PREDICTION EXERCISE FOR SUMMER 25 INTERN CANDIDATES

Your goal is to predict the finishing time of dogs in UK greyhound races based on the results of their most recent previous race. You should seek to create a method that will (using your best judgement) minimize mean-square error on similar unseen races.

Each row in the df.csv dataset contains data from both the greyhound's previous race (the ***1 variables) and data from the race you are trying to predict (the ***2 variables). For all rows included

1. The dog in question has raced within the last 60 days
2. Both races (the previous race and the current one) were run at the same stadium

You may assume that the unseen races that you are trying to predict have these properties as well.

A historical dataframe (df.csv) of results is included so that you can analyze the data and build a model. Your submission should be set up to read in a new csv of similar data (unseendf.csv) with the same columns except that time2 (the variable you are trying to predict) has been redacted.

What you need to submit: An R or python script that will read in unseendf.csv as a dataframe, add a column named “predtime” that is your prediction of each dog’s finishing time and write out this modified dataframe as a CSV named “**mypred.csv**”. You will need to make sure that we can run the script submitted. If your script does not run on its own, we may not be able to review it!

- Your code must load all libraries that are needed.
- You can assume that the file "unseendf.csv" will be in the working directory.
- If your script requires the original historical file "df.csv" (for example, if your code builds a model based on the dataset and then uses that in some capacity during the script), you can assume that "df.csv" is also in the working directory.

df.csv: the historical dataframe you may use to analyze data, construct a model, etc
unseendf_example.csv: a very short version of the unseen dataframe that you may use to ensure that your code processes correctly. This will be of the same format as "df.csv" but with the time2 variable (that you are trying to predict) withheld
submission_example.py: very short example submission using python
submission_example.r: very short example submission using R

Description of Variables

stadium: Stadium that the race is being run at (note that you can assume that previous race was also run at the same stadium/track)
birthdate: Birthdate of the dog whose times/data are in this row
date1: Date of previous race
distance1: Distance of previous race in meters
trap1: Which "trap" (starting position) the dog started in in the previous race
time1: Finishing time of the dog in previous race
comment1: A brief description of the dog's previous race
date2: Date of current race
distance2: Distance of current race in meters
trap2: Which "trap" (starting position) the dog started in in the current time
time2: Finishing time of the dog in the current race (what you are trying to predict)