### ASSIGNMENT 1 -- You may want to read ALL the questions before you begin.

### NOTE: FOR ALL QUESTIONS BELOW, ONLY CONSIDER PROJECTS WITH

### non-missing "Circulation.Date" >= 2008-01-01.

### EXCLUDE ALL OTHER PROJECTS FROM YOUR ANALYSIS.

### YOU MUST provide a link to your R code. ------ DON'T FORGET TO DO THIS!!!!!!!!!!!!

# Take note of the column names: i.e., you can type: names(foo)

# fyi: the column called "Rating" is the success rating at completion. 0 = lowest, 3 = highest.

# (1) When projects are approved, they are approved for a certain period of time (until the time of

# "original completion date"). While projects are active, this "original" completion date is

# often pushed out (extended), and then there is a "revised" completion date.

# You have been told that project duration at approval is generally about

# 2 years (24 months). In other words, (purportedly) when projects are approved, the difference

# between the original project completion date and the the approval date is (supposedly)

# approximately 24 months.

# (a) Is this claim true? Explain. (Remember, for this ENTIRE assignment, only consider

# projects with Circulation.Date >= 2008-01-01. This will be your only reminder...)

# Has project duration at approval changed over time (consider projects circulated earlier

# and circulated later). Be sure to discuss mean durations, median durations, and the

# interquartile range of durations (using the "quantile" function).

# Approximate suggested length: 3-5 sentences

# (b) How does original planned project duration differ from actual duration (if actual duration is

# measured as the duration between "ApprovalDate" and "RevisedCompletionDate"?) Once again, use

# means, medians, and interquartile ranges to explain your results.

# Approximate suggested length: 3-5 sentences

# (2) What % of projects that have ratings were rated 0?

# What % were rated 1? What % were rated 2? What % were rated 3? Answer these questions using a table

# or a figure. Provide a title AND an explanatory sentence or two that provides the numerical % results

# rounded to the nearest percentage-point.

# (3) Repeat problem 2, but this time exclude all PPTA projects. PPTA projects are more prone to

# negative ratings, because after a certain point in time only the low-rated PPTA projects required

# ratings. PPTA stands for "Project Preparatory Technical Assistance" and it is basically a project

# intended to set up a loan (often a very large multi-million-dollar loan). Only PPTAs that fail to

# "eventuate" to a loan are rated, which is why they are usually rated negatively.

# (4) Identify the top 25% of projects by "Revised.Amount" and the bottom 25% of projects by

# "RevisedAmount". ("RevisedAmount" shows the final project budget.)

# Compare the ratings of these projects. Can you draw a causal conclusion about the effect of

# budget size on ratings? Why or why not?

# Hint: Compare the characteristics of the two project groupings,

# e.g., "Dept", "Division", "Cluster", "Country"

# Approximate suggested length: 3-5 sentences.

# (5) Imagine your manager asks you to apply Jeremy Howard's drivetrain model to the

# problem of optimal budget-setting to maximize project success (i.e., "Rating").

# In such a situation, what would be the:

# (a) decision problem or objective?

# (b) lever or levers?

# (c) ideal RCT design?

# (d) dependent variable(s) and independent variable(s) in the modeler

# (e) And---Why would running RCTs and modeling/optimizing over RCT results be preferable

# to using (observational, non-RCT) "foo" data?

# Approximate suggested length: 1-3 sentences for each sub-question.