Assignment 3

Your name and student ID

Today's date

- Due date: Friday, September 20 5:00pm.
- Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for assignments submitted thereafter.
- This assignment is marked out of 47. Marks are indicated for each question. There are 19 questions total.
- Submission process: Please submit your assignment *directly* to Gradescope (see the last page for more details). Do not remove any `\newpage` tags from this file.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on data hub. Alternatively, you may wish to view the code in the condensed PDFs posted on bCourses (under Files > Lectures). Good luck!
- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily. This will save you and the GSIs from frustration!
- If your code runs off the page of the knitted PDF then you will LOSE POINTS! To avoid this, have a look at your knitted PDF and ensure all the code fits in the file. When it doesn't, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

```
library(readr)
library(dplyr)
library(ggplot2)
library(broom)
library(forcats)
```

Predicting insurance charges by age and BMI

Problem: Medical insurance charges can vary according to the complexity of a procedure or condition that requires medical treatment. You are tasked with determining how these charges are associated with age, for patients who have a body mass index (bmi) in the "normal" range (bmi between 16 and 25) who are smokers.

Plan: You have chosen to use tools to examine relationships between two variables to address the problem. In particular, scatter plots and simple linear regression.

Data: You have access to the dataset `A03-Insure.csv`, a claims dataset from an insurance provider. It is in the data folder.

Analysis and Conclusion: In this assignment you will perform the analysis and make a conclusion to help answer the problem statement.

1. [1 mark] Please type one line of code below to import these data into R. Assign the data to `insure_data`. Execute the code by hitting the green arrow and ensure the data set has been saved by looking at the environment tab and viewing the data set by clicking the table icon to the right of its name.

```
# replace this line with your code.
```

2. [2 marks] Write the code for four functions to get to know your dataset. Execute these functions line by line so you can look at their output, and examine these data.

```
# replace this line with your code.
# replace this line with your code.
# replace this line with your code.
# replace this line with your code.
```

3.a) [1 mark] How many individuals are in the dataset?: Replace this text with your answer.

3.b) [1 mark] How many nominal variables are in the dataset? What are they?: Replace this text with your answer.

3.c) [1 mark] How many ordinal variables are in the dataset? What are they?: Replace this text with your answer.

3.d) [1 mark] How many continuous variables are in the dataset? What are they?: Replace this text with your answer.

3.e) [1 mark] How many discrete variables are in the dataset? What are they?: Replace this text with your answer.

Run the following code by removing the "#" symbol in front of each of the six lines and executing the code chunk. Remind yourself what the `mutate()` function does in general, and notice that a new function `case_when()` is also being used.

```
#insure_data <- insure_data %>%
#  mutate(bmi_cat = case_when(bmi < 16 ~ "Underweight",
#                             bmi >= 16 & bmi < 25 ~ "Normal",
#                             bmi >= 25 & bmi < 30 ~ "Overweight",
#                             bmi >= 30 ~ "Obese")
#                  )
```

4.a) [1 mark] What did the above code achieve?:

Replace this text with your answer.

4.b) [1 mark] What type of variable is `bmi_cat`:

Replace this text with your answer.

5.a) [1 mark] Read the Problem statement proposed at the beginning of this exercise. Who belongs to the population of interest?:

Replace this text with your answer.

5.b) [1 mark] Using a dplyr function make a new dataset called `insure_subset` containing the population of interest

```
# replace this line with your code.
```

6. [3 marks] Make a scatter plot of the relationship between `age` and insurance `charges` for the population of interest. Give your plot an informative title.

```
# replace this line with your code.
```

7.a) [2 marks] Run a linear regression model on the relationship between `age` and `charges`. Think about which variable is explanatory (X) and which is response (Y). Assign the regression model to the name `insure_mod`. Then type `tidy(insure_mod)` below the model's code and execute both lines.

```
# replace this line with your code.
# replace this line with your code.
```

7.b) [1 mark] Interpret the slope parameter:

Replace this text with your answer.

7.c) [1 mark] Interpret the intercept parameter:

Replace this text with your answer.

7.d) [1 mark] Does the intercept make sense in this context?:

Replace this text with your answer.

7.e) [1 mark] Add the line of best fit to your scatterplot by copying and pasting the plot's code from Question 6 into the chunk below and adding a `geom` that can be used to add a regression line:

```
# replace this line with your code.
```

8. [2 marks] What do you notice about the fit of the line in terms of the proportion of points above vs. below the line. Why do you think that is?:

Replace this text with your answer.

Run the following `filter()` function by removing the "#" symbol in front of the two lines of code and executing the code chunk.

```
#insure_smaller_subset <- insure_subset %>%
  #filter(charges < 30000 & ! (charges > 25000 & age == 20))
```

9. [2 marks] How many individuals were removed? Who were they?:

Replace this text with your answer.

10. [2 marks] Run a regression model on `insure_smaller_subset` between `charges` and `age`. Assign it to an informative name and look at the output using the `tidy()` function, as was done with the previous linear model.

```
# replace this line with your code.
# replace this line with your code.
```

11. [2 marks] Add the new regression line to your ggplot. Keep the older regression line on the plot for comparison. To distinguish them, change the color, line type, or line width of the newly-added line.

```
# replace this line with your code.
```

12. [2 marks] Calculate r-squared for both linear models using a function learned in class (Hint: Chapter 4 lecture).

```
# replace this line with your code.
# replace this line with your code.
```

13. [4 marks] Calculate correlation between `age` and `charges` using the subset (`insure_subset`) and the even smaller dataset (`insure_smaller_subset`). Also calculate correlation squared for both data frames. What is the relationship between the correlation and r-squared values that you calculated in the previous question?

```
# replace this line with your code.
# replace this line with your code.
```

Replace this text with your answer.

PART B: Your supervisor asks you to extend your analysis to consider other smokers with BMIs classified as overweight or obese. In particular, she wanted to know if the relationship between age and medical charges is different for different BMI groups. You can use data visualization coupled with your skills in linear regression to help answer this question.

14. [1 mark] Make a new dataset called `insure_sub_smoke` that includes smokers of any BMI.

```
# replace this line with your code.
```

15. [1 mark] Make a scatter plot that examines the relationship between `age` and `charges` separately for normal, overweight, and obese individuals. A `facet_` command may help you.

```
# replace this line with your code.
```

Is there something out of order with your plot you just made? The issue is that the plot is automatically displayed by listing the BMI categories alphabetically. Run the following code by removing the "#" symbol in front of the two lines of code and executing the code chunks:

```
#insure_subset_smokers <- insure_subset_smokers %>%
  #mutate(bmi_cat_ordered = forcats::fct_relevel(bmi_cat, "Normal", "Overweight", "Obese"))
```

16. [1 mark] Re-run your plot code, but this time using `bmi_cat_ordered`.

```
# replace this line with your code.
```

17. [3 marks] Run a separate linear model for each BMI group. To do this, you will need to subset your data into the three groups of interest first. Call your models `normal_mod`, `overweight_mod`, `obese_mod`. Use the `tidy()` function to display the output from each model.

```
# replace this line with your subset code.
# replace this line with your subset code.
# replace this line with your subset code.

# replace this line with your model code.
# replace this line with your model code.
# replace this line with your model code.

# replace this line with your tidy output code.
# replace this line with your tidy output code.
# replace this line with your tidy output code.
```

18. Use the models to predict medical charges for a 20-year old by weight category. You don't need an R function to make these predictions, just the output from the model. Show your work for each calculation and round to the nearest dollar.

a) [1 mark] among normal BMI group: Replace this text with your answer.

b) [1 mark] among overweight BMI group: Replace this text with your answer.

c) [1 mark] among obese BMI group: Replace this text with your answer.

19. [3 marks] In three sentences maximum, (1) comment on the direction of the association, (2) comment on how much the slopes vary across the BMI groups, and (3) how much the prediction for a 20-year old varies.

Replace this text with your answer (sentence 1). Replace this text with your answer (sentence 2). Replace this text with your answer (sentence 3).

Submission

Please submit your lab *directly* to Gradescope. You can do this by knitting your file and downloading the PDF to your computer. Then navigate to Gradescope.com to submit your assignment. Here is a tutorial if you need help: https://www.gradescope.com/get_started. Scroll down on that page to "For students:submitting homework".