Lesson 5 Tidy(ing) Data - Part 1

Goals for Today

- Get more exposure and practice with standard `tidyverse` commands, such as `select`, `filter`, `mutate`, and others!

Readings for Today

Required

- Martin Nyamu, Data Cleaning in RStudio
  - This is a good analytical exercise you can follow along with to try out.
- Grolemund and Wickham, R for Data Science - Chapter 5, Data Transformations
  - Note: you don't have to do the exercises!

Optional

None!

Important Links and Files

- This cheatsheet will likely be invaluable!
- The file we worked on in class is available here

Written Assignment #1

Details

- **Due Date**:
  - 2021-10-01 (**this Friday!**) prior to the start of class (i.e., needs to be submitted before 1PM EST).
- **Format**:
  - You'll submit this via Google Classroom. It can be in whatever format you like - so long as it uploads!
- **Working Style**
  - You can do this individually or as a group - it is entirely up to you! If you do work in groups, please make note in the document that you did so and list everyone's name.

Data

Hungry? If you're not - you're about to be. We're going to be diving into a data set about delicious, sugary, probably-pretty-bad-for-you American breakfast cerals!

The context and background of this data can be found here. You'll need to look here to see what the columns mean (this is also known as a "data dictionary").

The data itself can be downloaded from here. If you would like to get the .csv file without having to unzip it, you can grab it from the course github in the /data directory here.

Questions

Answer these as thoroughly as you can and please provide the code that you've used to generate your answer.

1. What are the dimensions of this data set?
2. How many columns in this data set have a character data type?
3. Which manufacturer has the most cereals in this data? How many rows of the dataset does thie manufacturer represent?
4. What is the name of the cereal manufactured by American Home Food Products?
5. Which three breakfast cereals have the lowest calorie count per serving?

6. Which cold breakfast cereal from Quaker Oats has the highest overall rating?
7. If you sum the grams of sugar and grams of carbohydrates, which hot cereal brand has the highest combination?
8. How many rows in this data set contain a negative value?
9. Which cereal on the 1st display shelf has the highest amount of fiber *and* the lowest amount of sodium?
10. What percentage of the cereals in this dataset are hot breakfast cereals?
11. Which shelf has the highest mean (i.e., average) percentage of daily vitamins per serving?
12. Which manufacturer has the highest percentage of cereals found on the 3rd shelf?
13. What is the range between the highest and lowest rating?
14. Which cereal brand has the highest number of grams of fat per calorie?
15. If you had to guess which of these cereal brands may have had the most data input and/r data collection errors, which one would you guess and why?

Bonus question: which cereal brand has the most number of words in it's name? To answer this, try taking a peek at the function `str_count` with `?str_count` and see how it works. Just giving the answer to this doesn't count - showing me a valid command or series of commands that produces it will be worth some extra credit!