SAS University Edition Challenge

Day 3 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 analyse the songs that featured on triple J's hottest 100 list between 2014 and 2018.

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!

Today the company has given you 2 tables to work with: **H_100_2014_2018.xlsx** and **Master_Items_Track.sas7bdat.** Click the link below to access the data.

https://www.dropbox.com/sh/rqmbuxiqht2yqpc/AADnnlCC5pPEMUI67Tet60sma?dl=0

H_100_2014_2018.xlsx is an excel spreadsheet. It lists the songs from Triple J's Hottest 100 between 2014 to 2018. **Master_Items_Track.sas7bdat** is a SAS dataset containing information from SPOTIFY about the songs on **H_100_2014_2018.xlsx.** Some data is missing from **Master_Items_Track.sas7bdat**, and some needs to be removed.

You have been asked to add columns from **Master_Items_Track.sas7bdat** to **H_100_2014_2018.xlsx,** and to perform some analysis on the resulting table.

Summary of Skills Demonstrated

- Importing data
- Filtering data
- Joining tables

Task 1

Import **H_100_2014_2018.xlsx** into SAS University Edition. Give the imported table the name: **HOTTEST_100**.

Using **Master_Items_Track** as the source, create a new SAS dataset called '**TRACK_INFO**:

- Keep the columns **Year**, **Order**, **Track_id**, **Explicit**, and **List_id**
- Create 2 new columns: **Duration_S** and **Duration_M**
    - **Duration_S** contains the song duration in seconds
    - **Duration_M** contains the song duration in minutes

- Remove rows with **list_id** equal to **'201696', '201697', '201698', or '201669'**

Tips for Task 1:

1. Use the *Import Data* utility in SAS University Edition to generate SAS code that will bring the H_100_2014_2018.xlsx file into SAS University Edition, and convert it into a SAS table.
2. Use the *Query* utility to generate SAS code that will make **TRACK_INFO**. You will need to edit the code to create **Duration_S** and **Duration_M,** and to ensure that the syntax is formatted correctly.
3. The **HOTTEST_100** table should have 500 rows, and 4 columns. The **TRACK_INFO** table should have 484 rows, and 7 columns.

Task 2

Create a new table by joining the **TRACK_INFO** table to the **HOTTEST_100** table on both the **Year** and **Order** columns. Perform the join such that only rows that match in both the tables are included in the new table. Give the resulting SAS table the name: '**H_100_TRK_INFO**'

Tips for Task 2:

1. Use the *Combine Tables* Data Task in SAS University Edition to generate SAS code that will join **TRACK_INFO** to **HOTTEST_100**.
2. Before running the code, you may need to edit it to ensure that the syntax is formatted correctly.
3. The **H_100_TRK_INFO** should have 484 rows and 9 columns.

Question 1: Which PROC SQL logical operator allows the existence of multiple join conditions?

Task 3

Use the **H_100_TRK_INFO** table to work out the following:

1. Find the total number of songs with explicit lyrics in each year's list.
2. Find the average song duration in minutes, in each year's list.
3. Find the songs with the minimum and maximum durations, in each year's list
4. Find the average song duration in minutes, of songs that contain explicit lyrics, and those that do not, in each year's list.

NB: Don't worry that each year is missing some songs. Just use the songs that are present in the **H_100_TRK_INFO** table

Tips: Use the *Query* Utility in SAS University Edition. You will need to decide how to group the data and which summary functions to use. You may also need to edit the code generated by the *Query* Utility, before running it.

Question 2: Which year's top 100 list contained the highest number of songs with explicit lyrics?

Question 3: Which year's top 100 list had the shortest average song duration?

Question 4: Which is the longest song (in minutes) of 2016 that contains explicit lyrics?

Question 5: Do songs that contain explicit lyrics have a longer, or shorter average duration, than songs that do not contain explicit lyrics?