CptS 475/575: Data Science, Fall 2022
Assignment 3: Data Transformation and Tidying
Release Date: September 14, 2022 **Due Date:** September 21, 2022 (11:59 pm)

This assignment has two questions. What you will submit on Canvas will be a PDF file that contains your code, results, and any text explanation you provide as part of your solution. You are encouraged to use R Markdown to generate your report (in PDF).

For each of the two questions, the total points the question carries is indicated in parenthesis. This is further broken down into the subproblems the question has, and the weights/points are similarly indicated.

Good luck!

Question 1. (60 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 "WNBA Stats 21" called *WNBA_Stats_21.csv* (from the Modules page, under Datasets for Assignments on Canvas). This data was pulled from https://www.wnba.com/stats/player-stats/ website.

The dataset contains information about the Women's National Basketball Association games in 2021. It has 139 rows and 21 variables. Here is a description of the variables:

Variable	Description
PLAYER	Name of the player
TEAM	Name of the team
AGE	Age of the player
G	Games Played
MIN	Minutes Played
FGM	Field Goals Made
FGA	Field Goals Attempted
3PM	3 Point Field Goals Made
3PA	3 Point Field Goals Attempted
FTM	Free Throws Made
FTA	Free Throws Attempted
OREB	Offensive Rebounds
DREB	Defensive Rebounds
REB	Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TO	Turnovers
PTS	Points
DD2	Double Doubles
TD3	Triple Doubles

Load the data into R or Python, and check for abnormalities (NAs). You will likely notice several. All 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 containing the string "FG". (head(), head())

a) (10 pts) Count the number of players with Free Throws Made greater than 50 and Assists greater than 75. (filter(), query())
b) (10 pts) Print the PLAYER, TEAM, FGM, TO and PTS of the players with the 10 *highest* points, in descending order of points. (select(), arrange(), loc(), sort_values()). Which player has the second highest points?
c) (10 pts) Add two new columns to the dataframe; FGP (in percentage) is the ratio of FGM to FGA, FTP (in percentage) is the ratio of FTM to FTA. Note that the unit should be expressed in percentage (ranging from 0 to 100) and rounded to 2 decimal places (e.g., for Aja Wilson, FGP is 44.42). If you think they might be useful, feel free to extract more features than these, and describe what they are. (mutate(), assign()). What is the FGP and FTP for Tina Charles?
d) (14 pts) Display the average, min and max REB for each team, in descending order of the team average. (group_by(), summarise(), groupby(), agg()). You can exclude NAs for this calculation. Which team has the max REB?
e) (16 pts) In question 1c, you added a new column called FTP. Impute the missing (or NaN) FTP values as the FGP (also added in 1c) multiplied by the average FTP for that team. Make a second copy of your dataframe, but this time impute missing (or NaN) FTP values with just the average FTP for that team. 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. (40 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. The dataset is also available on the Modules page, under Datasets for Assignments, on Canvas. Note: if you are using Pandas you can perform these same operations, just replace the pivot_longer() function with melt() and the pivot_wider() 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 values_drop_na to true in the pivot_longer 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, 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) (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.

f) (10 pts) Suppose you have the following dataset called SchQtr (You can download this dataset from the Modules page, under Datasets for Assignments, on Canvas):

School	Year	Qtr.1	Qtr_2	Qtr.3	Qtr.4
UNI	2018	27	90	12	84
COL	2018	42	27	62	1
ACA	2018	6	51	58	8
UNI	2019	54	70	60	39
COL	2019	17	20	45	99
ACA	2019	39	91	78	38
UNI	2020	26	66	42	26
COL	2020	51	48	29	34
ACA	2020	71	31	30	56
UNI	2021	45	1	39	81
COL	2021	65	26	82	48
ACA	2021	22	69	48	38

The table consists of 6 columns; first showing the School Code, second representing the year and the last four columns provide the number of high excellence (summa cum laude) students graduating in each quarter of the year. Re-structure this table and show the code you would use to tidy this dataset (using gather()/pivot_longer() and separate()/pivot_wider() or melt() and pivot()) such that the columns are organized as:

School, Year, Interval_Type, Interval_ID and Student_Count.

Note: Here the entire Interval_Type column will contain value 'Qtr' since the dataset counts students every quarter. The Interval_ID will contain the quarter number.

Below is an instance of a row of the re-structured table:

School	Year	Interval_Type	Interval_ID	Student_Count
UNI	2018	Qtr	1	27

How many rows does the new dataset have?