Introduction to Data Science (11372 & G 11516)

Semester 1 2021

UNIVERSITY OF CANBERRA

INTRODUCTION TO DATA SCIENCE

Lecture 13

Dr. Ibrahim Radwan

DISTINCTIVE BY DESIGN

OUTLINE

UNIVERSITY OF CANBERRA

- Data Science definition and principles
- Data Science Ethical Concerns
- R for Data Science
- Data Reading
- Data Wrangling
- Data Tidying and Transforming
- Exploratory Data Analysis
- Correlation
- Data Modelling and Linear Regression
- Final Assignment

WHAT IS DATA SCIENCE?

UNIVERSITY OF CANBERRA

- "**Data**": is a raw, unstructured pieces or records that are needed to be processed to have a meaning,
- and, "**Information**" is where these data are processed and presented in a given context so as to make it useful.

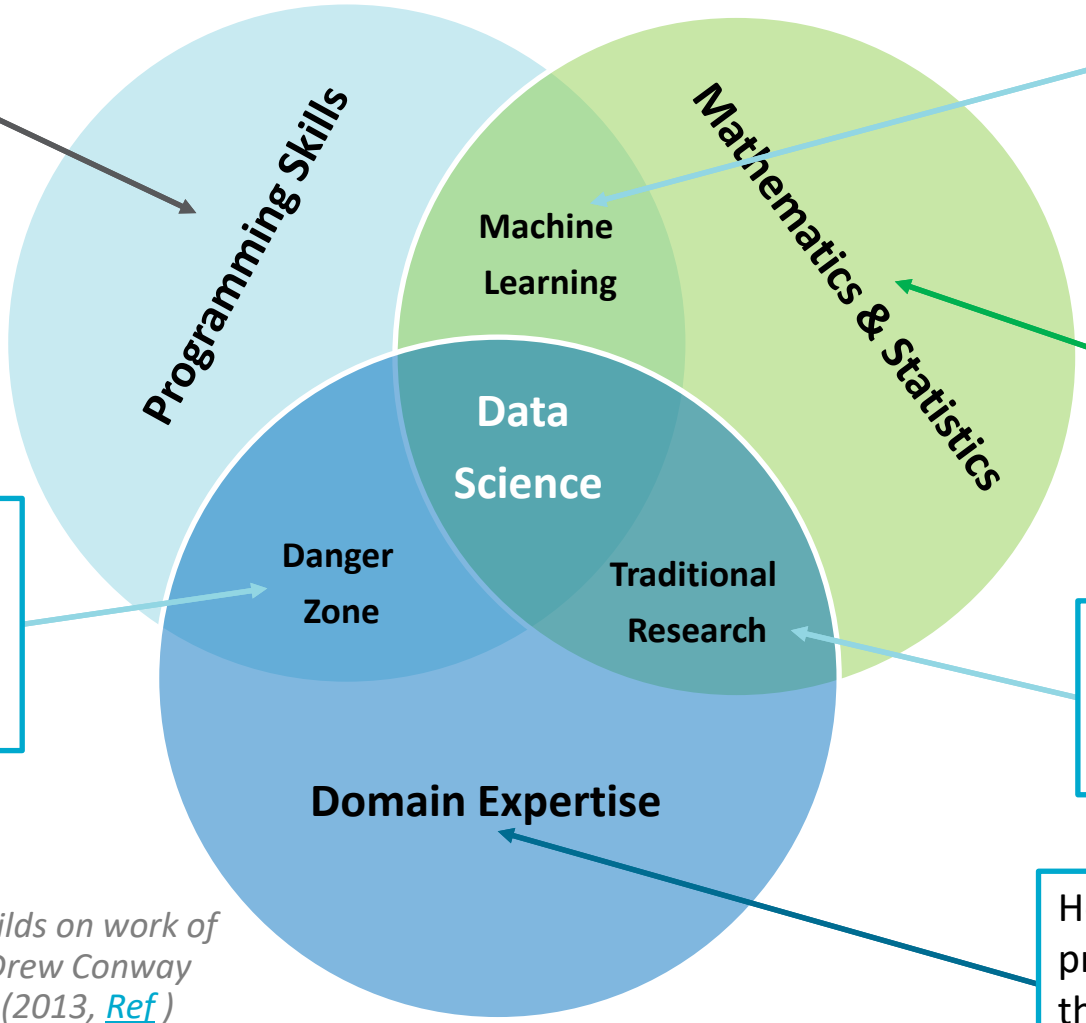
- Data Science is the science that encapsulates set of components that can be used to turn raw data into actionable insights.

WHAT IS DATA SCIENCE? (2)

UNIVERSITY OF CANBERRA

Computer science basics, such as algorithms, programming, etc.

Build models to predict, score or recommend output some values



Distributions of the variables, different modelling strategies, etc.

When we don't know why it works or why it is not. Lack of mathematical and statistical interpretation

Understanding the data records and the problem domain

builds on work of Drew Conway (2013, Ref)

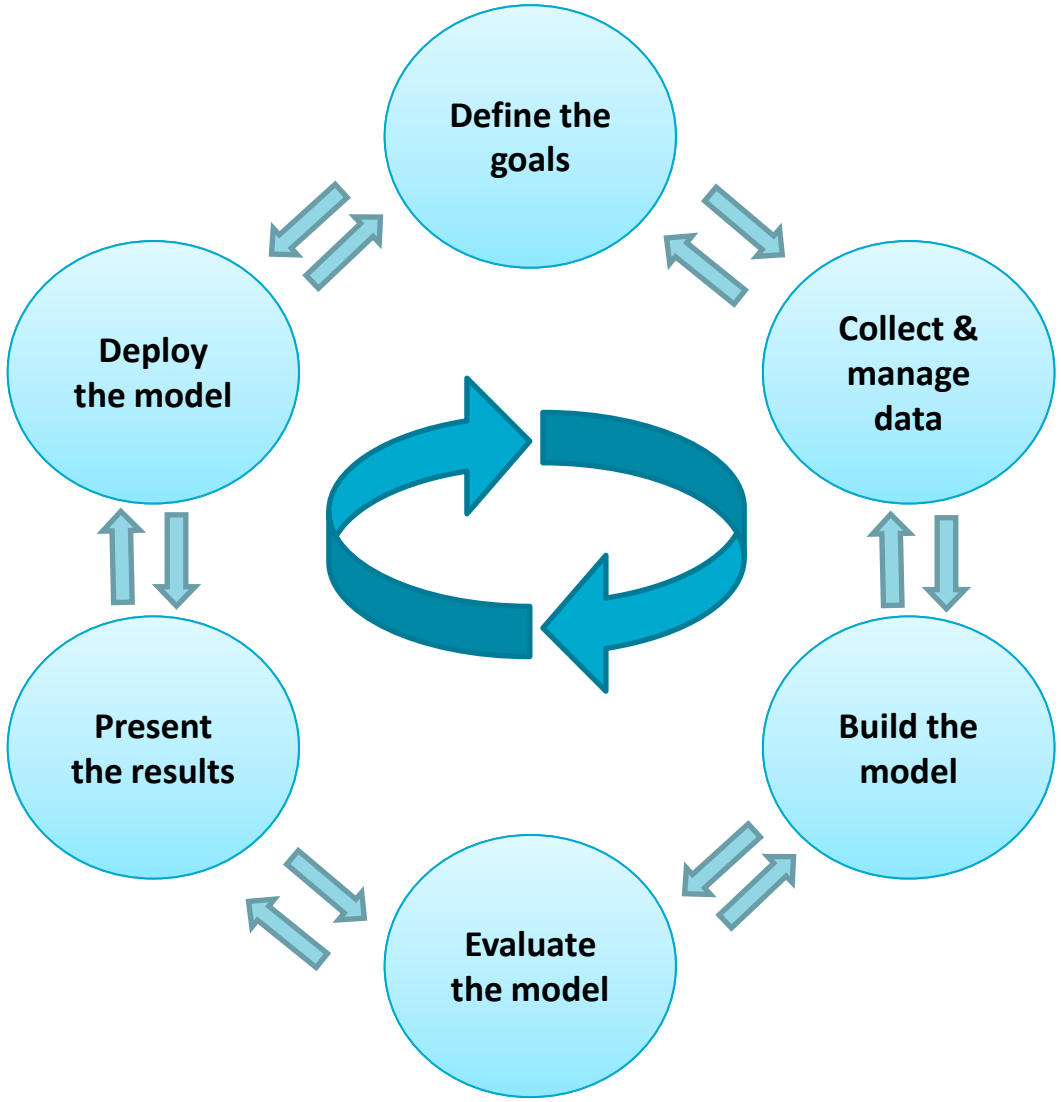
Helps you to understand the problem and the meanings of the data records

Data Science Venn diagram

STAGES OF A DATA SCIENCE PROJECT

UNIVERSITY OF CANBERRA

It is a project life cycle, so loops inside loops



The boundaries are fluid, so the activities are overlapping

You may move back and forth from any stage to any stage

Mostly, the project will be ended with a follow-up one

N. Zumel and J. Mount, "Practical Data Science with R"

CODE OF ETHICS FOR DATA SCIENTISTS

UNIVERSITY OF CANBERRA

- Data are everywhere and large volumes of data are collected every day
- Data science is meant to extract good things from the data, so it is all about doing what is good for the people and what is better for lives
  - Prediction of floods by analysing the satellite images
  - Preventing Suicide by understanding the causes from previous data and build programs to intervene when necessary
  - Different studies to look after the animals and the different species in our planet
- Data is an incredible tool for change, so we need to make sure that this change is what we all want to see

CODE OF ETHICS FOR DATA SCIENTISTS (2)

UNIVERSITY OF CANBERRA

- How ?
  - The code of ethics should represent the principles, values, and standards that govern our behaviour and actions
  - Community effort
  - We need to join in a global conversation about what the standards of dealing with the data should be.
  - In the level of your organisation, get your team of data scientists at a meetup every now and then and start talking about what a Code of Ethics would look like.
  - We need to make our own values and standards for data collection and analysis.

CODE OF ETHICS FOR DATA SCIENTISTS (3)

UNIVERSITY OF CANBERRA

- The following is an example of a "*code of conduct*", which has been put, after many discussions between data scientists, within the Data Science Association:
  - https://www.datascienceassn.org/code-of-conduct.html
- Mostly, organisations will be interested only in its minimum accountability to the law and to the regulations.
- Make your own standards and values when dealing with the clients (professional standards) and with the data you are collecting or analysing

CODE OF ETHICS FOR DATA SCIENTISTS (4)

UNIVERSITY OF CANBERRA

- Examples of these values could be:
  - Produce truthful, interpreted results:
    - Is the data analysis valid?
    - Is the interpretation of the results fair and making sense?
    - What are the social consequences of your outcome?
  - Respect privacy of the data,
    - *"Don't do data mining on someone else's data if you don't want it done on your data."*
    - Anonymise the data when possible
    - Don't use the data beyond the goals of the project or for any personal use

R FOR DATA SCIENCE

UNIVERSITY OF CANBERRA

- R Language (what and why?)
- Variables
- Operators
- Special Numbers
- Conditional Statements
- Loop Statements

PROGRAMMING IN R – BASICS

UNIVERSITY OF CANBERRA

Basic Data Types

Character	Numeric	Integer	Logical	Complex
"AA", "bb"	4.14, 2.65	7, 12	TRUE, FALSE	1+7i, 8+2i

Variables

- X <- 23
- naming conventions

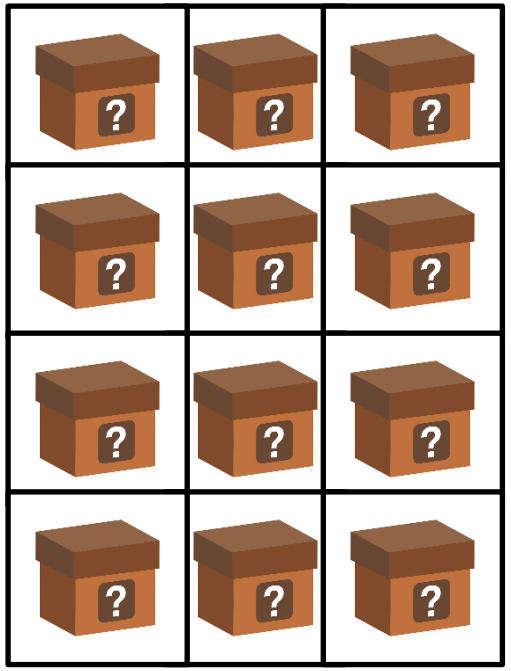
Operators

- Arithmetic
- Logical

DATA STRUCTURES

UNIVERSITY OF CANBERRA

- Collection of elements grouped under one name
  - e.g. container of boxes
    - What type of data to put in?
    - How to access these elements?
    - How to perform operations on these elements?
  - To answer these questions, the data structures can be organized into two categories:
    1. With **similar type** of elements
    2. With **dissimilar type** of elements



DATA STRUCTURES (2)

UNIVERSITY OF CANBERRA

Similar type of items

Homogeneous data structures

Atomic vector	1-D
Matrix	2-D
Array	n-D

Dissimilar type of items

Heterogeneous data structures

List	1-D
Data frame	2-D

DATA FRAMES

UNIVERSITY OF CANBERRA

- A data frame is a series of records represented by rows (observations), where each row contains values in several fields/columns (variables).
- They are like matrices in structure as they are also bi-dimensional.
  - All the operations, we have used for matrices can be applied on data frames as well
    - such as rbind(), cbind(), dim(), ...
- However, contrary to matrices, data frames may include data of a different type in each column.
- So, a data frame is a special case of the lists, where a list represents only one row (in-order) and a data frame can be one row or more.

DATA FRAMES (2)

UNIVERSITY OF CANBERRA

ID, Name, Age

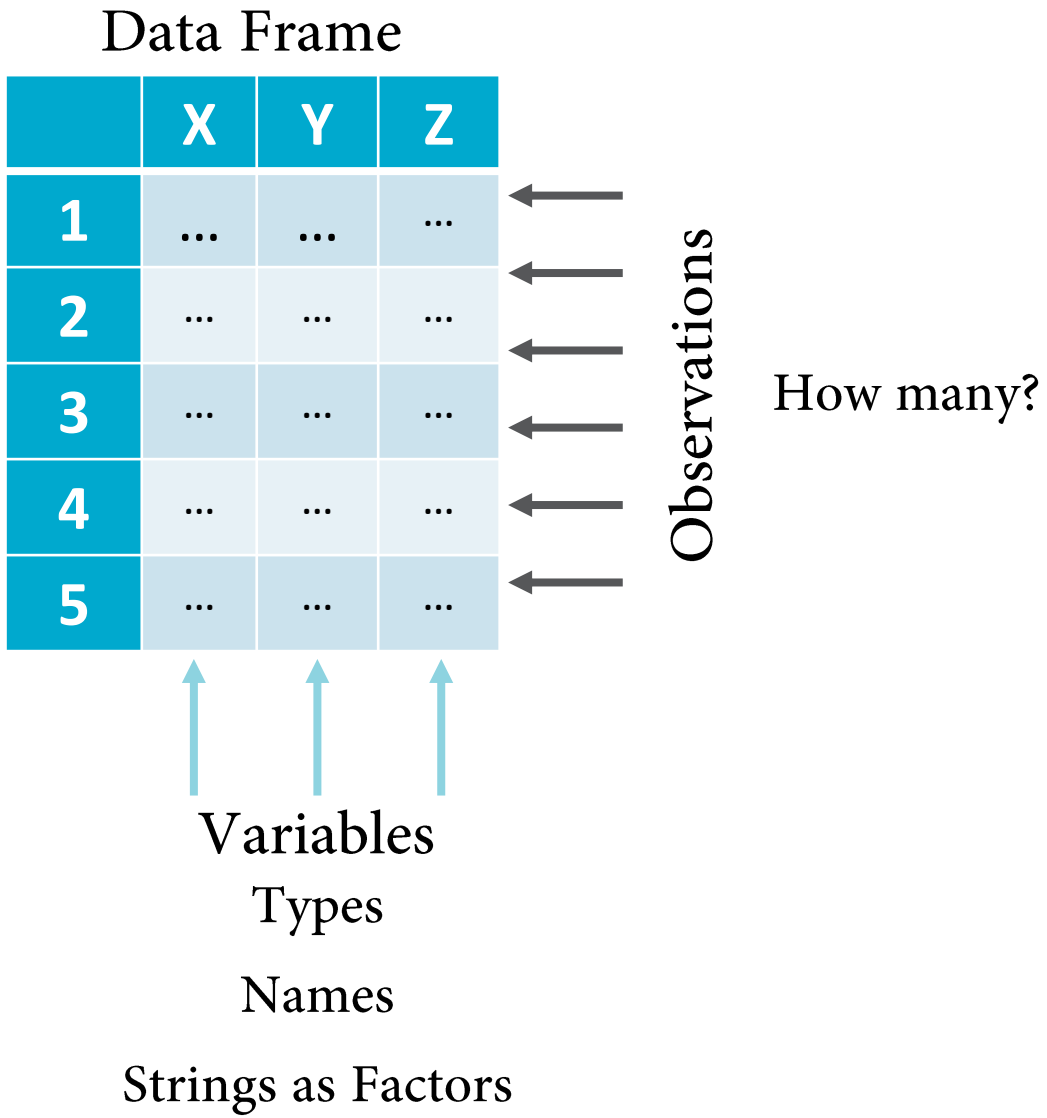
23424, Ana, 45

11234, Charles, 23

77654, Susanne, 76

data.csv

How to read this using the base functions in R?



TIBBLES – A CUSTOMIZED DATA FRAMES

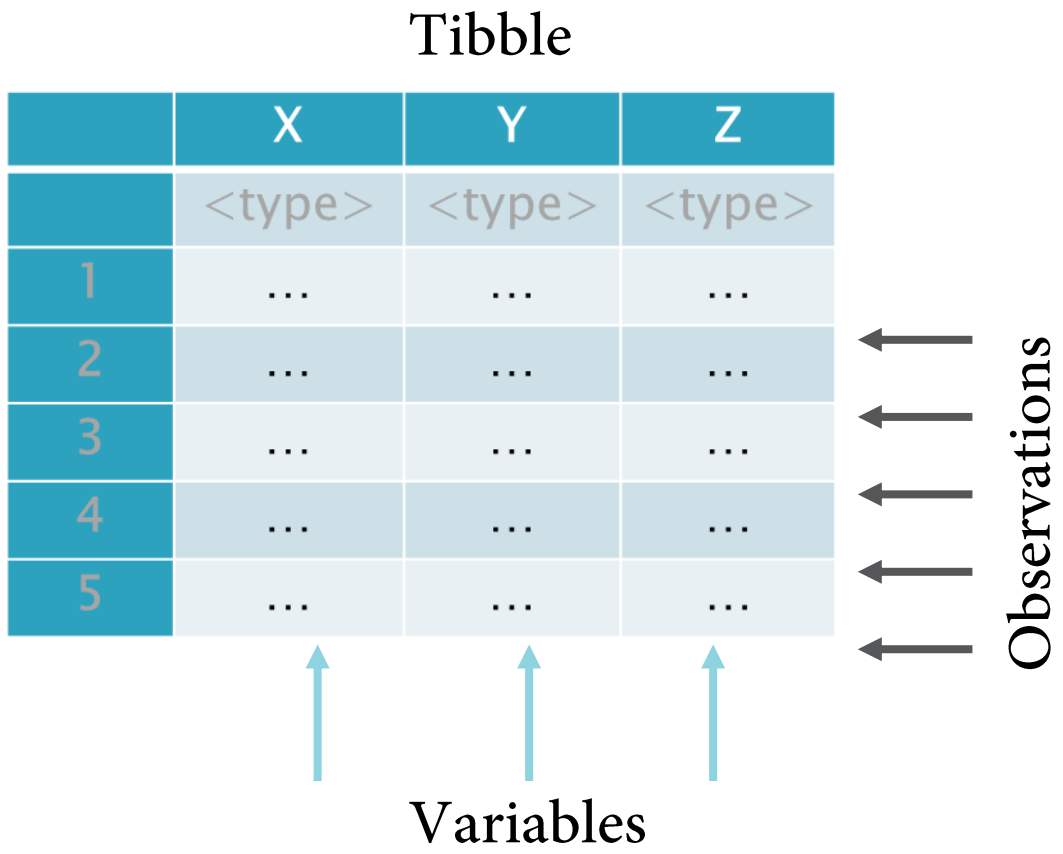
UNIVERSITY OF CANBERRA

- Tibbles are enhanced data frame objects where some of the features have been added and some others were dropped to make the life a bit easier when dealing with big datasets.

int, dbl, chr, factor, lgl, **Col specifications**

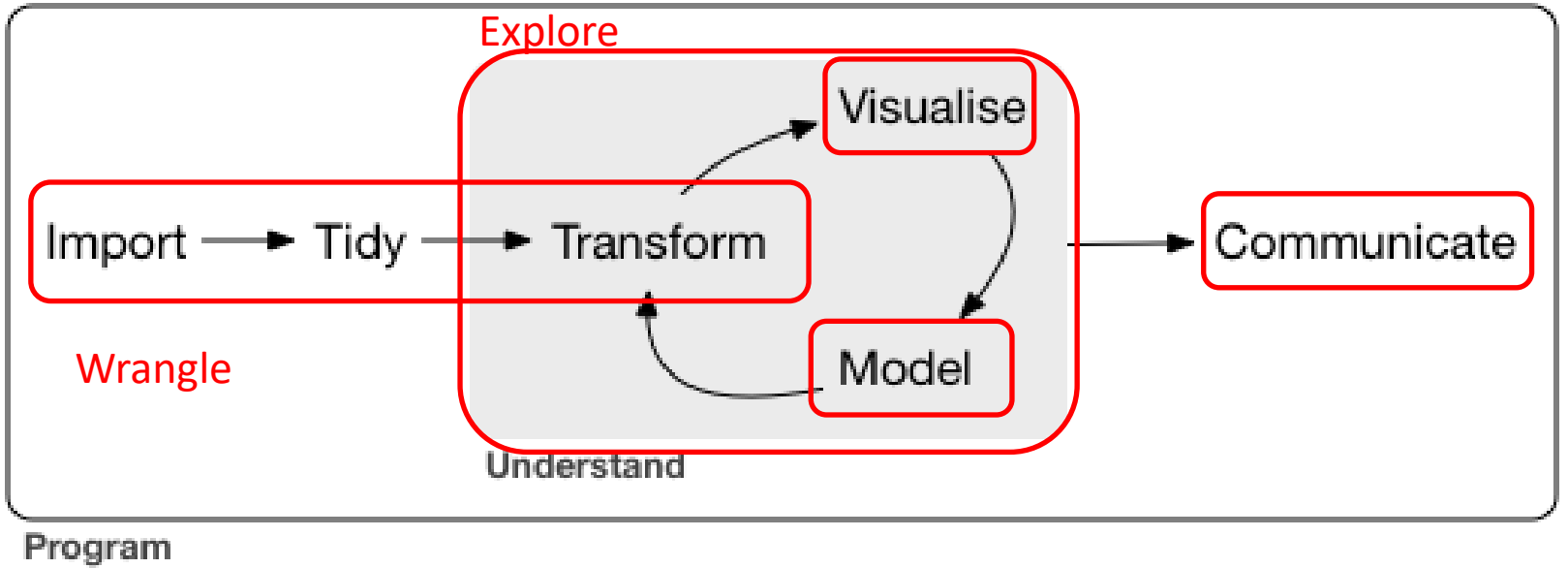
10 rows and the columns that fits on the screen **Nicely printed**

No need to set StringAsFactor **Strings are strings**



DATA SCIENCE; A PRACTICAL VIEW

UNIVERSITY OF CANBERRA



Program Steps

1- Reading Data

2- Data Wrangling

3- Data Exploratory

4- Modelling

5- Result Communication

R for Data Science, by *Garrett Grolemund and Hadley Wickham*

DATA WRANGLING

UNIVERSITY OF CANBERRA

- Data wrangling is the process of transforming the data into a suitable format to conduct the modelling or analysis processes.
- Data wrangling is important: without it, the data may not be ready to be analysed.
- Data wrangling involves the following five steps:

import

1. Gather your data from different sources

Transform

2. Clean up the duplicates, blanks, and other logical errors
3. Join all your data into a single table (e.g. data frame)
4. Adding new variables/records by calculating new fields and re-ordering

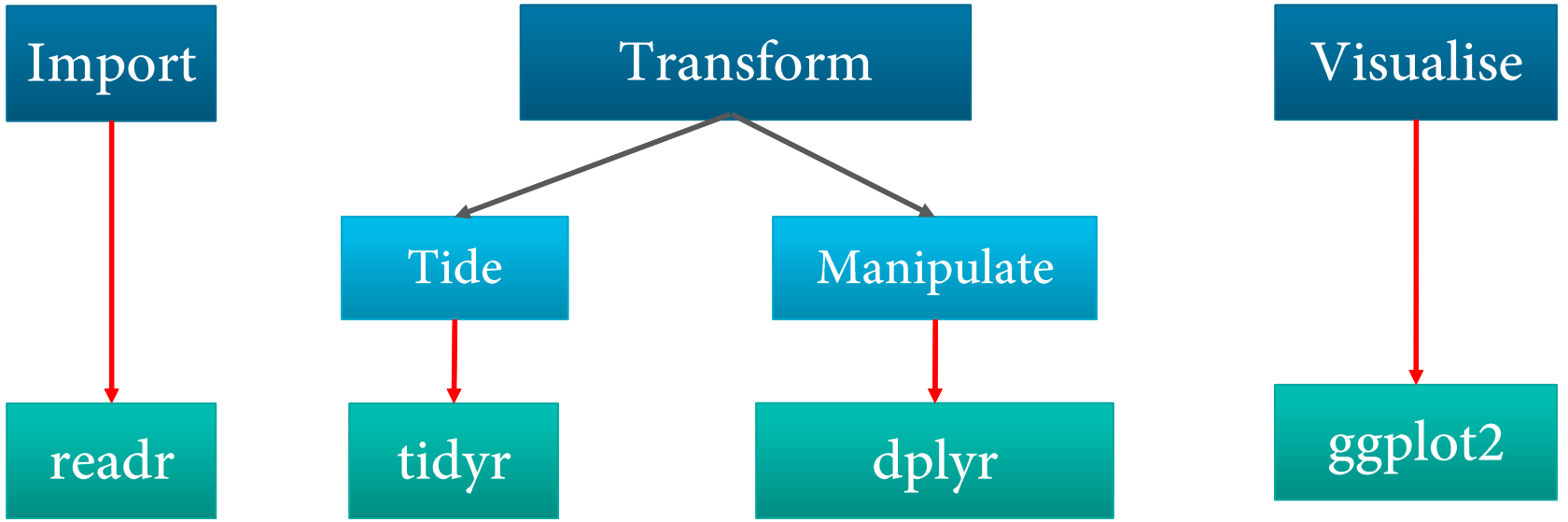
Visualise

5. View or visualize the data to remove outliers and illogical results

DATA WRANGLING (2)

UNIVERSITY OF CANBERRA

- Practically, we have three main processes to wrangle the data



- These four packages provide the *grammar* of the data importing, tidying, manipulation and data visualisation.

DATA IMPORT

UNIVERSITY OF CANBERRA

- Reading flat files such as CSV or text files
- Suppose we have the following CSV file:
  - We will use the `*readr*` library to read the csv files
  - Then, we can use the read_csv("filename.csv") function to read the contents.
  - The returned object is a tibble.
- It reads the data with 10x faster than the base R functions.

ID, Name, Age

23424, Ana, 45

11234, Charles, 23

77654, Susanne, 76

```
> read_csv("data_sample.csv")
Parsed with column specification:
cols(
  ID = col_double(),
  Name = col_character(),
  Age = col_double()
)
# A tibble: 3 x 3
     ID Name      Age
  <dbl> <chr>   <dbl>
1 23424 Ana        45
2 11234 Charles    23
3 77654 Susanne    76
>
```

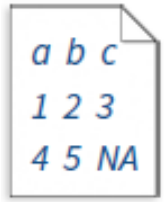
DATA IMPORT (2)

UNIVERSITY OF CANBERRA

```
read_*(file, col_names = TRUE, col_types = NULL, locale = default_locale(), na = c("", "NA"),
  quoted_na = TRUE, comment = "", trim_ws = TRUE, skip = 0, n_max = Inf, guess_max = min(1000,
  n_max), progress = interactive())
```

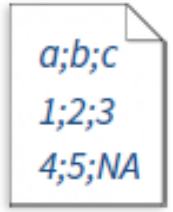
A	B	C
1	2	3
4	5	NA

Comma Delimited Files
read_csv("file.csv"**)**



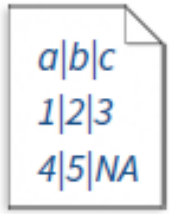
A	B	C
1	2	3
4	5	NA

Tab Delimited Files
read_tsv("file.tsv"**)**
Also **read_table()**



A	B	C
1	2	3
4	5	NA

Semi-colon Delimited Files
read_csv2("file2.csv"**)**



A	B	C
1	2	3
4	5	NA

Files with Any Delimiter
read_delim("file.txt", delim = "|"**)**

To save data into csv or txt file

Comma delimited file
write_csv(x, path, na = "NA", append = FALSE, col_names = !append**)**
File with arbitrary delimiter
write_delim(x, path, delim = " ", na = "NA", append = FALSE, col_names = !append**)**

DATA MANIPULATION

UNIVERSITY OF CANBERRA

- The `*dplyr*` package in `*tidyverse*` library presents five *verbs* for manipulating the data in data frames:
    1. filter() extracts a subset of the rows (i.e., observations) based on some criteria
    2. select() extracts a subset of the columns (i.e., features, variables) based on some criteria
    3. mutate() adds or modifies existing columns
    4. arrange() sorts the rows
    5. summarise() aggregates the data across rows (e.g., group them according to some criteria)
- Each of these functions takes a data frame as its first argument and returns a data frame.

THE PIPE OPERATOR %>%

UNIVERSITY OF CANBERRA

- In data wrangling, most likely, you need to perform series of operations (i.e. *verbs*) on the same data.
- This will need you to create intermediate tables temporarily to save the results to be processed with the next operations.
- R provides an elegant way to perform series of operations on the same data in one go via using the *pipe operator* %>%

original data ⟶ select ⟶ filter

```
F(x) is the same as
x %>% F
```

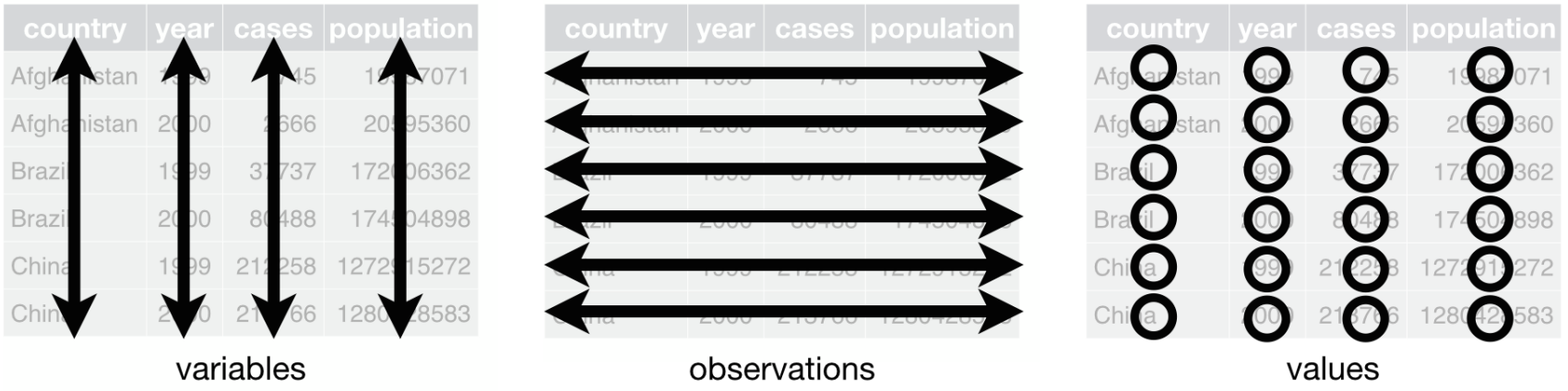
```
16 %>% sqrt() %>% log2()
[1] 2
```

TIDY DATA

UNIVERSITY OF CANBERRA

- There are three interrelated rules which make a dataset tidy:
  1. Each variable must have its own column.
  2. Each observation must have its own row.
  3. Each value must have its own cell.

Having your data in a tidy format is crucial for data manipulation and exploring



Credit: R for Data Science

TIDY DATA (2)

UNIVERSITY OF CANBERRA

- Example of non-tidy data:

	country	1960	1961	1962	1963	1964	1965
1	Germany	2.41	2.44	2.47	2.49	2.49	2.48
2	South Korea	6.16	5.99	5.79	5.57	5.36	5.16

The data are not tidy because:

1. Each row includes several observations and
2. One of the variables, year, is stored in the header.

TIDY DATA (3)

UNIVERSITY OF CANBERRA

To make the data in previous slide tidy, we need to convert it from wide to long. To do so, we first define the variables embedded in the data. Here we have 3 variables. Then we tabulate the data within their corresponding variables.

Now, this dataset is tidy because each row presents one observation with the three variables being county, year and fertility rate.

index	country	year	fertility
1	Germany	1960	2.41
2	South Korea	1960	6.16
3	Germany	1961	2.44
4	South Korea	1961	5.99
5	Germany	1962	2.47
6	South Korea	1962	5.79
7	Germany	1963	2.49
8	South Korea	1963	5.57
9	Germany	1964	2.49
10	South Korea	1964	5.36
11	Germany	1965	2.48

TIDY DATA GRAMMAR

UNIVERSITY OF CANBERRA

- The `*tidyr*` package presents four main *verbs/functions* to tide up the data:
    1. gather() collapses multiple columns into key-value pairs. It produces a "long" data format from a "wide" one.
    2. spread() takes two columns (key & value), and spreads into multiple columns: it makes "long" data wider. This is the reverse of gather.
    3. unite() unites multiple columns into one
    4. separate() takes a column and divides it into multiple columns
- Each of these functions takes a data frame as its first argument and returns a data frame.

DATA VISUALISATION

UNIVERSITY OF CANBERRA

- *"The simple graph has brought more information to the data analyst's mind than any other device."* — John Tukey

DATA VISUALISATION (2)

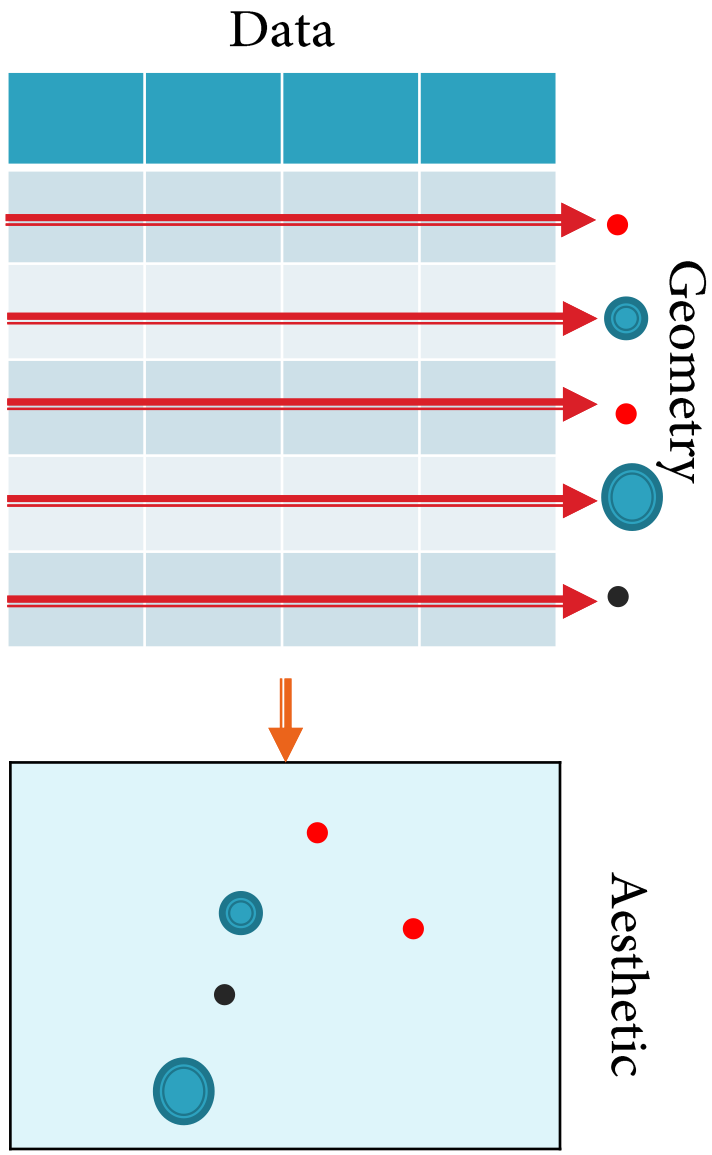
UNIVERSITY OF CANBERRA

- Presenting the data variables into a pictorial or a graphical format
- Visualising data provides a guide to:
  - Check changes in variables
  - Inspect the differences or relations between variables
  - Find patterns in the data
  - Grasp new concepts or insights from the data
- Data visualisation should be easy to the stakeholders
- Data visualisation aids data modelling processes

GRAPH COMPONENTS

UNIVERSITY OF CANBERRA

- To build a graph in R , you will need to specify three components:
  1. **Data**: the set of records/variables that we need to represent with a graph
  2. **Geometry**: the type of the plot, which will be generated, usually it is a function such as (scatterplot, boxplot, barplot, histogram, smooth density, etc.)
  3. **Aesthetic mapping**: the coordinate map and the other visual cues, such as size, scale and color.



BUILD GRAPHS IN R

UNIVERSITY OF CANBERRA

- There are two functions in ggplot2 library to build a graph in R:

qplot()

Quick **plot** that encapsulates the three graph components all in the call of the function.

ggplot()

Build the graph **layer by layer.** Start by defining an empty frame and then add the subsequent operations.

We will be focusing on using this method to build the graphs

BUILD GRAPHS IN R (2)

UNIVERSITY OF CANBERRA

- *Graph grammar* is an elegant way to use few functions to be able to build many graphs and plots layer by layer by combining these functions together.
- The template for building a ggplot graph:

```
ggplot(data = <DATA>) +
    <GEOM_FUNCTION>(mapping=aes(<MAPPINGS>))
```

- The ggplot2 is one of the core members of the `*tidyverse*` library, so you will be able to use its functions when loading the `tidyverse` library.
  - *library(tidyverse)*

EXPLORATORY DATA ANALYSIS (EDA)

UNIVERSITY OF CANBERRA

- EDA: *is the process of exploring the data variables toward discovering some trends or patterns from the data. This leads the modelling step toward fixing issues or guiding the decision making.*
- To understand the variables in a dataset, we may transform these variables into other format or extract their summaries (e.g. mean, variance, etc.) or to get insights about the distribution of these variables.
- The most elegant way to understand relationships with-in a variable or between variables is by ***visualising*** these relationships.

EXPLORATORY DATA ANALYSIS (EDA) – 2

UNIVERSITY OF CANBERRA

- To extract the relationships between variables or to discover the patterns/distributions of the variables, we need to check on the type of these variables.
- To conduct the data analysis on variables for sake of understanding their relationships, this analysis can be either:
  - Uni-variate analysis
    - Discover the variations of the data into **one** variable
  - Multi-variate analysis
    - Discover the co-variation of **multiple** variables
    - Bi-variate analysis is a special type of this analysis with only **two** variables.

EXPLORATORY DATA ANALYSIS (EDA) – 3

UNIVERSITY OF CANBERRA

- EDA
  - Uni-variate analysis
    - Discrete
    - Continuous
  - Bi-variate analysis (can be extended to multi-variate analysis)
    - Discrete
    - Continuous

We will start by using the visualisation to do the EDA for both of the univariate and bi-variate analysis.

EDA – UNIVARIATE

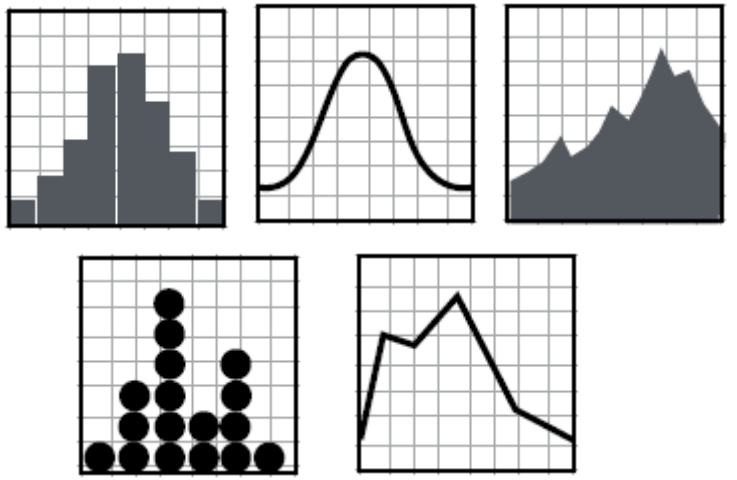
UNIVERSITY OF CANBERRA

- There are two types of visualization-based univariate analysis:
  - Visualising variation of continuous variable
  - Visualising variation of discrete variable
- Examples of univariate continuous :
  - Histograms, etc.
- Examples of univariate discrete:
  - bar plots, etc.

EDA – UNIVARIATE, CONTINUOUS

UNIVERSITY OF CANBERRA

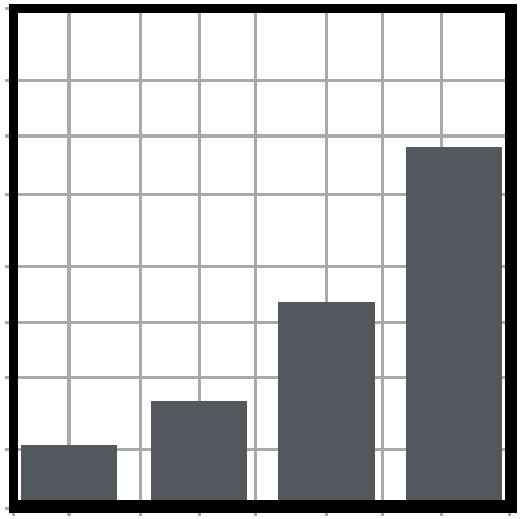
- The analysis is done based on just one variable, where it is a ‘numerical’ continuous variable.
- ggplot2 provides many functions to plot the variation of the univariate continuous variable such as:
  - *geom_histogram()*, for a histogram plot
  - *geom_density(),* for a density plot
  - *geom_area(),* for an area plot
  - *geom_dotplot(),* for a dot plot
  - *geom_freqpoly(),* for a frequency polygon



EDA – UNIVARIATE, DISCRETE

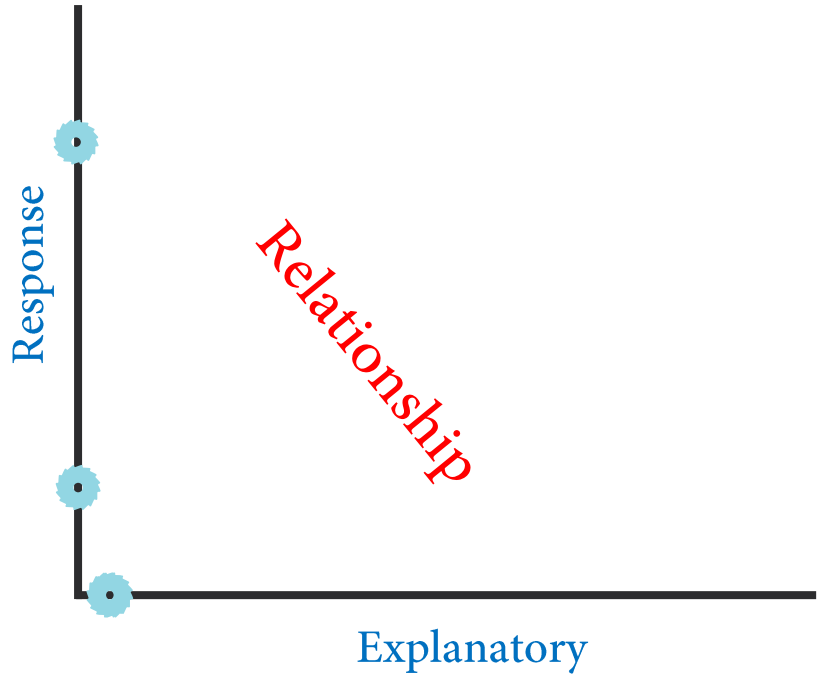
UNIVERSITY OF CANBERRA

- The analysis is done based on just one variable, where it is a discrete (i.e. categorical) variable.
- ggplot2 provides one functions to plot the variation of the univariate discrete, which is:
  - *geom_bar()*, for a bar plot



EDA – BI-VARIATE ANALYSIS

UNIVERSITY OF CANBERRA



EDA – BI-VARIATE ANALYSIS (2)

UNIVERSITY OF CANBERRA

- The bivariate analysis will be done for each of the following pairs:
  - Bivariate analysis of a continuous variable with respect to another continuous variable
  - Bivariate analysis of a continuous variable with respect to a discrete variable
  - Bivariate analysis of a discrete variable with respect to another discrete variable

EDA – BIVARIATE, CONTINUOUS

UNIVERSITY OF CANBERRA

- The analysis is done to inspect the relationship between two variables, where both of them are ‘numerical’, continuous variables.
- ggplot2 provides many functions to plot the variation between these two variables such as:
  - *geom_point()*, for a scatter plot
  - *geom_jitter(),* for scatter plot with displacing the overlapped points a bit away from each other
  - Etc.

BIVARIATE, CONTINUOUS & DISCRETE

UNIVERSITY OF CANBERRA

- The analysis is done to inspect the relationship between two variables, where the explanatory variable is discrete and the response variable is continuous.
- ggplot2 provides some functions to plot this variation such as:
  - *geom_bar()*, for a bar plot
  - *geom_boxplot(),* for visualising the quantiles of response variable
  - Etc.

BIVARIATE, DISCRETE

UNIVERSITY OF CANBERRA

- The analysis is done to inspect the relationship between two variables, where both of them are discrete.
- ggplot2 provides some functions to plot this variation such as:
  - *facet + geom_bar()*, for a bar chart for different categories as subplots
  - *geom_bin2d()*

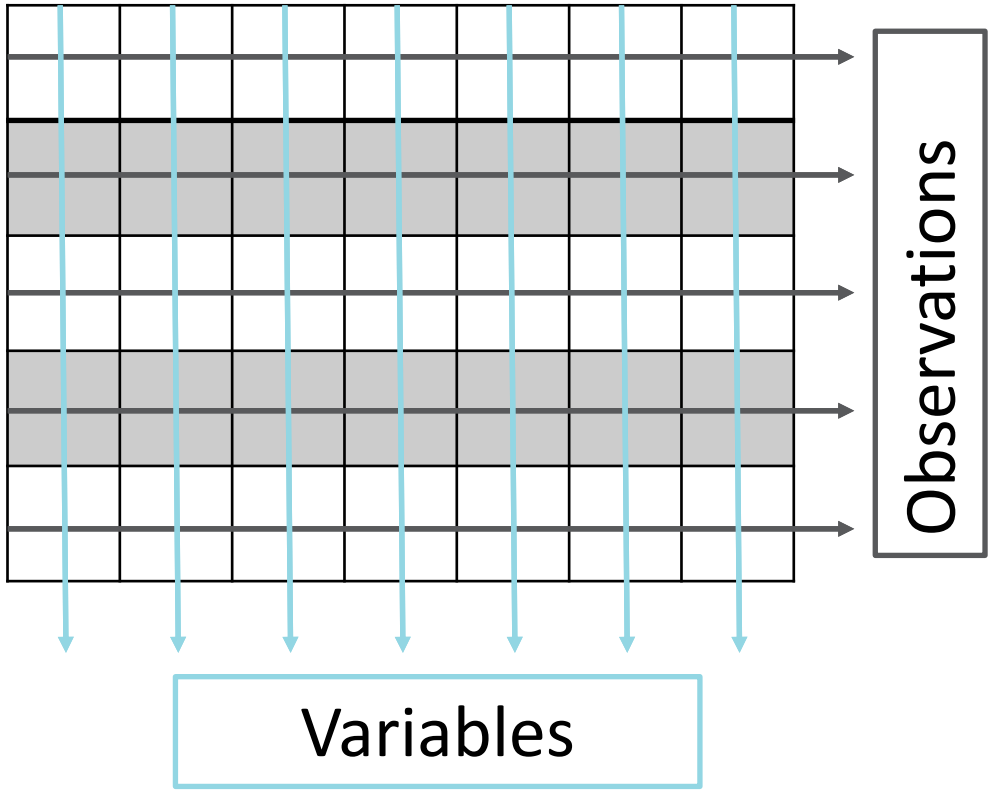
DATA MODELLING

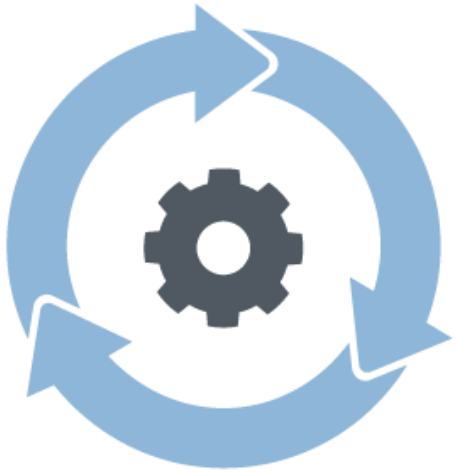
UNIVERSITY OF CANBERRA

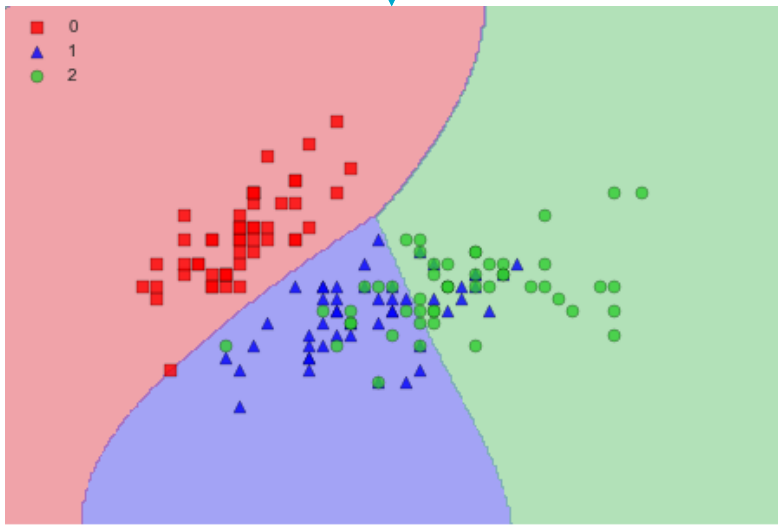
- Modelling is the process of teaching the machines to learn relationships between the data variables for sake of delivering a business value such as predicting an outcome or discovering potential issues or trends.
- Data Modelling makes assumptions about the data distributions & relationships.
- A data model provides a simple, low-dimensional summary of the data and encodes the variations between the data variables.

DATA MODELLING (2)

UNIVERSITY OF CANBERRA







DATA MODELLING (3)

UNIVERSITY OF CANBERRA

Supervised	Un-supervised	Reinforcement
Classification	Clustering	
Regression	Dimensionality Reduction	
Ranking		
Scoring		

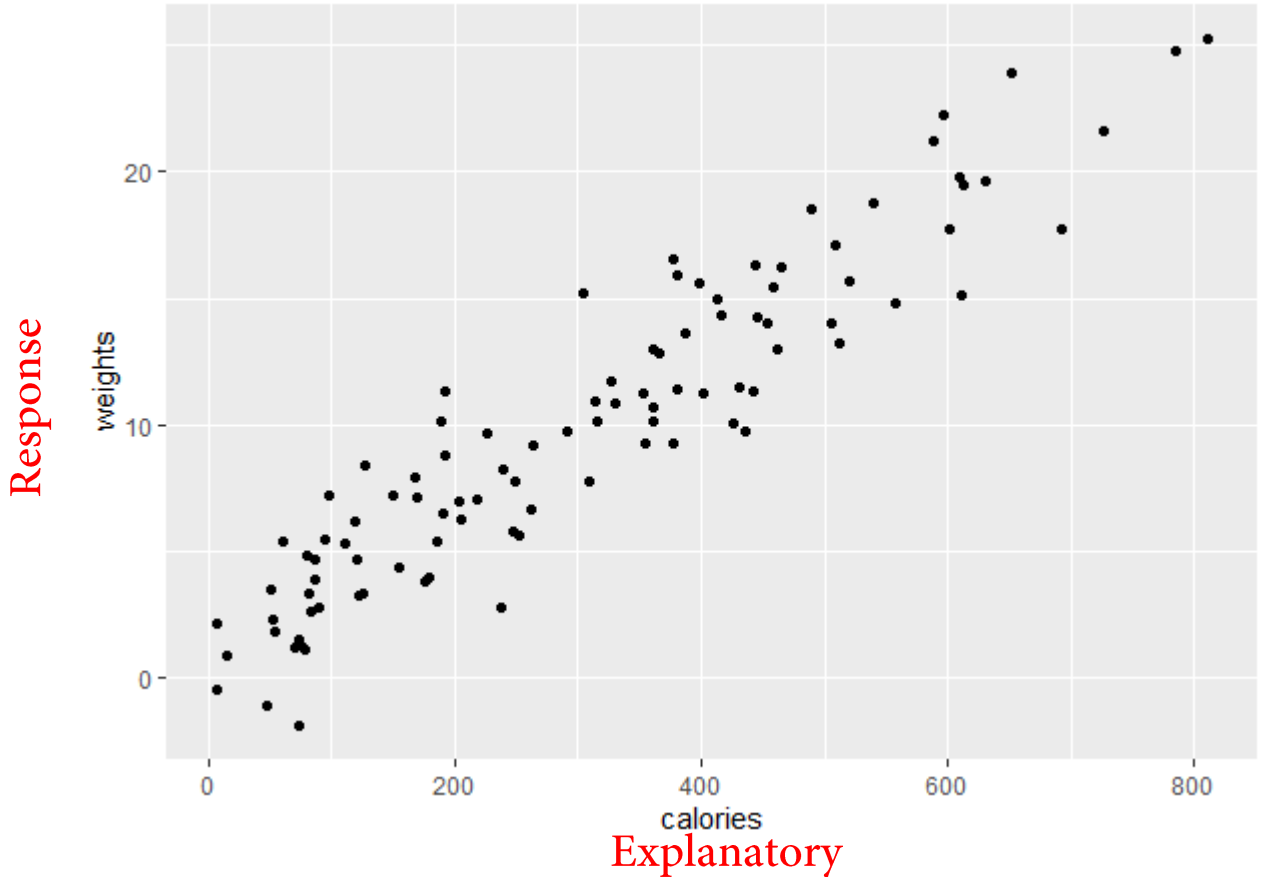
CORRELATION

UNIVERSITY OF CANBERRA

- It is used to measure the relationship between two numerical variables.
- It is actually a measurement for the strength of the linear association between numerical variables.

Relationships:

- Linear
- Positive
- Strong



CORRELATION (2)

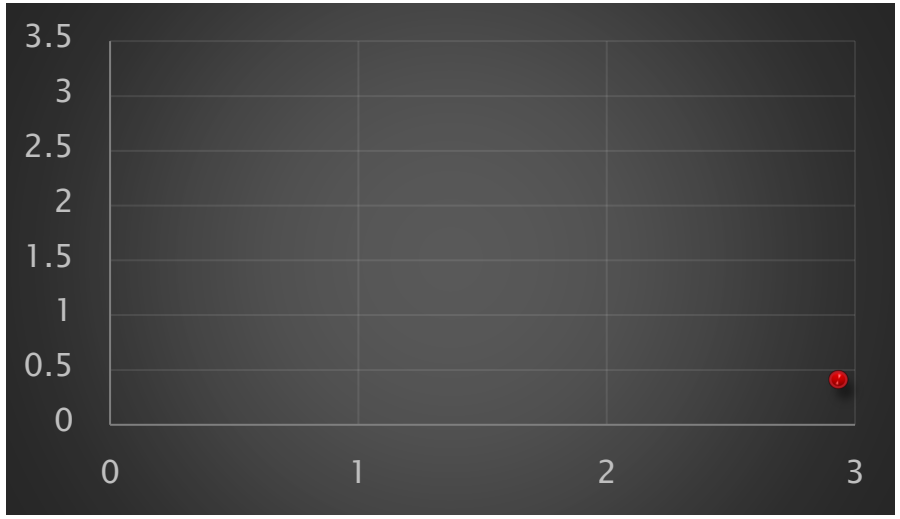
UNIVERSITY OF CANBERRA

- A correlation between variables indicates that as one variable changes, the other variable tends to change in a specific direction.
- Understanding this relationship is useful because we can use the value of one variable to predict the value of the other variable.
- …, for example, height and weight of people are correlated in average, *i.e.* as height increases, weight also tends to increase.
  - Consequently, if we observe an individual who is unusually tall, we may predict that his weight is also above the average.

CORRELATION (3)

UNIVERSITY OF CANBERRA

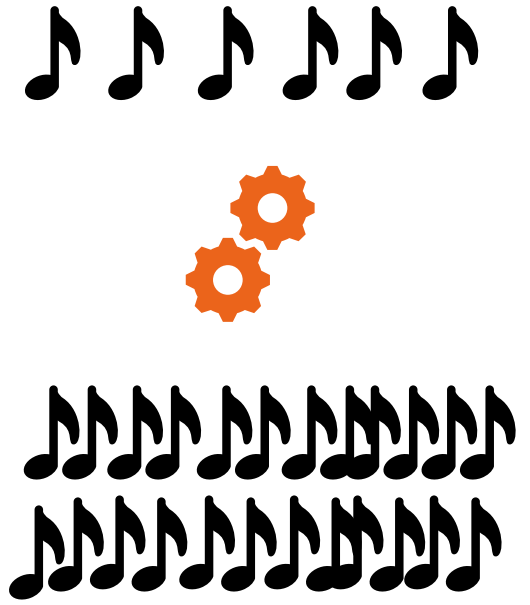
- A correlation coefficient is a quantitative measurement, which assesses the direction and the strength of tendency of the variables to vary.
- Scatterplot is a great way to check for the relationship between pairs of continuous variables, visually.
  - Each dot on the graph represents the relationship between the observational value of the two variables on the x axis and y axis.



LINEAR REGRESSION

UNIVERSITY OF CANBERRA

- Linear regression is a linear approximation of a causal relationship between two or more variables
- Process goes in the following steps:
    1. Get Sample Data
    2. Design a model that works for that sample
    3. Make predictions for the whole population
    4. Measure performance



LINEAR REGRESSION (2)

UNIVERSITY OF CANBERRA

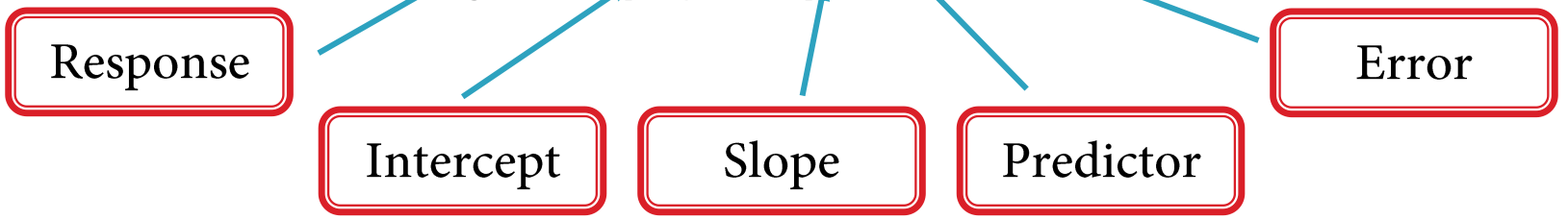
- Linear regression aims to establish a linear relationship between predictor (i.e. explanatory) variable(s) and predicted (i.e. response) variable.
- …, so we can estimate the values of the response variable when only the values of the predictor variable are available.
- The predicted variable is a continuous variable, while the predictor can be categorical or numeric variable.

LINEAR REGRESSION (3)

UNIVERSITY OF CANBERRA

- The linear regression can be estimated by following this formula:

$$y = \beta_0 + \beta_1 x + \epsilon$$



LINEAR REGRESSION (4)

UNIVERSITY OF CANBERRA

- Interpreting the parameters of the following equation with the previous example as following:

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0	β_1	ϵ
The intercept is the minimum or maximum value that can be achieved when the explanatory variable equals zero, i.e. the minimum wage of a job with zero education.	Quantifies the effect of the explanatory variable (x) on the response variable (y), i.e. for each year you invest in educating your self, how many dollars your income would increase?	The error estimation between the observation values and the predicted values of the response variable.

BUILDING MODEL IN R

UNIVERSITY OF CANBERRA

- In R, we use lm function to build the linear model
- When calling the lm function, the variable that we want to predict is put to the left of the ~ symbol, and the variables that we use to predict is placed at the right of the ~ symbol.

```
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

```
# fit regression line to predict cars's stopping distances
from cars's speed
model <- lm(dist~ speed, data = cars)
```

INTERPRETING MODEL OUTPUT

UNIVERSITY OF CANBERRA

```
model
## Call:
## lm(formula = dist ~ speed, data=cars)
##
## Coefficients:
##(Intercept)         speed
##    -17.579         3.932
```

- The estimated regression can be written as follow:
  - dist = -17.579 + 3.932 x speed
- The intercept (b0) is -17.579 . It can be interpreted as the predicted dist for a zero speed is -17.579 ft.
- The slope (b1), is 3.932 . This means that, for every 1 mph increase in the speed, the stopping distance is getting increased by 3.932 feet.

INTERPRETING MODEL OUTPUT (2)

UNIVERSITY OF CANBERRA

```
summary(model)
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,  Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

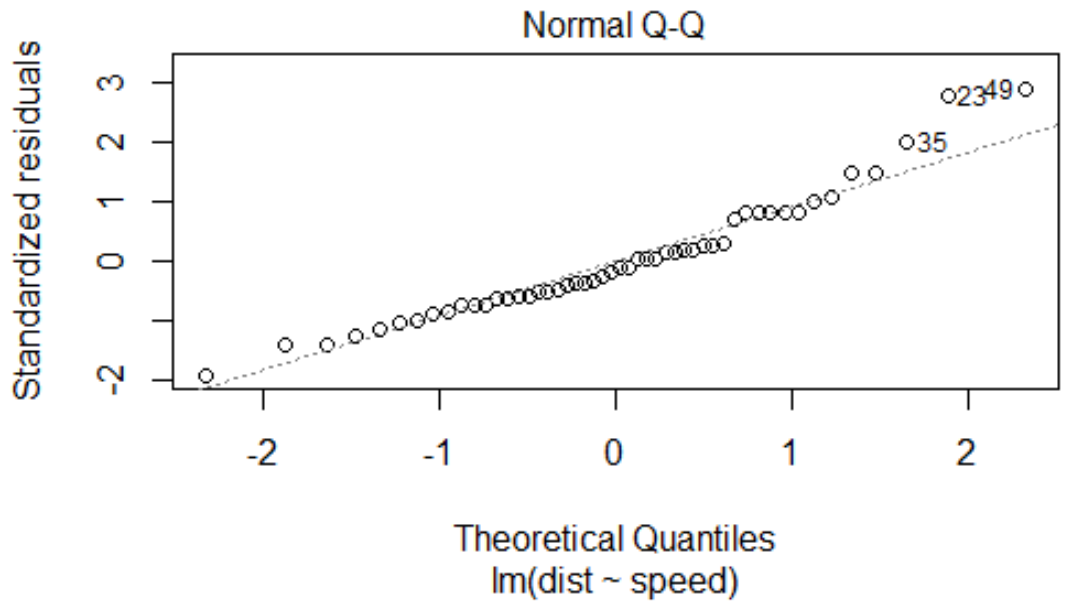
Overall p value on the basis of F-statistic, normally p value less than 0.05 indicates that overall model is significant

VISUALISING MODEL

UNIVERSITY OF CANBERRA

- To visualize the fitted model, you can use Q-Q Plot:

```
> plot(model,2)
```



A straighter line of the residual quantiles means better model fitting

THE PREDICTION

UNIVERSITY OF CANBERRA

- The predicted value is often denoted as $\hat{y}$
- The predict function in R can give us predictions directly.

```
# predict Y values
Y_hat <- predict(model, newdata=new_data se.fit = TRUE)
names(Y_hat)
```

LINEAR REGRESSION – ASSUMPTIONS

UNIVERSITY OF CANBERRA

Pre-building assumptions

1- Linearity

2- Normality

Post-building assumptions

1- Error distribution

2- Patterns of Residual vs. fitted values

FINAL ASSIGNMENT

UNIVERSITY OF CANBERRA

- What: "**take-home assignment**"
- When: "from Friday, 8th of May (17:00) to Friday, 14th of May (17:00)"
- Details:
  - "Represents 40% of the final grade",
  - "**Must achieve at least 50%** of this assignment to pass the unit", and
  - "Consists of 4 parts".

GOOD LUCK!

UNIVERSITY OF CANBERRA

- Thanks, and good luck in the final assessment!
- Any questions, please don't hesitate to post it on Canvas or to send it to me