LA4: Data Wrangling

```r
library(tidyverse)
library(summarytools)
library(rstatix)
library(kableExtra)
library(formatR)

# Read broadway data into R
b <- read_csv("broadway.csv")
```

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

1) Read all the instructions carefully.
2) Set up your R script as you did in LA3.
3) Include a comment for each line of code in your R script. This can be pseudocode or other comments.

Set up your R script for this assignment. Load the packages below and set your working directory. Remember to write pseudocode as comments for each line of code.

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

Hint

See LA3 for a reminder of how to set up your R script for assignments.

Step 2

Instructions

Download the broadway dataset and save it into your working directory.

Read the broadyway data into an object (a data frame) in R. Give the data frame a name of your choice; feel free to be creative. View the broadway data set in the Console and answer the questions below.

You should also take this opportunity to review the codebook.

a) How many cases are there in the broadway dataset?
b) How many variables are there in the broadway dataset?

Hint

Read this section of COMM 3710: Getting Started with R.

You can also open the documentation for the function, `read_csv()` in R. To do so, type `?read_csv()` into the Console.

Codebook

Table 1: Names and descriptions of variables in broadway dataset.

Variable	Description
Date.Day	The day of the month that this performance's week ended on.
Date.Full	The full date representation that this performance's week ended on in Month/Day/Year format.
Date.Month	The numeric month that this performance's week ended in (1 = January, 2 = February, etc.).
Date.Year	The year that this week of performances occurred in.
Show.Name	The name of the production.
Show.Theater	The name of the theater.
Show.Type	Whether it is a Musical, Play, or Special.
Statistics.Attendance	The total number of people who attended performances over the week.
Statistics.Capacity	The percentage of the theater that was filled during that week.
Statistics.Gross	The 'Gross Gross' of this performance, or how much it made in total across the entire week. Meas
Statistics.Gross Potential	The Gross Potential is the maximum amount an engagement can possibly earn.
Statistics.Performances	The number of performances that occurred this week.

Step 3

Instructions

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 the university
```

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 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 respectivly (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:

```
b %>% # Start with the broadway data frame
        filter(Show.Name == "Mamma Mia!") %>% # Select 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
```

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:

```
b %>% # Start with the broadway data frame
        filter(Show.Name == "Mamma Mia!") %>% # Select 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:

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

Hint

The answer to 3a can be found in the codebook. Answers to 3b, 3c, and 3d can be found in the Console panel in R once you have worked through the example.

Step 4

Use the same logic and process from Step 3 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?

Step 5

Instructions

Using `%>%`and the functions you just learned, write R code to determine the answers to the questions below.

a) Among shows in the **Studio 54 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?

Hint

Start by identifying the variables (which variable will you use to filter or select cases? Which variable will be used in the `mean()` function) that you will use in this code chunk. Then, write pseudocode detailing the steps you need to take to arrive at answers for a and b. Finally, write your R commands using the pipe operator. Remember that each line of code must have a comment.

If you need help with the `mean()` function, pull up the corresponding help documentation.

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.