PROBLEM STATEMENT: Suppose you’re an employee at Saavn as a Big Data professional working closely with the company's ML team. For better user engagement, you’re required to build a system that keeps the users updated based on their music preferences. Suppose, a new track of some particular artist has been released. Now, your responsibility would be to push the notification about this song to the appropriate set of audience. For instance, Baadshah’s new track “Tareefan” is released. Now, you would probably like to send notifications about this song to the users who prefer to listen to singers like Honey Singh and Raftaar than to users who prefer listening to singers like Jagjit Singh. Pushing a 'rap song' notification to an admirer of classical music is irrelevant. The user may get annoyed at some time and may even uninstall the app.

CODE FLOW DESCRIPTION:

1. Set the logging mechanism to log only errors in the console.

2. Create a spark session at local. Later when to be run on EC2 can be changed to master.

3. Load data from notification_clicks, newmetadata, notification_actor and the sample100mb.csv

4. We do a transformation and cleaning of data from the datasets.

5. Evaluate the Recency, that measures how recently a user last listened to a particular song.

6. Frequency, that measures frequency of the number of times a song was heard by performing a aggregation.

7. Create the dataset with song indexed.

8. We assemble features such as recency, frequency and last_lisen in the form a vector and storing it in assembler.

9. We initialize the K-Means model thereby indicating that we are building 5 clusters as requirement. This model then transforms dataframe to create a new dataframe named as predictions.

10.We evaluate the userId and artist Id on which cluster they get into.

11.We create a dataset that indcludes predictions with artitist_id its popularity followed by the windows rank.

12.Here we establish a dataset which includes artitist_id,count, the rank, prediction and the cluster user count.

13.Thereafter we establish a relationship between artitist_id, cluster user count, notification Id and user notification count.

14.We establish the Click through ratio by taking in effect prediction_user_count dividing it by user_notification_count

15.Finally evaluate clustering by computing Silhouette score and show the results.