Follow the Beat

Marcelo Pereira

May/2022

How can I help you?

Recommendation systems are part of consumers' life. These systems will recommend new products or services based on previous interactions or features of offered products. Using Spotify data, I explored 4 recommendation systems, comparing them against tracks suggested randomly.

1. Context

- E-commerce and streaming services rely heavily on recommendation engines. Even traditional businesses can take advantage of these models to improve their customers' experience
- Using collaborative filtering, I created recommendation systems based on the tracks' popularity and the co-occurrence of artists, albums, and tracks
- I selected playlists of Spotify users from the US created between 2018-2021 to perform this task.

2. Criteria for success

- For each playlist, a part of the tracks was used as a seed to select the remaining, so they were compared to the ones previously existing in the playlist, then we counted the matches

3. Scope of solution space

- The models are dedicated to songs suggestion and could be applied in any system with similar data available

4. Constraints within solution space

- Reduced data sample compared to the universe of information that is available for the music industry

5. Stakeholders to provide key insight

- Business leaders willing to improve the model of business by including data-driven decisions

6. Key data sources

- API Spotipy:
- https://spotipy.readthedocs.io/en/2.19.0/
- https://developer.spotify.com/documentation/web-api/

Data source

The data for the model was acquired using the API Spotipy:

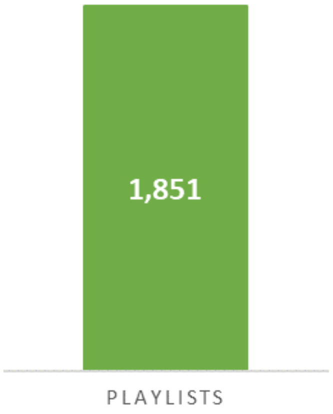
- https://spotipy.readthedocs.io/en/2.19.0/
- https://developer.spotify.com/documentation/web-api/

- Population target:
  - US users
  - Time period: 2018-2021

The data is hierarchical based on:

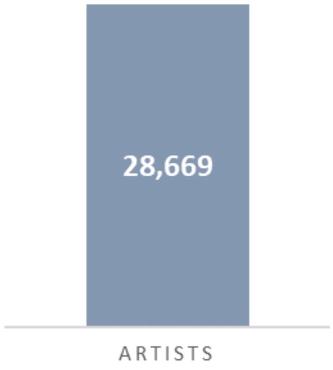
- User ID
  - Playlist ID
    - Tracks
      - Track ID
      - Track name
      - Artists
        - Artists ID
        - Artists names
      - Album
        - Album ID
        - Album name

Exploratory data analysis (EDA) – extracted data



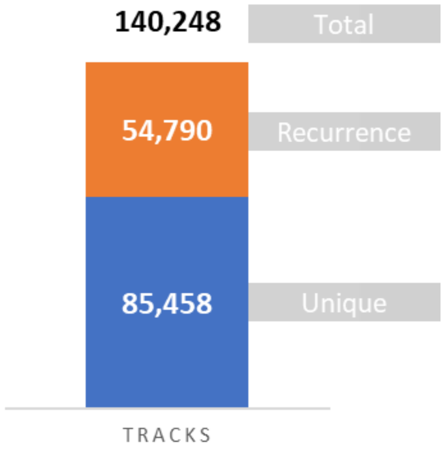
Playlists

- Tracks per playlist:
  - Max: 100 (Totaling: 48% of all playlist
  - Mean: 75.8
  - Median: 98.0
- 4% (75) with 20 or less tracks
- Artists/playlist: 15.5
- Albums/playlist: 33.5



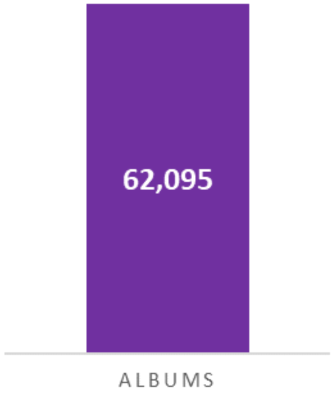
Artists

- 15,609 artists (54.4%) appeared only one time
- "Various Artists" the most frequent
- Drake is present in 877 playlists
- Albums/Artist: 2.2



Tracks

- 67,864 tracks (79.4%) appeared only one time
- Most popular track:
  - "Invisible" from the artists Andra and Lil Eddie - appeared 92 times
- On average each track appeared 1.6 times

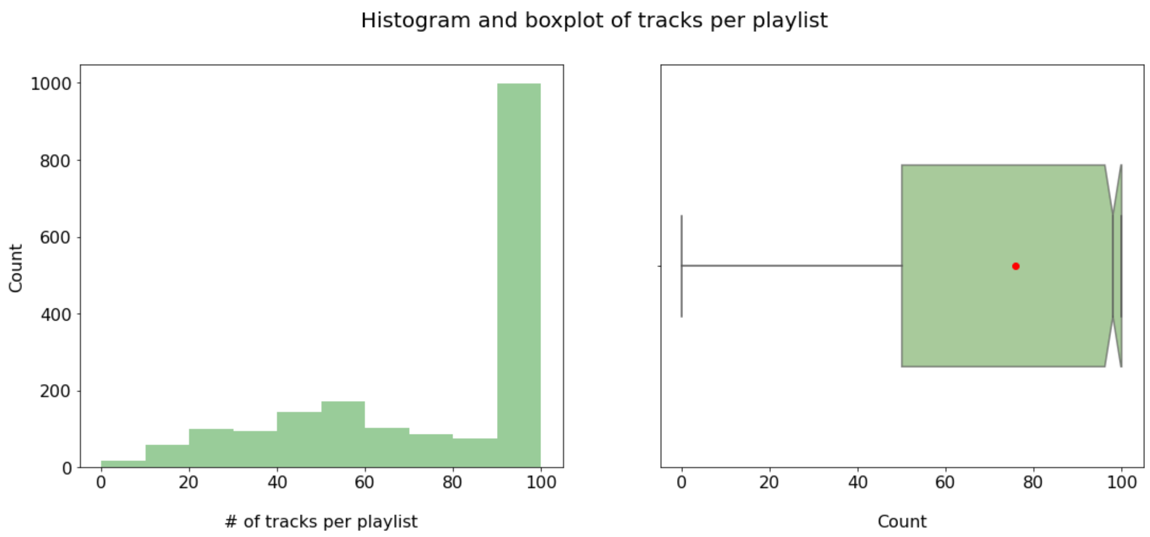


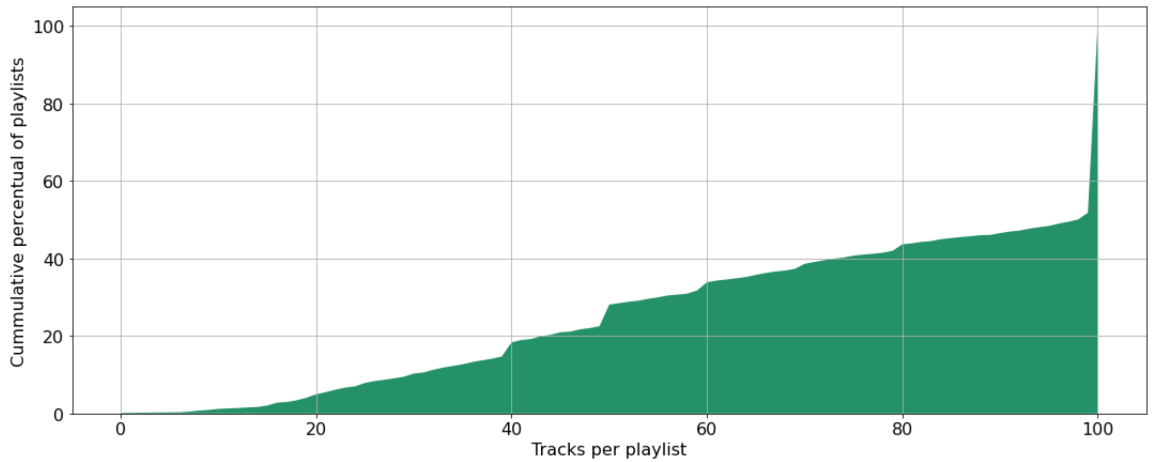
Albums

- 43,355 albums (69.8%) appeared only one time
- Album Beerbongs & Bentleys appeared 250 times

EDA – Playlists distributions

The majority of playlists have 100 tracks. Only 4% have less than 20 tracks



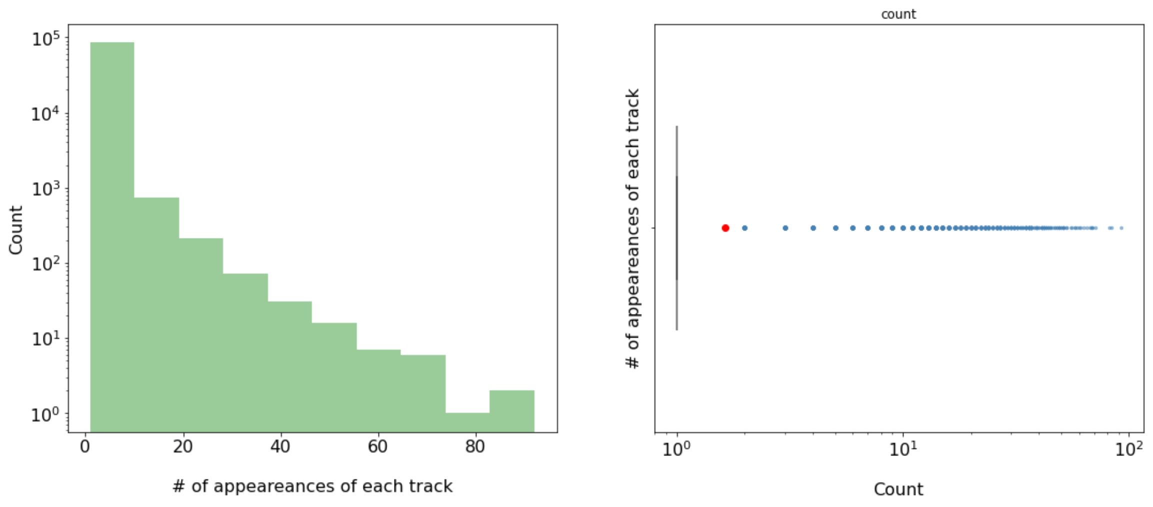


- The histogram and the cumulative percentage of tracks per playlist showcase that the playlist usually has 100 tracks
- The playlists with less than 20 tracks will be removed once they represent only 4% of the total

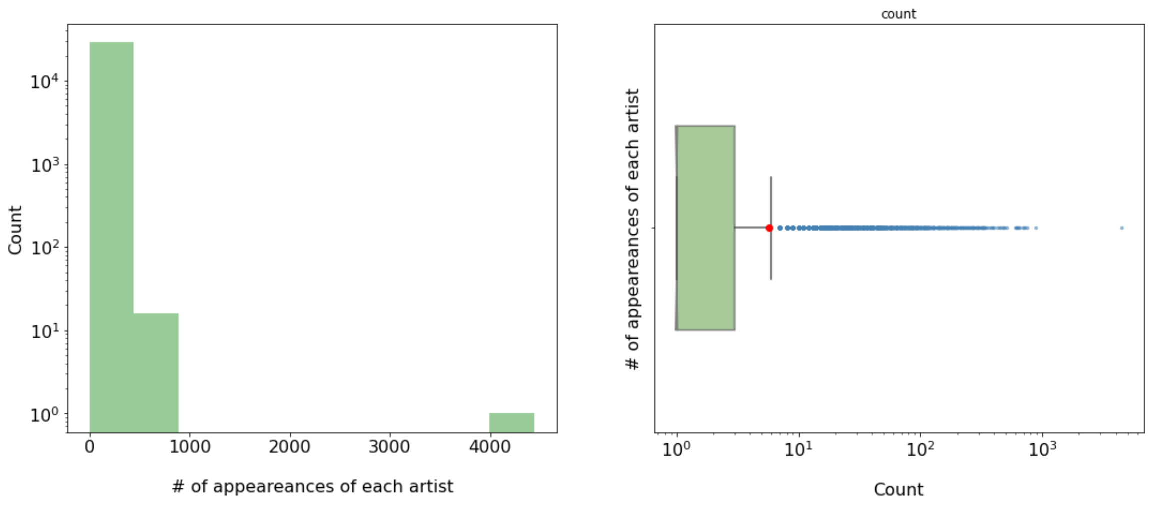
EDA – Tracks, artists, and albums

As seen in the histograms and boxplots (both in log scale), the tracks, artists, and albums appear just one time

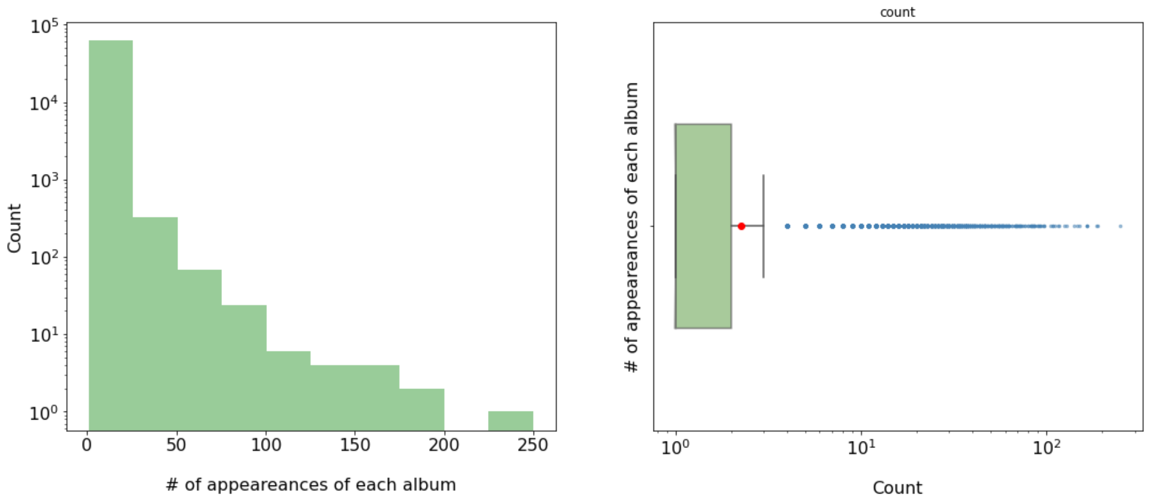
Histogram and boxplot of tracks count (log scale)



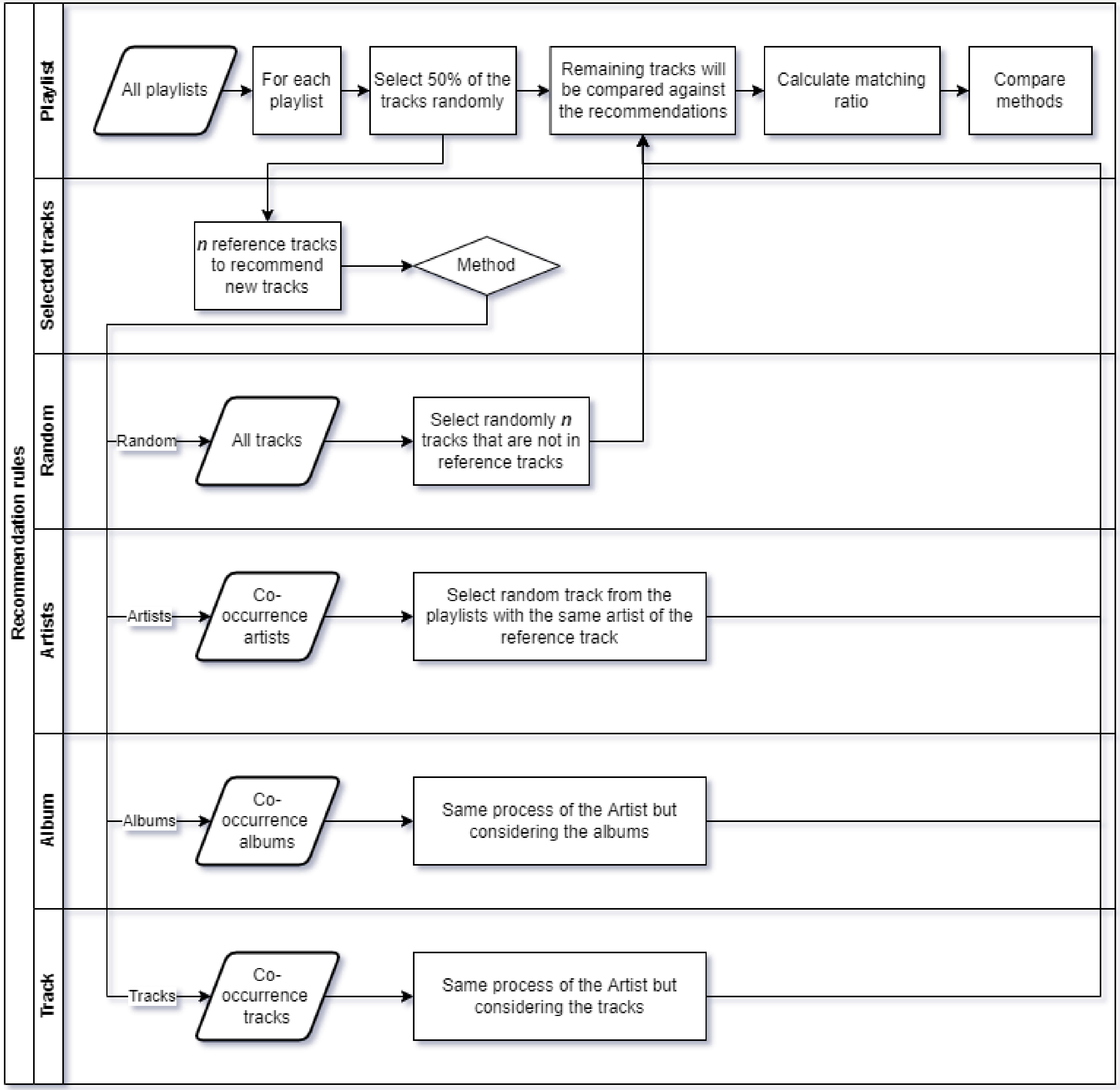
Histogram and boxplot of artists count (log scale)



Histogram and boxplot of albums count (log scale)

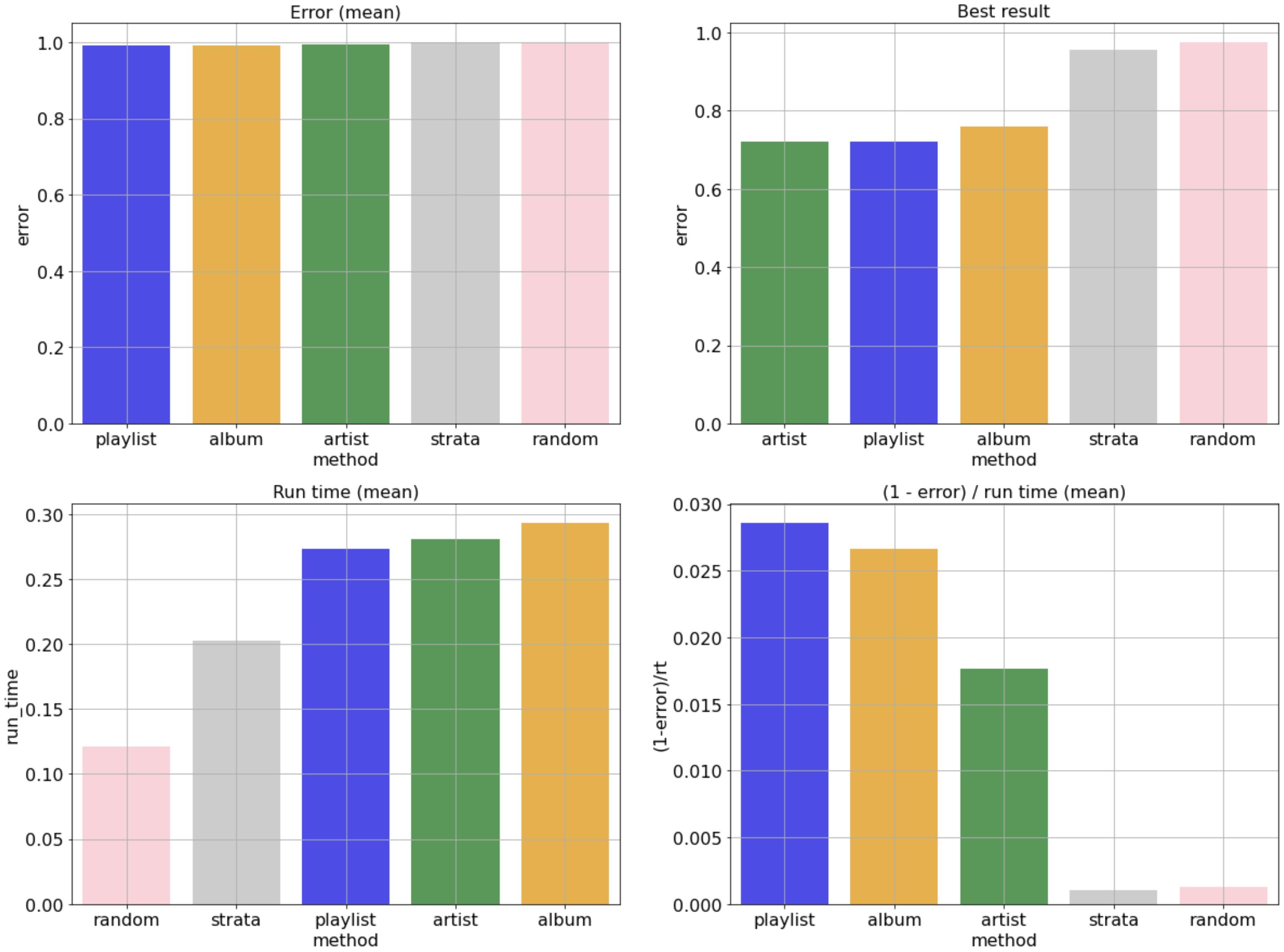


Model flowchart



Modeling - results and analysis

Playlist and album-based recommendation models with 99.1% of error tied in the lead



- Winner: **playlist** and album with 99.13% and 99.16% of error, respectively
- Single run best result: **playlist** and **artist** models with 72% of error
- The **random** model is the fastest
- The **playlist**-based model had the best accuracy per unit of time to run

Summary and conclusion

The performance of the models is poor, but the results pointed in the expected direction

Considerations

- The best performance for a single run was 38% of correct prediction
- Scenarios tested considered just one dimension at each time
- The selection criteria had a relatively reduced sample size, and no advanced ranking criteria were included
- Cold starts are considered the same way as repeating songs

Conclusion and alternatives

- The performance of the models is far from good, but they showcase that this is a direction that could be explored by using users' knowledge by selecting their playlists and the songs' features
- More complex strategies, including ranking techniques and combining multiple dimensions, must improve the results
- Some solutions can have a longer processing time if we consider creating pre-loaded selection lists
- Additional analysis is required to evaluate the initial sample size to the model performance

Contact:

Marcelo Pereira

https://www.linkedin.com/in/marcelo-alves-pereira/

map_fm@yahoo.com