# Rubric for Assessing 432 Project 1 Proposals

*Thomas E. Love*

*2019-02-09*

## Contents



## How Do We Grade The Proposals?

- Each project is evaluated with regard to ten separate parts, listed below. This includes the nine tasks specified in the Project Instructions for the Proposal, plus a tenth part (which is that all necessary materials - .Rmd, .HTML and .csv - are submitted to Canvas properly.)
- Students will receive 1 point for successful completion of each part of the proposal.
- Students receiving a grade lower than 10 will need to redo the problematic aspects of their proposal until they reach 10.
- Students will have 48 hours after the posting to Canvas of each redo request to resubmit their work addressing the stated concerns.
- The TAs will grade each proposal initially. Dr. Love will handle the redos personally, and will also review all proposals that receive a score of 10 initially.

## What is the Purpose of this Document?

There are two purposes.

1. Provide the teaching assistants at `431-help` with detailed instructions on how to evaluate each part of the proposal.
2. Provide the students in 432 with a clearer understanding of what they need to do to get their proposal approved.

# Evaluating Proposal Task 1 (Information on Data Source)

The instructions are:

> 1. Complete information on the source of the data: how did you get it, how was it gathered, by whom, in what setting, for what purpose, and using what sampling strategy.

Award the point for Task 1 only if the student (or students) has:

- specified the source of the data thoroughly (and provided a link if the data are available online)
- specified who gathered the data
- specified when the data were gathered
- specified the purpose for which the data were originally gathered
- provided some context as to the sampling strategy and study design (for instance, the data might be a survey, or the result of a case-control study, or the result of a randomized clinical trial)

all while using complete English sentences.

**Note**: Some students will have accessed data through a non-.csv data set, which they imported into R using the `haven` package or something similar. In this case, they may also include some initial "rawer" form of the data set, in which case, that must be explained as part of the explanation of the data source in Task 1. These students should provide a .csv of the tidied data after the iniital cleaning, as well, as part of their Task 3.

# Evaluating Proposal Task 2 (Code to load and tidy the data)

The instructions are:

> 2. Code to load the raw `.csv` file into a tibble, and tidy/clean up the data to be useful for your modeling work.

Award the point for Task 2 only if the student(s) have:

- provided code to read in the csv file as a tibble (usually with `read_csv`) **and**
- provided code which tidies the tibble with `filter` and `select` to include only the rows (subjects) and columns (variables) that they will work with in the project, so that there is no completely extraneous information left, **and**
- provided code which specifies the number of complete (non-missing) values in each column in the tibble after whatever tidying is complete. A common way to do this will be to use `colSums(!is.na(tibblename))` after substituting in the tibble name.
- stated in a complete English sentence that there are at least `XXX` complete observations (before any imputation) in each column of the tibble, where `XXX` is a number greater than 100 and no larger than 1000.

**Notes**:

- In the final version of the project, this code will likely need to be amplified to, for instance, create factor versions of some numerically coded categorical variables.
- This code must include whatever mutations are necessary to the data to identify the quantitative outcome proposed for linear regression modeling, and the binary outcome proposed for logistic regression modeling.

- Sanity checks and testing are an important part of verifying that the code does what you think it does. Include brief descriptions of whatever checks you do in this section of the proposal, in complete sentences between code chunks.
- For the proposal, the student should not impute, at all. That will come after the proposal.

## Evaluating Proposal Task 3 (Listing the tibble)

The instructions are:

> 3. A listing of the tibble, with all variables correctly imported (via your code) as the types of variables (factor/integer/numeric, etc.) that you need for modeling. Be sure that your listing specifies the number of rows and number of columns in your tidy data set.

Award the point for Task 3 only if

- an actual listing of the tibble is provided, and
- a sentence is provided that specifies the number of rows (observations) and the number of columns (variables) in the data, that accurately reflects what the tibble listing specifies, and
- there are a minimum of 100 and a maximum of 1000 rows in the data set, and
- there are a minimum of 7 and a maximum of 16 columns in the data set, and
- the left-most column in the tibble is a subject id code, which appears as a character variable. The best choice of name for this column for this proposal is either `subject` or `subj_id`.

**Notes**:

- A listing of the entire data set is not acceptable. This must be a tibble listing which includes the first 10 rows, only.
- The minimum size is 7 columns because that counts a subject ID code, a binary outcome, a quantitative outcome, and four predictors, which is the minimum size possible to complete the project.
- The reason that the maximum size is 16 columns is that the maximum number of potential predictors is 13 (and that's only for a data set with the full 1000 observations), to which we add a subject identifying code, and a binary outcome, and a quantitative outcome, making a total of 16 variables. Most students should, in fact, have considerably fewer than 16 variables in their data set.

## Evaluating Proposal Task 4 (Describing the Subjects)

The instructions are:

> 4. A description (one or two sentences) of who or what the subjects (rows) are in your data set.

Award the point for Task 4 only if:

- the student provides such a description, in a complete English sentence or two, and it makes sense to you.

## Evaluating Proposal Task 5 (The Code Book)

The instructions are:

> 5. A code book, which provides, for each variable in your tibble, the following information:
>
> - The name of the variable used in your tibble
> - The type of variable (binary, multi-categorical, quantitative)
> - The details for each variable

- if a categorical variable, what are the levels, and what % of subjects fall in each category
- if a quantitative variable, what is the range of the data, and what are the units of measurement
- if there are missing data, tell us how many observations are missing, and why, if you know why.

Award the point for Task 5 only if:

- an attractive Table 1 and follow-up complete English sentences as needed containing all of the required information is provided, and clear to you, and this includes the following:
- the student identifies the number of complete (or missing) observations associated with each variable, **and**
- there are no missing values in the `subject_id` variable and each value of `subject_id` is unique, **and**
- the student correctly identifies each variable as "quantitative", "binary", or multi-categorical" **and**
- if the variable is either "binary" or "multi-categorical" the student demonstrates that each category in the variable in question is endorsed by a minimum of 30 subjects, **and**
- at least two of the variables (not counting `subject_id`) are "binary", at least two are "quantitative" and at least one is "multi-categorical."

**Notes**:

- This is often the hardest point for students to earn, as there are several pieces here to keep in mind.
- Students also don't always accurately identify the type of variable - be sure to assess that carefully.
- If it's not absolutely clear what they're doing, provide comments to help them with their revision.
- Students can check that their `subject_id` values are unique by ensuring that `nrow(tibblename)` is equal to `n_distinct(tibblename$subject_id)`.

# Evaluating Proposal Task 6 (Describing the Variables)

The instructions are:

6. A sentence or two for each variable (column) providing a description of what the variable measures or describes, in English.

Award the point for Task 6 only if:

- the student provides such a description for each variable other than the `subject_id` code, in a complete English sentence or two, and it makes sense to you.

# Evaluating Proposal Task 7

The instructions are:

7. A sentence or two telling us what you will use your linear regression model to explain or predict, *followed by* a sentence or several telling us very precisely which (quantitative) variable will serve as your outcome in your linear regression model, and which four (or more) candidate predictors you intend to use for that model.

Award the point for Task 7 only if:

- the student provides such a description, in a complete English sentence or two, and it makes sense to you, and
- the student's proposed outcome variable is, in fact, a quantitative one, and
- the student's list of candidate predictors includes:
  - at least four predictors in total, **and**
  - at least one quantitative variable **and**
  - at least one multi-categorical variable (with three categories or more)

## Evaluating Proposal Task 8

The instructions are:

> 8. A sentence or two telling us what you will use your logistic regression model to explain or predict, *followed by* a sentence or several telling us very precisely which (binary) variable will serve as your outcome in your logistic regression model, and which four (or more) candidate predictors you intend to use for that model.

Award the point for Task 8 only if:

- the student provides such a description, in a complete English sentence or two, and it makes sense to you, and
- the student's proposed outcome variable is, in fact, a binary one, and
- the student's list of candidate predictors includes:
  - at least four predictors in total, **and**
  - at least one quantitative variable **and**
  - at least one multi-categorical variable (with three categories or more)

## Evaluating Proposal Task 9

The instructions are:

> 9. An affirmation that the data set meets all of the requirements specified here, most especially that the data can be shared freely over the internet, and that there is no protected information of any kind involved. You need to be able to write "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." If you are unsure whether this is true, select a different data set.

Award the point for Task 9 only if the student provides such an affirmation, and you have no concerns about this, either, based on their description of the study.

## Evaluating The Proposal: Part 10

Award the point for Part 10 only if

- the Canvas submission includes their R Markdown file, **and**
- there is a copy of the data set (this should be a .csv) used in building the R Markdown file posted to Canvas **and**
- the Canvas submission includes the knitted HTML result, produced with `code_folding = show` used, and produced without anywhere using `echo = FALSE`, so that all code in the entire document can be seen or hidden at the whim of the reader, **and**
- any graphs or tables are completely legible, and not (for instance) outside the size of the page, **and**
- an attempt at each of the nine tasks outlined above is part of the HTML, **and**
- the student has used the template provided, or something equivalent that maintains the same headings to facilitate finding the materials so you have no difficulties using the file.

**Note**: In general, no other information should be submitted, although some students may also include some initial "rawer" form of the data set, in which case, that must be explained within the proposal as part of the explanation of the data source in Task 1.

## Reporting the Results

Once the grading for a initial draft of a proposal is completed, the teaching assistants will:

- place the project title, and the student's grade regarding each of the ten points of the proposal (1 or 0) should be placed on the Google Sheet Dr. Love has made available to you, as well as providing any additional comments regarding specific issues that the student needs to address regarding any area where the student failed to get the point.

Once that has happened, Dr. Love will then:

- place the student's total score (0-10) on Canvas, and provide a comment in Canvas specifying either "Congratulations! Proposal Accepted." (for a score of 10) or "Redo requested on Parts X, Y, Z." (for a score of 0-9 - specifying as X, Y, and Z the Parts of the proposal need to be redone.)
- handle the grading of all revisions.

## Special Notes for Project Groups (rather than Individuals)

- Students working in groups of two will **each** need to submit something to Canvas.
  - One student will submit the proposal, and the other student will submit a single text document (Word is ideal) which states that their partner will be submitting their joint proposal.
  - Revisions (as needed) should continue to be submitted by the student who submits the initial proposal.