Comm1190 pre lecture and lecture notes

Data, Insights and Decisions (University of New South Wales)

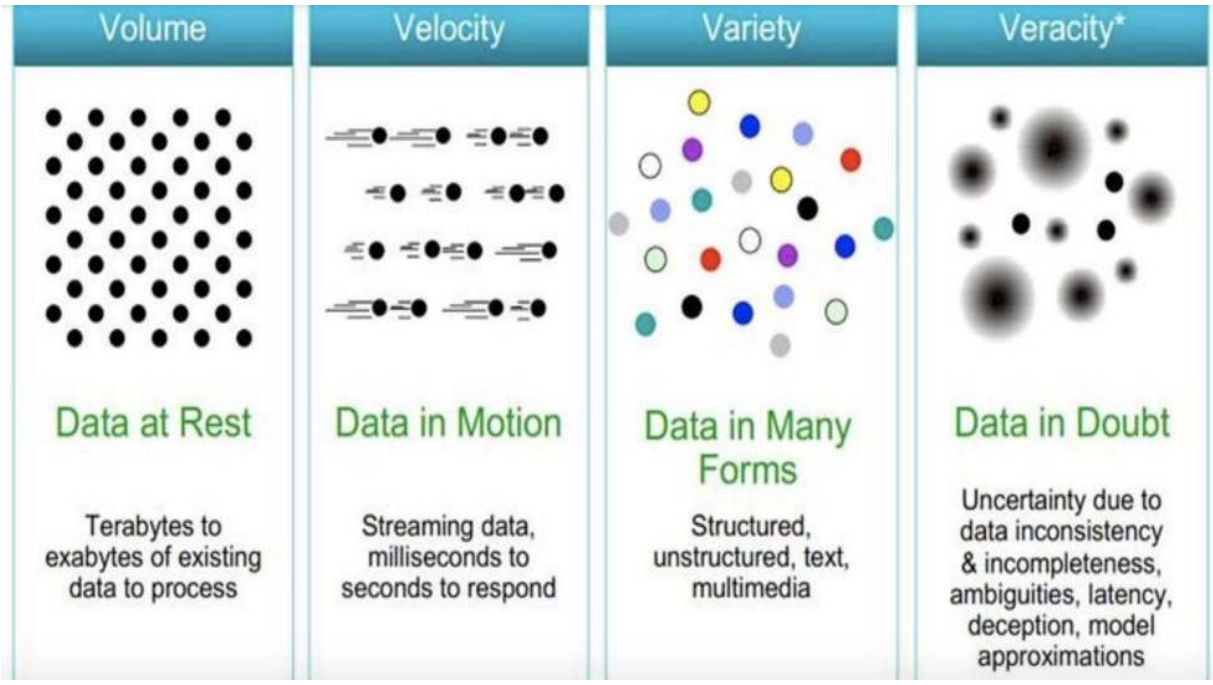
Lecture 1: Data, Analytics and Organisation

What is Business Analytics?

- Business analytics refers to the *extensive use of data, statistical and quantitative analysis, explanatory and and predictive models, and fact based management to drive decisions and actions*

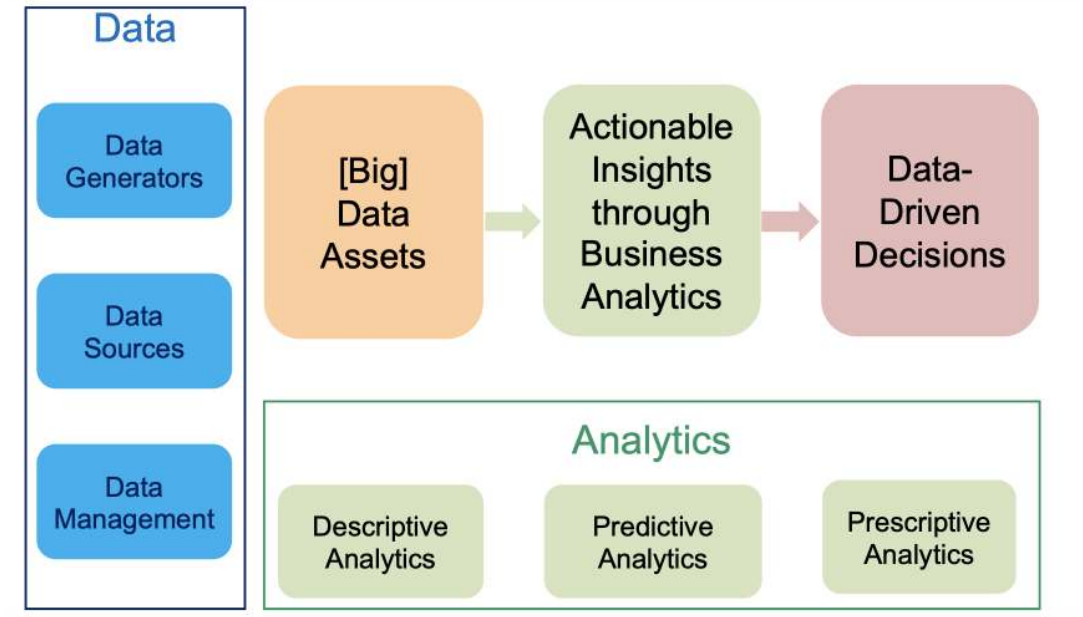
- Business intelligence refers to *"a broad category of tools, software and solutions for gathering, consolidating, analysing and providing access to data in a way that enable managers to make better business decisions"*

4Vs of Big Data - Volume, Velocity, Variety. Veracity



- R is case sensitive!!!
- Choose directory of dataset

Business Analytics Framework



Types of business analytics

- Descriptive analytics: *What has happened and why?*
    - Reports, Drill-down and visualisation
    - Create summary of historical data to yield information
    - "Describe" past or summarise raw data and interpret findings
    - Derive insights from past behaviours → influence on future outcomes

- Predictive analytics: *What will happen?*
    - Predicting future events
    - Understand future and provide marketers with actionable insights
    - Combine historical data to find out patterns and identify trends

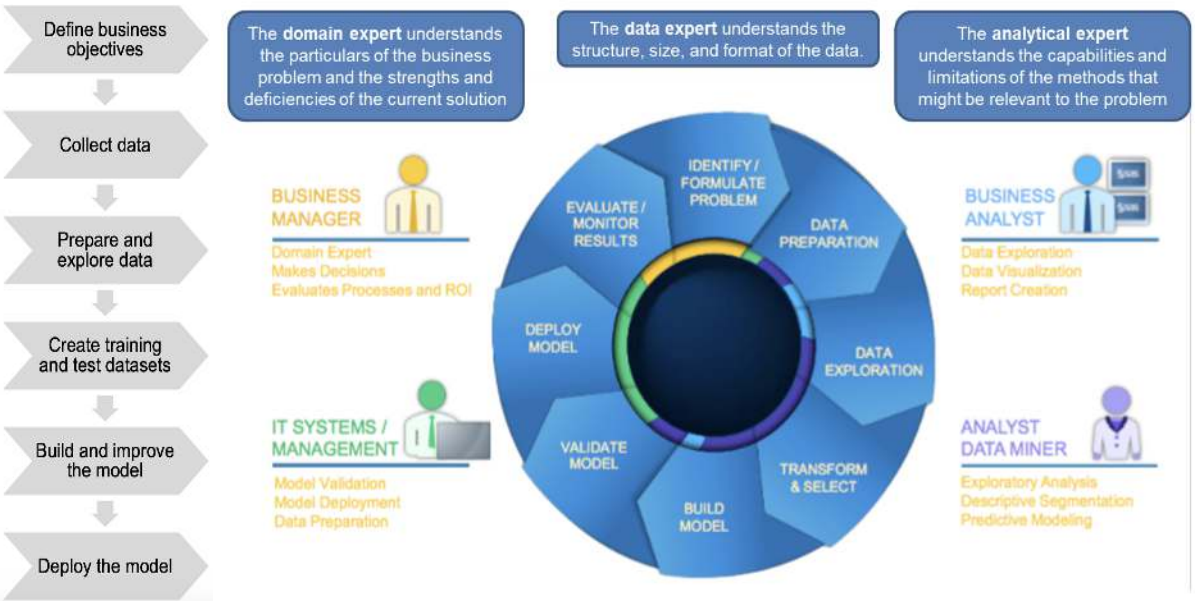
- Prescriptive analytics: *What can we do?*
    - Optimisation and Automated decision making
    - "Prescribe" decision recommendations
    - Predict both what and why it will happen

- R is case sensitive!!!
- Choose directory of dataset

Data Analytic Challenges

Ranking	Item	Description
1	Managing data quality	assuring data quality aspects, such as accuracy, data definitions, consistency, segmentation, timeliness, etc.
2	Using analytics for improved decision making	linking the analytics produced from big data with key decision making in the business
3	Creating a big data and analytics strategy	having a clear big data and analytics strategy that fits with the organisation's business strategy
4	Availability of data	the availability of appropriate data to support analytics (does the data exist?)
5	Building data skills in the organisation	the training and education required to upskill employees in general to utilise big data and analytics
6	Restrictions of existing IT platforms	existing IT platforms/architecture may make it difficult to migrate to and manage big data and analytics
7	Measuring customer value impact	can the real impact on the customer of managing big data be measured?
8	Analytics skills shortage	difficulty in acquiring the mathematical, statistical, visualisation skills for producing analytics
9	Establishing a business case	can 'tangible' benefits of big data be demonstrated (e.g., return on investment)?
10	Getting access to data sources	accessing appropriate data sources to produce and manage big data (can the data be accessed?)
11	Producing credible analytics	are the analytics produced from big data likely to be credible and trusted by the organisation?
12	Building a corporate data culture	e.g., are data and analytics taken seriously enough by the leaders at a strategic level in the business?
13	Making time available	will people have enough time to work with big data and analytics, over and above the 'day job'?
14	Managing data processes	managing the complexity of big data processes (e.g. generating, storing, cleaning data and producing analytics)
15	Technical skills shortage	difficulty in acquiring technical/IT skills for managing big data and operationalising analytics
16	Overcoming resistance to change	is there buy-in and engagement around the benefits of big data (the 'so what?')? Can barriers to change be overcome?
17	Managing and integrating data structures	data held in different business silos, systems and segmented in various ways is difficult to structure for analysis
18	Managing data security and privacy	ensuring that data is stored securely, only available to intended recipients, and anonymised as needed
19	Data visualisation	ability to display and visualise the data to communicate insights clearly within the organisation
20	Managing data volume	does the organisation have effective ways (systems) for storing and managing large volumes of data
21	Data ownership	who owns the big data? Inside (e.g., which department) and outside of an organisation (e.g., Government, partners)
22	Managing costs	ability to manage the costs associated with big data
23	Defining the scope	difficulty in defining the scope of big data projects in the organisation (where does it start and stop?)
24	Defining what 'big' data is	difficulty in defining what 'big data' actually is
25	Securing investment	ability to secure the investment needed to build big data and analytics (infrastructure, skills, training, etc.
26	Manipulating data	being able to process the data to produce analytic insight
27	Legislative and regulatory compliance	compliance with laws such as the Data Protection Act 1998/2003
28	Using the data ethically	using the data in an ethical way and ensuring all areas of the organisation are using it in acceptable ways
29	Performance management	ability to develop key indicators for big data and analytics performance reporting
30	Safeguarding reputation	e.g., reputation and brand damage caused by inappropriate use of data, data leakage, selling data
31	Working with academia	can the organisation build relationships and work effectively with academia?

Modern Data Scientists



- *The Domain Expert* understands particulars of business problem and strengths and deficiencies of the current solution
- *The Data Expert* understands the structure, size and format of the data

- *The Analytical Expert* understands the capabilities and limitations of the methods that might be relevant to the problem

- R is case sensitive!!!
- Choose directory of dataset

Pre-Workshop 1: Data, Analytics and Organisation

Introduction to R

- Three ways to define a value
  - X < - 3
  - X = 3
  - 3 - > X
  - Note: You cannot assign a value to a constant
    - X - > 3 or 3 = X will not work

- Assign a set of values into one variable (called vectors)
  - **var < - c(value 1, value 2, value 3,...)**
    - E.g. staff[c(1,10,20),1:5]

- Data types in R
  - Numeric: Integer or floating with decimals
  - Integer: natural or numbers without decimals
  - Logical: True or False (booleans)
  - Character: Assign character, phrases or words to a variable
    - Characters can be defined with single quotes or double quotes
  - ***Note****: numeric values can be both natural and floating point numbers while integers can ONLY be whole numbers with no decimals*
  - **You can check the data type using: class(variable)**

- Basic Calculations
  - sqrt(<number>)
  - log(value)
  - log(value,base)
  - You can do operations like addition, subtract, multiplication and division using +, -, * :
    - d= a + b
    - f= c * d
    - g= a / b
  - You can also do element-wise operations like addition, subtract, multiplication and division with two vectors
    - c(2, 3, 4) + d(5, 6, 7)

- Viewing data
  - Determine the number of rows and columns (*dimensions*) in the dataset using the ***nrow(dataframe) & ncol(dataframe)***

- Extracting values, rows & columns

- R is case sensitive!!!
- Choose directory of dataset

- Extract a **specific value at a certain row and column** using: **dataframe[row,column]**.
  - E.g. value at the first row and the third column: Staff [1,3]
- Extract an **entire row** using: dataframe[row,]
  - E.g. row1 = staff [1,]
- Extract a **column** using: dataframe[,column]
  - E.g. col3 = staff [,3]

- **$ → access a specific column of a dataset**
- Specifically, for any loaded dataset, you first put the name of the dataset and then the dollar sign and then the variable as one continuous word
  - E.g. variable = staff$S.Caramel

- Specify a **range of columns or rows** by specifying: start:end
  - E.g. Columns 2-5 : staff [,2:5]
  - E.g. Rows 6-10 : staff [6:10,]
- Specify **columns and rows**
  - E.g. columns 1,5,7,9 → Staff[,c(1,5,7,9)]
  - E.g. rows 2,4,8 → Staff[c(2,4,8),]
- You may know the columns of your dataset by name, and you may not know exactly their associated column numbers. You can **obtain columns by names** using c("name1", "name2", "name3", ...)
  - E.g. staff[c("week","Month","S.Caramel","Waffle.Cone")]

- **Subsets based on conditional expressions**
  - Extract a subset of a dataset using conditional expressions
  - E.g. only the rows of data for which sales of Chocolate flavour are above 200.
    - newdata <- staff[staff$Chocolate >= 200]
  - Rows of data for which sales of Chocolate flavour are above 200 and sales of Mango is below 100
    - newdata <- staff[staff$Chocolate >= 200 & staff$Mango <= 100,]
  - Extract the above rows of data and also a specific set of columns (e.g. "week","Month","S.Caramel","Waffle.Cone")
    - newdata <- staff[staff$Chocolate >= 200 & staff$Mango <= 100,c("week","Month","S.Caramel","Waffle.Cone")]
    - Alternatively: newdata <- subset(staff, subset=Chocolate >= 200 & Mango <= 100, select=c("week","Month","S.Caramel","Waffle.Cone")

- R is case sensitive!!!
- Choose directory of dataset

Lecture 2: Data Exploration & Visualisation

Data Types:

- Cross-Sectional
- Time Series
- Panel Data
- Textual Data
- Image Data
- Other Types of data

Variable Types

- Categorical
    - Values sorted into groups or categories
    - Bar charts, pie graphs used to represent data
    - Nominal values
        - Can be assigned a code in the form of a number (labels)
        - Can count but not order or measure nominal data
        - E.g. Sex, eye colour
    - Ordinal values
        - Can be ranked / ordered or have a rating scale attached but not measure
        - E.g. house numbers, swimming level

Data Quality: Six Dimension

- **Completeness - data is comprehensive and meets expectations**
    - E.g. Customer's first name and last name are mandatory but middle name is optional; so a record can be considered complete even if a middle name is not available
    - Questions: Is all required information available? Missing elements / unstable state of data?

- **Consistency - data across all systems / sourced from different places reflects same information**
    - E.g. (i) A business unit status is closed but there are sales for that business unit.
      (ii) Employee status is terminated but pay status is active.
    - Questions: Are data values the same across the data sets?
      Are there any distinct occurrences of the same data instances that provide conflicting information?

- R is case sensitive!!!
- Choose directory of dataset

- **Conformity - data is following the set of standard data definitions like data type, size and format**
    - Example: date of birth of customer is in the format “mm/dd/yyyy”
    - Questions: Do data values comply with the specified formats?
      If so, do all the data values comply with those formats?

- **Accuracy - data correctly reflects the real world object OR an event being described**
    - E.g. (i) Sales of the business unit are the real value.
      (ii) Address of an employee in the employee database is the real address
    - Questions: Do data objects accurately represent the “real world” values they are expected to model? Are there incorrect spellings of product or person names, addresses, and even untimely or not current data?

- **Integrity –all data in a database can be traced and connected to other data**
    - E.g. In a customer database, there should be a valid customer, addresses and relationship between them. If there is address relationship data without a customer then that data is not valid.
    - Questions: Is there any data missing that is important for relationship linkages?

- **Timeliness - information is available when it is expected and needed**
    - Example: (i) Companies that are required to publish their quarterly results within a given frame of time; (ii) Customer service providing up-to date information to the customers
    - Questions to ask yourself: Is this particular data relevant at this point in time?

Summary Statistics

- Mean - location measure
- Variance - dispersion measure
- Standard Deviation - measure of risk
- Minimum & Maximum
- Range
- Covariance
- Correlation

Correlation Coefficient

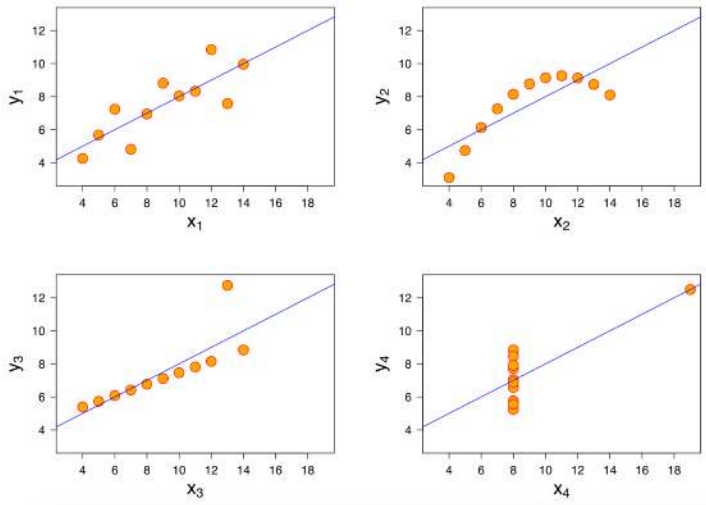
- A correlation of -1 indicates a perfect negative linear relationship
- A correlation of 1 indicates a perfect positive linear relationship
- A correlation of 0 implies no linear relationship
- The larger the correlation in absolute value, the stronger the (positive/negative) linear relationship.

- R is case sensitive!!!
- Choose directory of dataset

Dealing with outliers

- Drop the outlier record
    - Completely remove the record to avoid severe skewness
- Winsorization: Cap your outliers data
    - Limit extreme values in the statistical data to reduce the effect of possibly spurious outliers
- Imputation: Assign a new value
    - If an outlier is due to a mistake in your data, you try imputing a value
    - E.g., using the mean of a variable or utilising a regression model (week 4-5) to predict the missing value

Importance of data visualisation: Anscombe's quartet



Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	4.125
Correlation between x and y	0.816
Linear regression line	y = 3 + 0.5 x
R^2	0.67

UNSW

- R is case sensitive!!!
- Choose directory of dataset

Pre-Workshop 2: Data Visualisation

Data Visualisation using Line Charts

1. Firstly, extract a variable (sales_staff) using the $ sign
    a. E.g. sales_staff <- data$staff
    b. Here, the variable 'sales_staff' is using the data of the 'staff' column
2. Create a graph of sales staff
    a. plot(sales_staff)
    b. The function plot(variable) will plot the variable as the y component, and plot the associated indices on the x-axis.
3. Convert to Line Graph
    a. plot(sales_staff, type = "l")
4. Other Options
    a. E.g. Change the type of the line and the colour of the lines.
    b. plot(sales_staff,type = "l", lty=2, col = "red")
5. Plot Multiple Lines
    a. plot(sales_staff,type = "l", lty=2, col = "red")
       lines(data$student, col = "steelblue4")
6. Missing Data Points
    a. If data does not fit into graph
        i. Check to see the max value of student sales
        ii. Increase max value of y component
        iii. ylim = c(L,U), where L and U are the lower and upper bounds of the axis
    b. plot(sales_staff,type = "l", lty=2, col = "red", ylim = c(0,13000))
       lines(data$student, col = "steelblue4")
7. Legend
    a. Label lines to make clearer
    b. plot(sales_staff,type = "l", lty=2, col = "red", ylim = c(0,13000))
       lines(data$student, col = "steelblue4")
       legend(18, 13000, c("Staff", "Students"), col = c("red", "steelblue4"), lty = c(2,1))
        i. First two numbers (18, 13000) → legend position on the graph; (x,y)
        ii. The next option is the words that the legend says.
        iii. Next, is line colour
        iv. Finally, ensure lines have the same format as they do on the graph. Ensure legend is consistent with the graph → use code lty = c(2,1).
8. Labels
    a. title(xlab = <label>, ylab = <label>, main = <title>
        i. xlab sets the label x-axis

- R is case sensitive!!!
- Choose directory of dataset

ii. ylab sets the label y-axis
iii. main: label figure title

b. Note: x and y axis labels will be written overtop of the original labels. So, if creating new labels → remove the original labels first → option ann = FALSE
c. plot(sales_staff,type = "l", lty=2, col = "red", ylim = c(0,13000), ann = FALSE)
lines(data$student, col = "steelblue4")
legend(18, 13000, c("Staff", "Students"), col = c("red", "steelblue4"), lty = c(2,1))
title(xlab = "Weeks", ylab = "Sales", main = "Aggregate Sales")

Data Visualisation using bar plots, pie charts & box plots

1. Extract Data
2. Create a subset of sales from staff
3. Aggregate the data per flavour for staff
4. Create visualisations:
5. **Create Bar plot → barplot()**
    a. las: Changes the orientation of the labels. 2 means that the labels are perpendicular to the axis.
    b. col: Changes the colour of the bars
    c. barplot(sum_staff, las = 2, col = "lightblue") title(main = "Staff")
    d. To adjust the figure parameters to fit all texts on graph → use par():
        i. mar = c(<bottom>, <left>, <top>, <right>): adjust margins.
        ii. Here we set those margins to: par(mar = c(8, 4, 4, 4)).
        iii. cex = <size>: adjust the font size
        iv. par(mar = c(8, 4, 4, 4), cex = 0.7
        barplot(sum_staff, las = 2, col = "lightblue")
        title(main = "Staff")
    e. Try running the above figure without the line "par(mar = c(8, 4, 4, 4), cex = 0.7)". To do this simply put a hashtag in front of the code. This will make the line of code a comment. If you run the plot above without the code
    f. las – A numeric value indicating the orientation of the tick mark labels and any other text added to a plot after its initialization. The options are as follows:
        i. Always parallel to the axis (the default, 0)
        ii. Always horizontal (1)
        iii. Always perpendicular to the axis (2)
        iv. Always vertical (3)

- R is case sensitive!!!
- Choose directory of dataset

6. **Create Pie Chart**
    a. Create pie chart using → pie()
        i. par(cex=0.75)
        pie(sum_staff)
        title(main = "Staff")
7. **Box plot**
    a. par(mar = c(8, 4, 4, 1), cex = 0.65)
    boxplot(staff[, 4:ncols], las = 2, col = "lightblue");
    title(main = "Staff")
    b. Adjusting the width and height using:
        i. options(repr.plot.width = 5, repr.plot.height = 4)
        ii. E.g. options(repr.plot.width = 5, repr.plot.height = 4)
        par(mar = c(8, 4, 4, 1), cex = 0.65)
        boxplot(staff[, 4:ncols], las = 2, col = "lightblue");
        title(main = "Staff")

- R is case sensitive!!!
- Choose directory of dataset

Lecture 3: Data Exploration & Visualisation II

In addition to the mean, variance and standard deviation introduced in week 2, we consider measures of asymmetry (skewness) and peakine (kurtosis), computed using T sample observations, denoted $\{x_t\}_1^T$

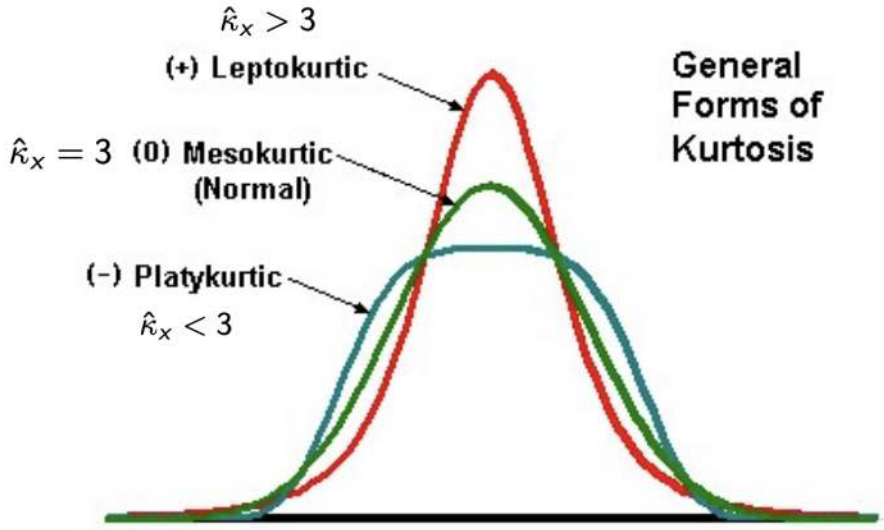
- **Skewness – symmetry measure**

$$\hat{\gamma}_x = \frac{1}{T-1}\sum_{t=1}^{T}\left(\frac{x_t - \hat{\mu}_x}{\hat{\sigma}_x}\right)^3$$

- **Kurtosis – peakidness measure**

$$\hat{\kappa}_x = \frac{1}{T-1}\sum_{t=1}^{T}\left(\frac{x_t - \hat{\mu}_x}{\hat{\sigma}_x}\right)^4$$

Summary statistics: kurtosis



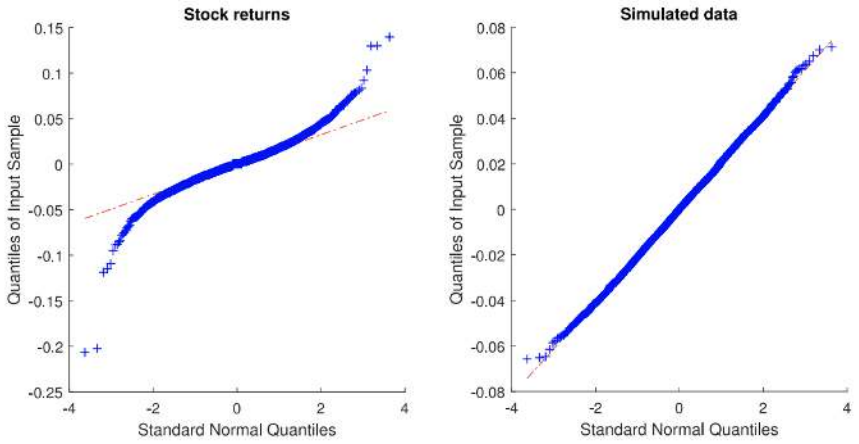
- R is case sensitive!!!
- Choose directory of dataset

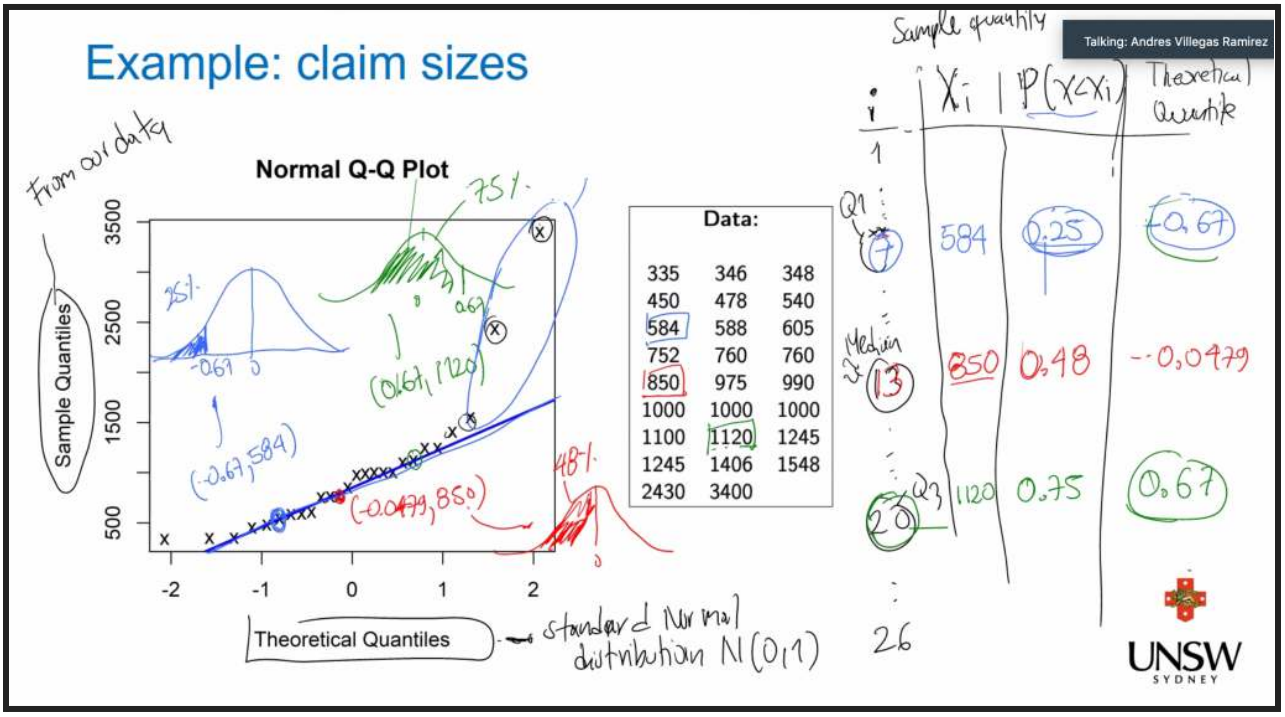
Normally Distributed Data

- Test for normality by comparing skewness and kurtosis with theoretical values under the normality assumption
    - Normal Distribution is
        - Bell shaped
        - Symmetrical (Skewness = 0)
        - Kurtosis = 3

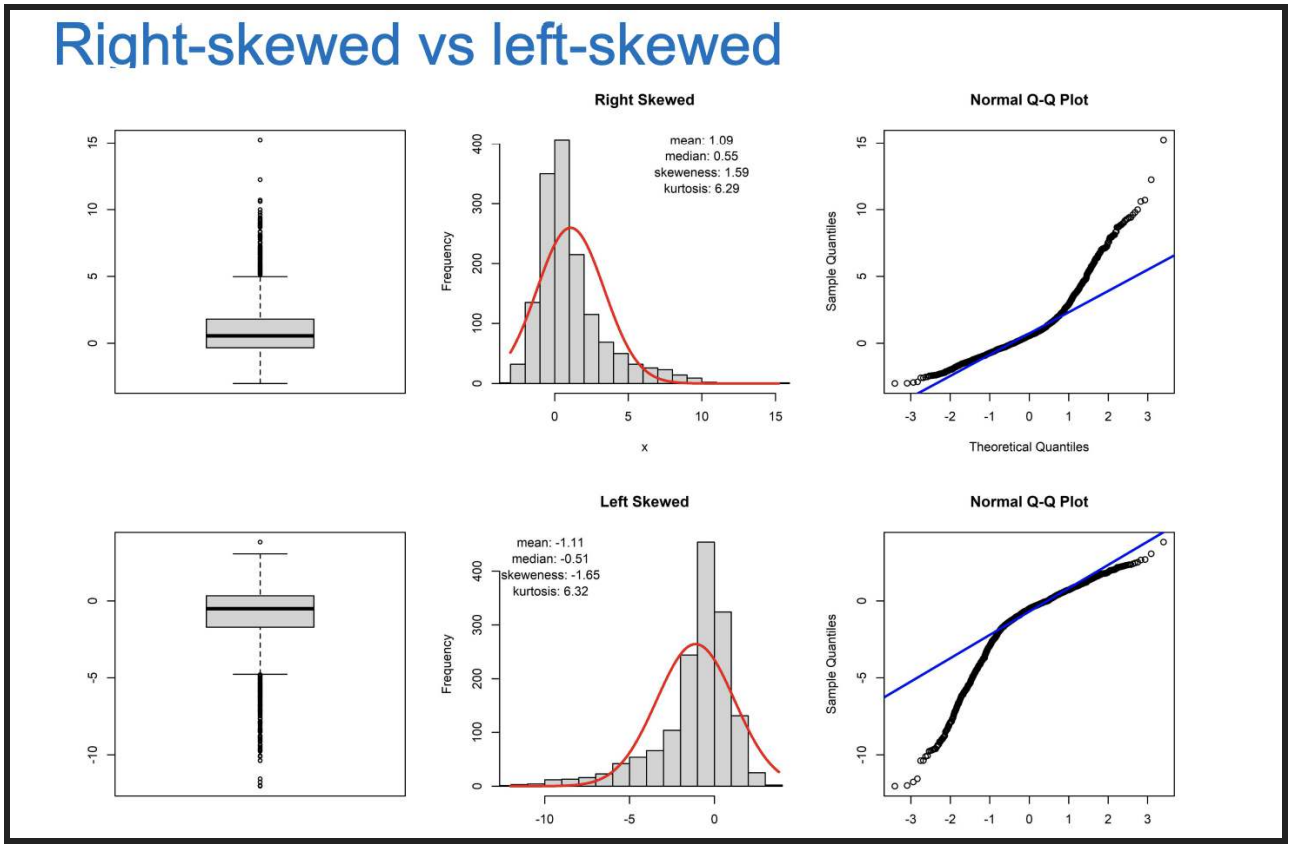
Quintile-to-Quintile (QQ Plot)

- A scatterplot created by plotting empirical quantiles from data against quantiles from normal distribution.





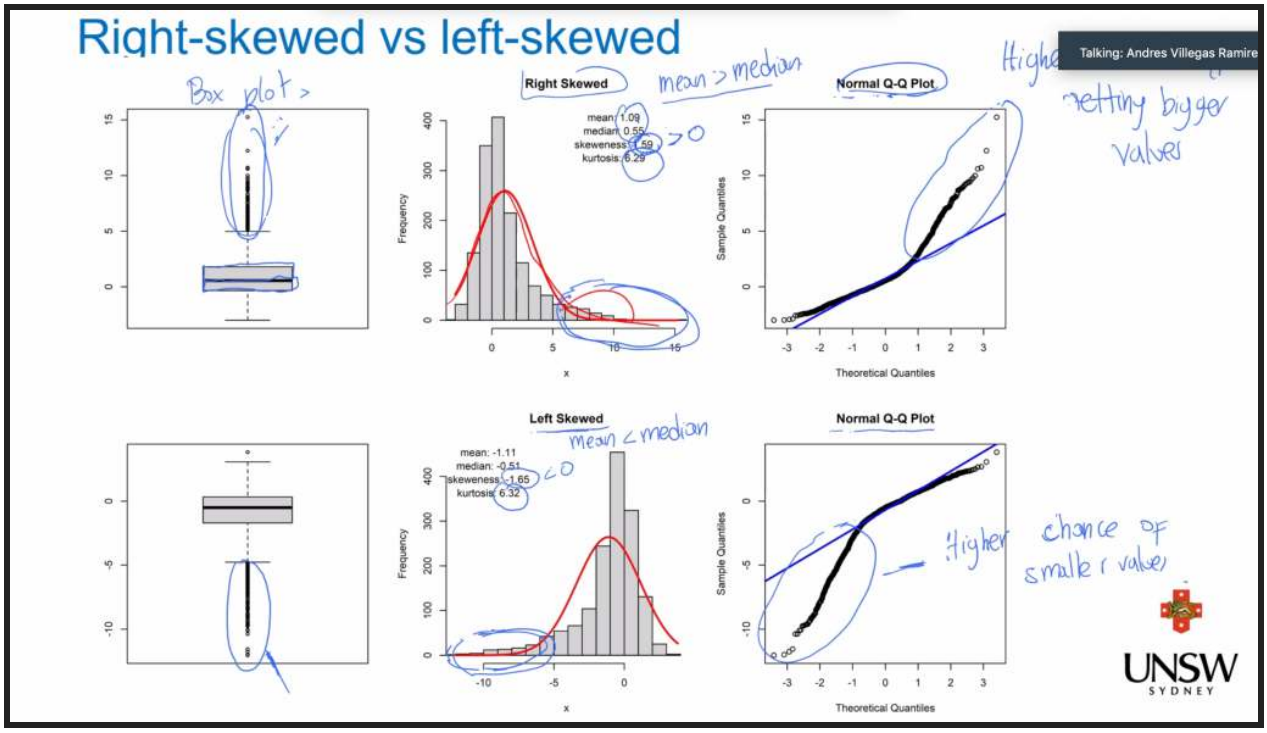
- R is case sensitive!!!
- Choose directory of dataset



- Leptokurtic
  - Kurtosis > 3
  - Extreme deviations at ends of dataset; many outliers on both ends
  - Fat tails

- Platykurtic
  - Kurtosis < 3
  - Opposite deviation to Leptokurtic at end points (Outliers)
  - Thin tails

- R is case sensitive!!!
- Choose directory of dataset



Data Transformation

- Scale: change the range of values
    - E.g. to a range of [0–1]. Generally, changing the scale (or scaling) won't change the shape of the data's distribution.
- Normalise/standardise: change the data's mean to zero and standard deviation to one
    - (E.g., unit variance). This tends to shift the shape of the data towards the shape of a standard normal distribution
- Power transformations: corrects skewness of the distribution

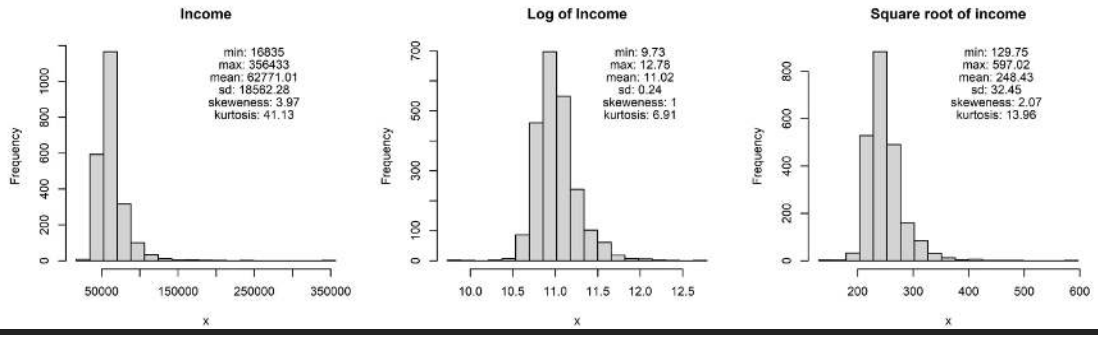
- R is case sensitive!!!
- Choose directory of dataset

Power transformations

- When analysing variables such as asset firms, wage of individuals, house prices, it is common to consider logarithmic units, i.e, $\log y$.
  - Pulled out extreme values
  - Symmetrise skewed distributions
- More generally we consider power transforms, y^{λ}
- *Log* and *square root* transforms are by far the most commonly used ones.

λ	Transformation
-2	$y = \frac{1}{y^2}$
-1	$y = \frac{1}{y}$
-0.5	$y = \frac{1}{\sqrt{y}}$
0	$y = \log y$
0.5	$y = \sqrt{y}$
1	$y = y$
2	$y = y^2$

Personal Income in Australia by Area (2017-2018)



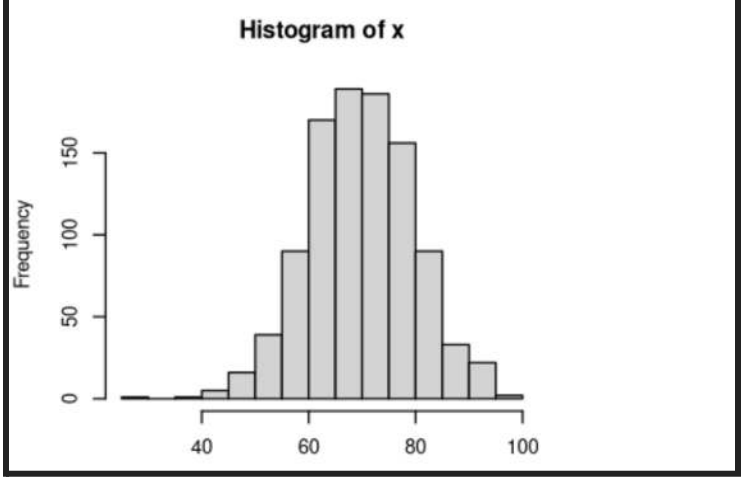
- R is case sensitive!!!
- Choose directory of dataset

Pre-Workshop 3: Statistics

Generate synthetic data using a normal distribution

- rnorm(<no. of samples>, mean = <mean>, sd = <standard deviation>)

```
x <- rnorm(1000, mean = 70, sd = 10)

options(repr.plot.width = 3, repr.plot.height = 3)
par(mar = c(5, 4, 4, 1), cex = 0.65)
hist(x)
```



Simple Statistics

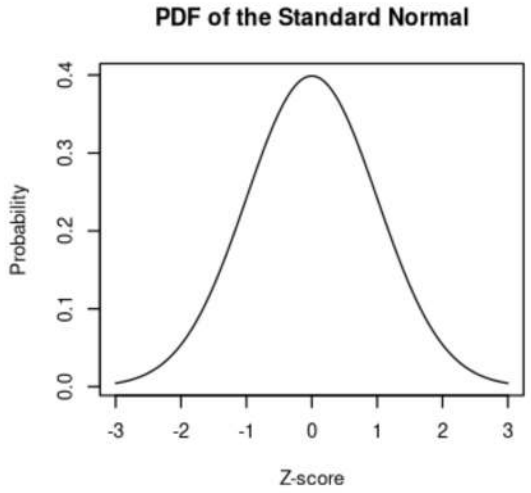
- mean(x), median(x)
- sd(x), var(x)
- min(x), max(x)
- skewness(x), kurtosis(x)
- range(x)
- quantile(x, 0.25), quantile(x, 0.5), quantile(x, c(0, 0.25, 0.5, 0.75, 1))
- Correlation between 2 sets of values
    - Create another set of values and store them in y before calculating the correlation between x and y.
    - y <- rnorm(1000, mean = 80 , sd = 5) cor(x, y)

- R is case sensitive!!!
- Choose directory of dataset

Probability Density Function

- Create PDF using; dnorm(<z scores>, mean = <mean>, sd = <standard deviation>)
- By default, mean = 0 and sd = 1, which will create a normal distribution

```
zscore <- seq(-3, 3, length = 1000)
probability <- dnorm(zscore)

par(mar = c(5, 4, 4, 1), cex = 0.65)
plot(zscore, probability, type = "l", ann = FALSE)
title(xlab = "Z-score", ylab = "Probability", main = "PDF of the Standard Normal")
```

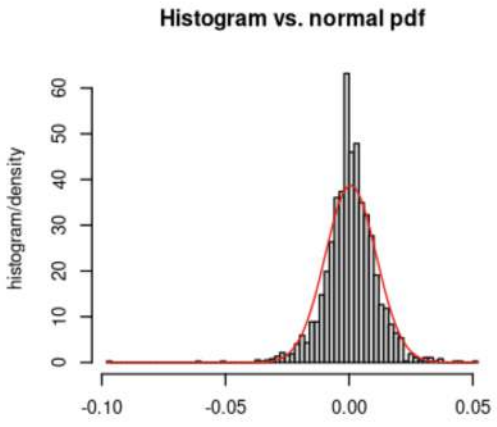


- R is case sensitive!!!
- Choose directory of dataset

Case Study: EuStockMarkets

1. Extract data for first 10 rows → EuStockMarkets[1:10,]
2. Generate correlation matrix → cor(EuStockMarkets)
3. Returns for **DAX Stock Price**
    a. DAX <- EuStockMarkets[, 1]
    b. DAX_returns <- diff(log(DAX))
4. Calculate the corresponding mean and standard deviation
    a. mu <- mean(DAX_returns)
    b. sigma <- sd(DAX_returns)
5. Fit a Normal Distribution
    a. To fit a normal distribution we use `dnorm()` with our `mu` and `sigma` values
    b. Then add this curve to our histogram → Curve ()

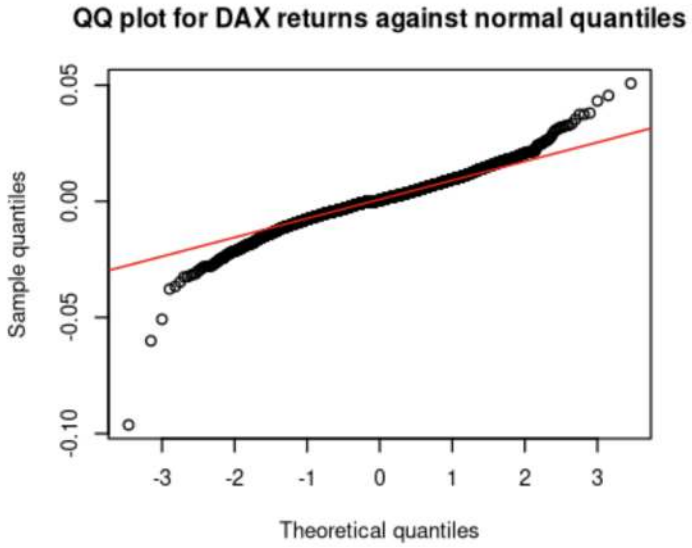
```
options(repr.plot.width = 3, repr.plot.height = 3)
par(mar = c(5, 4, 4, 1), cex = 0.65)

# Histogram
hist <- hist(DAX_returns, breaks = 100, prob = TRUE, ann = FALSE)
title(xlab = "data", ylab = "histogram/density", main = "Histogram vs. normal pdf")

# Fit a normal distribution
x <- hist$mids
curve(dnorm(x, mean = mu, sd = sigma), add = TRUE, col = "red")
```



- R is case sensitive!!!
- Choose directory of dataset

QQ Plot → create a Q-Q plot using `qqnorm()` and `qqline`

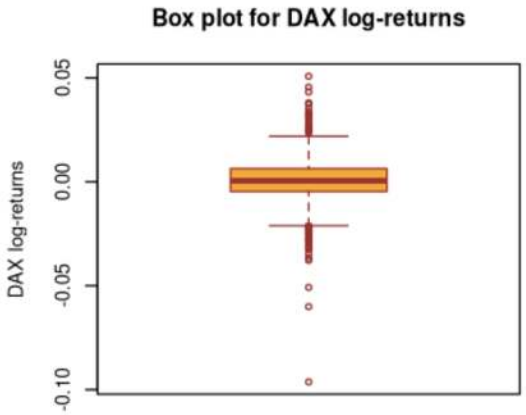
```
options(repr.plot.width = 4, repr.plot.height = 3)
par(mar = c(5, 5, 4, 5), cex = 0.65)

qqnorm(DAX_returns, ann = FALSE)
qqline(DAX_returns, col = "red")
title(xlab = "Theoretical quantiles", ylab = "Sample quantiles",
      main = "QQ plot for DAX returns against normal quantiles")
```



Box plot

```
[ ]
options(repr.plot.width = 3, repr.plot.height = 3)
par(mar = c(5, 4, 4, 1), cex = 0.65)

boxplot(DAX_returns, col = "orange", border = "brown")
title(ylab = "DAX log-returns", main = "Box plot for DAX log-returns")
```

[20]

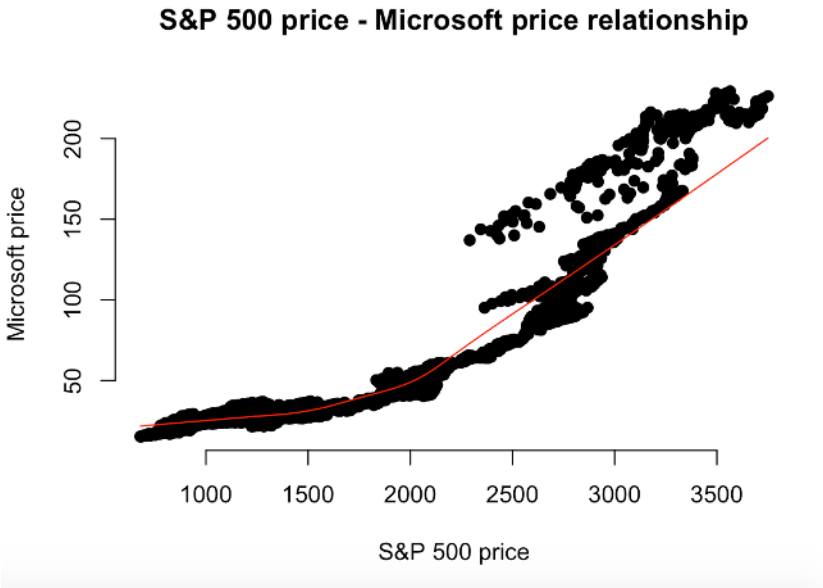


- R is case sensitive!!!
- Choose directory of dataset

Workshop 3 Shit

```
MSFT <- read.csv(file = "MSFT.csv")
MSFT_price <- MSFT[, 7]
MSFT_Returns <- diff(log(MSFT_price))
year_labels <- c("2001", "2003", "2005", "2007", "2009", "2011", "2013", "2015", "2017", "2019", "2021")
year_index <- c(1, 502, 1005, 1508, 2012, 2516, 3018, 3522, 4026, 4528, 5031)
MSFT[year_index, ]
plot(MSFT_price, type="l", ann=FALSE, xaxt= "n", col="red")
title(xlab="year", ylab= "Microsoft", main="Microsoft price")
install.packages("moments")
library(moments)
mean(MSFT_Returns)
var(MSFT_Returns)
sd(MSFT_Returns)
min(MSFT_Returns)
max(MSFT_Returns)
skewness(MSFT_Returns)
kurtosis(MSFT_Returns)
```

```
> mean(MSFT_Returns)
[1] 0.0004618773
> var(MSFT_Returns)
[1] 0.0003040446
> sd(MSFT_Returns)
[1] 0.01743688
> min(MSFT_Returns)
[1] -0.1132403
> max(MSFT_Returns)
[1] 0.1315893
> skewness(MSFT_Returns)
[1] 0.07863309
> kurtosis(MSFT_Returns)
[1] 9.097398
```



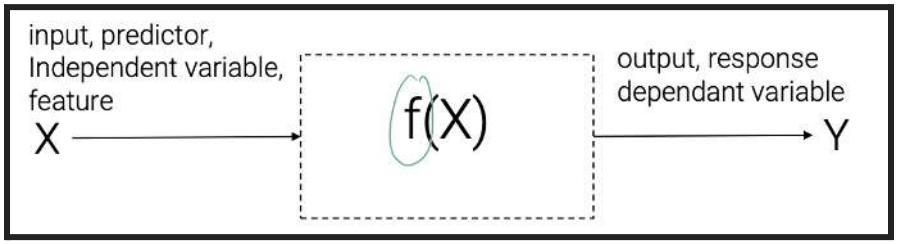
Note:

- Note what tutor says about deriving insights and recommendations based on log returns
- Also note or ask teacher how to read / analyse this stupid lookin relationship graph
- Thanks

- R is case sensitive!!!
- Choose directory of dataset

Lecture 4: Predictive Analysis I (Module 3)

- Estimating f(x)
    - Prediction
        - Predict outcomes of Y given X
        - What it means isn't as important, it just needs accurate predictions
        - Models tend to be more complex
    - Explanation / Inference
        - Understand how Y is affected by X
        - Which predictors do we add? How are they related?
        - Models tend to be simpler



Regression

- Assumption → **$Y = f(X) + \in$**
    - Y is the outcome, response, target variable
    - $X := X1, X2, \ldots, Xp$ are the features, inputs, predictors
    - ϵ is the error term capturing the measurement error and other discrepancies
- The Objective: Find an appropriate f for the problem at hand

How to estimate f(X)

- Parametric
    - Make an assumption about the shape of f
    - Problem reduced down to estimating a few parameters
    - Works fine with limited data, provided assumption are reasonable
    - Assumption strong: tends to miss some signal
    - More interpretable: works well in inference problems
- Non-Parametric
    - Make no assumption about the shape of f

- R is case sensitive!!!
- Choose directory of dataset

- Estimate f that gets as close to the data as possible without being too rough or too wiggly
- Need a large number of observations to obtain an accurate estimate for f
- Assumption weak: tends to incorporate some noise
- Less interpretable: may not work as well in inference problems

Simple Linear Regression

- Predict a quantitative response Y based on a single predictor variable X
- Approximately a linear relationship between X and Y

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Use (training data) to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- Make predictions given $X = x$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least Squares Method

- Most common approach to estimating $\hat{\beta}_0$ and $\hat{\beta}_1$
- Minimise the residual sum of squares (RSS)

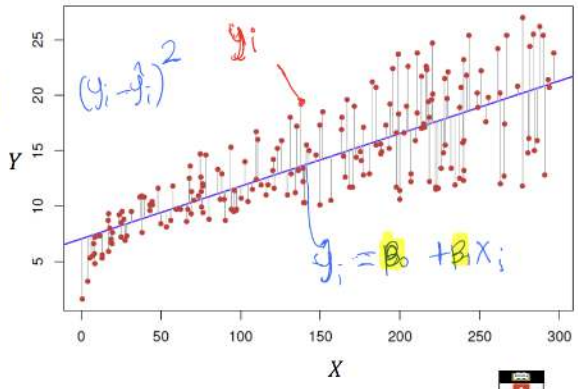
$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x)^2$$

- The least squares coefficient estimates are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

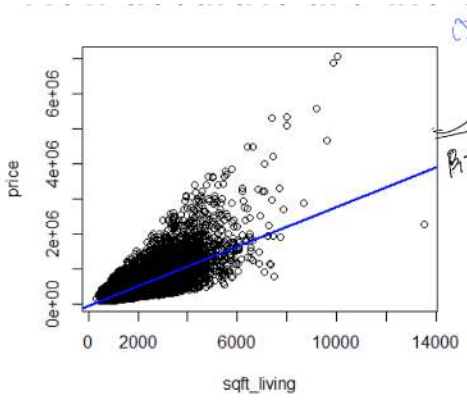
- where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$



UNSW SYDNEY

- R is case sensitive!!!
- Choose directory of dataset

Accuracy of Coefficient Measurements



Confidence interval

$\tilde{\beta}_1 \sim N(\beta_1, SE(\beta_1))$

A 95% confidence interval is given by

$$\left(\hat{\beta}_1 - 1.96 \times SE(\beta_1), \hat{\beta}_1 + 1.96 \times SE(\beta_1)\right)$$

So in this case

$$(281.251 - 1.96 \times 2.17, 281.251 + 1.96 \times 2.17)$$

$$(277, 285.5)$$

and for every additional square foot of living space the house prices will on average increase between 277 and 285 dollars.

Coefficient	Estimate	Standard error (SE)	p-value
β_0	-45461.104	4940.679	0.00
β_1	281.251	2.167	0.00

Relationship between X and Y

Hypothesis test

H_0: $\beta_1 = 0$ (There is no relationship between X and Y)

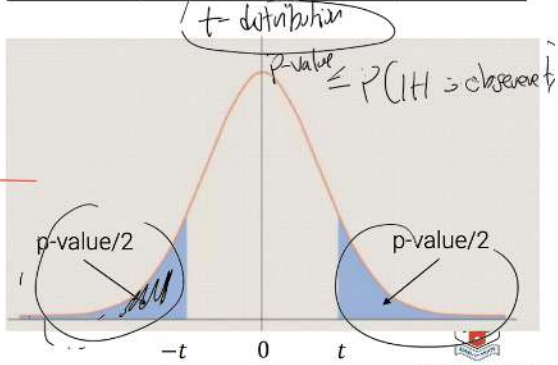
H_1: $\beta_1 \neq 0$ (There is some relationship between X and Y)

- Determine whether $\hat{\beta}_1$ is sufficiently far from 0
- Use t-statistic to determine how far is far enough

$$t = \frac{\hat{\beta}_1}{SE(\beta_1)}$$

- Under H_0 (i.e., if there is really no relationship between X and Y) the t-statistic will follow a t-distribution with $n-2$ degrees of freedom
- p-value: Assuming $\beta_1 = 0$, What is the probability of seeing any value equal to $|t|$ or larger

Coefficient	Estimate	Standard error (SE)	p-value
β_0	-45461.104	4940.679	0.00
β_1	281.251	2.167	0.00



- R is case sensitive!!!
- Choose directory of dataset

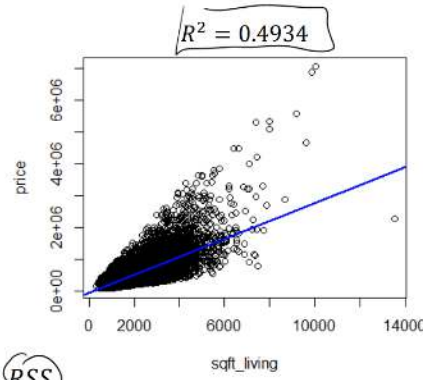
How well does the model fit the data?

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{TSS} = \underbrace{\left(\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right)}_{MSS} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{RSS}$$

Total sum of squares (TSS): Total variance of the response

Model sum of squares (MSS): Variability explained by the regression

Residual sum of Squares (RSS)



$$R^2 = \frac{\text{Variability explained by the regression}}{\text{Total variance of the response}} = \frac{MSS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

↳ proportion of variance explained by the model → (0,1)

UNSW SYDNEY

Qualitative predictors: Are houses with a waterfront more expensive?

dummy indicator $x_i = \begin{cases} 1 & \text{if the ith house has a waterfront} \\ 0 & \text{if the ith house does not have a waterfront} \end{cases}$

Average premium of house with a waterfront

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the ith house has a waterfront} \\ \beta_0 + \epsilon_i & \text{if the ith house does not have a waterfront} \end{cases}$$

- β_0 Average house price of houses without a waterfront
- $\beta_0 + \beta_1$: Average house price of houses with a waterfront

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

Coefficient	Estimate	Standard error (SE)	p-value
β_0	531684	2714	0.00
β_1	1187807	31296	0.00

p-value < 0.05 ⇒ reject H_0

UNSW SYDNEY

- R is case sensitive!!!
- Choose directory of dataset

Multiple Linear Regression

- Extend linear regression to accommodate multiple predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- β_j: Average effect on Y of a unit increase in X_j, holding all other predictors fixed

Multiple Linear Regression: House prices

Variable	estimate	std.error	p.value
(Intercept)	6229410.09	155782.56	0.00
bedrooms	-37775.03	2256.13	0.00
bathrooms	45090.30	3930.18	0.00
sqft_living	168.33	5.23	0.00
sqft_lot	-0.03	0.06	0.61
floors	31042.51	4252.11	0.00
waterfrontYes	636025.60	21032.02	0.00
view	39421.00	2561.74	0.00
condition	21117.38	2809.20	0.00
grade	120214.92	2531.63	0.00
sqft_above	-10.78	5.11	0.03
yr_built	-3597.09	79.92	0.00
yr_renovated	8.97	4.39	0.04
sqft_living15	27.73	4.04	0.00
sqft_lot15	-0.51	0.09	0.00

$R^2 = 0.6538$

F-Statistic: 23330, p-value=0

UNSW SYDNEY

- R is case sensitive!!!
- Choose directory of dataset

Multiple Linear Regression: House prices

Variable	estimate	std.error	p.value
(Intercept)	6229410.09	155782.56	0.00
bedrooms	-37775.03	2256.13	0.00
bathrooms	45090.30	3930.18	0.00
sqft_living	168.33	5.23	0.00
sqft_lot	-0.03	0.06	0.61
floors	31042.51	4252.11	0.00
waterfrontYes	636025.60	21032.02	0.00
view	39421.00	2561.74	0.00
condition	21117.38	2809.20	0.00
grade	120214.92	2531.63	0.00
sqft_above	-10.78	5.11	0.03
yr_built	-3597.09	79.92	0.00
yr_renovated	8.97	4.39	0.04
sqft_living15	27.73	4.04	0.00
sqft_lot15	-0.51	0.09	0.00

$R^2 = 0.6538$

F-Statistic: 23330, p-value=0 < 0.05 ⇒ Reject H_0

Some important questions:

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response? F-test
2. How well does the mode fit the data? R^2
3. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
4. How good is our model for making predictions on new data?

UNSW SYDNEY

Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?

hypothesis test

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ → no relationship. $Y = \beta_0 + \epsilon$

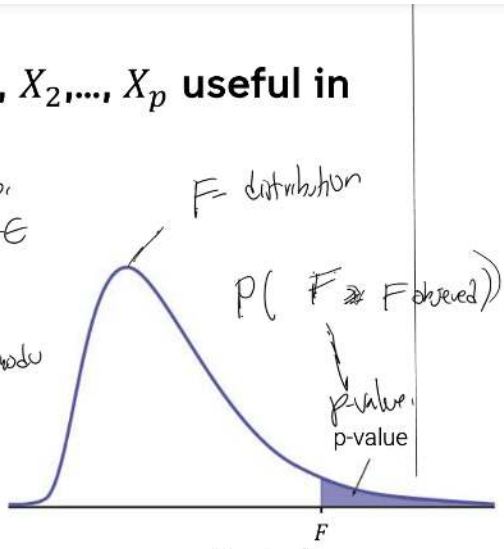
H_1: At least one of β_j is non-zero

- This hypothesis test is performed by computing the F-statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

RSS → variability explained by the model; RSS → unexplained variability

- The larger the value of F the more evidence there is to reject H_0
- p-value: Given $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, What is the probability of seeing any value larger than F



P-value < 0.05 ⇒ Reject H_0

UNSW SYDNEY

- R is case sensitive!!!
- Choose directory of dataset

Do all the predictors help to explain Y, or is only a subset of the predictors useful?

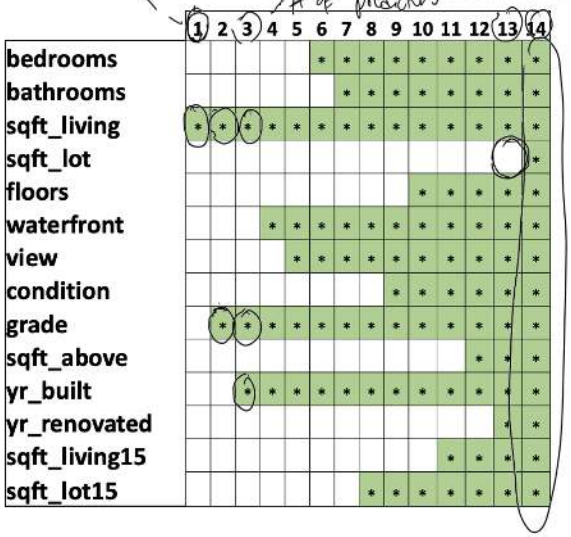
- Decide which variables to include in our model.
- Including all variables might result in an unnecessarily complex model.
- Try many different models an select the best model according to a given criterion
- Best subset approach:
  - Fit all possible models with p predictors
  - There are 2^p possible models
  - Feasible when p is not too big

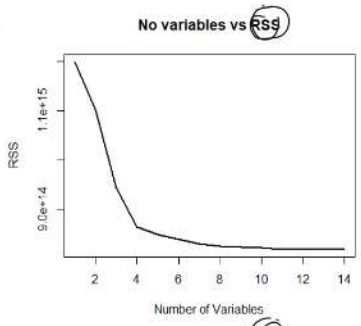
Possible models when $p = 3$

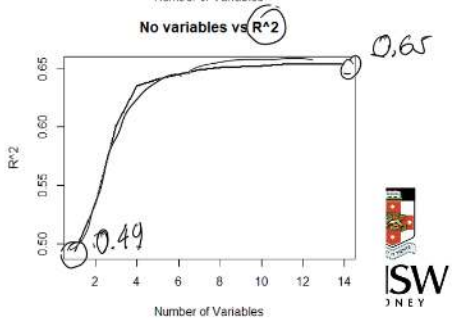
Number of Predictors	Models
0	$Y = \beta_0$
1	$Y = \beta_0 + \beta_1 X_1$, $Y = \beta_0 + \beta_2 X_2$, $Y = \beta_0 + \beta_3 X_3$
2	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3$, $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3$
3	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

UNSW SYDNEY

Best subset selection: House prices







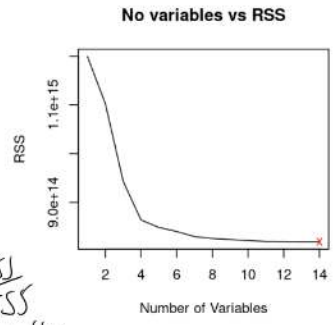
- R is case sensitive!!!
- Choose directory of dataset

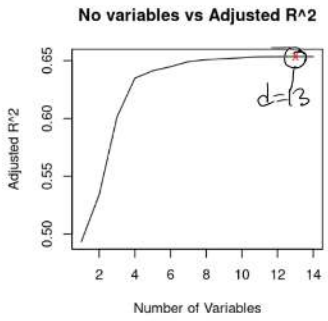
How to select the best Model?

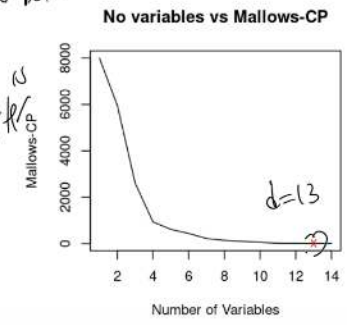
- The RSS and R^2 do not work as they always improve as we increase the number of variables
- Use alternative metrics which penalise the number of predictors (d)
- Adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$ } larger is better
- Mallows' $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$
- Bayesian informaion Criterion

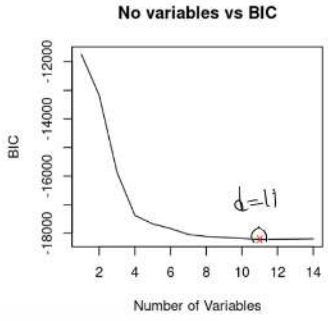
$$BIC = \frac{1}{n}(RSS + \log(n)\, d\hat{\sigma}^2)$$

} smaller is better



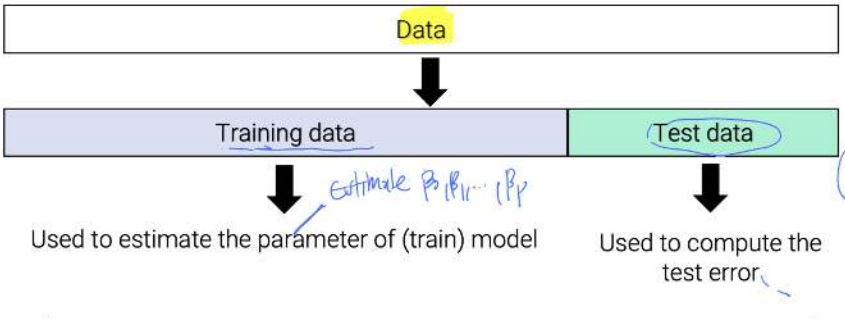






How good is our model for making predictions on new data?

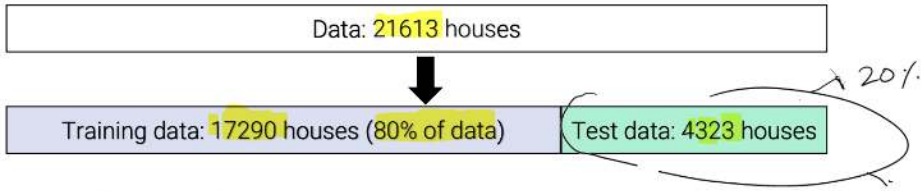
- In predicting modelling one of our objectives is to make predictions on new data
- We need to evaluate the prediction performance of the model on unseen data
- One approach to achieve this is using the validation set approach

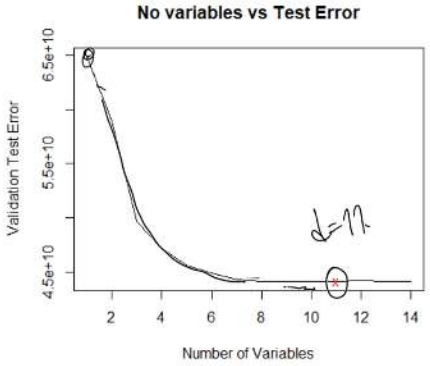


$$\text{Test mean square error } (MSE) = \frac{1}{\text{size of test data}} \sum_{i \in Test\ data} (y_i - \hat{y}_i)^2$$

- R is case sensitive!!!
- Choose directory of dataset

Validation set: House prices





- Having more variables doesn't necessarily improve the prediction performance
- Risk of overfitting the data
- The model with 11 variable produces the smallest test error

UNSW SYDNEY

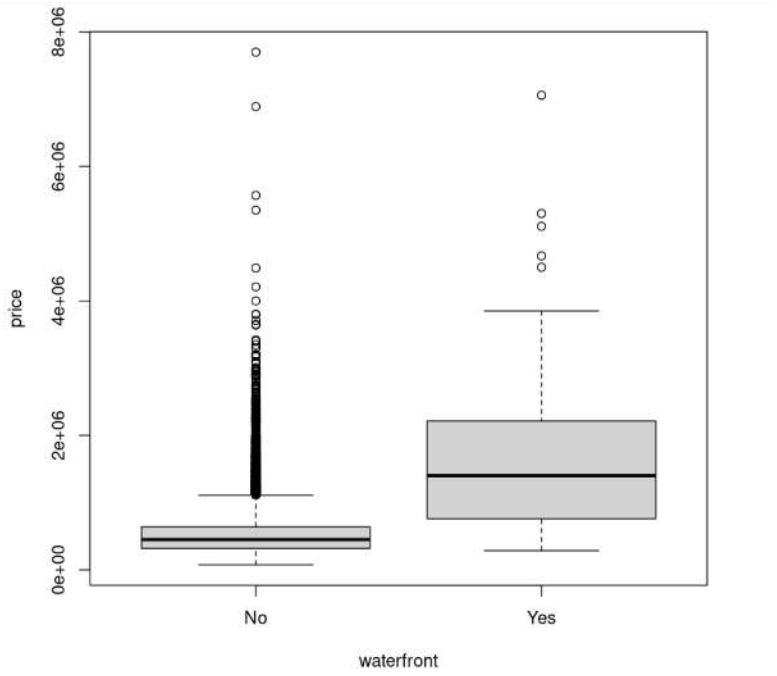
- Conclusion
  - Linear regression serves as the base for many more advanced predictive modelling and inference approaches
  - Important to have in mind what is the purpose of the modelling (prediction vs. inference)
  - There are many important things we haven't discussed
    - Allowing for non-linearities
    - Allowing for interactions
    - Checking modelling assumption (Especially important if the objective is inference)
    - How to fit and select the models when the number of variables is very very big (Especially important if the objective is prediction)

- R is case sensitive!!!
- Choose directory of dataset

Pre-Workshop 4: Predictive Analysis I (Module 3)

Exploratory Descriptive Analysis

- boxplot(price ~ waterfront, data = KCdata, xlab = "waterfront", ylab = "price")
    - price ~ waterfront: defines a *formula* telling R that we want to get a boxplot of the price per possible value of the variable waterfront.
    - data = KCdata: tells R that the data for the plots comes from KCdata



Define train and test data

- To validate the regression models and evaluate their predictive performance → divide the available data into a *train* dataset used to fit the models and a *test* dataset used for validation. We will use 80% of the data for training and 20% for validation:
    - ndata <- nrow(KCdata)

      set.seed(57)

      train <- sample(ndata, ndata*0.8)

      KCdataTrain <- KCdata[train,]

      KCdataTest <- KCdata[-train,]

- R is case sensitive!!!
- Choose directory of dataset

- Code will randomly select 80% of houses (17290) for *training* and 20% (4323) for *testing and validation*

Simple Linear Regression

- SLR of house price as a function of the house size

$$\texttt{price} = \beta_0 + \beta_1 \texttt{sqft\_living} + \epsilon$$

  - In R, linear regression models = *lm*

- lmArea <- lm(price ~ sqft_living, data = KCdataTrain)
  - price ~ sqft_living to specify that the model we are fitting is:

$$\texttt{price} = \beta_0 + \beta_1 \texttt{sqft\_living} + \epsilon$$

  - data = KCdataTrain to tell R that we are using the training data for model fitting.

- **summary(lmArea)** → Provides initial assessment
  - E.g., model coefficients, their standard errors, and the R^2 statistics

- Regression Coefficients
  - Regression coefficient = **coef(lmArea)**
  - Beta values found in summary (lmArea)
  - Confidence interval = **confint(lmArea)**
  - For Beta1, the 95% interval is (277,285.5) suggesting that Beta1 does not equal 0

We see that $\hat{\beta}_0 = -45461.104$ and $\hat{\beta}_1 = 281.251$ so that the fitted model is

$$\texttt{price} = -45461 + 281 \times \texttt{sqft\_living} + \epsilon.$$

The interpretation of $\hat{\beta}_1$ is that for every additional squeare foot of living space the house prices will on average increase by 281 dollars.

It is important to understand whether the coefficients are statistically significant.

Recall that a 95% confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm 1.96 SE(\beta_1)$$

where, $SE(\beta_1)$ is the standard error of β_1. $SE(\beta_1)$ can be found next to the coefficient estimate in the output of `summary(lmArea)`.

- R is case sensitive!!!
- Choose directory of dataset

Model Accuracy

- To determine model accuracy, first plot regression line
    - plot(KCdataTrain$sqft_living, KCdataTrain$price, xlab = "sqft_living", ylab = "price", main = "Train")
      abline(lmArea, col = "blue", lwd = 2)
- Abline function → draw any line, (not just the least square regression line)
- To draw a line with intercept a and slope b → abline(a, b)

- Assess the model fit quantitatively looking at the R^2, which represents the proportion of variance explained by the model.
    - Found at the bottom of summary(lmArea)
    - Also accessed by: *summary(lmArea)$r.squared*
    - Or square of correlation between X and Y: cor(KCdataTrain$sqft_living, KCdataTrain$price)^2

Qualitative Predictors

- Can create a dummy variable for qualitative data (E.g. Yes or No)

$$x_i = \begin{cases} 1 & \text{if the ith house has a waterfront} \\ 0 & \text{if the ith house does not have a waterfront,} \end{cases}$$

- Then use as predictor in regression equation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the ith house has a waterfront} \\ \beta_0 + \epsilon_i & \text{if the ith house does not have a waterfront,} \end{cases}$$

    - *β1* = average house price of houses without a waterfront
    - *β2* = average house price of houses with a waterfront

- R is case sensitive!!!
- Choose directory of dataset

- No need for dummy variables in R → automatically for qualitative variables in summary
    - lmWaterfront <- lm(price ~ waterfront, data = KCdataTrain)
      summary(lmWaterfront)

```
Call:
lm(formula = price ~ waterfront, data = KCdataTrain)

Residuals:
     Min       1Q   Median       3Q      Max
-1434491  -211684   -81684   108316  6358316

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    531684       2714  195.93   <2e-16 ***
waterfrontYes 1187807      31296   37.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 355500 on 17288 degrees of freedom
Multiple R-squared:  0.07692,   Adjusted R-squared:  0.07686
F-statistic:  1441 on 1 and 17288 DF,  p-value: < 2.2e-16
```

From this output we see that $\hat{\beta}_0 = 531684$ and $\hat{\beta}_1 = 1187807$, so houses without a waterfront have an average price of $531,684 and houses with waterfront a price of $531,684 + $1,187,807= $1,719,491.

- R is case sensitive!!!
- Choose directory of dataset

Multiple Linear Regression

- Case Study: Price as a function of area and waterfront

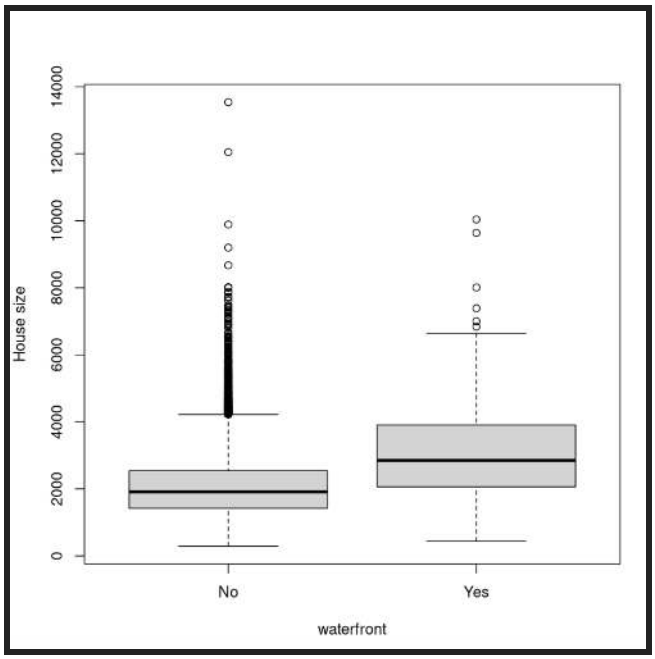
$$\texttt{price}_i = \beta_0 + \beta_1 \texttt{sqft\_living}_i + \beta_0 \texttt{waterfront}_i + \epsilon_i.$$

  - R Code: lmAreaWater <- lm(price ~ sqft_living + waterfront, data = KCdataTrain)
  - Assess the model: summary(lmAreaWater)

```
Call:
lm(formula = price ~ sqft_living + waterfront, data = KCdataTrain)

Residuals:
     Min       1Q   Median       3Q      Max
-1381623  -142238   -21921   105857  4231391

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -32937.221   4755.413  -6.926 4.47e-12 ***
sqft_living      272.148      2.095 129.918  < 2e-16 ***
waterfrontYes 857104.493  22407.125  38.251  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 252900 on 17287 degrees of freedom
Multiple R-squared:  0.5329,    Adjusted R-squared:  0.5329
F-statistic:  9863 on 2 and 17287 DF,  p-value: < 2.2e-16
```

- Sqft_living coefficient: every one sqft increase in the house size (*holding all other predictors constant*) → expect price increase by $272.
- Similarly expect the waterfront to increase house price by $857,104.
- In MLR, the coefficient associated with waterfront (857104.493) → smaller than SLR (1187807).
  - Confounding (one variable influences others), since houses with waterfront tend to be of a bigger size:
  - R^2 also higher in MLR model

- R is case sensitive!!!
- Choose directory of dataset



Models with All Variables

- lmAll <- lm(price ~ ., data = KCdataTrain)
  summary(lmAll)
- The smaller the p-value, the stronger the evidence that you should reject the null hypothesis. **A p-value less than 0.05 (typically ≤ 0.05) is statistically significant**.

Variable Selection with best subset approach

- Best Subset Approach (Number of predictors not too big)

Assumme that there are p predictors. In this approach, we fit all the possible models with 0 predictors, 1 predictor, 2 predictors, 3 predictors, ... and p predictors. For example, if $p = 3$ we consider the following $2^3 = 8$ models:

Number of Predictors	Models
0	$Y = \beta_0$
1	$Y = \beta_0 + \beta_1 X_1, Y = \beta_0 + \beta_2 X_2, Y = \beta_0 + \beta_3 X_3$
2	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2, Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3, Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3$
3	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

- R is case sensitive!!!
- Choose directory of dataset

To determine which model is best we can use different criteria including Mallow's C_p, the Bayesian Information criterion (BIC) and the adjusted R^2. These are defined as:

Mallow's C_p

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

BIC

$$BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$$

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

where n is the numbers of observations in the dataset, d is the number of predictors in the model, $\hat{\sigma}_2$ is an estimate of the variance of ϵ, and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares.

Essentially, all of these metrics penalise the model for the number of parameters included in the model. In the case of the Mallow's C_p and BIC, the smallest the value, the better. By contrast, in the case of the Adjusted R^2, the bigger the value the better.

- Best Subset Selection
    - The regsubsets function (part of the leaps library) performs best subset selection by identifying the best model that contains a given number of predictors, where best is quantified using *RSS*. The syntax is the same as for lm. The summary command outputs the best set of variables for each model size.
    - library(leaps)
      bestSubset <-regsubsets(price ~ ., data = KCdataTrain, nvmax = 14)
      bestSubsetSummary <- summary(bestSubset)
      bestSubsetSummary

- R is case sensitive!!!
- Choose directory of dataset

```
Selection Algorithm: exhaustive
          bedrooms bathrooms sqft_living sqft_lot floors waterfrontYes view
1  ( 1 )  " "      " "       "*"         " "      " "    " "           " "
2  ( 1 )  " "      " "       "*"         " "      " "    " "           " "
3  ( 1 )  " "      " "       "*"         " "      " "    " "           " "
4  ( 1 )  " "      " "       "*"         " "      " "    "*"           " "
5  ( 1 )  " "      " "       "*"         " "      " "    "*"           "*"
6  ( 1 )  "*"      " "       "*"         " "      " "    "*"           "*"
7  ( 1 )  "*"      "*"       "*"         " "      " "    "*"           "*"
8  ( 1 )  "*"      "*"       "*"         " "      " "    "*"           "*"
9  ( 1 )  "*"      "*"       "*"         " "      " "    "*"           "*"
10  ( 1 ) "*"      "*"       "*"         " "      "*"    "*"           "*"
11  ( 1 ) "*"      "*"       "*"         " "      "*"    "*"           "*"
12  ( 1 ) "*"      "*"       "*"         " "      "*"    "*"           "*"
13  ( 1 ) "*"      "*"       "*"         " "      "*"    "*"           "*"
14  ( 1 ) "*"      "*"       "*"         "*"      "*"    "*"           "*"

          condition grade sqft_above yr_built yr_renovated sqft_living15
1  ( 1 )  " "       " "   " "        " "      " "          " "
2  ( 1 )  " "       "*"   " "        " "      " "          " "
3  ( 1 )  " "       "*"   " "        "*"      " "          " "
4  ( 1 )  " "       "*"   " "        "*"      " "          " "
5  ( 1 )  " "       "*"   " "        "*"      " "          " "
6  ( 1 )  " "       "*"   " "        "*"      " "          " "
7  ( 1 )  " "       "*"   " "        "*"      " "          " "
8  ( 1 )  " "       "*"   " "        "*"      " "          " "
9  ( 1 )  "*"       "*"   " "        "*"      " "          " "
10  ( 1 ) "*"       "*"   " "        "*"      " "          " "
11  ( 1 ) "*"       "*"   " "        "*"      " "          "*"
12  ( 1 ) "*"       "*"   "*"        "*"      " "          "*"
13  ( 1 ) "*"       "*"   "*"        "*"      "*"          "*"
14  ( 1 ) "*"       "*"   "*"        "*"      "*"          "*"
```

- An asterisk (*) indicates that a given variable is included in the corresponding model. For instance, this output indicates that the best 1 variable model contains only sqft_living, while the best two-variable model contains sqft_living and grade
- By default, regsubsets only reports results up to the best eight-variable model. But the nvmax option can be used in order to return as many variables as are desired. In the code above we fitted up to a 14-variable model by setting nvmax = 14.

- R is case sensitive!!!
- Choose directory of dataset

- The summary function also returns the *RSS*, adjusted R^2, Cp and BIC. We can examine these to try to select the best overall model. Let's start by looking at the adjusted R^2
    - bestSubsetSummary$adjr2
    - We see that we start with a model with an adjusted R^2 of 0.49 when only one variable is included and end with a model with an adjusted R^2 of 0.65 when all 14 variables are included

- To find which model size **maximises** the adjusted R^2 we can use the function `which.max`:
    - best_adjr2 <- which.max(bestSubsetSummary$adjr2)
    - According to this criterion the best model is the one with 13 variables

- Similarly, we can use function `which.min` to find the model size which minimises the RSS, Cp and BIC:
    - best_rss <- which.min(bestSubsetSummary$rss) *(Residual sum of squares)*
    - best_cp <- which.min(bestSubsetSummary$cp) *(Mallow's CP)*
    - best_bic <- which.min(bestSubsetSummary$bic) *(BIC)*
    - best_rss; best_cp; best_bic

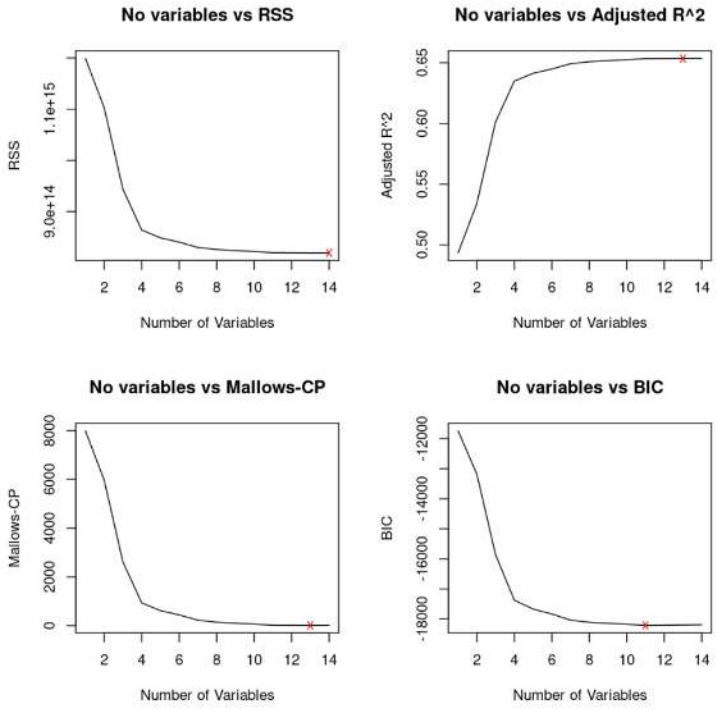
- R is case sensitive!!!
- Choose directory of dataset

We see that the C_p criterion selects also the model with 13 variables while the BIC criterion a simpler model with only 11 variables. Can you explain why the RSS is minimised for a model with 14 variables?

We can present graphically these results with the commands below:

[24]
```
#Plot selection criteria against model size
par(mfrow = c(2,2))
plot(bestSubsetSummary$rss, xlab = "Number of Variables", ylab = "RSS",
     main = "No variables vs RSS", type = "l")
points(best_rss, bestSubsetSummary$rss[best_rss], col = "red", pch ="x")
plot(bestSubsetSummary$adjr2, xlab = "Number of Variables", ylab = "Adjusted R^2",
     main = "No variables vs Adjusted R^2", type = "l")
points(best_adjr2, bestSubsetSummary$adjr2[best_adjr2], col = "red", pch ="x")

plot(bestSubsetSummary$cp, xlab = "Number of Variables", ylab = "Mallows-CP",
     main = "No variables vs Mallows-CP", type = "l")
points(best_cp, bestSubsetSummary$cp[best_cp], col = "red", pch ="x")

plot(bestSubsetSummary$bic, xlab = "Number of Variables", ylab = "BIC",
     main = "No variables vs BIC", type = "l")
points(best_bic, bestSubsetSummary$bic[best_bic], col = "red", pch ="x")
```

[24]



- We note that in the code above the command `par(mfrow = c(2,2))` is telling R that we want to have 2x2=4 plots in the same graph
- The candidate models are those with 11 and 13 variables. The variables included in each of this model can be seen with function;
    - coef(bestSubset, 11)
    - coef(bestSubset, 13)

- R is case sensitive!!!
- Choose directory of dataset

Predictive Performance with validations set approach

- Now that we have selected the models with 11 and 13 variables as our potential candidates, we want to evaluate their **predictive** performance using the test data:
    - First fit the two models using the training data (`KCdataTrain`):
    - lmPrice_11 <- lm(price ~ bedrooms + bathrooms + sqft_living + floors + waterfront + view + condition + grade + yr_built + sqft_living15 + sqft_lot15, data = KCdataTrain)
    - lmPrice_13 <- lm(price ~ bedrooms + bathrooms + sqft_living + floors + waterfront + view + condition + grade + sqft_above + yr_built + yr_renovated + sqft_living15 + sqft_lot15, data = KCdataTrain)

- Predictions
    - To obtain predictions for a model we can use the function `predict`, where the argument `newdata` specifies the data on which we want to make predictions:
    - predPriceTest_1 <- predict(lmArea, newdata = KCdataTest)
    - predPriceTest_11 <- predict(lmPrice_11, newdata = KCdataTest)
    - predPriceTest_13 <- predict(lmPrice_13, newdata = KCdataTest)
    - predPriceTest_All <- predict(lmAll, newdata = KCdataTest)

We can compute the test mean square error as:

$$\text{Test } MSE = \frac{1}{\text{size of test data}} \sum_{i \in \text{Test Data}} (y_i - \hat{y}_i)^2$$

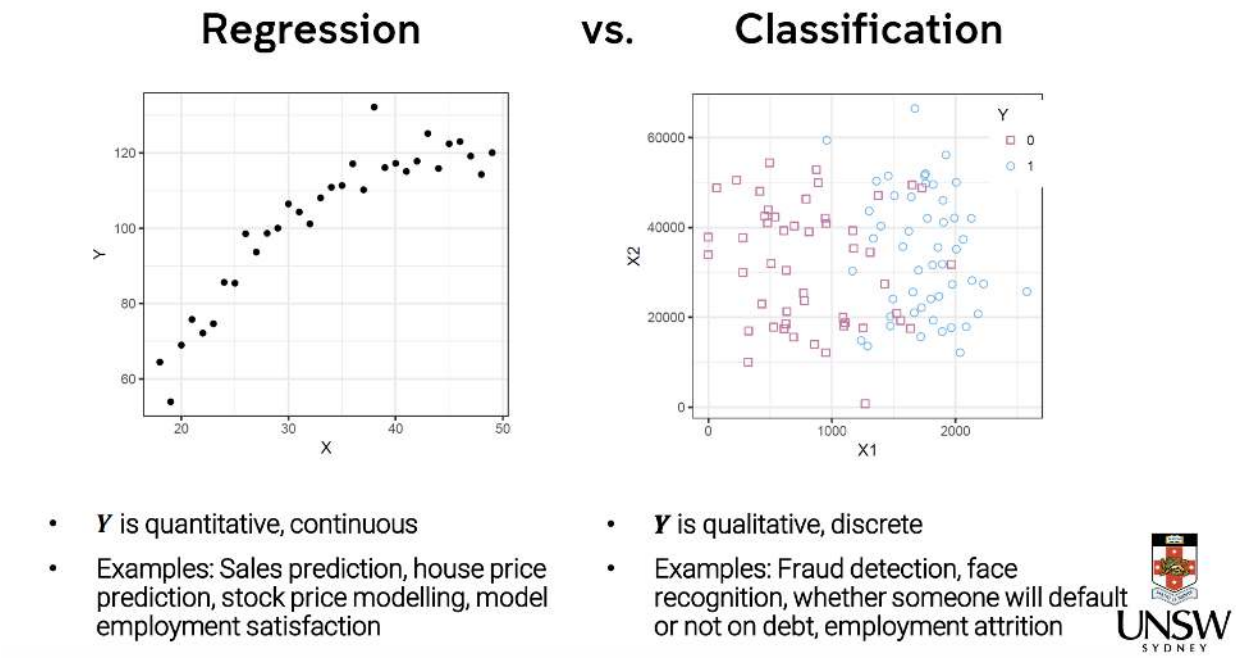
. In R, we can compute the test MSE for the models as follows:

```
[28] MSE_1 <- mean((KCdataTest$price - predPriceTest_1)^2)
     MSE_11 <- mean((KCdataTest$price - predPriceTest_11)^2)
     MSE_13 <- mean((KCdataTest$price - predPriceTest_13)^2)
     MSE_All <- mean((KCdataTest$price - predPriceTest_All)^2)
     MSE_1; MSE_11; MSE_13; MSE_All
```

- We see that the model with the smallest test MSE is the one with 11 variables. Therefore, for a prediction task, we would be inclined to select that model among the candidate models.

- R is case sensitive!!!
- Choose directory of dataset

Lecture 5: Predictive Analysis II (Module 3)



- Classification Problems
    - Medicine: Success/failure of a treatment, explained by dosage of medicine administered, patient's age, sex, weight and severity of condition, etc.
    - Politics: Vote for/against a political party, explained by age, gender, education level, region, ethnicity, geographical location, etc.
    - Business: Customer churns/stays depending on usage pattern, complaints, social demographics, etc

- R is case sensitive!!!
- Choose directory of dataset

- Coding in the binary case is easy

$$Y \in \{\bullet, \bullet\} \leftrightarrow Y = \{1, 0\}$$

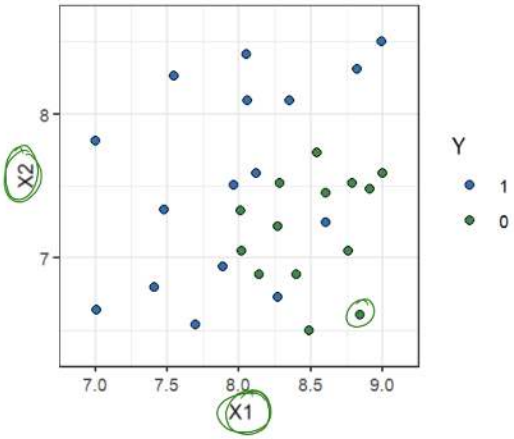
- Our objective is to find a good predictive model f that given X can

1. Classify observations

$$f(X) \rightarrow \hat{Y} \in \{\bullet, \bullet\}$$

2. Estimate the probability of being of a given class

$$f(X) \rightarrow$$

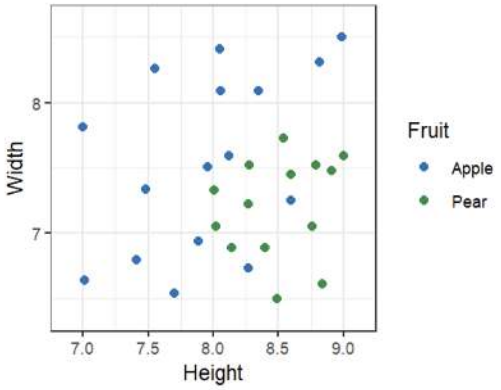


- Example: Distinguishing Apples and Pears

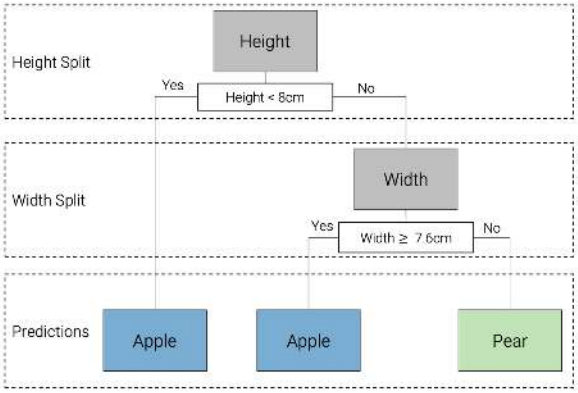
- Apples tend to be short and fat
- Pears are usually taller and more slight
- Design a method or algorithm to differentiate between apples and pears based only on their widths and heights.

$$Y \in \{\bullet, \bullet\} \leftrightarrow Y = \{\text{Apple}, \text{Pear}\}$$

$$f(\text{Height}, \text{Width}) \rightarrow \hat{Y} \in \{\bullet, \bullet\}$$

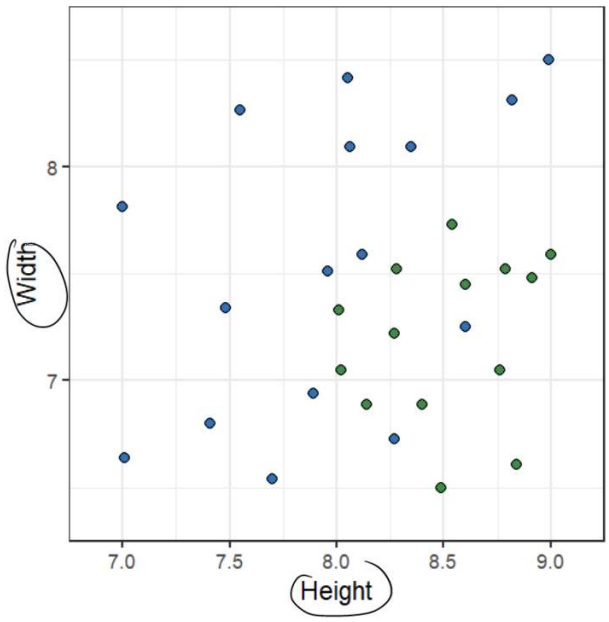


- **Classification Trees**
  - Set of rules based on input variables to make predictions summarised in a tree
  - Simple interpretation / Closely resemble human decision making
  - Useful for communicating predictions and results to non -technical stakeholders



- R is case sensitive!!!
- Choose directory of dataset

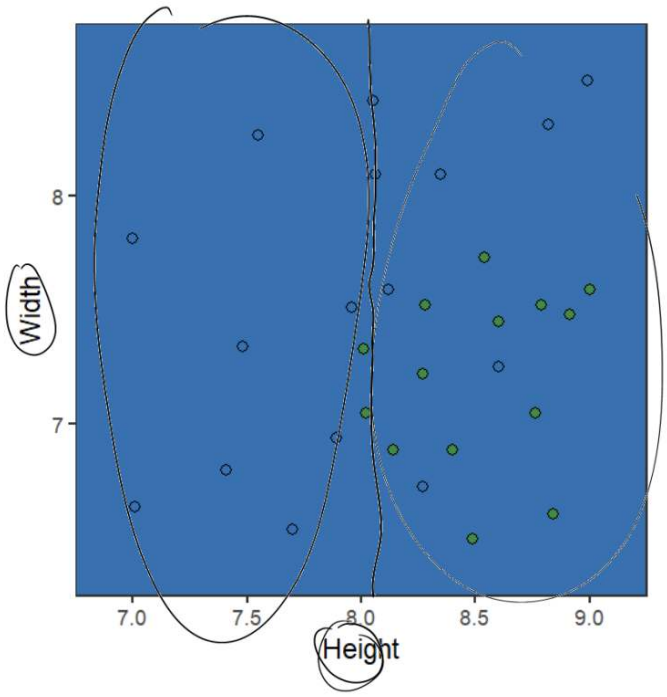
Tree-Building Process



30 Fruits
16 Apples → 16/30
14 Pears → 14/30

$\frac{16}{30} > \frac{14}{30}$

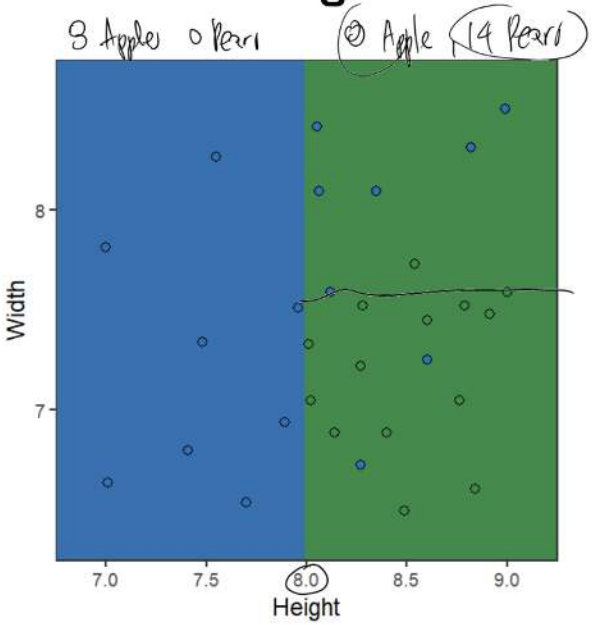
Tree-Building Process

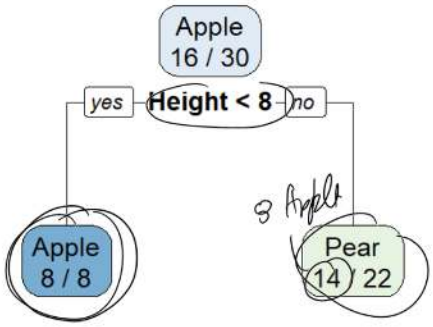


Apple
16 / 30

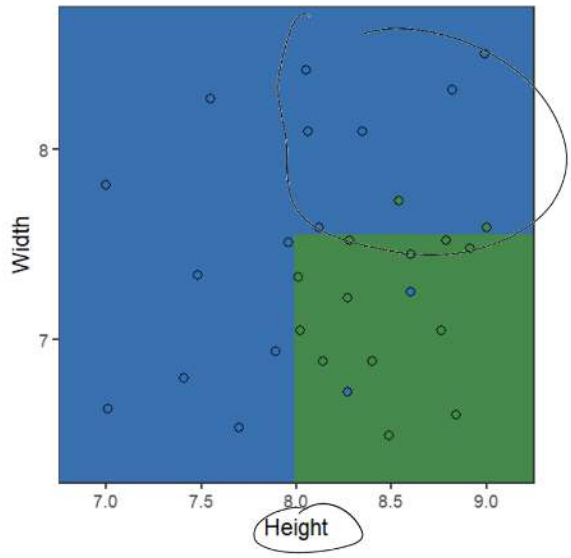
- R is case sensitive!!!
- Choose directory of dataset

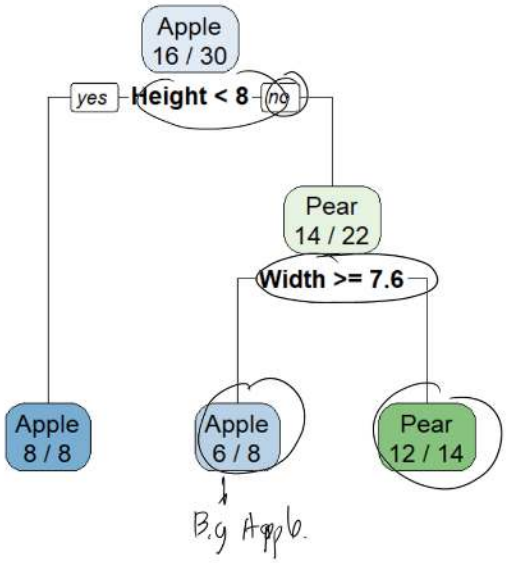
Tree-Building Process





Tree-Building Process





- R is case sensitive!!!
- Choose directory of dataset

Assessing Accuracy in Classification Problems

- In regression, we use Mean Squared Error
- We assess model accuracy using the accuracy rate:
  - $Accuracy\ Rate\ =\ \frac{Correct\ Predictions}{Number\ of\ Points}$
- In previous fruit example:
  - $Accuracy\ Rate\ =\ \frac{30-2-2}{30}\ =\ 86.7\%$

Confusion Matrix

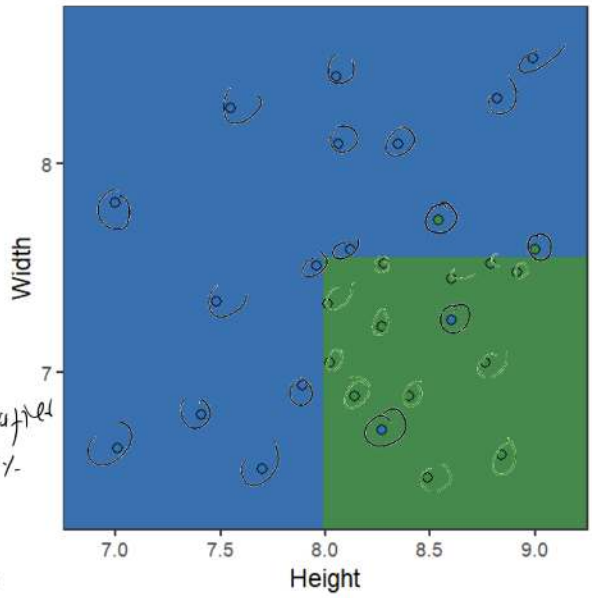
		Actual	
		Apple	Pear
Predicted	Apple	14	2
	Pear	2	12

16

$$\text{Accuracy rate} = \frac{14+12}{30} = \frac{26}{30} = 0.867 \approx 87\%$$

$$\text{Accuracy rate for Apple} = \frac{14}{16} = 0.875 \approx 87.5\%$$

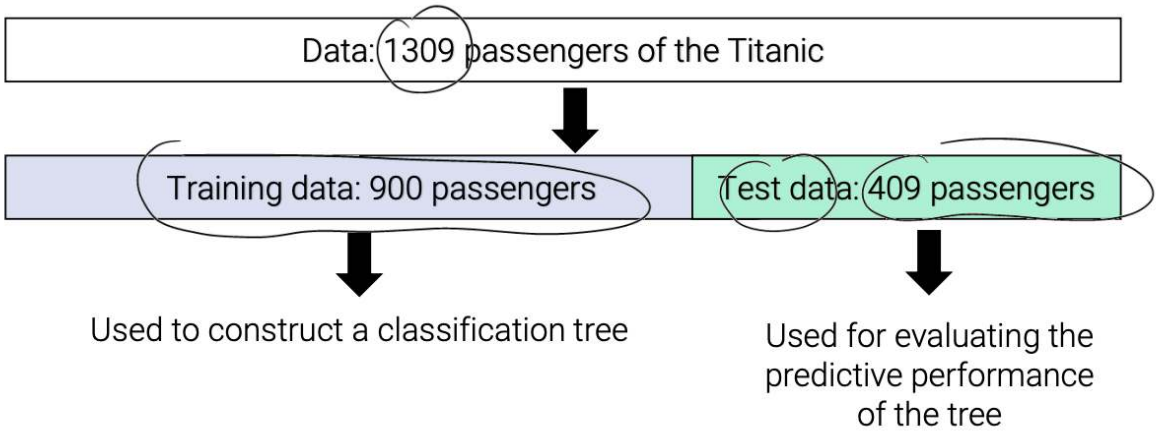
$$\text{Accuracy rate for Pear} = \frac{12}{14} = 0.857 \approx 86\%$$



Titanic Survival Case Study

- What people are likely to survive the Titanic
  - Output Y (Categorical):
    - Survived, perished
  - Input X
    - Ticket class
    - Sex
    - Age
    - # of siblings aboard
    - # of parents abroad
    - Fare
    - Cabin Number
    - Port of embarkation

- R is case sensitive!!!
- Choose directory of dataset

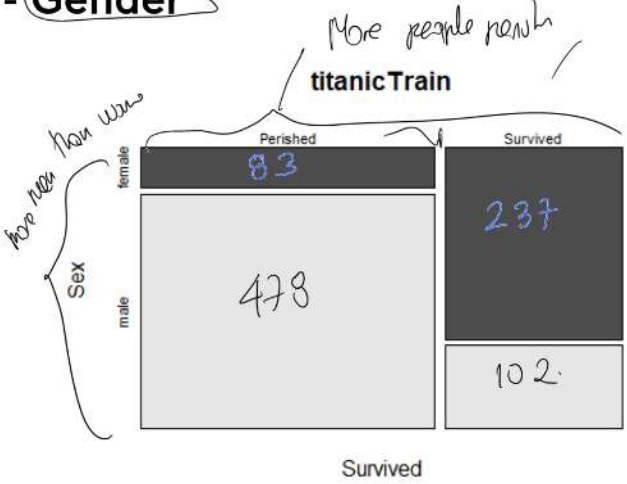


Exploratory data analysis - Gender

Passenger survival by gender

	Female	Male	Total
Perished	83	478	561
Survived	237	102	339
Total	320	580	900

Survival rate: $\frac{237}{320} > \frac{102}{580}$ $\frac{339}{900}$



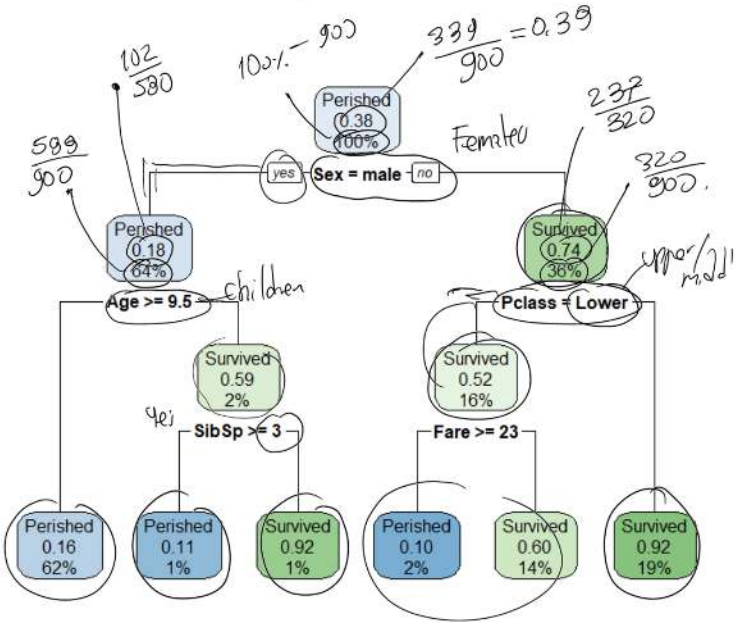
Mosaic plots are a good way to explore the relationship between two categorical variables

- R is case sensitive!!!
- Choose directory of dataset

What sorts of people were more likely to survive in the Titanic?

Passenger survival by gender

	Female	Male	Total
Perished	83	478	561
Survived	237	102	339
Total	320	580	900



Confusion matrix for Titanic survival – Classification tree

Training data (900 passengers)

		Actual	
		Perished	Survived
Predicted	Perished	496	92
	Survived	65	247

$$\text{Accuracy rate} = \frac{496 + 247}{900} = 0.826$$

$$\text{Accuracy rate for Perished} = \frac{496}{496 + 65} = 0.884$$

$$\text{Accuracy rate for Survived} = \frac{247}{247 + 92} = 0.729$$

Test data (409 passengers)

		Actual	
		Perished	Survived
Predicted	Perished	213	48
	Survived	34	114

$$\text{Accuracy rate} = \frac{213 + 114}{409} = 0.800$$

$$\text{Accuracy rate for Perished} = \frac{213}{213 + 34} = 0.862$$

$$\text{Accuracy rate for Survived} = \frac{114}{114 + 48} = 0.704$$

- R is case sensitive!!!
- Choose directory of dataset

Logistic regression

- Extend linear regression to model binary categorical variables

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

log-odds | linear model

Principles of Logistic Regression

$y = \begin{cases} 1 & \text{survived} \\ 0 & \text{if perishes} \end{cases}$

- The output is binary $Y = \{1, 0\}$
- Each case's Y variable has a probability between 0 and 1 that depends on the values of the predictors X such that

$$P(Y=1) + P(Y=0) = 1$$

$P(y=0) = 1 - P(y=1)$

- Probability can be restated as odds

$$\text{Odds}(Y=1) = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}$$

- Odds are a measure of relative probabilities — survived to die

- R is case sensitive!!!
- Choose directory of dataset

A logistic regression for Apples and Pears

$$Y = \begin{cases} 1 & \text{if Apple} \\ 0 & \text{if Pear} \end{cases}$$

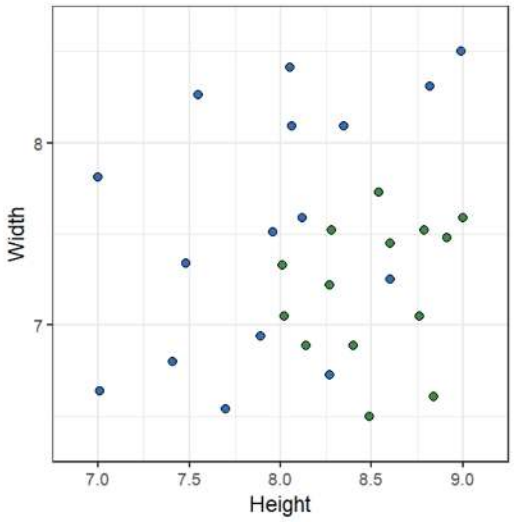
$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \times \text{Height} + \beta_2 \times \text{Width}$$

log odds = linear function of H, W

- The parameter estimates are

$$\hat{\beta}_0 = 13.621, \hat{\beta}_1 = -4.136, \hat{\beta}_2 = 2.803$$

- $\hat{\beta}_1 = -4.136$ implies that the taller the fruit the lower the chance it is an apple
- $\hat{\beta}_2 = 2.803$ implies that the wider the fruit the higher the chance it is an apple



$\frac{1}{1+e^{-}}$

A logistic regression for Apples and Pears

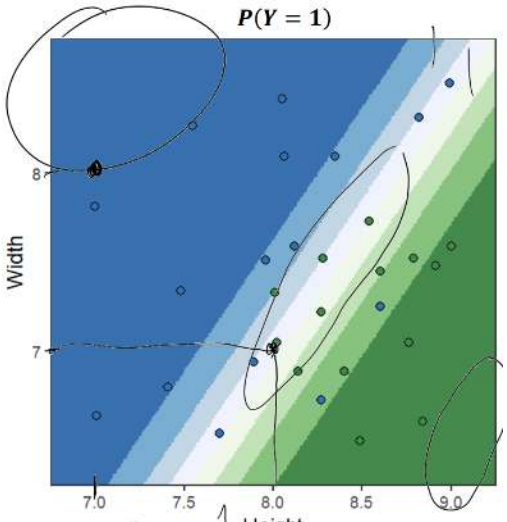
- The estimated logistic regression is

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 13.621 - 4.136 \times \text{Height} + 2.803 \times \text{Width}$$

- The probability of a given fruit being an apple is then

$$P(Y=1) = \frac{1}{1+\exp[-(13.621 - 4.136 \times \text{Height} + 2.803 \times \text{Width})]}$$

Height	Width	$13.621 - 4.136 \times \text{Height} + 2.803 \times \text{Width}$	$P(Y=1)$	Prediction
7	8	7.14	0.9992	Apple
8	7.5	1.60	0.8318	Apple
8	7	0.20	0.5492	Apple
8.5	7.5	-0.47	0.3847	Pear
9	7	-3.94	0.0191	Pear



$P(Y=1) = 0.9997 = \frac{1}{1-e^{7.14}}$ =

SYDNEY

- R is case sensitive!!!
- Choose directory of dataset

A logistic regression for Apples and Pears

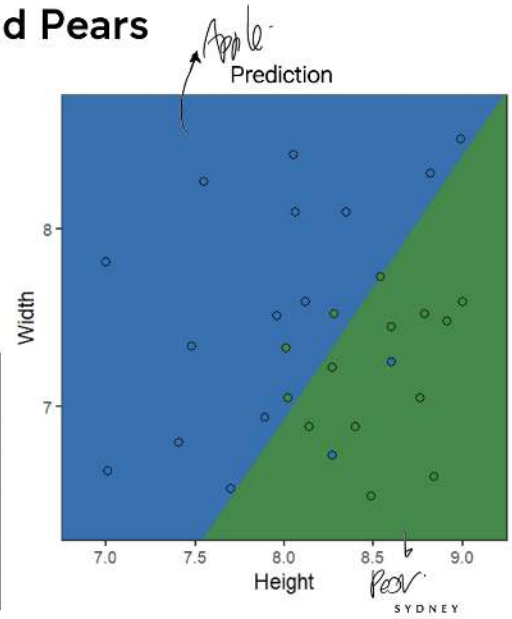
- The estimated logistic regression is

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 13.621 - 4.136 \times \text{Height} + 2.803 \times \text{Width}$$

- The probability of a given fruit being an apple is then

$$P(Y=1) = \frac{1}{1+\exp[-(13.621 - 4.136 \times \text{Height} + 2.803 \times \text{Width})]}$$

Height	Width	$\mathbf{13.621 - 4.136 \times Height + 2.803 \times Width}$	$\boldsymbol{P(Y=1)}$	Prediction
7	8	7.14	0.9992	Apple
8	7.5	1.60	0.8318	Apple
8.	7	0.20	0.5492	Apple
8.5	7.5	-0.47	0.3847	Pear
9	7	-3.94	0.0191	Pear



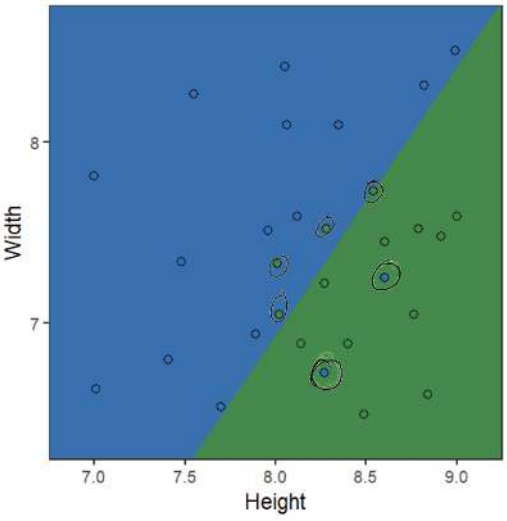
Confusion matrix

		Actual	
		Apple •	Pear •
Predicted	Apple •	14	4
	Pear •	2	10

$$\text{Accuracy rate} = \frac{14+10}{30} = \frac{24}{30} = 0.80$$

$$\text{Accuracy rate for Apples} = \frac{14}{16} = 0.875$$

$$\text{Accuracy rate for Pears} = \frac{10}{14} = 0.714$$



- R is case sensitive!!!
- Choose directory of dataset

A logistic regression for the survival of Titanic passengers

Response (Y)

$$Y = \begin{cases} 1 & \text{if Survived} \\ 0 & \text{if Perished} \end{cases}$$

Inputs (X)

Variable	Definition	Key	Variable type
Pclass	Ticket class	"Lower", "Middle", "Upper"	Categorical with 3 levels
Sex	Sex	"male", "female"	Categorical with 2 levels
Age	Age in years		Numerical
Sibsp	# of siblings		Numerical
Parch	# of parents		Numerical
Fare	Passenger fare		Numerical
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton	Categorical with 3 levels

How can we code the categorical inputs for use in the regression?

Coding of categorical variables

	Pclass m	Pclass u
lower	0	0
middle	1	0
Upper	0	1

Sex (Two levels)

$$\text{Sex} = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

Pclass (Three levels) We need two dummy variables.

$$\text{Pclass}_m = \begin{cases} 1 & \text{if Middle} \\ 0 & \text{Otherwise} \end{cases} \qquad \text{Pclass}_u = \begin{cases} 1 & \text{if Upper} \\ 0 & \text{Otherwise} \end{cases}$$

Embarked (Three levels)

$$\text{Embarked}_Q = \begin{cases} 1 & \text{if Queenstown} \\ 0 & \text{Otherwise} \end{cases} \qquad \text{Embarked}_S = \begin{cases} 1 & \text{if Southampton} \\ 0 & \text{Otherwise} \end{cases}$$

- R is case sensitive!!!
- Choose directory of dataset

A logistic regression for the survival of Titanic passengers

$\log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

Variable	estimate	std.error	p.value
(Intercept)	2.14	0.37	0.00
PclassMiddle	+0.98	0.23	0.00
PclassUpper	+2.05	0.31	0.00
Sexmale	-2.74	0.20	0.00
Age	-0.05	0.01	0.00
SibSp	-0.30	0.11	0.01
Parch	-0.10	0.12	0.38
Fare	0.00	0.00	0.14
EmbarkedEmbarkQ	-0.18	0.37	0.62
EmbarkedEmbarkS	-0.36	0.24	0.13

$\beta_j \neq 0$

$\beta_j = 0$

$H_0: \beta_j = 0$
$H_1: \beta_j \neq 0$
} p.value < 0.05 $\Rightarrow$ Reject H_0

- Positive parameter estimates imply that the odds of surviving increase
  - E.g. The odds of passengers in Middle class are higher than those of passengers in lower class
- Negative parameter estimates imply that the odds of surviving decrease
  - E.g. The odds of males surviving are lower than those of women
- As in linear regression standard errors can be to construct confidence intervals
- p-values can be used to evaluate significance of variables
  - Parch, Fare and Embarked are not statistically significant as they have high p-values

Confusion matrix for Titanic survival – Logistic regression

Training data (900 passengers)

		Actual	
		Perished	Survived
Predicted	Perished	479	101
	Survived	82	238

$$\text{Accuracy rate} = \frac{479 + 238}{900} = 0.797$$

Test data (409 passengers)

		Actual	
		Perished	Survived
Predicted	Perished	206	53
	Survived	41	109

$$\text{Accuracy rate} = \frac{206 + 109}{409} = 0.779$$

- Recall that the training and testing accuracy rate for the classification tree were 0.826 and 0.800, respectively
- A classification tree seems to perform better than a logistic regression

UNSW SYDNEY

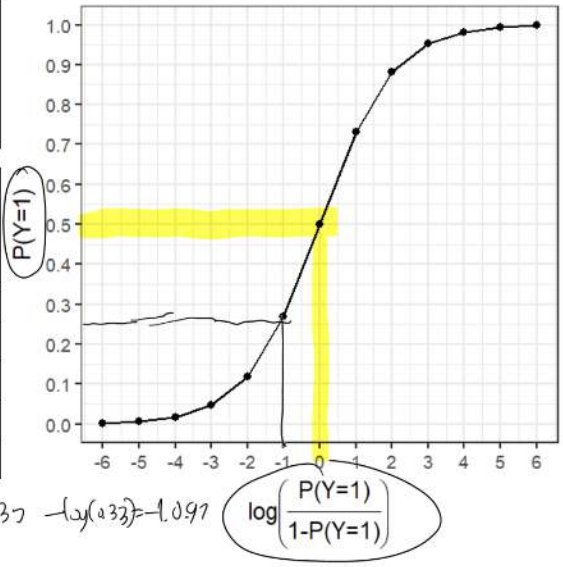
- R is case sensitive!!!
- Choose directory of dataset

Probabilities, odds and log-odds

Goal:

Transform a number between 0 and 1 into a number between $-\infty$ and ∞

Probability $P(Y=1)$	Odds $\frac{P(Y=1)}{1-P(Y=1)}$	Log-odds $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$
0.001	0.001	-6.907
0.25	0.333	-1.099
0.5	1	0
0.75	3	1.099
0.999	999.000	6.907



$P(Y=1)=0.25 \rightarrow P(Y=0)=0.75 \rightarrow \text{Odds}=\frac{0.25}{0.75}=0.33 \rightarrow \log(0.33)=-1.099$

- Using Logistic Regression on log-odds
  - Perform regression on log-odds

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p$$

  - Use data and maximum likelihood estimation to produce parameter estimates

$$\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_p$$

Different to OLS

  - Predict probabilities using

$$P(Y=1)=\frac{1}{1+\exp\left[-\left(\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\cdots+\hat{\beta}_pX_p\right)\right]}$$

UNSW SYDNEY

- R is case sensitive!!!
- Choose directory of dataset

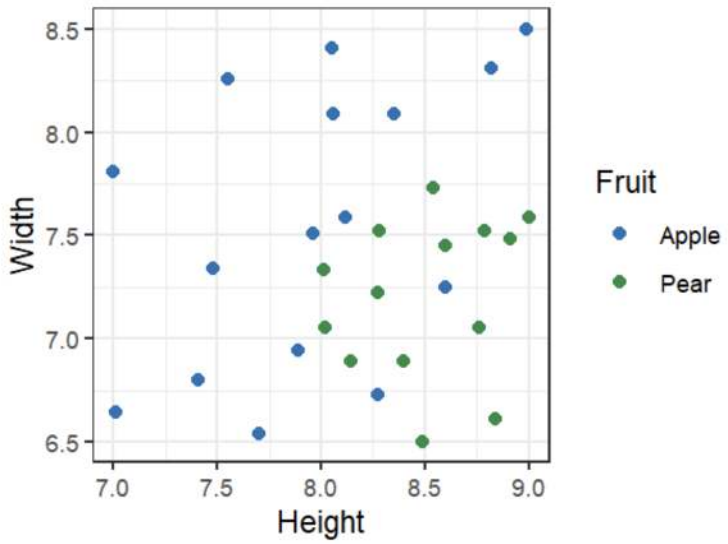
Pre-Workshop 5: Predictive Analysis II (Module 3)

Classification of fruits using classification trees

- ❖ *Construct a classification tree*
- ❖ *Plot and interpret the output of a classification tree*
- ❖ *Calculate the classification accuracy and confusion matrix of a classification tree*

- Distinguishing between Apples and Pears based on width and height
  - Design a method or algorithm to differentiate between apples and pears based only on their widths and heights
  - Data shows apples tend to be short and fat while pears are usually taller and more slight.

1. fruitdata <- read.csv("fruitdata.csv")
2. table(fruitdata$Fruit) → counts how many apples and pears in dataset

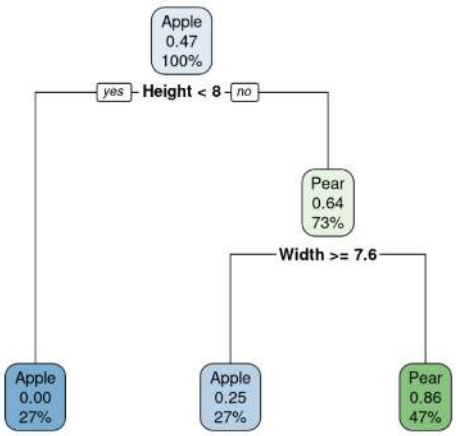


Classification Trees for Apples and Pears

- In `R` the library `rpart` is used to construct classification trees, while the library `rpart.plot` is used to plot such trees:
  - library(rpart)
  - library(rpart.plot)

- R is case sensitive!!!
- Choose directory of dataset

- Constructing the tree
    - `rpart()` function to fit a classification tree → predict type of fruit using the fruit's height and width
        - treeFruit <- rpart(Fruit ~ Height + Width, data = fruitdata)
    - Plot the classification tree using function `rpart.plot()`:
        - rpart.plot(treeFruit)



- From Table:
    - If the height is smaller than 8cm, we classify it as an apple.
    - If the height is bigger or equal to 8cm we then look at the width. If the width is bigger than 7.6cm then we classify it as an apple but if the width is bigger we classify it as a pear

1. Tells us that in the root node;
    - There are 30 fruits.
    - There are 14 "Pear" and 16 "Apple"
    - The proportion of apples is 16/30=0.53 and the proportion of pears is 14/30 = 0.46
2. Height< 7.985" (left branch of the tree)
    - 8 fruits which correspond to the 8/30=27% of the fruits
    - There are 0 "Pear" and 8 "Apple"
    - The proportion of apples is then 8/8=1 and the proportion of pears is 0/8 = 0.
    - Therefore fruits with height <7.985 should be classified as an Apple.

Predictions

- Assume fruits: height 6cm & width 6cm and another with a height 8.5cm & width 7cm

- R is case sensitive!!!
- Choose directory of dataset

    - Use code: *predict(treeFruit, newdata = data.frame(Height = c(6, 8.5), Width = c(6, 7)), type = "class")*
- Above, `newdata` specifies the data on which we want to make the predictions. For this we have used the function `data.frame()` to create a table with the speficifications of the fruits we want to predict:
    - *data.frame(Height = c(6, 8.5), Width = c(6, 7))*
- `type = "class"` → predict the actual class (type of fruit).
    - Alternatively, if we use `type = "prob"` → probabilities of pear or an apple:
    - *predict(treeFruit, newdata = data.frame(Height = c(6, 8.5), Width = c(6, 7)), type = "prob")*

- We see that in the case of the first fruit we are 100% sure that it is an apple. By contrast, in the case of the second fruit we have that there is a 14% chance that it is an apple and a 86% chance that it is a pear.

- **Classification accuracy and confusion matrix**
    - Evaluate classification tree in distinguishing between pears and apples
        - fruitPredTree <- predict(treeFruit, newdata = fruitdata, type = "class")
          fruitPredTree
    - The `table()` function can be used to produce a confusion matrix in order to determine how many fruits were correctly or incorrectly classified:
        - table(fruitPredTree, fruitdata$Fruit)
          (14+12)/30

```
fruitPredTree Apple Pear
        Apple    14    2
        Pear      2   12
```

    - Diagonal elements → correct predictions
    - Off-diagonals → incorrect predictions
    - 26 correct predictions → tree predicted the type of fruit correctly 86.7%
    - Model slightly better at predicting apples than pears:
        - Predicted apples correctly 14/16=87.5% of the times and pears 12/14=85.7% of the time.

- R is case sensitive!!!
- Choose directory of dataset

Classification of fruits using logistic regression

- A logistic regression for Apples and Pears

We are now going to use logistic regression as an alternative to predict the type of fruits.

For this we are going to code the type of fruit using a dummy variable:

$$Y = \begin{cases} 1 & \text{if Apple} \\ 0 & \text{if Pear} \end{cases}$$

We want to construct the following logistic regression for the probability of being an apple, as a function of the `Height` and `Width` of the fruit:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \times \texttt{Height} + \beta_2 \times \texttt{Width}$$

- Fitting the Logistic Regression
    1. To fit the logistic regression, first create the dummy variable Y → ifelse() function:
        a. **fruitdata$Y <- ifelse(fruitdata$Fruit == "Apple", 1, 0)**
    2. `glm()` function to fit logistic regression
        a. Fits *generalised linear models* which is a class of models that includes logistic regression
        b. Similar to lm(), except that we must pass in the argument family = binomial() to tell R to run a logistic regression rather than some other type of generalised linear model.
        c. **logisticFruit <- glm(Y ~ Height + Width, family = binomial(), data = fruitdata)**
    3. summary(logisticFruit)

- R is case sensitive!!!
- Choose directory of dataset

```
Call:
glm(formula = Y ~ Height + Width, family = binomial(), data = fruitdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6571  -0.5926   0.0692   0.5544   1.9224

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   13.671      9.896   1.381  0.16714
Height        -4.136      1.588  -2.604  0.00921 **
Width          2.803      1.240   2.260  0.02380 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 24.683  on 27  degrees of freedom
AIC: 30.683

Number of Fisher Scoring iterations: 6
```

This ouput is very similar to the one produced for linear regression.

For example, we see that $\hat{\beta}_0 = 13.671$, $\hat{\beta}_1 = -4.136$, and $\hat{\beta}_2 = 2.803$ so that the fitted model is:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 13.671 - 4.136 \times \texttt{Height} + 2.803 \times \texttt{Width}$$

- Interpret the signs of the coefficients
    - Positive coefficient → probability of Y=1 increases with the value of the corresponding variable

Negative coefficient → probability of Y=1 decreases with the value of the corresponding variable

Therefore, since $\beta_1 = -4.136$ is negative we conclude that the taller the fruit, the less probability there is of the fruit being an apple. By contrast, $\hat{\beta}_2 = 2.803$ means that the wider the fruit, the higher the chance that the fruit is an apple.

The p-values (`Pr(>|z|)`) associated with each of the two variables indicate that both Height and Width are significant in predicting the type of fruit.

Predictions

After a bit of a algebra we can show that in a our logistic regression the predicted probability of having an apple, $P(Y=1)$, is given by

$$P(Y=1) = \frac{1}{1+e^{-(13.671-4.136\times\texttt{Height}+2.803\times\texttt{Width})}}$$

- Thus, for a fruit with height=6 and width=6:

- R is case sensitive!!!
- Choose directory of dataset

    - 13.671-4.136*6+2.803*6 = 5.673
- The probability of it being an apple is:
    - 1/(1+exp(-5.673)) = 0.9965
- That is, we are almost sure that a fruit of height 6cm and width 6cm is an apple

- In R we can find the probabilities above with the predict() function. However, in the case of a logistic regression we need to set type = "response".
    - predict(logisticFruit, newdata = data.frame(Height = c(6, 8.5), Width = c(6, 7)), type = "response")

Confusion Matrix

- Create the confusion matrix for the logistic regression → first calculate the probabilities of each of 30 observation being an apple using the `predict()` function:
    - probsFruit <- predict(logisticFruit, newdata = fruitdata, type = "response")
      probsFruit
- In order to make a prediction as to whether a fruit is an apple or a pear, we must convert these predicted probabilities into class labels, `Apple` and `Pear`. The following two commands create a vector of class predictions based on whether the predicted probability of a fruit being an apple is greater than 0.5.
    - fruitPredLogistic <- rep("Pear", 30)
      fruitPredLogistic[probsFruit > 0.5] <- "Apple"
- The first command creates a vector of 30 `Pear` elements. The second line transforms to `Apple` all the elements for which the predicted probability of being an apple is bigger than 0.5. We can now use the `table()` function to compute the confusion matrix:
    - table(fruitPredLogistic, fruitdata$Fruit)
    - (14+10)/30

```
fruitPredLogistic Apple Pear
            Apple    14    4
            Pear      2   10
```

- We see that the logistic regression predicted the type of fruit correctly (14+10)/30=80% of the times, which is worse than the 86.7% of the classification tree.

Above and below 50 for Propensity to Stay, create dummy variable → Logistic Regression

- R is case sensitive!!!
- Choose directory of dataset

Lecture 7: Research Design and Experiments I

- Introduction
    - Organisations require answers to questions & input into decision-making
    - Research design addresses and maps out constituent parts of analysis
        - Subject matter theory
        - Appropriate data
        - Modelling approach that is appropriate for the data to deliver answers
    - Organisations need to answer what-if & evaluation type questions (prescriptive analysis)
        - Experiments with Random Control Trials (RCTs) obtains causal effects
        - Involves causal questions requiring estimates of causal effects
        - What if a change is made how will that affect future outcomes?
        - What impact did an intervention have? Was a policy change that was implemented effective?
    - Design → Prescriptive analysis
        - Research design: Can causal questions being asked be answered by available data & planned modelling approach?

- Case Study: Customer churning / retention problem
    - Descriptive: Is there a problem with customer churn
    - Predictive: What customers most at risk of churning
    - Prescriptive: Which customers are most likely to be retained if offered incentives to stay? Once implemented was the incentive cost effective?

- Other Examples - Obtain data through experiments:
    - a) Will is be profitable if on-line advertising is increased
        - Online A/B experiment (Split randomly into 2 groups)
    - b) Will a back-to-work intervention help people get a job?
        - require a field experiment
    - c) How much should homeowners living near to a chemical plant be compensated for a chemical spill?
        - Some questions not amenable to experiments → Too harmful
        - Obtain observational data
        - Design issues involved with natural experiments
    - Experimental mindset
        - Partial equilibrium approach - complex questions broken up into tractable components
        - Power of randomisation to control for confounders

- R is case sensitive!!!
- Choose directory of dataset

- Test, learn, adapt cycle (evidence based decision-making)

- Regression as an analytical tool

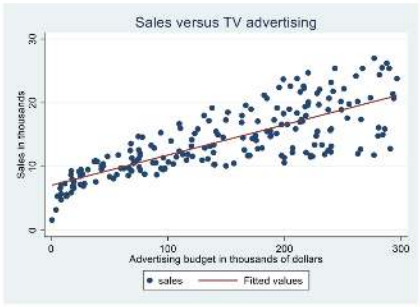
Linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Useful descriptive device to capture bivariate relationships

Consider sales (y_i) & TV advertising (x_i) (advertising.xlsx)

Q7.1 What key features of the data are revealed by the scatter plot?



Have specified a model

Sales in a market is a linear function of TV advertising

OLS provides best fit conditional on this model

Fine as a descriptive device providing stylized facts

Provides evidence of positive correlation (linear association) between sales & advertising

Prediction of sales for out-of-sample market?

$$\widehat{sales}_i = 7.033 + 0.048TV_i$$

A market where $TV = 100 \rightarrow \widehat{sales} = 11.833$

  - But may want models to do even more – causality & "what-if" counterfactuals
    - What happens to sales in a particular market if TV advertising were increased? Doesn't our regression model answer this question?

- <u>Threats to interpreting the regression results as causal</u>

1. **Confounding variables** leading to omitted variable bias:
    a. Is the estimated advertising effect biased? Maybe prices are varying across markets & these are correlated with advertising
2. **Reverse causality**
    a. What if markets with low sales increase advertising?

  - Prediction models aim to minimise prediction inaccuracy

- R is case sensitive!!!
- Choose directory of dataset

- Prediction questions may be solved using non-causal models

Sales at time t ($sales_t$) may be forecast well by past sales
$sales_t = \alpha_0 + \alpha_1 sales_{t-1} + u_t$

Causality & notion of ceteris paribus

- Definition of causal effect of *x* on *y*
  - How does variable *y* change if *x* is changed but all other relevant factors are held constant
  - In evaluating an intervention or policy change think of counterfactual outcomes & what-if questions
    - Sales with & without the increase in advertising
  - Requires (at a minimum) *x* & *u* to be unrelated (refer above table)

- Experiment 1:
  - Impact of back-to-work program on employment
    - "If a person is chosen from the population of those looking for work & given access to a back-to-work program, will that increase their chance of employment?"
    - Implicit assumption: all other factors that influence employment (experience, ability, local employment prospects,...) are held fixed
  - Experiment:
    - Choose a group of workers looking for work
    - Randomly assign them to access the program or not
    - Compare employment outcomes in next period
    - Experiment works because characteristics of people are unrelated to whether they receive program or not

- R is case sensitive!!!
- Choose directory of dataset

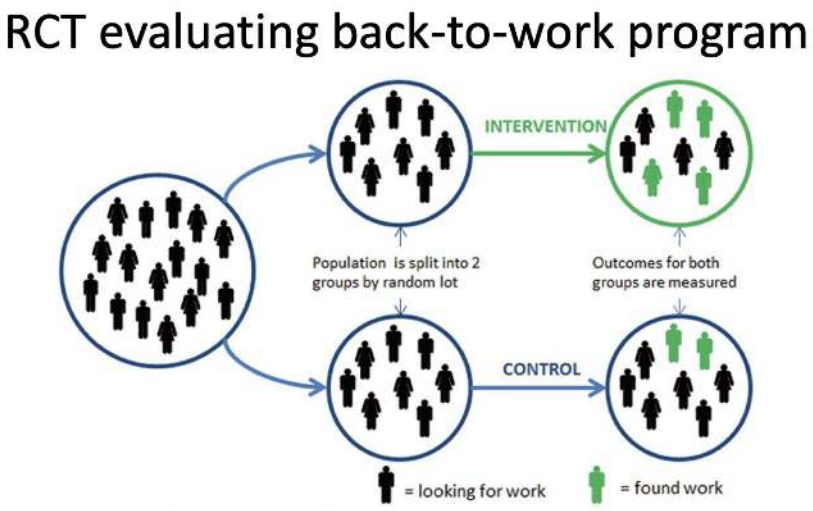


Figure 1. The basic design of a randomised controlled trial (RCT), illustrated with a test of a new 'back to work' programme.

- Experiment 2
    - A/B testing of a website landing page
        - "If a business rearranges its current website, how much will this change the conversion rate (new customers)?"
        - Implicit assumption: all other factors that influence who visits the website are held fixed
    - Experiment:
        - Design the new webpage
        - Randomly assign different users to old (A) & new (B) website
        - Compare conversion rates i.e. new customers
        - Experiment works because characteristics of users are unrelated to which website is seen
        - In online environments relatively easy to conduct

- Experiment 3: Policy Question → Importance of investing in education
    - Measuring returns to education
        - "If a person is chosen from the population & given another year of education, by how much will his or her wage increase?"
        - Implicit assumption: all other factors that influence wages such as experience, family background, intelligence etc. are held fixed
    - Experiment:
        - Choose a group of people

- R is case sensitive!!!
- Choose directory of dataset

        - Randomly assign different amounts of education to them!!!
        - Compare wage outcomes
        - Random assignment is infeasible in this case
        - Experiments are not always possible or ethical

- **Conducting RCTs**
    - Decide on form of intervention (new program/new website versus status quo)
    - Determine outcome of interest (employment/conversion rates)
    - Decide on randomisation unit (workers/customers/students)
    - Determine sample size & randomly assign units to treatment (new program/new website) & control (no program/old website) groups
    - Compare (average) outcomes to determine treatment effect
    - Any differences in outcomes can reasonably be attributable to the treatment as other aspects of data controlled by researcher
    - Decide on whether to adapt (implement program/use new website) or not on basis of findings

- Experimental evidence
    - Experimental evidence is input into decision-making
        - Even if the RCT yields a significant treatment effect this may not be enough to justify implementation of the intervention
        - Does the intervention represent value for money?
        - Interventions will be costly & so the size of any benefit needs to be weighed against the cost
    - <u>Null results are useful!</u>
        - A RCT that does not provide evidence of a treatment effect could avoid unnecessary costs
        - Informs us to use a different design and not invest!!!
    - Interventions may be intuitively appealing
        - But typically many intuitively appealing interventions
        - Which is better & whether they are value for money requires supporting evidence
    - Once implemented evidence should continue to be collected & interventions refined where appropriate
        - Interventions may work in one population but not another
        - Replication of core findings across several experiments represents more compelling evidence than a very significant effect in a single study
            - External validity

- R is case sensitive!!!
- Choose directory of dataset

- Observational data versus experimental data
    - Why experiment when observational data is available?
    - "…it is invariably true that the available data have **not been collected with research in mind**. … there is a mismatch between key concepts and available data or what is available suffers from sample selection problems. " Fiebig(2017)

    - In observational data, outcomes represent actual behaviour
        - Employment prospects depend on personal characteristics, employment conditions
        - Web access depends on personal characteristics, other advertising campaigns
        - No reason why factors impacting outcome are randomly assigned
            - Did people who were given the back-to-work program already have better employment prospects?
        - This is problem of confounding (lurking) variables

- Case Study: SAS most valuable career skill
    - What if workers choose to participate in the program?
        - If workers base their choice on likely benefits from the program then they are likely to provide an inflated (biased) estimate of the treatment effect
    - This is selection bias induced by an **endogenous treatment (Internal treatment)**
        - Q7.2 Does an SAS example with a huge sample & many controls avoid this selection problem? (NO)
        - Random assignment avoids this selection problem
    - Bottom line, useful to think with an experimental mindset

- R is case sensitive!!!
- Choose directory of dataset

Simple Linear Regression Model

Simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Useful special case if x_i represents a binary treatment

- Worker receives program ($x_i = 1$) or not ($x_i = 0$)
- Customer sees new website ($x_i = 1$) or old ($x_i = 0$)
- Patient received drug ($x_i = 1$) or placebo ($x_i = 0$)

More generally simply denotes group membership

- $x_i = 1$ if female, $x_i = 0$ otherwise
- Here gender is not a treatment!

Assume x_i & u_i are unrelated, $E(u_i|x_i) = 0$, then

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i + E(u_i|x_i) = \beta_0 + \beta_1 x_i$$

$$E(y_i|x_i = 0) = \beta_0$$
$$E(y_i|x_i = 1) = \beta_0 + \beta_1$$

$$\Rightarrow \beta_1 = E(y_i|x_i = 1) - E(y_i|x_i = 0)$$

Straightforward to show that OLS estimates are

$$\hat{\beta}_0 = \bar{y}(0) \;\&\; \hat{\beta}_1 = \bar{y}(1) - \bar{y}(0)$$

where $\bar{y}(x_i)$ is mean of y_i conditional on x_i

$\hat{\beta}_1$ is difference in means between the two groups

- R is case sensitive!!!
- Choose directory of dataset

Potential Outcomes

- Estimated β_1 represents difference in predicted y_i for treated versus those not treated
    - Simple regression of outcomes on binary treatment recovers difference in means (difference in proportions in Experiments 1 & 2)
    - Difference is descriptive but can it be interpreted as causal?

- Potential outcomes framework considers outcomes in two states of the world ($y_i(1)$ if treated or $y_i(2)$ if not)
    - Treatment effect is difference in two potential outcomes
    - $te_i = y_i(1) - y_i(0)$

- Q7.3 Without more structure the treatment effect can't be estimated. Why?

Fundamental problem of causal inference – need a credible way to infer unobserved counterfactual outcomes

If focus on average treatment effect (ATE)

$$\tau = E(te_i) = E[y_i(1) - y_i(0)]$$

… and assume random treatment assignment

… then in $y_i = \beta_0 + \beta_1 x_i + u_i,\ \beta_1 = \text{ATE}$

Implying estimated ATE is the difference-in-means estimator

- R is case sensitive!!!
- Choose directory of dataset

Case Study: Causal effect of Women as policy makers

- Hypothesis is that women leaders will support policies that women voters care about more
- 'reserved' indicates reserved for a female village leader
- 'female' indicates a female village leader
- irrigation & water are number of new or repaired facilities of this type – men tended to be more concerned about former & women the later

- Q7.4 Why would an observational study be problematic in determining whether women politicians promote different policies when in government?

First it seems that the policy was successfully applied

All 108 treated ($reserved = 1$) villages had a female head

Only 16 of the 224 control villages had a female head

Estimate regression models

$$water_i = \beta_0 + \beta_1 reserved_i + u_i$$

Estimated ATE is 9.25 increase in projects attributable to the policy & precisely estimated with 95% CI $[1.49, 17.02]$

$$irrigation_i = \alpha_0 + \alpha_1 reserved_i + v_i$$

Estimated ATE is -0.37 decrease in projects attributable to the policy but imprecise with 95% CI $[-2.58, 1.84]$

- ATEs indicate support for the hypothesis
  - Treated villages with female heads were much more likely to support water projects
  - Treated villages with female heads were less likely to support irrigation projects although this effect had a 95% CI that overlapped zero

- R is case sensitive!!!
- Choose directory of dataset

Workshop 7: Research Design & Experiments I

---

1. Critically evaluate the following experiment conducted by a bank that wanted to investigate the preferences of customers towards two features of their credit cards:
    a. the annual fee charged, and
    b. the annual percentage rate charged.

    The bank selected a large sample of people at random from a mailing list and randomly sent half the sample offers with a low rate and no card fee. While the other half of the sample received offers with a higher rate and a $50 annual fee.

    Identify the key problem with this experiment and explain how you would avoid it by redesigning the experiment?

    - Should be four different card options
        - 1. Low rate and low card fee
        - 2. Low rate and high card fee
        - 3. High rate and low card fee
        - 4. High rate and high annual fee
    - Customers will be inherently biassed towards choosing low rate and no card fee, this limits the effectiveness of the experiment
    - The experiment must be redesigned to control confoundment issues
        - Randomly assign customers to one of the four possible cards
        - Compare acceptance rates for the low-low versus low-high (1.V2.) card with the acceptance rates for the low-low versus high-low (3.V4.)

2. An energy company is considering a change to its pricing policy involving the introduction of time-of-day pricing whereby customers are charged more for electricity use at peak times of the day (e.g. 6-8pm) in an attempt to smooth out the high demand periods. Before they implement the policy, they decide to conduct an experiment to determine the impact of the change on electricity demand. They chose several regions of the state to be treated **(charged higher prices at peak times and off-peak prices remaining unchanged)** while in other areas all prices remained constant. Because management was concerned about possible customer complaints, they decided not to reveal to customers that the changes were being made.

- R is case sensitive!!!
- Choose directory of dataset

a) What are the problems with this as an experimental design?

- 1. Choice of regions and whether there was a good matching of treatment and control regions.
    - **Climate (main determinant of electricity demand) can vary quite a lot**. You do not want all treatment regions concentrated on one area.
- 2. Experiments where the participants didn't know they were in an experiment:
    - e.g. in Q1 customers didn't know that their offer is different from others but they do know the form of their offer. H
    - Here the treatment is a price change, but customers don't know the price has changed (well not until they get their next bill by which time it is too late to change behaviour.)

b) How might you overcome these problems using a different design?

a. Clearly make sure you have treatment and control regions that on average are similar
b. One way is to enlist customers into the treatment group and say that based on your bills for the experimental period in the past, if you do not change behaviour over the course of the experiment then your overall bill would stay the same (you would need lower off-peak rates to compensate for higher peak rates). Thus, they can only benefit from the change by shifting when they use their electrical appliances

3. Lewis & Reiley (LR) (2014) contrast their experimental approach to estimating the impact of online ads to that proposed by Abraham (2008) described as:
    - "Measuring the online sales impact of an online ad … in which a company pays to have its link appear at the top of the page of search results – is straightforward: We determine who has viewed the ad, then compare online purchases made by those who have not seen it."

a) What is the problem with the alternative research design of Abraham (2008) and how do LR avoid the problem?

- There is a clear self-selection problem. Those who click on and view an ad are more likely to purchase the product.
- In the experimental approach customers are randomly assigned to whether they see the ad or not. Treatment is randomly assigned and not determined by the customer. Thus, any difference in sales between the treatment and control group can reasonably be attributed to the ad as all other contributing factors are not related to whether anyone is treated or not.

- R is case sensitive!!!
- Choose directory of dataset

b) LR presents the following Table of summary statistics comparing the control and treatment groups of customers. What does this show and why is this important in terms of reporting results from a field experiment?

- In terms of respondent characteristics (female) and internet activity (Yahoo! page views) there is almost no difference between the control and treatment groups. Such comparisons confirm the success of the random assignment.
- The other measures describe the exposure of the treatment group to the ads and that the control group had zero exposure
- Notice that the treatment did not ensure everyone in the treatment group saw the ads just that if they did visit Yahoo! the ad would be there.

Table 2 Basic summary statistics for the campaign

	Control	Treatment
% Female	59.50 %	59.70 %
% Retailer ad views > 0	0.00 %	63.70 %
% Yahoo page views > 0	76.40 %	76.40 %
Mean Y! page views per person	358	363
Mean ad views per person	0	25
Mean ad clicks per person	0	0.056
% Ad Impressions Clicked (CTR)	-	0.28 %
% Viewers clicking at least once	-	7.20 %

c) LR notes that the difference in mean number of Yahoo! page views, 363 versus 358, was statistically significant. In other words, the 95% confidence interval for the difference does not include zero. Does this have any impact on your answer to (b)?
   - It remains the case that the actual size of the difference is small. We now know that because of the large sample size this difference is precisely estimated and turns out to be statistically significant.
   - In LR they do note that there were a few customers with excessively large numbers of views and that these happened to be in the treatment group. After accounting for these outliers, the significance of the difference disappears.

d) The initial baseline estimate provided by LR is the difference in mean sales of the treatment and control groups during the experiment. Because only 63.7% of the treatment group actually viewed the experimentally allocated ads, do you agree that this is a conservative estimate of the effect of advertising on sales.

   - Yes, it is conservative in that the difference could have been larger (it could not have been smaller) depending on how those in the treatment group would have responded if they had viewed the ads.

- R is case sensitive!!!
- Choose directory of dataset

e) In one of their additional analyses, LR redefined the treatment group to be only those in the treatment group who viewed the ads, while the rest of the treatment group who did not view the ads were added to the control group. What are the advantages and disadvantages of this modelling decision?

- Advantage: all of those in the treatment group who were not exposed are treated as controls so that we isolate the effect of treatment on those who received the treatment (in causal inference terminology this is the treatment effect on the treated). Potentially this is a more interesting treatment effect than the initial estimate (intention to treat effect).
- Disadvantage is that people choose not to be exposed and you worry that they are systematically different from others in the control group with whom they are pooled. LR check this and it seems not to be a worry.

4. Chattopadhyay and Duflo (2004)* is a highly cited field experiment that investigated the causal effects on policy outcomes of having female politicians in government. In India there was a period when one-third of village council heads were randomly reserved for female politicians. Part of the data from the study are provided as women.csv. The policy was implemented at the GP level of government, so some GPs were treated and others not. In the data the GP was said to be reserved or not and this indicator is in the data as the variable reserved. In the data there are 161 GPs and for each GP the study included two randomly selected villages giving a total of 322 village level observations. The variable female indicated whether the village in the GP in fact had a female leader or not

The outcomes of interest represent the number of new or repaired facilities of two types, irrigation, and water. The hypothesis being investigated is that women will support policies that women voters care about more. Previous research had indicated that women tended to complain about water quality while men tended to be more concerned with irrigation problems.

a) Why would an observational study be problematic in determining whether women politicians promote different policies when in government? (Q7.4 in lectures)
    - In an observational study the presence of female heads would not be randomly assigned. The women would need to choose to run and then be elected. The successful woman candidate could be very different from one willing to be assigned to a GP
    - The villages willing to elect a woman may be very different from those not willing.
    - In each of these cases you may be able to find appropriate variables that could explain these differences, but it is not obvious. Some of the key variables may be things that are difficult to measure
    - An even more pragmatic answer is that there may not be enough villages with female heads to draw any meaningful comparisons between the two groups. In the next question we see that in the control group villages that were not subject to the policy intervention only 7.5% had female heads.

- R is case sensitive!!!
- Choose directory of dataset

b) Confirm the analysis provided in lectures that checks whether the policy had been successfully applied by determining whether reserved GPs have female leaders?

- R output below in the form of a cross-tab confirms that all reserved GPs did have a female head and that of those villages in the control group 7.5% had female heads.

```
> res.fem.tab <- table(res = women$reserved, fem = women$female)
> addmargins(res.fem.tab)

     fem
res     0   1 Sum
  0   198  16 214
  1     0 108 108
  Sum 198 124 322

> res.fem.tab[1, 2] / sum(res.fem.tab[1, ])
[1] 0.07476636
```

c) Confirm the estimates of the ATE of the reservation policy for both irrigation, and water that were provided in lectures. Does your analysis support the hypothesis that women tend to support the preferences of women voters?

- R output below indicates support for the hypothesis – the estimated increase in projects attributable to the reservation policy was 9.25 for water but for irrigation there was a slight decrease of -0.37.

- R is case sensitive!!!
- Choose directory of dataset

- R output below in the form of a cross-tab confirms that all reserved GPs did have a female head and that of those villages in the control group 7.5% had female heads.

```
> res.fem.tab <- table(res = women$reserved, fem = women$female)
> addmargins(res.fem.tab)

     fem
res     0   1 Sum
  0   198  16 214
  1     0 108 108
  Sum 198 124 322

> res.fem.tab[1, 2] / sum(res.fem.tab[1, ])
[1] 0.07476636
```

```
Call:
lm(formula = water ~ reserved, data = women)

Residuals:
    Min      1Q  Median      3Q     Max
-23.991 -14.738  -7.865   2.262 316.009

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.738      2.286   6.446 4.22e-10 ***
reserved       9.252      3.948   2.344   0.0197 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.45 on 320 degrees of freedom
Multiple R-squared:  0.01688,  Adjusted R-squared:  0.0138
F-statistic: 5.493 on 1 and 320 DF,  p-value: 0.0197

> confint(water.women)
                2.5 %   97.5 %
(Intercept) 10.240240 19.23640
reserved     1.485608 17.01924
>
> irrigation.women <- lm(irrigation ~ reserved, data=women)
> summary(irrigation.women)
```

```
Call:
lm(formula = irrigation ~ reserved, data = women)

Residuals:
   Min     1Q Median     3Q    Max

-3.388 -3.388 -3.019 -1.019 86.612

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3879     0.6498   5.214 3.33e-07 ***
reserved     -0.3693     1.1220  -0.329    0.742
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.506 on 320 degrees of freedom
Multiple R-squared:  0.0003385,       Adjusted R-squared:  -0.002785
F-statistic: 0.1084 on 1 and 320 DF,  p-value: 0.7422

> confint(irrigation.women)
                2.5 %   97.5 %
(Intercept)  2.109436 4.666265
reserved    -2.576766 1.838103
```

- R is case sensitive!!!
- Choose directory of dataset

Lecture 8: Research Design and Experiments II

Gold Standard Analysis

- RCTs often though fo as a gold standard in estimating causal effects
    - But all data can misbehave
- Randomisation is powerful but the relevant population may be restrictive
    - Recall yahoo user case study
    - Are conclusions appropriate for different populations?
    - An RCT may have good internal validity (technical aspect) but lack external validity thus restricting its generalisability
        - Internal Validity: the extent to which the observed results represent the truth in the population we are studying
        - External Validity: The extent to which you can generalise the findings of a study to other situations, people, settings and measures
            - Can it be applied to a broader context?

- It is possible that finite samples treatment and control groups may differ by chance

  Good practice suggests adding pre-treatment controls x_{ik} in addition to treatment variable t_i

  $$y_i = \beta_0 + \beta_1 t_i + \delta_1 x_{i1} + \cdots + \delta_p x_{ip} + u_i$$

    - Even when randomisation does not fail such an approach should improve precision of B1 estimate
        - Important when estimating small effects because larger sample sizes might not guarantee precise estimates (Recall case study 2)

Challenge of Causal inference

- Sometimes direct experimentation is not possible / not ethical (E.g. oil spill)
- Role for non-experimental or observational data
    - These data comes with threats
    - Recall confoundment and selection into treatment

Legal AB Testing

- Whether proposed AB testing can **harm or disadvantage subjects** of a test
    - E.g. Trial of new drug

- R is case sensitive!!!
- Choose directory of dataset

        - Proportion of subjects receive placebo
        - If drug is life saving, then subjects who received placebo are likely to die

- AB testing **reputational risk**, needs to be one which is well understood by the corporation conducting the testing
    - E.g. Facebook testing users on happy and sad content → legal but negatively impacted emotionally disadvantaged → reputational harm for facebook
    - Platform manipulation (Google, facebook) → Likely to cause reputational harm

- Poor design of AB test:
    - Consumer law issues in Australia
        - E.g. Subjects in group B do not receive a service fit for the purpose → legal action
    - Ensure does not breach anti-discrimination laws
        - E.g. Experiments grouping by gender or sexual orientation if the service received was significantly different (disadvantaged) to the other group

- AB testing must take into account disadvantages of any subject and consider whether the alternatives are both acceptable to all subjects of the experiment (but could still lead to reputational risk)

- Consider regression model using observational data

$$y_i = \beta_0 + \beta_1 t_i + \delta_1 x_{i1} + \cdots + \delta_p x_{ip} + u_i$$

- Can we interpret the OLS estimate of β_1 as causal?
- In general β_1= change in y_i if compare $t_i = 0$ versus $t_i = 1$
    - For causal interpretation need **all other factors to be held constant** - otherwise identical in terms of x_{ik}
    - But also need disturbance to remain unchanged as t_i changes
    - Disturbance includes factors impacting y_i assigned to unobservables & if these correlated with t_i then confoundment

- R is case sensitive!!!
- Choose directory of dataset

- If z_i is an omitted variable (implicitly included in u_i) & it is correlated with t_i
  - OLS estimate of β_1 reflects both direct impact of t_i on y_i
  - But also impact of z_i through correlation of t_i & z_i
  - This confoundment problem means the estimator is **biased**
- With classical experimental data there may be omitted variables (included in u_i)
  - BUT because of random assignment these will be uncorrelated with binary treatment t_i
  - OLS estimate of β_1 reflects causal impact of treatment on y_i

- In summary, for observational data:
  - For OLS estimate of $\beta 1$ to be causal requires no omitted confounders /variables that are correlated with treatment
  - Need assignment of treatment to be as good as random given included controls – selection on observables
  - Recall observational data not collected for research
  - May be unobservable factors that make selection on observables difficult to assume/justify

- Data availability provides opportunities to exploit natural experiments (quasi experiments)
  - Here randomisation is accidental
  - Rely on some exogenous change (policy intervention) effects some individuals, families, firms, ... (treated) but not others (control)

- An organisation changes staff training but transition is staged: some branches make the change before others
  - Difference in timing yields treatment & control branches
  - Now reasonable to analyse as an experiment to estimate impact of training intervention

- R is case sensitive!!!
- Choose directory of dataset

Chemical Spill Project

- Hired as a data analyst by a consulting firm & you have your first project
    - A firm has admitted responsibility for a chemical spill
    - There is a court case determining damages for residents living near to the chemical plant
- Conceptually this is a causal problem that you need to solve
    - How much damage did residents incur because of the spill?
    - Consider the impact on housing prices, what is the difference in prices now & what it would have been with no spill?

- How can you estimate the spill's impact on house prices?
    - Clearly a designed experiment is not possible
    - Theoretical context, research design & associated data needs to be determined → these data will be observational

- Possible framework for your analysis is a **Lancastrian view of consumer demand**
    - Products viewed as bundles of characteristics or attributes
    - Consumers value attributes not the product per se
    - Each of these attributes has an implicit (shadow) price

- Hedonic Price Model
    - Hedonic regression allows estimation of attribute valuations despite there not being an explicit market
    - Houses valued for their characteristics
        - Number of bedrooms, bathrooms & car spaces, …
        - Location relative to city centre, beach, amenities

Hedonic regression using sample of housing data

$$\widehat{\log(price)} = \underset{(.024)}{6.71} + \underset{(.008)}{.143bed} + \underset{(.009)}{.190bath} + \underset{(.006)}{.040car} - \underset{(.002)}{.053dist}$$

$$n = 4663, R^2 = .378, se's\ in\ (\ )$$

$price$ = selling price; bed = no. of bedrooms; $bath$ = no. of bathrooms; car = $no.of\ car\ spaces$; $dist$ = distance from CBD

- R is case sensitive!!!
- Choose directory of dataset

Research Design : Chemical Spill

- Suggests approach using hedonic regression with similar data for oil spill area
  - But what type of data would be best?
  - Need to consider alternative research designs

- Research Design A

Get expert advice on affected area ($near_i = 1$ if house i is in area affected & 0 otherwise)

Collect data after spill for houses both affected & not

Compare average prices & attribute differences to the spill

Parameter of interest would be β_1 in following model:

$$\log(price_i) = \beta_0 + \beta_1 near_i + u_i$$

- Research Design B

Collect data before & after spill for houses in affected area

Compare prices before & after spill attributing differences to the spill

Define $after_i = 1$ if house i sold after spill & 0 otherwise

Parameter of interest would be θ_1 in following model:

$$\log(price_i) = \theta_0 + \theta_1 after_i + u_i$$

- Q8.1 Which design is preferred?

- Research Design C

- Collect data before & after spill for houses affected & not
- Compare prices before spill for houses in affected area & those not $\rightarrow \Delta_b$
- Repeat for after period $\rightarrow \Delta_a$
- Now calculate difference in these differences $(\Delta_a - \Delta_b)$ & attribute this to the spill
- Parameter of interest would be δ_3 associated with **interaction** term in following model:

$$\log(price_i) = \delta_0 + \delta_1 near_i + \delta_2 after_i + \delta_3 near_i \times after_i + u_i$$

- R is case sensitive!!!
- Choose directory of dataset

Just checking:

$\mathrm{E}[\log(price_i)] = \delta_0 + \delta_1 \quad \text{if } near_i = 1, after_i = 0$

$\mathrm{E}[\log(price_i)] = \delta_0 \quad \text{if } near_i = after_i = 0$

$\rightarrow \Delta_b = \delta_1$

$\mathrm{E}[\log(price_i)] = \delta_0 + \delta_1 + \delta_2 + \delta_3 \quad \text{if } near_i = after_i = 1$

$\mathrm{E}[\log(price_i)] = \delta_0 + \delta_2 \quad \text{if } near_i = 0, after_i = 1$

$\rightarrow \Delta_a = \delta_1 + \delta_3$

$\rightarrow \Delta_a - \Delta_b = \delta_3$

▸ This is the **difference in difference** (DiD) estimator

1. Note #1: Control variables not included in discussion purely for convenience,
    a. normally include control variables
    b. Treatments can often be viewed as if randomly assigned once we control for observable characteristics

- Suppose chemical plant is located in a relatively unattractive area
    - In A estimate of $\beta 1$ would in part reflect this difference that has nothing to do with the spill – a confoundment problem
    - Adding controls potentially avoids this source of bias

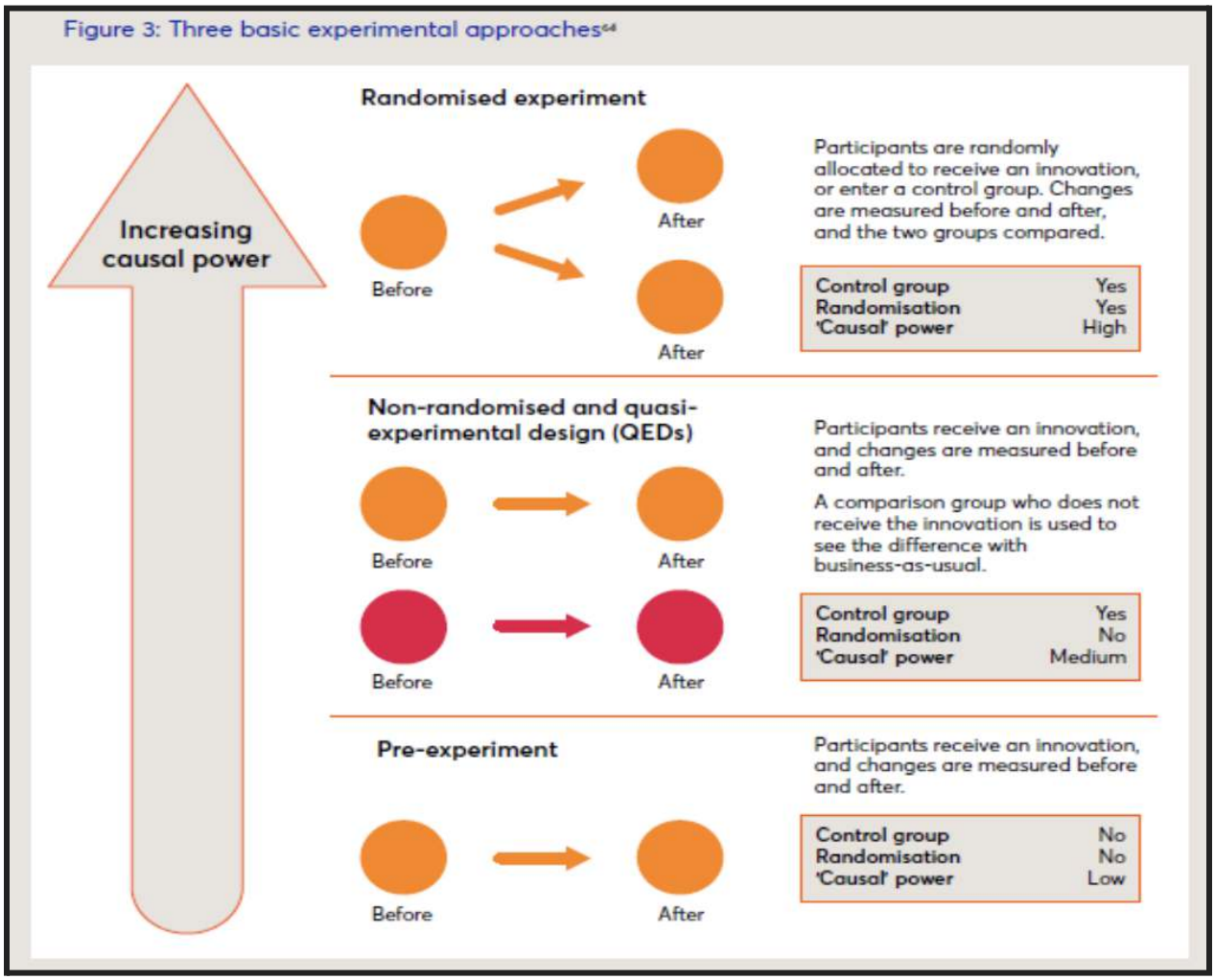
2. Note #2: Research design C is superior to A & B
    a. But does depend on a crucial assumption
    b. Without the spill, change is same for house prices in affected and unaffected areas
    c. Percentage change is the same in log assumption (Technical issue beyond scope of the course)

- Any pre-existing divergent trends in house prices in the two areas would invalidate design C
    - DiD estimator $\delta 3$ would in part capture this trend difference which should not be attributable to the spill

3. Note #3: Research design C would be even better if panel data were available
    a. Panel data is where individual observations (houses) are tracked over time
    b. Now differences in before & after sales could be calculated for the same house
    c. This controls for both observable & unobservable differences in specific houses
    d. Because of infrequency of repeat sales this type of data is unlikely to be available for such a study (but see next case study)

- R is case sensitive!!!
- Choose directory of dataset

Summary of experimental approaches



Case Study 3: Churn/ Retention Problem

- Consider the prominent business problem of retaining customers at risk switching providers.
    - E.g. credit cards, mobile phones, electricity supply, …
- Different types of analysis to answer different questions
    - Is there a problem with customer churn? - Descriptive analysis
    - Which customers are at most risk of churning? - Predictive analysis
    - Which customers are most likely to be retained if offered incentives to stay? - Prescriptive analysis requiring causal estimates

- R is case sensitive!!!
- Choose directory of dataset

Consider predicting which customers are most at risk of churning

- Define $y_i = 1$ if customer i churns within a specified time

$$Risk_i = Prob(y_i = 1 | x_{i1}, \dots, x_{ip}) = P(y_i = 1 | \boldsymbol{x}_i)$$

- Model-based approach using logistic regression

$$P(y_i = 1 | \boldsymbol{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}$$

- Other predictive approaches possible e.g. regression trees

- Predictive analytics provides a risk index (predictions from model for churning)
  - Useful but does it solve the retention problem?
- Management must decide how best to allocate resources to retain customers
  - Is a good predicted index of churning risk enough?
  - Treated as a prediction problem means allocating resources to persuade those at highest risk of churning to stay
  - Relies on strong assumption that customers most receptive to retention are those with highest risk

Ascarza (2018) argues you need to answer different question

- Would a customer at risk be receptive to incentives to stay?

Define

$$lift_i = P(y_i = 1 | \boldsymbol{x_i}, t_i = 0) - P(y_i = 1 | \boldsymbol{x_i}, t_i = 1)$$

- t_i = treatment dummy indicating if i^{th} customer offered an incentive to stay or not
- Common management strategy is to allocate incentive on the basis of those most at risk according to

$$risk_i = P(y_i = 1 | \boldsymbol{x_i}, t_i = 0)$$

- R is case sensitive!!!
- Choose directory of dataset

Incentives work if $lift_i > 0$ implying increased (decreased) probability of retention (churning)

- Suppose incentives imply proportionate decrease $(0 < \delta < 1)$ in risk of churning irrespective of customer characteristics
- Then targeting those with highest risk is an appropriate retention strategy

$$lift_i = P(y_i = 1|\boldsymbol{x_i}, t_i = 0) - \delta P(y_i = 1|\boldsymbol{x_i}, t_i = 0)$$
$$= (1 - \delta) risk_i$$

No a priori reason to expect homogeneous responses

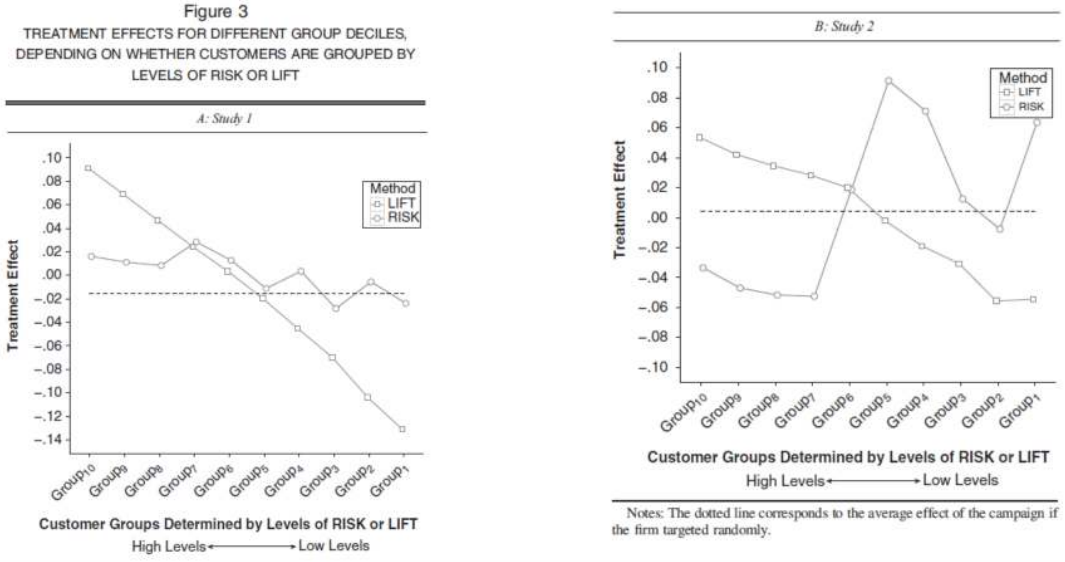
- Offering incentives to stay may even lead to increased churn in some customers

- R is case sensitive!!!
- Choose directory of dataset

- Ascarza (2018) runs experiments in two studies
  - Study 1: Mobile phone providers
    - Customers with pre-paid plans were offered extra credit when recharging
  - Study 2: Association membership renewal
    - Treatment was to add a gift to renewal communication
- In each case define two sets of 10 groups defined by deciles according to $risk$ & $lift$
  - For each decile-way combination calculate treatment effect, e.g.

$$TE_{R_{10}} = P(y_i = 1|R_{10}, t_i = 0) - P(y_i = 1|R_{10}, t_i = 1)$$
$$TE_{L_{10}} = P(y_i = 1|L_{10}, t_i = 0) - P(y_i = 1|L_{10}, t_i = 1)$$

Figure 3
TREATMENT EFFECTS FOR DIFFERENT GROUP DECILES, DEPENDING ON WHETHER CUSTOMERS ARE GROUPED BY LEVELS OF RISK OR LIFT



Notes: The dotted line corresponds to the average effect of the campaign if the firm targeted randomly.

- Figure 3: results not surprising that *riski* & *lifti* identify very different customers to target
  - Study 1: Of top 50% of *riski* customers only 52% are in top 50% of *lifti*
  - Study 2: Of top 50% of *riski* customers only 40% are in top 50% of *lifti*

- For two very different settings using *lifti* clearly dominates *riski* for use in retention management
  - Not “proof” but evidence that using *lifti* is better strategy

- R is case sensitive!!!
- Choose directory of dataset

- Estimating $risk_i$ is a prediction problem amenable to a predictive analytics
    - Estimating $lift_i$ is a causal not predictive problem relying on whatif counterfactuals
    - Firms need to run randomised experiments in combination with good predictive methods
    - Requires business to be able to interact individually with customers but this is relatively common

- Progress Report #2
    - Other causal problems that could be addressed in this framework (& not by data-driven decision-making)
        - What is the best type of incentive?
        - Does timing of the incentive matter?
        - How do you model the potential response of competitors?
        - How do you incorporate the expected future value of the customers into retention management?
        - See Athey (2017) for other examples & the case for combining predictive & causal approaches
    - Stressed role of regression as the primary tool of analysis
        - Have tried to be more explicit about what constitutes evidence of causation
        - Important role for experiments
    - Still a role for non-experimental or observational data
        - But using such data to estimate treatment effects comes with threats that need to be recognized
        - Key threats in using observational data are confounding & selection problems

- R is case sensitive!!!
- Choose directory of dataset

Workshop 8: Research Design & Experiments II

1. A certain school typically has two, year 7 classes that occupy one of two rooms. These rooms have different capacities (20 or 25 students) presenting the opportunity to exploit a natural experiment to investigate the impact of class size on student performance. Suppose you have data from this school's year 7 cohort collected over many years and comprising student grades and whether they were in the small or large class.

   Under what conditions would you expect the difference in year 7 grades between those in the large class relative to those in the small class to best reflect the causal effect of class size on student performance?

- Random allocation of students to each group
- Random allocation of teacher
- The performance should be measured on a standardised test that is comparable over time.
- The population of students needs to be relatively stable over time again to ensure comparability in results.

2. Jenny Craig is a weight-loss intervention. Their commercials show a photo of some celebrity before and after joining Jenny Craig.

   a) What are the controls and treatment in this experiment?
   b) Are you worried about any selection problems? Explain.

- R is case sensitive!!!
- Choose directory of dataset

a) This is a (within the person) before and after design. It is the weight of the same person being compared before (control) and after (treatment).

b)
   i) Biases selection based on only successful candidates to use in their ads
   ii) This is a sample selection problem. We need to distinguish this from the **endogenous** selection issue, where people choose whether to be treated or not.
      1) People who willingly join → more motivation to lose weight.

Note: Some students may worry about the sample size, n=1. But there is a history of self experimentation, especially in medicine. Researchers trying out their ideas on themselves. Often with success.

So that's not really the main problem, although in general, we do prefer more observations to increase the precision of our estimates. Also in an RCT version of testing whether Jenny Craig works or not, you would want a large sample to ensure that randomization has worked.

3. Your statistically naïve friend, Denzil, is very interested in the impact on housing prices of being located under the flight path. Given data on sales over a month, the regression of housing price on flightpath (a dummy variable indicating whether the house was under the Kingsford-Smith airport flightpath) provides a difference in means estimate.

a) Explain why this estimate would be a poor indication of the impact on sales of being under the flight path

b) After the Western Sydney International Airport at Badgerys Creek Sydney opens, suppose the decision is made to close the east-west runway at the Kingsford-Smith airport leaving just the two north-south runways in operation. This is purely hypothetical and highly unlikely but suppose it does. Denzil now revisits his original research question, but this time uses two samples of houses that were located under the east-west flight path. But one sample represented sales in a year before the closure of the runway while the second sample represented sales in a year after the runway was closed. Now he runs a regression of the following form:

$$Price_i = \beta_0 + \beta_1 after_i + u_i$$

- where $price$ = housing sale price; $after$ = 1 if the house was sold after the runway closure and zero otherwise

How do you interpret the regression parameters for this model?

c) Do you think this is a good research design to address this problem? Explain.

- R is case sensitive!!!
- Choose directory of dataset

a) This is a “difference-in-means” regression as the flight path is a binary variable that divides sales according to whether they were under the flight path or not. This is fine as a descriptive tool but not if you want to infer causality.
    i) There are likely to be many other confounding factors here that would be correlated with flightpath and would help explain the price difference.

b) This is a “difference-in-means” regression as after is a binary variable that divides sales according to whether they were sold after the closure of the runway or not.

   The estimated intercept will be the mean of the sales price for those houses sold in suburbs under the runway and before it was closed. The estimated after parameter will be the difference in means associated with being sold after the closure.

   Because the negative features of being under the flightpath (notably aircraft noise) have been removed you would expect house prices of those previously under the flightpath would increase and hence the estimate of $\beta 1$ to be positive

c) As always, the interpretation of $\beta 1$ relies on the likely unobservables captured in the disturbance and whether they are likely correlated with after.

   Here you would worry about what is happening in the housing market in general. You may find a positive estimate of $\beta 1$ but is that attributable to the reduced aircraft noise or simply that the housing market is active, and all houses are increasing in value?

   This is an example of a natural experiment. In such cases, you worry that the treatment is truly exogenous. For example, if the closure has been mooted for a while, then developers may buy up houses in the affected area and renovate them in anticipation of the expected post-closure appreciation in value. This again would distort the before and after estimate as a reflection of just being under the flight path or not

Note : Explore DiD as a solution. Ask students if they also had two samples of sales (before and after) for houses in control suburbs (i.e. close to the treated suburbs and hence similar but unaffected by aircraft noise either before or after). What could they do?

If you take the difference in mean sales for the control and treated groups before the closure and compare this to the same difference after, then this difference-in-difference estimate is a better estimate because any autonomous trends in housing prices are accounted for. It does rely on the assumption that these trends are the same in both the control and treated suburbs. The Varian (2016) paper is a reference for this DiD approach.

- R is case sensitive!!!
- Choose directory of dataset

4. The Australian Federal Government introduced a policy designed to provide an incentive for women to have more children: the so-called baby bonus. In the budget released in May 2004, the Government announced that families whose babies were born on or after July 1, 2004, would receive a sum of $3,000. No bonus was paid for births before this date. The amount of the bonus continued to rise and on July 1, 2008, rose to $5,000.

Here we only consider an analysis of Australian Bureau of Statistics birth data just before and immediately after the introduction of the new policy on July 1, 2004. In particular, the outcome variable of interest is the number of daily births in Australia for the period January 1 Page 4 to July 31 of 2004, which is 6 months before the introduction and one month after. The time series graph below depicts the birth data with t = 183 corresponding to July 1, 2004.

Suppose the primary objective of the analysis is to determine the impact of the introduction of the baby bonus. Did Australian women (and doctors) respond to the financial incentives and if so by how much? To address these issues, consider the following model:

$$births_t = \beta_1 + \beta_2 WH_t + \beta_3 A_t + u_t$$

where
$births_t$ = the number of births in Australia on day t;
$WH_t = 1$ if day t is a weekend or a public holiday (=0 otherwise);
$A_t = 1$ if day t is after June 30 (=0 otherwise).

(a) How do you interpret the regression parameters for this model?

b) Do you think this is a good research design and associated model specification to address this problem? Explain.

- R is case sensitive!!!
- Choose directory of dataset

a)

Again, this is a "difference-in-means" regression as A is a binary variable that divides births according to whether the time is after the introduction of the Baby Bonus (BB) or not. But here there is an added control for weekends and public holidays; a variable that is justified by observing the obvious weekly cycle in the data (doctors prefer not to work on weekends).

Because there are 4 possible regimes, this complicates the interpretation of the parameters:

$$E(births|WH = A = 0) = \beta_1$$
$$E(births|WH = 1, A = 0) = \beta_1 + \beta_2$$
$$E(births|WH = 0, A = 1) = \beta_1 + \beta_3$$
$$E(births|WH = 1, A = 1) = \beta_1 + \beta_2 + \beta_3$$

Thus, β_1 is the mean births on weekdays before the BB.

β_2 is the difference in mean births between weekdays & weekends irrespective of whether it is before or after the BB.

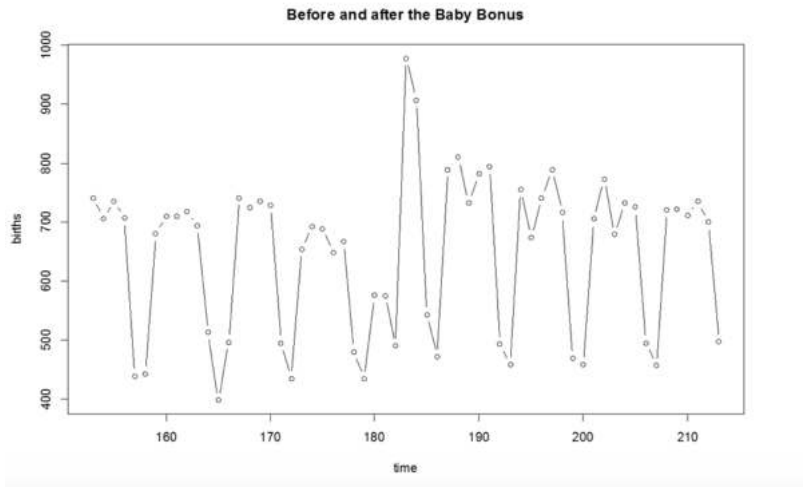
β_3 is the difference in mean births before & after the BB irrespective of whether it is a weekday or weekend.

b)

There are some problems with this design including the likelihood of an interaction effect between the BB and whether the day is a weekday or not. But more fundamentally it "imposes" a once-off permanent change (presumably increase) in births after the BB.

Because of the limited time interval here, every woman who gave birth in this period was already pregnant at the start of the period. The model is inconsistent with the data chosen.

With these data what can be identified is a shift in the birthdate of babies that would be born during this period. See Gans & Leigh (2009) for an extensive analysis of these data. They estimate a large "shift" in the timing of births of over 1000 births that can be attributable to the policy change.



Note: Salient message is that the model to be estimated needs to make sense! A useful general modelling lesson is to describe the model in words before doing any estimation and ask is it sensible/reasonable? It can never match reality exactly, but it should capture the main features of the process being modelled – think of models as maps.

- R is case sensitive!!!
- Choose directory of dataset

5. Upgrading infrastructure, such as freeways, has the potential to increase property values for impacted dwellings. Here the claim is that a major freeway upgrade has led to an increase in the value of residential houses.

a) Explain how you would design an experiment to test this causal claim. Is the experiment you have proposed feasible to conduct?

b) Instead of the experiment outlined in (a), consider investigating the claim by using available observational data contained in housing.xls. This is a sample of house sales over 12 months covering 9 months before and 3 months after the completion of the freeway upgrade. The data includes information on dwelling characteristics including a dummy variable, $upgrade$, that indicates whether the house was in an area with ready access to the freeway. Using only the data where $upgrade$ = 1, estimate the regression, $price = \beta 0 + \beta 1 after + u$, where $after$ is a dummy variable that is equal to one when the sale is made in the three months after the completion of the upgrade and zero otherwise. Interpret these results and discuss whether they are consistent with the claim of improved property values

c) Identify the potential problems associated with the approach in (b) by contrasting it with the experiment you described in (a).

d) Consider using a difference-in-difference approach using the entire sample to estimate following the regression model. What do you conclude from these results?

$$price = \beta_0 + \beta_1 upgrade + \beta_2 after + \beta_3 upgrade \times after + u$$

e) Someone evaluating your results notes that the freeway extension only impacts people living in the outer suburbs where prices are lower. They are concerned that the results in (d) may be biassed and your conclusions problematic. Run the following regression to explore this concern. What do you conclude?

$$price = \beta_0 + \beta_1 upgrade + \beta_2 after + \beta_3 upgrade \times after + \beta_4 distance + u$$

a)

- Randomly allocate houses to whether they benefit from the upgrade and compare average selling prices between the two groups.
- Upgrade the freeway and observe the selling price of a sample of houses. Now rewind the clock and don't upgrade the freeway and observe the selling prices of the same houses. Attribute any differences in average selling prices to the upgrade.
- Clearly, neither is feasible

b)

- R is case sensitive!!!
- Choose directory of dataset

property values.

$t_{stat} > t_{crit}(\alpha = .05)$

$\widehat{price} = 1212.8 + 122.4after$
$SE \quad (24.4) \quad (68.3)$
$n = 282,\ R^2 = .011, standard\ errors\ in\ (.)$

p-value < 0.05

$t_{stat} = \frac{\hat{\beta}_1}{SE} = \frac{122.4}{68.3} = 1.79$

- Houses sold after attracted a selling price that was $122,400 higher and you would reject the null hypothesis of no effect at the 5% level with a one-tail test.
- The results are consistent with the claim of improved property values BUT see next part.

c)

- Worry that this estimate reflects, at least in part, other factors. A confoundment problem.
- Maybe just an overall increase in property values irrespective of the upgrade (inflation?)
- Such problems were avoided by either winding back the clock or random assignments.

d)

$$\widehat{price} = 1300.7 - 87.9upgrade + 80.3after + 42.1upgrade \times after$$
$$(11.1) \quad (44.9) \quad (29.5) \quad (125.3)$$
$$n = 4{,}663,\ R^2 = .003, standard\ errors\ in\ (.)$$

The DiD estimate of $42,100 is still positive but much smaller than in (b).
Moreover, the estimate is very imprecise (CI is wide) and formally you would not reject the hypothesis of no effect at say 5%. -→ not statistically significant

e)

The results from the extended regression model are given below:

$$\widehat{price} = 1768.5 + 137.7upgrade + 74.6after + 63.5upgrade \times after - 52.7distance$$
$$(28.5) \quad (45.3) \quad (28.5) \quad (121.3) \quad (3.0)$$

$$n = 4{,}663,\ R^2 = .066, standard\ errors\ in\ (.).$$

It is true that distance from the CBD does imply lower prices (on average $52,700 less for each km ceteris paribus).
This impacts the estimated coefficient on the upgrade, which is now positive and statistically significant, but this is not the parameter of primary interest.
Controlling for distance does not change the qualitative conclusion, the DiD estimate of $63,500 is positive and imprecise.

Note: As well as being statistically insignificant, the Difference-in-difference estimates are relatively small although maybe not for a student. But relative to prices in the millions (see the first regression) these are small differences. It is good to remind students that the magnitude of an effect is often more important than its statistical significance.

- R is case sensitive!!!
- Choose directory of dataset

Lecture 9: Data Ethics

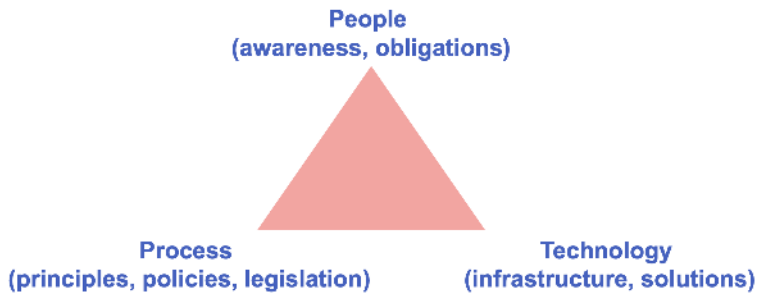
- Ethics: “moral principles that govern a person's behaviour or the conducting of an activity
- Data ethics:
    - Moral obligations of gathering, protecting, and using personally identifiable information and how it affects individuals”
    - “A new branch of ethics that studies and evaluates moral problems related to data, algorithms, and corresponding practices”

- A Tale of Two Schools

	Deontology	Utilitarian
Origins	Coined from Greek “Deon” meaning duty and care	Founder: Jeremy Bentham
Main Focus	Moral duties, irrespective of consequences	Do our actions maximise the positive outcome (utility) for most people?
Keywords	Duty for duty’s sake, Virtue is its own reward, Rule-based approach	Societal perspective, Public happiness, Minimum Pain, Consequentialism, Greatest Good

- Becoming a new source of competitive advantage
    - Responsible business practices – using data for good
    - Maintain trust between companies and customers and business partners
    - Comply with government and industry regulations
    - Enhance business reputation
    - Reduce cost

- Data ethics phenomena



- R is case sensitive!!!
- Choose directory of dataset

- Data Ethics – An analytics lifecycle perspective
    1. Define business objectives
    2. Collect data
    3. Prepare and explore data
    4. Create training and test datasets
    5. Build and improve the model
    6. Deploy the model

- Data Privacy Principles
    - "the claim of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others"
    - **Notice**: inform users about privacy policy, privacy protection procedures
        - e.g. who will be collecting data, how data will be collected, who owns data
    - **Choice and consent:** consent from individuals about the collection, use, disclosure, and retention of their information
    - **Use and retention:** data should be retained and protected according to law or business practices required
        - e.g. the length of data retention; avoid secondary use of data for other purposes
    - **Access:** provide access to individuals with the access to review, update, and modify the data about their personal information
    - **Protection:** data is used only for the purpose stated; de-identifiable of sensitive information; users have the right to opt out for the use of their data
    - **Enforcement and Redress:** provide channels for individuals to report, provide feedback, or complain

- Australian privacy principles
    1. Open and transparent management of personal information
    2. Anonymity and pseudonymity
    3. Collection of solicited personal information
    4. Dealing with unsolicited personal information
    5. Notification of the collection of personal information
    6. Use or disclosure of personal information
    7. Direct marketing
    8. Cross-border disclosure of personal information
    9. Adoption, use or disclosure of government related identifiers
    10. Quality of personal information
    11. Security of personal information
    12. Access to personal information
    13. Correction of personal information

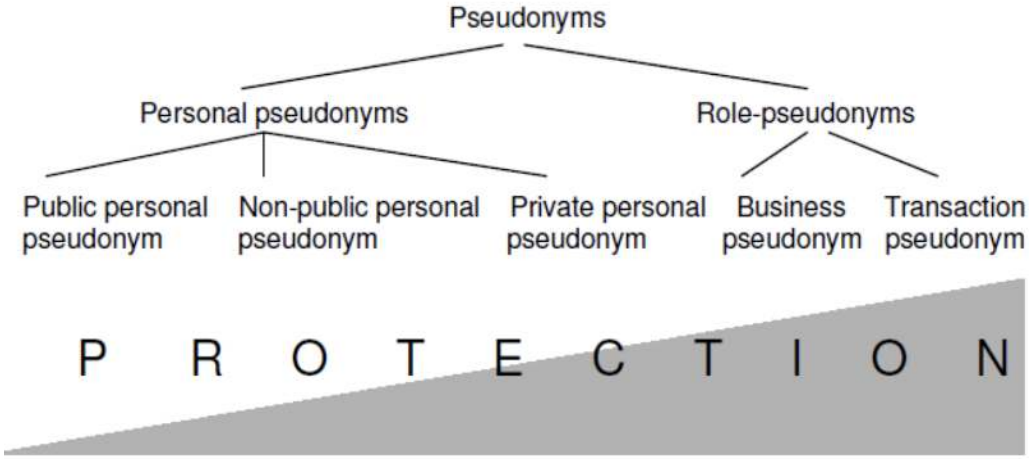
- Data types under protection

- R is case sensitive!!!
- Choose directory of dataset

    - Identify data – name, address, personal number
    - Demographic data – gender, age, education, religion, marital status
    - Analysis data –data attributes for which analysis is conducted such as diseases, habits

- Data protection dimensions
    - Anonymity – a user may use a resource or service without disclosing their identity
    - Pseudonymity - a user acting under a pseudonym may use a resource or service without disclosing their identity
    - Unobservability - a user may use a resource or service without others being able to observe that the resource or service is being used
    - Unlinkability - sender and recipient cannot be identified as communicating with each other

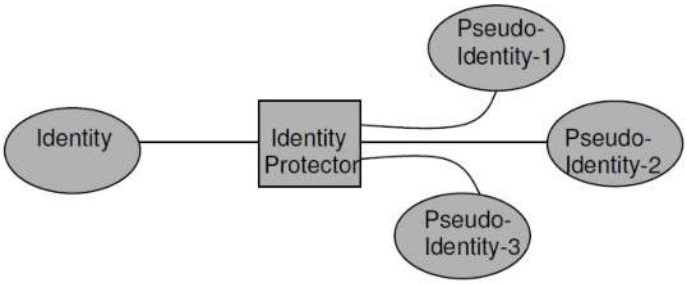
- A Classification of pseudonyms



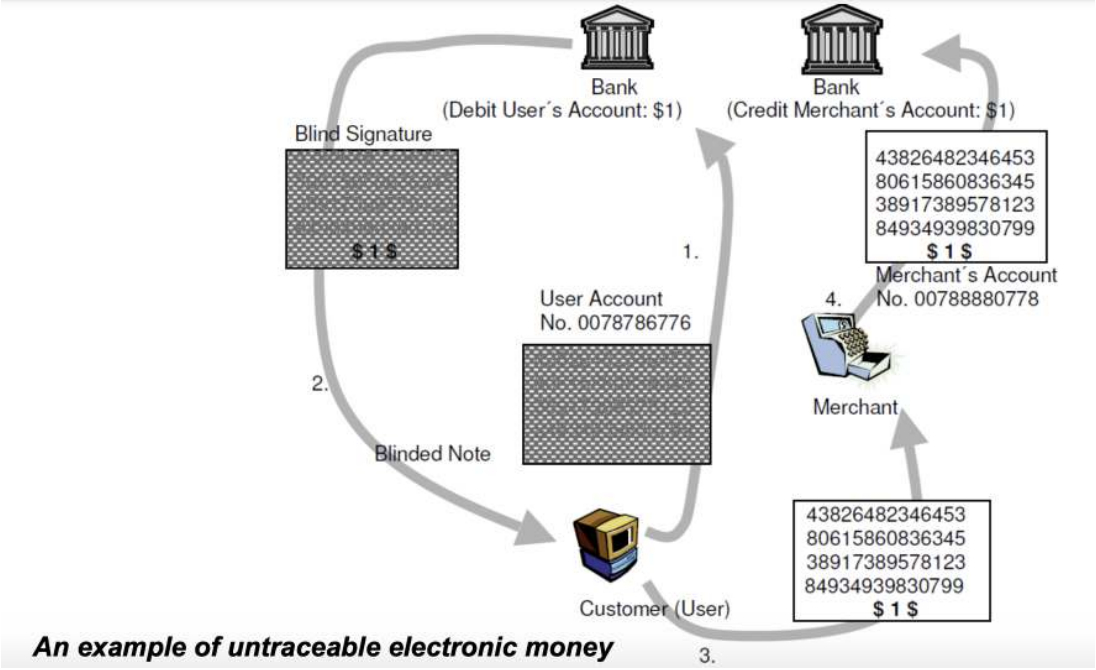
- Identity Protector (IP)
    - Reports and controls instances when identity is revealed
    - Generates pseudo-identities
    - Translates pseudo-identities into identities and vice-versa
    - Converts pseudo-identities into other pseudo-identified
    - Combats fraud and misuse of the system

- Application of IP at the database level

- R is case sensitive!!!
- Choose directory of dataset



- Application of IP at the application level



An example of untraceable electronic money

- Data Security
  - Protection of the data against accidental or intentional loss, destruction, or misuse from internal, external, and natural sources

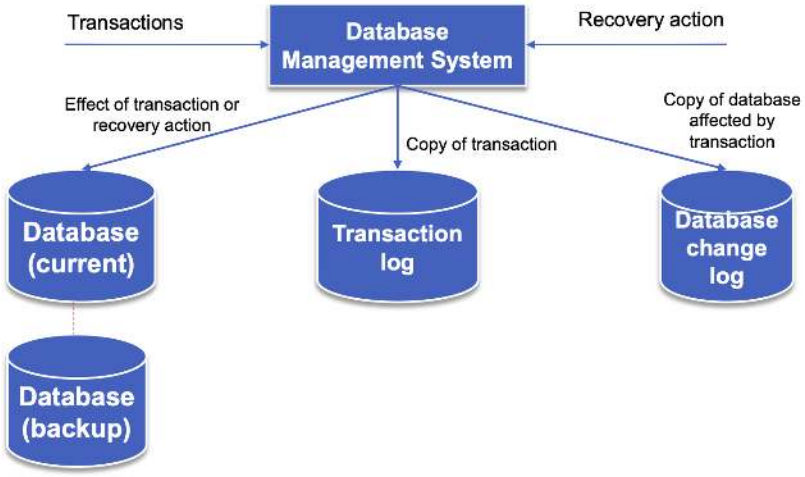
- Ethical Aspects of Data Security
  - Attracting, training, and retaining quality personnel to address ethical issues
  - A perceived potential conflict of interest also exists relative to ethical behaviours and technical knowledge

- Australian Security Principles

- R is case sensitive!!!
- Choose directory of dataset

1. Misuse
2. Interference
3. Loss
4. Unauthorised access
5. Unauthorised modification
6. Unauthorised disclosure

- Data Security Approaches
  - **Triggers:** a system defined rule to handle unexpected events
    - Prohibit inappropriate actions (e.g. changing salary records outside normal business days)
    - Cause special handling procedures to be executed (e.g., penalty applied if payment received after a certain due date)
    - Cause a log file to echo important information to review sensitive data (e.g., reminding users to double check where sensitive information change initiated)
  - Authorization: identify users and restrict the actions they may take against data
    - Identify users and restrict the actions (e.g., read, update, modify) they may take against a data
  - Authentication: identify persons attempting to gain access to data
    - Password or personal identification number
    - A smart card or a token
    - Unique personal characteristics, such as fingerprint or retinal scan
  - Audit trial: maintain the audit and the backup of data changes
    - Audit Trial Example:



Data Bias - Types and Mitigation Strategies

- R is case sensitive!!!
- Choose directory of dataset

- Confirmation Bias
    - People perform data analysis to prove predetermined assumptions
    - How to avoid
        - Record your beliefs and assumptions before starting your analysis
        - Resist the temptation to generate hypotheses or gather additional information to confirm your beliefs.
        - Revisit your recorded beliefs and assumptions at the conclusion of your analysis
- Outlier Bias
    - Uncomfortable truths are hidden behind a good-enough average
    - Outliers can be useful to detect fraud or risks
    - How to avoid
        - Examine the distribution of the sample
        - Use median instead of average
        - Identify and analyse outliers

- Selection Bias
    - Sample is not representative of the population (e.g., A/B testing)
    - How to avoid
        - Randomization
        - Make sure sampling techniques are appropriate
    - E.g. face recognition software fails to detect certain races

- Survivorship Bias
    - Focus on one side of a story e.g. focus on positives only
    - How to avoid
        - Develop thorough understanding of phenomenon before data collection

- Historical Bias
    - Socio-cultural prejudices and beliefs are mirrored into analytics process
    - Certain groups of people are privileged in credit rating systems
    - How to avoid
        - Identify biases in historical sources
        - Develop inclusive data governance frameworks
    - E.g. Gender biassed apple credit card algorithm

- Data Transparency
    - *The principle of enabling the public to gain information about the operations and structures of a given entity* (Heald 2006)
    - Understanding how data was selected, recorded, analysed, and used
    - Being able to access, update, and modify the information
- **How far should we go? - Moral vs Legal**

- R is case sensitive!!!
- Choose directory of dataset

	Ethics	Law
Meaning	Ethics is a branch of moral philosophy that guides people about the basic human conduct	The law refers to a systematic body of rules that governs the whole society and the actions of its individual members
Objective	Ethics are made to help people to decide what is right or wrong and how to act.	Law is created with an intent to maintain social order and peace in the society and provide protection to all the citizens.
Governed by	Individual, Legal and Professional norms	Government
Violation	There is no punishment for violation of ethics.	Violation of law is not permissible which may result in punishment like imprisonment or fine or both.
Binding	Ethics do not have a binding nature	Law has a legal binding.

- **Risk Management for Data Ethics**
    - Identify risk
    - Assess the vulnerability of critical assets to specific threats
    - Determine the expected likelihood and consequences of specific types of outcomes on specific assets
    - Identify ways to reduce those risks
    - Prioritise risk reduction measures

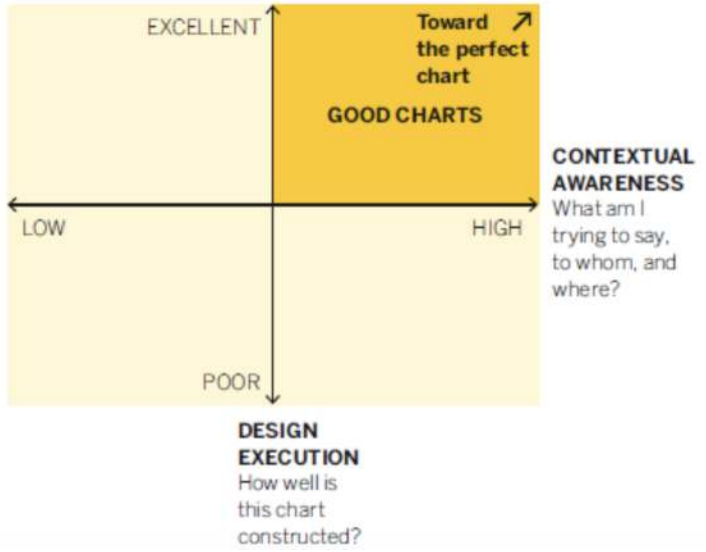
Workshop 9: Data Ethics

- R is case sensitive!!!
- Choose directory of dataset

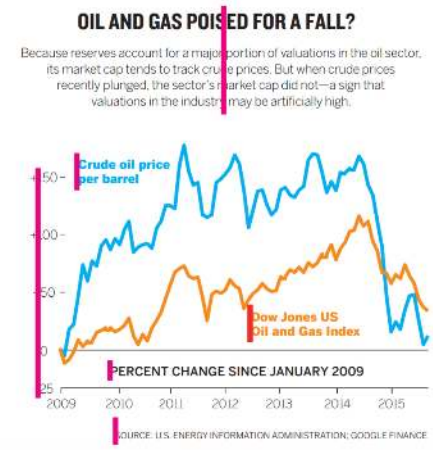
Lecture 10: Data Communication

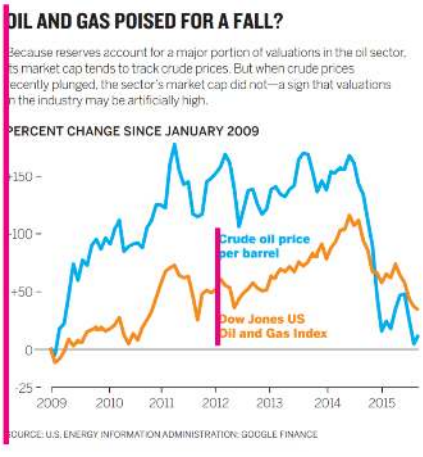
- Rules for good charts
    - Never use pie charts
    - Do not use geomap plots unless geography is relevant
    - Line charts work best for trends
    - Do not focus on whether a chart is “right” or “wrong” but rather focus on whether the chart is good

- Good charts matrix



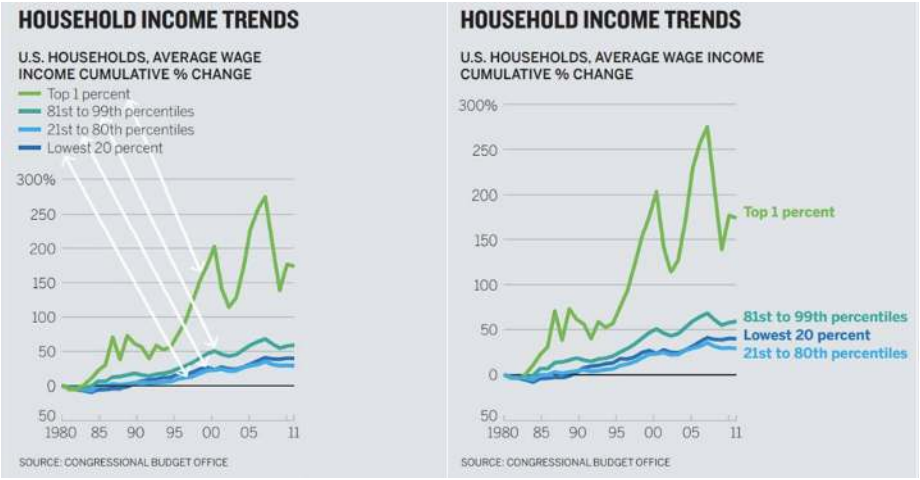
- **Good charts keep elements aligned**



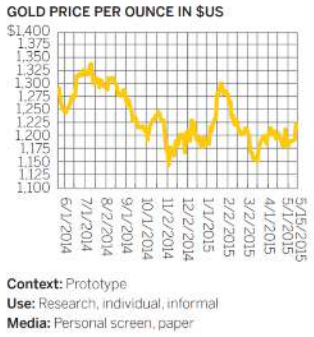


- **Good charts limit eye travel**

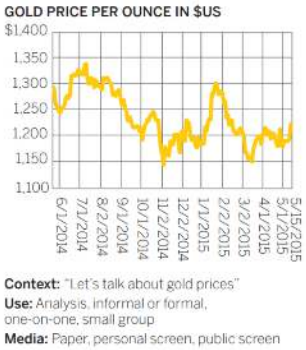
- R is case sensitive!!!
- Choose directory of dataset



- When is minimalism valuable?



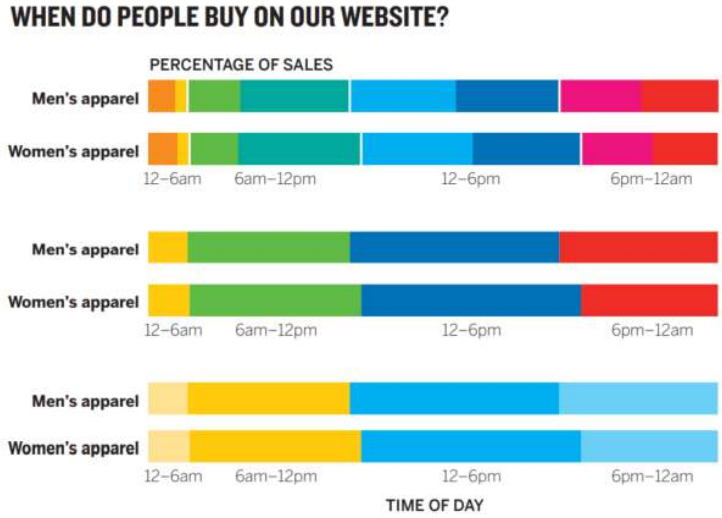
Context: Prototype
Use: Research, individual, informal
Media: Personal screen, paper



Context: "Let's talk about gold prices"
Use: Analysis, informal or formal, one-on-one, small group
Media: Paper, personal screen, public screen

Context: "Gold prices are dropping this year"
Use: Presentation, formal, small or large group
Media: Paper, small screen or large screen

- Remove redundancy within key elements



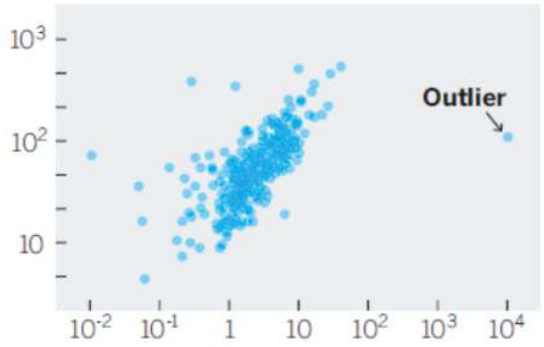
- **How does a chart hit your eyes?**

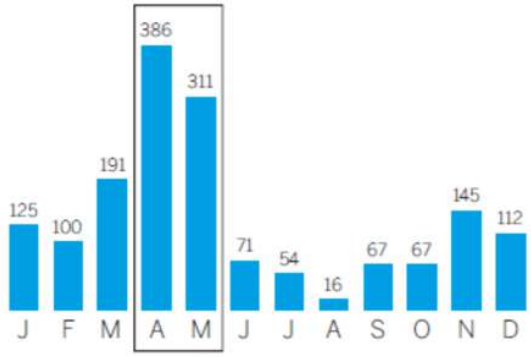
- R is case sensitive!!!
- Choose directory of dataset

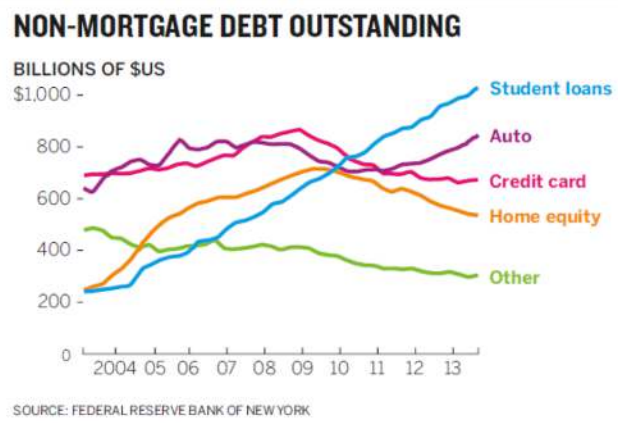
    - Ordered colour schemes
    - We DO NOT see in order (i.e., left to right) → We first see what stands out

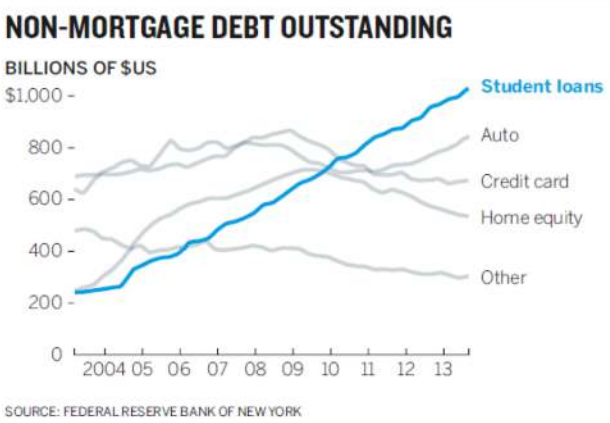
- Good Charts Make a Case:
    - Competing:
        - Attention, resources, financing
    - Persuading:
        - Pitching clients, swaying opinions, recruiting customers
    - Lead to actions

- Make your point stand out by **emphasising, isolating, removing or adding info** (In order of the graphs below)

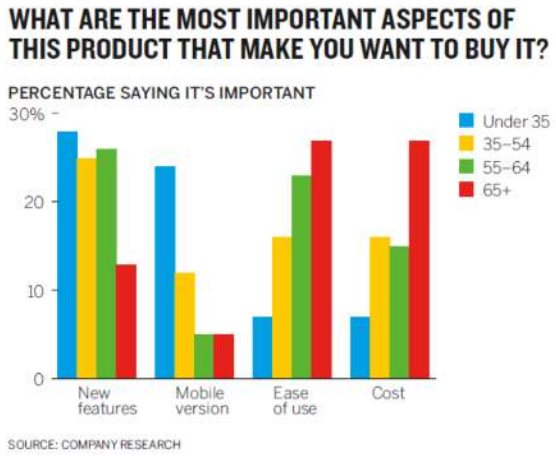


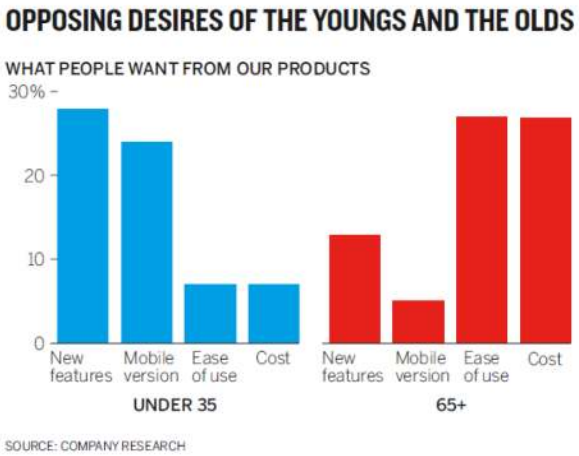






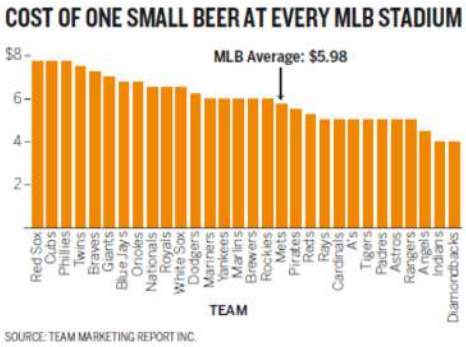
- R is case sensitive!!!
- Choose directory of dataset





What am I trying to say or show?

I am trying to show the distribution of costs of buying a beer at baseball stadiums.



I need to convince them that . . .

I need to convince them that beer is unreasonably expensive at every single baseball stadium.

- R is case sensitive!!!
- Choose directory of dataset