Problem Set 2: Modules 2 & 3

YOUR NAME HERE

due INSERT DATE HERE

Structural stuff:
1. Be sure to change the "author" above to your name. Also insert the due date for this term/assignment.
2. Save your .Rmd file as LastName_FirstName.Rmd (do this before you knit).
3. You need to submit your .Rmd code file AND a knit file (upload both simultaneously to the course webpage; you can't upload them one-by-one). You will only receive full credit if you upload both files. 4. Below I have set up the file for you with the libraries you'll need. I have also inserted code chunks for you (note, I won't do this every time).
5. I expect that the .Rmd file you submit will run cleanly, and that the knit file won't contain any errors (LOOK at the knit file after you create it - if questions/text are running into each other, if you see error messages, etc., you're not done).
6. You can use comments to tell me what you are doing either in text or in code chunks, but remove "old" code that didn't run/work.

```
knitr::opts_chunk$set(echo = TRUE)

library(yardstick)
```

```
## For binary classification, the first factor level is assumed to be the event.
## Use the argument `event_level = "second"` to alter this as needed.
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x readr::spec()   masks yardstick::spec()
```

This assignment includes questions from Module 2 (conditional means) and Module 3 (descriptive plots). The questions are interspersed below - all plotting questions are from Module 3 and everything else is from Module 2.

For this research question, we're interested in understanding more about what predicts alumni earnings for a set of postsecondary institutions. We have a limited dataset (only 125 institutions) but we hope that

(eventually) we could use the information about potential predictors from this study to inform predictions about a larger set of colleges. This analysis serves as our exploration of the data and our outcome of interest (md_earn_wne_p6 - the median earnings of graduates 6 years post graduation).

1. Import/load the "sc_debt.Rdata" dataset here.
2. Plot the distribution of our outcome: `md_earn_wne_p6`. Make sure you choose an appropriate univariate plot.

2a. What does this distribution show you about the outcome?

3. Calculate the unconditional mean of the outcome: `md_earn_wne_p6`
4. Use your mean as a prediction (i.e., create a new variable that consists of the unconditional mean of the outcome and add it to your dataset).
5. Calculate the summary measure of the errors for each observation—the difference between your prediction and the outcome (hint: RMSE).

5a. Provide one sentence interpreting this RMSE.

6. Let's look at one potential predictor: public vs. private (`control`). We know that, on average, private schools tend to cost more. If a school costs more, we would hope its alumni eventually earn more, right? First, let's examine the distribution of this predictor. Plot this variable using the appropriate univariate plot.

6a. Describe the distribution of this potential predictor variable.

7. Calculate the mean of our outcome for each level of this predictor variable.

7a. What does the data tell you about this variable? Do you think it might be a good predictor? Why or why not?

8. Use these conditional means as a prediction for our outcome (i.e., add them to your dataset so that we can provide a "best guess" as to each college's level of the outcome).
9. Calculate a summary measure of the error in your predictions using this one predictor.

9a. Interpret this RMSE.

9b. Did your updated conditional/one-predictor model show an improvement over the unconditional model? How do you know?

10. Now, I want you to add in a second predictor of your choosing. For this question, list the variable and why you think it might be a good predictor for our outcome.
11. Plot your new predictor variable using the appropriate *univariate* plot (you're just plotting the new predictor alone here; nothing combined yet)

11a. Describe the distribution of this potential predictor variable.

11b. Does it seem like this might be a good predictor? How come? (Note: if your answer is no, then try a different variable, repeating the steps for questions 10-11a above until you get one that you think is a good one. For your answers here, just include the "good" one, but you can let me know others you tried).

12. Calculate the mean of our outcome for each level of our *combined* predictor variables.

12a. What does the data tell you about these variables? Do you think, together, they might give us good predictions? Why or why not?

13. Use these conditional means as a prediction for our outcome (i.e., add them to your dataset so that we can provide a "best guess" as to each college's level of the outcome).

14. Calculate a summary measure of the error in your predictions using these two predictors.

14a. What does this RMSE tell you?

14b. Did your updated two-predictor conditional model show an improvement over your last two models (unconditional and one-predictor)? How do you know?