CptS 475/575: Data Science, Fall 2024
Assignment 3: Data Transformation and Tidying
Release Date: September 13, 2024 **Due Date:** September 21, 2024 (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 or Quarto 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.

For all plots and visualizations, make sure your plot has a relevant title, and the x and y axes are labeled appropriately. Ensure that the labels are clear, legible, and easy to read.

Good luck!

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 *NBA_Stats_23_24.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 2023 - 2024. It has 735 rows and 30 variables. Here is a description of the variables:

Variable	Description
Rk	Rank
Player	Player's name
Pos	Position
Age	Player's age
Tm	Team
G	Games played
GS	Games started
MP	Minutes played per game
FG	Field goals per game
FGA	Field goal attempts per game
FG%	Field goal percentage
3P	3-point field goals per game
3PA	3-point field goal attempts per game
3P%	3-point field goal percentage
2P	2-point field goals per game
2PA	2-point field goal attempts per game
2P%	2-point field goal percentage
eFG%	Effective field goal percentage

FT	Free throws per game
FTA	Free throw attempts per game
FT%	Free throw percentage
ORB	Offensive rebounds per game
DRB	Defensive rebounds per game
TRB	Total rebounds per game
AST	Assists per game
STL	Steals per game
BLK	Blocks per game
TOV	Turnovers per game
PF	Personal fouls per game
PTS	Points per game

Load the data into R or Python, and check for missing values (NaN). 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 "FG". (head(), head())

a) (5 pts) Count the number of players with Free Throws per game greater than 0.5 and Assists per game greater than 0.7. (filter(), query())
b) (10 pts) Print the Player, Team, Field goals per game, Turnovers per game, and Points per game of the players with the 10 *highest* points, in descending order of points. (select(), arrange(), loc(), sort_values()). Which player has the seventh highest points?
c) (10 pts) Add two new columns to the dataframe: FGP (in percentage) is the ratio of FG to FGA, FTP (in percentage) is the ratio of FT 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 Jamal Cain, FGP is 43.33) (mutate(), assign()). What is the FGP and FTP for Josh Giddey?
d) (10 pts) Display the average, min and max Offensive rebounds per game 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 Offensive rebounds per game?
e) (15 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. (50 pts total) For this question, you will first need to read section 5.3.1 in the R for Data Science book (https://r4ds.hadley.nz/data-tidy#sec-billboard). Grab the dataset "billboard" from the tidyr package (tidyr::billboard), 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(week = parse_number(week))
```

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 (artist, track, date.entered, week, rank) 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) (5 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) (5 pts) Generate a line plot showing the rank progression of a specific song over time. You can choose a song you like best from the dataset. (Hint: higher ranks are better so reverse your axis appropriately). Briefly describe what the plot shows.

g) (8 pts) Produce a barplot to show the count of songs per artist in the dataset. Limit the plot to the top 15 artists by number of songs. What are your thoughts about this top 15 list? Were you surprised by the presence of any particular artist?

h) (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	2022	61	24	81	70
1	2023	30	92	96	84
1	2024	84	97	33	12
2	2022	31	62	11	97
2	2023	39	47	11	73
2	2024	69	30	42	85
3	2022	67	31	98	58
3	2023	68	51	69	89
3	2024	24	71	71	56
4	2022	71	60	64	73
4	2023	12	60	16	30
4	2024	82	48	27	13

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	2022	Qtr	1	61

How many rows does the new dataset have?