# BIOSTAT 707 Homework 1

*In this homework, the objectives are to*

1. Run bash commands in RMarkdown to answer questions.
2. Work with dates in R.
3. Learn to perform exploratory data analysis and visualization
4. Practice data preprocessing and transformation

**Assignments will only be accepted in electronic format in RMarkdown (.rmd) files and HTML files.** Commented versions of any code to produce analyses will be required. RMarkdown and html homework files should be uploaded to Sakai with the naming convention date_lastname_firstname_HW[X].Rmd. For example, Dr. Dunn's first homework assignment would be named 20210831_Dunn_Jessilyn_HW1.Rmd and 20210831_Dunn_Jessilyn_HW1.html. When you submit to Gradescope, please convert the html file to pdf. It is important to note that **5 points** will be deducted for every assignment that is named improperly.

```
library(tidyverse) # R package for data science in an easy to use way
library(ggplot2) # # R package for visualizing your data
library(lubridate) # package for working with dates in R more easily
library(patchwork) # package to simplify combining separate ggplots
library(gridExtra) # package to arrange multiple grid-based plots on a page
library(psych) # multivariate analysis and scale construction using factor analysis, pri
ncipal component analysis, cluster analysis and reliability analysis
library(corrplot) # provides a visual exploratory tool on correlation matrix
```

# Shell Scripting

Note: throughout this homework, we refer to the terms "variables", "features", and "predictors" interchangeably.

Working in R is sometimes limited by the amount of data that R can load and allow the user to efficiently work with. Hence, sometimes we interact with datasets using shell scripting, or bash commands, especially when the datasets are too large to work efficiently in R.

To receive full credits for each of the questions, you should copy the shell command you used to achieve the answer and paste it here in this RMD file and also the printed answer from the shell terminal after each question.

We will work with a dataset that comes from this Our World in Data (https://ourworldindata.org/coronavirus-source-data). It shows many metrics such as the new case count, the total death count, and death count per million population from every region and country since the end of last year.

To complete this homework, you will need to access the Duke Compute Cluster (DCC) and use the shell environment directly available from there. Follow the these steps to access DCC:

- Connect to DukeVPN following the instructions in this website: https://oit.duke.edu/what-we-do/services/vpn (https://oit.duke.edu/what-we-do/services/vpn)
- Log into Duke Compute Cluster (DCC) by typing below in your terminal:

1. Upload the folder named “owid-covid-data.csv.gz” that you can download from the Sakai HW1 folder to your DCC home directory.

2. Look at the first 5 lines and last 5 lines of the data file owid-covid-data.csv.gz without unzipping it (This is not a very large dataset per se, but we use it to simulate a working with a very large dataset that would be difficult or even impossible to open due to its size). Find out what “gzcat” does by running “man gzcat” and use “gzcat” to complete this task.

3. How many lines are there in this file? Find out without unzipping the file.

4. Unzip owid-covid-data.csv.gz without deleting the original zipped file. Use “man gunzip” to find out how to do that.

5. Find out how many days of COVID-19 data *Italy* has in this dataset. Use “grep” to find the lines that contain the work “Italy” and count the lines.

6. Look at “owid_covid_codebook.csv”, which is a data dictionary for “owid_covid_data.csv”. This tells us what each of the fields (columns) of the dataset means. Each row is a COVID-19 case count entry. How many columns and how many rows are there in this dataset? Find out which column indicates the country from the codebook. Use bash command to determine how many unique countries there are data in this file. I recommend using “awk” to parse the file.

7. Use “awk” to filter by country “iso_code” so that we are only looking at COVID data from the USA. Find out the date that has the largest number of single day new cases in the USA using “sort”.

# Working with Dates and EDA

8. Unzip the file named “owid-covid-data.csv.gz” and load this dataset into this R Markdown file using **read.csv()**. Select the following variables from the original data file and save just these fields into a dataframe named *covid_df*:

iso_code continent location date population total_cases new_cases total_deaths new_deaths total_cases_per_million new_cases_per_million total_deaths_per_million new_deaths_per_million positive_rate total_vaccinations people_vaccinated people_fully_vaccinated

9. The **date** column in *covid_df* was automatically designated by R as a factor (which is a categorical variable designation) or a string (which is a character variable designation) and is therefore not understandable in R as a “datetime object” in its current form. The datetime objects are useful because we will be able to plot and perform relevant computations a lot easier.

- Use the package **lubridate** to create a new vector named **Date** that contains the transformed **date** column structured as a POSIXct date-time object.
- Replace the **date** column in the existing dataframe *covid_df* with the updated POSIXct date-time formatted **Date** column that you’ve just created

10. Use **ggplot2** to create a grid of marker+line plots containing 4 subplots in a 4 by 1 grid (4 rows by 1 column) to show the change in COVID-19 infection over time for the USA.

- The order of the three plots from the top to bottom of the grid should be *new_cases_per_million*, *new_deaths_per_million*, *positivite_rate*, *people_fully_vaccinated*.
- Clearly label your y-axes and x-axes with the names of the variables and their units. (You can refer to the codebook for details about each column.)
- Adjust the size of the plot display so that you can see all the facets clearly when you knit.

- I recommend using the package “patchwork” for making multiple plots, but you can also use other packages for this purpose.

11. What are the trends you observe for these 4 different measures? E.g. When do the highest values occur for each metric? Are these metrics decreasing or increasing at the latest recorded date? Clearly explain your answer for full credit. You can also comment on how these observations correspond with public health efforts (e.g. mask-wearing, physical distancing, and stay-at-home orders). Be sure to reference trusted sources about the public health effort timing!

---

# Exploratory Data Analysis and Data Preprocessing

12. Load the dataset named “student_performance.csv” that can be downloaded from Sakai. This data covers student achievement in secondary education of two Portuguese schools. You can find information about this dataset from the text file, also available in Sakai, named “student_performance_info.csv”, where the column names and variables are explained.

13. Use ggplot and facet_warp() to plot a bar plot for the following categorical variables in this dataset:

Pstatus famsup paid activities nursery higher internet romantic

- When using facet_wrap(), you are encouraged to use 2 plots in each row.
- Adjust the fig.height and fig.width variables in your R chunk definition to show the plots more clearly.

14. Using ggplot, draw 2 histograms plus density function curves of *absences* and *G3* in a 1 by 2 grid

- I recommend using the library “patchwork” for making the 2-plot grid. Label your axis and legend appropriately for full credit.
- You should also choose a suitable binwidth for each variable.

15. Answer the following questions:

- What is the average final grade of subjects in this dataset whose parents are separated?

- What is the average final grade of subjects in this dataset whose parents remain together?

- What is the average number of absences of subjects in this dataset who want to participate in higher education?

- What is the average number of absences of subjects in this dataset who don’t plan to participate in higher education?

16. Now, using a similar 1 by 2 grid, overlay two separate density curves for the two groups with different preferences for higher education, for each variable (feature) from question 3. For these figures, plot just the density curve (without the density histogram bins). Add a vertical line for each of the two density curves on each plot at the mean value for that feature for each group (grouped by preference for higher education). For example, in the first entry of the grid, which is row 1 and column 1 of the 1-by-2 grid, there should be two density plots drawn in two colors for the two different choices in whether one wants to pursue higher education or not.

17. What are your observations from the previous plots? E.g. Do any of the features visually show a strong difference in means or in shape of the data between the students who want to pursue higher education and those who don’t? For each feature, which diagnosis group seems to have a higher average value? Clearly explain your answer to receive full credit.

18. By observing the histogram and density plots of the variables “absences” and “G3”, answer the following questions for each of the variables “absences” and “G3”:

  a. Is the data for each predictor skewed?
  b. Does it have positive skewness or negative skewness?
  c. Compute the skewness using the definition from the lecture.
  d. According to the criterion introduced in the lecture, is the dataset moderately skewed or highly skewed?

19. Correlation plots are a way to visualize multivariate relationships. Using the corrplot package, make a correlation plot of the numeric fields in the dataframe. Clearly label your axis and legend for full credit.

20. Which factor has the strongest correlation with the final grade “G3”? Does this observation make sense to you, and why or why not?

21. Calculate the Z scores of the *absences* and determine if there are any outliers for z > 3. Plot a histogram of *absences*. Remove the outliers (defined by z>3) and plot *absences* again.

22. Perform winsorization on the “absences” variable. This is a technique that was not covered in lecture, but is another type of transformation to limit the effect of outliers on your analyses. Typically, you can decide a threshold (often a specific range of percentiles) and replace all of the data points outside of the threshold with the closest value from within the threshold. An example from Wikipedia may be a helpful demonstration: “a 90% winsorization would see all data below the 5th percentile set to the 5th percentile, and data above the 95th percentile set to the 95th percentile.” Conduct a 90% winsorization on the “absences” variable and plot a histogram of the winsorized data. What is your observation?