Milestone 1

Music Recommendation System for Spotify

Anoop Singhal

Problem Definition

Objective

- Improving customer satisfaction of Spotify users
- Increasing popularity of Spotify

Context

- Providing relevant recommendations as per user’s existing likings as well allowing them to explore new songs
- Recommending most trendy songs to the new & existing users.
- Addressing cold start problem

Key questions

- What should be threshold rating above which item recommendation is made
- Rating Data is 7% of the total possible ratings by all user for all items, this will affect the quality of ratings estimated
- How to reduce amount of data for decreased model training time
  - Remove users with less than 90 songs listened
  - Remove songs listened by less than 120 users
  - Play count
    - Remove play counts more than 5
    - Flatten play count greater than 5 to 5 - Instead of dropping all interations (play_count) > 5 we can cap the values to 5. This way we do not have to reduce the dataset and also keeping the relevant interactions without biasing the recommendations towards songs played too many times by a user.

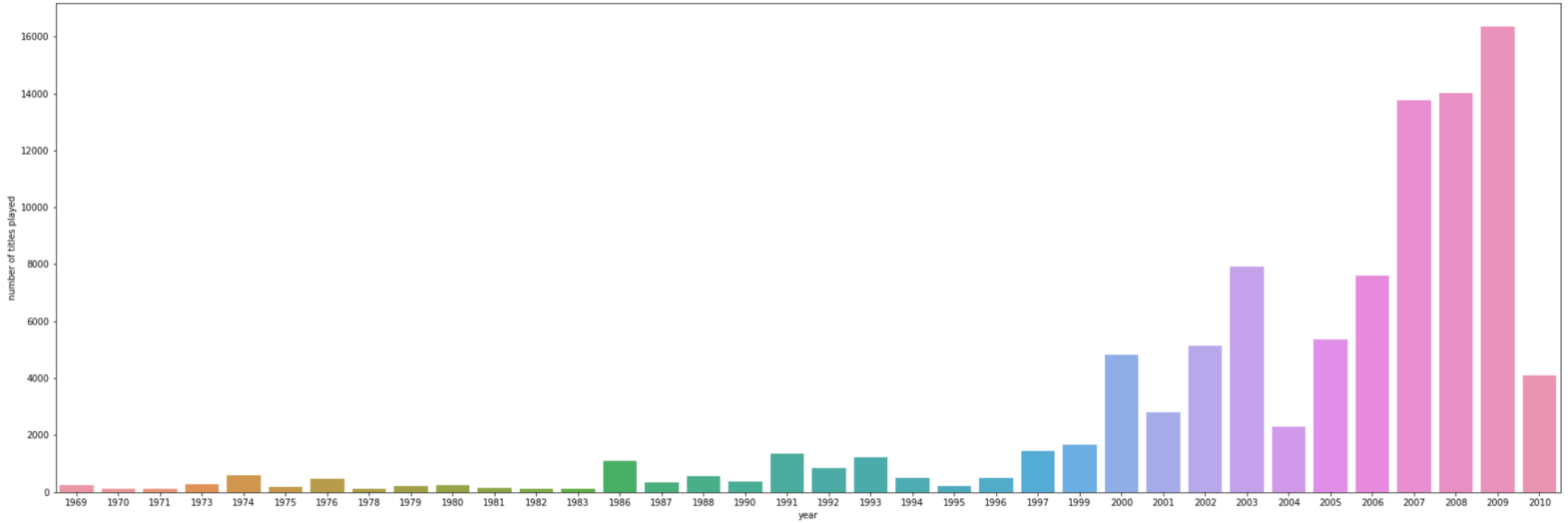
Data Exploration

Data Description

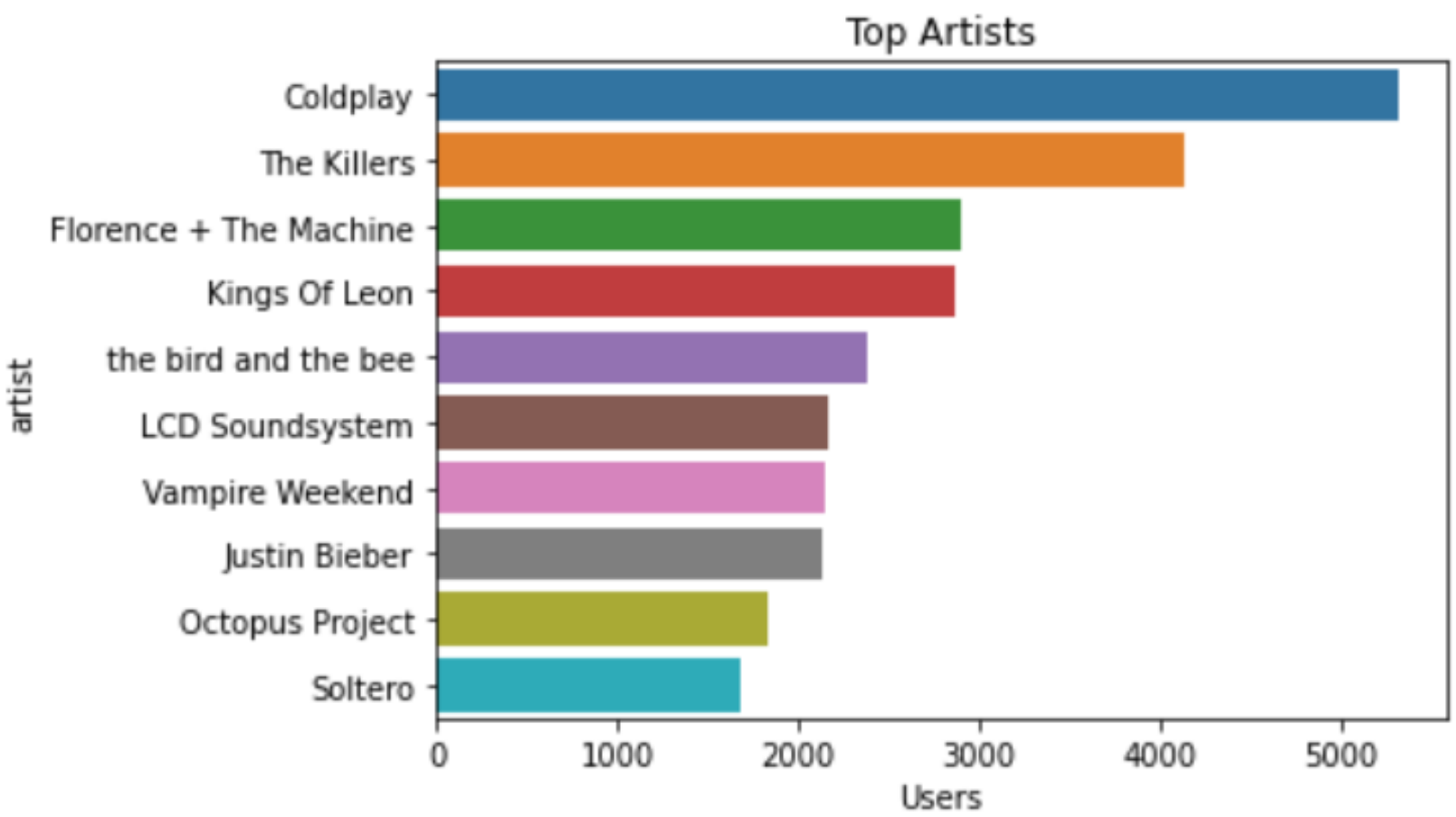
- Total number of user 3155
- Total number of songs 563
- Possible number of rating 3155 * 563 = 1,776,265 Vs. available ratings 117,876 which is about 7%

Play count by Year

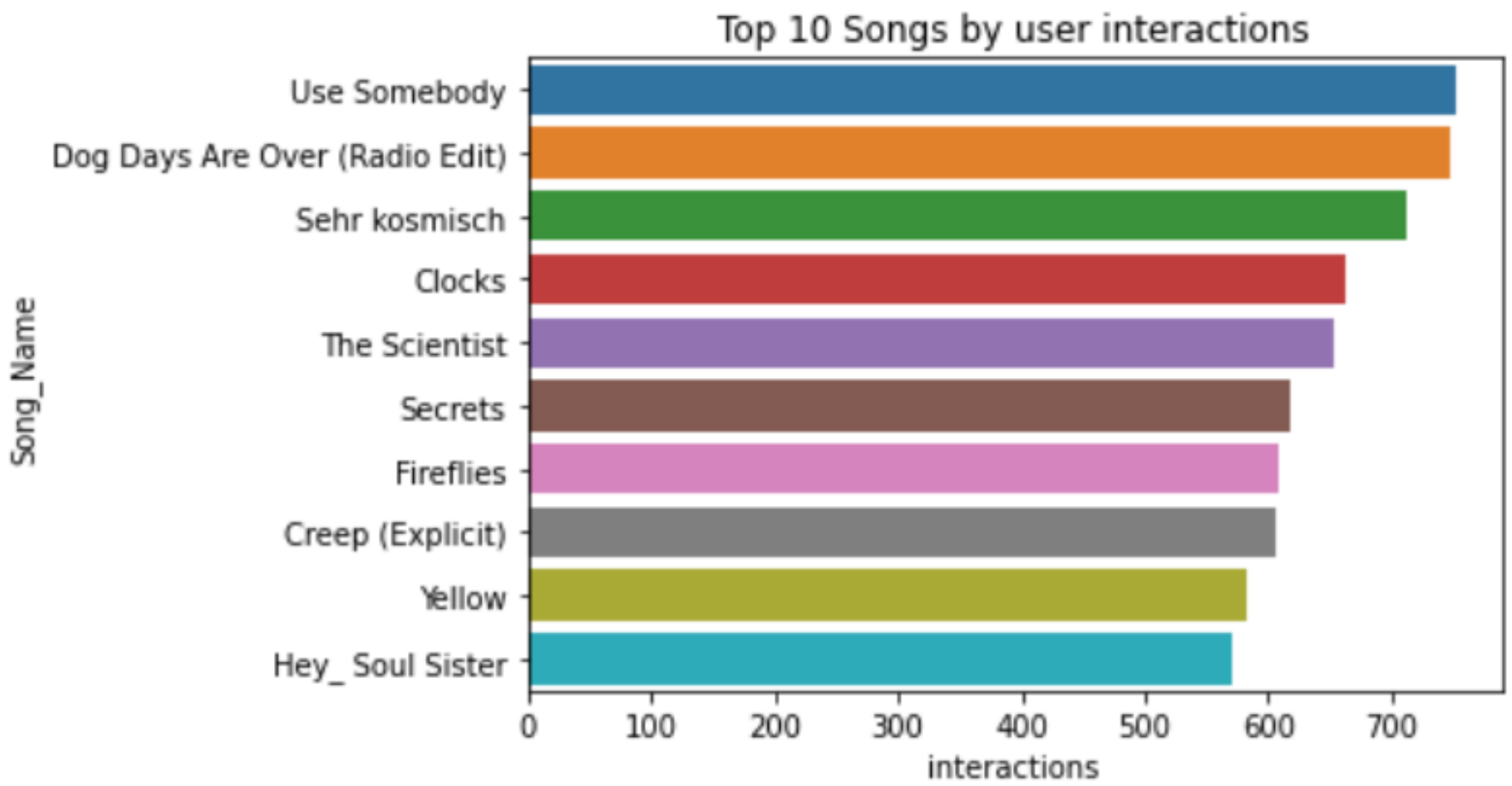
- Titles/albums released in recent years are more listened as compared to songs listened to old releases
  - This could mean that people are liking the latest song releases more than the older albums/titles
  - Spotify songs app is probably used mostly by younger generation. This observation is based on the fact that younger generation tend to like latest songs.



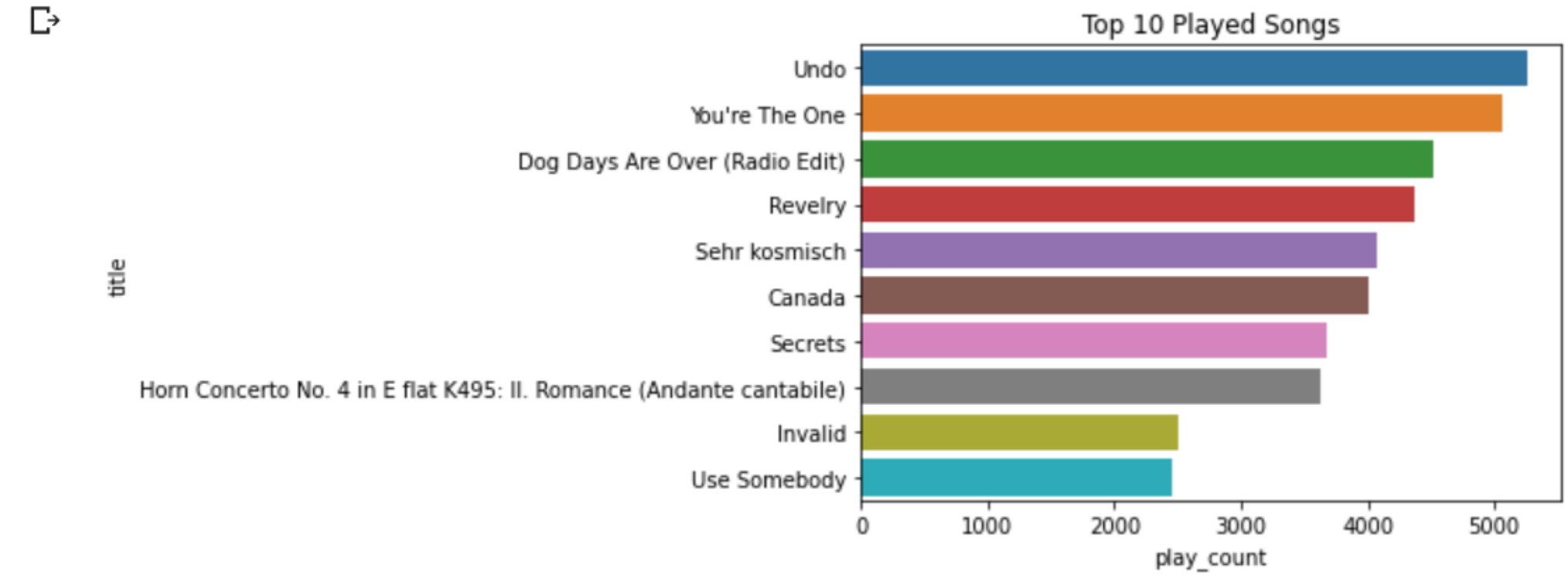
Top 10 Artists by Play Count



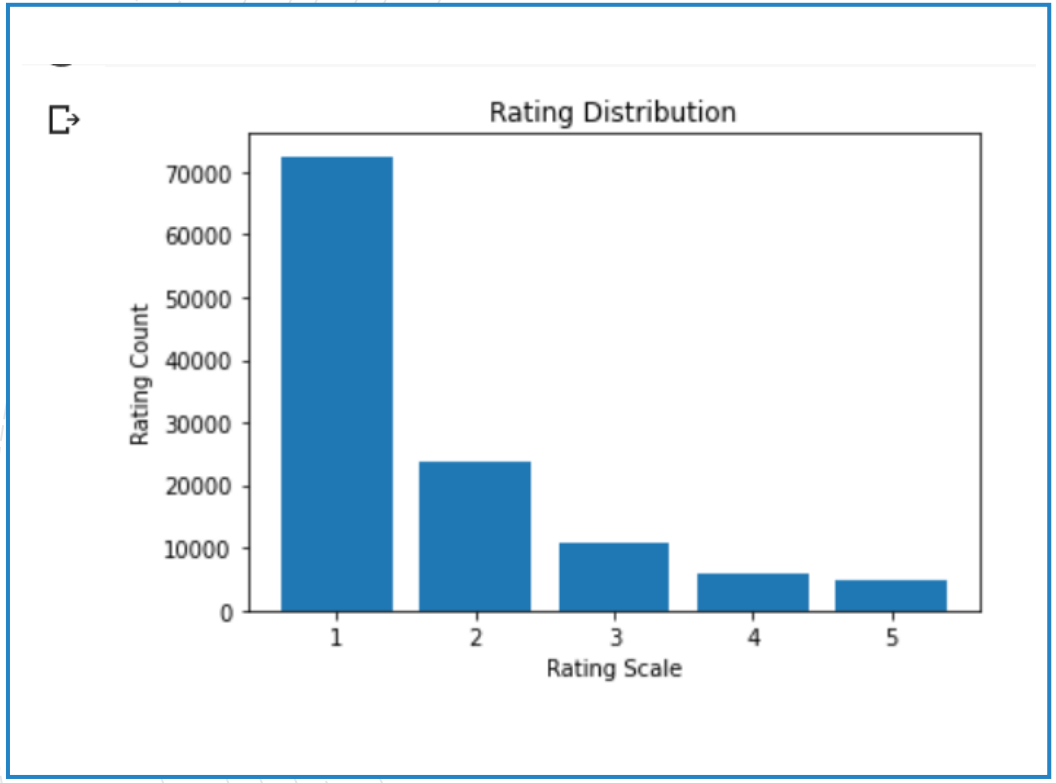
Top 10 Popular Songs



Most Played Songs



Rating Distribution



- Ratings are more skewed towards lower side. It means users are tough raters on Spotify app
- Threshold rating for recommending a song must be about 1.5 based on the rating distribution plot

Proposed Approach

Potential Techniques

- The play count is considered as rating by the users. More the play count more it is liked by the user

- The **rating scale** is set to (1,5). This is done by dropping the interactions where play count is greater than 5.

- **Threshold** rating for recommending a song must be about 1.5 based on the rating distribution plot

Potential Techniques

- How to remove effect personal rating scale
  - Zero mean normalization
  - Z-Score normalization in addition also considers the variation in ratings given by user
- Finding k similar neighbors for rating prediction
  - Pearson Baseline similarity measure (instead of cosine similarity) in CF removes the personal effect on rating.

Overall Solution Design

- We will deal this as a regression problem (predicted rating 1-5) rather than a classification problem (like/dislike). This would help us to recommend songs in decreasing order of predicted rating
- Following models can be developed based on the available data for recommending songs
  - Collaborative Filtering models
    - User-User similarity CF
    - Item - Item similarity CF
- Model Based approach
- Clustering based approach
- Content Based Filtering (using song information)

Measures of Success

- RMSE to be minimized
- F-Score to be maximized – as both recall and precision are important in our case
  - Recall - TP/(TP + FN) - %age of relevant recommendations made
  - Precision - TP/(TP + FP) - %age of recommendations made that are relevant