Lab 3 Instructions

BSTA 512/612

2024-02-29

⚠ IMPORTANT TO READ

- Please do not delete the rubric from your `.qmd` file. I will use it to circle the grades!
- There is an intructions file and a file for you to edit and turn in. Please only work in the latter file!!

```
# PLEASE DO NOT REMOVE THIS CODE CHUNK!!!
### ADD YOUR LIBRARIES HERE!!! ####

library(tidyverse)
library(gtsummary)
library(here)
if(!require(lubridate)) { install.packages("lubridate"); library(lubridate) }
load("~/Library/CloudStorage/OneDrive-OregonHealth&ScienceUniversity/Teaching/Datasets/Imp
```

Directions

Please turn in your `.html` file on Sakai. Please let me know if you greatly prefer to submit a physical copy.

You can download the `.qmd` file for this lab here.

This is the **instructions** file. The link above will take you to the **editing** file where you can add your work!! Please do not remove anything from the editing file!!

Purpose

The main purpose of this lab is to perform some quality control on our data, recode some of the multi-selection categorical variables, continue data exploration, and start analyzing the main relationship of our research question.

Grading

This lab is graded out of 12 points. Nicky will use the following rubric to assign grades.

Rubric

	4 points	3 points	2 points	1 point	0 points
Formatting	Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with no major grammatical nor spelling errors. With little editing, the answer can be incorporated into the project report.	Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with grammatical or spelling errors. With editing, the answer can be incorporated into the project report.	Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with major grammatical or spelling errors. With major editing, the answer can be incorporated into the project report.	Lab submitted on Sakai with `.html` file. Answers are bulletted or do not use complete sentences.	Lab *not* submitted on Sakai with `.html` file.

	4 points	3 points	2 points	1 point	0 points
Code/Work	All tasks are directly followed or answered. This includes all the needed code, in code chunks, with the requested output.	All tasks are directly followed or answered. This includes all the needed code, in code chunks, with the requested output. In a few tasks, the code syntax or output is not quite right.	Some tasks are directly followed or answered. This includes all the needed code, in code chunks, with the requested output.	Some tasks are directly followed or answered.This includes all the needed code, in code chunks, with the requested output. In a few tasks, the code syntax or output is not quite right.	More than a quarter of the tasks are not completed properly.
Reasoning*	Answers demonstrate understanding of research context and investigation of the data. Answers are thoughtful and can be easily integrated into the final report.	Answers demonstrate understanding of research context and investigation of the data. Answers are thoughtful, but lack the clarity needed to easily integrate into the final report.	Answers demonstrate some understanding of research context and investigation of the data. Answers are fairly thoughtful, but lack connection to the research.	Answers demonstrate some understanding of research context and investigation of the data. Answers seem rushed and with minimal thought.	Answers lack understanding of research context and investigation of the data. Answers seem rushed and without thought.

*Applies to questions with reasoning (like target population, choosing variables, revisiting research question)

Lab activities

0. Restate your research question

> ! Task
>
> Please restate your research question below using the following format. It's repetitive, but it helps me contextualize my feedback as I look through your lab.

How is implicit anti-fat bias, as measured by the IAT score, associated with "insert main independent variable here"?

1. Quality Control

There are a few more issues with the data that we need to look into. First, there is another coding for `NA` values in the race variable: `-999`. We will need to filter out these observations.

We will also need to look at individuals who have potentially answered the survey questions untruthfully. We cannot catch everything, but a good place to start is by looking at individuals who have done more than one of the following:

- selected the earliest or latest possible birth year
- selected the lowest or highest possible education
- selected all gender identities (for those using gender identity)
- selected all races (for those using multiple selection race)
- selected the lowest or highest weight (for those looking at BMI)
- selected the lowest or highest height (for those looking at BMI)

I want to take a second to mention that any of the above selections, and combinations of the above selections, are valid. However, we should start to flag the possibility that someone has **not** gone through the survey properly if we notice that most or all of the respondent's answers are the first answer choice, last answer choice, or selected all options. Additionally, not all of these carry the same importance in discerning validity. For example, a recorded age of 111 years old is the most striking to me. When paired with other selections that are the maximum or minimum (or first or last) option, then I will record it for future investigation. If this observation looks to be an outlier or high leverage point in our analysis, that is when I'll decide to remove it.

Tasks for 1.1

1. Filter out observations with a value of -999 in the race variable.
2. Glimpse at the observations that may indicate a respondent who has not properly completed the survey portion. This will require filtering **for** specific answer choices. Please see examples of `filter()` on it's documentation page.

1. Working with multi-selection variables

In the list of variables that we may choose to work with (in Lab 2), there are two that allowed respondents to select multiple categories. The two variables are `genderIdentitiy` and `raceombmulti`. **If you did not choose these variables to work with, you may skip this section.**

If you chose one or both of these variables, then we need to make new variables that correspond to indicators for each possible selection in the respective variable.

Let's start with the `grepl` function. For this function, we can input one of our column names and a value, then it will output, for each row, if the value is in the column. For example, in `genderIdentity` an individual may identify as a "Trans female/Trans woman" and "Gender queer/Gender nonconforming." In our dataset in R, this would show as `[4,5]` in `genderIdentity`. If we want to create two separate indicators for anyone who identifies as "Trans female/Trans woman" then I need to look for the value `4` in the column `genderIdentity`. I will run a separate indicator to find individuals who identify as "Gender queer/Gender nonconforming." Here is an example code of how I would use `grepl` to do this:

```r
iat_prep_new = iat_prep_old %>%
  mutate(ind_tf_tw = grepl(4, genderIdentity),
         ind_gq_gnc = grepl(5, genderIdentity))
```

You will need to extend this to all other gender identities.

For race, `raceombmulti` is also the follow up question to `raceomb_002`. So our indicators need to reflect both variables. In this case, we need to use `grepl` on both columns at once. For example, if I want to create an indicator for individuals who identify as American Indian/Alaskan Native then I need to find individuals who identify as American Indian/Alaskan Native only and individuals who identify as American Indian/Alaskan Native in addition to another race. For example, my code might look like:

```r
iat_prep_new = iat_prep_old %>%
  mutate(ind_AIAN = grepl(1, raceomb_002) | grepl(1, raceombmulti))
```

I suggest only searching for 1-7 in both `raceomb_002` and `raceombmulti`. Note that if `raceomb_002 = 8` , then individuals identified as “multiracial” and will select values in `raceombmulti`.

> ❗ Task for 1.2
>
> - If you are using `genderIdentity` or `raceombmulti`, create indicator variables for each possible selection.

2. Continuing data exploration

In this section

- Look at all other relationships between IAT score and each covariate.
  - For categorical variables, is there an inherent order? Does the ordered values follow a linear relationship? Are the categories evenly spaced? Think education - is there a natural place to divide the categories up?? multivariate data exploration

3. Make a Table 1

4. Fit the simple linear regression