

SAS University Edition Challenge

Day 2 of 5

Challenge Overview

You've just concluded a series of successful interviews for a data scientist role at your dream job: A tech-startup that designs bespoke music playlists for businesses and individuals using freely available music data from Spotify.

As one final test of your suitability for the role, the company has asked you to spend one week helping them out with some tasks.

You read an article last week about SAS University Edition and believe it would be the perfect tool to get the job done. You'd be correct in that belief.

What are you waiting for? Let's get to work!

Day 2: **Based on a data set of the albums of the top 100 songs from 2014-2018, analyse what time of year is the worst for an album to be released in order for one of its songs to be in the top 100.**

Click the link below to access the data.

https://www.dropbox.com/sh/jfeh0bo2toohm36/AAD_FoSW2KbCkWa13_cpzHPAa?dl=0

Summary of Skills Demonstrated

- Filtering data
- Feature engineering
- Categorical variable analysis

Task 1

You want analyse the frequency of top 100 songs by release month. However, you first want to create a cleaner dataset to work with:

1. Filter your input data set for `release_date_precision = "day"`.
2. Remove the variables `uri`, `href` and `list_id`.
3. As there is no release month variable on the data, you need to create it using the `release_date` variable. Call this new column `release_month`.

Tips: All of the above can be done in a [DATASTEP](#).

1. Filter the input data set with a WHERE statement.
2. Use a DROP statement to remove certain variables from your output data set.
3. Either of the SAS functions [SCAN](#) and [SUBSTR](#) can be used to extract the release month from the character variable `release_date`.

Question 1: After filtering your data set for the above specifications, removing the unwanted variables and creating the *release_date* variable, how many observations and how many variables does your data set have?

Task 2

You now want to analyse what month, over the 2014-2018 period, has the least number of observations in our dataset.

Tips: If you prefer to analyse categorical data graphically, consider using [PROC SGPLOT](#) with a **vbar** statement. Or if you prefer to analyse categorical data using frequency tables then consider using [PROC FREQ](#).

Question 2: Based off our data, from 2014-2018 what is the worse month to release an album if you want one of its songs to make the top 100?

Task 3

You now want to group your analysis by year the album was released and analysis the amount of albums in each half of each year (group first 6 months and last 6 months together).

1. First you want to dedup your data set on **album_id** and **year** (i.e. you want to remove multiple occurrences of the same album_id-year combination from your data set).
2. Build off your code used for the previous step (or use alternate methods if you prefer), incorporating the *year* variable to group the data.

Tips:

1. [PROC SORT](#) can be used to remove duplicates of certain variable combinations using the **nodupkey** option.
2. You can format the values 01 – 06 to represent something like “H1” and values 07 – 12 to represent “H2” by creating your own format using [PROC FORMAT](#). This can then be incorporated into your PROC SGPLOT or PROC FREQ using a *format* statement.

Question 3: In what year (from 2014 – 2018) had the least number of albums in the first six months of the year that contained top 100 songs (after deduping on album_id for each year)?