# STATS 201 Data Analysis
# Assignment 1, NEFU, 2021

**Instructions concerning this assignment:**

A major purpose of this assignment is to ease you into the assignment procedures and the use of the statistical package **R**. We will be doing this through the use of **R Studio** and using **R Markdown**.

We are providing you an R Markdown document called **`STATS201_2021_NEFU_A1.rmd`** which will have some answers already filled in. You will need to fill in and complete the rest of the document. The data files you will be using for the assignment are described in the questions and are available online. Make sure you put these data files in the same place you put the R Markdown document because it is going to look for them there. The first change you need to make to the markdown document is put your name and ID number at the top.

**Question 1. [15 Marks]**

We wish to investigate the relationship between electricity consumption and the gross domestic product (GDP) for countries of the world. GDP is an indicator of a country's economic performance adjusted for purchasing power parities to account for between-country differences in price levels. Information was obtained for a selection of 28 of the most populous countries in the world. The data is stored in "`electricity.csv`" and contains the variables

| | |
|---|---|
| `Electricity` | electricity consumption in billions of kilowatt-hours. |
| `GDP` | gross domestic product (GDP) in billions of dollars (US). |
| `Country` | name of the country. |

A lot of the analysis for this question is already filled in for you. There are several parts and additional questions for you to address:

- Comment on the initial plot of the data.
- We fitted an initial linear model, but there were issues with the two countries with the highest GDP. Identify these two countries, replot the data eliminating these two countries, comment on this plot, and then refit the simple linear regression model (including assumption checks) without these two countries.
- Create a scatter plot with the fitted line from the new fitted model superimposed over it.
- Write an appropriate **Executive Summary**.
- Use the final model to provide a prediction for the electricity usage of a country with a GDP of 1000 billion dollars. Interpret this interval. Based on this interval, how useful is the model for prediction?

**Question 2. [8 Marks]**

A researcher speculates that the average life expectancy among countries of the world is about 68 years. He collected data from a random sample of 55 countries and wishes to use this data to test his hypothesis. The data is stored in "`countries.csv`" and contains the variable

| | |
|---|---|
| `Life` | life expectancy for each country (in years). |

- Comment on the plots/exploratory data analysis.
- Manually calculate the *t*-statistic for comparing the mean life expectancy to 68 years and the corresponding 95% confidence interval.
- Why are the P-values from the *t*-test output and null linear model different?
- Write an appropriate **Executive Summary**.

**Question 3. [16 Marks]**

A researcher in the city Eugene in the state of Oregon, USA, was interested in how the sale price of a house is influenced by the age of the house. In 2005, she took a random sample of 76 single-family homes. The data is stored in "`homes.csv`" and contains the variables

| | |
|---|---|
| `Price` | the sale price of a house (in thousands of dollars). |
| `Age` | age of the house, defined as 2005 minus the year the house was built). |

- Comment on the question of interest/goal of the analysis.
- Comment on the initial plot of the data.
- Fit an appropriate linear model, including model checks.
- Plot the data with your estimated model superimposed over it.
- Write appropriate **Methods and Assumption Checks** and **Executive Summary.**

# Assignment Notes

For a lot of assignment questions we will simply be giving you descriptions of how and why data was collected and minimal guidance. You will see how to analyse the data in the case studies in class, but here is a general approach to answering open data questions:

- Comment on the question(s) of interest or the goal(s) of the analysis.
- Look at the data (plot it, get summary statistics) and comment on it.
- Fit a model to the data
  - Check the model assumptions.
  - Change model and repeat checks as needed. You may have to do this more than once.
- Generate inference output from your final model.
- Write a **Method and Assumption Checks** section.
  - This will detail the steps you took and why you took them in building the model.
  - It will include brief descriptions of the model assumption checks.
  - It will include a mathematical statement for the final model you fitted.
- Write an **Executive Summary**.

## Make sure you read the notes on the next page!

**Some very important notes:**

- When using case studies as guides, DO NOT blindly follow one case study. All data sets have their own individual attributes and are not likely to perfectly match a case study you find. Instead INTELLIGENTLY use the case studies to guide you.

- When commenting on plots, keep the comments brief and relevant. When commenting on assumptions, you do not need to go into great detail. If the plot shows no problems, it is ok to say that. If the plot shows problems, briefly describe the problems and then say what can be done about them.

- When writing **Executive Summaries**:
  - We want the main conclusions in terms of the original questions asked.
  - If there is a key question or goal for the data, make sure you answer it directly, then go into details as needed. For example, if a study is asking if a tablet affected blood pressure you could have a sentence along the lines of "we have evidence that the tablet increased blood pressure", then back this up with appropriate quantification (for example, giving a range of numbers for the increase in mean blood pressure after taking the tablet). Don't leave us to infer the results from the quantification.
  - Point out any unusual steps or changes made to your model in easy to understand terms.
  - You should be using easy to understand, natural language. You should be avoiding using variable names, lots of decimal places and unnecessary detail.
  - State units when known.
  - This should be a brief, easy to read summary of your analysis that someone with little understanding of statistics can comprehend without having read through the analysis.

- If you want to check your Executive Summary, get a friend who hasn't done statistics to read it and tell you what they think it means.

# Obtaining a Copy of **R** and **R Studio** for Use at Home

First, download and install the latest version of **R** with the latest s20x library:

Go to: www.stat.auckland.ac.nz
Click on **CRAN** under the **HOSTING** heading at the top right of the screen
In the box labelled **Download and Install R**, click on **Linux** or **(Mac) OS X** or **Windows** depending on your computer system

**What is R Markdown?**

**R Markdown** is a relatively new markup language built into **R Studio**. It is an authoring format that enables easy creation of dynamic documents and reports from **R**.

This means you work on a master document including your R commands and then "knit" this together, with R Studio running R in the background and creating a final document with all your graphs and computer output inserted in the correct place. If you want to make changes in your output document, you just "knit" the document again; you don't have to redo the analysis as it is automatically rerun for you. You also don't need to repeatedly copy and paste output.

**Getting Started:**

The best way to see this in action is to try this.

Run R studio. Under the File menu, go Open File and navigate to find and Open STATS201_2021_NEFU_A1.rmd.

Next, find the Knit button (next to an icon of some wool and knitting needles) and choose **Knit Word** (unless you are at home and do not have Word installed in which case choose **Knit HTML**). This will immediately "knit" the document into a Word (or HTML) file, running the R code and creating a file of the output.

Congratulations – you have created your first document!

**Triple Knitting:**

You can knit documents to 3 formats: HTML, Word or PDF. We will accept assignments handed in using any of the three formats, though we prefer Word as it is cleaner and you can easily resize plots to tidy it up and reduce the number of pages in your assignment, if you wish.

To knit to Word, you must have Microsoft Word installed on your computer.

To knit to PDF, you must have LateX installed on your computer. LateX is a free, extremely powerful mathematical typesetting language. Unfortunately we cannot support you installing this, but you are welcome to do so.

**Editing the Document:**

Now you should try editing the document to see what happens.

- Change the title to include your name and ID number.
- Edit some of the comments.
- See what happens when you try changing the formula for the equations.
- Add some more R code.

Make sure you have saved the original document and remember to back up often. You learn about new software tools by experimenting with them. You can always undo mistakes.

Equations take a little more practice. You will note we have provided you the equations for Assignment 1 to make your life easier.

On the following page we have provided some information on how you can change the formatting of your document such as adding headers, lists, bold, italics. It is all fairly easy with a little practice.

**Some formatting options:**

To start a new paragraph, end a line with two spaces.

*italics* produces *italics*  **bold** produces **bold**

# Header produces **Header**

## Header produces **Header**

### Header produces **Header**

...

###### Header produces **Header**

To make an unordered list:

```
* Item 1
* Item 2
    + sub item 1
    + sub item 2
```

gives

- Item 1
- Item 2
    - sub item 1
    - sub item 2

To make an ordered list:

```
1. Item 1
2. Item 2
    + sub item 1
    + sub item 2
```

gives

1. Item 1
2. Item 2
    - sub item 1
    - sub item 2

To write equations, enclose them between $ signs.
$\beta_0=55$ gives $\beta_0 = 55$

To get Greek letters: \sigma produces σ etc.

superscripts^2 gives superscripts$^2$  subscripts_1 gives subscripts$_1$

\ne gives ≠  \times gives ×

Want to put a hat on a symbol? $\hat{\beta}$ gives $\hat{\beta}$

Want to take a square root? $\sqrt{n}$ gives $\sqrt{n}$

Want to make a fraction? $\frac{this}{that}$ gives $\frac{this}{that}$

Want to make a silly complicated formula?

$\hat{\mu} \ne \sqrt{\frac{\beta_0^2}{\sigma \times \bar{x}}}$

gives $\hat{\mu} \neq \sqrt{\frac{\beta_0^2}{\sigma \times \bar{x}}}$