STA130H1S – Fall 2022

Week 3 Tutorial Handout

Today's agenda (5 min):

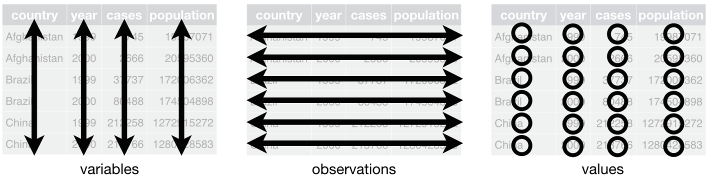
- Q&A/vocabulary list
- Group Discussion
- Writing prompt

This Week's Vocab (15-20 min) :

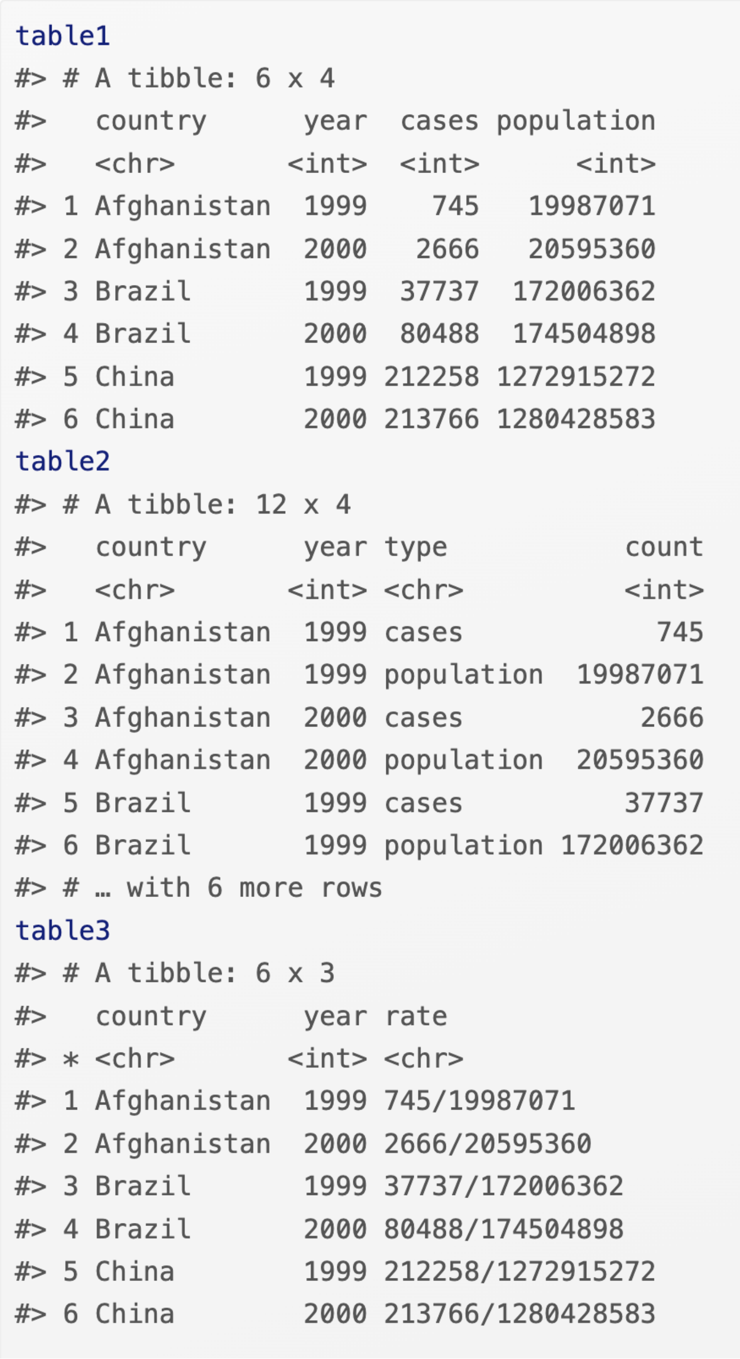
- Cleaning data
- Tidy data
- Removing a column
- Extracting a subset of variables
- Filtering a tibble based on a condition (e.g. based on the values in one or more of the variables/columns)
- Sorting data based on the values of a variable
- Renaming the variables
- Grouping categories
- Defining new variables
- Producing new data frames
- Handling missing values (NAs)
- Creating summary tables

Discussion (20 min) :

- What are the three interrelated rules that make a dataset tidy?
  - Each variable must have its own column.
  - Each observation must have its own row.
  - Each value must have its own cell.



- Are these data below tidy? Why or why not? If not, how to make them tidy?



```
table1
#> # A tibble: 6 x 4
#>   country      year  cases population
#>   <chr>       <int>  <int>      <int>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
table2
#> # A tibble: 12 x 4
#>   country      year type           count
#>   <chr>       <int> <chr>          <int>
#> 1 Afghanistan  1999 cases            745
#> 2 Afghanistan  1999 population  19987071
#> 3 Afghanistan  2000 cases           2666
#> 4 Afghanistan  2000 population  20595360
#> 5 Brazil       1999 cases          37737
#> 6 Brazil       1999 population 172006362
#> # … with 6 more rows
table3
#> # A tibble: 6 x 3
#>   country      year rate
#> * <chr>       <int> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

- Why is data visualization so important?
- Why does your audience matter? (Think about what message you want to portray and the types of data/ visualizations that you'll use)
- Why is it important for data visualizations to be intuitive? How can you ensure your figures are intuitive to the intended audience?
- What might happen to a data visualization project if you failed to clean the data?

Writing prompt (30 min) : You have just been hired by a consultancy company. Congratulations! They are doing a report on each Olympics for the past 10 years. Given your recent experience in STA130, you ask to be responsible for the 2012 summary. Write a short report to your boss on information that can be gleaned about the ages of the athletes (since your boss' favourite sports are badminton and weightlifting, you know she will be happy if your summary talks about these sports specifically, but you can talk about other interesting of features athletes' ages which can be learned from your plots and tables.)

Important Features to Include

- Start off with a small introduction. You should include 1 or 2 sentences to draw your reader in, and then explain what you will be discussing.
- Make sure to include at least 1 figure to help your reader visualize what you are speaking about.
- You want to show off all the knowledge you gained in STA130 so you must include at least 2 vocabulary words. However, your boss isn't a statistician, therefore you must define any vocabulary terms you used.
- Make sure to finish with a conclusion to remind your boss of the key take home points from your summary about the athletes' ages.

Some general reminders

- Try to not spend more than 20 minutes on the prompt.
- Aim for more than 200 words but less than 400 words.
- Use full sentences.
- Grammar is not the main focus of the assessment, but it is important that you communicate in a clear and professional manner. Remember, this is meant to be a small report to your boss! So you should not include any slang or emojis.
- Remember that you have only conducted a preliminary analysis and therefore you may not have a definitive answer. That is totally ok! It is hard to ever say something is 100% one way or another. Therefore, you will want to try to incorporate some hedging language into your writing. You can find more information about hedging in a short video here: https://q.utoronto.ca/courses/253019/pages/writing-skills-videos

Vocabulary

- Cleaning data
- Tidy data
- Removing a column
- Extracting a subset of variables
- Filtering a tibble based on a condition (e.g. based on the values in one or more of the variables/columns)
- Sorting data based on the values of a variable
- Renaming the variables
- Grouping categories
- Defining new variables
- Producing new data frames
- Handling missing values (NAs)
- Creating summary tables

You may also find these vocabulary words from last week useful with your writing this week

- Where are the data centered (towards the left, right, middle)
- How much spread (relative to what?)
- Shape: symmetric, left-skewed, right-skewed
- The tails of the distribution (heavy-tailed or thin-tailed)
- Modes: where, how many, unimodal, bimodal, multimodal, uniform

- Outliers, extreme values
- Frequency (which category occurred the most or least often; data concentrated near a particular value or category)
- Mean (average), median, mode
- standard deviation, interquartile range