Music Recommendation System

Milestone : 1

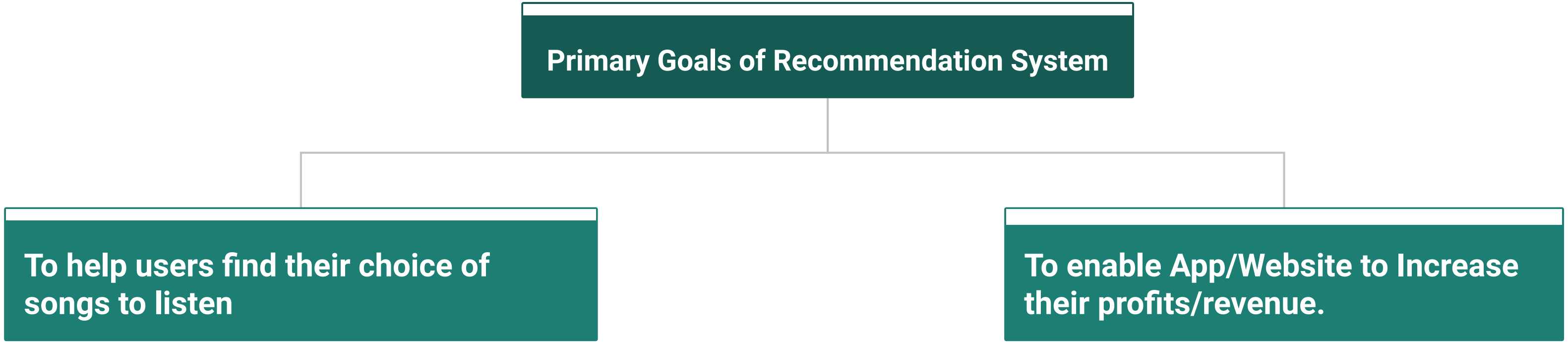
Capstone Project By:
Rutuja Sanjay Dhanawade

Problem Definition & Background

According to spotify, every year approximately 22 million tracks are generated and 60,000 new tracks are ingested by its platform every single day!

On the other hand, users have a variety of choice for music.

Under such circumstances, a best way for a company to attract it’s customer is by providing personalized music recommendation service to their customers.

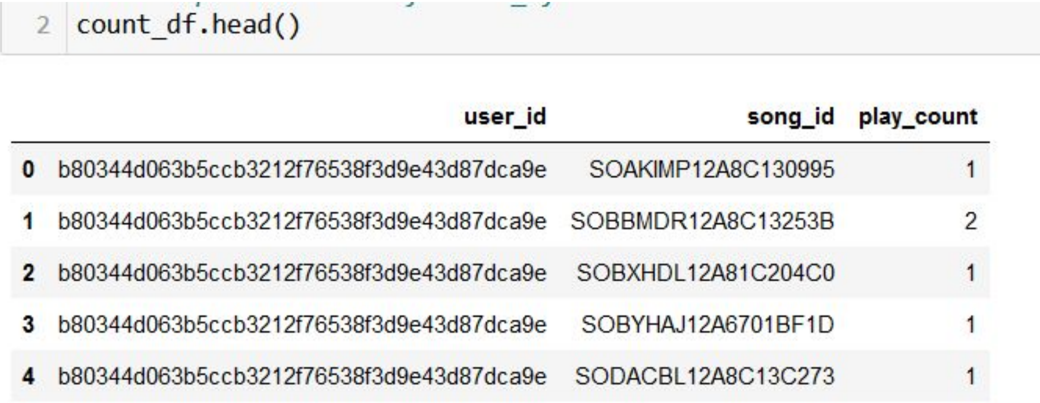


Primary Goals of Recommendation System

To help users find their choice of songs to listen

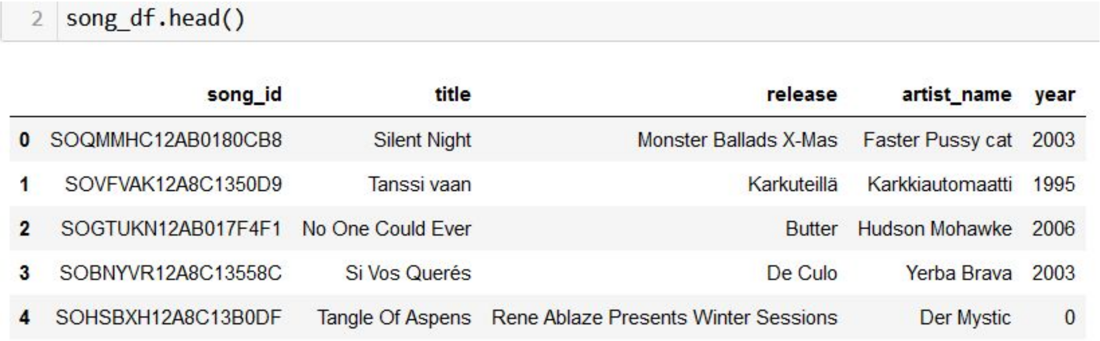
To enable App/Website to Increase their profits/revenue.

Exploratory Data Analysis



```
count_df.head()
```

	user_id	song_id	play_count
0	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOAKIMP12A8C130995	1
1	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBBMDR12A8C13253B	2
2	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBXHDL12A81C204C0	1
3	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBYHAJ12A6701BF1D	1
4	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SODACBL12A8C13C273	1



```
song_df.head()
```

	song_id	title	release	artist_name	year
0	SOQMMHC12AB0180CB8	Silent Night	Monster Ballads X-Mas	Faster Pussy cat	2003
1	SOVFVAK12A8C1350D9	Tanssi vaan	Karkuteillä	Karkkiautomaatti	1995
2	SOGTUKN12AB017F4F1	No One Could Ever	Butter	Hudson Mohawke	2006
3	SOBNYVR12A8C13558C	Si Vos Querés	De Culo	Yerba Brava	2003
4	SOHSBXH12A8C13B0DF	Tangle Of Aspens	Rene Ablaze Presents Winter Sessions	Der Mystic	0

There are 2 tables:

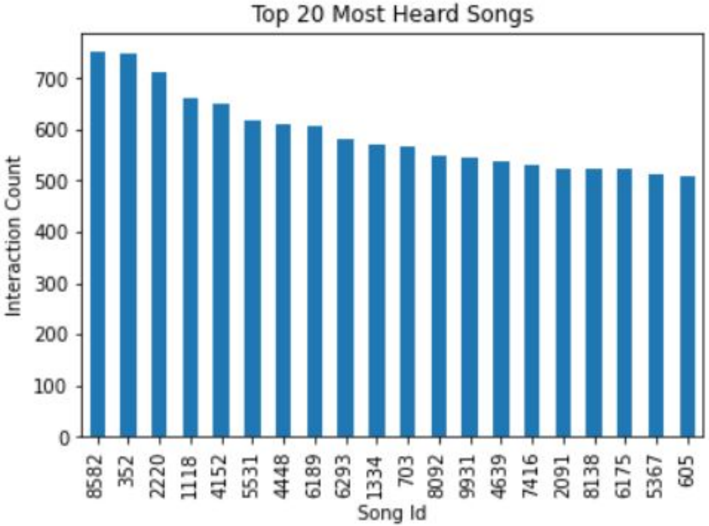
count_df with a total number of 2000000 records.

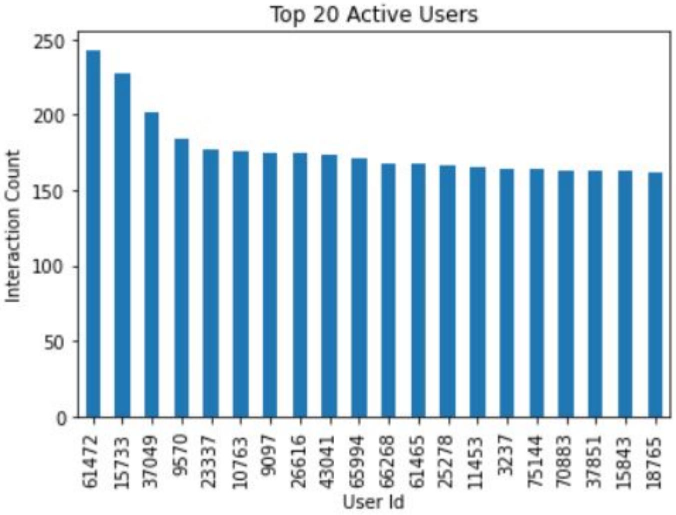
This dataset contains interaction of songs and users.

Song_df with a total of 1000000 records

This dataset contains all the song details

Exploratory Data Analysis





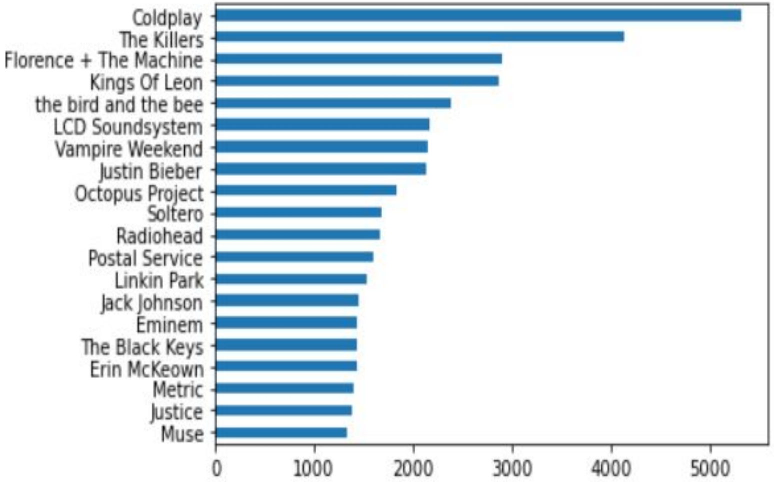
This figure indicates top 20 most heard songs.

Note: Song Id is encoded.

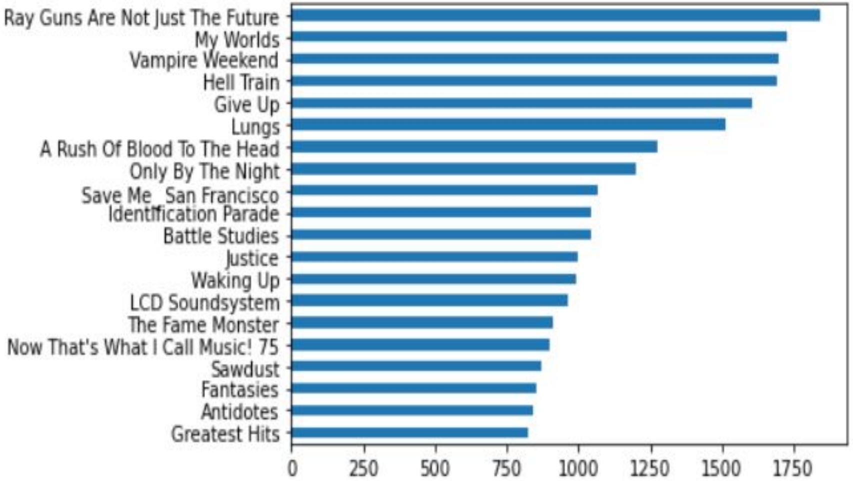
This figure indicates top 20 most active users

Note: User Id is encoded.

Exploratory Data Analysis



Top 20 Most Popular Artists



Top 20 Most Popular Albums

Exploratory Data Analysis

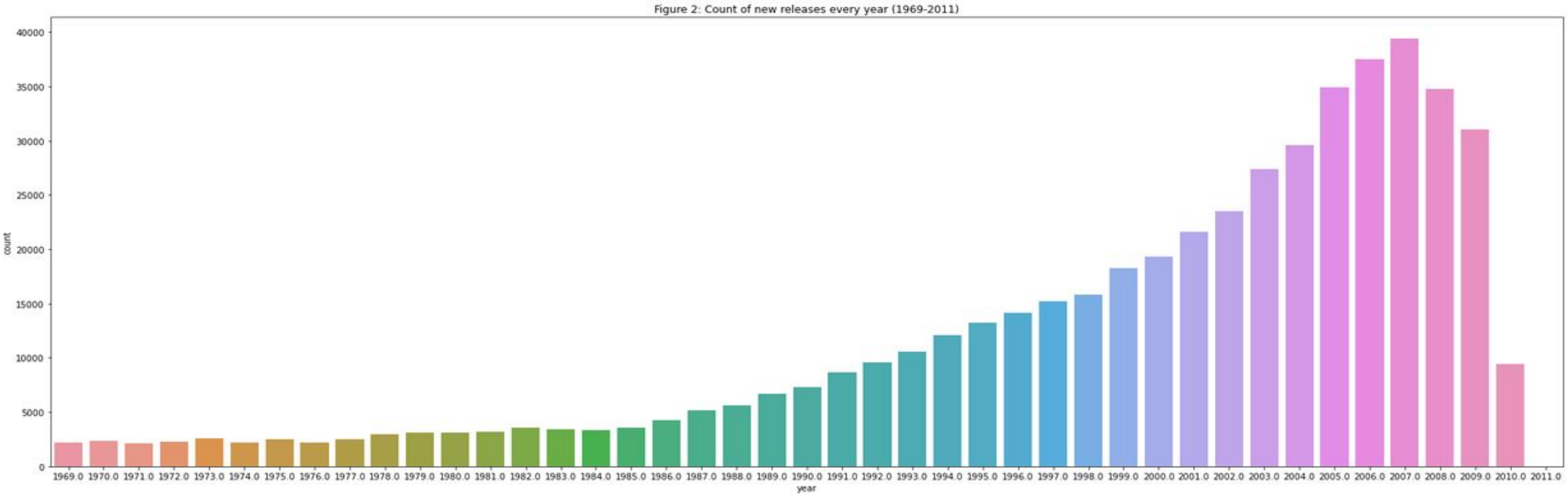


Figure 2: Count of new releases every year (1969-2011)

The above figure represents the number of albums released in each year.

Till 2007, one can see a steady growth in the number of albums released every year, but there is a drop in number of albums released from 2008 to 2010.

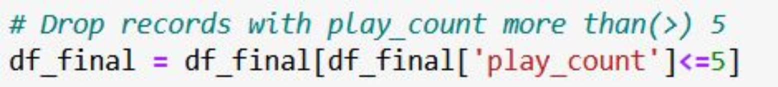
This can be due to the Great Recession of 2007-2009

Data Pre-Processing Methods

In this dataset, we focus on the ‘play_count’ column as it captures the interaction between the users and the songs they listen. The detail explanation is provided in the jupyter notebook for these pre-processing methods.

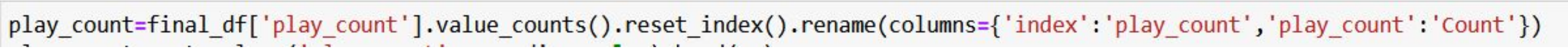
Now I propose 3 main ways to process this ‘play_count’ column:

1. Dropping records with play_count value > 5:

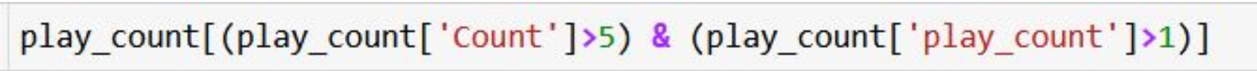


```
# Drop records with play_count more than(>) 5
df_final = df_final[df_final['play_count']<=5]
```

2. Filtering data based on count (occurrence) of play_count values and considering only ‘play_count’>1

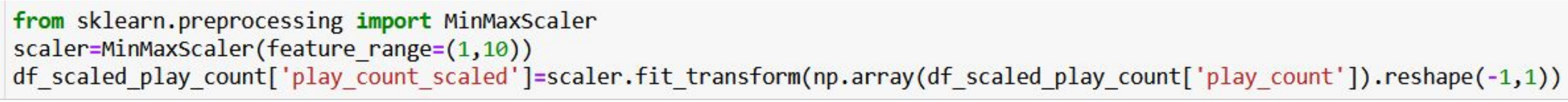


```
play_count=final_df['play_count'].value_counts().reset_index().rename(columns={'index':'play_count','play_count':'Count'})
```



```
play_count[(play_count['Count']>5) & (play_count['play_count']>1)]
```

3. Normalizing the play_count values which lie in the range of (1-2213) to (1-10)



```
from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler(feature_range=(1,10))
df_scaled_play_count['play_count_scaled']=scaler.fit_transform(np.array(df_scaled_play_count['play_count']).reshape(-1,1))
```

Approaches for Recommendation System

The given data majorly contains information related to user interaction with respect to listening songs. Therefore, one can implement various types of **Collaborative Filtering** with this kind of dataset.

There is no information provided based on song genre/ the users features. In such a case, one could proceed with **Content based Recommendation System** approaches as well.

I intend to solve this problem, by evaluating the performance of following algorithms after pre-processing data the data by above 3 methods. Each of the above 3 pre-processing techniques would be evaluated against following algorithms.

1. **Rank Based Recommendation System (Useful to deal with the problem of Cold Start)**
2. **Collaborative Filtering Method**

   a. User-User Based Similarity
   b. Song-Song Based Similarity

3. **Model based Recommendation System using Matrix Factorization (SVD)**
4. **Graph Neural Networks (Optional)**

The combination of data pre-processing and algorithm which gives maximum R-squared/RMSE and Precision would be selected as a final model.

Thank You!!

Any Questions?

Rutuja Sanjay Dhanawade
dhanawade.r@northeastern.edu
+1 857 437 9192