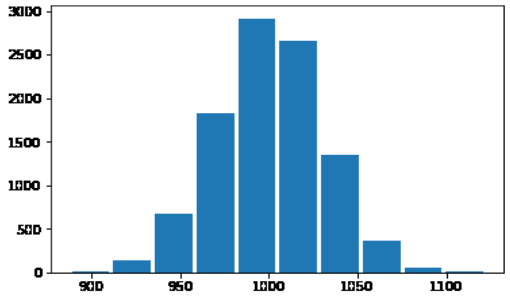
(Standard) normal distribution $N(\mu, \sigma)$, pdf

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where μ denotes means, σ denotes standard deviation

For standard normal, μ=0, σ=1

Distributions in datasets

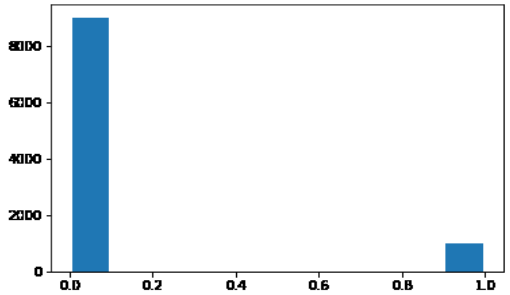


The binomial distribution $B(n, p)$, pmf

$$P(x = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

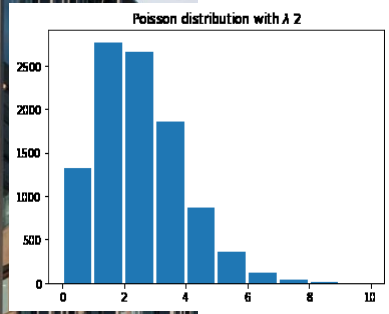
p the chance of success; n the number of draw

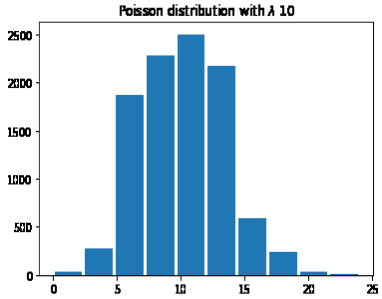
When n large, approximate to normal around np



The Bernoulli distribution is a special case of binomial distribution with $n = 1$

Distributions in datasets

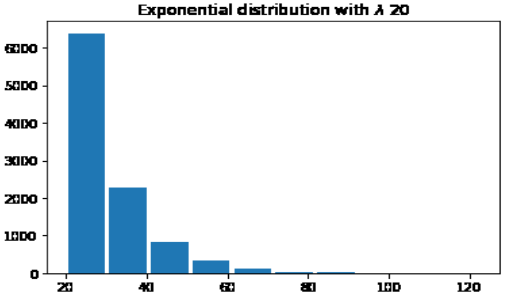




The Poisson distribution $\mathrm{POIS}(\lambda)$, pmf

$$P(x = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

- the number of independent events occurring within a time frame, who happen at a fixed rate λ (for example, 10/hour)
- $P(x = k)$ represents the chance of k events happening.
- For a sufficiently large number of observations and a λ larger than 10, the Poisson distribution looks like a normal distribution.



The exponential distribution $\mathrm{EXP}(\lambda)$

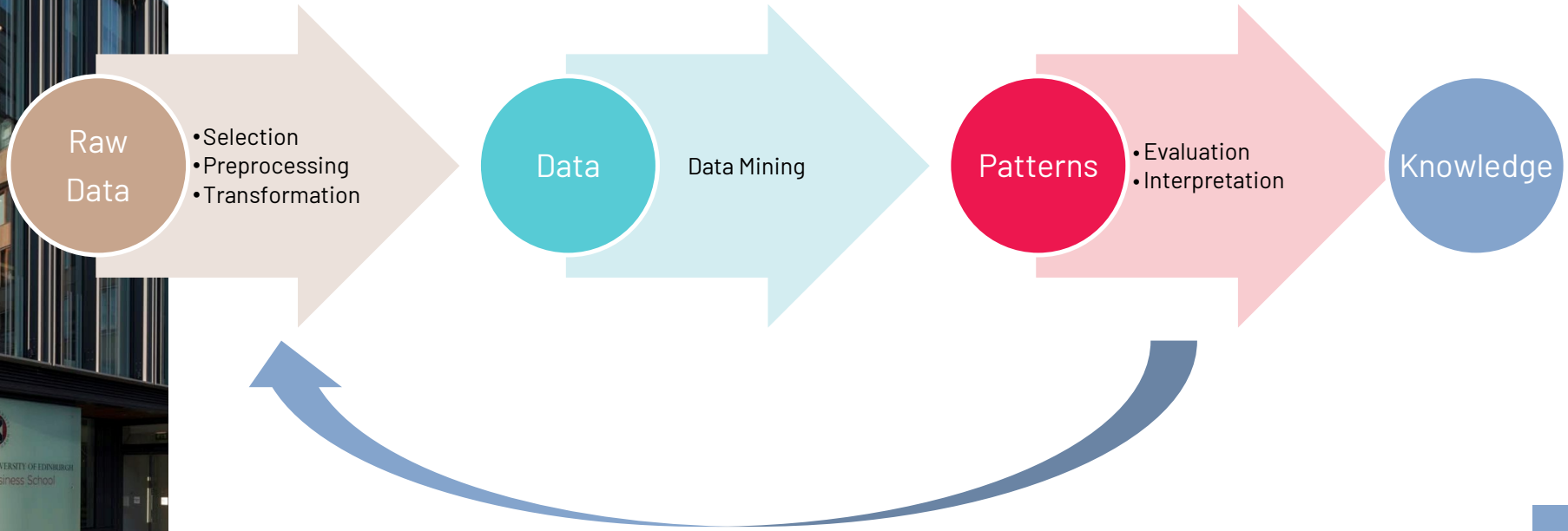
$$P(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, x \geq 0 \\ 0, \quad x < 0 \end{cases}$$

The waiting time between independent events, average $\frac{1}{\lambda}$

Others: the Gamma and Beta distributions, Weibull distribution…

The Modelling Process

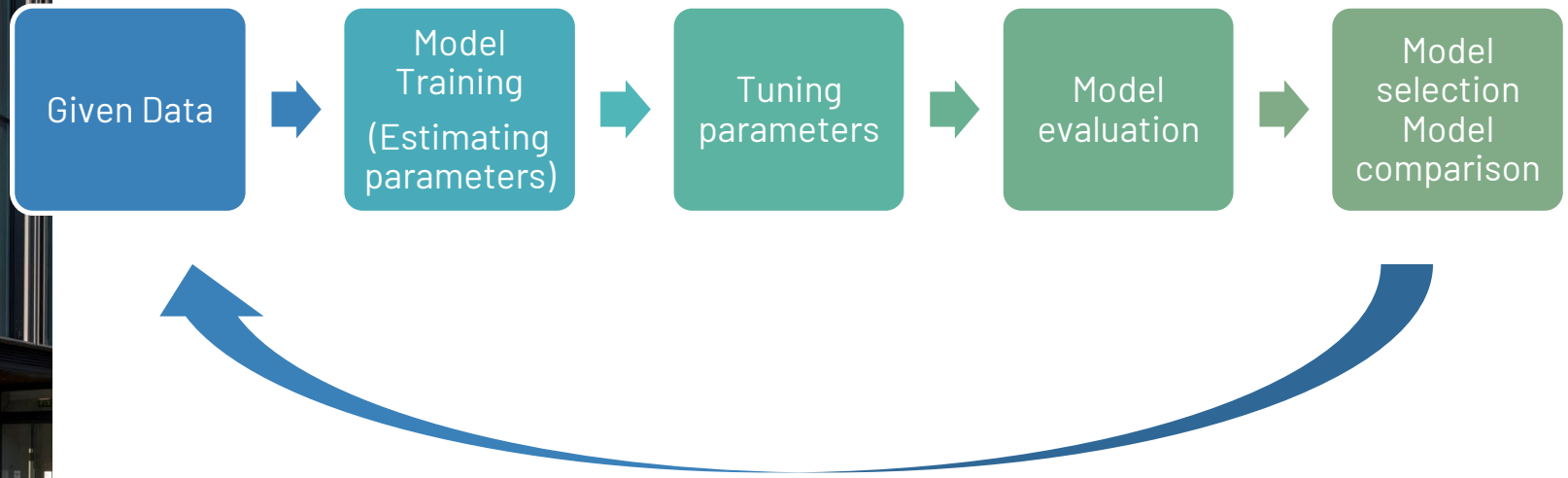
A general data mining process



For more details, read the following article about Knowledge Discovery in Database (KDD)

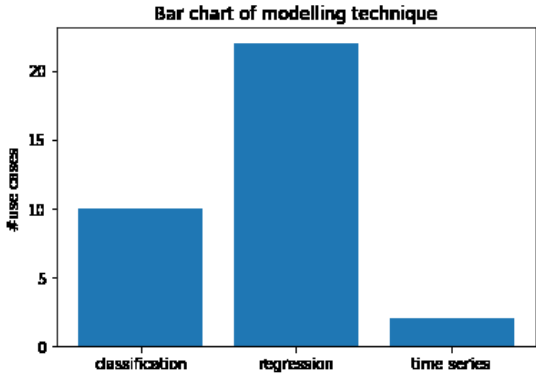
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

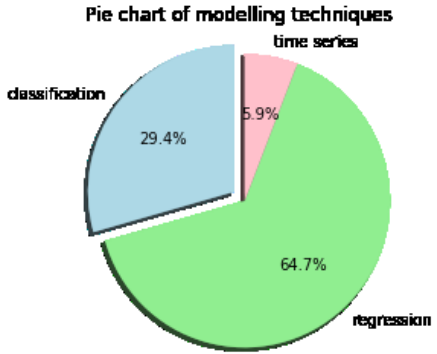
A general modelling process

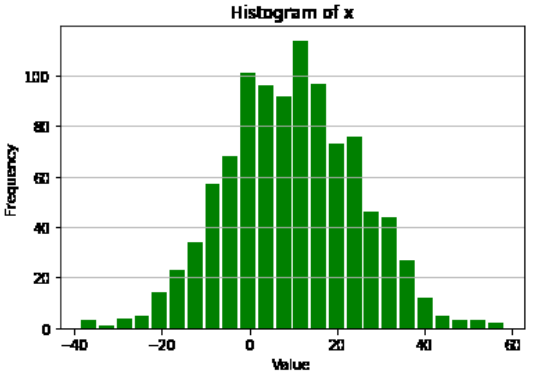


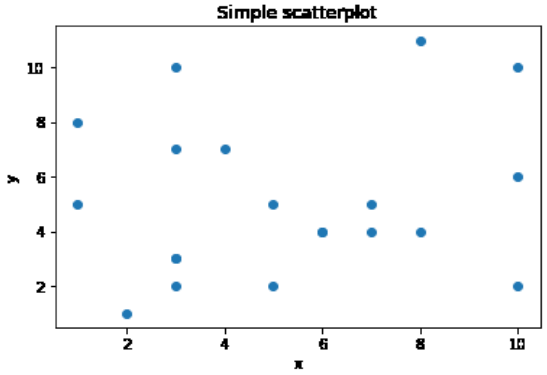
Visual VAR and Coding

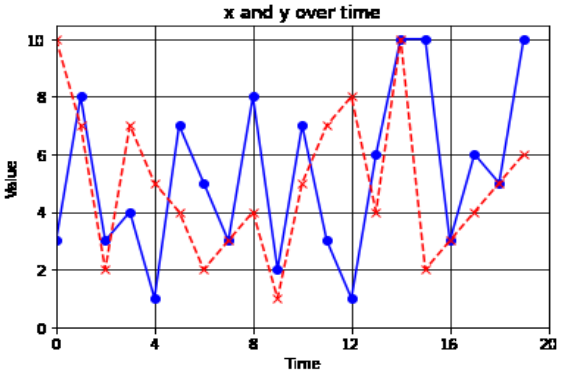
The importance of visualisation





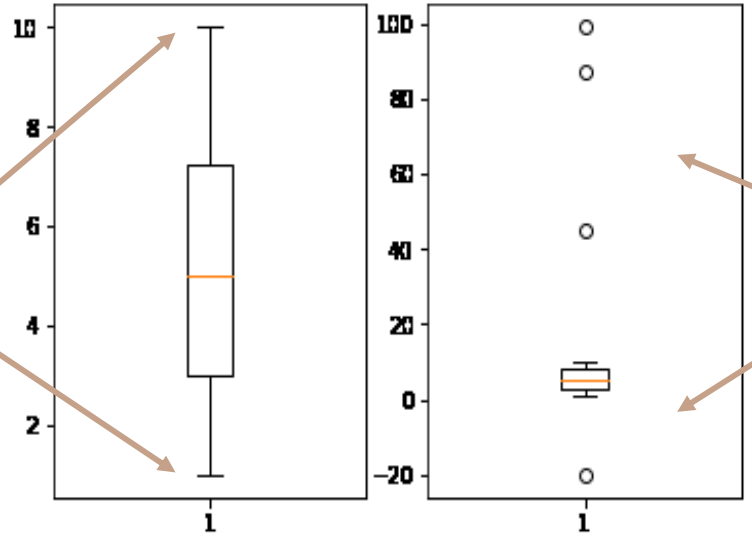






The importance of visualisation

Boxplots



Whiskers:
1.5 interquartile range

Outliers are identified as observations lying beyond the whiskers

Quiz: Visual VAR and Coding

How can the shape of the variables be important for building predictive modelling?

- A. You need to know the shape of the dependent variable's distribution to build a model.
- B. You need to know the shape of the independent variable(s') distribution to build a model.
- C. You need to know the shape of both the independent and dependent variables' distribution to build a model.
- D. The need to understand the shape of the variables' distribution will depend on the algorithm.

- **Matplotlib (plt), Numpy (np) and Pandas (pd)**

- **Study files**
  7 - visualisation_types1_reading.ipynb
- 8 - visualisation_types2_reading.ipynb
- **Try the following exercise (optional)**
  9 - obtaining_insights_into_the_housing_dataset_activity4.ipynb

- Other Python packages: Seaborn, ggplot2, Bokeh, Plotly

Activity: Visualisation and insights

UNIVERSITY OF EDINBURGH
Business School