CptS 475/575: Data Science, Fall 2020

Assignment 3: Data Transformation and Tidying

Release Date: September 16, 2020 **Due Date:** September 23, 2020 (11:59 pm)

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

Question 1. (50 pts total) For this question you will be using either the dplyr package from R or the Pandas library in python 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 or Python, and check the first few rows for abnormalities. You will likely notice several. All of the tasks in this assignment can be hand coded, but the goal is to use the functions built into dplyr or Pandas to complete the tasks. Suggested functions for Python will be shown in blue while suggested R functions are shown in red. Note: if you are using Python, be sure to load the data as a Pandas DataFrame.

Below are the tasks to perform. Before you begin, print the first few values of the columns with a header including "sleep". (head(), head())

a) (8 pts) Count the number of animals which weigh under 1 kilogram and sleep more than 14 hours a day. (filter(), query())
b) (8 pts) Print the name, order, sleep time and bodyweight of the animals with the 6 *longest* sleep times, in order of sleep time. (select(), arrange(), loc(), sort_values())
c) (8 pts) 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. (mutate(), assign())

d) (12 pts) Display the average, min and max sleep times for each order. (group_by(), summarise(), groupby(), agg())

e) (14 pts) 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 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. (group_by(), mutate(), groupby(),assign())

Question 2. (50 pts total) 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 (tidyr::who), and tidy it as shown in the case study before answering the following questions. Note: if you are using pandas you can perform these same operations, just replace the gather() function with melt() and the spread() function with pivot(). However, you may prefer to use R for this question, as the dataset is from an R package.

a) (5 pts) 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) (5 pts) How many entries are removed from the dataset when you set na.rm to true in the gather command (in this dataset)?

c) (5 pts) 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) (5 pts) Looking at the features (country, year, var, sex, age, value) 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) (10 pts) Explain in your own words what a gather operation is, and briefly describe an example of a situation when it might be useful. Do the same for spread.

f) (10 pts) Generate an informative visualization, which shows something about the data. Give a brief description of what it shows, and why you thought it would be interesting to investigate.

g) (10 pts) Suppose you have the following dataset called siteDemo:

Site	**U30.F**	**U30_M**	**O30.F**	**O30.M**
facebook	30	35	66	58
myspace	1	2	3	6
snapchat	6	5	3	2
twitter	18	23	12	28
tiktok	44	60	2	7

You know that the U30.F column is the number of female users under 30 on the site, O30.M denotes the number of male users 30 or older on the site, etc. Construct this table, and show the code you would use to tidy this dataset (using gather() and separate() or melt() and pivot()) such that the columns are organized as: Site, AgeGroup, Gender and Count.