**Spotify Project**

**Team 20**

**Ankit Bhargava, Erik Subatis, Mark McDonald**

**TA Advisor**: Rashmi ([rjain29@gmail.com](mailto:rjain29@gmail.com))

**Project Repo**: <https://github.com/subatis/CS109a_finalproject_group20>

**Project Database**: <https://drive.google.com/drive/folders/14OBw3t3gKwPgxX3tx_ogHHpXpx6iWXRC>

**Table of Contents:**

[Summary](#summary)

[Data Structure](#data_structure)

[EDA](#eda)

[Models](#models)

[Scores and Comparisons](#scores_comparisons)

[Summary](#summary)

**Summary** [(back to index)](#table_of_contents)

<brief summary of the project >

**Data Structure** [(back to index)](#table_of_contents)

<explain data received and how it was organized>

**Database**

The project data has been stored in a database for uniformity, speed and simplicity of access. In addition to the playlists provided, tables were added to include additional data for tracks and artists.

**Utility Functions and Project API**

A set of API wrappers were created to streamline access to Spotify data and the Team Database. These API’s simplify Spotify authentication and avoid the need to navigate arguments that are not relevant to the project.

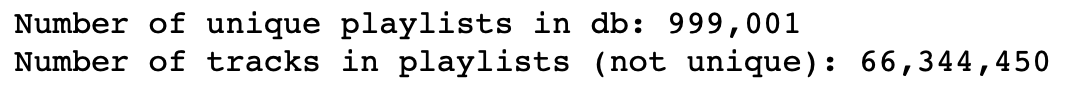
**Data Inconsistencies and Corrections**

Various records needed to be deleted or changed after reviewing their validity. Some artist URL’s had changed in Spotify and some tracks were no longer available. These changes were not extensive and did not have a major impact on the usefulness of the data.

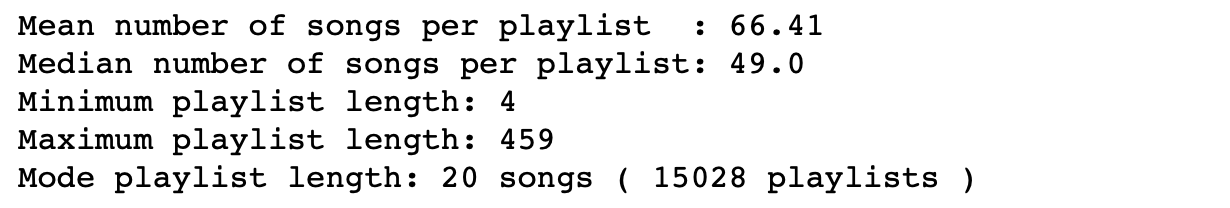
**EDA** [(back to index)](#table_of_contents)

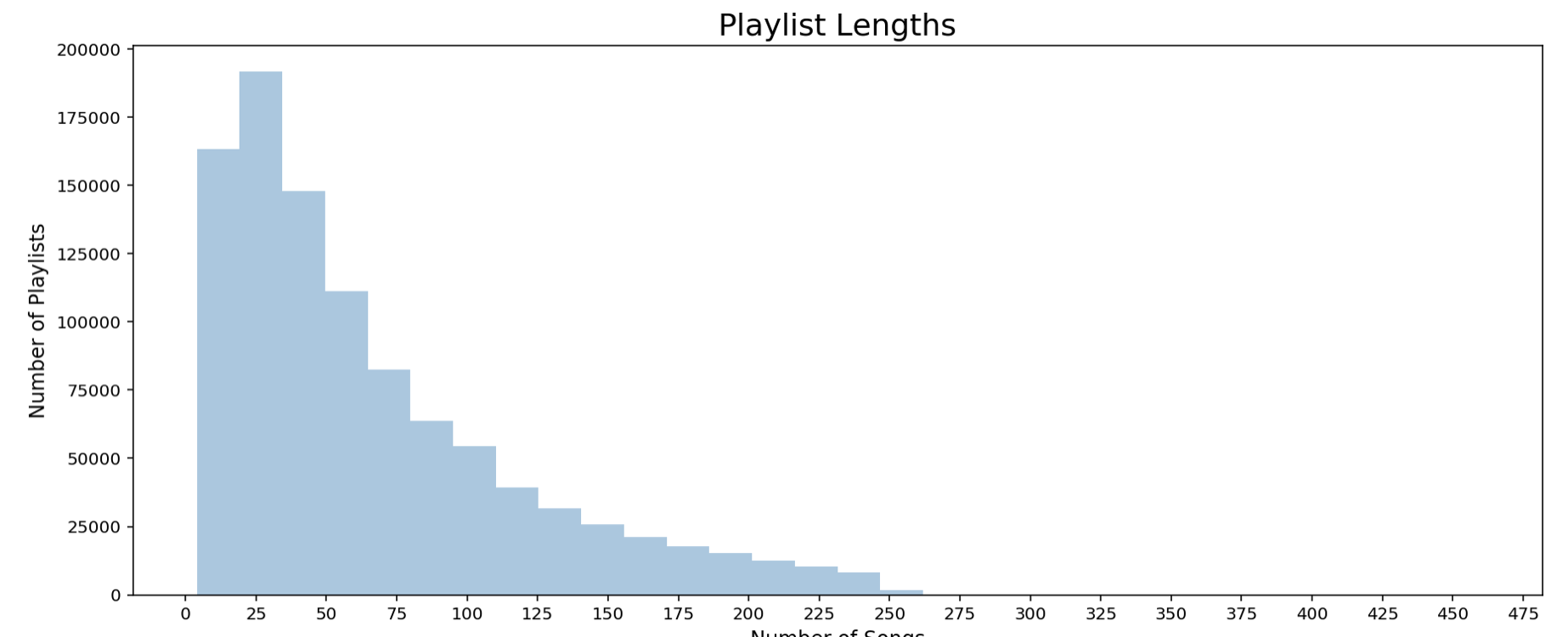
**Playlists**

The Playlists table is extensive with 999k playlists and 66M tracks



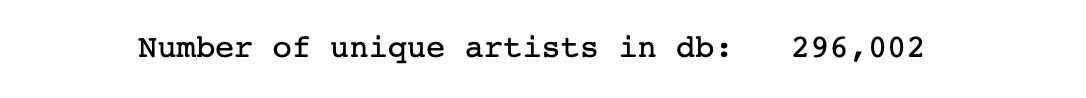
There are some outliers with very long lengths, but the average playlist is 50 songs long and the most common length is 20.



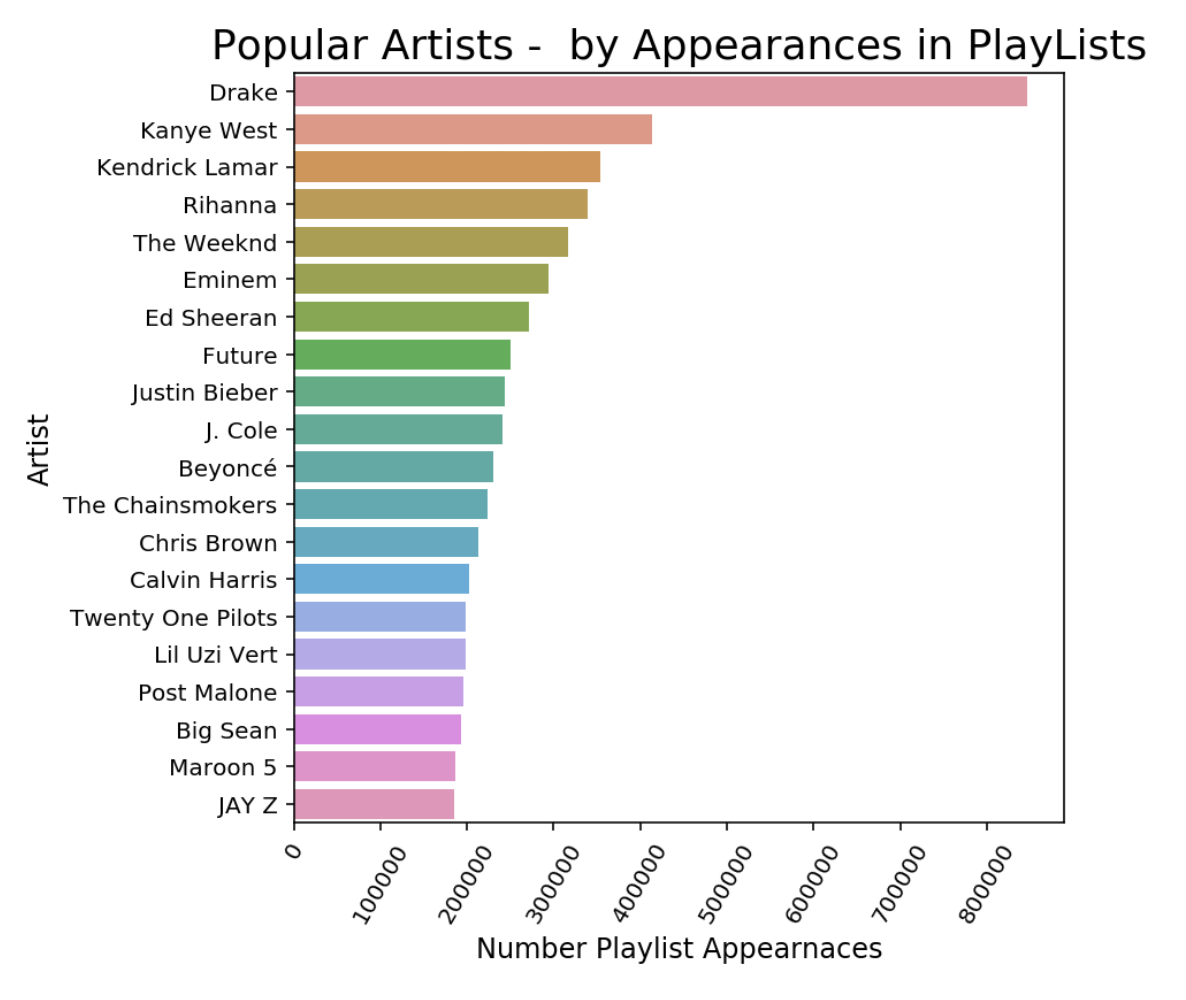


**Artists**

In the playlists, 296k unique artists exist:



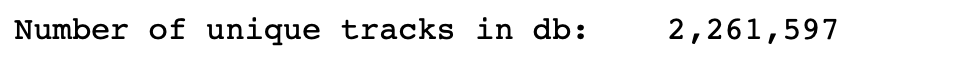
By examining the number of appearances in playlists, we are able to determine the popularity of Artists based on our dataset. Spotify also supplies a field called ‘artist popularity’; however, we found that a majority of Artists in our dataset had a popularity of 0, so we will not rely on the Spotify popularity data.



Spotify supplies genres by artist. After extracting this data, we determined that over 60% of artists had no genre assigned by Spotify, so we will not rely on this data.

**Tracks**

2.2M unique tracks can be found in the playlists.



After querying data from Spotify, various additional useful fields are available for each track. Values are assigned to a significant portion of the population making these features useful for building recommendation lists. Distributions of these features are available in the accompanying notebook.

**Models** [(back to index)](#table_of_contents)

<summary of models and explanations>

**Scores and Comparisons** [(back to index)](#table_of_contents)

<metrics for each model and comparisons>

**Conclusion** [(back to index)](#table_of_contents)

<findings>

**NOTE**:

To convert this to Git markdown:

pandoc -f docx -t gfm Spotify\ Project-ProjectDocument.docx -o README.md