**R Mini Project 4: R Script Covering Outliers and Graphs**

Data Exploration 5101 - Fall 2024

**Due**: Week 12, November 21st, 2024

**Background:**

As the data scientist, you have been asked to create a report on food-borne illness outbreaks related to Norovirus and Escherichia coli (*E. coli*). You are given the dataset in Excel format, NORS\_20241106.xlsx (Available on Blackboard),

This project uses data from the National Outbreak Reporting System (NORS). NORS is a web-based platform designed to support reporting to the US Centers for Disease Control and Prevention by local, state, and territorial health departments in the United States of all waterborne disease outbreaks and enteric disease outbreaks transmitted by food, contact with environmental sources, infected persons or animals, or unknown modes of transmission. A dashboard is available from the CDC at <https://www.cdc.gov/ncezid/dfwed/beam-dashboard.html> and the data was downloaded from <https://data.cdc.gov/Foodborne-Waterborne-and-Related-Diseases/NORS/5xkq-dg7x/about_data> on November 6, 2024.

If you do not like food-borne diseases (come on! No one wants to get sick from a delicious cookie!), you can pretend the data are on anything. This application applies to any situation where you are asked to explore a dataset with quantitative measurements on three different types of items. In this specific case, it is enteric illnesses.

**Instructions:**

Write an R script with inline comments that successfully:

* Write your name, date, and assignment at the top
* (1 point) Loads all necessary libraries, including readxl (but you need more than that one library)
  + library(readxl)
* (1 point) Reads in the data from the Excel spreadsheet
  + illness ← read\_excel('NORS\_20241106.xlsx', sheet = 'NORS\_20241106')
* (1 point) Describes the data using:
  + At a minimum – glimpse() and summary() **and** document anything you think significant from the results.
  + Any other way that helps you understand more about the data to explore are encouraged.
* (14 points) Explores and cleans the data:
  + (1 point) Check for NAs or blanks. There are many NAs but do you need to take appropriate action to address the missing values? Document why there are the NAs to explain why you believe they are there.
  + (1 points) The dates are problematic, as we have seen previously. There are separate columns for year and month. Create a new column called Date that is a datetime field using the year, month, and assume the day of the month is 1 (or whatever day of the month you like).
  + (1 points) Select all the rows which have “Norovirus” stated in any way in the Etiology. Note, there is “Norovirus”, “Norovirus unknown”, “Norovirus Genogroup I” and many other instances. You want **any** row that has “Norovirus” in the Etiology.
  + (2 points) Present statistics on Norovirus that you think would be of interest to your project’s sponsor. Suggestions are how many Norovirus outbreaks per state with how many sickened, how many per year with how many sickened, what is the most common Primary Mode based on number sickened, most common Food Vehicle based on number sickened, or whatever you think important.
  + (2 points) To support your statistics on Norovirus, create a graph using ggplot() (adhere to good visualization practices).
  + (1 points) Select all the rows with “Escherichia” stated in any form in the Etiology field. Determine how to handle instances where multiple etiologies are listed and document what you do (or do not do) about it.
  + (2 points) Similarly, present statistics that you think would be of interest on Escherichia coli.
  + (2 points) To support your *E. coli* statistics, present a graph of your selection using ggplot() (adhere to good visualization practices).
  + (2 points) Select at least one statistic and present a graph showing both Escherichia and Norovirus together. Examples might be the increase of sickened by illnesses over time? Or the comparison by state. Whatever you think is of interest, and document why you are presenting the graph you select.
* (3 points) A project summary (At least 1 page)
  + Submitted with your script as .rmd, or .pdf.
  + Avoid writing in the first person (~~First I did this, then I did this, etc.~~)
  + Focus on presenting your results.
    - Include your graphs adhering to good visualization practices
      * Title, axis titles, captions, and legend as needed.

Partial credit is given for code that does not work IF you include a comment on what you were trying to do! On the other hand, significant points are deducted on code that works IF there is no comment to explain what the attempted code does. Also, since data exploration requires plenty of critical thinking, we want you to explore on your own. A script with just the minimum may get 1 point instead of 2 points for any item. To receive the full credit of 20 points, you should treat this project as a true data exploration and explore various options to present a quality “report.”

It’s important to remember that your R script may look different than a classmate’s script. There are multiple different ways to complete the assignment with the same outcome.

The deliverables are R scripts saved with the name R4\_ YourLastName**.r** or R4\_ YourLastName**.Rmd** submitted to Blackboard for Week 12, November 21st. This mini project is **20%** of the course grade.