MATHS 7107 Data Taming Tutorial

Questions

You work as a data scientist at the multi-million dollar Australian jewellery company *sparkles and glitter*. 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* 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.

The price of the diamonds has already been converted to US dollars.

Specifically you will need to complete the following for your boss:

1. Create a file - some sort of reproducible report - that can incorporate your explanations, code and output (analysis and plots etc).
2. Load the `diamonds` dataset. This is saved in the `tidyverse` package.
3. Check the data to see if there are any entries missing (i.e. are there any NA's?).
4. Determine how many types of `cut` there are. What are they? Show how many diamonds there are of each particular `cut`.
5. Your boss wants to know whether the price of the diamonds depends more on `cut` or `color` Using `ggplot`, produce two side-by-side boxplots of `price`, one using `cut` and one using `color`. Which variable appears to affect price more, `cut` or `color`?
6. If a customer wants to buy a `Premium` diamond, with `color` rating `J`, how much should they expect to pay on average?
7. Write a short summary outlining exactly what you did so your boss is prepared when his colleague from America zooms next week. This will mean your research is reproducible to the sister company and your boss won't get cranky when he doesn't know an answer!