HW 11

David Gerard

2020-04-08

Instructions

- Write your solutions in this starter file. You should modify the "author" field in the YAML header.
- Do not modify the paths of any files.
- Only commit R Markdown and HTML files (no PDF or Word files). Make sure you have knitted to HTML for your final submission.
- **Make sure to commit each time you answer a question.** Lack of informative and frequent commits will result in point deductions.
- Only include the necessary code, not any extraneous code, to answer the questions.
- For this homework, **use only data.table and not the tidyverse**.
- Learning objectives:
  - Manipulating data.tables.
  - Tidying data.tables.

Bob Ross

The data in "bob.csv" contains information on episodes hosted by Bob Ross, a painter who hosted a popular television series on PBS. Each episode would consist of him completing an entire painting. His paintings almost always consisted heavily of elements from nature: trees, clouds, mountains, lakes, etc. These data consist of indicators for what elements are in each episode. This dataset was taken from the excellent crew at fivethirtyeight. See here for their article. The variables are:

- `EPISODE`: The season and episode number of the episode.
- `TITLE`: The title of the painting.
- Every other variable is an indicator for whether the episode contains the element described by the variable name. For example, `BARN` is `0` if the episode does not have a barn in the painting and `1` if the episode does have a barn in the painting.

1. Use data.table and `fread()` to read these data into R.

2. Tidy these data by having new columns called "`ELEMENT`" and "`PRESENT`". `ELEMENT` should contain the name of the element included in the painting and the variable `PRESENT` should be `1` if that element is present and `0` otherwise. Your new `data.table` should look like this:

```
##         EPISODE                       TITLE     ELEMENT PRESENT
##      1:  S01E01    ""A WALK IN THE WOODS"" APPLE_FRAME       0
##      2:  S01E02           ""MT. MCKINLEY"" APPLE_FRAME       0
##      3:  S01E03           ""EBONY SUNSET"" APPLE_FRAME       0
##      4:  S01E04            ""WINTER MIST"" APPLE_FRAME       0
##      5:  S01E05           ""QUIET STREAM"" APPLE_FRAME       0
##     ---
## 26997:  S31E09       ""EVERGREEN VALLEY"" WOOD_FRAMED       0
## 26998:  S31E10            ""BALMY BEACH"" WOOD_FRAMED       0
## 26999:  S31E11      ""LAKE AT THE RIDGE"" WOOD_FRAMED       0
```

```
## 27000:  S31E12 ""IN THE MIDST OF WINTER"" WOOD_FRAMED       0
## 27001:  S31E13         ""WILDERNESS DAY"" WOOD_FRAMED       0
```

3. The variable `EPISODE` contains both the season number and the episode number. Create two new variables called "`SEASON`" and "`EPISODE_NUM`". `SEASON` should be a numeric containing the season number and `EPISODE_NUM` should be a numeric containing the episode number within the season. Do this by reference. You can use readr parsers here. Your data.table should look like this:

```
##         EPISODE                      TITLE     ELEMENT PRESENT SEASON
##      1:  S01E01    ""A WALK IN THE WOODS"" APPLE_FRAME       0      1
##      2:  S01E02           ""MT. MCKINLEY"" APPLE_FRAME       0      1
##      3:  S01E03           ""EBONY SUNSET"" APPLE_FRAME       0      1
##      4:  S01E04            ""WINTER MIST"" APPLE_FRAME       0      1
##      5:  S01E05           ""QUIET STREAM"" APPLE_FRAME       0      1
##     ---
## 26997:  S31E09       ""EVERGREEN VALLEY"" WOOD_FRAMED       0     31
## 26998:  S31E10            ""BALMY BEACH"" WOOD_FRAMED       0     31
## 26999:  S31E11      ""LAKE AT THE RIDGE"" WOOD_FRAMED       0     31
## 27000:  S31E12 ""IN THE MIDST OF WINTER"" WOOD_FRAMED       0     31
## 27001:  S31E13         ""WILDERNESS DAY"" WOOD_FRAMED       0     31
##         EPISODE_NUM
##      1:           1
##      2:           2
##      3:           3
##      4:           4
##      5:           5
##     ---
## 26997:           9
## 26998:          10
## 26999:          11
## 27000:          12
## 27001:          13
```

4. Recode the `PRESENT` variable to have values that are more human-readable. Do this by reference.

5. Calculate the proportion of episodes each season that have each element. Save these summaries in a new data.table. Arrange the rows in decreasing order of proportion. Your data.table should look like this:

```
##            ELEMENT SEASON PROPORTION
##    1:        TREE      2          1
##    2:        TREE      4          1
##    3:        TREE     13          1
##    4:       TREES      2          1
##    5:       TREES     13          1
##   ---
## 2073: WOOD_FRAMED     26          0
## 2074: WOOD_FRAMED     27          0
## 2075: WOOD_FRAMED     28          0
## 2076: WOOD_FRAMED     29          0
## 2077: WOOD_FRAMED     31          0
```

6. Did Bob draw trees less frequently over the seasons. Explore with an appropriate plot. You can use ggplot2 or base R.

7. Does Bob prefer conifers or deciduous trees? Justify your conclusions with numerical summaries and a statistical test.