MATHS 7107
Data Taming
Practical 3

1 Background

You work as a data scientist at the multi-million dollar Australian jewellery company *Sparkles and Glitter Pty Ltd*. Your boss has asked you to do some research on diamonds to better understand which diamonds have a higher price so eventually the Company can increase profits (and hopefully pay you more money!!)

Your boss has specifically told you that your work must be in a report form so it can be forwarded to the sister company *Shine and Shimmer Inc.* located in the United States of America. Your boss wants them to be able to run your analysis on the data they have collected on diamonds they have sold. They use `R Markdown`, and so you will have to write your report in `R Markdown`. They also strictly adhere to the principles of Tidy Data and Tame Data.

1.1 Deliverable specifications

Here are some important instructions that you must adhere to:

- Your `R Markdown` must produce a Portable Document Format file.
- You will then also need an identical HTML file.
- Your report must knit on their machine, otherwise it will be rejected. (Your work is of no use if it doesn't work.)
- Make sure you use the Tidy Data and Tame Data conventions.
- Use sections and subsections to make it easier to read.
- You must label your code chunks.
- Show the `R` code that you are using, but don't show any warnings. Your boss doesn't want to see all of that.
- Make sure you put text in your report to explain what each piece of output means.

IMPORTANT!

Since you have been employed as a data scientist, you are expected to be able to satisfy these conditions. **If you do not, you'll be fired!** So it's time to tame some data.

2 Questions

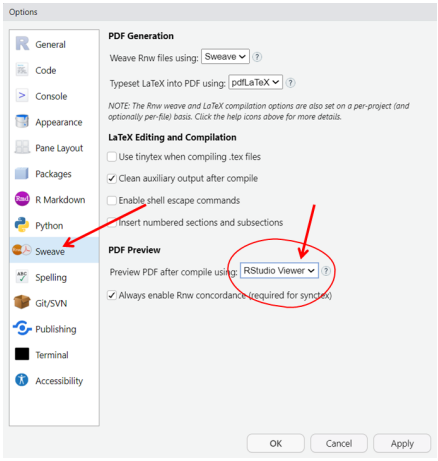
You will need to complete the following for your boss:

1. First setup an `RStudio` project for this report.

Hint

Turn on the Sweave viewer, so that you can see the PDF output straightaway. Click on Tools → Global Options... → Sweave → Preview PDF after compile using: RStudio Viewer.

See the picture below.



2. Create an `R Markdown` file. Make sure you give it a sensible name.
   - **Hint**: Click the "Knit on Save" button, then use `Ctrl-s` to save and knit at the same time.
3. Load the `diamonds` dataset. This is saved in the `tidyverse` package.
4. Check the data to see if there are any entries missing (i.e. are there any `NA` values?).
5. Determine how many types of `cut` there are. What are they? Show how many diamonds there are of each particular `cut`. (For this section you could try using "inline code" in your R Markdown document.)
6. Your boss wants to know whether the price of the diamonds depends more on `cut` or `color`, and they want you to use a scatterplot. Using `ggplot`, produce two scatterplots of `price`, one using `cut` and one using `color`.
7. Since your boss is not a data analyst, they don't know that the scatterplot will be useless. So also produce two boxplots of the same data. (Use the `aes` option `fill=x_variable_name` to add some colour to your boxplots.) From these plots, which variable appears to affect price more, `cut` or `color`?
8. If a customer wants to buy a `Premium` diamond, with `color` rating `J`, how much should they expect to pay on average?