POWERING POTENTIAL

L3 Data Essentials

- Module 4
- Statistics for Decision-Making
- 3-day Live event learning

Overview of 3-day learning

Module 4, revisit online content, knowledge check, task activities, and review.

DAY 2

Statistics and Data Modelling

SCHEDULE: DAY 2

AM

- Inferential vs. descriptive statistics
- Statistics in business
- Measures of central tendency + Activity
- Measures of variation + Activity
- What to do about outliers? + Activity

PM

- Correlation + Activity
- Linear regression + Activity
- Activity: GP practice scenario

Day 2 learning

DAY 1 RECAP: ALGORITHMS

How much can you remember?

1. How would you define an algorithm?
2. What are the benefits and disadvantages of using automated algorithms in business?
3. What do Search and Sort algorithms do?
4. What are the main types of algorithms discussed?

STATISTICS

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numeric data.

- **Descriptive statistics:** describes or summarises the characteristics of a dataset.

  e.g., sums and counts, mean, standard deviation, visualisations.

- **Inferential statistics:** makes inferences or predictions about a dataset based on a representative sample.

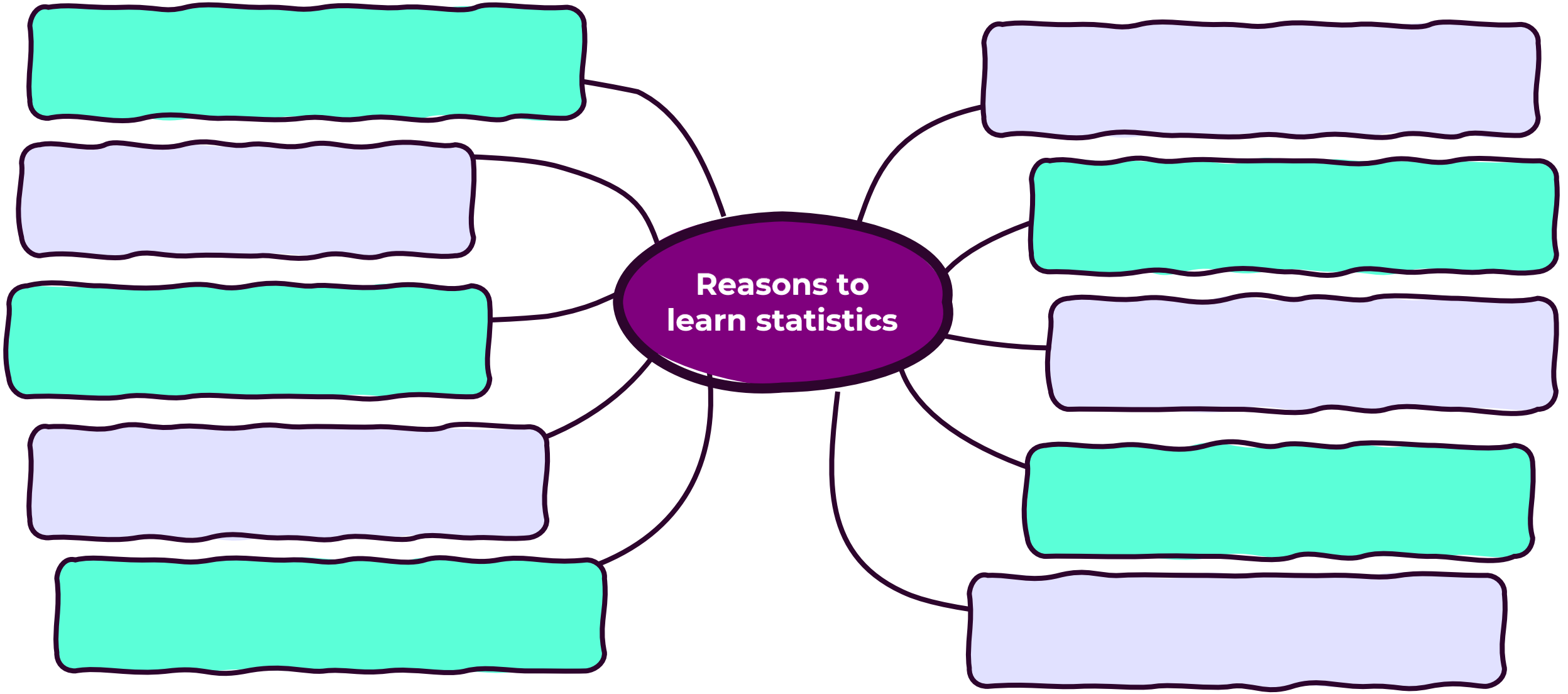
  e.g., linear regression, time-series forecasting, hypothesis testing.

STATISTICS IN BUSINESS

Statistics can be used to understand datasets and make sense of data, to inform and guide the organisation's strategy, and in general, aid day-to-day operational and strategic decision making.

Reflect on what the main reasons to learn and use statistics in your organisation and industry are.

‘It’s easy to lie with statistics. It’s hard to tell the truth without statistics.’ — Andrejs Dunkels



Activity: How businesses use statistics

https://thebossmagazine.com/how-businesses-use-statistics/

- Read through the online article.
- Produce notes on how statistics can be used to improve operations, marketing, finances, and support trend analysis.
- Discuss your findings in breakout sessions and feedback to the group.

Q: ‘Produce notes on how statistics can be used to improve operations, marketing, finances, and support trend analysis.’

Operations

Marketing

Finances

Trend analysis

MEAN, MEDIAN, AND MODE

Measures of central tendency

Mean: The ‘average’ value of the list. All values added and divided by the number of items.

Median: The number in the middle if all the values were to be laid out in natural order.

Mode: The most popular value in the dataset.

MEASURES OF CENTRAL TENDENCY

e.g.,
2, 5, 4, 5, 3

Mean

$\bar{x}$ = (2+5+4+5+3)/5 = 3.8

Mode

	Frequency
2:	1
3:	1
4:	1
5:	2

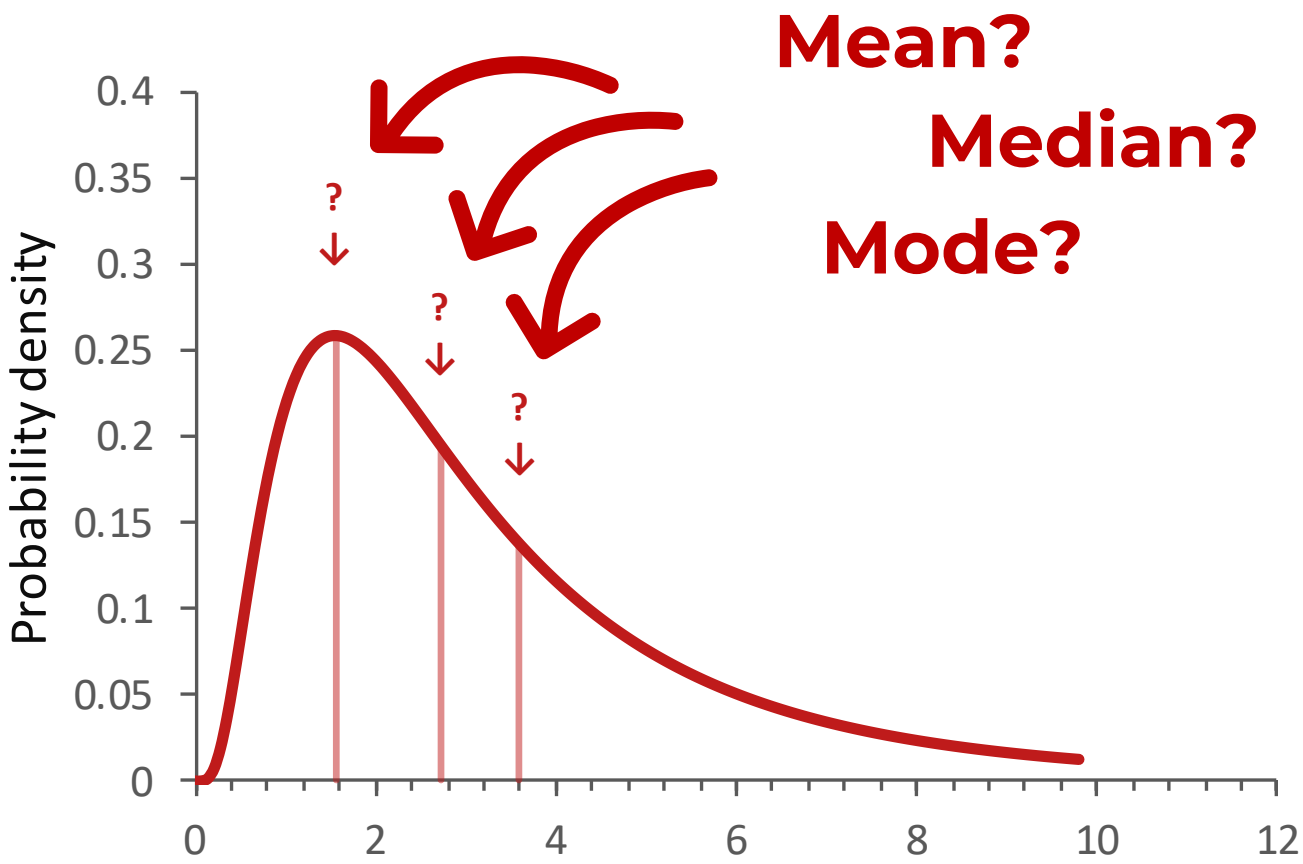
Median

2, 3, 4, 5, 5

Mid-range

(2+5)/2 = 3.5

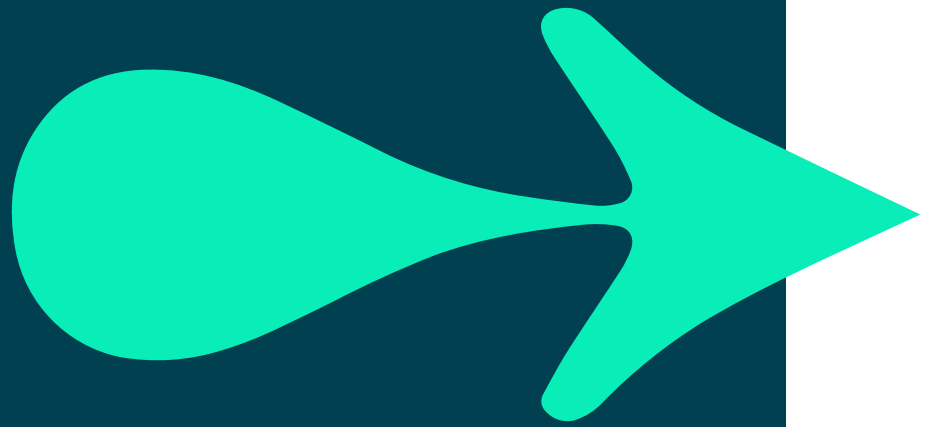
ACTIVITY: MEAN, MEDIAN, AND MODE



Open file

'Mean, Median, and Mode v{#.#.#}.xlsx'

EXERCISE AND DISCUSSION: WHICH MEASURES TO USE?



Which measure would best answer the question **‘What’s a typical value’** in each of these scenarios?

1. Test scores in a class
2. Eye colour in a survey of a population
3. Income of employees in a company
4. Age of participants in a marathon
5. Transportation used by people in a city
6. Height of individuals in a study
7. Products purchased in a store

MEASURES OF VARIATION:

RANGE AND IQR

e.g.,
2, 3, 4, 6, 7, 7, 8, 19

Range:

Width of spread for the whole data set (largest – smallest)

19 – 2 = 17

Inter-Quartile Range

Width of spread for the middle half of the data set.

Similar to finding the median, we find the first (or lower) quartile and the third (or upper) quartile before finding the difference.

IQR = Q3 – Q1
= 7.5 – 3.5
= 4

MEASURES OF VARIATION:

STANDARD DEVIATION AND VARIANCE (POPULATION)

e.g.,
2, 3, 4, 6, 7, 7, 8, 19

$x_i - \bar{x}$ is the distance from each piece of data to the mean.

Some distances are positive, and some are negative. Before calculating the average distance of the data from the mean, we need to deal with the signs. For standard deviation and variance this is done by squaring each of the distances first.

Variance, s^2, is the average of these squared distances from the mean.

Standard Deviation, s, is the result of completing the calculation by square rooting.

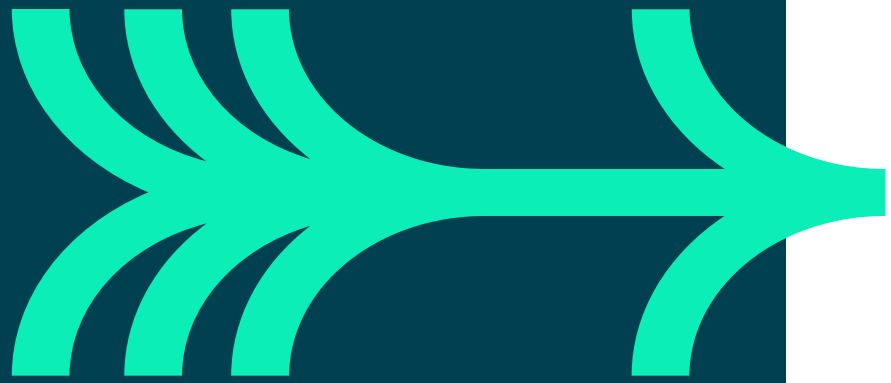
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = 24.5$$

$$s = 4.95$$

ACTIVITY: COMPARE TWO INVESTMENT STRATEGIES

1. View and analyse the data contained in the file investment_performance.csv .
2. The scores are results from two investment strategies – the control group (using the current method) and the experimental group (using a new investment strategy).
3. Compare one measure of the ‘typical’ value for the two groups and one measure of how consistent or spread out the data is within each group.
4. Report and discuss the results in class.

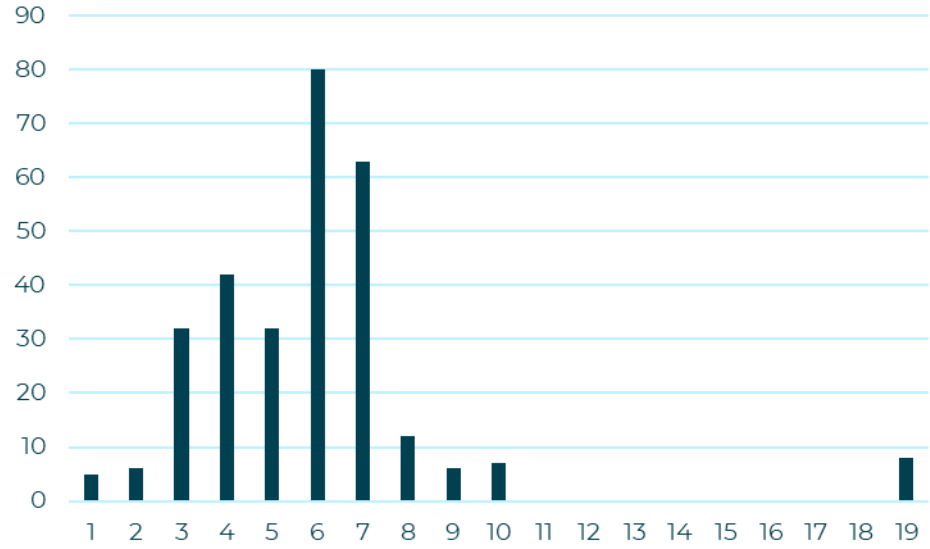
DISCUSSION: WHAT DO THESE SUMMARY STATISTICS TELL US?



```
Control Scores:
Mean: 78.08751663789478
Mode: 61.135907012612826
Median: 77.38615374332923
Mid-Range: 78.24753427263798
Range: 34.2232545200503
IQR: 15.577896589435682
Population Variance: 89.78049627551594
Population Standard Deviation: 9.4752705590703

Experimental Scores:
Mean: 85.31905514596525
Mode: 66.15394543335037
Median: 84.37171508463862
Mid-Range: 88.38986419909233
Range: 44.47183753148393
IQR: 17.49185513359525
Population Variance: 119.82727219584062
Population Standard Deviation: 10.946564401484176
```

OUTLIERS: KEEP, CORRECT, OR REMOVE?



The importance of investigating the outliers

Keep it, Correct it, or Remove it?

1) The values are ages of people at a party
2) The values are scores from a test out of 10

OUTLIERS: CALCULATED

e.g.,
2, 3, 4, 6, 7, 7, 8, 19

These two methods are designed, assuming the data is normally distributed, to create boundaries which capture 95% of the data – leaving just 5% to be investigated.

$$\bar{x} \pm 2s$$

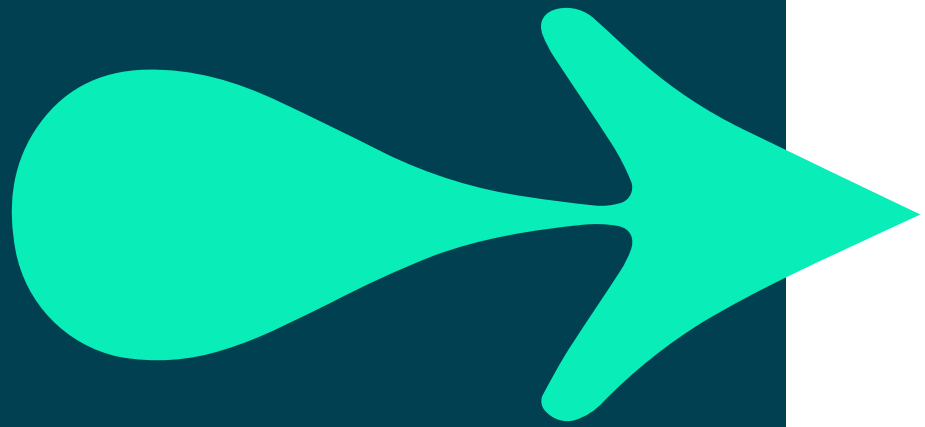
= -3.58, 17.6 => outlier: 19

$$Q_1 - 1.5IQR$$
$$Q_3 + 1.5IQR$$

= -2.5, 13.5 => outlier: 19

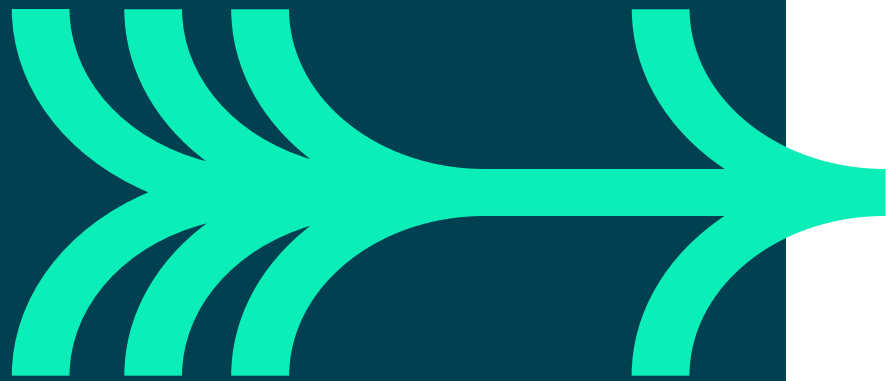
Either could be used to automatically detect or highlight outliers.

ACTIVITY ON FUNDS: OUTLIERS



- Select the Excel file returns.xls which describes the monthly returns of a basket of financial securities (mutual funds).
- Use the Excel function mean(), median() and stdev() to calculate the mean, median, and standard deviation of each time series.
- Calculate the outlier boundaries using mean – 2 * standard deviations and mean + 2 * standard deviations.
- Check if there are any outliers in your dataset (values that are smaller than the lower bound or larger than the upper bound).
- Compare those findings with differences between mean and median. Is there any correlation?

LEARNING CHECK



Think about your answers to these questions:

- What are the different ways of identifying a ‘typical’ value?
- What does standard deviation tell us?
- Once outliers are identified, what might we do with them?

Activity: Understanding statistic types

In your online learning content for Module 4, you were introduced to different types of statistics:

- **Correlation and Covariance**
- **Linear Regression**

Task

1. The class is split in break-out sessions to review and analyse the different statistics, and report back to the class a brief description of the same, supported by notes (i.e., one pager or short presentation).

2. Discuss as a class what is the use of each statistics and provide examples of practical use of the same, particularly and ideally within your working organisations.

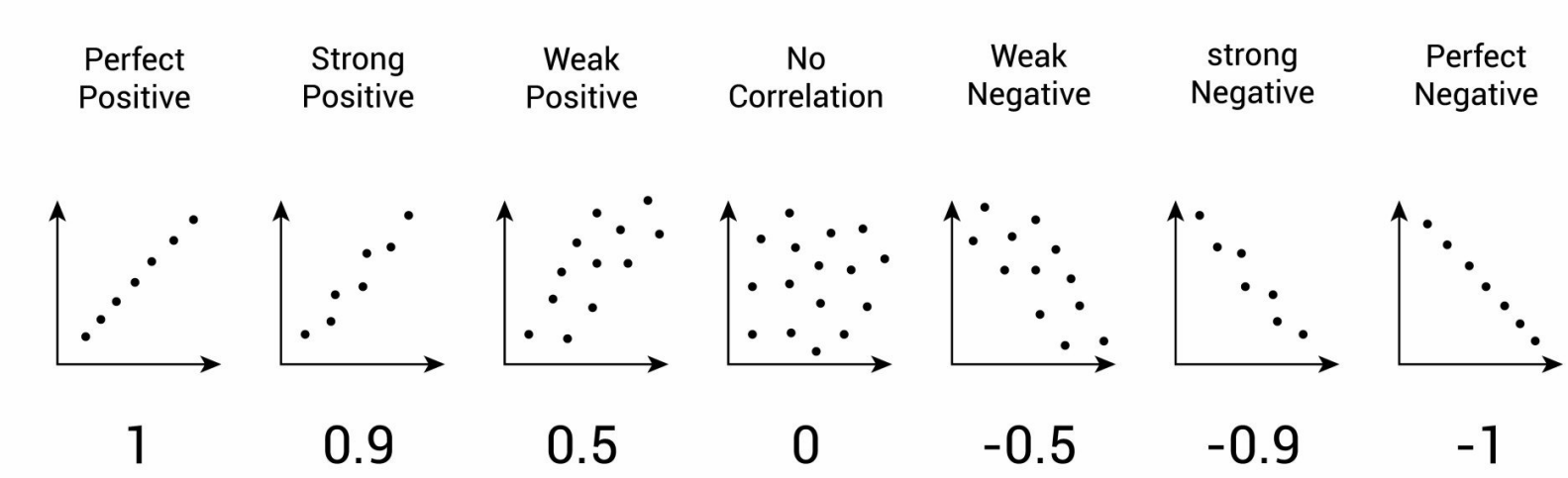
WHAT IS CORRELATION?

Correlation is a statistical measure that quantifies the strength and the direction of the linear relationship between two variables.

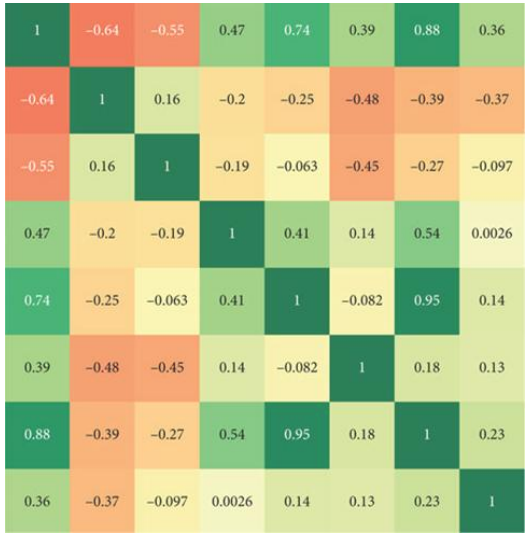
Correlation is used to determine whether changes in one variable are associated with changes in another variable.

Correlation does not imply causation. It only indicates if there is some form of relationship between variables

LET'S VISUALISE CORRELATIONS

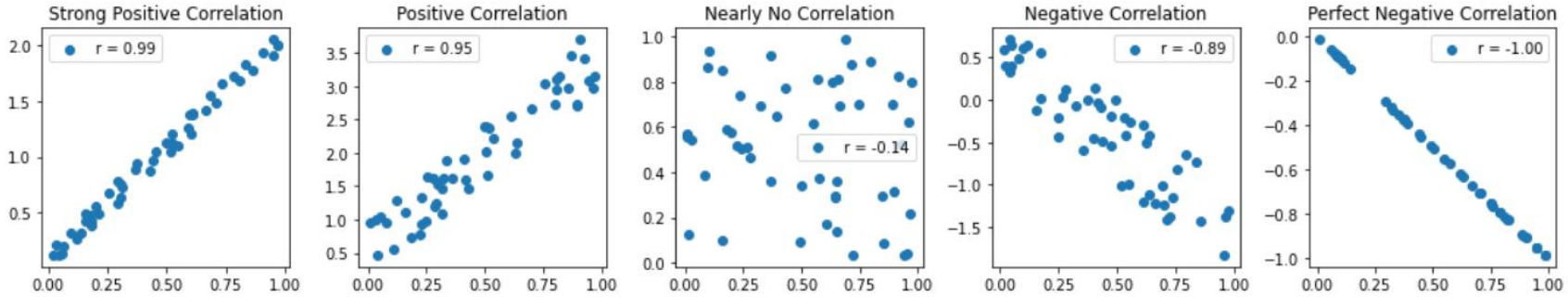


The correlation structure existing between a set of variables can be represented and visualised with a **correlation matrix**, where all the correlation values between couples of variables in the set are represented.



CORRELATION FROM SCATTER PLOTS

Checking what a scatter plot looks like is another way of checking for outliers or anomalies, but we also might be looking to see if a line of best fit (or regression line) might be appropriate.



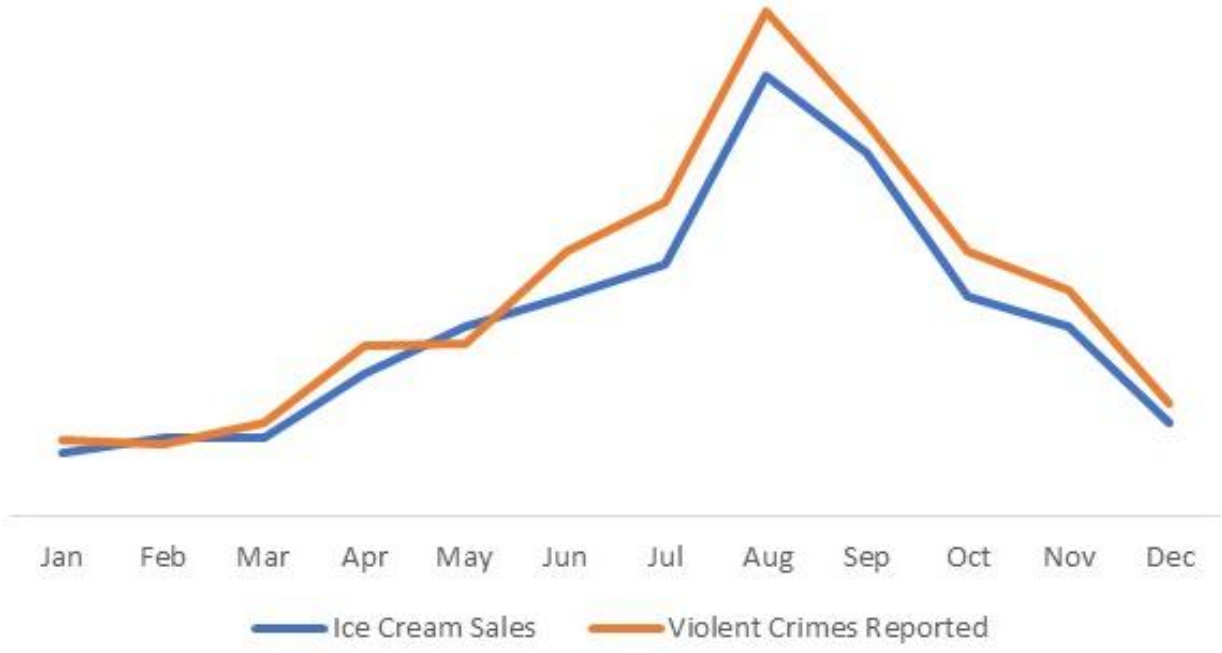
The more squashed the 'ball' of data is, the stronger the correlation.

Correlation or association can tell us if two data features are linked, but one hasn't necessarily caused the other – there may be a third unseen feature that is a **causal link**.

CORRELATION AND CAUSATION

Although the values of the two variables show a remarkable correlation, it is hard to think that ice cream sales have a direct impact on crime reported, or the other way round.

Possibly a common variable is influencing both? Or just an example of **spurious correlation**?



WHAT IS A LINEAR REGRESSION?

Linear regression is used to model the relationship existing between a dependent variable and one, or more, independent variables.

It aims to find a linear equation that best fits the data points to explain, or even to predict, the behaviour of the dependent variable, assuming that this is explainable with the independent variable(s).

In linear regression the relationship is modelled using a linear function, estimated from known values of the independent and the dependent variables

LINEAR REGRESSION MODEL AND VARIABLES

Simple Linear Regression

$$y = \beta_0 + \beta_1 x_1$$

β_0 - intercept

β_1 - slope

Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

y – dependent variable

$x_1, \ldots, x_n$ - independent variables

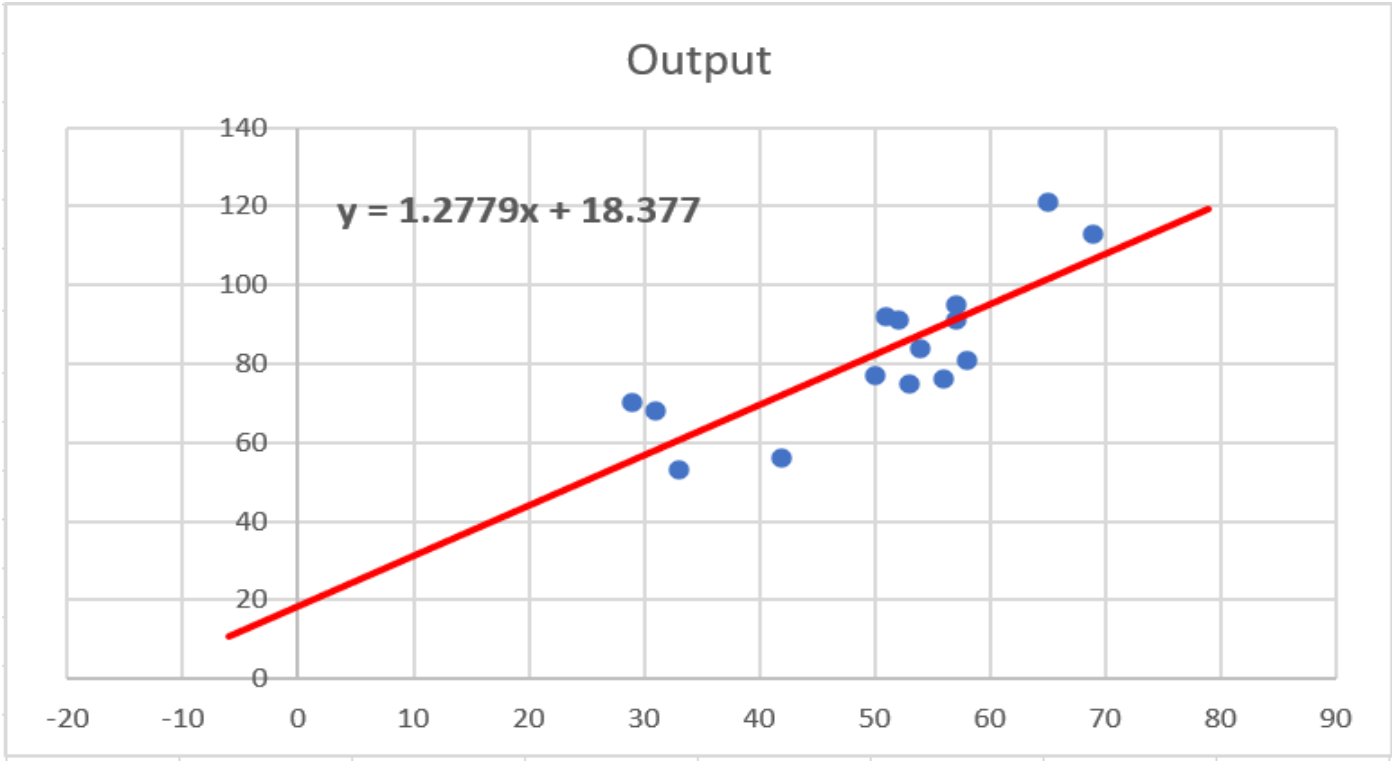
USE CASES FOR LINEAR REGRESSION

Linear regression is a widely used model and type of inferential statistics. Its application fits several situations of practical application

For example:

- Quarterly demand for new cars depending on level of unemployment, interest rates, and GDP
- Personal income depending on age, education, profession, and area of residence
- Demand for a product depending on its price and the prices of related products

THE ROLE OF THE MODEL COEFFICIENTS

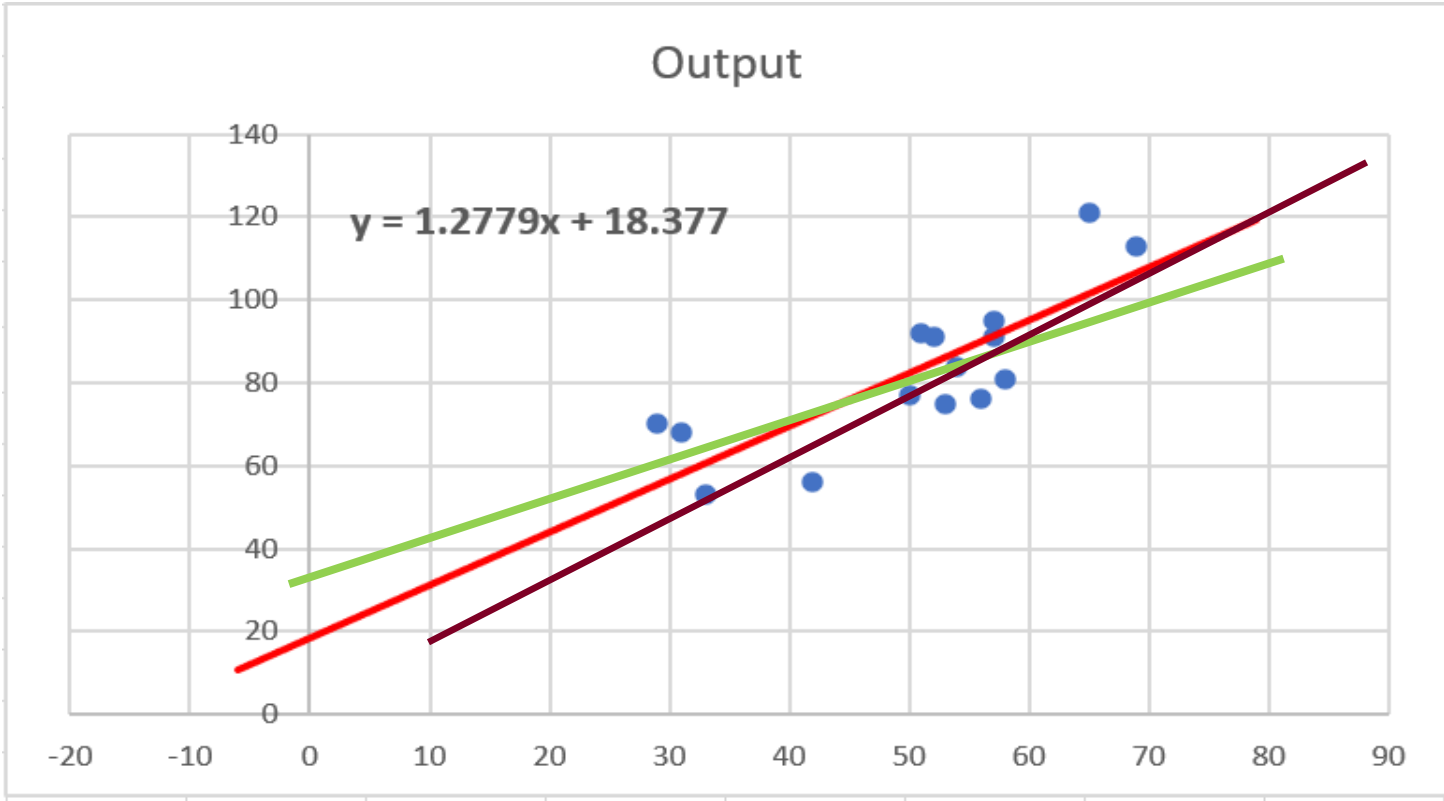


Intercept –the value of the dependent variable when the independent variables is zero.

Slope –the change of the dependent variable y when the independent variable for that slope increases by 1.

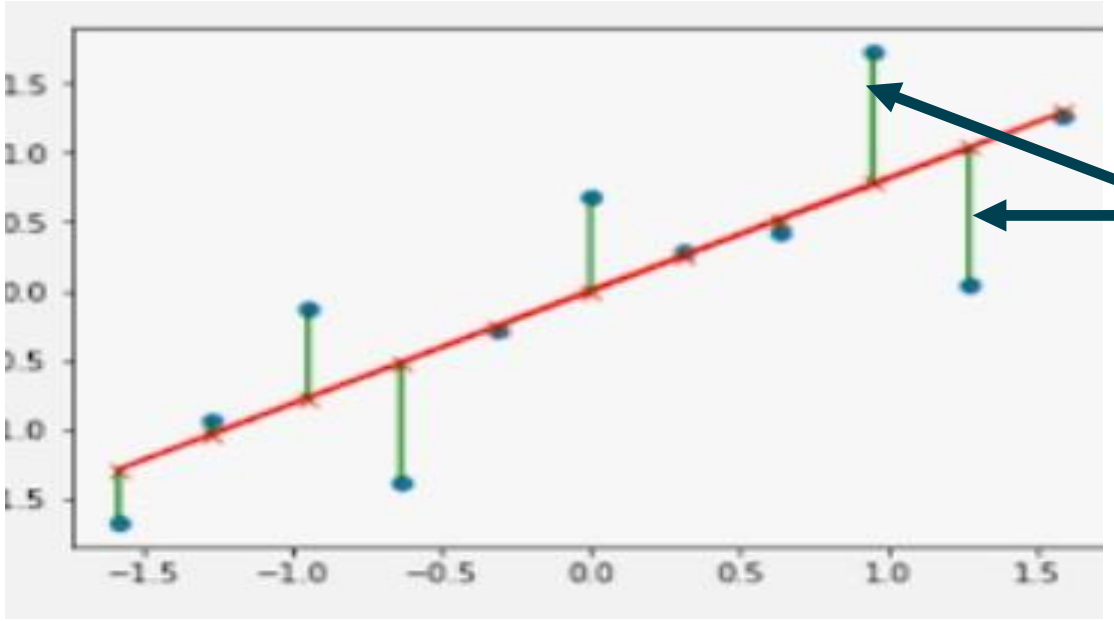
LINE OF BEST FIT

Why the red line and not any other?



RESIDUALS

No line passes through all data points.



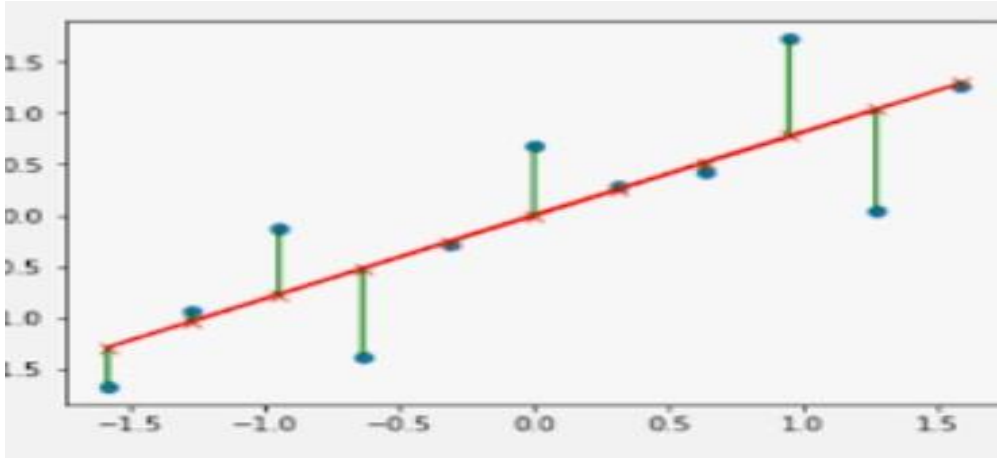
Residuals

Residual = actual – predicted by the model

We are looking for the line which is associated with the least total residuals.

ORDINARY LEAST SQUARES

What if we add the residuals for each possible line and compare?



The residuals can be positive or negative, and their sum will be around zero. Raising to the power two removes the sign.

Ordinary Least Squares (OLS)

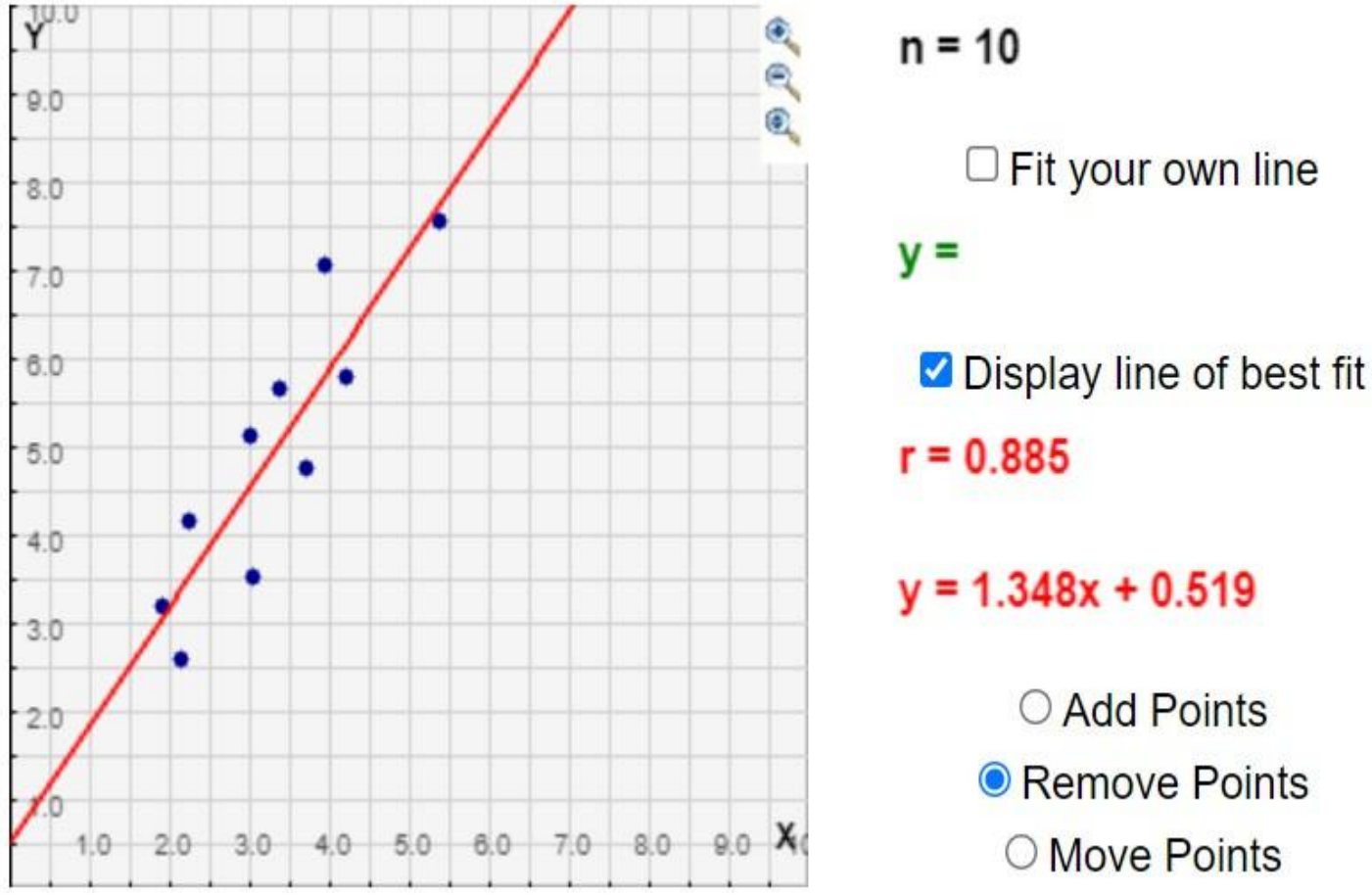
Find the values of $\beta_0, \beta_1, ..., \beta_n$ that minimise the sum of the squares of the residuals. These are the coefficients of the line of best fit.

THE EFFECT OF OUTLIERS

OLS aims to minimise the sum of squares of the residuals for all data points. Outliers worsens the fit.

Simple Linear Regression interactively
http://www.shodor.org/interactivate/activities/Regression/

No outliers



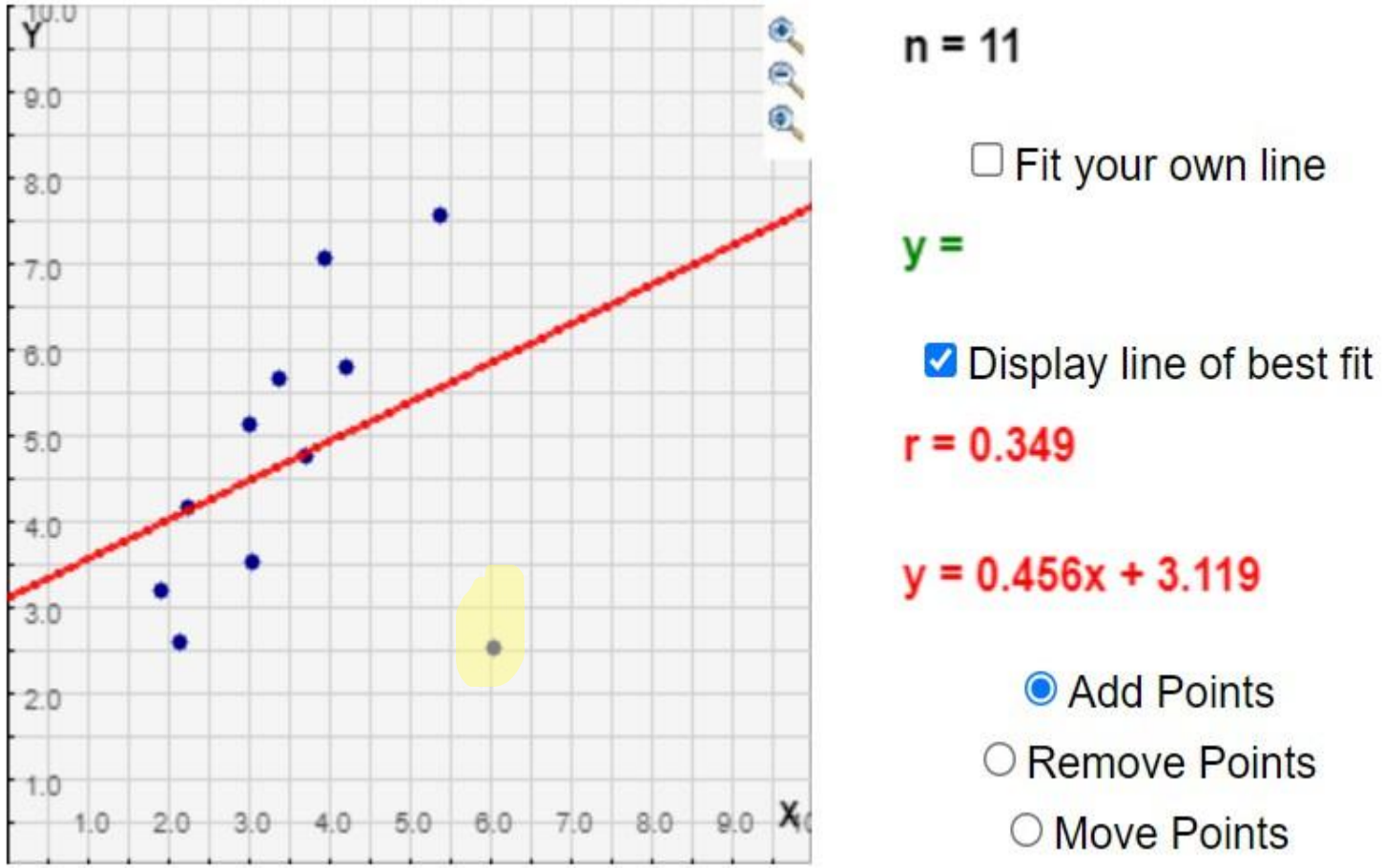
THE EFFECT OF OUTLIERS

OLS aims to minimise the sum of squares of the residuals for all data points. Outliers worsens the fit.

Simple Linear Regression interactively

http://www.shodor.org/interactivate/activities/Regression/

With outliers

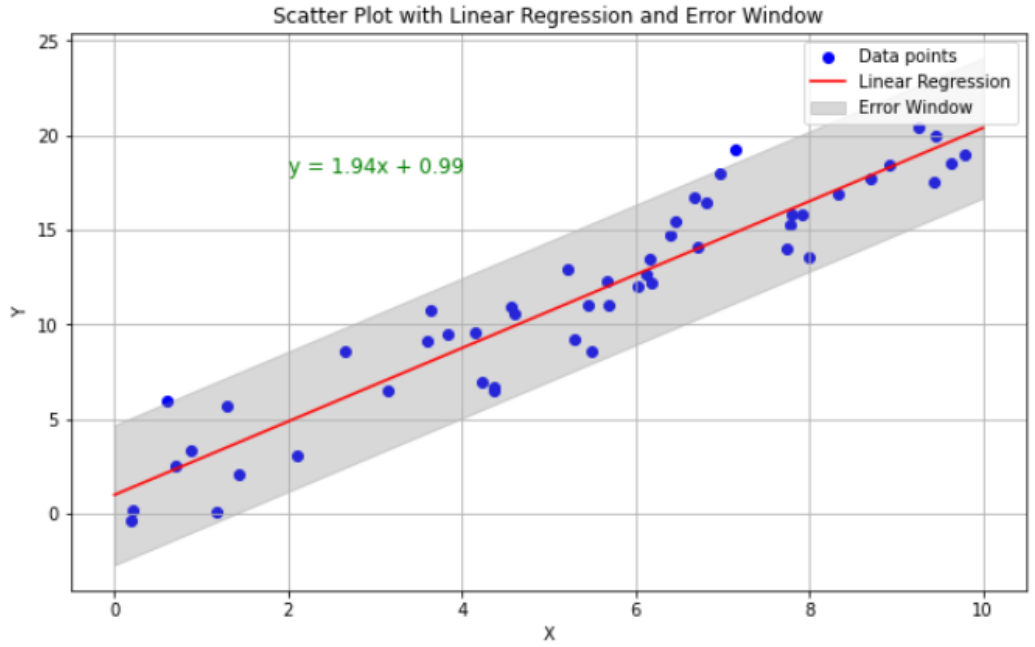


PREDICTING FROM A LINEAR MODEL

The equation for the regression line is calculated using existing data.

Suppose a new value of x is requested and we want to predict the value of y that goes with it.

If x=6, the equation gives us y=12.63, but we can see from the error window that it could be as big as about 17 or as small as about 8.5.

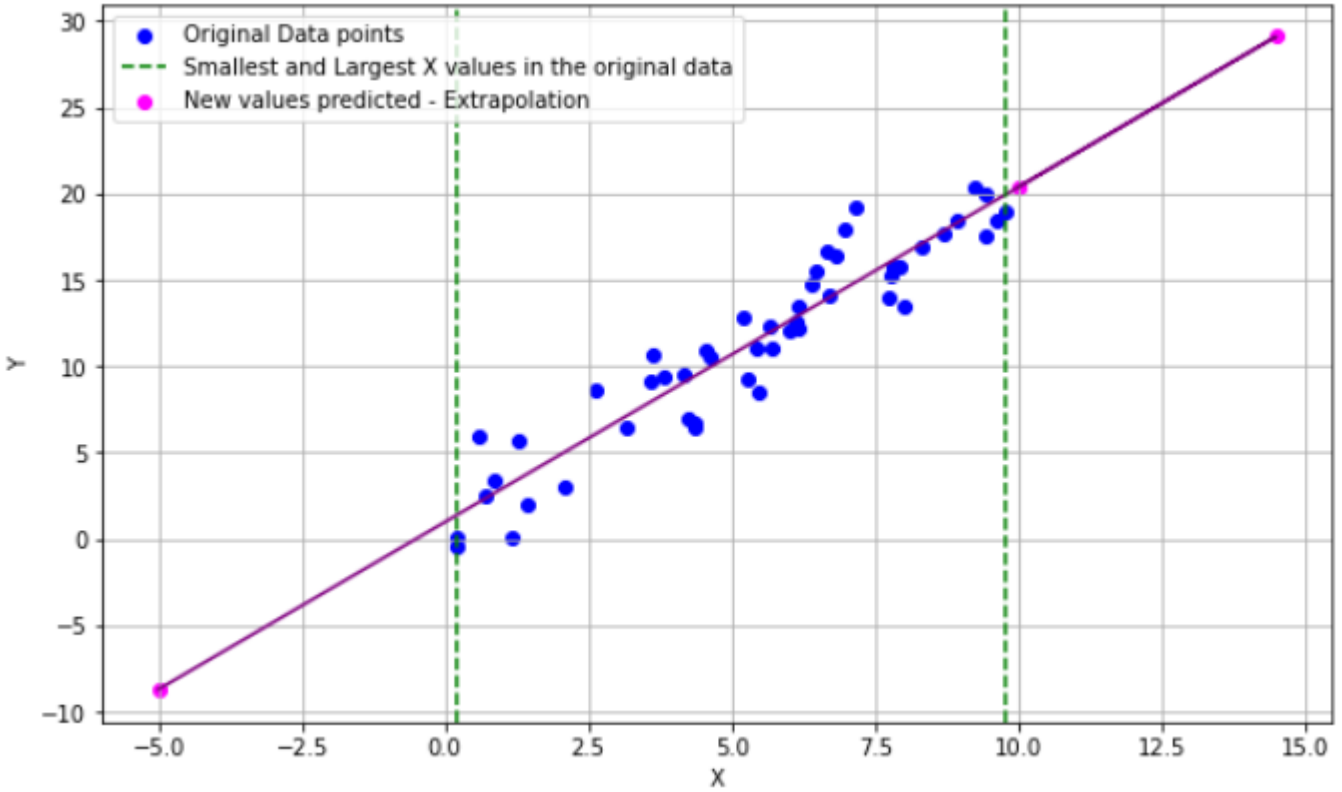


EXTRAPOLATION

Interpolation / Extrapolation

In between the points = reliable

Outside the points = unreliable



ACTIVITY ON FUNDS: ANALYSIS

Select the Excel file returns.xls (which we previously used):

1. Use data analysis in Excel (you might need to load the add-ins Analysis ToolPack) to estimate the correlation matrix of the funds' returns. What insight you get from the same?

2. Draw a scatter plot of the returns for: funds 1 and 7, funds 9 and 10, funds 7 and8. Are the charts confirming the indications obtained from the correlation matrix?

3. Create a histogram for the returns of funds 1, 2, 3, 4, and5 (use the same number of bins). Discuss the shape of each histogram. Can you guess what are mean, median, and mode values? Can you identify funds where the outliers' impact is more relevant? You might use the Excel function skew() to support your findings.

4. Use data analysis to calculate a linear regression of each fund's returns (focus on the first 5 funds) vs. explanatory variables (available in the file). Analyse their impact in each case .

5. Report and discuss the results in class.

Activity: GP scenario analysis

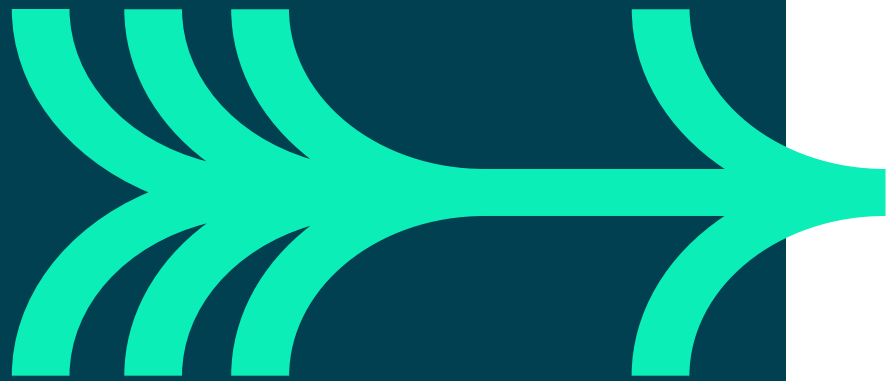
Scenario description:

- They hired you as a data analyst in a GP Practice. You will support the finance team extrapolating useful insight from available data to improve the decision-making process on budget and resources in managing the GP.
- You can use a dataset covering the period 2010-2023 with information available about patients and medical expenses linked to diagnostic checks, medications, and any type of assistance provided over the period.
- You immediately realize that trends and statistics describing the ‘typical’ operational functioning of the GP over the period were largely influenced and skewed by the impact of Covid Pandemic over the last years.
- You need to propose a budget plan for the next 3 years. Would you include in the planning the data skewed by the pandemic? What kind of adjustments might be implemented to the dataset, if any, to improve the robustness of the planning?

Task:

- The class is split in break-out sessions to review and discuss the scenario proposed. Report back to the class a brief description of your considerations and conclusions about the same.

LEARNING CHECK



Think about your answers to these questions:

- How would you define correlation?
- In which situations you would analyse the correlation between two variable?
- In which situations you would use a linear regression model?