UNIVERSITY OF COLORADO - BOULDER
Robotics Program

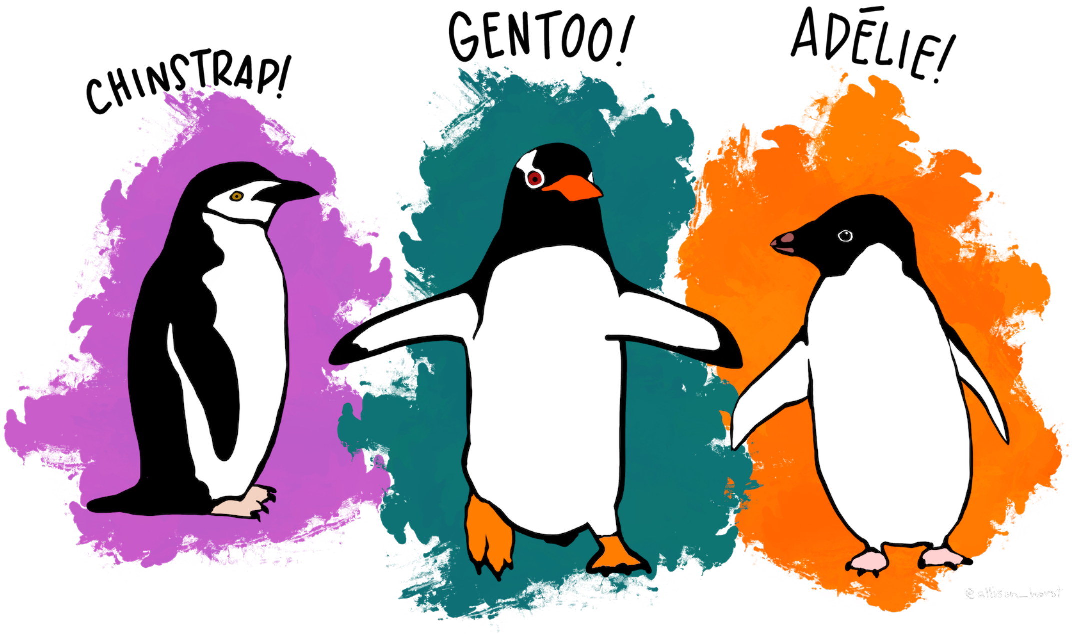
ROBO 5008 - Intro to Research

Assignment: Data & Figures (Assigned: 10/15, Due: 11/5 11:59pm on *Canvas*)
Data Analysis and Figure Generation

Instructions

What you turn in should be neatly prepared, and concisely address all parts of the assignment in a clearly organized manner. Be sure to explain your analysis and document all decisions (particularly for data cleaning). Code-related questions should be done in *Python*. Do not include in your final report, but rather include a link to a github repository **You are encouraged to discuss your work with other students, but submitted work must be your own. Please indicate on the first page of your answer sheet who your collaborators for this assignment are.**

All figures require axes labels (including units where applicable), legends where appropriate, and a clear title or caption. Points will be deducted if these elements are not included.



In this assignment, you will work with the Palmer Penguins dataset, which contains real-world biological measurements for three species of penguins observed in the Palmer Archipelago, Antarctica. The dataset contains both clean and raw versions - in this assignment, you will work from the raw data to practice realistic research workflows. You will clean the dataset, explore key patterns through visualizations, build a simple model, and produce publication-quality figures. This exercise is designed to help you gain experience in managing messy data, making informed analysis decisions, and effectively communicating results - essential skills for your research career.

1. **Dataset Access & Overview** Download the raw Palmer Penguins dataset:

   - Raw CSV: `https://raw.githubusercontent.com/allisonhorst/palmerpenguins/main/inst/extdata/penguins_raw.csv`

   The raw dataset contains measurements for three penguin species (Adelie, Gentoo and Chinstrap), with missing values, inconsistent categorical entries, and potential outliers.

2. **Data Cleaning (25%)** Prepare a cleaned version of the dataset. Your cleaning process should:

   - Identify and handle missing values. Support your approach to handling missing data in your report.
   - Standardize categorical variables (species, sex, island names).
   - Identify outliers and decide whether to remove, adjust, or retain them (justify your choice).

   Deliverables:

   - Cleaned CSV file.
   - Brief written description ($\leq$200 words) of all cleaning steps and your rationale.

3. **Exploratory Data Analysis (25%)** Explore relationships in the cleaned dataset:

   - Compare measurements (bill length, flipper length, body mass, etc.) by species and sex.
   - Explore correlations between numerical measurements.
   - Visualize distributions and relationships with appropriate plots.

   Deliverables:

   - 2–5 exploratory figures with captions. These figures need not be as polished as your final figures.
   - 2–3 sentences of interpretation for each figure.

4. **Basic Modeling (25%)** Choose **one** modeling path:

   - **Regression:** Predict body mass from bill length, flipper length, and sex.
   - **Classification:** Predict species from numerical measurements.

   Deliverables:

   - Model summary table (coefficients, accuracy, and/or R^2 as appropriate).
   - A paragraph discussing the modeling approach and potential shortcomings.
   - 1–2 sentence interpretation of results.

5. **Final Figures (25%)** Select two figures from your analysis that best communicate your findings. Refine them for publication-quality presentation (this may require post-processing in 3$^{\text{rd}}$ party software such as Adobe Illustrator or Affinity Designer):

    - Clear, consistent labeling and color scheme.
    - Informative captions.
    - Axes labeled with units where applicable.

    Deliverables:

    - Two final polished figures with captions.
    - Brief paragraph explaining why these figures were chosen, what data they portray and what your conclusions are from the figures.
    - 2–3 sentences on your approach and workflow to generate these final figures. Are there any improvements you envision for the future?

6. **Submission**

    - Single PDF containing: cleaning summary, data analysis, modeling, final figures. This PDF has to be generated using LaTeX. Note that this document is expected to have correct formatting suitable for a research publication. Pay close attention to items such as page numbers, figure axes, titles, etc.
    - The code you used to clean data, generate models and produce figures in the form of a GitHub link should be submitted. This should also include a copy of your cleaned CSV file.