HW 4 - Do you even lift?

Packages

```
library(tidyverse)
library(tidymodels)
```

Data

Today, we will be working with data from www.openpowerlifting.org. This data was sourced from tidy tuesday and contains international powerlifting records at various meets. At each meet, each lifter gets three attempts at lifting max weight on three lifts: the bench press, squat and deadlift. The data dictionary for this dataset from tidytuesday is reproduced below:

Dictionary

variable	class	description
name	character	Individual lifter name
sex	character	Binary gender (M/F)
event	character	The type of competition that the lifter entered.Values are as follows:- SBD: Squat-Bench-Deadlift, also commonly called "Full Power".- BD: Bench-Deadlift, also commonly called "Ironman" or "Push-Pull"- SD: Squat-Deadlift, very uncommon.- SB: Squat-Bench, very uncommon.- S: Squat-only.- B: Bench-only.- D: Deadlift-only.

variable	class	description
equipment	character	The equipment category under which the lifts were performed.Values are as follows:- Raw: Bare knees or knee sleeves.- Wraps: Knee wraps were allowed.- Single-ply: Equipped, single-ply suits.- Multi-ply: Equipped, multi-ply suits (includes Double-ply).- Straps: Allowed straps on the deadlift (used mostly for exhibitions, not real meets).
age	double	The age of the lifter on the start date of the meet, if known.
age_class	character	The age class in which the filter falls, for example `40-45`
division	character	Free-form UTF-8 text describing the division of competition, like `Open` or `Juniors 20-23` or `Professional`.
bodyweight_kg	double	The recorded bodyweight of the lifter at the time of competition, to two decimal places.
weight_class_kg	character	The weight class in which the lifter competed, to two decimal places.Weight classes can be specified as a maximum or as a minimum. Maximums are specified by just the number, for example `90` means "up to (and including) 90kg." minimums are specified by a `+` to the right of the number, for example `90+` means "above (and excluding) 90kg."
best3squat_kg	double	Maximum of the first three successful attempts for the lift.Rarely may be negative: that is used by some federations to report the lowest weight the lifter attempted and failed.
best3bench_kg	double	Maximum of the first three successful attempts for the lift.Rarely may be negative: that is used by some federations to report the lowest weight the lifter attempted and failed.
best3deadlift_kg	double	Maximum of the first three successful attempts for the lift.Rarely may be negative: that is used by some federations to report the lowest weight the lifter attempted and failed.

variable	class	description
place	character	The recorded place of the lifter in the given division at the end of the meet.Values are as follows:- Positive number: the place the lifter came in.- G: Guest lifter. The lifter succeeded, but wasn't eligible for awards.- DQ: Disqualified. Note that DQ could be for procedural reasons, not just failed attempts.- DD: Doping Disqualification. The lifter failed a drug test.- NS: No-Show. The lifter did not show up on the meet day.
date	double	ISO 8601 Date of the event
federation	character	The federation that hosted the meet. (limited to IPF for this data subset)
meet_name	character	The name of the meet.The name is defined to never include the year or the federation. For example, the meet officially called `2019 USAPL Raw National Championships` would have the MeetName `Raw National Championshps`.

Exercise 1:

Let's begin by taking a look at the squat powerlifting records.

- Part A: Load the data from "ipf_lifts.csv" into a dataframe called `ipf`. Then remove any observations that are negative for squat and create a new column called `best3_squat_lbs` that converts the record from kg to lbs (you may have to google the conversion formula from kg to pounds). Save your transformed data frame as `ipf_squat`.
- Part B: Using `ipf_squat`, create a scatter plot to investigate the relationship between squat (in lbs) and age. Age should be on the x-axis. Add a linear trend-line. Be sure to label all axes and give the plot a title. Comment on what you observe (no more than 2 sentences).

Exercise 2:

- Part A: Write down the linear regression model that predicts/analyzes lift squat (lbs) from age (years) in X, Y, β notation. What is X? What is Y?
- Part B: Next, fit the linear model using the `ipf_squat` data frame.

- Part C: Re-write your previous equation replacing β with the fitted numeric estimates. This is called the "fitted" linear model.
- Part D: Interpret each estimate of β. Do you find the interpretations reasonable? (no more than 2 sentences) (Hint: look at the intercept)

Exercise 3:

- Part A: Building off of your `ipf_squat` data frame, create a new column called `age2` that takes the age of each lifter and squares it. Save your data frame with an appropriate name.
- Part B: Visualize squat in lbs vs `age2` and add a linear best fit line. Does this model look like it fits the data better? Is this still a linear model?

Exercise 4:

One metric to assess the fit of a model is the correlation squared, also known as R^2.

- Part A: Fit the age^2 model and save the object as `age2Fit`. Report the corresponding R^2 value.
- Part B: Report the R^2 value for the model that you fitted in question 2B. Do you prefer (based on R^2) the model in question 2B, or the one in question 4A?
- Part C: If you were to add body weight as a second predictor to the age^2 model, would R^2 increase or decrease? Explain. (This is a puirely conceptual question, no need to run any code with an actual second predictor.)

Exercise 5:

We want to use the `bodyweight_lbs` variable and the `sex` variable to try to predict best bench press (in lbs).

- Part A: Perform the necessary steps to convert any of bodyweight or bench press measurements from kg to pounds using the mutate function. Also convert the variable `sex` from a character to a factor using the `as.factor` function. (Look it up if you are unclear what `as.factor` does).
- Part B: Fit an interaction effects model with bodywieght (in lbs) and sex as predictors of best bench press (in lbs). Display the resulting coefficients.
- Part C: Interpret all the coefficients $\hat{\beta}$.

Exercise 6.

Do lifters who fail a drug test perform better or worse at bench press than other lifters? We'll answer this question in two parts.

- Part A: Remove all observations from the data frame that have `NA` listed under bench press. Create a new column called `doping_status` that takes value `doping` if the lifter failed a drug test and `not doping` otherwise. Save this data frame as `ipf_dope`. (you might want to consult the dataframe description of the variable `place` at the beginning of this assignment to find out about how to look for doping status)
- Part B: Using `ipf_dope` from the previous exercise, compute the 5%, 50%, 95% quantiles for bench press across both `sex` and `doping_status`. Use bench press in lbs here. With this information, answer the question "Do lifters who fail a drug test perform better or worse at bench press than other lifters? Is this consistent across sex and quantiles?"

Reminder:

- All plots should follow the best visualization practices: include an informed title, label axes, and carefully consider aesthetic choices.
- All code should not exceeding the 80 character limit.

Submission

- Go to Gradescope
- Click on your STA 199 course.
- Click on the assignment, and you'll be prompted to submit it.
- Mark all the pages associated with exercise. All the pages of your homework should be associated with at least one question (i.e., should be "checked"). If you do not do this, you will be subject to lose points on the assignment.
- For the "Workflow & formatting" question, select nothing.
- Make sure you push your finalized qmd file onto Github as well.

Rubric

- Ex 1: 6 pts.
- Ex 2: 8 pts.
- Ex 3: 5 pts.
- Ex 4: 6 pts.

- Ex 5: 12 pts.
- Ex 6: 8 pts.
- Workflow and formatting - 5 pts

> **Note**
>
> The "Workflow & formatting" grade is to assess the reproducible workflow. This includes:
>
> - linking all pages appropriately on Gradescope
> - committing the submitted version of your `.qmd` to GitHub
> - Are you under the 80 character code limit? (You shouldn't have to scroll to see all your code).