LA-4: Data Wrangling (15 points)

Learning Outcomes

In this assignment, you will learn how to:

- Read/Load data into R (in LA-3, you learned how to enter data into R manually).
- Use the pipe operator, `|>`.
- Filter cases from a data frame using the `filter()` function.
- Summarize variables using the `summarise()` function.

> **Tip**
>
> Read all the instructions carefully before starting the assignment. Set up your R script as you did in LA-3.

Instructions

1) Set up your R script for this assignment (name, uNID, date, assignment number). Install (if necessary) and load the packages below.

- `tidyverse`
- `summarytools`
- `rstatix`

2) Download the data file called `broadway.csv` from Canvas and upload it to your project for this assignment on Posit Cloud. If you are using R on your own machine, be sure you know to location to which the data file is downloaded. Create a folder on your Desktop called `3710` and move your data file into this folder. You should also save your .R files into this folder for your assignments.

3) Read the data into an object in R using the `read_csv()` function. For example, you might call your object, which will be a data frame, `bway` or `b`. Feel free to be creative with your object name; keep in mind that you will likely have to type the name of the data frame in your code many times. To read the data into R, use the following code (note there are differences in the code depending on whether you are working on Posit Cloud online or RStudio on your own laptop):

```
# On your own laptop:
# install.packages("here") # you will only need to install this package if you
# are working from your personal laptop

bway <- read_csv(here::here("Desktop", "3710","bway.csv"))

# On Posit Cloud

bway <- read_csv("broadway.csv")
```

4) Let's get a sense of our data using the `glimpse()` function. This function, which is included in the `tidyverse` suite of packages for R, allows us to get a glimpse of our data. To use this function, the command is: `glimpse(df)` where `df` is the name of your data frame.

5) Using the results of `glimpse()` in your Console, answer the following questions (as comments in your R script).

   a) How many cases (hint: rows) are there in the broadway dataset?
   b) How many variables (hint: columns) are there in the broadway dataset?

6) Now that we know the dimensions of our data frame (i.e., rows/cases, columns/variables), let's get a little more familiar with these data. Download the **codebook** from Canvas (.csv file) and take a look at the names and descriptions of the variables. In your R script, answer the following questions:

   a) What does the variable, `Statistics.Performances`, measure?
   b) What is the name of the variable that describes the maximum amount that a show can earn?

7) In the next step, we will learn to use the pipe operator. The pipe operator, `|>`, is used to perform sequential functions in R. It is part of the package, `magrittr` and also included in the `tidyverse` package. You can also think of the pipe operator, `|>`, as "then." If we were to use `|>` to describe a daily routine, it might look something like this:

```
Woke up |>
Took a shower |>
Got dressed |>
Made breakfast |>
Ate breakfast |>
Went to class
```

   Now, we will work on an example using the broadway data. We want to figure out the average number of people who attended the show for *Mamma Mia!* and *The Lion King*. Answer the question below as comments in your R script (be sure to label this clearly in the script and use `#` to include your answers as comments in your script).

   a) What is the name of the variable that tells us the name of the show?
   b) What does the variable, `Date.Month`, measure and what is the format of the data in this variable?
   c) What is the name of the variable that measures the attendance per show?

   Next, write the steps (in plain language as comments, i.e., pseudocode) to determine the average attendance per show for *Mamma Mia!* and *The Lion King*, respectively. An example of pseudocode is shown below.

```
# Start with the broadway data frame, which I called bway
# Filter for cases that only contain Mamma Mia! from the broadway data using
# the Show.Name variable
# Calculate the mean of the attendance among these cases
```

   Translate the pseudocode to R functions using the pipe operator and calculate the mean attendance for *Mamma Mia!*.

```
bway |> # Start with the broadway data frame
        filter(Show.Name == "Mamma Mia!") |> # Filter for only Mamma Mia! cases
        summarise(M_attd = mean(Statistics.Attendance, na.rm = TRUE)) # Find mean

# Note that na.rm = TRUE removes NAs (i.e., missing values) from the calculation
# of the mean, if there are any in your dataset. We don't have those in this
# dataset, though.
```

8) Let's say we also want to know the minimum and maximum number of people who attended a Mamma Mia! show. To do so, add arguments to the `summarise()` function.

```r
bway |> # Start with the broadway data frame
        filter(Show.Name == "Mamma Mia!") |> # Filter for only Mamma Mia! cases
        summarise(M_attd = mean(Statistics.Attendance, na.rm = TRUE), # Find mean
                  min_attd = min(Statistics.Attendance, na.rm = TRUE), # Find minimum
                  max_attd = max(Statistics.Attendance, na.rm = TRUE)) # Find maximum}
```

Using the calculations performed by the R commands above, answer the following questions:

a) What is the minimum number of people who attended Mamma Mia!?
b) What is the maximum number of people who attended Mamma Mia!?
c) What is the mean number of people who attended Mamma Mia!?

> **Note**
>
> There is a function that is part of the package, `summarytools`, that will enable you to obtain these statistics with a single function. For **2 bonus points**, try to figure out this function and use it in this exercise.

9) Use the same logic and process from the example above to answer the following questions.

   a) What is the minimum number of people who attended The Lion King?
   b) What is the maximum number of people who attended The Lion King?
   c) What is the mean number of people who attended The Lion King?
   d) Compare the statistics you just calculate for *Mamma Mia!* and *The Lion King*. Overall, which show would you say is more successful?

10) Using `|>` and the functions you just learned, write R code to determine the answers to the questions below. Round your answer to the nearest whole dollar.

   a) Among shows in the `Majestic` theater that are `Musicals`, what is the minimum, maximum, and mean "Gross Gross" in dollars?
   b) Among shows in the `Vivian Beaumont` theater that are `Musicals`, what is the minimum, maximum, and mean "Gross Gross" in dollars?
   c) Comparing the statistics you just calculated, is the Vivian Beaumont or Majestic theatre more successful?

Submission

Submit your R script (named `LA-#_FirstName-LastName.R`) to Canvas.

Your R script should:

1) Include commands and functions that are necessary to address all the questions in the assignment.
2) Contain comments that answer the questions in the assignment.
3) Run in its entirety without errors.

To ensure that your R script runs without errors, you should:

- Save your script.
- Navigate back to Your Workspace on Posit Cloud.
- Reopen your project.
- Run the entire script line-by-line without editing it to ensure there are no errors.

! Important

These standards apply to all submissions in this course that require R scripts. You should follow these instructions for preparation, naming, and saving of your R script for all of your individual lab assignments.