Previous Questions

Data Source (URL web address with hyperlink)
https://spotifycharts.com/regional/global/weekly/2021-05-14–2021-05-21

Context of Data and Variables
This is the top 200 hot songs in the world in Spotify Charts from May 14, 2021, to May 21, 2021. It contains the following variables: song popularity ranking; artist; song title; song stream.

Output from str() function applied to the data object

```
str(spotify_top_200)

tibble [200 x 4] (S3: tbl_df/tbl/data.frame)
 $ Rank   : num [1:200] 1 2 3 4 5 6 7 8 9 10 ...
 $ Song   : chr [1:200] "good 4 u" "MONTERO (Call Me By Your Name)" ...
 $ Singer : chr [1:200] "Olivia Rodrigo" "Lil Nas X" "Doja Cat" ...
 $ Streams: num [1:200] 43029667 40805633 37723231 35908730 30358414 ...
```

Research Questions
The relationship between the number of singers on the charts and the song stream.
Who is the hottest singer or group this week?

Statistical Analysis Plan

1 Population

• From May 14, 2021 to May 21, 2021, the top 200 hits on the global Spotify charts.

2 Primary Objective

• Estimate the relationship between the sum of the song streams of each artist's songs on the chart this week and the number of songs on the chart.

3 Secondary Objectives

• Considering the number of songs on the chart and the weight of the song stream, evaluate which singer is the hottest singer this week.
• Estimate the change between song popularity ranking and song streaming.

4 Data Collection methods

• Crawl data through the browser
• The main purpose is to crawl the data from May 14, 2021 to May 21, 2021.
• Use R language for data crawling, which uses R language 'robotstxt'; 'rvest'; 'curl'; 'stringr'; 'tidyverse' data package.
• At the beginning of crawling data, the paths_allowed ('data source') command is used to observe whether crawling the data source complies with the webpage regulations.

5 Variables under consideration

• Steams(continuous variable) - Spotify classifies a single stream of a song when it has been listened to for 30 seconds or more. If you restart the song, whether by having it on repeat or clicking it again, it will count as another play after 30 seconds have been listened to again. If you listen to a song offline that you have saved in your library then each play will still count after 30 seconds the same. Spotify will track your offline plays and add them to their servers when you next connect to the internet. So if you end up skipping the end of a song, as long as you've listened for over 30 seconds it will still count.- **Primary outcome variable**
• Singer(categorical variable) - By counting the number of times the same singer has been on the charts during the week, the number of times each singer's songs have been on the chart. - **Primary explanatory variable**
• Ranks(order variable) - Popularity level between songs - **Secondary explanatory variable**
• Songs(categorical variable) - Song name

6 Missing data procedures

• If the two data of artist name and song stream are missing, the data will be deleted.

7 Numerical and graphical summaries to be presented

• Counts of number of missing observations to be given for each variable
• Count the number of times the singer appears, and calculate the sum of the corresponding song streams.
• The average and standard deviation of the song stream
• Box plot of song stream
• A scatter plot between the number of tracks on the chart by the same artist (x-axis) and logarithms of the sum of the corresponding song streams (y-axis) in a week.

8 Models to be fitted

• Primary objective:
Linear model (ANOVA)
logarithms of the sum of the corresponding song streams $\sim$ the number of tracks on the chart
After the model is fitted, observe whether the model significantly passes the coefficient of determination and F test.
• Secondary Objectives:
The hottest artist of the week is calculated by weighting, where different songs have different weights in the total song stream to calculate the artist level.
Song streams $\sim$ Ranks, model regression is used to predict the predicted values of music streams of different levels on the lower axis.