Social Sciences Intro to Statistics

Pset 1: Due MONTH, DATE, YEAR at 11:59pm

Belle Lee

2024-06-15

Overview

Welcome to Social Sciences Introduction to Statistics using R. This problem set is intended to give you some practice becoming familiar with using R. In this problem set, we are asking you to: create an R project, render your file, load and investigate an R data frame that is stored on the web, and apply some basic functions to atomic vectors.

- Note: Change the values of the YAML header above to your name and the date.

Question 1: Creating an R project

Create an R project

- Create a folder where you want to save files associated with problem set 1. Let's call that folder "problemset1", but you can name it whatever you want.
  - For instance, it could be SSS » problem_sets » problemset1.
- In RStudio, click on "File" » "New Project" » "Existing Directory" » "Browse".
- Browse to find and select your problem set 1 folder.
- Click on "Create Project".
  - An R project file has the extension ".Rproj".
  - The name of the file should be "problemset1.Rproj", or whatever you named the folder.

Save this problemset1.Rmd file anywhere in the folder named problemset1.

- Use this naming convention "lastname_firstname_ps#" for your .qmd files (e.g. lee_belle_ps1.qmd).

  - If you want, you can change the name of this file to include your first and last name.

- Run the `getwd()` function and the `list.files()` function in the code chunk below.
- What is the output? Why?

```
getwd()
list.files()
```

ANSWER:

ANSWER KEY: The output shows "/Users/bellelee/Documents/SSS, Fall 2024/problem_sets/problemset1" since that is the working directory I'm currently in. The getwd() code asks r studio to get or show the working directory. The output for list.files() shows "lee_belle_ps1.qmd" "problemset1.Rproj" since those are the two files in working directory.

Question 2: Render to pdf

- At the top of this .qmd file, type in your first and last name in the appropriate place in the YAML header (e.g. "Belle Lee").
- in the date field of the YAML header, insert the date within quotations (any date format is fine).
- Now click the "Render" button near the top of your RStudio window (icon with blue arrow sign) or drop down "File" and select "Render Document".
  - Alternatively you can use the shortcut: **Cmd/Ctrl + Shift + k**.
  - *Note*: One goal of this assignment is to make sure you are able to render without running into errors.

Question 3: Load .Rdata directly with url and then investigate the data frame

1. Load the package(s) we will use today: tidyverse

If package not yet installed, then must install before you load. Install in "console" rather than .qmd file

- Generic syntax: `install.packages("package_name")`
- Install "tidyverse": `install.packages("tidyverse")`

Note: when we load package, name of package is not in quotes; but when we install package, name of package is in quotes:

- `install.packages("tidyverse")`
- `library(tidyverse)`

2. This question asks you to load a dataframe by specifying the `read_csv()‘` function.

- Url link for data frame: https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/netflix_data/Netflix%20Movies%20Dataset%20All.csv

Load the dataframe within this code chunk below.

```
# ANSWER KEY
#library(tidyverse)

#load netflix data
#netflix_data <- read_csv("https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/mai
```

3. Print the data frame `netflix_data` by typing its name.

```
# ANSWER KEY
#netflix_data
```

3. Use the `typeof()` function to investigate the type of data frame `netflix_data`.

```
# ANSWER KEY
#typeof(netflix_data)
```

4. Apply the `dim()` function to the data frame `netflix_data`. What does this output mean in your own words?

```
#ANSWER KEY
#dim(netflix_data)
```

ANSWER:

ANSWER KEY: The output of 246 and 22 means that our data frame netflix_data, contains 246 observations of 22 rows, variables, or number of items.

Question 4: Investigating variable type/structure

Show the steps to isolate all the productions from the year 2018

1. Show how we can see the year that productions were released

```
#ANSWER KEY
#str(netflix_data$RELEASE_YEAR)
```

2. Create frequency table to identify possible values of `RELEASE_YEAR`

```
#ANSWER KEY
#table(netflix_data$RELEASE_YEAR, useNA = "always")
```

3. Isolate all production from 2018 (output omitted)

```
#ANSWER KEY
#filter(netflix_data, RELEASE_YEAR == "2018")
```

Question 5: How to order data

- `arrange()` sorts in **ascending** order by default
- use `desc()` to sort a column by descending order

1. Order our 'netflix_data' data frame by descending order based on number of votes. What is the title with the third highest number of votes and how many votes did they receive?

```
#ANSWER KEY
#arrange(netflix_data, desc(NUMBER_OF_VOTES))
```

ANSWER:

ANSWER KEY: The Walking Dead with 945,125 votes.

2. Now let's sort our data frame with multiple variables. Sort by ascending by IMDb score and descending by votes and by title; combine with select

```
# ANSWER KEY
#select(arrange(netflix_data, SCORE, desc(NUMBER_OF_VOTES), TITLE), TITLE, SCORE, NUMBER_OF_
```

Question 6: Pipes

Do task with and without pipes

1. Count the number of productions with 7.2 score that were produced in Great Britain

Without pipes

```
#ANSWER KEY
#count(filter(netflix_data, SCORE == "8.5", MAIN_PRODUCTION == "GB"))
```

With pipes

```
#ANSWER KEY
#netflix_data %>% filter(SCORE == "8.5", MAIN_PRODUCTION == "GB") %>% count()
```

2. Using `netflix_data` select the following variables (TITLE, RELEASE_YEAR, SCORE) and assign `<-` them to object `netflix_data_temp`.

```
#ANSWER KEY
#netflix_data_temp <- netflix_data %>%
  #select(TITLE, RELEASE_YEAR, SCORE, MAIN_PRODUCTION)
```

3. Using the object you just created `netflix_data_temp`, create a frequency table of `SCORE` for productions from the U.S. from 2020.

```
#ANSWER KEY
#netflix_data_temp %>%
  #filter(MAIN_PRODUCTION == "US", RELEASE_YEAR == "2020") %>% count(SCORE)
```

Question 7: Scatterplot

Using the `els_parphd` dataset, create a scatterplot of the relationship between hours/week spent on homework (`hw_time`) on the x-axis and 2011 earnings (`earn2011`) on the y-axis, with the color of points determined by high school region (`byregion`)

```
#ANSWER KEY
load(url("https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/els/output_
#Students might need to add libraries if they were not loaded earlier in the pset

#els <- df_els_stu_allobs_fac
#els_parphd <- els_v2 %>% filter(f1pared=="Completed PhD, MD, other advanced degree")
#ggplot(data= els_parphd, aes(x = hw_time, y = earn2011, color = byregion)) + geom_point()

#Students may choose to add better labels for the x and y axis or title the graph. Wouldn't c
```

Question 8: Bar Chart

Using the `els_v2` dataset, create a bar chart with the variable "ever attended postsecondary education" (`f2evratt`) as x-axis and number of students as y-axis

- Hint: Essentially, you are being asked to create a bar chart from the following frequency count:

```
#ANSWER KEY
#ggplot(data = els_v2, aes(x = f2evratt)) + geom_bar()
```

Render to pdf and submit problem set

Render to pdf by clicking the "Render" button near the top of your RStudio window (icon with blue arrow) or drop down "File" and select "Render to PDF"

- Go to the [class website] (Need to fill in classwebsite) and under the "Readings & Assignments" » "Week 1" tab, click on the "Problem set 1 submission link"
- Submit both .qmd and pdf files
- Use this naming convention "lastname_firstname_ps#" for your .Rmd and pdf files (e.g. lee_belle_ps1.qmd & lee_belle_ps1.pdf)