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, `|>`.
- Select variables from a data frame using the `select()` function.
- 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.

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.

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 are there in the broadway dataset?

   b) How many variables 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 this 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 (we will use the variable, `Show.Name`). Answer the question below in a comment in your R script (be sure to label this clearly in the script); step-by-step instructions to help you answer this question are shown below.

```
a) What does the variable, `Date.Month`, measure and what is the format of the data in this variable?

b) What is the name of the variable that measures the attendance per show?
```

In your R script, write pseudocode to determine the average attendance per show for *Mamma Mia!* and *The Lion King* respectively (remember that lines that begin with `#` are comments in R):

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

Next, translate the pseudocode to R functions using the pipe operator:

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

```
# A tibble: 1 x 1
   Mattd
   <dbl>
1 10436.
```

```
# 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.
```

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.