Prompt:
The `sleep.csv` contains the following data:

- The total number of hours slept (h) denoted `Duration` in the dataset;
- The number of minutes (minutes) engaged in physical activity during the day denoted `Physical_Activity` in the dataset;
- The (self-reported) stress level experienced ranked on a scale of 1 to 10, where 10 is high and 1 is low, denoted `Stress` in the dataset;
- The (self-reported) quality of sleep reported on a scale of 1 to 10, where 10 is high and 1 is low, denoted `Quality` in the dataset;

for 374 people.

In your groups, construct a linear regression model to predict the duration of sleep from the other columns in the data. In your model construction, you may find it necessary to "transform" some of the input variables. For this data set, the transformations that may be helpful are:

- Taking the square root of the Stress values;
- Taking the square of the Stress values;
- Taking the natural log of the Quality values;
- Taking the square of the Quality values;
- Taking the natural log of the Duration values;
- Taking the square root of the Duration values;

It may be helpful to review the Lecture 19 notes to see how to "transform" columns in python for linear regression. In your construction, you may want to use one, multiple, or none of the above transformations. As part of your groups, you will need to determine the **best** linear regression model to predict the Sleep Duration. If you choose a model with transformed variables, it is expected that you will compare the transformed model results with the results from a standard multiple linear regression (i.e., without transformed data). **Be sure to clearly state which linear regression your group deems to be the 'best' and explain why you have chosen this model.**

Note: Since some of the data is ranked on a scale of 1 to 10, we will see "columns" of data points when looking at some of the diagnostic plots at the integer or half-integer values (i.e., the points on the rank scale). In these cases, when we are looking for "random scatter", we look vertically in the column to see if the points have any pattern and to see if the points are "equally" distributed over the horizontal $y = 0$ line (if applicable). If we do not see a equal distribution in the columns of the plots, then we may be concerned that one (or more) of the linear regression assumptions are violated.

Do NOT split the data into a training and testing set: i.e., use all of the data to construct the linear regression model.

Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs)
Learning to think critically is an important learning goal of this course. Therefore, you (or your group) are expected to create your own ideas, without the use of generative artificial intelligence tools, including ChatGPT or other similar tools (e.g., Gemini, CoPilot, Claude, etc), to complete any assignment/assessment in this course. Specifically:

- Do NOT submit any text, code, and/or images created by GenAI or LLMs.
- Do NOT submit any work that is paraphrasing output generated by GenAI or LLMs.
- Do NOT cite GenAI or LLMs as a reference. This is equivalent to citing Wikipedia, which is not a trustworthy or appropriate reference. Any references should be from peer-reviewed journal articles or published books.

Utilizing GenAI or LLMs to complete work in this course is academic misconduct. Submitting work that is not your own will receive a grade of zero and will be reported to the Undergraduate Chair.

In addition, redistribution of course materials (e.g., course notes, instructional materials, the Math 360 online textbook, Modelling Assignments), including posting them online and sharing them with GenAI or LLM chatbots (e.g., ChatGPT, Gemini, CoPilot, Claude, Bard, etc) without permission of the instructional team is NOT acceptable.

Bibliography
You may need to find values from the literature to use in your model and/or other information to inform the construction of your model. Be sure to include a bibliography in the format shown below in your write up and provide in text citations using the format [last name, first initial. publication year]. As a reminder, Wikipedia is not a reliable reference and should not be used in bibliographies.

If you are citing information from the course notes and/or course textbook, then please indicate the corresponding lecture number or the textbook section.

The bibliography does NOT count towards the page count – see the next page for submission guidelines.

[1] Daniels, L., Scott, M., & Miskovic, Z. L. *The role of Stern layer in the interplay of dielectric saturation and ion steric effects for the capacitance of graphene in aqueous electrolytes*. J. Chem. Phys., 146(9), (2017).

What to Submit:

- Groups submit ONE Jupyter notebook and PDF presenting all 6 steps above to the Canvas dropbox and a PDF to comPAIR.
- This means that ONE person in your group should submit the Jupyter notebook and PDF to Canvas and ONE person in your group should submit the PDF to comPAIR.
- The Jupyter notebook should:
  - Contain the names of your group members (it is OK to remove these for the comPAIR review).
  - Contain all steps of the modelling process **clearly labelled**. See the examples from class for the expected model formatting.
  - Follow the mathematical communication guidelines.
  - Not exceed 6 pages as a PDF. **Any work after page 6 will NOT be read or graded.**
  - Not exceed 4 plots/diagrams/figures. Note that the summary table (i.e., the OLS results table) from the linear regression does NOT count as a plot/diagram/figure, however, the space that the summary table occupies WILL count towards the 6 page limit. **Any analysis related to subsequent plots/diagrams/figures after the initial 4 will NOT be read or graded.**
    - Note that figures with subplots will count as the number of subplots. For example, if there is one figure with 3 subplots, then this will count as 3 plots. The only exception is the `sm.graphics.plot_regress_exog( , )` plot, which outputs a figure with 4 subplots. We will count this as 1 plot for this modelling assignment. Note that this is per covariate, so if you produce the `sm.graphics.plot_regress_exog( , )` plot for 3 covariates, then this counts as 3 plots.
  - A bibliography, if needed.
    - The bibliography does NOT count towards the 6 page limit.
  - Not require any packages that require a pip/pip 3 install. We will be running the code in each notebook, so codes should only depend on the packages used in class and the standard packages in Jupyter.
- Each group member will complete the Group Work Activity Quiz to outline their contributions to the constructed model.
  - Groups members that do NOT complete the form will receive a grade of ZERO for their group contribution.
  - Group members may receive different grades depending on their active contributions to their group's work.