

Project 2 Instructions

Please use the project template Quarto document to complete your project. The rendered document (as a pdf) *and* the raw Quarto file (as .qmd) must be submitted to Canvas by 11:00pm on Thurs., **Mar. 13, 2025**. These two documents will be graded jointly, so they must be consistent (as in, don't change the Quarto file without also updating the pdf document!).

All results presented *must* have corresponding code, and the code should be visible in the final generated pdf for ease of grading. **Any answers/results given without the corresponding R code that generated the result will be considered absent.** All code reported in your final project document should work properly. Please do not include any extraneous code or code which produces error messages. (Code which produces warnings is acceptable, as long as you understand what the warnings mean and explain this.)

For this project, you will be using a dataset about Himalayan expeditions, taken from the Himalayan Database, a compilation of records for all expeditions that have climbed in the Nepal Himalaya. The dataset members contains records for all individuals who participated in expeditions from 1905 through Spring 2019 to more than 465 significant peaks in Nepal.

Each record contains information including the name of the mountain (peak_name), the year of the expedition (year), the season (season), the age of the expedition member (age), their citizenship (citizenship), whether they used oxygen (oxygen_used), and whether they successfully summited the peak (success). More information about the dataset can be found at <https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-22/readme.md> and <https://www.himalayandatabase.com/>.

The project structure will be similar to Project 1. However, this time you will define the questions that you will then answer. Also, you will have to do some data wrangling in addition to data visualization. The final project should be structured as follows:

- Questions (2 specific questions you will answer)
- Introduction (1–2 paragraphs)
- Approach (2–3 paragraphs)
- Analysis (2–4 code blocks, 2 figures total, 1 for each question, text/code comments as needed)
- Discussion (1–3 paragraphs)

We encourage you to be concise. A paragraph should typically not be longer than 5 sentences. (But don't count sentences, if you need a few more that's fine!)

You are **not** required to perform any statistical tests in this project, but you may do so if you find it helpful to answer your question.

Instructions

First state the two questions you will answer. The questions should be conceptual and open-ended and not prompt a specific analysis. In particular, make sure you understand the difference between a question and an instruction.

This is a question: *Does the total number of attempted climbs change over time?*

This **is not** a question; it is an instruction: *Make a scatter plot of the total number of attempted climbs versus time.*

This is a question that prompts a specific analysis; it is actually an instruction pretending to be a question: *In a scatter plot of the total number of attempted climbs versus time, do the points fall onto a line with positive slope?*

Note that you are not permitted to use the question used here as example, whether the total number of attempted climbs changes over time.

In the Introduction section, write a brief introduction to the dataset, the questions, and what parts of the dataset are necessary to answer the questions. You may repeat some of the information about the dataset provided above, paraphrasing on your own terms. Imagine that your project is a standalone document and the grader has no prior knowledge of the dataset. You do not need to describe variables that are never used in your analysis.

In the Approach section, describe what type of data wrangling you will perform and what kind of plot you will generate to address your questions. For each plot, provide a clear explanation as to why this plot (e.g. boxplot, barplot, histogram, etc.) is best for providing the information you are asking about. (You can draw on the materials provided here for guidance.) **The two plots should be of different types, and at least one plot needs to use either color mapping or faceting or both.**

Across your two questions, your data wrangling code needs to use at least three different data manipulation functions that modify data tables, such as `mutate()`, `filter()`, `arrange()`, `select()`, `summarize()`, etc.

In the Analysis section, provide the code that performs required data wrangling and then generates your plots. You may find it helpful to compute and output summary tables in addition to making plots. Use scale functions to provide nice axis labels and guides. Also, use theme functions to customize the appearance of your plot. **For full points, you will have to apply some unique styling to your plots;** you cannot rely exclusively on preexisting theme functions. All plots must be made with `ggplot2`. Do not use base R plotting functions.

In the Discussion section, interpret the results of your analysis. Identify any trends revealed (or not revealed) by the plots. Speculate about why the data looks the way it does.