Homework 1 | **Soc 723**

Regression and probability review | **Social Statistics II**

Instructions

For this first problem set we are going to work with data from the 2024 NYC marathon. It comes from here. It includes information on the city and country of each runner, their gender, age, number of races they have run before and, most importantly, their finishing time. For the probability questions you can work with the numbers in the table below (rounded from the original to simplify calculations), but for the regression questions you will have to load the dataset (which I set on Slack) in R. **To make your work reproducible use set.seed(1) at the top of your script.**

Top \ Country	USA	Kenya	Other
Top 10	3	4	3
Below top 10	38000	9	18000

Problem Set

1. **Probability review.** The table presents the **joint distribution** for two variables: country of origin and whether a runner ended up in the top 10. Compute the following (don't just show the final result, show how you arrived at it):
    a. The **marginal probability** of a runner being Kenyan P(Kenya) and of being American P(USA).
    b. The **joint probability** of being Kenyan and ending up in the top 10 P(Kenya, Top10) and the joint probability of being Kenyan and ending up outside the top 10 P(Kenya, ~ top 10).

c. Calculate the **conditional probability** of being in the top 10 given that a runner is Kenyan P(top10 | Kenya) and the conditional probability of being Kenyan given that a runner is in the top 10 P(Kenya | top10).
d. If you wanted to compare the performance of US and Kenyan runners using **conditional probabilities**, which one would you choose? Why?

2. **Regression review.** Using the dataset that I sent on slack ("nyc_marathon.rds") do the following:
    a. Plot the finishing times in seconds as a histogram (make sure to use enough "bins" if you use ggplot or "breaks" in base R to show the true shape of the distribution, you will need at least 200). Describe the shape of this distribution. Is there anything weird about it?
    b. Now take a simple random sample of 500 runners and plot the relationship between finishing time and age on the sample.
    c. Using the same sample, do a linear regression to predict finishing time using age. What does the coefficient for age tell you? And the intercept?
    d. Using the same regression, predict the finishing time (in minutes) for a runner who is: 5, 20, 50 and 110 years old.
    e. Do the same regression, but this time using the data for all the runners. Plot the two regressions as lines in the same plot and compare them. Under what conditions will they become more similar?
    f. Take repeated samples of size 500 (at least 1000 samples). For each sample, run the same regression and extract the coefficient for age. Plot the distribution of coefficients and compare it to the age coefficient in the whole sample, what do you observe?