Tidyverse Challenge Problems

SIBDS @ Columbia

2022-06-24

Getting Started

1. Create a new project in your RWorkspace named `tidyverse_challenge`, or something similar.
2. Click into the RProject and create a new RMarkdown document within the project named `tidyverse_challenge`, or something similar.
3. Modify the YAML header to your preference and knit the RMarkdown document to make sure it runs as expected.
4. Run this line of code in your console to get the datasets you need:

```
devtools::install_github("p8105/p8105.datasets")
```

Problem 1

Load the `instacart` dataset with the following code:

```
library(p8105.datasets)
data("instacart")
```

1. How many aisles are there, and which aisles are the most items ordered from?
2. Make a plot that shows the number of items ordered in each aisle, limiting this to aisles with more than 10000 items ordered. Arrange aisles sensibly, and organize your plot so others can read it.
3. Make a table showing the three most popular items in each of the aisles "baking ingredients", "dog food care", and "packaged vegetables fruits". Include the number of times each item is ordered in your table.
4. Make a table showing the mean hour of the day at which Pink Lady Apples and Coffee Ice Cream are ordered on each day of the week; format this table for human readers (i.e. produce a 2 x 7 table).

Problem 2

Load the `BRFSS` dataset with the following code:

```
data("brfss_smart2010")
```

1. First, do some data cleaning:

- Format the data to use appropriate variable names
- Focus on the "Overall Health" topic include only responses from "Excellent" to "Poor" organize responses as a factor taking levels ordered from "Poor" to "Excellent"

2. Using this dataset, do or answer the following (commenting on the results of each):

- In 2002, which states were observed at 7 or more locations? What about in 2010?

- Construct a dataset that is limited to Excellent responses, and contains, year, state, and a variable that averages the data_value across locations within a state. Make a "spaghetti" plot of this average value over time within a state (that is, make a plot showing a line for each state across years – the geom_line geometry and group aesthetic will help).
- Make a two-panel plot showing, for the years 2006, and 2010, distribution of data_value for responses ("Poor" to "Excellent") among locations in NY State.

Problem 3

Accelerometers have become an appealing alternative to self-report techniques for studying physical activity in observational studies and clinical trials, largely because of their relative objectivity. During observation periods, the devices measure "activity counts" in a short period; one-minute intervals are common. Because accelerometers can be worn comfortably and unobtrusively, they produce around-the-clock observations.

This problem uses five weeks of accelerometer data collected on a 63 year-old male with BMI 25, who was admitted to the Advanced Cardiac Care Center of Columbia University Medical Center and diagnosed with congestive heart failure (CHF). The data can be downloaded from CourseWorks (Recitations > Problem Sets > data). In this spreadsheet, variables activity.* are the activity counts for each minute of a 24-hour day starting at midnight.

Load, tidy, and otherwise wrangle the data. Your final dataset should include all originally observed variables and values; have useful variable names; include a weekday vs weekend variable; and encode data with reasonable variable classes. Describe the resulting dataset (e.g. what variables exist, how many observations, etc).

Traditional analyses of accelerometer data focus on the total activity over the day. Using your tidied dataset, aggregate accross minutes to create a total activity variable for each day, and create a table showing these totals. Are any trends apparent? Accelerometer data allows the inspection activity over the course of the day. Make a single-panel plot that shows the 24-hour activity time courses for each day and use color to indicate day of the week. Describe in words any patterns or conclusions you can make based on this graph.