Project A Plan Evaluation

Thomas E. Love, Ph.D.

2023-02-06

Table of contents

There are nine elements (described below) that the teaching assistants are going to check in the Project A plans as they are submitted. Should a project pass at least eight of these nine elements, it will then go to Dr. Love for his approval (which is the tenth and final element.)

Students will have to redo their project A plan until they pass all 10 elements of this review. As we go, we will inform students which elements they have completed successfully, and which they haven't, as part of our feedback to them in response to each submission.

1 Element A: Submission Check

One student in each project group needs to submit the following to Canvas:

- a Quarto file
- their tidied data in an .Rds file
- an HTML file
- if they are working with a partner, one student has submitted all of this material, and the other has submitted a one-page note indicating that they are working with the partner, and that other person will submit the material.

1.1 Grading Element A

If all of this material is submitted on time, then the plan receives one point for Element A, and we'll move on to Element B.

If all of this material is not complete by the deadline, then the TAs should inform the student that the material is incomplete, and ask them to (re)submit as necessary until everything is complete. The TAs shouldn't grade the project or review the other Elements until Element A is OK.

2 Element B: Format is Acceptable

The HTML file should have:

- one unnumbered section called R Packages and Setup
- ten numbered sections, labeled as follows:

1. Data Source
2. The Subjects
3. Loading and Tidying the Data
4. The Tidy Tibble
5. The Code Book

6. Linear Regression Plans
7. Logistic Regression Plans
8. Affirmation
9. References
10. Session Information

There should also be a proper **title** that does not contain the words 432, Project, Proposal or Plan, and should be no more than 80 characters long, not including any subtitle.

The **author(s)** should also be specified clearly (with both first and last names) in the right place at the top of the HTML file, and there should be an appropriate **date** listed, as well.

The table of contents should be visible in the HTML, correct and complete, and the HTML document should include the appropriate set of code tools at the top to allow us to download the source file (.qmd) and to show or hide the code. All of this is just part of the template.

2.1 Grading Element B

If all 11 of these sections appear in the table of contents within the HTML file (and no other main sections appear and numbering of sections is correct), and an appropriate title, author and date are provided in the right places, then the plan receives one point for Element B, and we'll move on to Element C.

If the set of headings, or title/author(s)/date is incomplete or incorrect, or if there are any problems meeting the standards for Element B listed above, the TAs will specify the problem in our feedback, and require the project to be resubmitted with this fixed, but will proceed to grade as many of the other elements as possible in the current version.

3 Element C: Review of R Packages and Setup

The student should have a complete list of all R packages used in the work in the R Packages and Setup section. This means that the word `library()` should not appear in code after this section of the document, and that the student does not use `::` in the document after this section, either. (For example, if they want to use `favstats()`, then the `mosaic` package should be loaded here, and `favstats()` should be used rather than `mosaic::favstats()` in their analytic work.)

- The student should not load any of the core elements of the `tidyverse` or `tidymodels` separately, and should instead simply load those two meta-packages, and load those two meta-packages (in either order) after loading all other packages they plan to use.
  - The core packages in the `tidyverse` meta-package (that load automatically when one loads `tidyverse`) are:

    - ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, forcats
  - The core packages in the tidymodels meta-package (that load automatically when one loads tidymodels) are:
    - broom, rsample, parsnip, recipes, workflows, tune, yardstick, dials

- **Exception**: it's 100% fine if they load broom rather than tidymodels if that's the only package from the tidymodels meta-package that they plan to use.
- The student should load the rms package (which automatically loads Hmisc) but we won't worry about whether they also load Hmisc or not if they've loaded rms.
- The student's list of packages should also definitely include naniar and janitor.

The student should have used `#| message: false` as part of their code chunk where the packages are listed so that there are no HTML messages about when packages were built or how objects were masked.

3.1 Grading Element C

If the student has done all the things listed above for Element C, then the plan receives one point for Element C, and we'll move on to Element D.

If there is some problem, the TAs will specify the problem in feedback, and require the project to be resubmitted with this fixed, but will proceed to grade as many of the other elements as possible in the current version.

4 Element D: Review of Data Source and Subjects (Sections 1-2)

Section 1 should include complete information on the source of the data.

This includes descriptions (in complete, clear English sentences) of:

- how the student got the data including a clear link (with all necessary details) to the URL we can use to obtain the raw data freely,
- how the data were gathered and by whom and in what setting and using what sampling strategy (details here can be somewhat brief, and should cite the original source - which should also appear as part of the references in section 9.)

After reading what the student provides, the TA should be able to clearly understand the purpose for which the original data were gathered.

The TA should make sure that the data source (URL) provided allows them (the TA) to access the data.

Section 2 should include a description (we hope two sentences will be sufficient) of the subjects (rows) in the data set in sufficient detail that the TA feels they understand who the subjects are, and how the subjects were sampled from the population of interest.

4.1 Grading Element D

If the student has met all the standards listed above for Element D, then the plan receives one point for Element D, and we'll move on to Element E.

If there is some problem, the TAs will specify the problem in our feedback, and require the project to be resubmitted with this fixed, but will proceed to grade as many of the other elements as possible in the current version.

5 Element E: Review of Data Loading and Tidying (Section 3)

In this section, the student should start by ingesting the data and creating a tibble. They should then clean the data as described in the project A instructions.

There should be subsections to delineate the work in Section 3, although they don't absolutely have to use all of the subsections listed in the template under this section, which are:

- Loading the Data
- Cleaning the Data
  - Selecting Variables We'll Use
  - Changing Variable Names
  - Sampling the Data
  - Converting Variable Types
  - Working with Categorical Predictors
  - Arranging the Tibble

For example, the "Sampling the Data" subheading is only necessary if they actually need to sample the data down to 1200 rows. We do strongly prefer, though, that they use all of these subheadings that are relevant to their work, so that we can find things easily.

In addition, the TAs will specifically check the following six things:

- They should not filter to complete cases, but have missing values (if any exist in the raw data) at the end of their cleaning.
- They should wind up with a data set with no more than 1200 rows
- All multi-categorical variables should have no more than 6 categories.
- All variables should be converted to appropriate types (factors for predictors with multiple categories, character for the ID variable, etc.)

- There should be no warnings or error messages in this section (or, for that matter, anywhere in the document.)
- The tibble should not be listed, nor should there be any large strands of R output in this section.

5.1 Grading Element E

If the student has met all the standards listed above for Element E, then the plan receives one point for Element E, and we'll move on to Element F.

If there is some problem, the TAs will specify the problem in our feedback, and require the project to be resubmitted with this fixed, but will proceed to grade as many of the other elements as possible in the current version.

6 Element F: Review of Tidied Tibble and Codebook (Sections 4-5)

In Section 4, the student will present their tidied tibble. The TAs will check:

- that this is just a listing, not a glimpse or anything else - the command should just be the name of the tibble to print it, and
- that the tibble is a tibble (and not just a data frame), which prints the first 10 rows, and has appropriate dimensions (in terms of number of rows and number of columns).
- that a complete sentence in Section 4 specifies the number of rows and columns in the data
- that there is a sentence in Section 4 identifying the "identifying variable" and there is R code to prove that it is different for every row in the data set
- that the tibble is then saved as an .Rds file (probably using write_rds) and that the name of the file is the same as the file submitted to Canvas as the tidy data file.

In Section 5, the student will provide their codebook and the result of running `describe` from `Hmisc`

- The TAs will verify that the codebook includes each variable included in the tidy tibble from Section 4.
- The TAs will check that each variable has a role (identifier, outcome or input), a type (categorical - with # of levels, or quantitative), and a brief description that makes sense to the TA.
- The TAs will also check that each variable in the codebook and tibble also appears in the `describe` results, in the same order.

6.1 Grading Element F

If the student has met all the standards listed above for Element F, then the plan receives one point for Element F, and we'll move on to Element G.

If there is some problem, the TAs will specify the problem in our feedback, and require the project to be resubmitted with this fixed, but will proceed to grade as many of the other elements as possible in the current version.

7 Element G: Review of Linear Regression Plan (Section 6)

The TAs will verify that the student has presented:

- subsections labeled:
  - My First Research Question
  - My Quantitative Outcome
  - My Planned Predictors (Linear Model)
- A coherent and clear research question that ends with a question mark.
- A specific statement as to what the linear regression outcome variable will be, and how many rows in the data set have complete information on this outcome.
- A graph of the distribution of the outcome variable, with appropriate commentary on the nature of that distribution in at least one complete sentence.
- A demonstration using R code that the outcome variable has at least 10 different, ordered, observed values.
- A set of at least four clearly identified candidate predictors, where at least one is quantitative, and at least one is categorical with 3-6 categories.
- A description of the student's guesses as to the expected direction of the relationship between the outcome and each predictor individually.

7.1 Grading Element G

If the student has met all the standards listed above for Element G, then the plan receives one point for Element G, and we'll move on to Element H.

If there is some problem, the TAs will specify the problem in our feedback, and require the project to be resubmitted with this fixed, but will proceed to grade as many of the other elements as possible in the current version.

8 Element H: Review of Logistic Regression Plan (Section 7)

The TAs will verify that the student has presented:

- subsections labeled:
  - My Second Research Question
  - My Binary Outcome
  - My Planned Predictors (Logistic Model)
- A coherent and clear research question that ends with a question mark.
- A specific statement as to what the logistic regression outcome variable will be, and a count of the number of rows in the data with each of the two possible values of this outcome, with an indication as to missing values of this variable.
- A set of at least four clearly identified candidate predictors, where at least one is quantitative, and at least one is categorical with 3-6 categories. If they're re-using variables from the linear model, they need only to list them here. If there are new predictors here, they should be evaluated as we did in the previous element.
- A description of the student's guesses as to the expected direction of the relationship between the outcome and each predictor individually.

8.1 Grading Element H

If the student has met all the standards listed above for Element H, then the plan receives one point for Element H, and we'll move on to Element I.

If there is some problem, the TAs will specify the problem in our feedback, and require the project to be resubmitted with this fixed, but will proceed to grade as many of the other elements as possible in the current version.

9 Element I: Affirmation, References & Session Info (Sections 8-10)

- Check that the affirmation meets the specifications of the project, and ensure that the statement below appears, and that the TA agrees with it:

  I am certain that it is completely appropriate for these data to be shared with anyone, without any conditions. There are no concerns about privacy or security.

- There should be at least one reference (to the source of the data) in the References section.

- The session information should be in a section of its own, and be run using `xfun::session_info()`. Be sure you can identify the R version used (and inform the students of a problem if it's not at least R version 4.2.2.) and the system used (listed as `Running under:`)

9.1 Grading Element I

If there is some problem with Element I, the TAs will specify the problem in our feedback.

- If the TA has a concern about whether these data can be shared, they should make sure Dr. Love reviews that concern before feedback goes to the student, to see if he shares the concern.

If the student has met all the standards listed above for Element I, then the plan should receive the point for Element I.

After Reviewing All Nine Elements

Sum the points for Elements A-I (which should be a number between 1 and 9 since everyone will have a successful Element A)

- If the evaluation is now at either 8 or 9 points (as a result of meeting at least 8 of Elements A-I) the evaluation should then be forwarded to Dr. Love for his review.
- If the evaluation is below 8 points, the TAs will send the evaluation to Dr. Love and he will then inform the student of their status, with a deadline to revise and resubmit the materials with these elements fixed.

10 Element J: Dr. Love's Review

Once the TA evaluation is at either 8 or 9 points out of 9, Dr. Love will review the plan. If he is satisfied that the remaining adjustments are small, he will approve the project and give the student the following grade on the Project A plan:

- 20 points if the project is approved on the first try, after an on-time submission to Canvas of all materials.
- 18 points if the project is approved on the second try after an on-time initial submission, or after the first try in a late initial submission.
- 16 points if the project is approved after the second try following an on-time initial submission, or after the first try following a late initial submission.

If he's not yet satisfied, he will request a revision (again) due at a time he will specify.