Week 5 Assignment

PA 434

2/9/2020

#dplyr - Part 2

Perform the following actions and answer the following questions using dplyr. Use pipes when writing your code. Note that these questions might require you to combine several dplyr functions.

A few suggestions:

- Start by writing down what information you need to retrieve to answer these questions
- Break down the steps that you need to do in order to retrieve this information
  - Which variables are involved in the analysis? Which observations should be included?
  - Is this information readily available in the dataset or do you need to create a new column? (mutate)
  - Do you need to use the entire datasets or filter your data? (filter)
  - Do you need to group your data according to any variable? (group_by)
  - Do you need to summarize results by looking at multiple rows at the same time? (summarize)

Questions

1. What is the number of Migrant by region in 2015?
2. How many countries have a share of immigrants that it's greater than or equal to 50% of their population in 2015?

Hint:

```
data %>%
  mutate( ............ = Migrants / ...........) %>%
  filter( ............ > 0.5, year == .........) %>%
  summarise(.......... = n())
```

3. In which world regions are most of these countries located?
4. How many countries have a share of immigrants greater than or equal to 50% of their population if we consider the average from 1990 to 2015? Which is the country with the highest average across all year?

Hint: Questions 2 and 3 are variations of question 1. Think about where you need to filter and group_by

5. Which region saws the largest increase in the number of refugees from 1990 to 2015? And which region saws the largest decrease? (Note: Pay attention to unexpected results. Do some digging if needed!).

Hint: You can perform multiple operations with mutate. Just separate each operation with a comma. See example below

```
summarise(perc_male = MaleMigrants / Migrants,
          perc_female = FemaleMigrants / Migrants)
```

6. Has there been any changes in the immigration patterns by gender over time? Which country has the lowest female immigration? Which region has the highest female immigration? Consider the average across all the years.

Missing values

If you were to conduct this analysis for your organization, you most likely would want to report some information about missing information in the dataset.

Even if you end up removing your missing values from your analysis (as we did here), it is important to acknowledge how many and which countries were not included in the analysis because you did not have information about them. We haven't checked so far, so it is even possible that most countries miss observations, or that some years have more missing values than others!

Note that this step is generally done at the very beginning, before starting your analysis. In this way, you know what to expect from your data and you can take appropriate decisions.

1. Use the is.na functions to calculate how many missing values in the refugees, population, and migrant columns.
2. For the column(s) with missing values, conduct some additional analysis - are those random by region, year, or country? Which regions, years, or countries have the largest amount of missing values? Can you think of an explanation of why there are these missing values? Would you classify these missing values as MCAR, MAR, or MNAR?
3. Write a short wrap-up of what you discovered on missing values. Note any surprising findings and you think missing values might affect any analysis contacted with the dataset.