# Class3_Exercise

PA 434

1/25/2020

## INSTRUCTIONS

Create a new R project where to store your files and data.

Download the datasets for this assignment from Blackboard.

There are four datasets:

- **MigrationFlows** reports the total number of immigrants present in the country (total and by gender)
- **Population** reports the total population of a country (total and by gender)
- **Refugees** reports the total number of refugees within a country
- **Origin** reports the origin (by continent) of immigrants in a country

All data include 6 years: 1990, 1995, 2000, 2005, 2010, 2015

The total number of countries is 232 (when all countries are included in a dataset)

## DATA SOURCES

Migration stocks are the numbers of migrants living in a country or region at a given point in time. Source: https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates15.asp (I did some slight modifications to this original dataset)

232 countries - http://www.madore.org/~david/misc/countries.html

## YOUR TASK #1 - TIDY DATA

Datasets are stored in different formats. Your boss is asking you to make each of these dataset "tidy" if they are not tidy already.

He asks you to produce two outcomes:

1. A tidy dataset containing all common observations (e.g., countries) across all four datasets
2. A tidy dataset contaning as much information as possible across all four datasets.

As you work on your R script to clearn your data, consider these few questions and provide your answers into the script. You can answer them at any point in your script, just report the number so that we can easily identify them.

1. What is the unit of analysis of the dataset?
2. What are the issues with each dataset? How do you make your data tidy?
3. What is expected number of rows for each of the two datasets that you need to produce?
4. Make sure to report the final size of each new dataset and confirm that it matches with your expectations.

# YOUR TASK 2 - REVIEW OF DOLLAR-SIGN SYNTAX

Using the dataset containing the most information, your boss asks you to look at the data and extract a few information:

5. What is the average percentage of migrants on the total population of a country in 2015?
6. Did the average percentage of refugees on the total of immigrants increased or decreased from 1990?
7. What is the highest percentage of immigrants in a country in 2010? What is the smallest?
8. What is the median percentage of immigrants from the different continents/geographical areas in 2015?

# A COUPLE OF SUGGESTION

When starting to "tidy" your data, if you want to take a progressive approach start from the origin dataset, refugees dataset, move to population and then migrationflows. If you like a challenge, start with migrationflows.

Make sure that your column are named appropriately

If you get NA as a result when answering questions in "Your Task 2" try to use the argument na.rm = T in your mean function. We haven't talked about it yet, but give it a try if you want.

To analyze your data, use the subset syntax []. We will learn how to extract this information with tidyverse next week.