STA130H1S – Fall 2022

Problem Set 2

() and STA130 Professors

Instructions

Complete the excercises in this .Rmd file and submit your .Rmd and .pdf output through Quercus on September 22 by 5:00 p.m. ET.

Question 1

The Week 1 Problem Set included the following code.

```
my_answers <- c(r1,r2,c1,c2)
square_answers <- c(10,-1,3,12)
```

For the first three questions below choose the correct answer from the following.

(A) A single value counting how many correct rows and columns you calculated.
(B) A numeric vector of the differences between the math square answers and your answers (should be all 0s if you got them all right).
(C) A character vector of 'TRUE' and 'FALSE', 'TRUE' for each answer that matches and 'FALSE' for any that don't.
(D) A logical vector of `TRUE` and `FALSE`, `TRUE` for each answer that matches and `FALSE` for any that don't.
(E) A single logical value `TRUE` or `FALSE`, `TRUE` if all the values match, `FALSE` if any of the values don't match.

a) Which of the above best describes what `my_answers == square_answers` is?

b) Which of the above best describes what `sum(my_answers == square_answers)` is?

c) Which of the above best describes what `all(my_answers == square_answers)` is?

d) What is the sequence of steps involved in getting the answer for `sum(c(TRUE,FALSE))`? What additional step is required to get the answer for `sum(my_answers == square_answers)`?

REPLACE THIS TEXT WITH YOUR ANSWER

Hints

- Your answer should include the word ***coercion***.
- The `sum` function works only on `numeric` data types and does not itself directly know anything about `logical` data types.

Question 2

The data for this question will be based on a sample of Superbowl ads.

- The data is stored in the file `superbowl_ads.csv` in the same directory as this file, and includes the following variables:

- `year` (double) Superbowl year
- `brand` (character) Brand for commercial
- `funny` (logical) Contains humor
- `show_product_quickly` (logical)) Shows product quickly
- `celebrity` (logical) Contains celebrity
- `danger` (logical) Contains danger
- `view_count` (double) Youtube view count
- `like_count` (double) Youtube like count
- `dislike_count` (double) Youtube dislike count
- `superbowl_ads_dot_com_url` (character) Superbowl ad URL

This data was posted on github by the data-oriented reporting outlet FiveThirtyEight and subsequently featured on Tidy Tuesday. For more information see the above links.

```
library(tidyverse) # Load the tidyverse functionality so it is available to use
superbowl <- read_csv("superbowl_ads.csv")
```

(a) Use the `glimpse()` function to view properties of the `superbowl` data set. How many rows and columns are there? How many observations does it include? How many variables are measured for each observation?

REPLACE THIS TEXT WITH YOUR ANSWER

(b) Create 3 histograms to explore the distribution of `view_count` at the same time: (i) one with 2 bins, (ii) one with 8 bins, and (iii) one with 50 bins; make sure to specify meaningful axis labels where appropriate. Which of these histograms is most appropriate to describe the distribution of the average number of players at the same time? Why? Write a few sentences describing the distribution based on the histogram you chose as most appropriate.

```
# You can put multiple plots in the same code chunk
```

```
# Or your can put different plots in separate code chunks
```

```
# Feel free to add or remove code chunks as desired
```

REPLACE THIS TEXT WITH YOUR ANSWER

(c) Construct two plots to visualize the distribution of `brand` and one of these other categorical variables: `funny`, `danger` or `celebrity` from the superbowl ads data and describe the distribution in 1-2 sentences; make sure to specify meaningful axis labels where appropriate. Hint: If you choose a categorical variable with many different categories, you may find it useful to use `coord_flip()` to flip the bars horizontally and/or change the options in the R code chunk to make the plot large (ex: {r, fig.height=15, fig.width=5}).

```
# One reason to use use different code chunks for different figures is
# to assign different figure aspect ratio controls to different figure
```

REPLACE THIS TEXT WITH YOUR ANSWER

(d) Construct a set of two boxplots showing visual summaries of the distribution of number of likes (`like_count`) for whether ads included a celebrity or not (`celebrity`); make sure to specify meaningful axis labels where appropriate. Write 3-4 sentences comparing these distributions.

```
# This should be a single plot, NOT TWO... boxplots can be put in the same plot!
```

REPLACE THIS TEXT WITH YOUR ANSWER

Question 3

The `births` data set is part of the `openintro` package. It consists of random sample of 100 births for babies in North Carolina where the mother was not a smoker and another 50 where the mother was a smoker. Type `?births` in the R console for more information about the data and to see the definition of each variable. The code below loads the required libraries for this question and provides a glimpse of the `births` data frame.

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
glimpse(births)
```

```
## Rows: 150
## Columns: 9
## $ f_age     <int> 31, 34, 36, 41, 42, 37, 35, 28, 22, 36, 27, 35, 25, 36, 27, ~
## $ m_age     <int> 30, 36, 35, 40, 37, 28, 35, 21, 20, 25, 19, 34, 19, 33, 27, ~
## $ weeks     <int> 39, 39, 40, 40, 40, 40, 28, 35, 32, 40, 32, 40, 41, 38, 39, ~
## $ premature <fct> full term, full term, full term, full term, full term, full ~
## $ visits    <int> 13, 5, 12, 13, NA, 12, 6, 9, 5, 13, 5, 15, 13, 10, 11, 13, 1~
## $ gained    <int> 1, 35, 29, 30, 10, 35, 29, 15, 40, 34, 32, 20, 47, 20, 5, 22~
## $ weight    <dbl> 6.88, 7.69, 8.88, 9.00, 7.94, 8.25, 1.63, 5.50, 2.69, 8.75, ~
## $ sex_baby  <fct> male, male, male, female, male, male, female, female, male, ~
## $ smoke     <fct> smoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker, smoker, ~
```

(a) Choose two categorical variables and plot the distribution of each one (in seperate plots). Identify whether each of these variables is a nominal or ordinal categorical variable. Write one or two sentences interpreting each plot.

REPLACE THIS TEXT WITH YOUR ANSWER

(b) Choose a quantitative variable and plot its distribution. Identify whether the variable you selected is continuous or discrete, and write 2-3 sentences describing the distribution.

REPLACE THIS TEXT WITH YOUR ANSWER

(c) Construct a plot that shows the relationship between birth weight (`weight`) and mother's smoking status (`smoke`); make sure to specify meaningful axis labels where appropriate.

```
# To figure out how to do this google "ggplot2 scatter plot", or check out
# - https://ggplot2.tidyverse.org/#usage
#   - https://ggplot2.tidyverse.org/#cheatsheet
#   - https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf
```

Additional Recommended Study Material

Did you finish quickly? Do you still have an unused course study time allocation? Would you like to cover this material a little bit more?

ggplot2

- Official Cheatsheet
  - Finding Answers
- Learning Resources
  - Official Usage
  - R4DS Textbook
  - DoSS Toolkit

Markdown

Markdown supports efficiency and productivity, and it's needed for our class.

- RStudio Markdown Cheatsheet
  - R4DS Introduction
  - RStudio Introduction
- Markdown Tutorial

For Reference Only: ***NOT a READING RECOMMENDATION***

- knitr Documentation
- .Rmd Documentation