POLS0008
INTRODUCTION TO QUANTITATIVE RESEARCH METHODS

WEEK ONE: UNDERSTANDING DATA

Dr Andreas Mastrosavvas (a.mastrosavvas@ucl.ac.uk)

About the course

Description of the POLS0008 Course

- This module will introduce students to the key tenets of quantitative methods (or statistics) in the social sciences. It assumes no knowledge of quantitative methods or statistical software. Hence, it caters for students from diverse disciplinary backgrounds and adopts a practical hands-on approach to learning, with tutor supported computer tutorials.

- This module covers descriptive statistics (central tendency and variation), data visualisation, data access, probability, sampling, hypothesis testing, inferential statistics and ends with an introduction to simple linear regression. Students will be introduced to the R statistical software and work with real-world data

Objectives of POLS0008 Course

To develop your understanding

- To introduce foundational concepts and key tenets of quantitative methods
- Statistical methodology (i.e., exploratory and inferential)
- To get you thinking critically about quantitative research findings

To provide you with practical experience with Statistics

- To provide practical experience in working with secondary data and quantitative methods
- Confidence building – tackling ‘statistics anxiety’
- Become an R/RStudio programmer

Develop transferable skills

- Coding, presenting research finding and writing

Meet the module tutors

Professor Stephen Jivraj (Convenor)

- stephen.jivraj@ucl.ac.uk
- Office: 317, 1-19 Torrington Place
- Office Hours: See Moodle

Dr Andreas Mastrosavvas (Co-convenor)

- a.mastrosavvas@ucl.ac.uk
- Office: Room 345, 1-19 Torrington Place
- Office Hours: See Moodle

Meet the Postgraduate Seminar Tutors

Angel Torres Guevara

a.torresguevara@ucl.ac.uk

European and International Social and Political Studies

Giovanni Hollenweger

giovanni.hollenwger.21@ucl.ac.uk

European and International Social and Political Studies

James Rice

james.rice@ucl.ac.uk

Department of Political Science

General format of the course

- **2-hour lectures**
  - Weekly sessions every Tuesday 12:00pm – 02:00pm
  - Delivered by the module tutors

- **1-hour computer practical seminar**
  - 9 groups on Thursday from 10:00am, up to 03:00pm
  - These sessions are facilitated by the postgraduate seminar tutors

NOTE: All lectures and tutorials sessions are compulsory & will be delivered in-person.

Module Outline & Assessment

Exploratory Statistics (Andreas)

- **WK01**: Understanding Data & Introduction to RStudio
- **WK02**: Examining Data I (Descriptive Statistics)
- **WK03**: Examining Data II (Descriptive Statistics)
- **WK04**: Sourcing Data

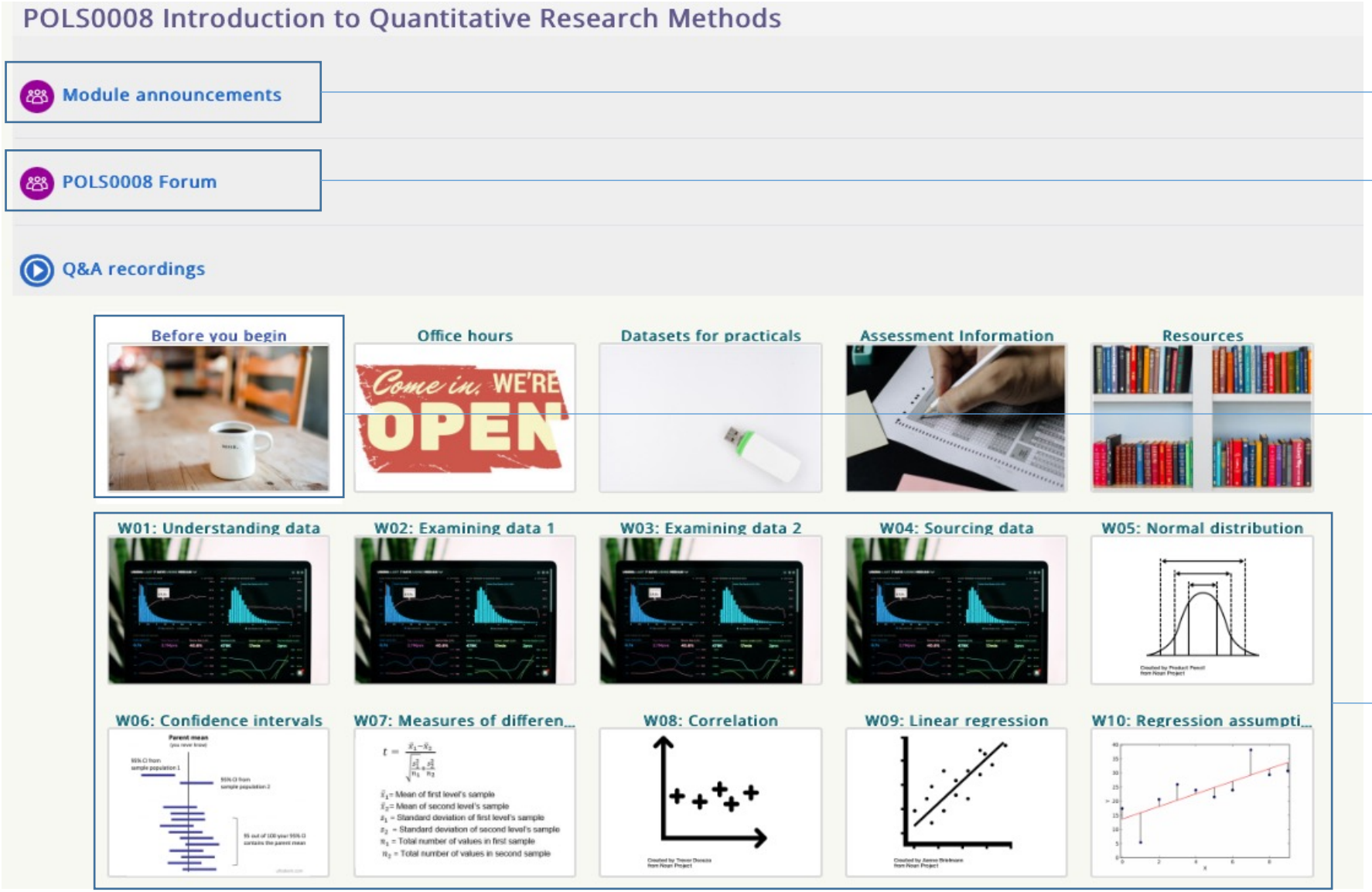
Hypothesis testing & Inferential Statistics (Stephen)

- **WK05**: Normal Distributions
- **WK06**: Confidence Intervals
- **WK07**: Measures of Difference
- **WK08**: Correlation
- **WK09**: Linear Regression
- **WK010**: Regression Assumptions

Details about assessment

It's a 3,000 words essay which will be based on secondary analysis of survey data. It is worth 100% of your final marks for this course. The deadline will be confirmed soon.

Access to the content on Moodle [1]



For module and seminar tutors for making an important announcement

For students to post questions on problems etc.,

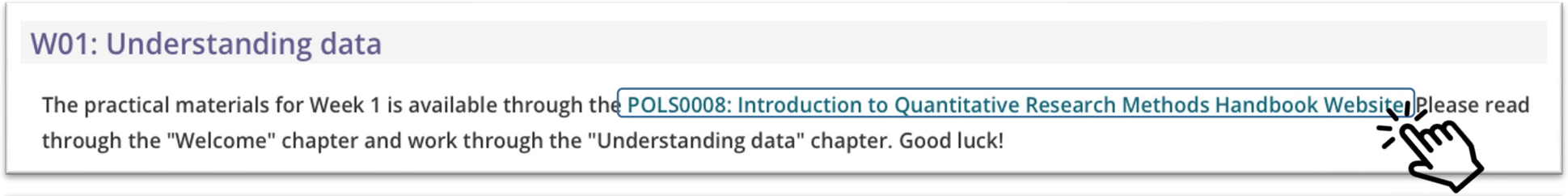
Module outline document is contained here. Everything you need to know included read list etc is found there.

All our lecture and teaching materials are hosted in these sections. Click to access them.

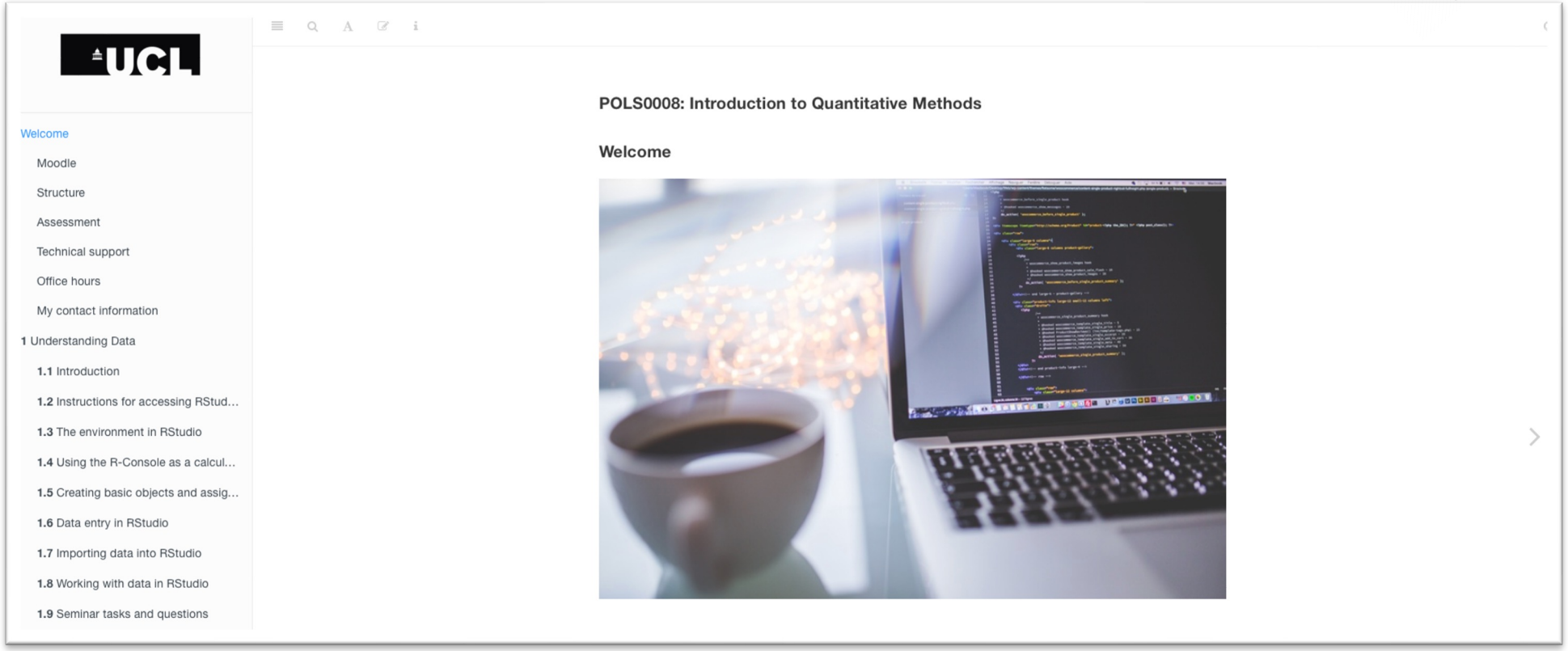
Access to the content on Moodle [2]

1.

2.

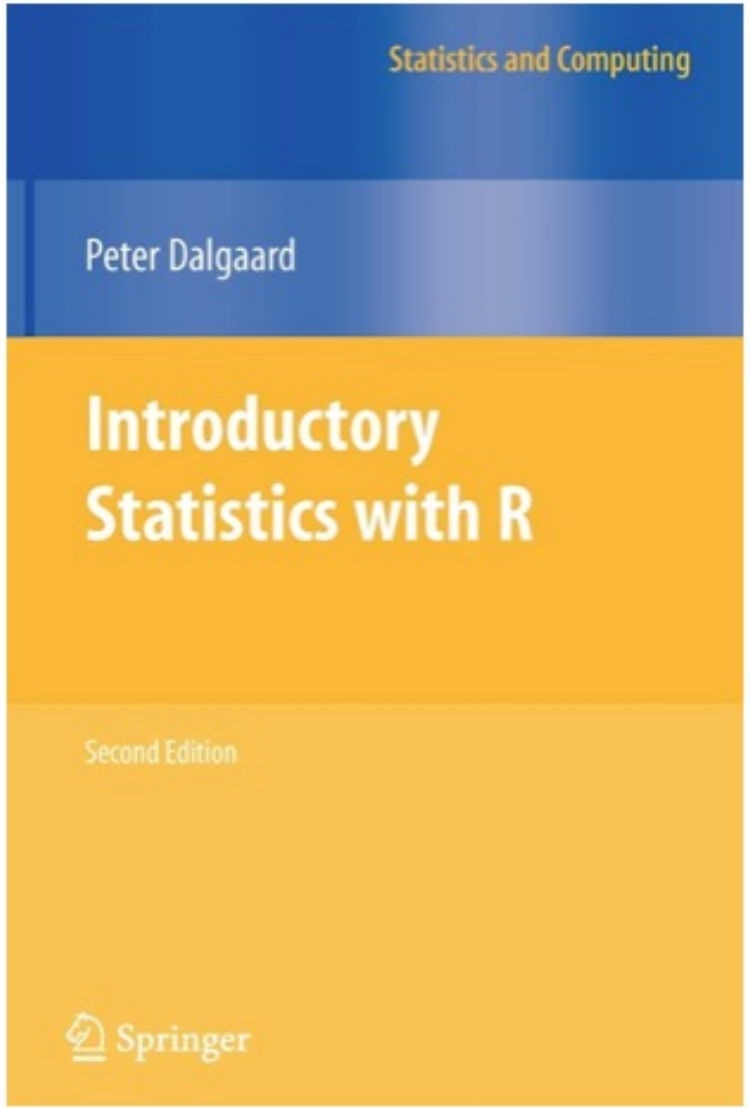


3.



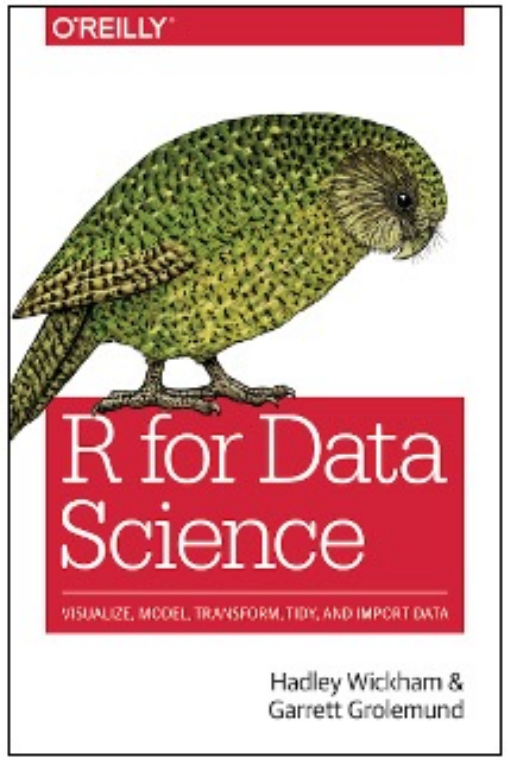
All my lecture notes, reading lists, as well as guided videos and datasets for seminars for computer seminars are hosted on a GitHub website

Recommended books for POLS0008



High recommendation for the mastery of basic theory, principles and R/RStudio coding for statistical analysis

High recommendation for learning the basics and attaining mastery of the 'base-R' coding etiquettes of R/Studio.

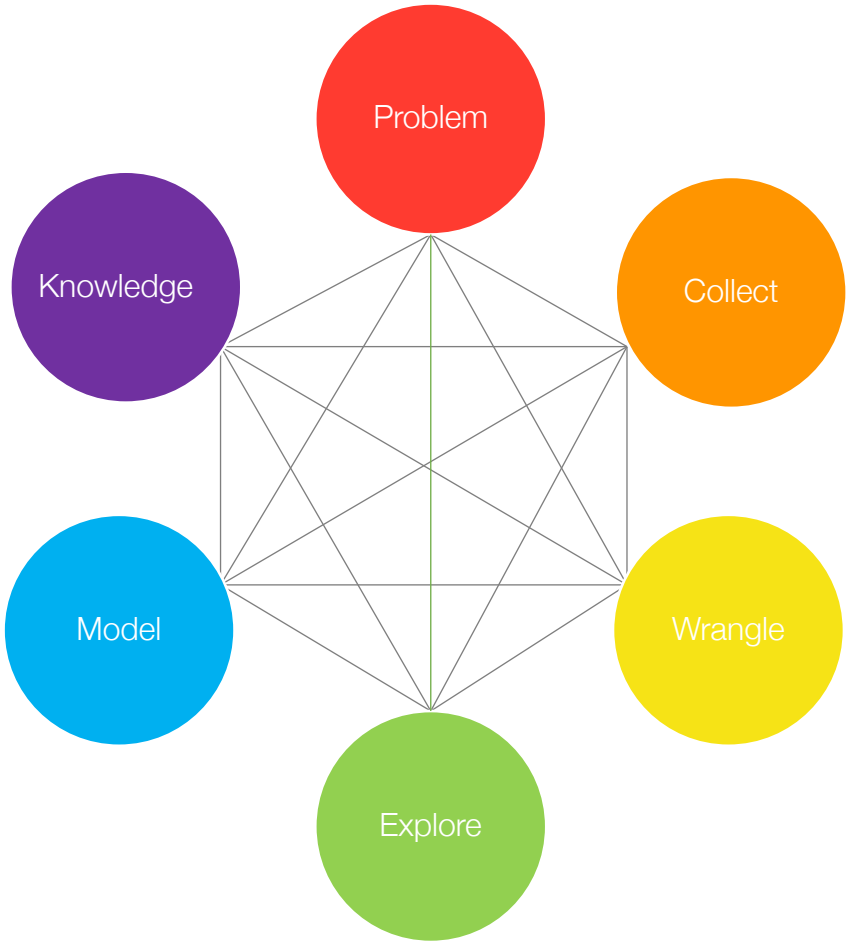


High recommendation for learning the basics and attaining mastery of the 'tidyverse-R' coding etiquettes for R/Studio.

Lecture

Contents

- Why is statistical methods important?
- What is statistics as subject?
- Basic building blocks: Understanding the different data types
- What is descriptive statistics and introduction to summary measures and frequency distributions?
- RStudio and live demonstrations



Format of today's lesson goes…

Theory & Application

15-minute break

Live demonstration in RStudio

Why are statistical methods important?

Why are Statistical methods important?

- **It's an interdisciplinary subject with many applications**
  - **Social sciences** (e.g., quantitative criminology, political science, housing markets and demography etc.,)
  - **Epidemiology** (e.g., finding associations between modifiable lifestyle risk factors and cancers; quantifying of risks environmental pollutants and respiratory tract disease in children etc.,)
  - **Business and Marketing** (e.g. understanding customer mobility and patronage at stores; customer behaviour etc.,)

- **Used for evidence-based research**
  - Certain types of statistical methods are best for establishing evidence of a particular phenomena. It's always best to use these methods with a observational study design framework (i.e., case-control, cohort or Randomised Control Trial (RCTs)).

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (100)

RELATED

Executive Summary

ALSO AVAILABLE
- Buy PDF

Artwork: **Tamar Cohen, *Andrew J Buboltz,*** 2011, silk screen on a page from a high school yearbook, 8.5" x 12"

Download a free chapter from Thomas H. Davenport's book *Keeping Up with the Quants*.

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Why knowing statistics and data science is important?

Well… its incredibly promising in terms of career!

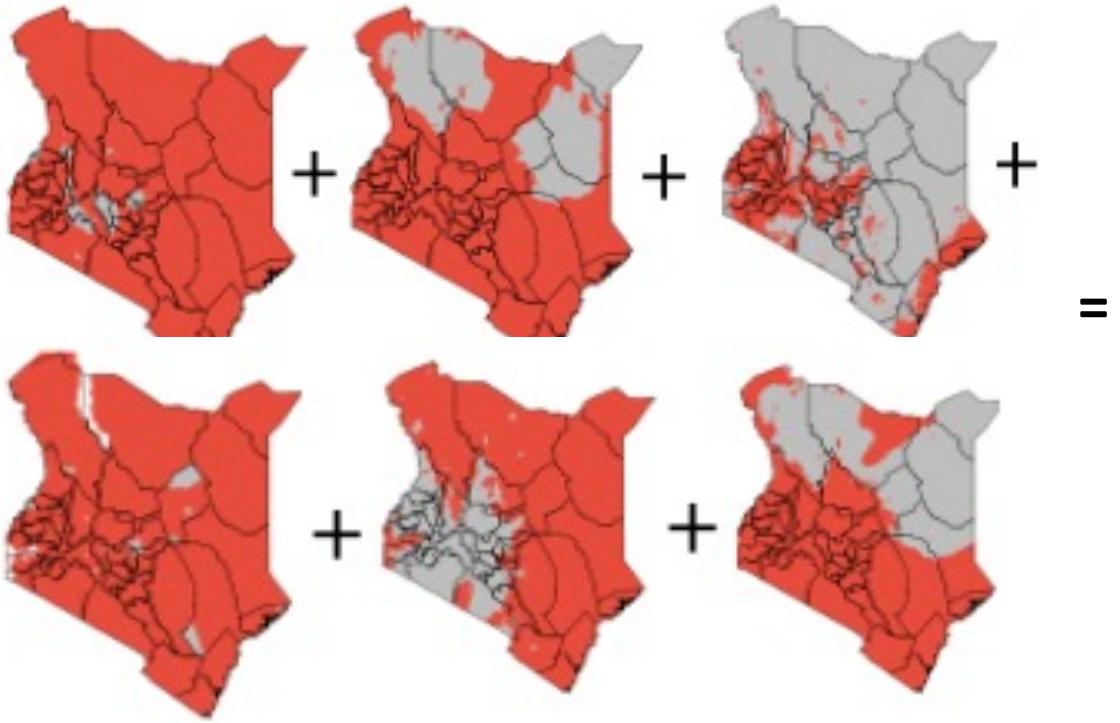
The "hottest skill set" that got people hired in 2014 (and since) is statistical analysis – and this is thanks to the growth of "Big Data". Believe it or not, jobs which demands for statistics and data analytical skills are expected grow an astounding 34% by 2024. Therefore, by studying statistics, the science of studying & analysing from data, you are putting yourself in the ideal position for a career in one of the world's most demanded.

Source: Harvard Business Review | [LINK]: http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

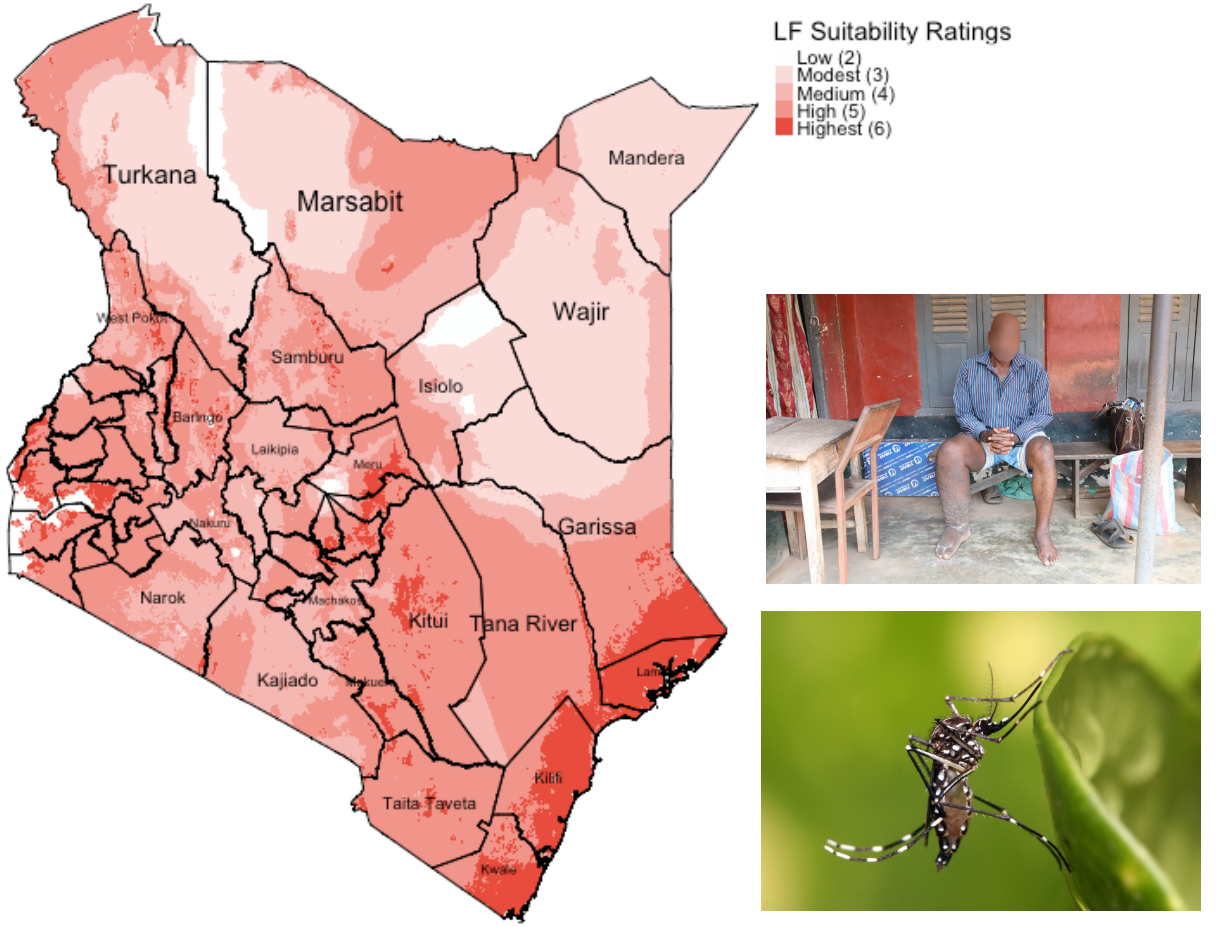

BIG
DATA

“Big” Analysis: identify suitable areas for Lymphatic Filariasis (LF) transmission in Kenya.

Combining 6 different environmental gridded data set to predict the spatial extents of where infectious disease are more likely to transmitted in Kenya



Sources:
1. **Global Atlas for Helminths Infection (http://www.thiswormyworld.org)**
2. **ESPEN (https://espen.afro.who.int)**



Example: Map generated in R to illustrate areas that are environmentally suitable for the spread of neglected tropical disease called ‘Lymphatic Filariasis (LF)’ in Kenya.

Source: Cheshire, J., Uberti, O. (2014). London: The Information Capital, DNA of the City (Census variables reveal London's genetic code), chapter 2: Who we are (Page 76).

To think big, start small

This module is all about high quality data analysis.

- “Small data” can be just as powerful.
- “Big data” may contain a lot of noise.

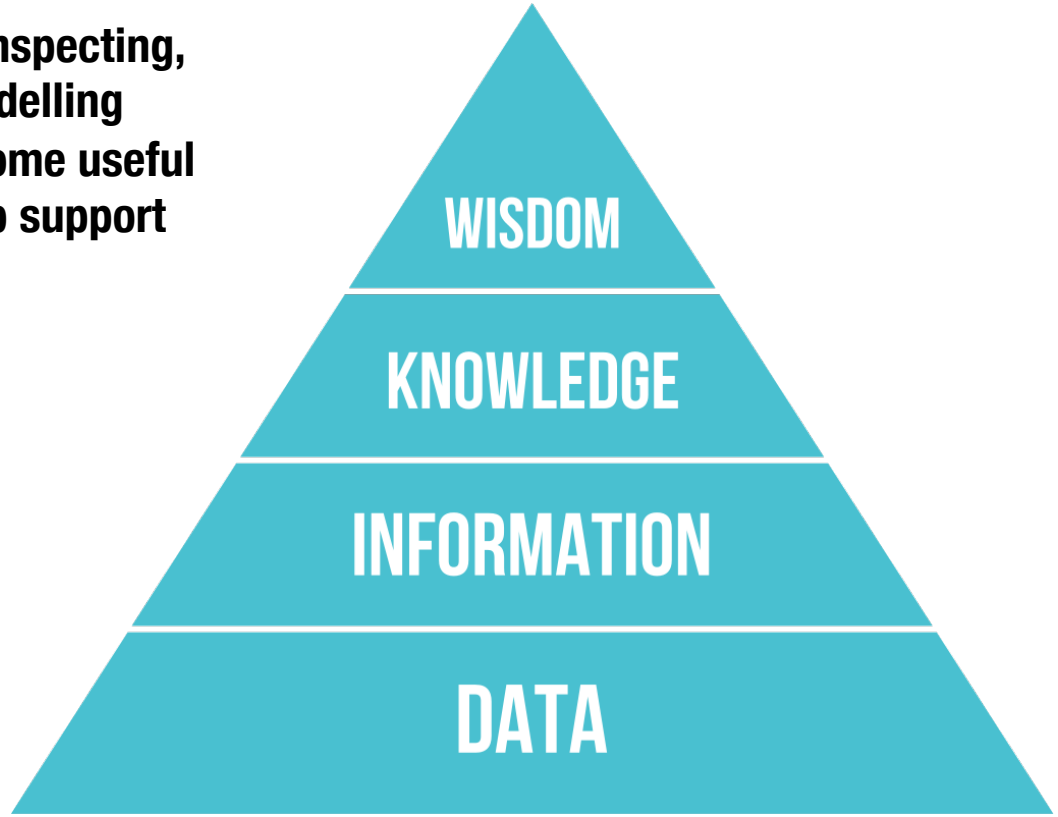
Many of the methods are scalable

- That is, they can be applied to datasets small and big.

All analysis should start with the same basic checks and descriptions of the data.

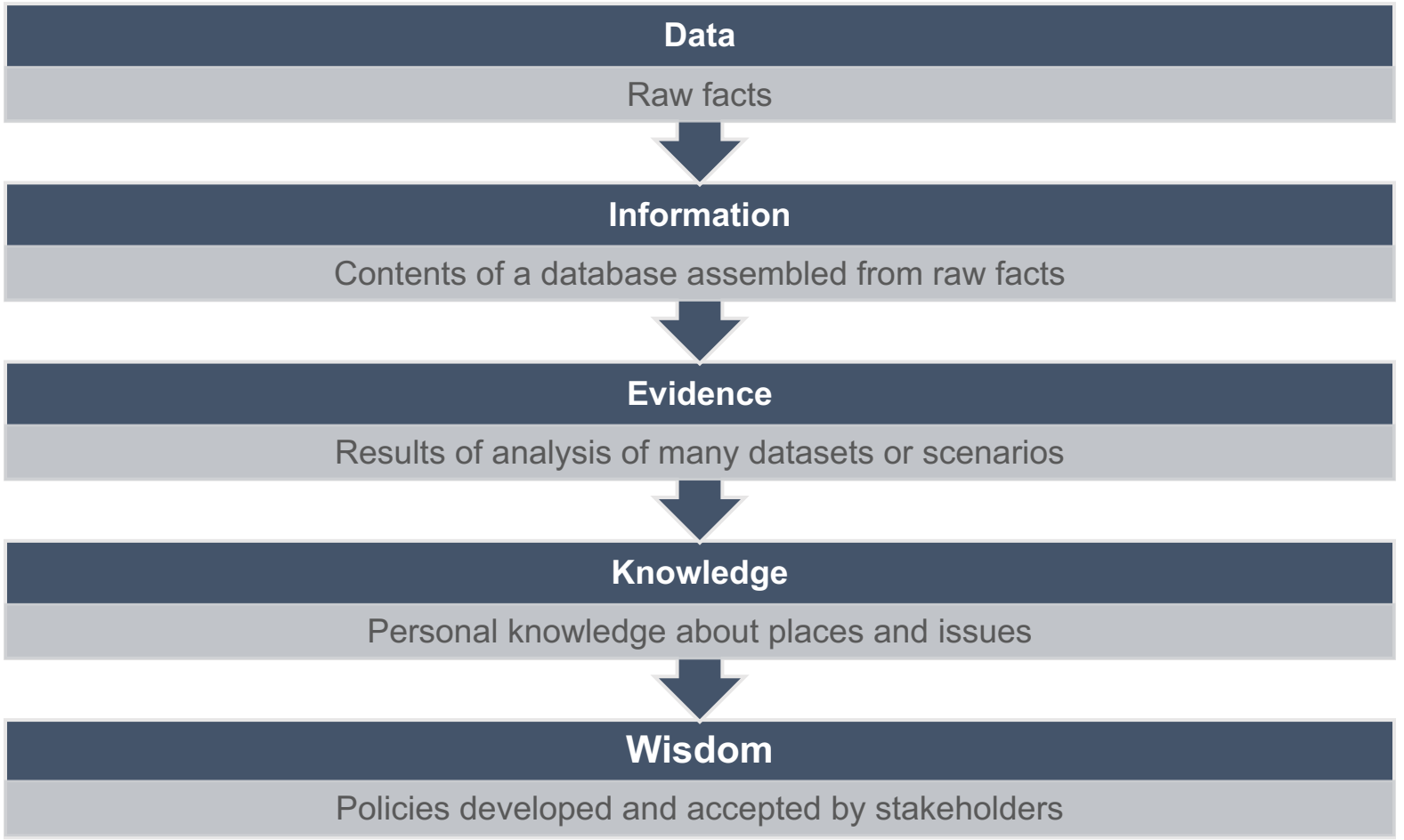
DIKW Pyramid

Data analysis is the process of inspecting, cleansing, transforming, and modelling **DATA** with the aims of gaining some useful insight (or **INFORMATION**) to help support decision making



DIKW is a useful framework for describing the relationship, or structural 'stages' (aka rites of passage) one must go through to gain KNOWLEDGE and WISDOM.

Data vs Information vs Wisdom



What is Statistics?

What is Statistics?

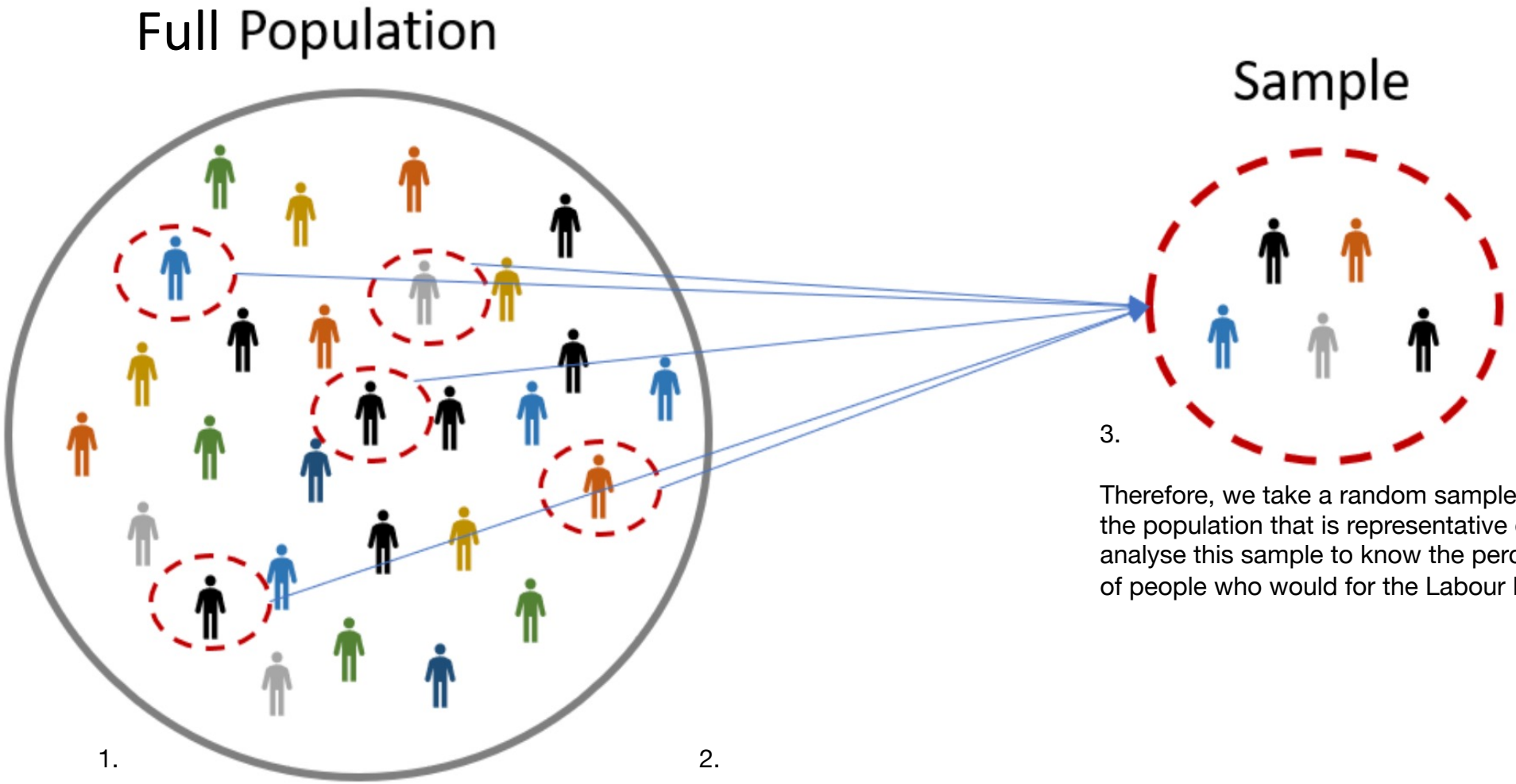
- **Composed of three main facets**
- **Design**
  - How to collect the data (i.e., probabilistic sampling approaches)
- **Description**
  - Describing the way, the data looks
  - Summarizing the data that has been collected
- **Inference**
  - Making predictions about the wider population or about the future
  - Specifically, statistical inference
- **The “descriptive” and “inferential” elements are what we refer to as statistical analysis**

1st Statistical Terminology – Populations vs. Samples [1]

- **Population**
  - The entire possible set of subjects we wish to study
  - These could be states, individuals, businesses, organizations, etc.

- **Sample**
  - The subset of subjects chosen for study through data collection

1st Statistical Terminology – Populations vs. Samples [2]



1.

Say, we want to know the percentage of people in the whole of the UK population who would vote for the Labour Party

2.

It is impossible, or it very tedious for us to collect data on everyone in the UK population

3.

Therefore, we take a random sample from the population that is representative of UK, analyse this sample to know the percentage of people who would for the Labour Party

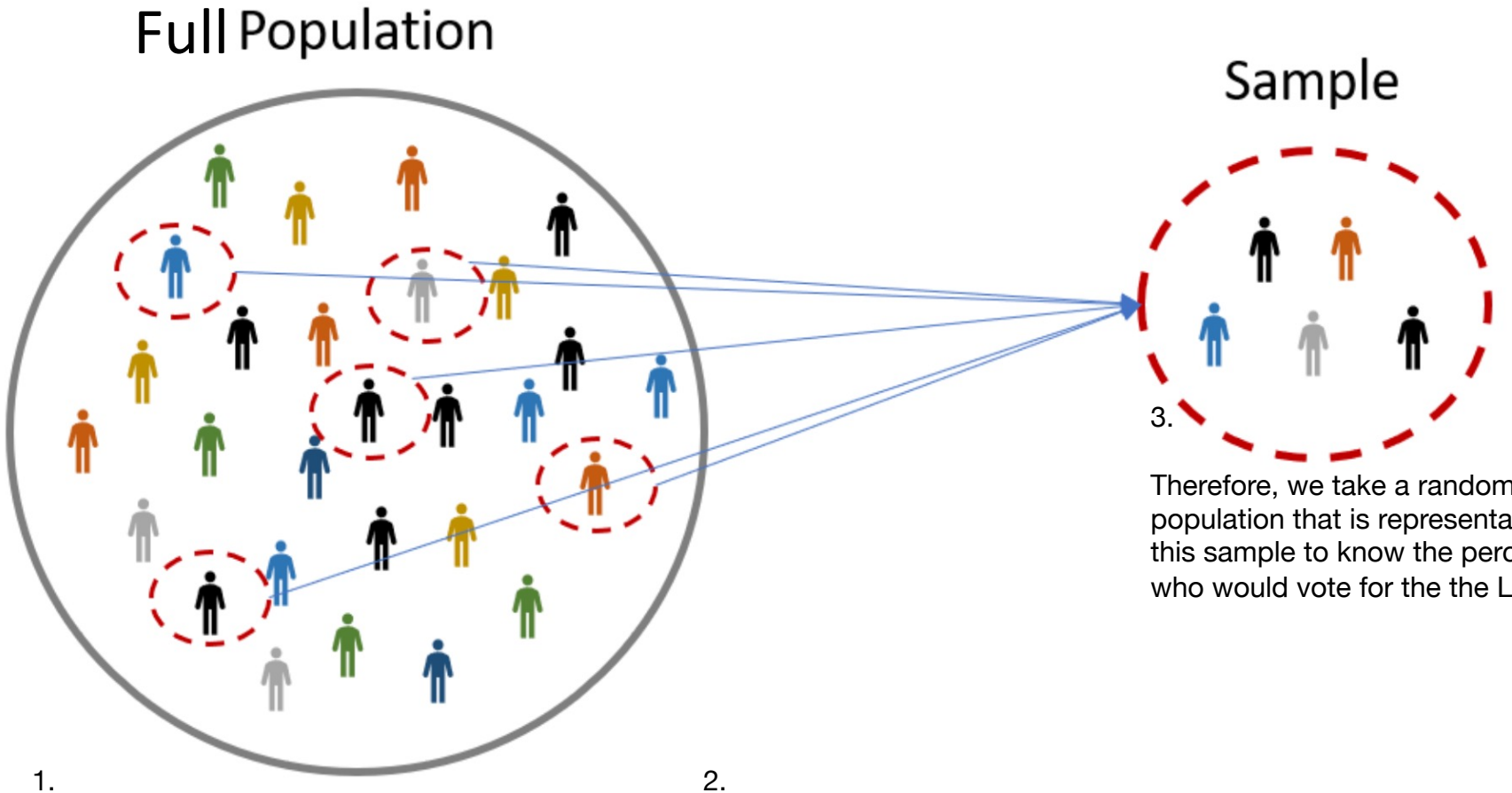
2nd Statistical Terminology – Parameter and Statistic [1]

- **Parameter:**
  - A numerical summary about the **OVERALL** population
  - We rarely know true population parameters, so we use:

- **Statistic(s) (or estimate(s)):**
  - A numerical summary of the sample data
  - Statistics generally contain two bits of information: 1) a measure of central tendency and 2) a measure of variability
  - We could use what we know about how the sample would vote to make inferences to the population

2nd Statistical Terminology – Parameter and Statistic [2]

Parameter: The proportion of people in the UK population who voted for the Labour Party (p)
Statistic: The sample proportion of people from the UK population voted for the Labour Party ($\hat{p}$)



1.

Say, we want to know the percentage of people in the whole of the UK population who voted for the Labour Party

2.

It is impossible, or it very tedious for us to collect data on everyone in the UK population

3.

Therefore, we take a random sample from the population that is representative of UK, analyse this sample to know the percentage of people who would vote for the the Labour Party

More (very basic) Statistical Notation

sample statistic	population parameter	description
n	N	number of members of sample or population
$\bar{x}$ "x-bar"	μ "mu" or μ_x	mean
s (TIs say Sx)	σ "sigma" or σ_x	standard deviation For variance, apply a squared symbol (s^2 or σ^2).
$\hat{p}$ "p-hat"	p	proportion

Basic building blocks: Understanding the different data types and variables

Type of Variables

- A variable is anything that we can measure about the subjects in our sample
- Variables vary, that is they take on a range of values
- There are two classifications of variables

Continuous Variables	Categorical Variables
Interval Ratio	Nominal Ordinal

- Levels of measurement: (Lowest) Nominal << Ordinal << Interval << Ratio (Highest)

Broad types: Continuous & Categorical Variables

- **Discrete variables contain data with countable items:**
  - Number of crimes in London in the last month
  - Number of students in a class
  - Number of languages spoken

- **Continuous variables contains data with measurable items:**
  - Age (in years: 25, 57, 45, 34, 38 etc.)
  - Monthly Income (in £££: 2399.68, £5569.89, £1123.10, £1,450.99, £3847.12 etc.)
  - Height (in meters)
  - Weight (in kg)

- **Categorical variables has categories or groups:**
  - e.g., gender, ethnicity, employment status etc.,

Data types [1]: Nominal

Notes

- Categorical measure
- Discrete set of categories with no natural order
- To distinguish groups with labels
- May be referred to a qualitative or categorical variable
- It is the lowest level of measurement

Examples

- **Gender**
  - 0 = Female
  - 1 = Male

  (Dichotomous or Binary)
- **Race**
  - 1 = Asian
  - 2 = Black
  - 3 = White

  (Polychotomous)
- **Party membership**
  - 1 = Liberal Democrats
  - 2 = Tory
  - 3 = Labour

  (Polychotomous)

Data types [2]: Ordinal

Notes

- Categorical measure
- Discrete set of categories that have some natural order
- There categories have rankings but difference between rankings is not known
- Order matters!
- It is the 2nd lowest level of measurement (above Nominal)

Examples

- **Likert scale** (strongly disagree, disagree, neutral, agree, strongly agree)
- **Socioeconomic status**
  - 1 = Working class (Low)
  - 2 = Middle class
  - 3 = Upper class (High)
- **Size**
  - 1 = Small (Low)
  - 2 = Medium
  - 3 = Large (High)
- e.g., Size, ranking of favourite sports, wellness rankings etc.,

Data types [3]: Interval

Notes

- Continuous measure
- Unlike ordinal variable, difference between categories are known and equal (-must be known to calculate an interval)
- Zero is arbitrary (meaning that whatever observation you measure it does not indicate that its non-existent)
- 2nd best level of measurement
- measurement (above Nominal)

Examples

- Example: Temperature in degree Celsius
  - The difference between 78 degrees and 79 degrees (i.e., is 1 degree) is the SAME as the difference between 45 and 46 degrees (i.e., is 1 degree again).
- Measure of zero degrees Celsius indicates that it does mean that there is no temperature – it only means that its temperature at zero is at freezing point
  - In addition, it does not represent the absolute lowest value, since there are negative values for temperature

Data types [4]: Ratio

Notes

- Continuous measure
- Most precise
- Exact value
- Unlike interval measure, a zero value that there's "nothing" there (not arbitrary)

Examples

- Weight
- Height
- Income
- House price
- Crime rate
- Age

3rd Statistical Terminology [1] – Dependent variable(s)

- **Dependent Variables**

  - The variable to be explained, described or understood.
    - ❖Sometimes called the "**outcome**", "**event**", or "**criterion**" variable
    - ❖Mathematically, it's most often denoted with the letter ***Y***
    - ❖The DV should be dependent upon something else (i.e., independent variables) and…
    - ❖… it should **NOT** affect the independent variable (though this can happen and needs to be considered and accounted for)

  - Like most variables, DVs should vary, if you have a constant DV, you will not be able to explain the effect of other variables on it

3rd Statistical Terminology [2] – Independent variable(s)

- **Independent variables**

  - It should be independent from the effects of the dependent variable (though maybe not always other independent variables)
  - Presumed as the determinant or cause, or something that impacts the dependent variable

    - ❖ It is interchangeable with the term “explanatory” or “predictor” variables. In epidemiology we often say, “risk factors”, in social sciences we say, “social-risk factors” etc.,
    - ❖ Always **antecedent** to the dependent variable being explained
    - ❖ Mathematically, it is often denoted as the letter ***X***

- Again, since it is a variable, the values measured should vary. For instance, think of age, income, employment rates, gender, level of democracy as socio-demographics risk factors for Gross Domestic Product.

Here, we are saying GDP is the dependent variable (or outcome of interest). We are asking ourselves how do these independent variables impact GDP?

What is Descriptive Statistics?

What are Descriptive Statistics?

Univariable analysis

- Analysis of only one variable on some characteristic
  - ❖Frequency Distributions – essentially a count or distribution of values on some single variable
  - ❖Other descriptive statistics – some summary measure that describes the data in a way not obvious by looking at the frequency distribution

Bivariable analysis

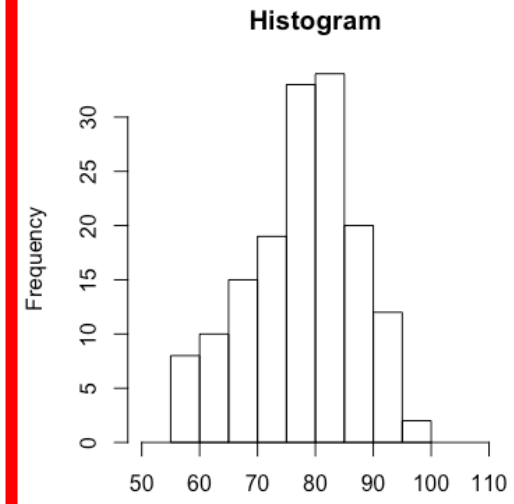
- Analysis of two variables – can be simple scatter plots or cross-tabulations

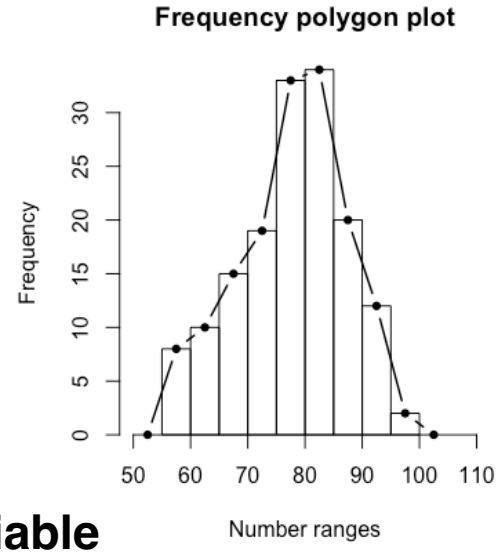
Multivariable analysis

- Analysis of three or more variables

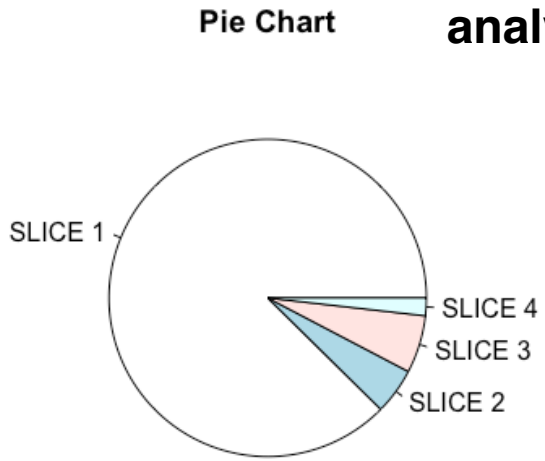
Continuous data

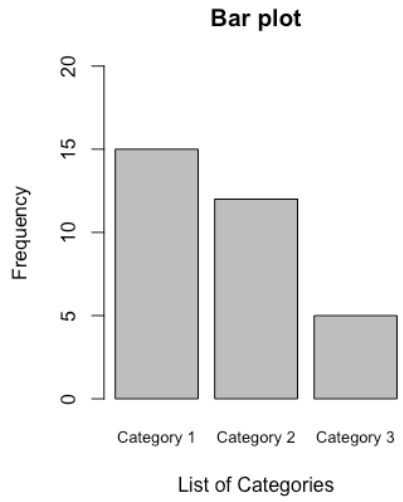
Categorical data

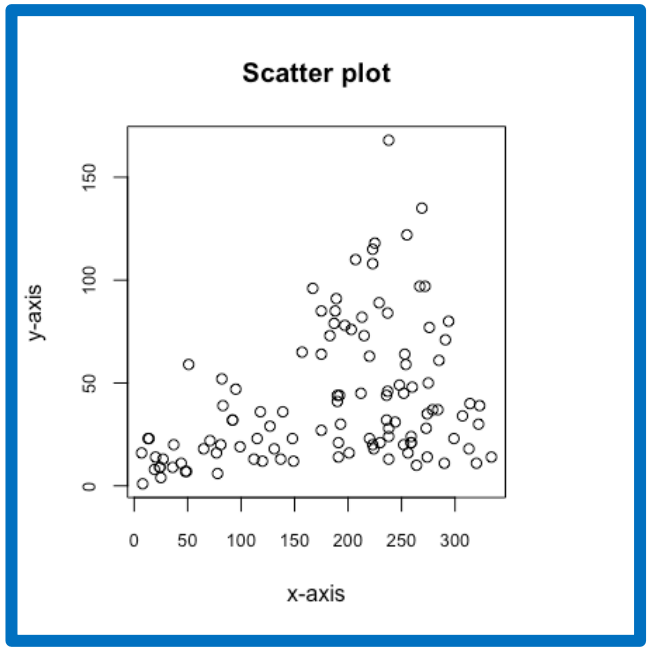




Univariable analysis







Bivariable analysis

Summary Measures

What kinds of analysis and summary statistics can you perform on a particular type of dataset?

1. Frequency distribution

You can group the data by according to categories and perform the following:

Numerical data grouped into categories

- Compute the **Frequencies (counts)**
- Compute the **Percentages (or Relative Frequency)**
- Calculate the **Cumulative Frequencies** or **Cumulative Percentages**

Proper categorical data etc.

- Graphical approaches also include **bar plots** and **pie charts**
- The **Mode** (category with that occurs most)

2. Central tendency measures

You can perform the following analysis:

- Compute the **mode (value that occur most)**
- Compute the **median**
- Compute the **mean**
- **Lowest (Minimum) & Highest (Maximum)**
- **Percentiles**
- **Variance**
- **Standard deviation**
- **Range**
- **Quartiles and Interquartile ranges**

Frequency Distribution

- Can be represented as instances or frequency of an outcome of interest.

Example:

Heights (cm) of 21 (*n*) kids entering reception (aged 4 years). Health-wise, the normal height of a 4-year-old child should be on average 101cm and above:

Dataset: 94, 95, 97, 97, 100, 100, 101, 102, 103, 105, 105, 108, 108, 109, 109, 112, 113, 113, 118, 119, 121

We want to understand the frequency or the occurrence in heights among this cohort of kids.

We create a table that contains group categories for height (of 5cm) measurement, and compute the frequency and proportions. In addition, we compute the cumulative frequency and its cumulative proportion as well.

94, 95, 97, 97, 100, 100, 101, 102, 103, 105, 105, 108, 108, 109, 109, 112, 113, 113, 118, 119, 121

Height groups	Frequency	Relative Frequency or percentage (%)	Cumulative Frequency	Cumulative Relative Frequency
90-95	2	0.09523810 (9%)	2	0.09523810 (9%)
96-100	4	0.19047619 (19%)	6	0.2857143 (28%)
101-105	5	0.23809524 (24%)	11	0.5238095 (52%)
106-110	4	0.19047619 (19%)	15	0.7142857 (71%)
111-115	3	0.14285714 (14%)	18	0.8571429 (85%)
116-120	2	0.09523810 (9%)	20	0.9523810 (95%)
120+	1	0.04761905 (4%)	21	1.0000000 (100%)

We group the data points accordingly

94, 95

97, 97, 100, 100

101, 102, 103, 105, 105

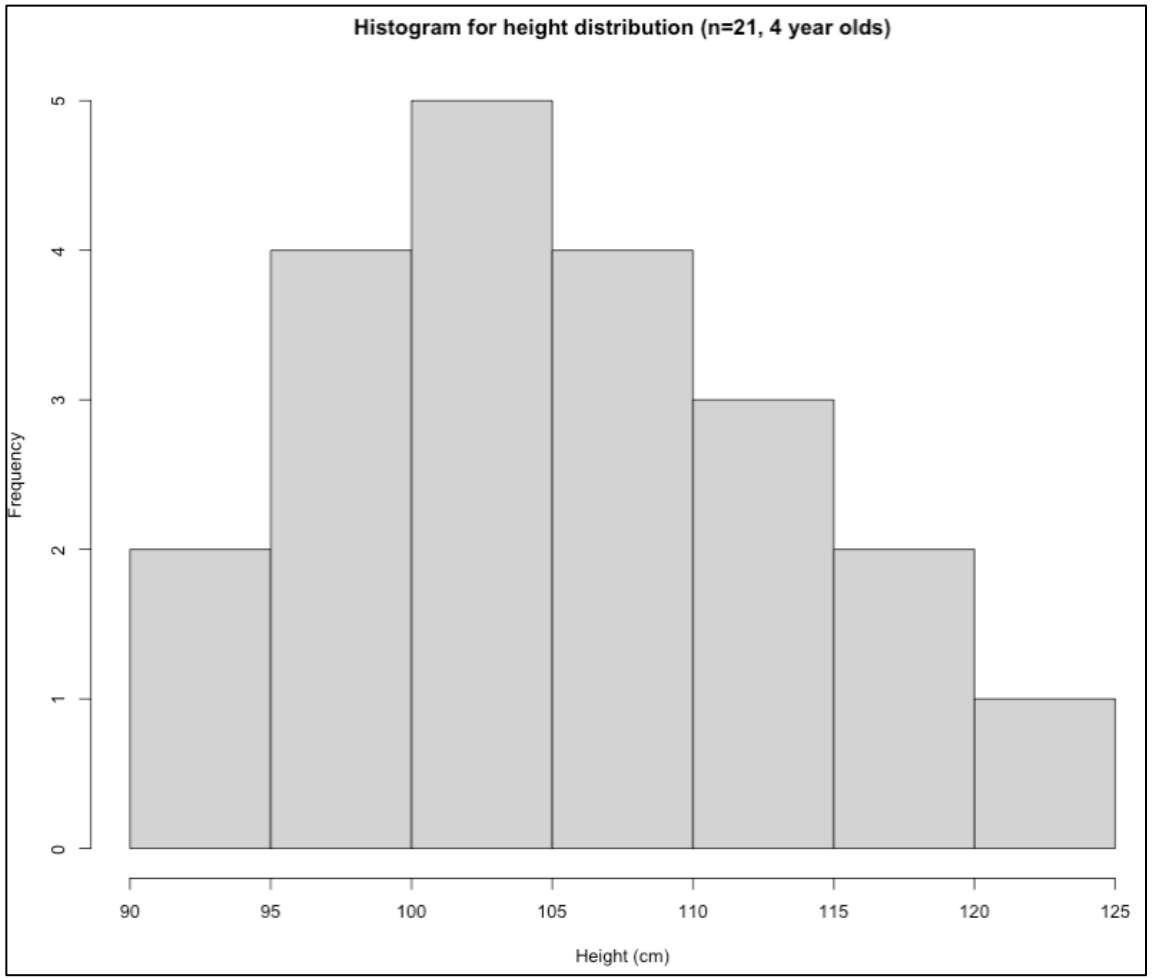
108, 108, 109, 109

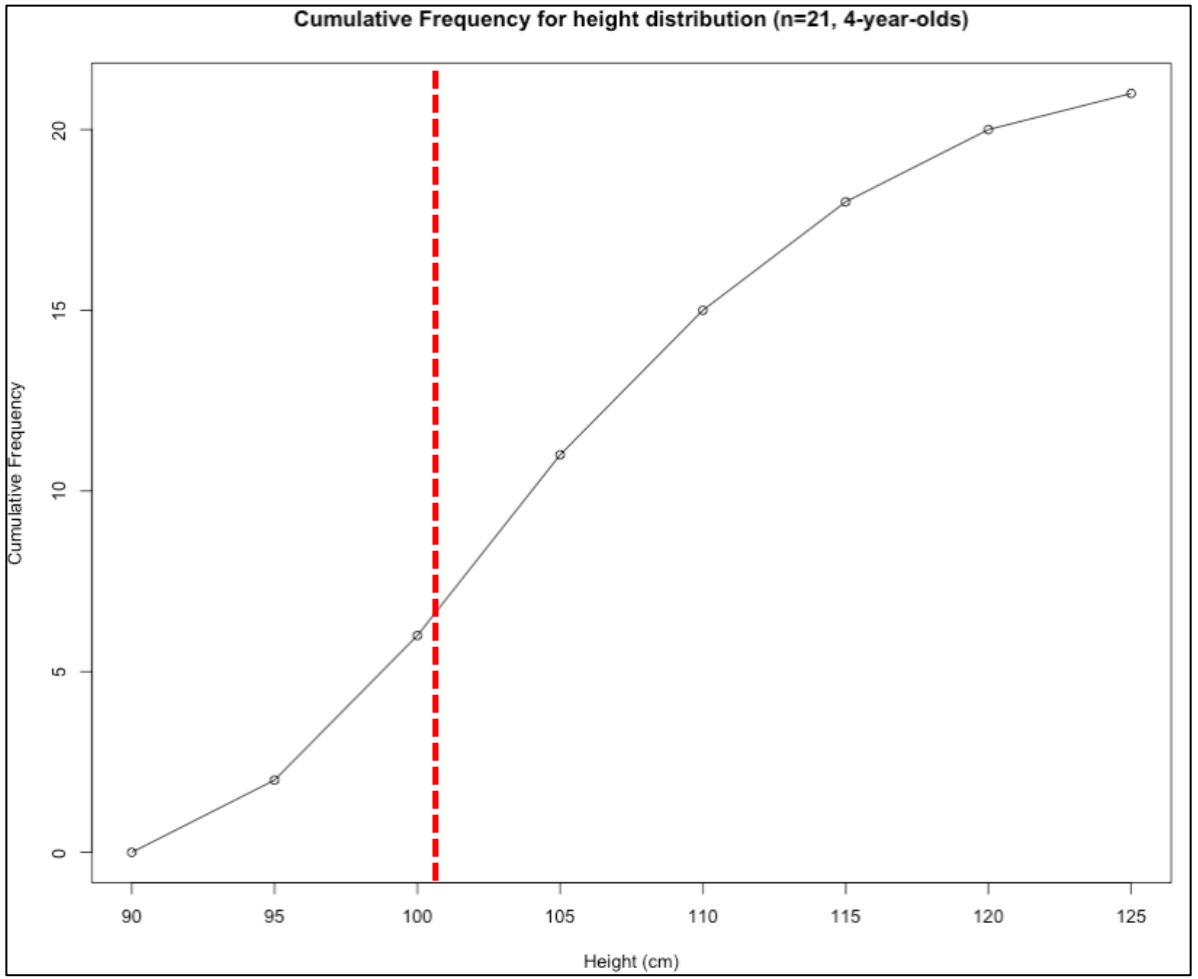
112, 113, 113

118, 119

121

This output is called a "**Frequency Distribution table**", it's visual representation is a **histogram** for the data's **relative frequency** and **cumulative frequency plot** for the **cumulative frequency**.





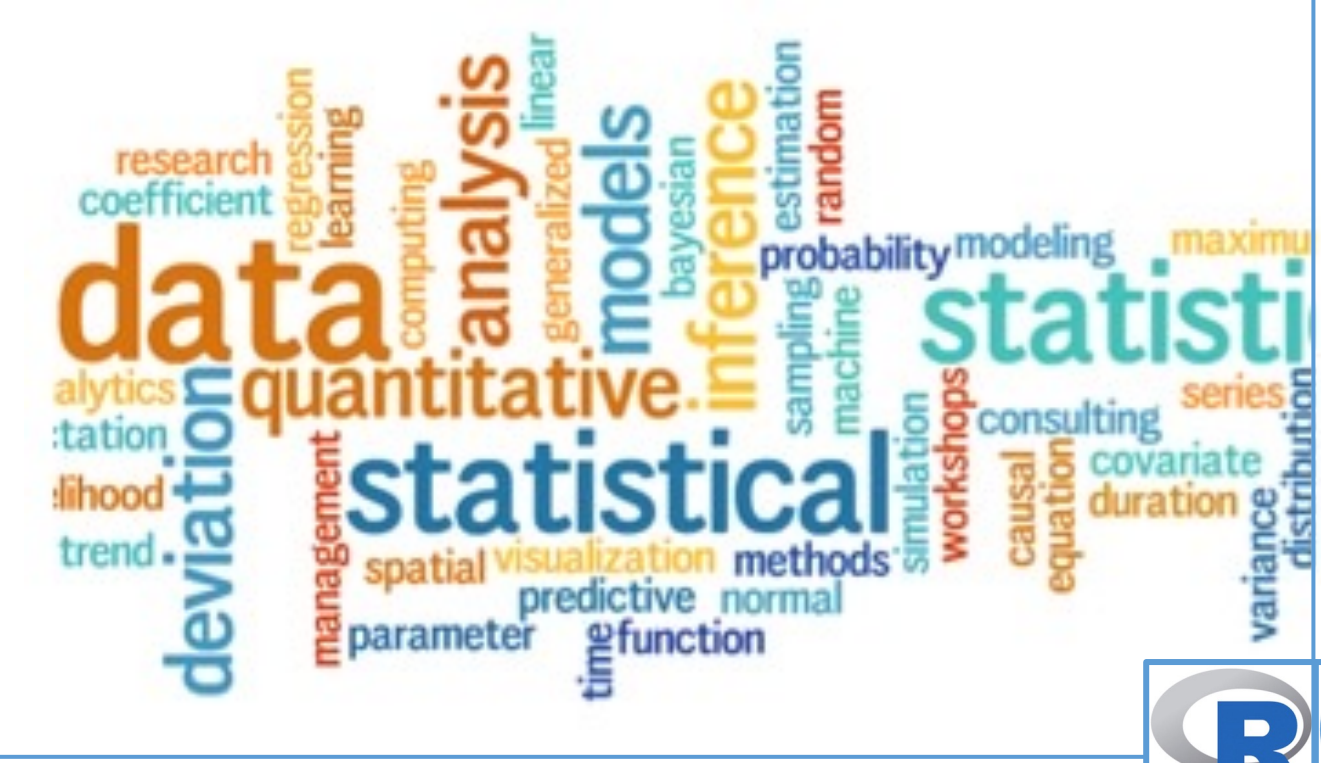
Interpretation: The above table output show the frequency distribution of heights (in cm) in kids who are 4 years of age entering in reception. The group with the highest frequency was 101-105 cm which accounts for 24% of the data. Health-wise, we can see from the cumulative frequency results that there are 6 kids with height values that are less than 101 cm. This corresponds to 0.2857 (29%) of the data – descriptively, these 6 kids growth is a cause for concern.

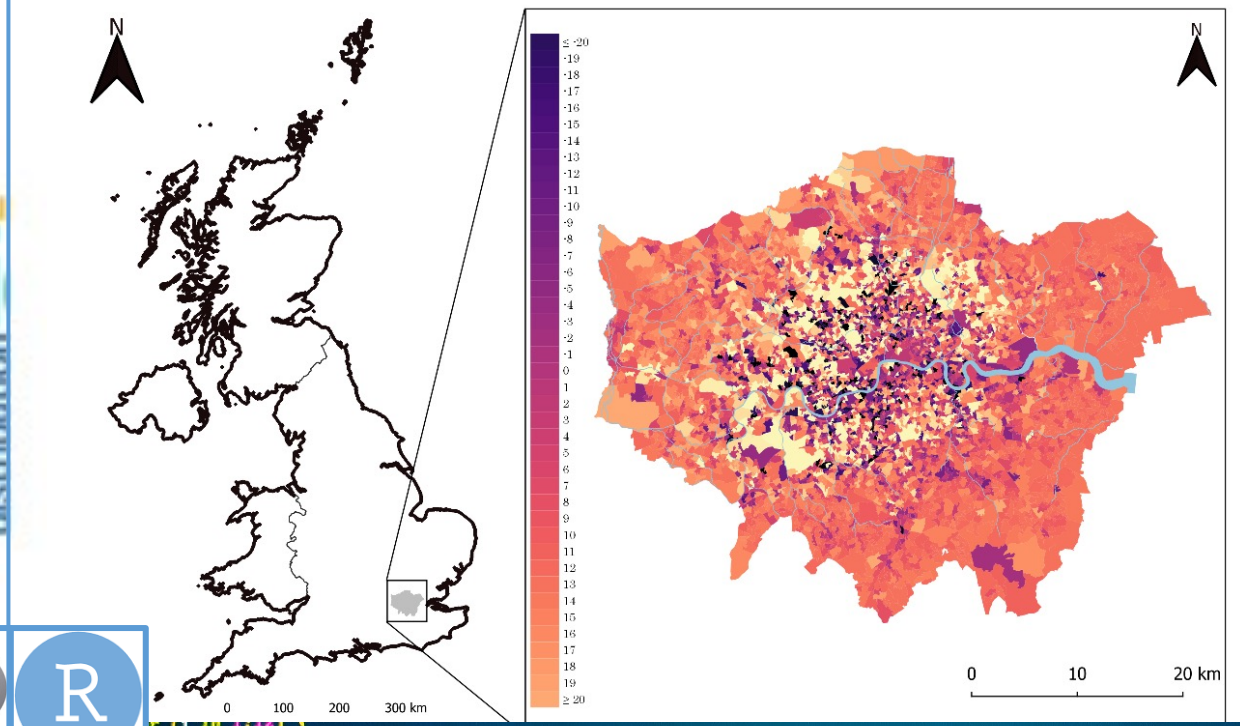
Break

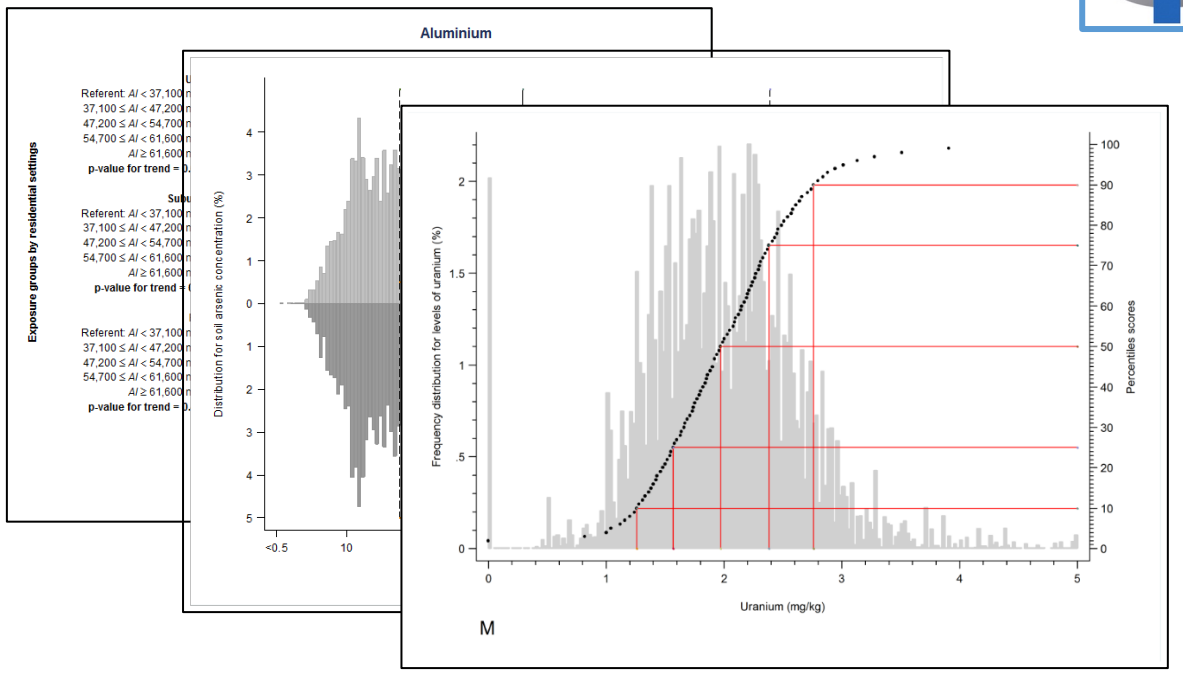
RStudio and live demonstration

What is R/RStudio

The r-project for statistical computing: https://www.r-project.org
Open-source RStudio: https://www.rstudio.com
The R Journal: https://journal.r-project.org







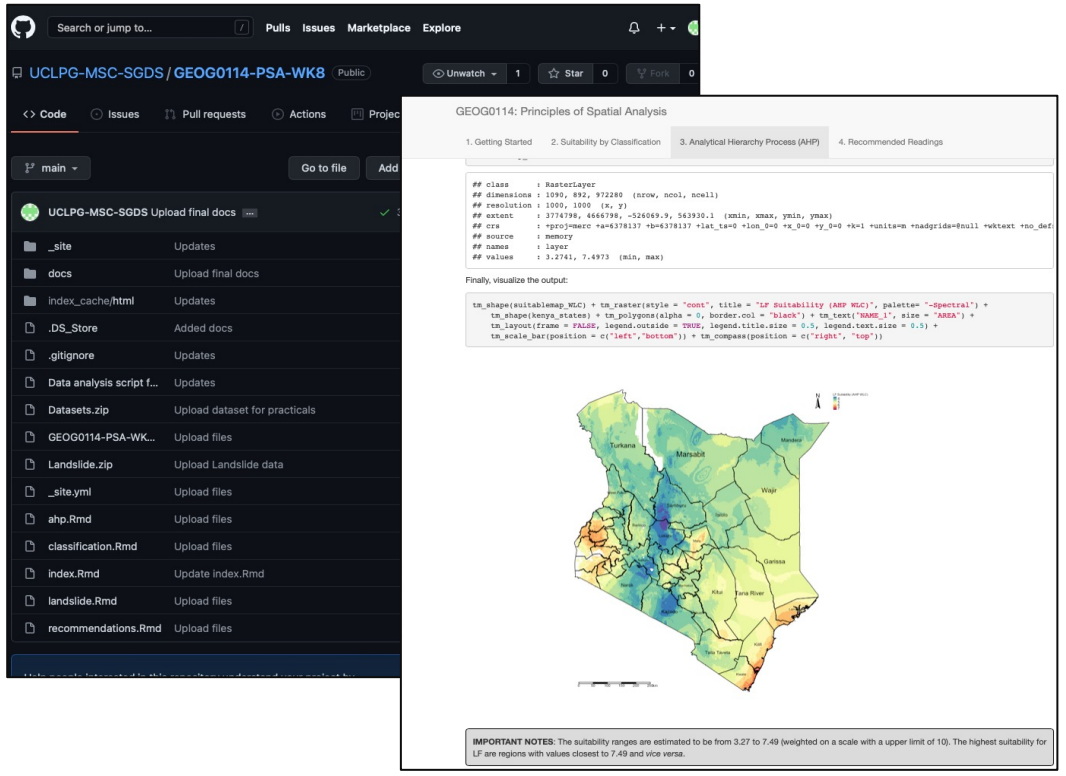

```
180     file <- paste0("/Users/anwarmusah/Desktop/AM_Zika2019/Data/Brazil/Climatic/Temperatu
181     raster_file <- raster(file)
182     recife_temperature_cropped <- crop(raster_file, recife_extent)
183     recife_temperature_masked <- mask(recife_temperature_cropped, bra_recife_outline)
184     recife_temperature_masked <- projectRaster(recife_temperature_masked, crs=pcs)
185     recife_temp_aggr <- extract(recife_temperature_masked, bra_recife_areas, fun=mean,
186     recife_temp_aggr$districtID<-bra_recife_areas$ID
187     colnames(recife_temp_aggr)[1] <- "fid"
188     colnames(recife_temp_aggr)[2] <- "temperature"
189     colnames(recife_temp_aggr)[3] <- "district_id"
190     recife_temp_aggr$year <- i
191     recife_temp_aggr$month <- j
192     recife_temperature <- recife_temp_aggr[,c(1,3,4,5,2)]
193   }
194   else {
195
196     file <- paste0("/Users/anwarmusah/Desktop/AM_Zika2019/Data/Brazil/Climatic/Temperatu
197     raster_file <- raster(file)
198     recife_temperature_cropped <- crop
199     recife_temperature_masked <-
```

Why learn how to code in RStudio?

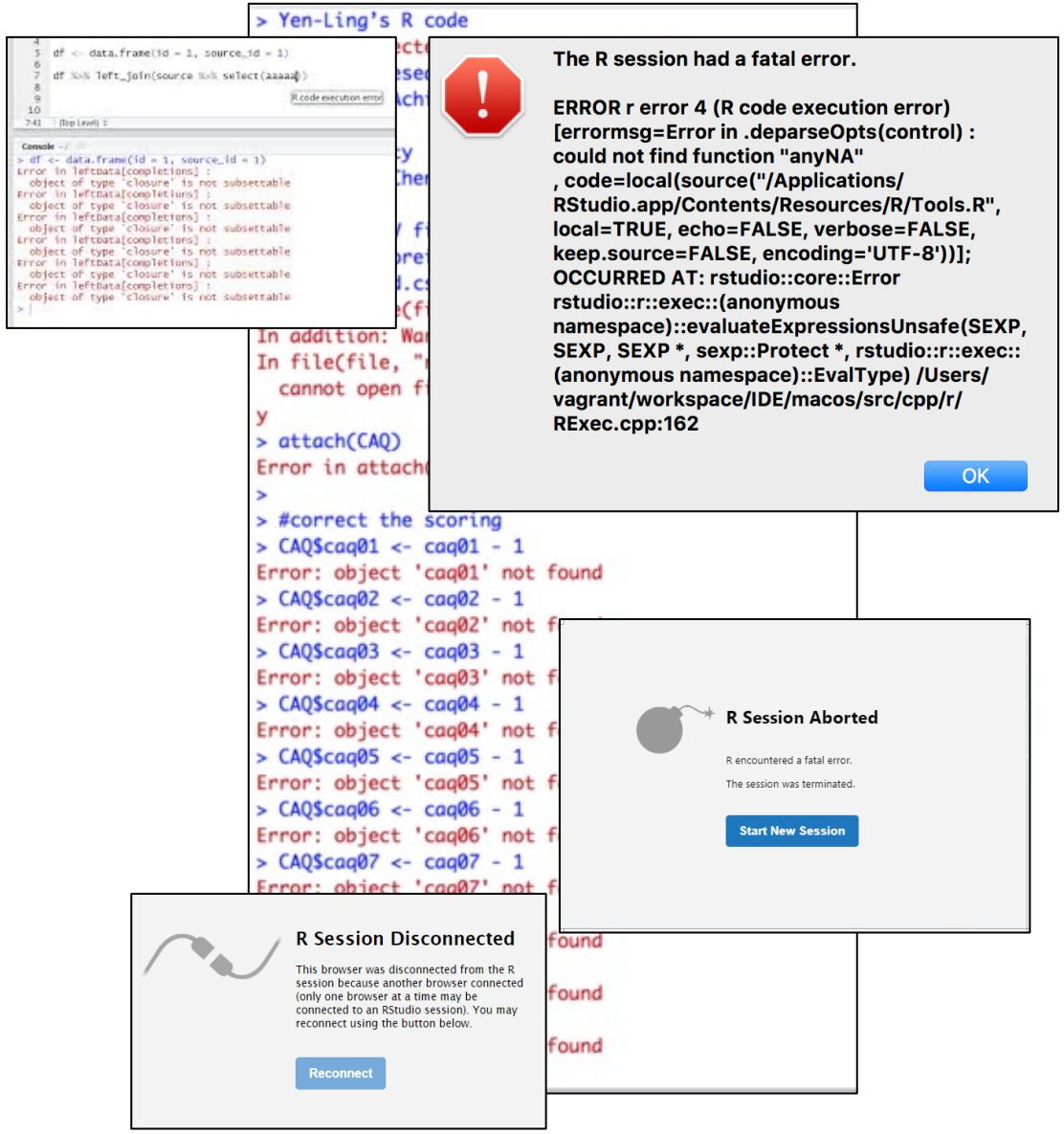
1. **Efficiency**
   - Automated tasks and data managing
   - Can recycle & reuse code scripts for new projects

1. **Fosters good scientific practice**
   - Transparency and replication (AKA reproducible research)
   - Creates log so anyone can follow in your footstep (i.e., github, gitlab etc.,)

2. **You can literally pull-off some really creative stuff like generating websites, accessing tools via APIs etc.**



Example: Working in RStudio and synchronising it with GitHub to not only use as a cloud back-up, but to generate a website through RStudio and GitHub for teaching MSc Students.


> Yen-Ling's R code
df <- data.frame(id = 1, source_id = 1)
df %>% left_join(source %>% select(aaaaa))
R code execution error
Console
> df <- data.frame(id = 1, source_id = 1)
Error in leftData[completions] :
object of type 'closure' is not subsettable
The R session had a fatal error.
ERROR r error 4 (R code execution error) [errormsg=Error in .deparseOpts(control) : could not find function "anyNA" , code=local(source("/Applications/RStudio.app/Contents/Resources/R/Tools.R", local=TRUE, echo=FALSE, verbose=FALSE, keep.source=FALSE, encoding='UTF-8'))]; OCCURRED AT: rstudio::core::Error rstudio::r::exec::(anonymous namespace)::evaluateExpressionsUnsafe(SEXP, SEXP, SEXP *, sexp::Protect *, rstudio::r::exec::(anonymous namespace)::EvalType) /Users/vagrant/workspace/IDE/macos/src/cpp/r/RExec.cpp:162
OK
In addition: Wa
In file(file, "
cannot open f
y
> attach(CAQ)
Error in attach
>
> #correct the scoring
> CAQ$caq01 <- caq01 - 1
Error: object 'caq01' not found
> CAQ$caq02 <- caq02 - 1
Error: object 'caq02' not f
> CAQ$caq03 <- caq03 - 1
Error: object 'caq03' not f
> CAQ$caq04 <- caq04 - 1
Error: object 'caq04' not f
> CAQ$caq05 <- caq05 - 1
Error: object 'caq05' not f
> CAQ$caq06 <- caq06 - 1
Error: object 'caq06' not f
> CAQ$caq07 <- caq07 - 1
Error: object 'caq07' not f
found
found
found
R Session Aborted
R encountered a fatal error.
The session was terminated.
Start New Session
R Session Disconnected
This browser was disconnected from the R session because another browser connected (only one browser at a time may be connected to an RStudio session). You may reconnect using the button below.
Reconnect

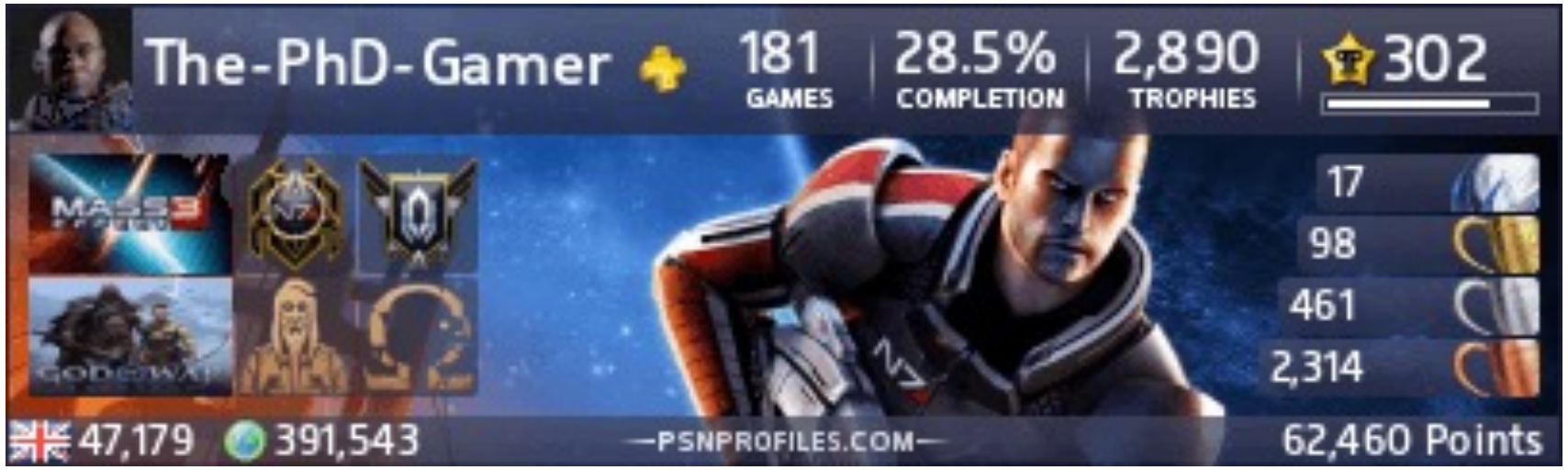
Live demonstration time – RStudio Server

RStudio Server: https://rstudio.data-science.rc.ucl.ac.uk/

To access RStudio Server with your personal computer, you will need the following conditions to hold:

- Must be connected to the UCL Network i.e., WiFi (EDUROAM) or be inside UCL remote environment.
- Must sign into RStudio Server with your UCL username and password with the above link.

A Video gamer's statistics



We have compiled the following information about the gaming habits of **Anwar Musah** (aka **The-PhD-Gamer**) across 3 console generations i.e., PlayStation 3, 4 and 5.

There are 161 game titles (~6,000 hours of game time) listed in the shared dataset (last updated 01/2024).

PSNProfiles (https://psnprofiles.com/)

Data Descriptor: [Downloadable Dataset]

Variables Names	Descriptor
Number	[Numeric] Unique Identifier
GameTitle	[String] Name of the video game
Genre	[String] Type of genre (9 categories)
Platform	[String] Type of console (3 categories)
HourPlayed	[Numeric] Total number of hours invested in a game
CompletionRate	[Numeric] Percentage of completed content in a game
Status	[String] Current status of the game in terms of play is 'in-progress' or 'quit', or in 'hiatus' or 'done' as in completed (4 categories)
PlatinumTrophy	[Binary] 1 = 'Attained platinum trophy' and 0 = 'No platinum trophy'
DLCTrophies	[Binary] 1 = 'Yes' and 0 = 'No'. Are there any DLC trophies present (annoying feature as it effects the completion rate)?

We are going to use this data as a test dummy. We are just going to access RStudio server and then show how to import this dataset.

Then set you folks going for the practical material to prepare for the seminars on Thursday.

Summary

- Importance of quantitative research methods (or statistics)
- We have acquainted ourselves with some basic statistical terminologies (i.e., population, samples, parameters and statistics)
- Types of variables & data types (i.e., continuous, categorical, dependent and independent variables)
- Descriptive statistics (frequency distributions)
- Introduced to RStudio

Any questions?