# Homework 2 - Data Viz, Wrangling and Advanced Regression

## Due Date

Friday February 14, 2025 by midnight.

## Assignment Description

For this assignment, we will be analyzing data on alcohol consumption and life expectancy. The learning objectives are to conduct data wrangling and visualization keeping key questions in mind. We will also do a regression analysis.

The primary questions of interest are:

- Is there an association between life expectancy and alcohol consumption?
- Is there a difference in the association between life expectancy and alcohol for males and females?
- How have life expectancy and alcohol consumption changed over time?

## Data Wrangling

1. (3 points) Download two datasets: life expectancy and alcohol consumption. These data sets contain information for male and female life expectancy, and alcohol consumption per capita for several countries. Before merging, prepare both datasets as follows:

a. Put the life expectancy data in "tidy" format by creating a new column "Sex". You may want to use `pivot_longer` function from the `tidyr` package.
b. Filter the alcohol consumption data to exclude rows with data for "Both sexes".
c. For convenience, you may rename any variables which have complicated names.
d. Merge these datasets by country name and year.

Write a short paragraph describing your steps and include a summary of the variables in the dataset and the number of observations in each before merging, after filtering, and after merging.

2. (3 points) Using the `kable` package, create a summary table of the merged dataset showing the mean and sd of life expectancy and alcohol consumption by year, and sex. Briefly summarize.

3. (3 points) Create a new categorical variable named "consumption_level" using the alcohol total per capita variable. For female and male *separately*, calculate the quartiles of alcohol consumption. Categorize consumption level as low (0-q1) medium (q1-q3), and high (q3+). To make sure the variable is correctly coded, create a summary table that contains the minimum total alcohol consumption, maximum alcohol consumption, and number of observations for each category.

## Exploring the Data

4. (6 points) Conduct EDA, following the checklist from week 3 and the homework 1. Focus on the key variables pertaining to the primary questions (above). Briefly summarize your steps.

## Visualization

5. (6 points) Create the following figures and interpret them. Be sure to include easily understandable axes, titles, and legends.

a. Stacked histogram of alcohol consumption by sex. Use different color schemes and transparancies so you can see both distributions.
b. Facet plot by year for 2000, 2010, and 2019 showing scatterplots with linear regression lines of life expectancy and alcohol consumption.
c. A linear model of life expectancy as a function of time, adjusted for sex. Compare the summary for Canada, and a second country of your choice.

6. (2 points) Create a faceted barplot of male and female life expectancy for the 10 countries with largest discrepancies (between males and females) in 2000 and 2019.

7. (2 points) Create a boxplot of life expectancy by alcohol consumption level and sex for the year 2019.

8. (2 points) Choose a visualization to examine the association life expectancy between males and females with alcohol consumption over time.

## Advanced Regression

9. (8 points) Construct a multiple linear regression model to examine the association between life expectancy and alcohol consumption level adjusted for population, year, and sex. Note you may want to scale population since the values are so large. First use population as a linear predictor variable, and then fit another model where you put a cubic regression spline on population. Provide the following in your analyses:

a. Summaries of your models, including overall model fit and interpretation of the parameter estimates (linear and non-linear).
b. Plot of the smooth (gam) model and interpretation.
c. Take the average alcohol consumption and life expectancy by country and sex (i.e. remove time). Re-run the same two regressions (except without year), summarize and interpret.
d. Plot average alcohol consumption and life expectancy and color by sex. Add a linear regression line and a smooth regression line.