Lecture 7 problem set

INSERT YOUR NAME HERE

Contents

Reading

Your required reading is the "Data Clearning Guidelines" document

- The goal of this document is to describe data investigation/cleaning standards for new research assistants who join our research team.
- This is a early beta/working draft of the document. We will be revising and adding sections in the future.
- Several examples, use NLS 72 postsecondary course transcript data, so we recommend reading this document prior to completing problem set
- Some topics in this document (e.g., acquiring data, merging data, reshaping data) have not yet been covered in class. you can skip/skim these sections.

Problem Set instructions

Overview

- Using the NLS72 course-level dataset, your assigmnent is to create the following GPA variables:
  - institution-level (i.e., transcript-level) GPA variable
  - term-level GPA variable
- Finally, we will ask you to provide one substantive recommendation for improving our beta "Data Cleaning Guidelines" document.

NLS72 Codebook and Supplemental Addendum

- https://github.com/ozanj/rclass/blob/master/data/nls72/NLS72_codebook.pdf
- https://github.com/ozanj/rclass/blob/master/data/nls72/NLS72_suppaddendum.pdf

General Instructions

- Don't make changes to "input" variables; instead, create a new variable(s)
- You are responsible for deciding what data investigations to conduct (e.g., conditional statements, frequency counts, etc.)
- Keep the data investigations you want me to see; though you might want to comment out very long lists of observations
- Reference the NLS72 codebook and supplemental addendum when needed. Document your rationale for data decisions via comments and provide reference page numbers from codebook or addendum when useful.
- Whenever you create a new variable, run checks to make sure variable created correctly (e.g., counts, cross-tabulations, assertions)
- As you work towards creating the gpa variable(s) you will create several new "input" variables
- Below, you will find some "header" instructions/hints in regards to important steps you should be completing in process of creating these gpa variables
- Whenever relevant, you can insert code you developed from the previous problem set as part of your answers for this problem set

Load libraries and data

Load libraries

```
#install.packages("tidyverse") #uncomment if you haven't installed these packaged
#install.packages("haven")
#install.packages("labelled")
library(tidyverse)
#> -- Attaching packages ---------------------------------------------------------------------------
#> v ggplot2 3.2.1     v purrr   0.3.2
#> v tibble  2.1.3     v dplyr   0.8.3
#> v tidyr   1.0.0     v stringr 1.4.0
#> v readr   1.3.1     v forcats 0.4.0
#> -- Conflicts ------------------------------------------------------------------------------------
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()    masks stats::lag()
library(haven)
library(labelled)
```

Open data, run the code chunk below

```
rm(list = ls()) # remove all objects
#getwd()
#list.files("../../../documents/rclass/data/nls72") # list files in directory w/ NLS data

#Read Stata data into R using read_data() function from haven package
nls_crs <- read_dta(file="https://github.com/ozanj/rclass/raw/master/data/nls72/nls72petscrs_v2.dta", e
```

Part I: Investigate data

First stage of creating an analysis dataset is conducting a thorough investigation of the "input" dataset(s) and an investigation of key variables. This part should include preliminary investigation of the data frame, one-way investigations of following key input variables: `transnum`, `termnum`, `crsecred`, `gradtype`, `crsgrada`, `crsgradb`, and investigations of the relationships between the following pairs of variables: (1) `gradtype` and `crsgrada`, (2) `gradtype` and `crsgradb`, (3) `crsecred` and `gradtype`.

Part II: Write out plan

Write a plan for how you will create your GPA variables

This plan should include your general conceptual definition for how to calculate GPA.

- The general definition of GPA is quality points (course credit multiplied by numerical grade value) divided by total credits.

- The plan should describe how you will apply this general definition to actual variables in the NLS course-level data

- The plan should also describe how you plan to deal with idiosyncracies in the value of "input" variables (e.g., missing values, strange values) and your rationale for treating the variable values this way.

- Note: you will almost certainly update this plan as you make progress.

Some guidelines/hints for creating gpa variable (several of these steps you did in previous problem set)

- You will have to create a new version of course credit called that is missing (`NA`) for strange values of `crsecred`
- You will have to create a new course grade variable that has numeric grade for each course
  - the primary input variables for will be `crsgrada`,`crsgradb`, `gradtype`, and your new course credit variable
  - Use this key to assign numeric values to letter grades from `crsgrada` - A+=4; A=4; A-=3.7; B+=3.3; B=3; B-=2.7; C+=2.3; C=2; C-=1.7; D+=1.3; D=1; D-=.7; F=0; E=0; WF=0
    - Note: WF refers to "Withdrawal with a failing grade"
    - Note: other letter grades will have missing values for numeric grade
      * your new course grade variable should be missing for observations where your new course credit variable equals `NA`
      * your new course grade variable should be missing if `gradtype` indicates that the grade is numeric (rather than letter) but the value of the numeric grade (`crsegradb`) is greater than `4`
- After you create the variable `numgrade` you may want to create a new course credit variable that is missing (`NA`) for observations where your new numeric course grade variable is missing
- For creating institution-level GPA variable, calculate institutional level quality points and total credit variables by summing across observations within `id` and `transnum`. Finally, divide the institutional level quality points by insitutional total credits to generate the institutional level GPA.
- Make sure that the denominator for your GPA variable (i.e., total course credits) excludes courses where course credit is non-missing but have missing values for the new numeric course grade variable you created

Write Your plan here:

Part III: Clean data

Clean data: create new versions of variables that will be inputs to your GPA variable

Prior to creating any new variable, you should be conducting investigations of the input variable (either here or in Part I). After creating any new variable, conduct investigations of the value of the new variable and check the value of the new variable against values of the input variable(s).

Part IV: Create institution-level GPA variable

Create institution-level GPA variable and save as a new object

After you create the gpa variable, conduct some basic investigations/descriptive statistics to check whether it looks reasonable

Part V: Create term-level GPA variable

Create term-level GPA variable and save as a new object

After you create the gpa variable, conduct some basic investigations/descriptive statistics to check whether it looks reasonable

Part VI: Provide one substantive recommendation for improving "Data Cleaning Guidelines" document

Finally, we will ask you to provide one substantive recommendation for improving our beta "Data Cleaning Guidelines" document. This could be a recommendation for revising content currently in the document or a recommendation for new content that should be added to the document. Think about what information would be helpful for new research assistants to know. Your answer need not be longer than a few sentences.

Once finished, knit to (pdf) and upload both .Rmd and pdf files *Remeber to use this naming convention "lastname_firstname_ps5"*