CptS 475/575: Data Science, Fall 2018

Assignment 3: Data Transformation and Tidying

Release Date: September 19, 2018 **Due Date:** September 27, 2018 (11:59 pm)

This assignment has two questions. You are encouraged to use R Markdown to generate your report (in PDF).

Question 1. For this question you will be using the dplyr package to manipulate and clean up a dataset called msleep (mammals sleep) that is available on the course webpage (at https://scads.eecs.wsu.edu/wp-content/uploads/2017/10/msleep_ggplot2.csv). The dataset contains the sleep times and weights for a set of mammals. It has 83 rows and 11 variables. Here is a description of the variables:

Name	Description
name	common name
genus	taxonomic rank
vore	carnivore, omnivore or herbivore?
order	taxonomic rank
conservation	the conservation status of the mammal
sleep_total	total amount of sleep, in hours
sleep_rem	rem sleep, in hours
sleep_cycle	length of sleep cycle, in hours
awake	amount of time spent awake, in hours
brainwt	brain weight in kilograms
bodywt	body weight in kilograms

Load the data into R, and check the first few rows for abnormalities. You will likely notice several.

Below are the tasks to perform. Use select() to print the head of the columns with a title including "sleep".

a) Use filter() to count the number of animals which weigh over 50 kilograms and sleep more than 6 hours a day.
b) Use piping (%>%), select() and arrange() to print the name, order, sleep time and bodyweight of the animals with the top 6 sleep times, in order of sleep time.
c) Use mutate to add two new columns to the dataframe; wt_ratio with the ratio of brain size to body weight, rem_ratio with the ratio of rem sleep to sleep time. If you think they might be useful, feel free to extract more features than these, and describe what they are.
d) Use group_by() and summarize() to display the average, min and max sleep times for each order. Remember to use ungroup() when you are done.
e) Make a copy of your dataframe, and use group_by() and mutate() to impute the missing brain weights as the average wt_ratio for that animal's order times the animal's weight. Make a

second copy of your dataframe, but this time use group_by() and mutate() to impute missing brain weights with the average brain weight for that animal's order. What assumptions do these data filling methods make? Which is the best way to impute the data, or do you see a better way, and why? You may impute or remove other variables as you find appropriate. Briefly explain your decisions.

Question 2. For this question, you will first need to read section 12.6 in the R for Data Science book, here (http://r4ds.had.co.nz/tidy-data.html#case-study). Grab the dataset from the tidyr package, and tidy it as shown in the case study before answering the following questions.

a) Explain why this line

```
> mutate(key = stringr::str_replace(key, "newrel", "new_rel"))
```

is necessary to properly tidy the data. What happens if you skip this line?

b) How many entries are removed from the dataset when you set na.rm to true in the gather command (in this dataset). How else could those NA values be handled? Among these options, which do you think is the best way to handle those missing values for this dataset, and why?

c) Explain the difference between an explicit and implicit missing value, in general. Can you find any implicit missing values in this dataset, if so where?

d) Looking at the features (country, year, var, sex, age, cases) in the tidied data, are they all appropriately typed? Are there any features you think would be better suited as a different type? Why or why not?

e) Explain in your own words what a gather operation is, and give an example of a situation when it might be useful. Do the same for spread.

f) Generate an informative visualization, which shows something about the data. Give a brief description of what it shows, and why you thought it was interesting.