Homework 2: Data Wrangling

This homework will forcus on improving your data wrangling skills. In particular, this homework assignment will have an emphasis on working with the `dplyr` functions `mutate`, `group_by`, `summarize`, `filter`, and `count`. Moreover, there is also an emphasis on using the `lubridate` packages to work with dates.

For this week, the data you will be using is Michael Topper's Spotify data. The data is the universe of all streaming data from Michael's Spotify account for a one-year period. Here is a quick description of some of the columns:

- `end_time` - the time that the stream of a specific track ended.
- `artist_name` - the name of the artist.
- `track_name` - the title of the song.
- `ms_played` - the number of milliseconds the song was played for (NOT song length!).

We will be do some simple analysis on the data and explore some of Michael's streaming habits.

Wrangling

1. Read in the data using the `readr::read_csv` function.

2. Using the `lubridate` package, create 4 new columns: `end_time` which is the original `end_time` column, but in ymd_hm format, `hm` which will be the time of the `end_time` column in hours/minutes format (00:00), `end_date` which will be only the date of the `end_time` column in YYYY-MM-DD format, `seconds_played` which is the amount of seconds the song streamed for, and `minutes_played` which is the amount of minutes the song streamed for. *Hint*, for the `hm` column, you may want to use the package `hms`.

3. Find which artist Michael streamed the most in terms of number of songs streamed (doesn't matter whether the song played for only a couple seconds).

4. Find which artist Michael streamed the most in terms of minutes played. How many minutes played was the 5th most streamed artist?

5. What was Michael's favorite hour of the day to stream? What about his top 5 favorite hours?

6. Hopefully you noticed something weird in problem 5; the time seems incorrect unless Michael is an extreme night-owl. Note that the time zone is incorrect. Go here to figure out what time zone the data is using.

7. We need to change the time zone to Pacific Time. Using the `lubridate` package, change the `end_time` column and the corresponding `hm` and `end_date` columns to have the correct time zone.

8. Now find Michael's favorite hours of the day to stream. Do these seem more reasonable?

9. Let's define a "skip" as any song that Michael started listening to, but decided to stop listening to after less than 5 seconds. Create a new column called `skip` which is equal to 1 if Michael skipped the track, and 0 otherwise.

10. Re-analyse Michael's favorite hours of the day to stream and favorite artist (in terms of tracks played) while taking into filtering out these skips.

11. While we haven't covered graphing yet, I encourage you to try and make a fun graph of Michael's streaming habits.