Data Scientist Challenge

*Please read over the following questions and get back to us in your preferred format. We review*

*challenges anonymously so please do not include personally identifiable information in your submission.*

*The purpose of this challenge is not for you to come up with perfect answers but to show us your process*

*of tackling real-life, ambiguous data questions.*

**Question 1**

Listening behavior on playlists is analyzed using the following tables:

● **playlists** table with columns:

○ *user\_id* : the user who uploaded the playlist

○ *playlist\_id* : the unique identifier of the playlist

**● listings** table with columns:

○ *playlist\_id* : the unique identifier of the playlist

○ *track\_id* : the unique identifier of the track

○ *position* : the position of this track in the playlist

○ *track\_duration* : the length of the track recording (in milliseconds)

**● plays** table with columns:

○ *timestamp* : the time the play occurred

○ *user\_id* : the user who played the given track

○ *track\_id* : the unique identifier of the track

○ *playlist\_id* : the unique identifier of the playlist (set to 0 if the play did not happen

in a playlist)

○ *listening\_duration* : the duration that the user listened to the track (in milliseconds)

Using this schema, write or describe a SQL query that would help you find out the following:

● Given a user X, what are their 3 most listened to playlists?

● What is the percentage of users that create playlists only containing dj-sets? Assume a

dj-set is a track with a duration longer than 10min.

● For a user X, recommend 5 playlists they would like that they haven’t heard before. Go

for a practical over elaborate solution and stick to something that works well in SQL.

Please explain your logic.

**Question 2**:-

JSON-formatted dataset describing a sample of users’ listening

behavior on queues of tracks generated by some of SoundCloud’s recommender systems. We

would like to know what factors affect a listener’s choice to skip a recommended track.

In this dataset, each record (i.e., each sub-list of the “data” field) represents the total amount of

time the user spent listening to a track that had been recommended before a different track was

played or the listening session ended.

The provided attributes (in the “columns” field) are:

● *ts* : the time the user listened to the recommended track (as Unix timestamp)

● *country\_code* : the two-character country abbreviation where the user was located at the

time of listening

● *client\_version* : the version of the client that the listener was using

● *listening\_context* : the section of the SoundCloud web app where the user listened to the

recommended track

○ “collection”: any of the pages in the Collection section

○ “you”: the listener’s own profile page

○ “search”: any of the pages in the Search experience

○ “stream”: the page of the Stream experience

○ “personal-recommended’: the page of the Discover experience for personalized

recommendations

○ “charts”: the page of the Charts experience for trending content discovery

○ “tracks”: the dedicated page of an uploaded track

○ “users”: the profile page of another user (not the one listening)

○ “playlists”: the dedicated page of a playlist

● *recommender\_algorithm\_name* : the name of the algorithm that produced the track

recommendation for the listener

○ “content-based”: an implementation of content-based filtering

○ “collaborative”: an implementation of collaborative filtering

○ “hybrid”: a hybrid implementation of content-based and collaborative filtering

○ “fallback”: a static list of the popular tracks, used only as a fallback strategy

● *track\_id* : the unique identifier of the recommended track

● *track\_genre\_category* : the broad genre category describing the recommended track

● *track\_upload\_date* : the date the recommended track was uploaded (as Unix timestamp)

● *track\_duration* : the length of the recommended track recording (in milliseconds)

● *listen\_duration* : the duration that the user listened to the recommended track (in

milliseconds)

● *listener\_id* : the unique identifier of the user who listened to the recommended track

● *listener\_signup\_date* : the date the listener created their SoundCloud account (as Unix

timestamp)

● *listener\_top\_genre\_category\_listened* : the genre category that the user has most

frequently listened to (based on all their past listening in all contexts)

● *listener\_prev\_month\_listening\_time* : the total duration that the user listened in the

previous month (in all contexts)

● *listener\_prev\_month\_avg\_daily\_tracks\_listened* : the daily mean number of distinct tracks

that the user listened in the previous month (in all contexts)

Using this dataset and a programming language of your choice, analyse user behaviour with

regards to **skipping** a track.

We expect you to provide us with two outputs:

● First, a thorough exploratory analysis of the data. Make sure you elaborate on your

approach in reasonable detail, especially on your assumptions and simplifications. We

expect the output to be appropriate for a fellow data scientist to peer review, rerun your

code and understand your thought process.

● Second, a short presentation with recommendations for the product manager.