CptS 475/575: Data Science, Fall 2023
Assignment 3: Data Transformation and Tidying
Release Date: September 13, 2023 **Due Date:** September 20, 2023 (11:59 pm)

This assignment has two questions. What you will submit on Canvas will be a PDF or HTML 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/HTML) if you used R to solve the problems. If you used Python, Jupyter notebook would be convenient to produce your PDF or HTML, but you are free to use whatever IDE you are comfortable with.

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 called *NBA_Stats_22_23.csv* (available in the Modules section on Canvas under the folder Datasets for Assignments). This data was pulled from https://www.nba.com/stats website.

The dataset contains information about the Men's National Basketball Association games in 2022 - 2023. It has 539 rows and 25 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
GP	Games Played
W	Wins
L	Losses
MIN	Minutes Played
PTS	Points
FGM	Field Goals Made
FGA	Field Goals Attempted
X3PM	3 Point Field Goals Made
X3PA	3 Point Field Goals Attempted
FTM	Free Throws Made
FTA	Free Throws Attempted
OREB	Offensive Rebounds
DREB	Defensive Rebounds
REB	Rebounds
AST	Assists
TOV	Turnovers
STL	Steals
BLK	Blocks

PF	Personal Fouls
FP	Fantasy points
DD2	Double Doubles
TD3	Triple Doubles

Load the data into R or Python. 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 are 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 "X3". (head(), head())

a) (10 pts) Count the number of players with Free Throws Made greater than 60 and Assists greater than 80. (filter(), query())
b) (10 pts) Print the PLAYER, TEAM, W, L, FGM, TOV 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 AJ Griffin, FGP is 46.53) (mutate(), assign()). What is the FGP and FTP for Joe Harris?
d) (14 pts) Display the average, min and max PF 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 PF?
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 (http://r4ds.had.co.nz/tidy-data.html#case-study). Grab the dataset "who" 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 by just replacing the pivot_longer() function with melt() and the pivot_wider() function with pivot().

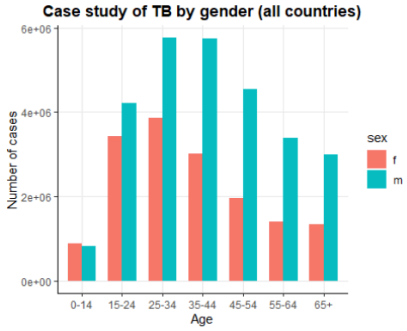
a) (5 pts) Explain why this line

```
> 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) (8 pts) Produce a barplot to show the count of TB cases by gender for all countries. You can create side by side bars for the two genders. Your resulting plot is expected to look like the one shown below:



f) (12 pts) Suppose you have the following dataset called RevQtr (You can download this dataset from the Modules page, under Datasets for Assignments, on Canvas):

Group	Year	Qtr.1	Qtr_2	Qtr.3	Qtr.4
1	2019	27	90	12	84
2	2019	42	27	62	19
3	2019	26	51	58	8
1	2020	54	70	60	39
2	2020	17	20	45	99
3	2020	39	91	78	38
1	2021	26	66	42	26
2	2021	51	48	29	34
3	2021	71	31	30	56
1	2022	45	11	39	81
2	2022	65	26	82	48
3	2022	22	69	48	38

The table consists of 6 columns. The first shows the Group code, the second shows the year and the last four columns provide the revenue for each quarter of the year. Re-structure this table and show the code you would write to tidy the dataset (using gather()/pivot_longer() and separate()/pivot_wider() or melt() and pivot()) such that the columns are organized as:

Group, Year, Interval_Type, Interval_ID and Revenue.

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

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

Group	Year	Interval_Type	Interval_ID	Revenue
1	2019	Qtr	1	27

How many rows does the new dataset have?