ENVS225 Assignment 1

Gabriele Filomena and Zi Ye

2025-09-29

Table of contents

Welcome

This is the website for "Exploring the Social World - Quantitative Block: Statistics" (module **ENVS225**) at the University of Liverpool. This block of the module is designed and delivered by Dr. Gabriele Filomena and Dr. Zi Ye from the Geographic Data Science Lab at the University of Liverpool. The module seeks to provide hands-on experience and training in introductory statistics for human geographers.

- As an html website.
- As a GitHub repository.

Contact

Gabriele Filomena - gfilo [at] liverpool.ac.uk Lecturer in Geographic Data Science Office 1xx, Roxby Building, University of Liverpool - 74 Bedford St S, Liverpool, L69 7ZT, United Kingdom.

Zi Ye - zi.ye [at] liverpool.ac.uk Lecturer in Geographic Information Science Office 107, Roxby Building, University of Liverpool - 74 Bedford St S, Liverpool, L69 7ZT, United Kingdom.

Overview

Aim and Learning Objectives

This sub-module aims to provide training and skills on a set of basic quantitative research methods for data collection, analysis, and interpretation. You will learn how to define coherent, relevant research questions, utilise various research quantitative methods, and identify appropriate methodologies to tackle your research questions. **This block serves as the foundation for the dissertation and fieldwork modules.**

Background

Data and research are key pillars of the global economy and society today. We need rigorous approaches to collecting and analysing both the statistics that can tell us 'how much' and if there are observable relationships between phenomena; and the information gives us a nuanced understanding of cultural contexts and human dynamics. Quantitative skills enable us to explore and measure socio-economic activities and processes at large scales, while qualitative skills enable understanding of social, cultural, and political contexts and diverse lived experiences. Rather than being in opposition, qualitative and quantitative research can complement one another in the investigation of today's pressing research questions.

To these ends, this block will help you develop your quantitative (statistical) skills, as critical tools. This course will help you understand what quantitative statistical researchers use and develop a set of research techniques that can be used in your field classes and dissertations.

Learning objectives:

- Understand how to explore a dataset, containing a number of observations described by a set of variables.
- Demonstrate an understanding in the application and interpretation of commonly used quantitative research methods.
- Demonstrate an understanding of how to work with quantitative data to address real-world research questions.

Module Structure

Staff: Dr Zi Ye and Dr Gabriele Filomena

Where and When

Week 1, 2, 4, 6 Tuesday @ Central Teaching Hub PCTC

- Lecture (10 – 11 am).
- Practical PC session (11 am – 1 pm).

Week 3

- Lecture (3 – 4 pm) Thursday @ Central Teaching Hub – Lect. Theatre C.
- Practical PC session (9 – 11 am) Friday @ Central Teaching Hub PCTC.

Week 5

- Lecture (3 – 4 pm) Thursday @ Central Teaching Hub – Lect. Theatre C.
- Practical PC session (10 – 12 am) Friday @ Central Teaching Hub PCTC.

Lectures will introduce and explain the fundamentals of quantitative methods, with the opportunity to apply the method introduced in the labs later in the week.

The computer practical sessions, will give you the chance to use and apply quantitative methods to real-world data. These are primarily self-directed sessions, but with support on hand if you get stuck. Support and training in R will be provided through these sessions. Weekly sessions will be driven by empirical research questions.

Week	Topic	Format	Staff
1	Introduction & Review	Lecture and Computer Lab Practical	GF
2	Single & Multiple Linear Regression	Lecture and Computer Lab Practical	GF
3	Multiple Linear Regression with Categorical Variables	Lecture and Computer Lab Practical	ZY
4	Logistic Regression	Lecture and Computer Lab Practical	ZY
5	Data Visualisation	Lecture and Computer Lab Practical	GF
6	Summary and Assessment Support	Lecture and Computer Lab Practical	ZY

Software and Data

For quantitative training sessions, ensure you have installed and/or have access to **RStudio**. To run the analysis and reproduce the code in R, you need the following software installed on your machine:

- R-4.2.2 (or later)
- RStudio 2022.12.0-353 (or later)

To install and update:

- R, download the appropriate version from The Comprehensive R Archive Network (CRAN).
- RStudio, download the appropriate version from here.

This software is already installed on University Machines. But you will need it to run the analysis on your personal devices.

Data

Example datasets could be accessed through Canvas or (some) on GitHub Repository of the module. These include:

- 2021 UK Census Data.
- 2021 Annulation Population Survey (APS) - only on Canvas.
- 2016 Family Resource Survey (FRS) - only on Canvas.
- 2011 Sample of Anonymised Records (SAR).

Note: The Annual Population Survey requires the completion of a quiz prior to its usage, as it is licensed.

Assessment

Deadline: Monday 3rd November 2025 - 2pm. **Word count**: 2000 words - including tables, excluding references.

The assignment **Data Exploration and Analysis** consists of writing a research report using one of the regression techniques learned during the module. The basic idea is to put in practice the methods learned during the quantitative block of the module. You are required to apply a linear or logistic regression model to the data provided for the module. The report needs to include the following sections (in brackets, % of the whole lenght):

- Introduction (5%).
- Literature Review (20%).
- Methods and data (30%).
- Results and discussion (40%).
- Conclusion (5%).
- Reference List.

Required Report Structure

1. **Introduction**
    - Context: Why is the topic relevant or worth being investigated?
    - Brief discussion of existing literature.
    - Knowledge gap and Aim.
    - 1 Research question.
2. **Literature review**
    - More detailed Literature review, i.e. what do we already know about this subject
    - Rationale for including certain predictor variables in the model.
    - What knowledge gap remains that this article will address? (includes "not studied before in this area"). *Note: there is no expectation on totally original research. The focus is on a clean, sensible, data analysis situated in existing ideas.*
3. **Methodology:**
    - A brief introduction to the dataset being analysed (who collected it? When? How many responses? etc.)

- A description of the variables chosen to be analysed.
- A description of any transformation made to the original data, i.e. turning a continuous variable of income into intervals, or reducing the number of age groups from 11 to 3.
- A description and justification of the statistical techniques used in the subsequent analysis (i.e. the Multivariate regression model: Multiple or Logistic Linear Regression).

4. **Results and Discussion**

   - Descriptive statistics and summary of the variables employed.
   - Correct interpretation of correlation coefficients.
   - Usage and results of an appropriate multivariate regression model.
   - Interpretation of the results, including links and contrasts to existing literature.
   - Selective illustrations (graphs and tables) to make your findings as clear as possible.

5. **Conclusion**

   - Summary of main findings.
   - Limitations of study (self-critique).
   - Highlight any implications derived from the study.

Follow this structure and include **ALL** these points, do not make your life harder.

How to get there?

The first stage is to identify **ONE** a relevant research question to be addressed. Based on the chosen question, you will need to identify a dependent (or outcome) variable which you want to explain, and at least two relevant independent variables that you can use to explain the chosen dependent variable. The selection of variables should be informed by the literature and empirical evidence.

To detail in the Methods Section: Once the variables have been chosen, you will need to describe the data and **appropriate** type of regression to be used for the analysis. You need to explain any transformation done to the original data source, such as reclassifying variables, or changing variables from continuous to nominal scales. You also need to briefly describe the data use: source of data, year of data collection, indicate the number of records used, state if you are using individual records or geographical units, explain if you are selecting a sample, and any relevant details. You also need to identify type of regression to be used and why.

To detail in the Results and Discussion Section: Firstly, you need to provide two types of analyses. First, you need to provide a descriptive analysis of the data. Here you could use tables and/or plots reporting relevant descriptive statistics, such as the mean, median and standard deviation; variable distributions using histograms; and relationships between

variables using correlation matrices or scatter plots. Secondly, you need to present an estimated regression model or models and the interpretation of the estimated coefficients. You need a careful and critical analysis of the regression estimates. You should think that you intend to use your regression models to advice your boss who is expecting to make some decisions based on the information you will provide. As part of this process, you need to discuss the model assessment results for the overall model and regression coefficients. Remember to substantiate your arguments using relevant literature and evidence, and present results clearly in tables and graphs.

How is it graded?

Grade	Score Range	UG	Descriptor	Assignment Expectations
Fail	0-34%	Fail	**Inadequate**	**Literature Review**: Lacks relevance and fails to justify variable choice. Evidence is irrelevant or missing, providing no support to the research question. **Methods**: Data is not described, and the regression model is entirely missing. No appropriate statistical method is applied. **Results and Discussion**: No descriptive statistics, graphs, or tables are provided. Model results and interpretation are absent. **Structure and References**: Report is disorganized with significant referencing and citation errors throughout.

Grade	Score Range	UG	Descriptor	Assignment Expectations
Narrow Fail	35-39%	Fail	**Highly Deficient**	**Literature Review**: Review is present but lacks coherence and fails to justify variable choice. Evidence is poorly aligned with the research question and mostly irrelevant. **Methods**: Minimal data description; the regression model is missing but some statistical methods are mentioned. **Results and Discussion**: Few or no descriptive statistics or visuals are present. Statistical methods are unclear or incorrectly applied. Results are vague and lack meaningful interpretation. **Structure and References**: Report structure is poor, with referencing errors in multiple sections.

Grade	Score Range	UG	Descriptor	Assignment Expectations
Third / Fail	40-49%	Third (UG)	**Deficient**	**Literature Review**: Relevant literature is partially addressed but lacks depth, with limited justification for variable choice. Evidence is minimally aligned with the research question. **Methods**: A very basic data description is provided, but the selected regression model is deeply inadequate or incorrect (e.g., multiple linear regression for a categorical outcome; logistic regression for a continuous outcome). **Results and Discussion**: Descriptive statistics or visuals may be present but insufficient. Model results are presented with little to no interpretation. **Structure and References**: Report structure is present but lacks clarity, with inconsistencies

Grade	Score Range	UG	Descriptor	Assignment Expectations
2.2 / Pass	50-59%	2.2 (UG)	**Adequate**	**Literature Review**: Addresses relevant literature but with limited justification of variable choices. Evidence generally supports the research question but lacks detail. **Methods**: Data description is present but brief; a regression model is included but applied illogically or incorrectly (e.g., multiple linear regression for a categorical outcome; logistic regression for a continuous outcome) and with little explanation. **Results and Discussion**: Basic descriptive statistics, graphs, or tables are presented; the regression model is applied with some inaccuracies and/or interpretation is minimal. **Structure and References**: Report is mostly organized

Grade	Score Range	UG	Descriptor	Assignment Expectations
2.1 / Merit	60-69%	2.1 (UG)	**Good**	**Literature Review**: Relevant literature is discussed, with some justification for variable choice. Evidence supports the research question well. **Methods**: Data is described with some detail, though potential data transformations are under-explored. The regression model is appropriate for the selected variable types. **Results and Discussion**: Descriptive statistics and visuals are provided. Model results are discussed, though interpretation lacks depth. Findings are compared to existing literature. **Structure and References**: Report is logically structured and clear, with mostly correct citations.

Grade	Score Range	UG	Descriptor	Assignment Expectations
First / Distinction	70-79%	First (UG)	**Very Good**	**Literature Review**: Strong grasp of relevant literature, with well-justified variable selection. Evidence aligns well with the research question. **Methods**: Data is comprehensively described with consideration of relevant transformations. The regression model is appropriate and well-justified. **Results and Discussion**: Descriptive statistics and clear visuals support findings. Model results are accurately interpreted with strong connections to existing literature. **Structure and References**: Report has a coherent, professional structure with only minor referencing errors.

Grade	Score Range	UG	Descriptor	Assignment Expectations
High First / High Distinction	80-100%	High First (UG)	**Excellent to Outstanding**	**Literature Review**: Critical and thorough literature review with strong, well-justified variable selection. Evidence fully supports the research question with insightful connections. **Methods**: Detailed data description and transformation steps are clearly articulated. Regression model is expertly applied and justified. **Results and Discussion**: Comprehensive descriptive statistics, graphs, and tables are provided. Model results are innovatively interpreted with strong links to existing research. **Structure and References**: Report is professionally structured, with flawless citations and a high standard of organization.

In summary:

1. **Introduction**: Should establish the topic's relevance, present a concise literature overview, identify a knowledge gap, and outline research questions.
2. **Literature Review**: Requires an in-depth review of relevant studies, justification for chosen independent variables, and identification of a potential knowledge gap or unexplored area aligned with the chosen research question.
3. **Methods and Data**: Should describe the dataset, variable transformations, and justify the regression technique. Key transformations, such as reclassifying variables, should be explained with clarity and relevance.
4. **Results and Discussion**: Involves presenting descriptive statistics, followed by a clear regression analysis. Discussion should interpret results, compare findings with existing literature, and include meaningful tables and graphs.
5. **Conclusion**: Summarize findings, discuss limitations, and suggest future directions.
6. **Referencing**: Requires correct and consistent citations and a well-structured reference list.

Employing a novel dataset, i.e. not employed during the practical sessions, for the assignment will be awarded with a higher grade. For example, see other quantitative dataset from Secondary datasets for Human Geography.

Assessment: How to submit

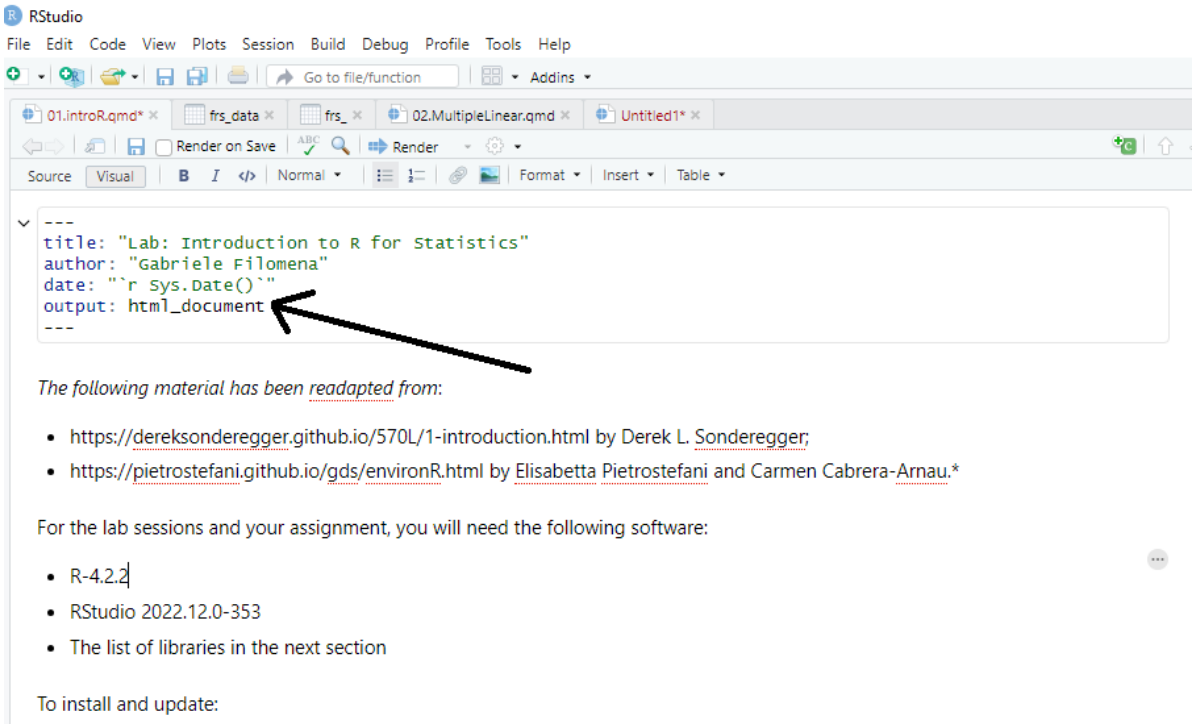
You must submit a `.pdf` file, that is a rendered version of a Quarto Markdown file (`qmd` file). This will allow you to write a research paper that also includes your working code, without the need of including the data (rendered `.qmd` files are executed before being converted to R).

How to get a PDF?

1. **Install Quarto**: Make sure you have Quarto installed. You can download it from quarto.org.
2. **LaTeX Installation**: For PDF output, you'll need a LaTeX distribution like **TinyTeX** from R, by executing this in the R console:

```
install.packages("tinytex")
tinytex::install_tinytex()
```

3. **Open the Quarto File**: Open your `.qmd` file in RStudio.
4. **Set Output Format**: In the YAML header at the top of your Quarto file, specify `pdf` under `format`:




```
title: "Your Document Title"
author: "Anonymous" # do not change
format: pdf
```

5. Click the **Render** button in the RStudio toolbar (next to the Knit button).

Assessment Template

Suggested Report Structure — This template follows the requested sections. Each section contains guidance text and an empty R code cell for your analysis. Download it from hereand work on it with your own Rstudio.

Setup

Load commonly used libraries for statistical analysis and data manipulation.

```
# Core tidyverse
library(tidyverse)    # includes ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, forcat

# Tables & reporting
library(knitr)
library(kableExtra)

# Stats helpers
library(broom)

## These need to be installed on your PC.
```

Once all the chunks execute correctly (no errors when running all the code, `Ctrl +Alt + R`) you can render the `qmd` file. Follow the instructions here to render the document to a PDF. Remove this message and any irrelevant text from the final submission.

Introduction

Aim. State clearly what you want to investigate (e.g., association between X and Y).
Relevance. Explain why the topic matters (theory, policy, practice).

Gap & Research Question (RQ). Identify what is missing in the current knowledge and pose 1 RQ (avoid yes/no questions; ask *how*, *to what extent*, *which factors*).
Structure. Briefly preview how the rest of the article is organized.

Literature Review

What we know. Summarise key findings from prior studies relevant to your RQ.
Predictor rationale. Justify the inclusion of variables (theory-driven, prior empirical evidence).
Remaining gap. Specify precisely the gap this report addresses (e.g., population, geography, time period, variable, method).
Note: The goal is a clean, sensible analysis grounded in existing ideas—not necessarily novel theory.

Methodology

Data. Briefly describe the dataset (who collected it, when, sample size, key variables).
Transformations. Describe any recoding/aggregation (e.g., bin income into bands; reduce age groups from 11 to 3). Provide justification.
Techniques. Outline and justify the statistical methods used (e.g., GLM/LM, logistic regression, mixed models, matching).

Important: do not include statistics here, or graphs, or tables.

Results and Discussion

```
# summary variables, distribution, plots, etc
```

Present the variables Start with descriptive statistics for your variables and distribution. If relevant, include discuss one-to-one associations (briefly, e.g. correlation plots). You can provide correlation plots and/or boxplots but do not overload the report with graphs.***

```
# run the model, show tables summarising the Regression model.
```

Results + interpretation. Present estimates and interpret them in plain language, tying back to the RQ. The Regression model should the core of this section!! **Link to literature.** Compare/contrast with prior findings. **Selective visuals.** Use clear charts/tables to highlight key results only (avoid clutter). ———————————————————————————

Conclusion

Summary. Recap the main findings vis-à-vis the RQs (do not introduce new results).
Limitations. Offer a brief, honest self-critique (data, measurement, design, external validity).
Implications. Indicate what the findings suggest for practice, policy, or future research.

Working Directories and Paths

Start clean (optional but handy)

```
# Clear the environment (objects in memory)
rm(list = ls())
```

Know where you are (working directory)

Your working directory (WD) is the folder R uses as the starting point for relative paths.

```
# Show current working directory
getwd()
```

```
[1] "C:/Users/gfilo/OneDrive - The University of Liverpool/Teaching/stats/general"
```

A relative path is specified from the working directory (e.g., `data/myfile.csv`).

An absolute path starts at the drive root (e.g., `C:/Users/you/Documents/stats/data/myfile.csv`).

Recommended folder layout (ENVS225)

Download the module materials, unzip them them wherever you prefer. Then un-nest the folder with the material (you will have one folder called `stats-main` containing another folder inside, with the same name. Aim for:

```
stats-main/
   data/
   labs_img/
   labs/
```

You can delete other sub-folders (e.g., docs) if you don't need them. It is strongly advised to move this folder into your M: drive on university machines, so it's accessible from every other PC on campus. Go to `This PC` and access `M:\(UoL userName)`

Set the working directory in RStudio (every time you use a new PC)

Windows: `RStudio > Tools > Global Options > General > Default working directory` > Browse to your `stats-main` folder.

macOS: `RStudio > Preferences > General`(older versions: under `Appearance/Pane Layout`) set Default working directory to `stats-main/`, this is the folder with all the materials inside.

Then restart RStudio and verify:

```
getwd()
```

```
[1] "C:/Users/gfilo/OneDrive - The University of Liverpool/Teaching/stats/general"
```

Tip: In addition always save your projects inside the folder `stats-main`. This keeps paths stable and prevents issues.

Loading data (CSV, Excel) with relative paths

Option a) Your `.qmd` project is in `stats-main`

```
library(readr)
library(readxl)
df <- read_csv("data/survey.csv")
xls <- read_excel("data/survey.xlsx", sheet = 1)
```

Option b) Your `.qmd` project is in `stats-main\subfolder`

```
df <- read_csv("../data/survey.csv")
xls <- read_excel("../data/survey.xlsx", sheet = 1)
```

In summary:

- `read_csv(MyFile.csv)` RStudio looks in the working directory for the file `MyFile.csv`.
- `read_csv(MyFolder/MyFile.csv)` RStudio looks inside the Working Directory (WD), for the folder `MyFolder`, then the file `MyFile.csv`.
- `read_csv(data/survey.csv)` RStudio looks inside the WD, looks inside `data`, then `survey.csv`.

- `read_csv(``../data/survey.csv` RStudio goes up one folder from the WD, then looks for the folder `data`and the file `surve.csv`.

You can always inspect where you are by

```
getwd()
```

```
[1] "C:/Users/gfilo/OneDrive - The University of Liverpool/Teaching/stats/general"
```

```
list.files()        # What's in the working directory?
```

```
 [1] "about.qmd"           "assessment.html"     "assessment.qmd"
 [4] "howSubmit.html"      "howSubmit.qmd"       "overview.html"
 [7] "overview.qmd"        "reportTemplate.html" "reportTemplate.qmd"
[10] "setup.qmd"           "wdPaths.html"        "wdPaths.qmd"
[13] "wdPaths.rmarkdown"   "wdPaths_files"
```

```
list.files("data")  # Do we see the data folder?
```

```
character(0)
```

1 Lab: Introduction to R

The following material has been readapted from:

- https://dereksonderegger.github.io/570L/1-introduction.html by Derek L. Sonderegger;
- For the lab sessions and your assignment, you will need the following software:
- R-4.2.2 (or higher)
- RStudio 2022.12.0-353 (or higher)
- The list of libraries in the next section

To install and update:

- R, download the appropriate version from The Comprehensive R Archive Network (CRAN)
- RStudio, download the appropriate version from Posit

1.1 R?

R is an open-source program that is commonly used in Statistics. It runs on almost every platform and is completely free and is available at www.r-project.org. Most of the cutting-edge statistical research is first available on R.

R is a script based language, so there is no point and click interface. While the initial learning curve will be steeper, understanding how to write scripts will be valuable because it leaves a clear description of what steps you performed in your data analysis. Typically you will want to write a script in a separate file and then run individual lines. This saves you from having to retype a bunch of commands and speeds up the debugging process.

1.2 R(Studio) Basics

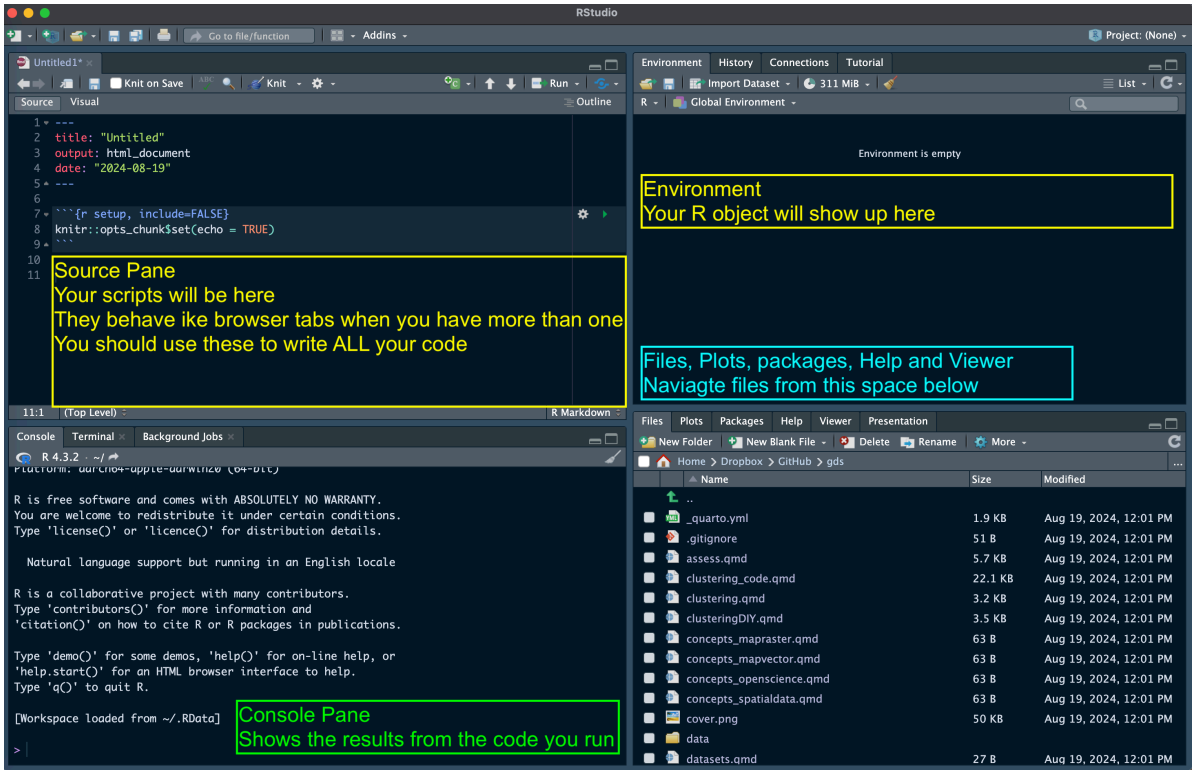
We will be running R through the program RStudio which is located at rstudio.com. When you first open up RStudio the console window gives you some information about the version of R you are running and then it gives the prompt `>`. This prompt is waiting for you to input a command. The prompt + tells you that the current command is spanning multiple lines. In a script file you might have typed something like this:

```
for( i in 1:5 ){
    print(i)
}
```

Finding help about a certain function is very easy. At the prompt, just type `help(function.name)` or `?function.name`. If you don't know the name of the function, your best bet is to go the the web page www.rseek.org which will search various R resources for your keyword(s). Another great resource is the coding question and answer site stackoverflow.

1.2.1 Starting a session in RStudio

Upon startup, RStudio will look something like this.



Note: the **Pane Layout** and **Appearance** settings can be altered:

- on Windows by clicking RStudio>Tools>Global Options>Appearance or Pane Layout
- on Mac OS by clicking RStudio>Preferences>Appearance or Pane Layout.

You will also have a standard white background; but you can choose specific themes.

Source Panel (Top-Left)

This is where you write, edit, and view scripts, R Markdown/Quarto documents, or R scripts. It allows:

- Editing Scripts: Write and edit R scripts or documents (`.R`, `.Rmd`, `.qmd`).
- Executing the Code: Run lines, blocks, or the entire script directly from the editor.

Console Panel (Bottom-Left)

The Console is the main place to run R commands interactively. It allows:

- Executing the Code: Type and run R commands directly.
- Viewing outputs, warnings, and errors for immediate feedback.
- Browsing and reusing past commands (History Tab).
- Toggling between the R Console, and the Terminal (yuo don't really need the latter).

Environment Panel (Top-Right)

This panel helps track variables, functions, and the history of commands used. It contains:

- Environment Tab: Shows all current variables, datasets, and objects in your session, including their structure and values.
- History Tab: Provides a record of past commands. You can re-run or move commands to the console or script.

Files / Plots / Packages / Help Panel (Bottom-Right)

This multifunctional panel is for file navigation, plotting, managing packages, viewing help, and managing jobs. It contains:

- Files Tab: Navigate, open, and manage files and directories within your project.
- Plots Tab: Displays plots generated in your session. You can export or navigate through multiple plots here.
- Packages Tab: Lists installed packages and allows you to install, load, and update packages.
- Help Tab: Displays help documentation for R functions, packages, and other resources. You can search for documentation by typing a function or package name.

At the start of a session, it's good practice clearing your R environment (console):

```
rm(list = ls())
```

In R, we are going to workin with **relative paths**. With the command `getwd()`, you can see where your working directory is currently set.

```
getwd()
```

1.2.2 Using the console

We can use the console to perform few operations. For example type in:

```
1+1
```

```
[1] 2
```

Slightly more complicated:

```
print("hello world")
```

```
[1] "hello world"
```

If you are unsure about what a command does, use the "Help" panel in your Files pane or type `?function` in the console. For example, to see how the `dplyr::rename()` function works, type in `?dplyr::rename`. When you see the double colon syntax like in the previous command, it's a call to a package without loading its library.

1.2.3 R as a simple calculator

You can use R as a simple calculator. At the prompt, type 2+3 and hit enter. What you should see is the following

```
# Some simple addition
2+3
```

```
[1] 5
```

In this fashion you can use R as a very capable calculator.

```r
6*8
```

```
[1] 48
```

```r
4^3
```

```
[1] 64
```

```r
exp(1)   # exp() is the exponential function
```

```
[1] 2.718282
```

R has most constants and common mathematical functions you could ever want. For example, the absolute value of a number is given by `abs()`, and `round()` will round a value to the nearest integer.

```r
pi     # the constant 3.14159265...
```

```
[1] 3.141593
```

```r
abs(1.77)
```

```
[1] 1.77
```

Whenever you call a function, there will be some arguments that are mandatory, and some that are optional and the arguments are separated by a comma. In the above statements the function `abs()` requires at least one argument, and that is the number you want the absolute value of.

When functions require more than one argument, arguments can be specified via the order in which they are passed or by naming the arguments. So for the `log()` function, for example, which calculates the logarithm of a number, one can specify the arguments using the named values; the order woudn't matter:

```r
# Demonstrating order does not matter if you specify
# which argument is which
log(x=5, base=10)
```

```
[1] 0.69897
```

```
log(base=10, x=5)
```

```
[1] 0.69897
```

When we don't specify which argument is which, R will decide that `x` is the first argument, and `base` is the second.

```
# If not specified, R will assume the second value is the base...
log(5, 10)
```

```
[1] 0.69897
```

```
log(10, 5)
```

```
[1] 1.430677
```

When we want to specify the arguments, we can do so using the `name=value` notation.

1.2.4 Variables Assignment

We need to be able to assign a value to a variable to be able to use it later. R does this by using an arrow `<-` or an equal sign `=`. While R supports either, for readability, I suggest people pick one assignment operator and stick with it.

Variable names cannot start with a number, include spaces, and they are case sensitive.

```
var <- 2*7.5       # create two variables
another_var = 5   # notice they show up in 'Environment' tab in RStudio!
var
```

```
[1] 15
```

```
var * another_var
```

```
[1] 75
```

As your analysis gets more complicated, you'll want to save the results to a variable so that you can access the results later. if you don't assign the result to a variable, you have no way of accessing the result.

1.2.5 Working with Scripts

Normally you would use the Console for quick calculations or executions. **In this module, though, we are going to work with Quarto Markdown Scripts (`.qmd` files).**

The R Markdown is an implementation of the Markdown syntax that makes it extremely easy to write webpages or scientific documents that include code. This syntax was extended to allow users to embed R code directly into more complex documents. Perhaps the easiest way to understand the syntax is to look at an at the RMarkdown website. The R code in a R Markdown document (`.rmd` file extension) can be nicely separated from regular text using the three backticks (3 times ‘, see below) and an instruction that it is R code that needs to be evaluated. A code chunk will look like:

```
    for (i in 1:5) {print(i)}
```

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
```

** `.qmd - the type of scripts we use -` are just a more flexible development of `.rmd` files.**

Markdown files present several advantages compared to writing your code in the console or just using scripts. You’ll save yourself a huge amount of work by embracing Markdown files from the beginning; you will keep track of your code and your steps, be able to document and present how you did your analysis (helpful when writing the methods section of a paper), and it will make it easier to re-run an analysis after a change in the data (such as additional data values, transformed data, or removal of outliers) or once you spot an error. Finally, it makes the script more readable.

1.2.6 R Packages

One of the greatest strengths about R is that so many people have developed add-on packages to do some additional function. To download and install the package from the Comprehensive R Archive Network (CRAN), you just need to ask RStudio it to install it via the menu `Tools -> Install Packages...`. Once there, you just need to give the name of the package and RStudio will download and install the package on your computer.

Once a package is downloaded and installed on your computer, it is available, but it is not loaded into your current R session by default. To improve overall performance only a few

packages are loaded by default and the you must explicitly load packages whenever you want to use them. You only need to load them once per session/script.

```
library(dplyr)   # load the dplyr library, will be useful later
```

This is just a quick intro to R, now move to the actual practical of week 1.

2 Lab: Exploring a Dataset

The following material has been readapted from:

- https://pietrostefani.github.io/gds/environR.html by Elisabetta Pietrostefani and Carmen Cabrera-Arnau
- https://raw.githubusercontent.com/dereksonderegger/570L/master/07_DataImport.Rmd

The lecture's slides can be found here.

Before completing the practical, please take this quiz.

For the lab sessions and your assignment, you will need the following software:

- R-4.2.2 (or higher) The Comprehensive R Archive Network (CRAN)
- RStudio 2022.12.0-353 (or higher) Posit

IMPORTANT: Before starting: verify that you have set the directory correctly and familiarise yourself with working with directories and loading files. These are the instructions.

Now create a .qmd project. We use the `File -> New File -> Quarto Document..` dropdown option; a menu will appear asking you for the document title, author, and preferred output type. You can select HTML. Call it "*week1*" and save it in your main folder.

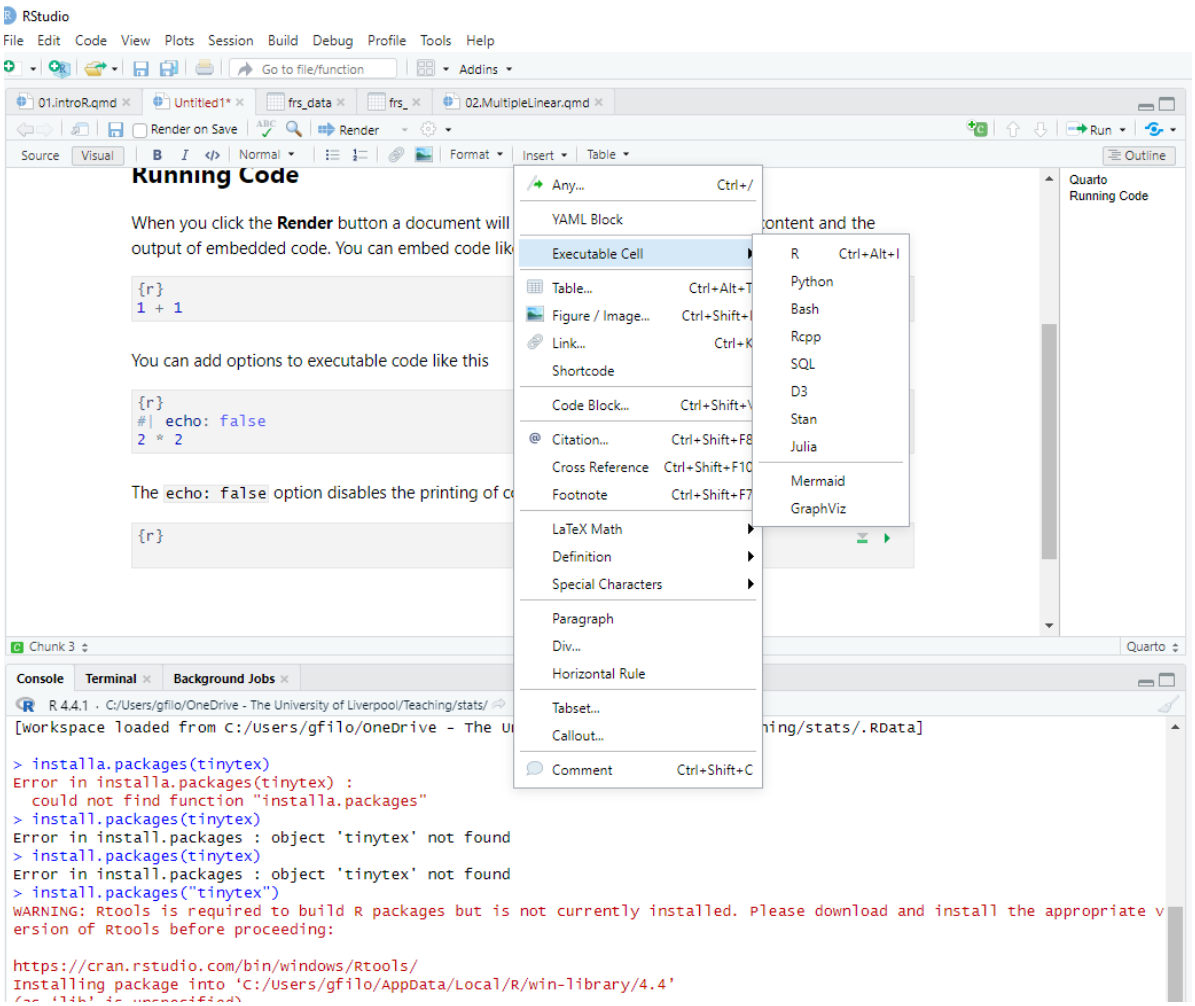
Follow the practical below. You can describe what you are doing in normal text. See here for how to format normal text in Markdown documents

Remember, when you want to write code in a markdown document you have to enclose it like this:

```{r}

```

or you can insert it manually:



2.1 Practice: Dataset and Dataframes

Within this module we will be working with data stored in so-called datasets. A dataset is a structured collection of data points that represent various measurements or observations, often organized in a tabular format with rows and columns. A dataset might contain information about different locations, such as neighborhoods or cities, with each row representing a place and each column detailing characteristics like population density, average income, or number of green parks. For example, a dataset could be compiled to study patterns in urban mobility, where the data includes the number of daily commuters, the distance they travel, and the mode of transport they use. Datasets provide the essential building blocks for statistical analysis; they enable exploring relationships, identifying patterns, and drawing conclusions about certain phenomena.

Examples of everyday datasets:

- **Premier League Standings**: Each row represents a team, with columns for points, games played, wins, draws, and losses.
- **Movie Dataset**: Each row represents a movie, with columns showing its title, genre, release year, director, and rating.
- **Weather Dataset**: Each row shows a day's weather in a city, with columns for temperature, humidity, wind speed, and precipitation.

Usually, data is organized in

- **Columns** of data representing some trait or variable that we might be interested in. In general, we might wish to investigate the relationship between variables.
- **Rows** represent a single object on which the column traits are measured.

For example, in a grade book for recording students scores throughout the semester, their is one row for every student and columns for each assignment. A greenhouse experiment dataset will have a row for every plant and columns for treatment type and biomass.

2.1.1 Datasets in R

In R, we want a way of storing data where it feels just as if we had an Excel Spreadsheet where each row represents an observation and each column represents some information about that observation. We will call this object a `data.frame`, an R represention of a data set. The easiest way to understand data frames is to create one.

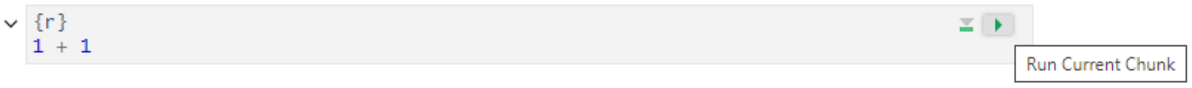
Task: Copy the code below in your markdown. Create a `data.frame` that represents an instructor's grade book, where each row is a student, and each column represents some sort of assessment.

```
library(dplyr)

Grades <- data.frame(
  Name  = c('Bob','Jeff','Mary','Valerie'),
  Exam.1 = c(90, 75, 92, 85),
  Exam.2 = c(87, 71, 95, 81)
)
# Show the data.frame
# View(Grades)  # show the data in an Excel-like tab.  Doesn't work when knitting
Grades          # show the output in the console. This works when knitting
```

```
     Name Exam.1 Exam.2
1     Bob     90     87
2    Jeff     75     71
3    Mary     92     95
4 Valerie     85     81
```

To execute just one chunk of code press the green arrow top-right of the chunk:



R allows two differnt was to access elements of the `data.frame`. First is a matrix-like notation for accessing particular values.

Format	Result
`[a,b]`	Element in row `a` and column `b`
`[a,]`	All of row `a`
`[,b]`	All of column `b`

Because the columns have meaning and we have given them column names, it is desirable to want to access an element by the name of the column as opposed to the column number.

Task: Copy and Run:

```
Grades[, 2]       # print out all of column 2
```

```
[1] 90 75 92 85
```

```
Grades$Name       # The $-sign means to reference a column by its label
```

```
[1] "Bob"     "Jeff"    "Mary"    "Valerie"
```

2.1.2 Importing Data in R

Usually we won't type the data in by hand, but rather load the data from some package. Reading data from external sources is a necessary skill.

***Comma Separated Values* Data**

To consider how data might be stored, we first consider the simplest file format: the comma separated values file (`.csv`). In this file type, each of the "cells" of data are separated by a comma. For example, the data file storing scores for three students might be as follows:

```
Able, Dave, 98, 92, 94
Bowles, Jason, 85, 89, 91
Carr, Jasmine, 81, 96, 97
```

Typically when you open up such a file on a computer with MS Excel installed, Excel will open up the file assuming it is a spreadsheet and put each element in its own cell. However, you can also open the file using a more primitive program (say Notepad in Windows, TextEdit on a Mac) you'll see the raw form of the data.

Having just the raw data without any sort of column header is problematic (which of the three exams was the final??). Ideally we would have column headers that store the name of the column.

```
LastName, FirstName, Exam1, Exam2, FinalExam
Able, Dave, 98, 92, 94
Bowles, Jason, 85, 89, 91
Carr, Jasmine, 81, 96, 97
```

Reading (`.csv`) files

To make R read in the data arranged in this format, we need to tell R three things:

1. Where does the data live? Often this will be the name of a file on your computer, but the file could just as easily live on the internet (provided your computer has internet access).
2. Is the first row data or is it the column names?
3. What character separates the data? Some programs store data using tabs to distinguish between elements, some others use white space. R's mechanism for reading in data is flexible enough to allow you to specify what the separator is.

The primary function that we'll use to read data from a file and into R is the function `read.csv()`. This function has many optional arguments but the most commonly used ones are outlined in the table below.

Argument	Default	Description
`file`	Required	A character string denoting the file location.
`header`	`TRUE`	Specifies whether the first line contains column headers.
`sep`	`","`	Specifies the character that separates columns. For `read.csv()`, this is usually a comma.

Argument	Default	Description
`skip`	`0`	The number of lines to skip before reading data; useful for files with descriptive text before the actual data.
`na.strings`	`"NA"`	Values that represent missing data; multiple values can be specified, e.g., `c("NA", "-9999")`.
`quote`	`"`	Specifies the character used to quote character strings, typically `"` or `'`.
`stringsAsFactors`	`FALSE`	Controls whether character strings are converted to factors; `FALSE` means they remain as character data.
`row.names`	`NULL`	Allows specifying a column as row names, or assigning `NULL` to use default indexing for rows.
`colClasses`	`NULL`	Specifies the data type for each column to speed up reading for large files, e.g., `c("character", "numeric")`.
`encoding`	`"unknown"`	Sets the text encoding of the file, which can be useful for files with special or international characters.

Most of the time you just need to specify the file. |

Task: Let's read in a dataset of terrorist attacks that have taken place in the UK:

```
attacks <- read.csv(file   = '../data/attacksUK.csv')  # where the data lives
View(attacks)
```

2.2 Practice: Descriptive Statistics

2.2.1 Summarizing Data

It is very important to be able to take a data set and produce summary statistics such as the mean and standard deviation of a column. For this sort of manipulation, we use the package `dplyr`. This package allows chaining together many common actions to form a particular task.

The fundamental operations to perform on a data set are:

- **Subsetting** - Returns a dataframe with only particular columns or rows
  - `select` - Selecting a subset of columns by name or column number.
  - `filter` - Selecting a subset of rows from a data frame based on logical expressions.
  - `slice` - Selecting a subset of rows by row number.
- `arrange` - Re-ordering the rows of a data frame.
- `mutate` - Add a new column that is some function of other columns.
- `summarise` - calculate some summary statistic of a column of data. This collapses a set of rows into a single row.

Each of these operations is a function in the package `dplyr`. These functions all have a similar calling syntax:

- The first argument is a data set.
- Subsequent arguments describe what to do with the input data frame and you can refer to the columns without using the `df$column` notation.

All of these functions will return a data set.

Let's consider the `summarize` function to calculate the mean score for `Exam.1`. Notice that this takes a data frame of four rows, and summarizes it down to just one row that represents the summarized data for all four students.

```
library(dplyr) # load the library
Grades %>%
  summarize( Exam.1.mean = mean( Exam.1 ) )
```

```
  Exam.1.mean
1        85.5
```

Similarly you could calculate the **standard deviation** for the exam as well.

```
Grades %>%
  summarize( Exam.1.mean = mean( Exam.1 ),
             Exam.1.sd   = sd(   Exam.1   ) )
```

```
  Exam.1.mean Exam.1.sd
1        85.5  7.593857
```

Task: Write the code above in your markdown file and run it. Do not to copy it this time.

The `%>%` operator works by translating the command `a %>% f(b)` to the expression `f(a,b)`. This operator works on any function `f`. This is useful when we want to start with `x`, and first apply a function `f()`, then `g()`, and then `h()`; the usual R command would be `h(g(f(x)))` which is hard to read. Using the pipe command `%>%`, this sequence of operations becomes `x %>% f() %>% g() %>% h()`.

Below, the code takes the Grades dataframe and calculates a column for the average exam score, and then sorts the data according to the that average score

```
Grades %>%   mutate( Avg.Score = (Exam.1 + Exam.2) / 2 ) %>%   arrange( Avg.Score )
```

```
     Name Exam.1 Exam.2 Avg.Score
1    Jeff     75     71      73.0
2 Valerie     85     81      83.0
3     Bob     90     87      88.5
4    Mary     92     95      93.5
```

You don't have to memorise this.

Let's go back to the terrorist attacks dataframe. There are attacks perpetrated by several different groups. Each record is a single attack and contains information about who perpetrated the attack, what year, how many were killed and how many were wounded. You can get a glimpse of the dataframe with the function head

```
head(attacks, n = 10)
```

```
   nrKilled nrWound year        country                         group
1         0       0 2005 United Kingdom   Abu Hafs al-Masri Brigades
2         0       0 2005 United Kingdom   Abu Hafs al-Masri Brigades
3         0       0 2005 United Kingdom   Abu Hafs al-Masri Brigades
4         0       0 2005 United Kingdom   Abu Hafs al-Masri Brigades
5         0       1 1982 United Kingdom Abu Nidal Organization (ANO)
6         0       0 2014 United Kingdom                  Anarchists
7         0       0 2014 United Kingdom                  Anarchists
8         0       0 2014 United Kingdom                  Anarchists
9         0       0 2014 United Kingdom                  Anarchists
10        0       0 2014 United Kingdom                  Anarchists
                          attack                  target
1              Bombing/Explosion             Transportation
2              Bombing/Explosion             Transportation
```

```
3               Bombing/Explosion              Transportation
4               Bombing/Explosion              Transportation
5                   Assassination     Government (Diplomatic)
6  Facility/Infrastructure Attack                    Business
7  Facility/Infrastructure Attack                    Business
8  Facility/Infrastructure Attack                    Business
9  Facility/Infrastructure Attack Private Citizens & Property
10 Facility/Infrastructure Attack                      Police
                      weapon
1  Explosives/Bombs/Dynamite
2  Explosives/Bombs/Dynamite
3  Explosives/Bombs/Dynamite
4  Explosives/Bombs/Dynamite
5                    Firearms
6                  Incendiary
7                  Incendiary
8                  Incendiary
9                  Incendiary
10                 Incendiary
```

We might want to compare different actors and see the mean and standard deviation of the number of people wound, by each group's attack, across time. To do this, we are still going to use the `summarize`, but we will precede that with `group_by(group)` to tell the subsequent `dplyr` functions to perform the actions separately for each breed.

```
attacks %>%
  group_by( group) %>%
  summarise( Mean = mean(attacks$nrWound),
             Std.Dev = sd(attacks$nrWound))
```

```
# A tibble: 38 x 3
   group                                                Mean Std.Dev
   <chr>                                               <dbl>   <dbl>
 1 Abu Hafs al-Masri Brigades                          0.963    7.22
 2 Abu Nidal Organization (ANO)                        0.963    7.22
 3 Anarchists                                          0.963    7.22
 4 Animal Liberation Front (ALF)                       0.963    7.22
 5 Animal Rights Activists                             0.963    7.22
 6 Armenian Secret Army for the Liberation of Armenia 0.963    7.22
 7 Black September                                     0.963    7.22
 8 Continuity Irish Republican Army (CIRA)             0.963    7.22
 9 Dissident Republicans                               0.963    7.22
```

```
10 Informal Anarchist Federation                    0.963    7.22
# i 28 more rows
```

Task: Write the code above in your markdown file and run it. Try out another categorical variable instead of `group` (e.g. `year`) and `nrKilled` instead of `nrWound`.

Let's now move to another dataset to address a research question.

For illustration purposes, we will use the **Family Resources Survey (FRS)**. The FRS is an annual survey conducted by the UK government that collects detailed information about the income, living conditions, and resources of private households across the United Kingdom. Managed by the Department for Work and Pensions (DWP), the FRS provides data that is essential for understanding the economic and social conditions of households and informing public policy.

Consider questions such as:

- How many respondents (persons) are there in the 2016-17 FRS?
- How many variables (population attributes) are there?
- What types of variables are present in the FRS?
- What is the most detailed geography available in the FRS?

Task: To answer these questions, download from Canvas, save in the data folder, load and inspect the dataset.

```r
# the FRS dataset should be already loaded, otherwise
frs_data <- read.csv("../data/FRS/FRS16-17_labels.csv")

# Display basic structure
glimpse(frs_data)
```

```
Rows: 44,145
Columns: 45
$ household        <int> 6087, 6101, 6103, 6122, 6134, 6136, 6138, 6140, 6143,~
$ family           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ person           <int> 5, 3, 3, 3, 2, 4, 4, 3, 3, 4, 4, 3, 4, 2, 5, 3, 4, 3,~
$ country          <chr> "England", "England", "England", "Northern Ireland", ~
$ region           <chr> "London", "South East", "Yorks and the Humber", "Nort~
$ age_group        <chr> "05-10", "05-10", "05-10", "05-10", "05-10", "05-10",~
$ sex              <chr> "Female", "Male", "Male", "Female", "Female", "Female~
$ marital_status   <chr> "Single", "Single", "Single", "Single", "Single", "Si~
$ ethnicity        <chr> "Mixed / multiple ethnic groups", "White", "White", "~
$ hrp              <chr> "Not HRP", "Not HRP", "Not HRP", "Not HRP", "Not HRP"~
$ rel_to_hrp       <chr> "Son/daughter (incl. adopted)", "Son/daughter (incl. ~
```

```
$ lifestage        <chr> "Child (0-17)", "Child (0-17)", "Child (0-17)", "Chil~
$ dependent        <chr> "Dependent", "Dependent", "Dependent", "Dependent", "~
$ arrival_year     <chr> "UK Born", "UK Born", "UK Born", "UK Born", "UK Born"~
$ birth_country    <chr> "Dependent child", "Dependent child", "Dependent chil~
$ care_hours       <chr> "0 hours per week", "0 hours per week", "0 hours per ~
$ educ_age         <chr> "Dependent child", "Dependent child", "Dependent chil~
$ educ_type        <chr> "School (full-time)", "School (full-time)", "School (~
$ fam_youngest     <chr> "7", "4", "0", "7", "0", "9", "10", "0", "3", "10", "~
$ fam_toddlers     <int> 0, 1, 1, 0, 2, 0, 0, 2, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1,~
$ fam_size         <int> 4, 4, 4, 3, 4, 4, 3, 5, 4, 4, 3, 4, 4, 3, 5, 4, 4, 4,~
$ happy            <chr> "Dependent child", "Dependent child", "Dependent chil~
$ health           <chr> "Not known", "Not known", "Not known", "Not known", "~
$ hh_accom_type    <chr> "Terraced house/bungalow", "Detached house/bungalow",~
$ hh_benefits      <int> 10868, 0, 1768, 8632, 8372, 1768, 1768, 1768, 0, 0, 1~
$ hh_composition   <chr> "Three or more adults, 1+ children", "One adult femal~
$ hh_ctax_band     <chr> "Band D", "Band F", "Band A", "Band B", "Band A", "Ba~
$ hh_housing_costs <chr> "4316", "10296", "5408", "Northern Ireland", "5720", ~
$ hh_income_gross  <int> 54236, 180804, 26936, 19968, 17992, 76596, 31564, 366~
$ hh_income_net    <int> 44668, 120640, 23556, 19968, 17992, 62868, 29744, 287~
$ hh_size          <int> 5, 4, 4, 3, 4, 4, 4, 5, 4, 4, 4, 4, 4, 5, 5, 4, 4, 4,~
$ hh_tenure        <chr> "Mortgaged (including part rent / part own)", "Mortga~
$ highest_qual     <chr> "Dependent child", "Dependent child", "Dependent chil~
$ income_gross     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ income_net       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ jobs             <chr> "Dependent child", "Dependent child", "Dependent chil~
$ life_satisf      <chr> "Dependent child", "Dependent child", "Dependent chil~
$ nssec            <chr> "Dependent child", "Dependent child", "Dependent chil~
$ sic_chapter      <chr> "Dependent child", "Dependent child", "Dependent chil~
$ sic_division     <chr> "Dependent child", "Dependent child", "Dependent chil~
$ soc2010          <chr> "Dependent child", "Dependent child", "Dependent chil~
$ work_hours       <chr> "Dependent child", "Dependent child", "Dependent chil~
$ workstatus       <chr> "Dependent Child", "Dependent Child", "Dependent Chil~
$ years_ft_work    <chr> "Dependent child", "Dependent child", "Dependent chil~
$ survey_weight    <int> 2315, 1317, 2449, 427, 1017, 1753, 1363, 1344, 828, 1~
```

and summary:

```
summary(frs_data)
```

```
   household         family          person       country
 Min.   :    1   Min.   :1.000   Min.   :1.00   Length:44145
 1st Qu.: 4816   1st Qu.:1.000   1st Qu.:1.00   Class :character
```

```
Median : 9673   Median :1.000   Median :2.00   Mode  :character
Mean   : 9677   Mean   :1.106   Mean   :1.98
3rd Qu.:14553   3rd Qu.:1.000   3rd Qu.:3.00
Max.   :19380   Max.   :6.000   Max.   :9.00
   region           age_group            sex            marital_status
Length:44145       Length:44145       Length:44145       Length:44145
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


 ethnicity            hrp             rel_to_hrp         lifestage
Length:44145       Length:44145       Length:44145       Length:44145
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


 dependent          arrival_year       birth_country       care_hours
Length:44145       Length:44145       Length:44145       Length:44145
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


  educ_age           educ_type          fam_youngest        fam_toddlers
Length:44145       Length:44145       Length:44145       Min.   :0.0000
Class :character   Class :character   Class :character   1st Qu.:0.0000
Mode  :character   Mode  :character   Mode  :character   Median :0.0000
                                                         Mean   :0.2557
                                                         3rd Qu.:0.0000
                                                         Max.   :4.0000
   fam_size         happy              health          hh_accom_type
Min.   :1.000   Length:44145       Length:44145       Length:44145
1st Qu.:2.000   Class :character   Class :character   Class :character
Median :2.000   Mode  :character   Mode  :character   Mode  :character
Mean   :2.599
3rd Qu.:4.000
Max.   :9.000
 hh_benefits    hh_composition     hh_ctax_band       hh_housing_costs
Min.   :    0   Length:44145       Length:44145       Length:44145
1st Qu.:    0   Class :character   Class :character   Class :character
Median : 1768   Mode  :character   Mode  :character   Mode  :character
```

```
Mean   : 5670
3rd Qu.:10192
Max.   :54080
hh_income_gross   hh_income_net        hh_size      hh_tenure
Min.   :-326092   Min.   :-334776   Min.   :1.00   Length:44145
1st Qu.:  22256   1st Qu.:  20748   1st Qu.:2.00   Class :character
Median :  35984   Median :  31512   Median :3.00   Mode  :character
Mean   :  46076   Mean   :  37447   Mean   :2.96
3rd Qu.:  57252   3rd Qu.:  47008   3rd Qu.:4.00
Max.   :1165216   Max.   :1116596   Max.   :9.00
highest_qual        income_gross       income_net          jobs
Length:44145       Min.   :-354848   Min.   :-358592   Length:44145
Class :character   1st Qu.:     52   1st Qu.:      0   Class :character
Mode  :character   Median :  12740   Median :  12012   Mode  :character
                   Mean   :  17305   Mean   :  14204
                   3rd Qu.:  23712   3rd Qu.:  20384
                   Max.   :1127360   Max.   :1110928
life_satisf          nssec           sic_chapter        sic_division
Length:44145       Length:44145       Length:44145       Length:44145
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


  soc2010           work_hours         workstatus        years_ft_work
Length:44145       Length:44145       Length:44145       Length:44145
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


survey_weight
Min.   :  221
1st Qu.: 1097
Median : 1380
Mean   : 1459
3rd Qu.: 1742
Max.   :39675
```

2.2.2 Understanding the Structure of the FRS Datafile

In the FRS data structure, each row represents a person, but:

- Each person is nested within a family.
- Each family is nested within a household.

Below is an example dataset structure:

household	family	person	region	age_group	sex	marital_status	rel_to_hrp
1	1	1	London	40-44	Female	Married/Civil partnership	Spouse
1	1	2	London	40-44	Male	Married/Civil partnership	Household Representative
1	1	3	London	5-10	Male	Single	Son/daughter (incl. adopted)
1	1	4	London	5-10	Female	Single	Son/daughter (incl. adopted)
1	1	5	London	16-19	Male	Single	Step-son/daughter
2	1	1	Scotland	35-39	Male	Single	Household Representative
3	1	1	Yorks and the Humber	35-39	Female	Married/Civil partnership	Household Representative
3	1	2	Yorks and the Humber	35-39	Male	Married/Civil partnership	Spouse
3	1	3	Yorks and the Humber	5-10	Male	Single	Step-son/daughter
4	1	1	Wales	0-4	Male	Single	Son/daughter (incl. adopted)
4	1	2	Wales	60-64	Male	Married/Civil partnership	Household Representative
4	1	3	Wales	55-59	Female	Married/Civil partnership	Spouse
4	2	3	Wales	30-34	Female	Single	Son/daughter (incl. adopted)

The first five people in the FRS all belong to the same household (household 1); they also

all belong to the same family. This family comprises a married middle-aged couple plus their three children, one of whom is a stepson.

The second household (household 2) comprises only one person – a single middle-aged male.The third household comprises another married couple, this time with two children.

Superficially the fourth household looks similar to households 1 and 2: a married couple plus their daughter. The difference is that this particular married couple is nearing retirement age, and their daughter is middle-aged. Consequently, despite being a child of the married couple, the middle-aged daughter is treated as a separate 'family' (family 2 in the household). This is because the FRS (and Census) define a 'family' as a couple plus any 'dependent' children. A dependent child is defined as a child who is either' aged 0-15 or aged 16-19, unmarried and in full-time education. All children aged 16-19 who are married or no longer in full-time education are regarded as 'independent' adults who form their own family unit, as are all children aged 20+.

The inclusion of all persons in a household allows us more flexibility in the types of research question we can answer. For example, we could explore how the likelihood of a woman being in paid employment `WorkStatus` is influenced by the age of the youngest child still living in her family (if any) `fam_youngest`.

In the FRS (and Census), a "family" is defined as a couple and any "dependent" children. Dependent children are defined as those aged 0–15, or aged 16–19 if unmarried and in full-time education.

2.2.3 Explore the Distribution of Your Outcome Variable

Before starting your analysis, it is critical to know the type of scale used to measure your outcome variable: is it categorical or continuous? Here we will start off by exploring a continuous variable which can then turn into a categorical variable (e.g. top earners: yes or no). We explore the income distribution in the UK by first looking at the low and high end of the distribution ie. What sorts of people have high (or low) incomes?

In the FRS each person's annual income is recorded, both gross (pre-tax) and net (post-tax). This income includes all income sources, including earnings, profits, investment returns, state benefits, occupational pensions etc. As it is possible to make a loss on some of these activities, it is also possible (although unusual) for someone's gross or net annual income in a given year to be negative (representing an overall loss).

Task: Load the FRS dataset into your R environment, if it's not already loaded, and inspect the data.

Open the dataset in RStudio's **Data Viewer** to explore its structure, including the `income_gross` and `income_net` variables.

```
 # Open the data in the RStudio Viewer
View(frs_data)
```

in the **Data Viewer** tab, scroll horizontally to locate the `income_gross` and `income_net` columns. If columns are listed alphabetically, they will appear near other attributes that start with "income."

You should notice two things:

- Incomes are recorded to the nearest £, NOT in income bands.
- Dependent children almost all have a recorded income of £0.

This second observation highlights the somewhat loose wording of our question above (*What sorts of people have high (or low) incomes?*). To avoid reaching the somewhat banal conclusion that those with the lowest of all incomes are almost all children, we should re-frame the question more precisely as *What sorts of people (excluding dependent children) have low incomes?*

Task: Determine the Scale of the Outcome Variable.

```
# Summarize income variables
summary(frs_data$income_gross)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-354848      52   12740   17305   23712 1127360 
```

```
# Summarize income variables
summary(frs_data$income_net)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-358592       0   12012   14204   20384 1110928 
```

Task: Exclude Dependent Children.

You need to select all cases (persons) that are independent, that is where the variable `dependent` has values different from `!=` "Dependent" or equal `==` "Independent".

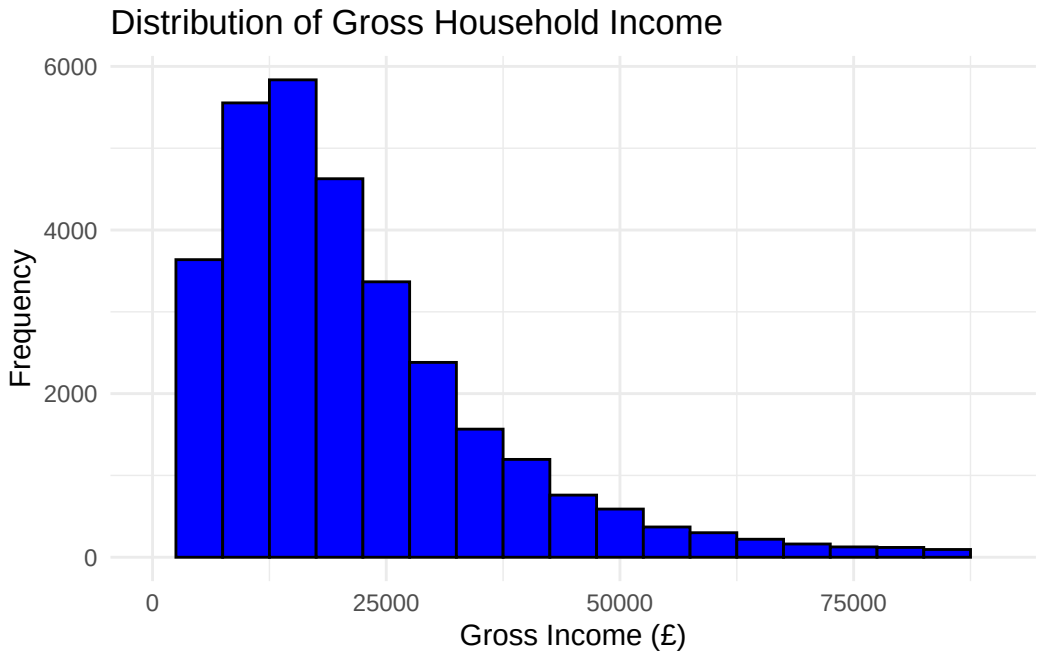
```
# Filter to include only independent persons
frs_independent <- frs_data %>% filter(dependent != "Dependent")
```

Task: Create a basic histogram (a visualisation lecture is scheduled later on).

The income variables in the FRS are all scale variables so a good starting point is to examine its distribution looking at a histogram of `income_gross`.

```r
library(ggplot2)

 ggplot(frs_independent, aes(x = income_gross)) +
   geom_histogram(binwidth = 5000, fill = "blue", color = "black") +
   labs(
     title = "Distribution of Gross Household Income",
     x = "Gross Income (£)",
     y = "Frequency"
   ) +
   xlim(0, 90000) +
   theme_minimal()
```



You should see the histogram below. It reveals that the income distribution is very skewed with few people earning high salaries and the majority earning just over or less 35,000 annually.

Task: Adopt a regrouping strategy.

You can also cross-tabulate gross (or net) income with any of the other variables in the FRS to your heart's content – or can you?

Again, here is important to recall that the income variables in the FRS are all 'scale' variables; in other words, they are precise measures rather than broad categories. Consequently, every single person in the FRS potentially has their own unique income value. That could make for a table c. 44,000 rows long (one row per person) if each person has their own unique value.

The solution is to create a categorical version of the original income variable by assigning each person to one of a set of income categories (income bands). Having done this, cross-tabulation then becomes possible.

But which strategy to use? Equal intervals, percentiles or 'ad hoc'. Here I would suggest that 'ad hoc' is best: all you want to do is to allocate each independent adult to one of three arbitrarily defined groups: 'low', 'middle' and 'high' income. **Define `Low` and `High` Income Thresholds**

Define thresholds for income categories:

- Low-income threshold: £_____
- High-income threshold: £_____

Task: Create a New Variable Based on Regrouping of Original Variable.

Recode `income_gross` into categories based on the chosen thresholds.

```
# Define thresholds for income categories
LOW_THRESHOLD <- 10000 # Replace with the upper limit for low income
HIGH_THRESHOLD <- 50000 # Replace with the lower limit for high income

# Define income categories based on thresholds
frs_independent <- frs_independent %>%
    mutate(income_category = case_when(
        income_gross <= LOW_THRESHOLD ~ "Low",
        income_gross >= HIGH_THRESHOLD ~ "High",
        TRUE ~ "Middle" ))
```

The `mutate()` function in R, from the **dplyr** package, is used to add or modify columns in a data frame. It allows you to create new variables or transform existing ones by applying calculations or conditional statements directly within the function.

Explanation of the code

- **`frs_independent %>%`**: The pipe operator `%>%` sends `frs_independent` into `mutate()`, allowing us to apply transformations without reassigning it repeatedly.
- **`mutate()`**: Starts the transformation process by defining new or modified columns.
- **`income_category = case_when(...)`**:
  - This creates a new column named `income_category`.
  - The `case_when()` function defines conditions for assigning values to this new column.
- **`case_when()`**:
  - `case_when()` is used here to assign categorical labels based on conditions.

- `income_gross <= LOW_THRESHOLD ~ "Low"`: If `income_gross` is less than or equal to `LOW_THRESHOLD`, `income_category` will be labeled "Low."
- `income_gross >= HIGH_THRESHOLD ~ "High"`: If `income_gross` is greater than or equal to `HIGH_THRESHOLD`, `income_category` will be labeled "High."
- `TRUE ~ "Middle"`: Any values not meeting the previous conditions are labeled "Middle."

Task: Add some Metadata.

Define metadata for the new variable by labeling income categories.

```
# Add metadata by converting to a factor and defining labels

frs_independent$income_category <- factor(frs_independent$income_category,
                          levels = c("Low", "Middle", "High"), labels = c("<= £10,000", "£10,0
```

Task: Check your work.

Examine the frequency distribution of the variable you have just created. Both variables should have the same number of missing cases, unless:

- Missing cases in the old variable have been intentionally converted into valid cases in the new variable.
- You forgot to allocate a new value to one of the old variable categories, in which case the new variable will have more missing cases than the old variable.

```
# Frequency distribution of income categories
table(frs_independent$income_category)
```

```
<= £10,000 £10,001 - £49,999        >= £50,000
      8584             22981              2271
```

After preparing the data, use cross-tabulations to compare income levels across demographic groups.

```
# Cross-tabulate income category by age group, nationality, etc.
table(frs_independent$income_category, frs_independent$age_group)
```

```
                   16-19 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64
  <= £10,000         373   680   492   558   474   511   554   652   781   826
  £10,001 - £49,999  263  1241  1802  2056  2052  1948  1995  1967  1749  1772
  >= £50,000           1     8    59   186   314   331   334   356   237   177

                   65-69 70-74  75+
  <= £10,000         773   744 1166
  £10,001 - £49,999 2073  1554 2509
  >= £50,000         144    56   68
```

Explore income distribution across different regions.

```
# Cross-tabulate income category by region
table(frs_independent$income_category, frs_independent$region)
```

```
                   East Midlands East of England London North East North West
  <= £10,000                 562             665    740        357        878
  £10,001 - £49,999         1550            1855   1850        979       2347
  >= £50,000                 135             245    367         48        174

                   Northern Ireland Scotland South East South West Wales
  <= £10,000                    874     1212        895        588   399
  £10,001 - £49,999            2305     3234       2563       1707   971
  >= £50,000                    123      322        367        149    63

                   West Midlands Yorks and the Humber
  <= £10,000                 744                  670
  £10,001 - £49,999         1892                 1728
  >= £50,000                 164                  114
```

Tips for Cross-Tabulation

- Place the income variable in the columns.
- Add multiple variables in the rows to create simultaneous cross-tabulations.

3 Lab: Correlation, Single, and Multiple Linear Regression

IMPORTANT: Before starting: verify that you have set the directory correctly and familiarise yourself with working with directories and loading files. These are the instructions.

Now create a .qmd project. We use the `File -> New File -> Quarto Document..` dropdown option; a menu will appear asking you for the document title, author, and preferred output type. You can select HTML. Call it "*week2*" and save it in your main folder.

Follow the practical below. You can describe what you are doing in normal text. See here for how to format normal text in Markdown documents

In this week's practical, we will review how to calculate and visualise correlation coefficients between variables. This practical is split into two parts. The first part focuses on measuring and visualising the relationship between between **continuous variables**. The second part goes through the implementation of a Linear Regression Model, again between continuous variables.

Before getting into it, have a look at this resource, it really helps understand how regression models work.

The lecture's slides can be found here.

Before completing the practical, please take this quiz.

Learning Objectives:

- Visualise the association between two continuous variables using a scatterplot.
- Measure the strength of the association between two variables by calculating their correlation coefficient.
- Build a formal regression model.
- Understand how to estimate and interpret a multiple linear regression model.

Note on file paths: When calling the `read.csv` function, the path will vary depending on the location of the script it is being executed or your working directory (WD):

1. **If the script or your WD is in a sub-folder** (e.g., `labs`), use `"../data/Census2021/EW_DistrictPerc`
   The `..` tells R to go one level up to the main directory (`stats/`) and then access the `data/` folder. **Example**: `read.csv("../data/Census2021/EW_DistrictPercentages.csv")`.
2. **If the script is in the main directory** (e.g. inside `stats/`), you can access the data directly using `"data/Census2021/EW_DistrictPercentages.csv"`. Here, no `..` is necessary as the `data/` folder is directly accessible from the working directory. **Example**: `read.csv("data/Census2021/EW_DistrictPercentages.csv")`

3.1 Part I. Correlation

3.1.1 Data Overview: Descriptive Statistics:

Let's start by picking **one dataset derived from the English-Wales 2021 Census data**. You can choose one dataset that aggregates data either at a) county, b) district, or c) ward-level. Lower Tier Local Authority-, Region-, and Country-level data is also available in the data folder.

see also: https://canvas.liverpool.ac.uk/courses/77895/pages/census-data-2021

```
# Load necessary libraries
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
options(scipen = 999, digits = 4)  # Avoid scientific notation and round to 4 decimals globa

# load data - you main need to remove "../"
census <- read.csv("../data/Census2021/EW_DistrictPercentages.csv") # District level
```

We're using a (district/ward/etc.) level census dataset that includes:

- % of population with poor health (variable name: `pct_Very_bad_health`).
- % of population with no qualifications (`pct_No_qualifications`).
- % of male population (`pct_Males`).
- % of population in a higher managerial/professional occupation (`pct_Higher_manager_prof`).

First, let's get some descriptive statistics that help identify general trends and distributions in the data.

```r
# Summary statistics
summary_data <- census %>%
  select(pct_Very_bad_health, pct_No_qualifications, pct_Males, pct_Higher_manager_prof) %>%
  summarise_all(list(mean = mean, sd = sd))
summary_data
```

```
  pct_Very_bad_health_mean pct_No_qualifications_mean pct_Males_mean
1                    1.173                      17.9          48.97
  pct_Higher_manager_prof_mean pct_Very_bad_health_sd pct_No_qualifications_sd
1                        13.22                 0.3401                    3.959
  pct_Males_sd pct_Higher_manager_prof_sd
1       0.6605                     4.73
```

Q1. Complete the table below by specifying each variable type (continuous or categorical) and reporting its mean and standard deviation.

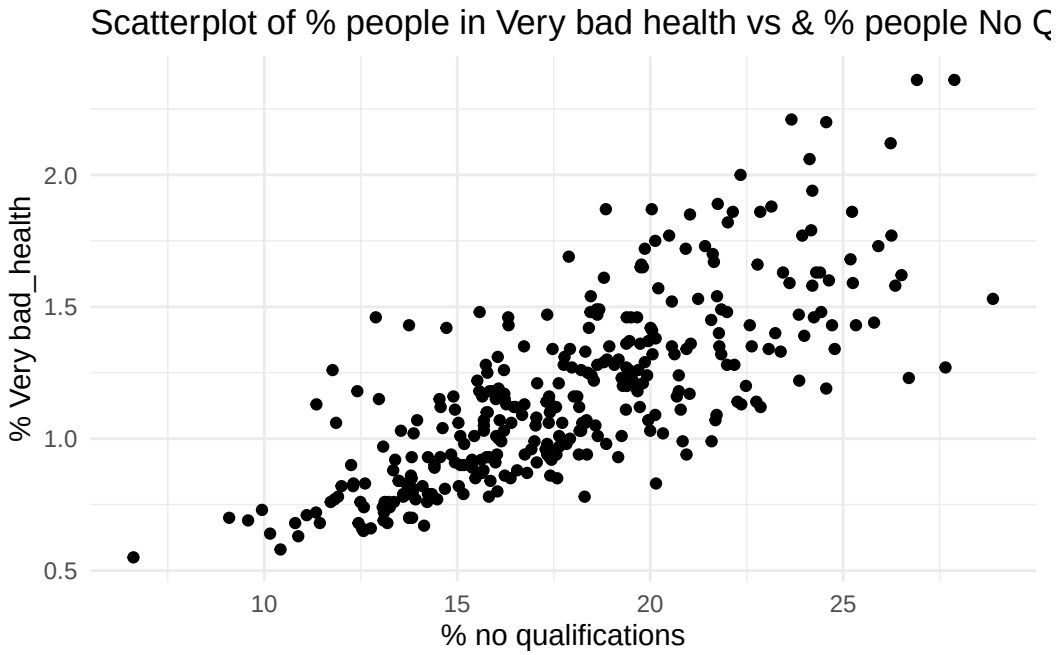
Variable Name	Type (Continuous or Categorical)	Mean	Standard Deviation
pct_Very_bad_health			
pct_No_qualifications			
pct_Males			
pct_Higher_manager_prof			

3.1.2 Simple visualisation for continuous data

You can visualise the relationship between two continuous variables using a scatter plot. Using the chosen census datasets, visualise the association between the % of population with bad health (`pct_Very_bad_health`) and each of the following:

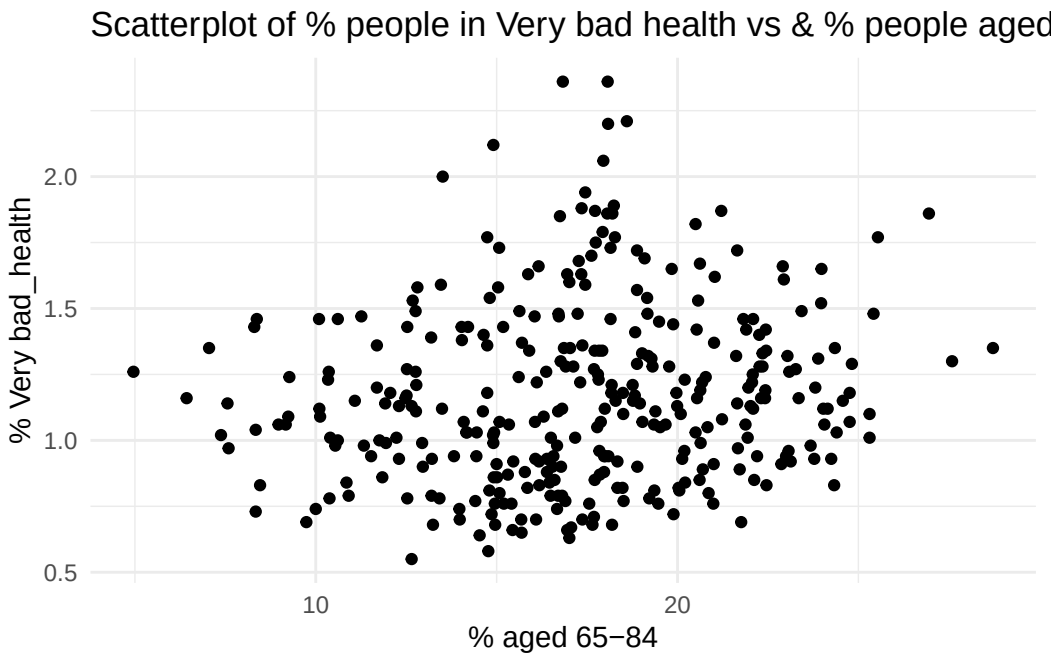
- the % of population with no qualifications (`pct_No_qualifications`);
- the % of population aged 65 to 84 (`pct_Age_65_to_84`);
- the % of population in a married couple (`pct_Married_opposite_sex_couple`);
- the % of population in a Higher Managerial or Professional occupation (`pct_Higher_manager_prof`).

```
# 1
ggplot(census, aes(x = pct_No_qualifications, y = pct_Very_bad_health)) +
    geom_point() +
    labs(title = paste("Scatterplot of % people in Very bad health vs & % people", "No Quali
         x = "% no qualifications", y = "% Very bad_health") +
    theme_minimal()
```

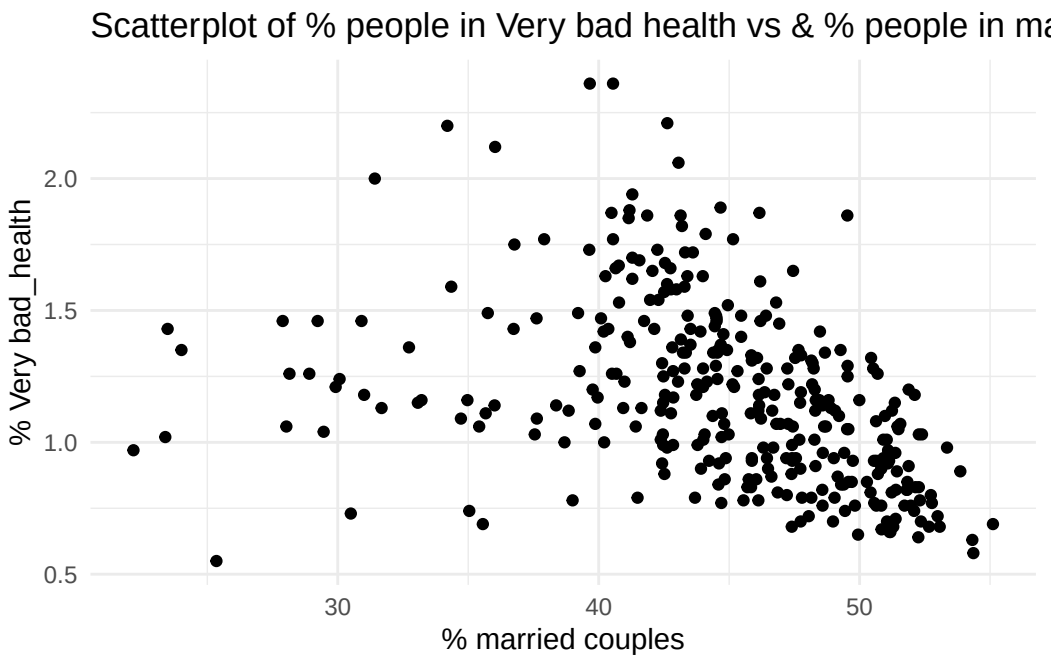



```
# 2
ggplot(census, aes(x = pct_Age_65_to_84, y = pct_Very_bad_health)) +
    geom_point() +
    labs(title = paste("Scatterplot of % people in Very bad health vs & % people", "aged betw
```

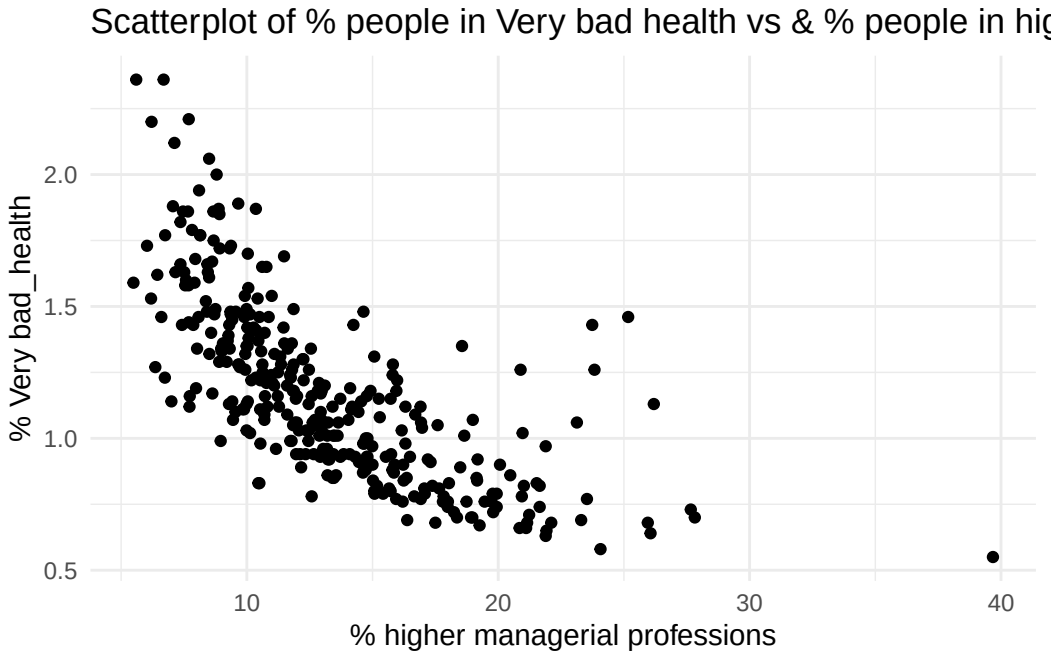
```
        x = "% aged 65-84", y = "% Very bad_health") +
    theme_minimal()
```




```
# 3
ggplot(census, aes(x = pct_Married_opposite_sex_couple, y = pct_Very_bad_health)) +
    geom_point() +
    labs(title = paste("Scatterplot of % people in Very bad health vs & % people", "in marri
         x = "% married couples", y = "% Very bad_health") +
    theme_minimal()
```




```
# 4
ggplot(census, aes(x = pct_Higher_manager_prof, y = pct_Very_bad_health)) +
    geom_point() +
    labs(title = paste("Scatterplot of % people in Very bad health vs & % people", "in highe
         x = "% higher managerial professions", y = "% Very bad_health") +
    theme_minimal()
```



Q2. Which of the associations do you think is strongest, which one is the weakest?

As noted, before, an observed association between two variables is no guarantee of causation. It could be that the observed association is:

- simply a chance one due to sampling uncertainty;
- caused by some third underlying variable which explains the spatial variation of both of the variables in the scatterplot;
- due to the inherent arbitrariness of the boundaries used to define the areas being analysed (the 'Modifiable Area Unit Problem').

Q3. Setting these caveats to one side, are the associations observed in the scatterplots suggestive of any causative mechanisms of bad health?

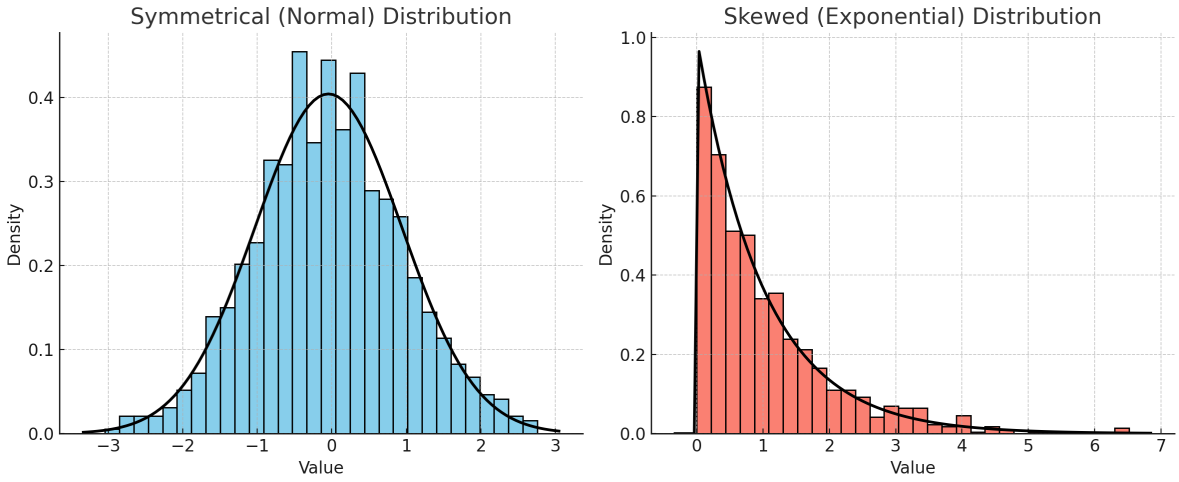
Rather than relying upon an impressionistic view of the strength of the association between two variables, we can measure that association by calculating the relevant correlation coefficient. The Table below identifies the statistically appropriate measure of correlation to use between two continuous variables.

Variable Data Type	Measure of Correlation	Range
Both symmetrically distributed	Pearson's	-1 to +1
One or both with a skewed distribution	Spearman's Rank	-1 to +1

Different Calculation Methods: Pearson's correlation assumes linear relationships and is suitable for symmetrically distributed (normally distributed) variables, measuring the strength of the linear relationship. Spearman's rank correlation, however, works on ranked data, so it's more suitable for skewed data or variables with non-linear relationships, measuring the strength and direction of a monotonic relationship.

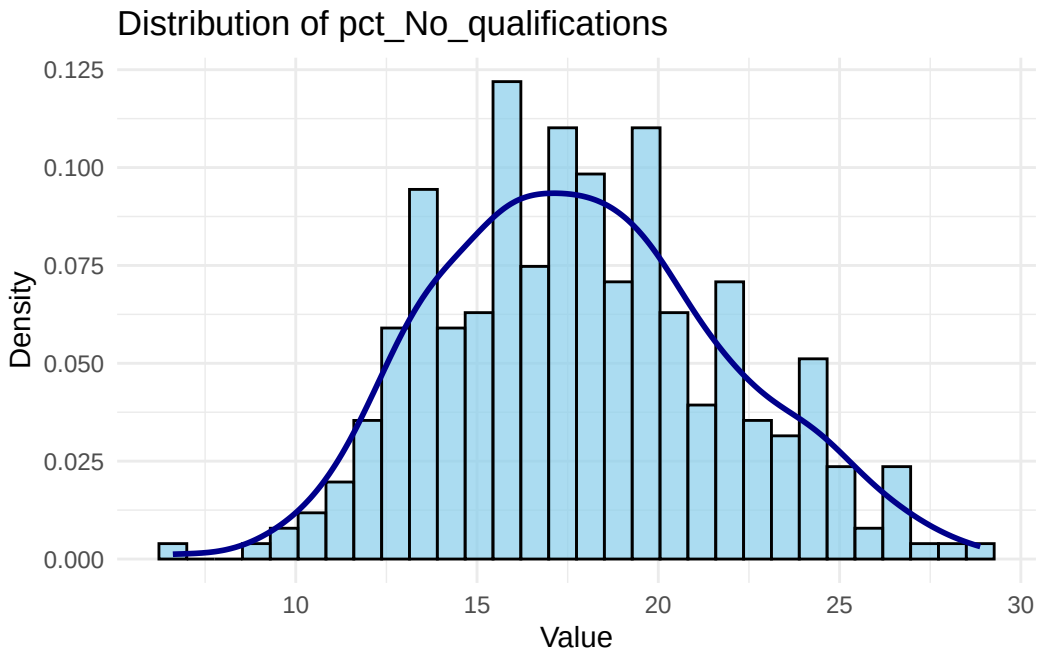
When calculating correlation for a single pair of variables, select the method that best fits their data distribution:

- Use **Pearson's** if both variables are symmetrically distributed.
- Use **Spearman's** if one or both variables are skewed.



You can check the distribution of a variable (e.g. `pct_No_qualifications` like this):

```
# Plot histogram with density overlay for a chosen variable (e.g., 'pct_No_qualifications')
ggplot(census, aes(x = pct_No_qualifications)) +
    geom_histogram(aes(y = after_stat(density)), bins = 30, color = "black", fill = "skyblue
    geom_density(color = "darkblue", linewidth = 1) +
    labs(title = "Distribution of pct_No_qualifications", x = "Value", y =  "Density") +
  theme_minimal()
```



When analyzing multiple pairs of variables, using different measures (Pearson for some pairs, Spearman for others) creates inconsistencies since Pearson and Spearman values aren't directly comparable in size due to their different calculation methods. To maintain consistency across comparisons, calculate **both Pearson's and Spearman's correlations** for each pair, e.g. do the trends align (both showing strong, weak, or moderate correlation in the same direction)? This consistency check can give confidence that the relationships observed are not dependent on the correlation method chosen. While in a report you'd typically include only one set of correlations (usually Pearson's if the relationships appear linear), calculating both can validate that your observations aren't an artifact of the correlation method.

Research Question 1: Which of our selected variables are most strongly correlated with % of population with bad health?

To answer this question, complete the Table below by editing/running this code:.

Pearson correlations

```
pearson_correlation <- cor(census$pct_Very_bad_health,
    census$pct_No_qualifications,use = "complete.obs", method = "pearson")

# Display the results
cat("Pearson Correlation:", pearson_correlation, "\n")
```

```
Pearson Correlation: 0.7619
```

Spearman correlations:

```
spearman_correlation <- cor(census$pct_Very_bad_health,
    census$pct_No_qualifications, use = "complete.obs", method = "spearman")

cat("Spearman Correlation:", spearman_correlation, "\n")
```

```
Spearman Correlation: 0.7781
```

Covariates	Pearson	Spearman
pct_Very_bad_health - pct_No_qualifications		
pct_Very_bad_health - pct_Age_65_to_84		
pct_Very_bad_health - pct_Married_opposite_sex_couple		
pct_Very_bad_health - pct_Higher_manager_prof		

What can you make of this numbers?

If you think you have found a correlation between two variables in our dataset, this doesn't mean that an association exists between these two variables in the population at large. The uncertainty arises because, by chance, the random sample included in our dataset might not be fully representative of the wider population.

For this reason, we need to verify whether the correlation is statistically significant,

```
# significance test for pearson, for example
pearson_test <- cor.test(census$pct_Very_bad_health,
    census$pct_No_qualifications, method = "pearson", use = "complete.obs")
pearson_test
```

```
    Pearson's product-moment correlation

data:  census$pct_Very_bad_health and census$pct_No_qualifications
t = 21, df = 329, p-value <0.0000000000000002
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7127 0.8037
sample estimates:
   cor
0.7619
```

Look at https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test for details about the function. But in general, when calculating the correlation between two variables, a **p-value** accompanies the correlation coefficient to indicate the statistical significance of the observed association. This p-value tests the null hypothesis that there is no association between the two variables (i.e., that the correlation is zero).

When interpreting p-values, certain thresholds denote different levels of confidence. A p-value less than 0.05 is generally considered statistically significant at the 95% confidence level, suggesting that we can be 95% confident there is an association between the variables in the broader population. When the p-value is below 0.01, the result is significant at the 99% confidence level, meaning we have even greater confidence (99%) that an association exists. Sometimes, on research papers or tables significance levels are denoted with asterisks: one asterisk (*) typically indicates significance at the 95% level ($p < 0.05$), two asterisks (**) significance at the 99% level ($p < 0.01$), three asterisks (***) significance at the 99.99% level ($p < 0.01$).

Typically, p-values are reported under labels such as "Sig (2-tailed)," where "2-tailed" refers to the fact that the test considers both directions (positive and negative correlations). Reporting the exact p-value (e.g., $p = 0.002$) is more informative than using thresholds alone, as it gives a clearer picture of how strongly the data contradicts the null hypothesis of no association.

In a nutshell, lower p-values suggest a stronger statistical basis for believing that an observed correlation is not due to random chance. A statistically significant p-value reinforces confidence that an association is likely to exist in the wider population, though it does not imply causation.

3.1.3 Part. 2: Implementing a Linear Regression Model

A key goal of data analysis is to explore the potential factors of health at the local district level. So far, we have used cross-tabulations and various bivariate correlation analysis methods to explore the relationships between variables. One key limitation of standard correlation analysis is that it remains hard to look at the associations of an outcome/dependent variable to multiple independent/explanatory variables at the same time. Regression analysis provides a very useful and flexible methodological framework for such a purpose. Therefore, we will investigate how various local factors impact residents' health by building a multiple linear regression model in R.

We use `pct_Very_bad_health` as a proxy for residents' health.

Research Question 2: How do local factors affect residents' health?

Dependent (or Response) Variable:

- % of population with bad health (`pct_Very_bad_health`).

Independent (or Explanatory) Variables:

- % of population with no qualifications (`pct_No_qualifications`).
- % of male population (`pct_Males`).
- % of population in a higher managerial/professional occupation (`pct_Higher_manager_prof`).

Load some other Libraries

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
v readr     2.1.5
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(broom)
```

and the data (if not loaded):

```
# Load dataset
census <- read.csv("../data/Census2021/EW_DistrictPercentages.csv")
```

Regression models are the standard method for constructing predictive and explanatory models. They tell us how changes in one variable (the target variable or independent variable, Y) are *associated with* changes in explanatory variables, or dependent variables, X_1, X_2, X_3 (X_n), etc. Classic linear regression is referred to *Ordinary least squares* (OLS) regression because they estimate the relationship between one or more independent variables and a dependent variable Y using a hyperplane (i.e. a multi-dimensional line) that minimises the sum of the squared difference between the observed values of Y and the values predicted by the model (denoted as $\hat{Y}$, Y-hat).

Having seen **Single Linear Regression** in class - where the relationship between one independent variable and a dependent variable is modeled - we can extend this concept to situations where more than one explanatory variable might influence the outcome. While single linear regression helps us understand the effect of **ONE** variable in isolation, real-world phenomena are often influenced by multiple factors simultaneously. Multiple linear regression addresses

this complexity by allowing us to model the relationship between a dependent variable and multiple independent variables, providing a more comprehensive view of how various explanatory variables contribute to changes in the outcome.

Here, regression allows us to examine the relationship between people's health rates and multiple dependent variables.

Before starting, we define two hypotheses:

- **Null hypothesis** (H_0): For each variable X_n, there is no effect of X_n on Y.
- **Alternative hypothesis** (H_1): There is an effect ofX_n on Y.

We will test if we can reject the null hypothesis.

3.1.4 Model fit

```r
# Linear regression model
model <- lm(pct_Very_bad_health ~ pct_No_qualifications + pct_Males + pct_Higher_manager_pro
summary(model)
```

```
Call:
lm(formula = pct_Very_bad_health ~ pct_No_qualifications + pct_Males +
    pct_Higher_manager_prof, data = census)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4911 -0.1357 -0.0368  0.0985  0.7669

Coefficients:
                        Estimate Std. Error t value             Pr(>|t|)
(Intercept)              4.00293    0.87981    4.55            0.0000076 ***
pct_No_qualifications    0.05283    0.00591    8.94 < 0.0000000000000002 ***
pct_Males               -0.07353    0.01785   -4.12            0.0000479 ***
pct_Higher_manager_prof -0.01318    0.00494   -2.67                0.008 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.213 on 327 degrees of freedom
Multiple R-squared:  0.61,  Adjusted R-squared:  0.607
F-statistic:  171 on 3 and 327 DF,  p-value: <0.0000000000000002
```

Code explanation

`lm()` Function:

- `lm()` stands for "linear model" and is used to fit a linear regression model in R.
- The formula syntax `pct_Very_bad_health ~ pct_No_qualifications + pct_Males + pct_Higher_manager_prof` specifies a relationship between:
  - **Dependent Variable**: `pct_Very_bad_health`.
  - **Independent Variables**: `pct_No_qualifications`, `pct_Males`, and `pct_Higher_manager_prof`. The model is trained on the `data` dataset.

Storing the Model: The `model <-` syntax stores the fitted model in an object called `model`.

`summary(model)` provides a detailed output of the model's results, including:

- **Coefficients**: Estimates of the regression slopes (i.e., how each independent variableaffects `pct_Very_bad_health`).
- **Standard Errors**: The variability of each coefficient estimate.
- **t-values** and **p-values**: Indicate the statistical significance of the effect of each independent (explanatory) variable.

- **R-squared** and **Adjusted R-squared**: Show how well the independent variables explain the variance in the dependent variable.
- **F-statistic**: Tests the overall significance of the model.

We can focus only on certain output metrics:

```
# Regression coefficients
coefficients <- tidy(model)
coefficients
```

```
# A tibble: 4 x 5
  term                    estimate std.error statistic  p.value
  <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)               4.00     0.880        4.55 7.58e- 6
2 pct_No_qualifications     0.0528   0.00591      8.94 2.99e-17
3 pct_Males                -0.0735   0.0178      -4.12 4.79e- 5
4 pct_Higher_manager_prof  -0.0132   0.00494     -2.67 7.97e- 3
```

These are:

- **Regression Coefficient Estimates.**
- **P-values.**
- **Adjusted R-squared**.

3.1.5 How to interpret the output metrics

3.1.5.1 Regression Coefficient Estimates

The `Estimate` column in the output table tells us the rate of change between each dependent variable X_n and Y.

Intercept: In the regression equation, this is β_0 and it indicates the value of Y when X_n are equal to zero.

Slopes: These are the other regression coefficients of an independent variable, e.g. β_1, i.e. estimated average changes in Y for a one unit change in an independent variable, e.g. X_1, when all other dependent or explanatory variables are held constant.

There are two key points worth mentioning:

- **The unit of** X and Y: you need to know what the units are of the independent and dependent variables. For instance, one unit could be one year if you have an age variable, or a one percentage point if the variable is measured in percentages (all the variables in this week's practical).
- **All the other explanatory variables are held constant**. It means that the coefficient of an explanatory variable X_1 (e.g. β_1) should be interpreted as: a one unit change in X_1 is associated with β_1 units change in Y, keeping other values of explanatory variables (e.g. X_2, X_3) constant – for instance, X_2= 0.1 or X_3= 0.4.

For the independent variable X, we can derive how changes of 1 unit for the independent are associated with the changes in `pct_Very_bad_health`, for example:

- The association of `pct_No_qualifications` is positive and strong: each increase in 1% of `pct_No_qualifications` is associated with an increase of 0.05% of very bad health rate.
- The association of `pct_Males` is negative and strong: each decrease in 1% of `pct_Males` is associated with an increase of 0.07% of `pct_Very_bad_health` in the population in England and Wales.
- The association of `pct_Higher_manager_prof` is negative but weak: each decrease in 1% of `pct_Higher_manager_prof` is associated with an increase of 0.013% of `pct_Very_bad_health`.

3.1.5.2 P-values and Significance

The ***t tests*** of regression coefficients are used to judge the statistical inferences on regression coefficients, i.e. associations between independent variables and the outcome variable. For a t-statistic of a dependent variable, there is a corresponding ***p-value*** that indicates different levels of significance in the column `Pr(>|t|)` and the asterisks `*`.

- `***` indicates "changes in X_n are significantly associated with changes in Y at the <0.001 level".
- `**` suggests that "changes in X_n are significantly associated with changes in Y between the 0.001 and (<) 0.01 levels".
- Now you should know what `*` means: The significance is between the 0.01 and 0.05 levels, which means that we observe a less significant (but still significant) relationship between the variables.

P-value provide a measure of how significant the relationship is; it is an indication of whether the relationship between X_n and Y found in this data could have been found by chance. Very small p-values suggest that the level of association found here might **not** have come from a random sample of data.

In this case, we can say:

- Given that the p-value is indicated by `***`, changes in `pct_No_qualifications` and `pct_Males` are significantly associated with changes in `pct_Very_bad_health` at the <0.001 level; the association is highly statistically significant; we can be confident that the observed relationship between these variables and `pct_Very_bad_health` is not due to chance.
- Given that the p-value is indicated by `**`, changes in `pct_Higher_manager_prof` are significantly associated with changes in `pct_Very_bad_health` at the 0.001 level. This means that the association between the independent and dependent variable is not one that would be found by chance in a series of random sample 99.999% of the time.

In both cases we can then confidently reject the **Null** hypothesis (H_0: no association between dependent and independent variables exist).

Remember, If the *p-value* of a coefficient is smaller than 0.05, that coefficient is statistically significant. In this case, you can say that the relationship between this independent variable and the outcome variable is *statistically* significant. Contrarily, if the *p-value* of a coefficient is larger than 0.05 you can conclude that there is no evidence of an association or relationship between the independent variable and the outcome variable.

3.1.5.3 R-squared and Adjusted R-squared

These provide a measure of model fit. They are calculated as the difference between the actual value of Y and the value predicted by the model. The **R-squared** and **Adjusted R-squared** values are statistical measures that indicate how well the independent variables in your model explain the variability of the dependent variable. Both R-squared and Adjusted R-squared help us understand how closely the model's predictions align with the actual data. An R-squared of 0.6, for example, indicates that 60% of the variability in Y is explained by the independent variables in the model. The remaining 40% is due to other factors not captured by the model.

Adjusted R-squared also measures the goodness of fit, but it adjusts for the number of independent variables in the model, accounting for the fact that adding more variables can artificially inflate R-squared without genuinely improving the model. This is especially useful when comparing models with different numbers of independent variables. If Adjusted R-squared is close to or above 0.6, as in your example, it implies that the model has a **strong explanatory power** while not being overfit with unnecessary explanatory variables.

A high R-squared and Adjusted R-squared indicate that the model captures much of the variation in the data, making it more reliable for predictions or for understanding the relationship between Y and the explanatory variables. However Low R-squared values suggest (e.g. 0.15) that the model might be missing important explanatory variables or that the relationship between Y and the selected explanatory variables is not well-captured by a linear approach.

An R-squared and Adjusted R-squared over 0.6 are generally seen as signs of a **well-fitting model** in many fields, though the ideal values can depend on the context and the complexity of the data.

3.1.6 Interpreting the Results

```
coefficients
```

```
# A tibble: 4 x 5
  term                    estimate std.error statistic  p.value
  <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)               4.00     0.880        4.55 7.58e- 6
2 pct_No_qualifications     0.0528   0.00591      8.94 2.99e-17
3 pct_Males                -0.0735   0.0178      -4.12 4.79e- 5
4 pct_Higher_manager_prof  -0.0132   0.00494     -2.67 7.97e- 3
```

Q4. Complete the table above by filling in the coefficients, t-values, p-values, and indicating if each variable is statistically significant.

Variable Name	Coefficients	t-values	p-values	Significant?
pct_No_qualifications				
pct_Males				
pct_Higher_manager_prof				

From the lecture notes, you know that the Intercept or Constant represents the estimated average value of the outcome variable when the values of all independent variables are equal to zero.

Q5. When values of `pct_Males`, `pct_No_qualifications` and `pct_Higher_manager_prof` are all *zero*, what is the % of population with very bad health? Is the intercept term meaningful? Are there any districts (or zones, depending on the dataset you chose) with zero percentages of persons with no qualification in your data set?

Q6. Interpret the regression coefficients of `pct_Males`, `pct_No_qualifications` and `pct_Higher_manager_prof`. Do they make sense?

3.1.7 Identify factors of % bad health

Now combine the above two sections and identify factors affecting the percentage of population with very bad health. Fill in each row for the direction (positive or negative) and significance level of each variable.

Variable Name	Positive or Negative	Statistical Significance
pct_No_qualifications		
pct_Higher_manager_prof		
pct_Males		

Q7. Think about the potential conclusions that can be drawn from the above analyses. Try to answer the research question of this practical: How do local factors affect residents' health? Think about causation *vs* association and consider potential confounders when interpreting the results. How could these findings influence local health policies?

3.2 Part C: Practice and Extension

If you haven't understood something, if you have doubts, even if they seem silly, ask.

1. Finish working through the practical.
2. Revise the material.
3. Extension activities (optional): Think about other potential factors of very bad health and test your ideas with new linear regression models.