Part 1: Data Background and Variables [~3 pages in length]

Copied from the overview document.

This writeup should be approximately **3 pages** in length and is worth **4%** of your final grade.

In this component, you will introduce your dataset and provide context for your reader. This is where you demonstrate that you understand how your data was collected and what the variables represent, before diving into analysis.

Your writeup should include:

- **Dataset Introduction**: Where did the data come from? What is it about? Why is it interesting or important?
- **Variable Descriptions**: Describe the key variables in your dataset. What do they measure? What are their units? Are they categorical or quantitative[1]?
- **Research Context**: What kinds of questions could be answered with this dataset? What makes this dataset suitable for linear regression analysis?
  - You could be thinking about questions like:
    - Are there quantitative predictor and response variables?
    - Are there categorical variables that could be used to define groups for comparison?
    - Is the sample size large enough to support your analysis?
    - Is it reasonable to assume that the observations are *independent*?
- **Data Quality Considerations**: Are there missing values? Outliers? Any data cleaning that is required?
  - This would not be a result of EDA, but rather an initial look at the data file (if applicable) and any information provided about the data collection process.

What I'm looking for:

- Clear writing that would help a reader unfamiliar with your dataset understand it.
- Thoughtful selection of which variables to highlight (you don't need to describe every single variable if there are many).
- Evidence that you've explored and understood the structure of your data.

[1] It is here that I typically find many analyses are lacking. We have a tendancy to want to jump right into the modelling or "fun" part of analysis, but if you do not know *what* you are modelling, then your interpretation is not going to be even remotely as useful.

Part 1: Data Background and Variables (Due October 30, 2025)

Format:

- Approximately 3 pages in length
- PDF format rendered using Quarto
- All code displayed in an appendix *only* (no code present in the writeup's main body)
- References / bibliography as appropriate
- Professional formatting and presentation

Groups and Data Selection:

- Maximum of 4 members (groups already formed)
- Select a dataset that is of interest to your group

Writeup Content:

- **Dataset Introduction**: Origin, context, importance (with citations)
- **Variable Descriptions**: Key variables, their measurements, units, and types
- **Research Context**: Potential questions, suitability for linear regression
- **Data Quality Considerations**: Missing values, outliers, data cleaning needs
  - You are not expected to perform EDA yet. This discussion should be based on viewing the raw data file and reading the data documentation and description.

Writeup Structure:

- Title page
  - Must contain: title of project, group members' names, group number, date
- Main content
- References
- Appendix with code (if any exploratory code is included)

> AI Tool Use
>
> I am fine with groups using AI to **help** them with their projects. You are responsible for writing the document and anything that you submit should be created by your group members (this does *not* include AI).
> AI is useful for help with literature searchs, grammar and spelling, and help with code and document formatting.
> You are creating the content; AI can help you with this process, but it should not be doing the work for you.
> **As always, if you are unsure, please talk to Dave!**

Marks [Total = 4% of final grade]

- **Document Format and Presentation** [1.5%]: Professional formatting, clear structure, appropriate length
- **Dataset Introduction** [1%]: Clear context, proper citations, importance established
- **Variable Descriptions** [1%]: Comprehensive coverage of key variables
- **Research Context and Data Quality** [0.5%]: Thoughtful discussion of research potential and data considerations

Getting Started - Finding and Choosing a Dataset

Finding datasets can be challenging.

You must check with Dave before using any datasets from Kaggle or the UC Irvine Machine Learning Repository.

Data Suitability

If you are unsure if the dataset will be appropriate for your project, please ask me! I would be more than happy to talk to you about your projects.

Possible sources for datasets:

1. TidyTuesday R Project on GitHub

- Contains data for various topics and from many domains and sources.
- Be careful to examine the data and ensure that it will be suitable for *this* project!

2. Statistics Canada

- Canadian government data on a variety of topics.
- Many datasets are available for free download.

3. Canada's Open Government Portal

- Another source of Canadian government data (similar to Data.gov below).
- Searchable by topic, format, and more.

3. Propublica

- Archived datasets from 2013 to 2023.
- From their About Us page: "*ProPublica is an independent, nonprofit newsroom that produces investigative journalism with moral force.*"

4. FiveThirtyEight

- Datasets used in their articles.
- Topics include politics, sports, science, economics, and culture.

5. Data.gov

- The U.S. government's open data site.
- Thousands of datasets on a wide range of topics.

6. World Bank Open Data

- Global development data.
- Economic, social, and environmental indicators.

7. Google Dataset Search

- A search engine for datasets across the web.
- Can help you find datasets on specific topics.

8. Flowing Data

- Datasets used in their visualizations.
- Topics include health, demographics, and social trends.
- Datasets referenced in articles.

Document Format and Submission

Use Quarto and submit a rendered PDF document with mostly the default settings for margins, line spacing, etc.. You are welcome to make minor changes as you see fit.

Part 1 should be **approximately 3 pages** of content[2]. **More is NOT better.** You should explain what you need/want to explain. We are not "filling up space".

A goal of this project is to get practice scoping your writing and determining what is most important for your reader to understand.

Since Part 1 focuses on understanding your dataset rather than creating visualizations, you likely won't have plots or figures in this component. If you do include any exploratory tables or figures, they should be labelled appropriately with figure captions or table captions[3].

The submission should be in a **PDF format** and should be submitted to Crowdmark along with the .qmd file used to create the PDF.

You should hide all of the code chunks present in the main body of the report. You should include all code in an appendix.

There is a file on Blackboard to help with this as well as material at the end of this document.

[2] The 3 pages do not include a title page, reference pages, or appendix.

[3] But seriously, you're moving into the Part 2 if you start including figures and tables, so be careful here.

Main Content - Data Background and Variables

The goal of Part 1 is to demonstrate that you understand your dataset before diving into exploratory data analysis and statistical modeling. You should introduce your reader to the data and provide sufficient context about where it came from, what it measures, and why it's interesting.

Your Part 1 writeup should include:

1. Dataset Introduction

Provide context for your dataset:

- **Origin**: Where did the data come from? Who collected it? When?
- **Purpose**: Why was this data collected? What was the original research question or goal?
- **Importance**: Why is this dataset interesting or relevant?
- **Citations**: Use information from the reference documents on Github if provided. Cite the GitHub page, the data source, and any other sources you used.
  - I am indifferent as to the citation style that you use, however please be consistent.
  - You can (and probably should?) use Quarto's built-in ability to cite sources: Quarto Citations.
  - More info is available at the end of this document.

2. Variable Descriptions

Describe the key variables in your dataset. You don't need to describe every single variable if there are many, but focus on those that are most relevant or interesting:

- **What does each variable measure?**
- **What are the units?** (e.g., dollars, kilograms, years)
- **Variable type**: Is it quantitative (continuous or discrete) or categorical (ordinal or nominal)?
- **Variable role**: Which variables might serve as response variables? Which might be predictors?

> ℹ Why Variable Types Matter
>
> Understanding whether a variable is quantitative or categorical will determine what statistical methods you can use. This is foundational for your later analysis.

3. Research Context

Discuss what makes this dataset suitable for the kinds of analyses you'll be doing in this course:

- **Potential research questions**: What kinds of questions could be answered with this dataset?
- **Suitability for linear regression**:
  - Are there quantitative predictor and response variables?
  - Are there categorical variables that could be used to define groups for comparison?
- **Sample size**: Is the sample size large enough to support your analysis?
- **Independence**: Is it reasonable to assume that the observations are independent?

4. Data Quality Considerations

Discuss any data quality issues you've identified from your initial examination:

- **Missing values**: Are there variables with missing data?
- **Outliers**: Are there unusual or extreme values that might need investigation?
- **Data cleaning**: Will any data cleaning or transformation be required?
- **Data collection concerns**: Are there any issues with how the data was collected that might affect your analysis?

! Not Full EDA Yet

This is not yet a full exploratory data analysis with plots and numerical summaries. This is about understanding the **structure** and **context** of your data before you begin analyzing it. You may look at the data file and summary output to inform this discussion, but detailed visualizations and numerical summaries belong in Part 2.

Implementation Notes

Chunk Label Names

Do not use special characters or spaces in your chunk labels.
Use hyphens instead. (Yes, I know that hyphens are special characters …)

Figure and Table Captions

You can add figure captions in addition to any number of additional properties, to R plots by adding YAML (chunk options, `#|`) at the top of code chunks:

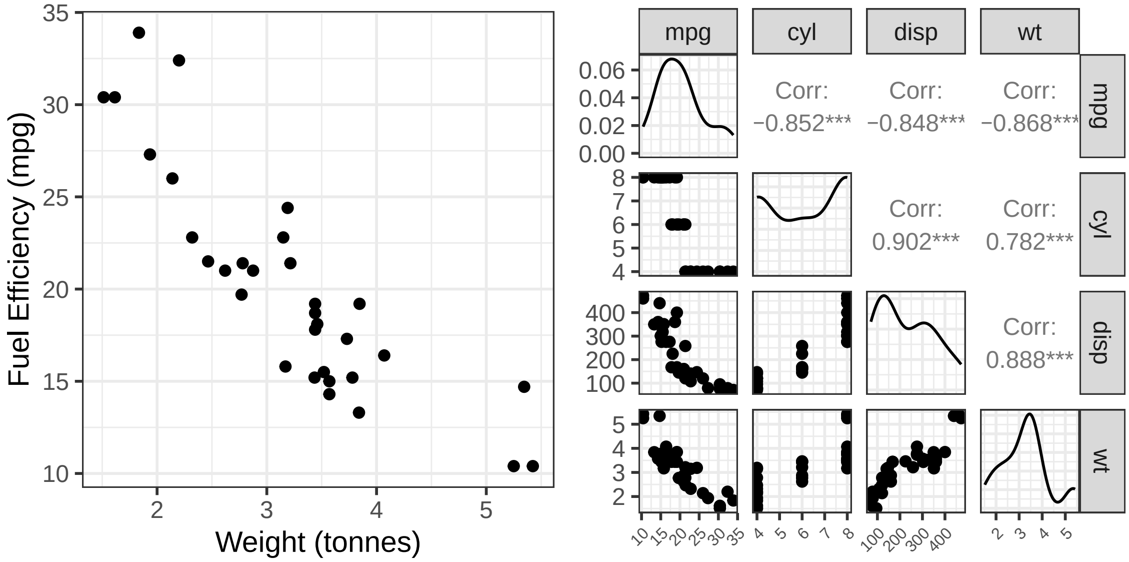
~~~
```{r}
#| message: false
#| warning: false
#| label: fig-my-plot
#| fig-cap: "Super interesting plots!"
#| fig-subcap:
#|   - "Scatter plot of weight vs. fuel efficiency."
#|   - "Pairs plot of some variables."
#| fig-height: 3
#| fig-width: 3
#| layout-ncol: 2
#| fig-align: center

ggplot(data = mtcars, aes(x = wt, y = mpg)) +
  geom_point() + labs(x = "Weight (tonnes)", y = "Fuel Efficiency (mpg)")
ggpairs(mtcars %>% select(mpg, cyl, disp, wt),
        upper = list(continuous = wrap("cor", size = 3))) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 6))
```
~~~

You can then reference the figure in the text using the label you provided using `@fig-my-plot`. For example, `"The plot in @fig-my-plot shows that ..."` will output “The plot in Figure 1 shows that . . .” in your rendered document. Notice that the word “Figure” is inserted for you.

Captions and cross-referencing are the same for tables, except we start the labels with `tbl-` which provides “Table XX” instead of “Figure XX”.

For example, `@tbl-my-table` can be used to reference the table below: “See Table 1 for more information.” (NOTE: the word “Table” is added automatically).



(a) Scatter plot of weight vs. fuel efficiency.

(b) Pairs plot of some variables.

Figure 1: Super interesting plots!

~~~
```{r}
#| label: tbl-my-table
#| tbl-cap: "A table of interesting data."

kable(mtcars %>% select(mpg, cyl, disp, wt) %>% head(n = 3)) %>%
  kable_styling(latex_options = "hold_position")
```
~~~

Table 1: A table of interesting data.

	mpg	cyl	disp	wt
Mazda RX4	21.0	6	160	2.620
Mazda RX4 Wag	21.0	6	160	2.875
Datsun 710	22.8	4	108	2.320

Figure Placement

You may notice that your figures move around once you add captions. The reason is that they stop being "images" and start being "floats". The name is pretty descriptive, actually; they float around to try to minimize white space.

If you want your figures to show up *where you put them in your document* (which is usually the intention), then you can add the `header-includes:` content to your YAML chunk at the top of your document:

```
title: "Untitled"
format: pdf
header-includes:
    - \usepackage{float}
    - \floatplacement{table}{H}
    - \floatplacement{figure}{H}
```

Bibliography and Citing

To create the `.bib` file, you would create a text file with a `.bib` file extension. This file must be in the same directory as your Quarto document, just like with data files!

Each reference has a form similar to:

```
  @article{smith201X,
    title         = {An interesting article},
    author        = {John Smith},
    year          = {201X},
    journal       = {Journal of Interesting Articles}
  }
```

For example, an article that showed up for me on Google Scholar would have the form:

```
@article{oliveira202410,
  title={The 10 October 2024 geomagnetic storm may have caused
 the premature reentry of a Starlink satellite},
  author={Oliveira, Denny M and Zesta, Eftyhia and Nandy, Dibyendu},
  journal={arXiv preprint arXiv:2411.01654},
  year={2024}
}
```

To get the bibtex entry:

1. Go to Google Scholar
2. Search for the article
3. Click on the "Cite" button below the article
4. Click on the "BibTeX" link
5. Copy the text in the box and paste it into your own `.bib` file.
6. You cite using `@oliveira202410` in your Quarto file.

Including Code in an Appendix

Hiding Code Chunks:
To hide your code chunks from the main body of your document, include the following chunk option at the top of your code chunks:
`#| echo: false`
This chunk option will suppress the code from being displayed but will still run the contents of the code chunk (i.e., to output a table or plot!).

Including All Code in an Appendix:
Create a new page `\newpage` and include all of the code chunks on that page with a header of "Appendix" (i.e., `# Appendix` to get the header).

```
#| echo: true
#| eval: false
#| ref-label: !expr knitr::all_labels()
```

Using the above chunk options will, for this document, produce:

```r
library(tidyverse)
library(knitr)
library(kableExtra)
library(GGally)

theme_set(theme_bw())
ggplot(data = mtcars, aes(x = wt, y = mpg)) +
  geom_point() + labs(x = "Weight (tonnes)", y = "Fuel Efficiency (mpg)")
ggpairs(mtcars %>% select(mpg, cyl, disp, wt),
        upper = list(continuous = wrap("cor", size = 3))) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 6))
kable(mtcars %>% select(mpg, cyl, disp, wt) %>% head(n = 3)) %>%
  kable_styling(latex_options = "hold_position")
```

Acknowledgements

- Claude Code was used to help modify a previous semester's document to better align with the goals of this semester.
- Some autocomplete use from GitHub Copilot.
- Perplexity AI used to help search for data sources.