Qualifying exam questions

Instructions

You will need to load in the follwoing packages:

```
library(tidyverse)
library(Lahman)
```

Question 1

In this question you will be tasked with creating a data set to analyze in the next question. This data set will be constructed from the two data sets described below:

- `dat_chadwick`: A data set containing batting information on all balls in play from 2017-2022 with the 2020 season excluded. In particular we have the launch angle (angle of ball off bat) and the exit velocity (velocity of ball off bat) for each ball in play. This data set includes batter's name, the team they played for, the year. It also includes a unique batter identifier called `key_bbref`.
- `batters_fulltime`: A data set containing batting outcome information for full-time players from 2017-2022 with the 2020 season excluded. These outcomes are aggregates over a season. In particular we have the number of home runs hit by a player and the number of at bats (at bats are chances to hit a home run). This data set includes batter's name, the team they played for, the year. It also includes a unique batter identifier called `playerID`. This unique identifier is the same as the `key_bbref` identifier in the `dat_chadwick` data set.

These data sets are loaded in below:

```
dat_chadwick = read_csv("dat_chadwick.csv")
#View(dat_chadwick)

batters_fulltime = Batting %>%
  filter(yearID >= 2017, AB >= 550, yearID != 2020) %>%
  left_join(People %>% dplyr::select(playerID, nameFirst, nameLast)) %>%
  mutate(name = paste(nameFirst, nameLast)) %>%
  dplyr::select(name, playerID, teamID, yearID, HR, AB) %>%
  rename(year = yearID)
#View(batters_fulltime)
```

You need to transform the `dat_chadwick` data set to create features of the exit velocity and launch angle distributions for each player season. This transformed data set will be joined with the `batters_fulltime`. The goal of the analysis in the next question is to use the features of the exit velocity and launch angle distributions that you construct in this question to model home runs. You should name the final data set that you created in this question `bat`. You may restrict attention to complete cases, although that is neither required and it might not be necessary.

```
bat = bat[complete.cases(bat), ]
```

Question 2

In this question you are to model home runs. Your model should include year and features of the exit velocity and launch angle distributions that you constructed in the previous question. You should also add a term that isolates the Colorado Rockies (teamID: COL) because the high altitude of the stadium that this team plays its games allows for batters to hit home runs more often.