CptS 475/575: Data Science, Fall 2021
Assignment 3: Data Transformation and Tidying
Release Date: September 15, 2021 **Due Date:** September 22, 2021 (11:59 pm)

This assignment has two questions. What you will submit on Canvas will 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 point 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 flights (from the library nycflights13 in R) that can be loaded (for R users) from the 'data' folder on the github repository

https://github.com/tidyverse/nycflights13

The dataset flights.csv is also made available on canvas at (via the Module titled Datasets):

https://wsu.instructure.com/courses/1494050/modules/items/13783382

The dataset contains information about the flights departing New York City in 2013. It has 336,776 rows and 19 variables. Here is a description of the variables:

Name	**Description**
year	Year of departure
month	Month of departure (1-12 for Jan-Dec)
day	Day of departure
dep_time	Actual departure times (format HHMM or HMM), local tz
sched_dep_time	Scheduled departure times (format HHMM or HMM), local tz
dep_delay	Departure delays, in minutes. Negative times represent early departures
arr_time	Actual arrival times (format HHMM or HMM), local tz
sched_arr_time	Scheduled arrival times (format HHMM or HMM), local tz
arr_delay	Arrival delays, in minutes. Negative times represent early arrivals
carrier	Two letter carrier abbreviation
flight	Flight number
tailnum	Plane tail number
origin	Origin airport
dest	Destination airport
air_time	Amount of time spent in air, in minutes
distance	Distance between airports, in miles
hour	Time of schedule broken into hours

minute	Time of schedule broken into minutes
time_hour	Scheduled date and hour of the flight as a POSIXct date

Load the data into R or Python, and check for abnormalities (NAs). 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 containing the string "time". (head(), head())

a) (10 pts) Count the number of flights that departed NYC in the first week (first 7 days) of January and February combined. (filter(), query())
b) (10 pts) Print the year, month, day, carrier and air_time of the flights with the 6 *longest* air times, in descending order of air_time. (select(), arrange(), loc(), sort_values())
c) (10 pts) Add a new column to the dataframe; speed (in miles per hour) is the ratio of distance to air_time. Note that the unit of speed should be miles per hour. If you think they might be useful, feel free to extract more features than these, and describe what they are. (mutate(), assign())
d) (14 pts) Display the average, min and max air_time times for each month. (group_by(), summarise(), groupby(), agg()). You can exclude NAs for this calculation.
e) (16 pts) Impute the missing air_times as the distance divided by the average speed of flights for that destination (dest). Make a second copy of your dataframe, but this time impute missing air_time with the average air_time for that destination. 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. 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 package13232222332.

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 qtrRev:

Group	**Year**	**Qtr.1**	**Qtr.2**	**Qtr.3**	**Qtr.4**
1	2006	15	16	19	17
1	2007	12	13	27	23
1	2008	22	22	24	20
1	2009	10	14	20	16
2	2006	12	13	25	18
2	2007	16	14	21	19
2	2008	13	11	29	15
2	2009	23	20	26	20
3	2006	11	12	22	16
3	2007	13	11	27	21
3	2008	17	12	23	19
3	2009	14	9	31	24

The table consists of 6 columns; first showing the group number, second representing the year and the last four columns provide the revenue generated 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: Group, Year, Time_Interval, Interval_ID and Revenue.
Note: Here the entire Time_Interval column will contain value 'Qtr' since the dataset measures revenue every quarter. The Interval_ID will contain the quarter number.

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

Group	Year	Time_Interval	Interval_ID	Revenue
1	2006	Qtr	1	15