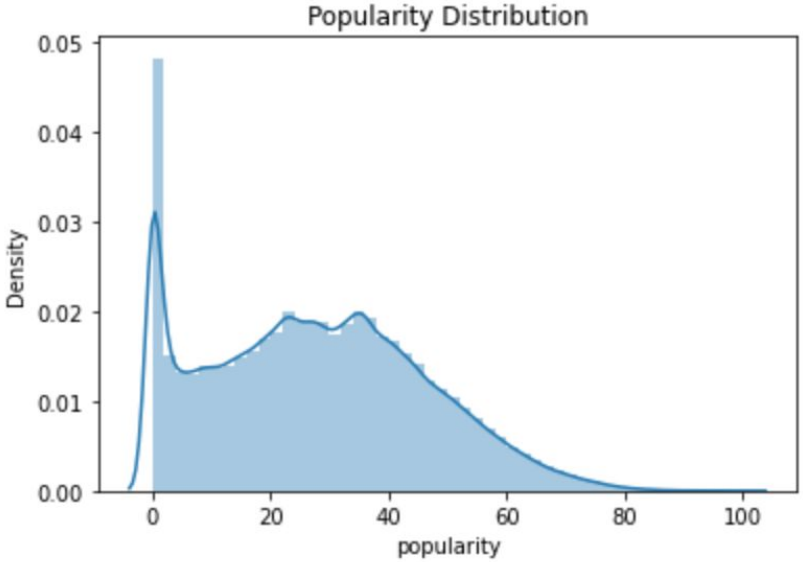
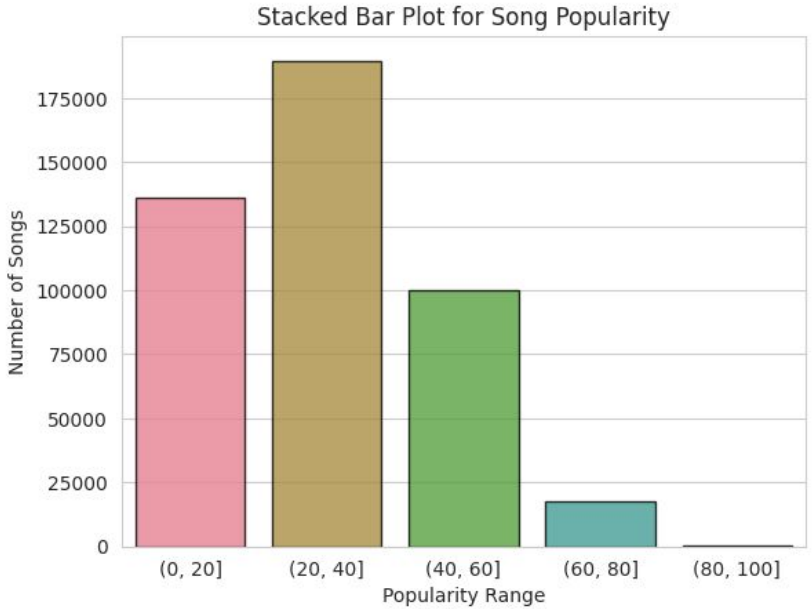
Data Analysis & Visualization

Capucine Philippine Leroux: cpl2146
Chris Berniel Cobashatse: cc4536
Narjes Al-Zahli: na2852
Vedant Gannu: vg2565
Xiangming Huang: xh2550

Data Exploration - Popularity

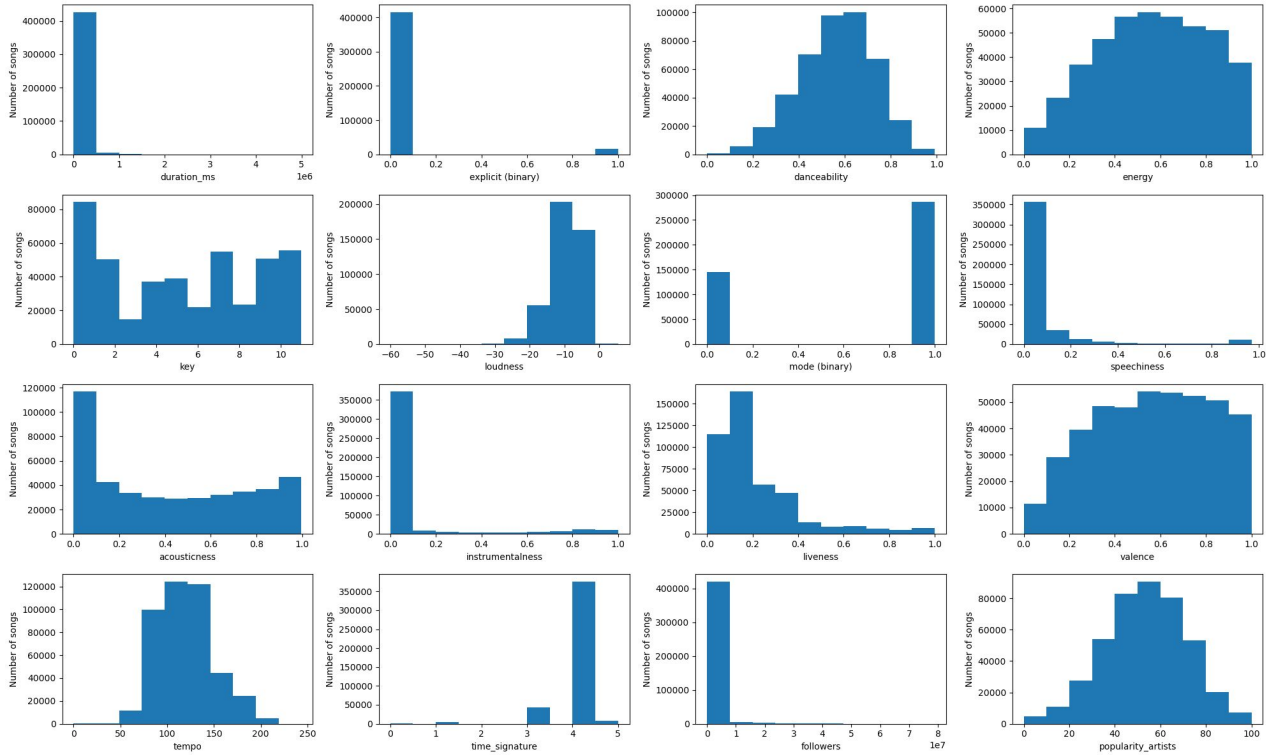




Right-skewed distribution indicating an imbalance dataset:

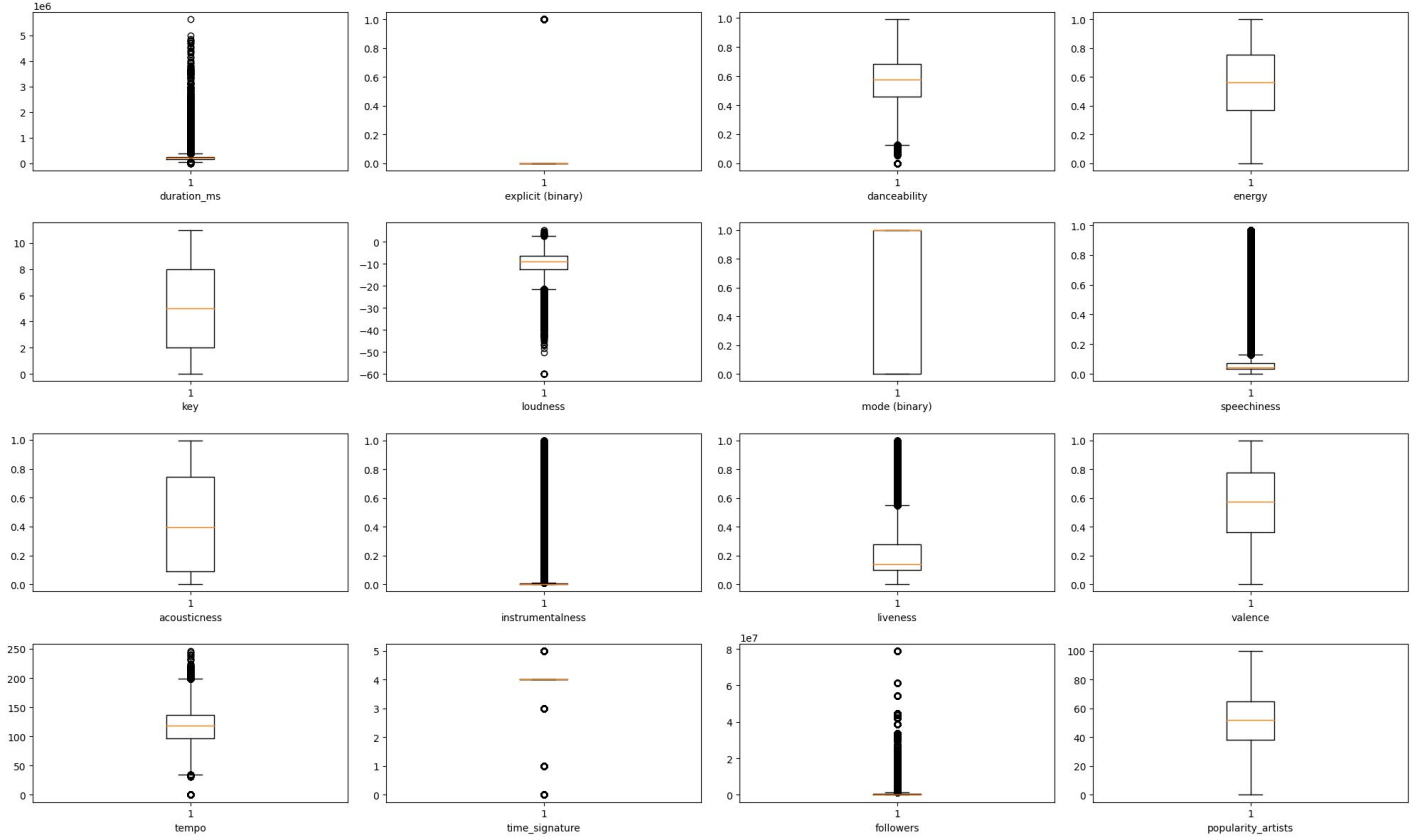
Less than 12.5% of the songs have a popularity of 60 or higher

Number of Tracks vs Features Histograms

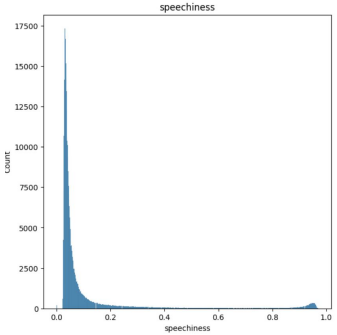


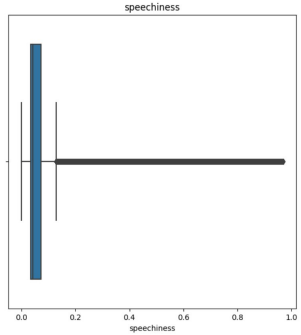
1. Features with a normal distribution:
   a. danceability
   b. energy
   c. valence
   d. tempo
   e. popularity_artists
2. Highly skewed (to the left or right) features:
   a. loudness
   b. speechiness
   c. acousticness (more or less)
   d. instrumentalness
   e. liveness,
   f. followers
   g. time_signature (more or less).
3. Binary features:
   a. explicit
   b. mode
4. Randomly distributed feature
   a. key
5. Feature with very high values and need representation with logarithmic scale
   a. followers

Feature Boxplots

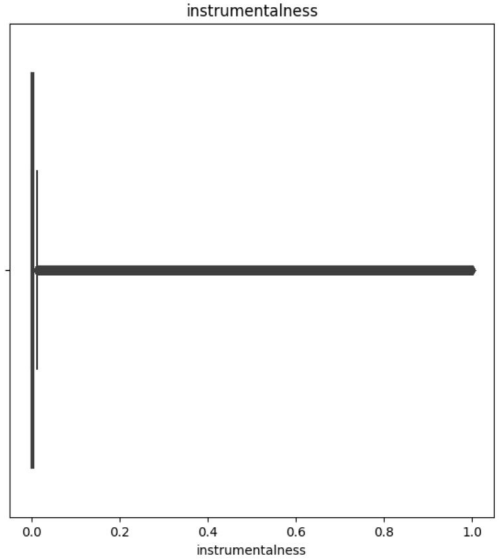


Features with a lot of outliers: **duration_ms, danceability, loudness, speechiness, instrumentalness, liveness, tempo, followers**

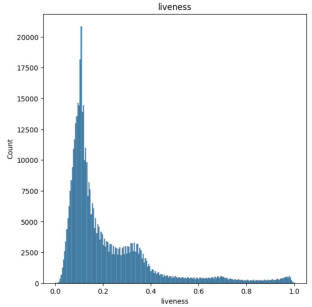


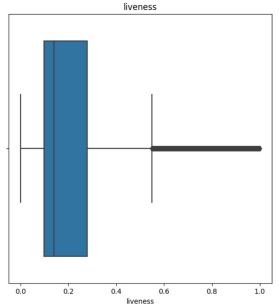


Without Log transform: 64,769 outliers

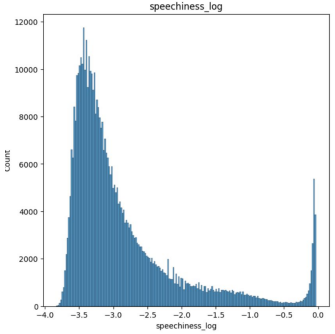


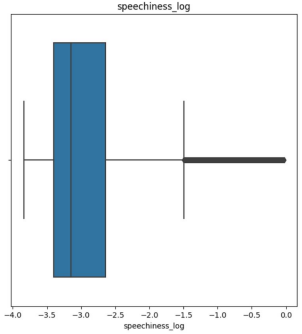
101,821 outliers in the **instrumentness** column, but this column might be worth keeping since purely vocal songs also might also be associated with high popularity. Also this is ¼ of the dataset rows



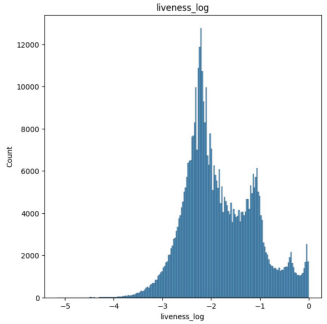


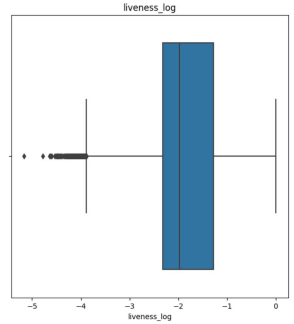
Without Log transform:
32625 outliers





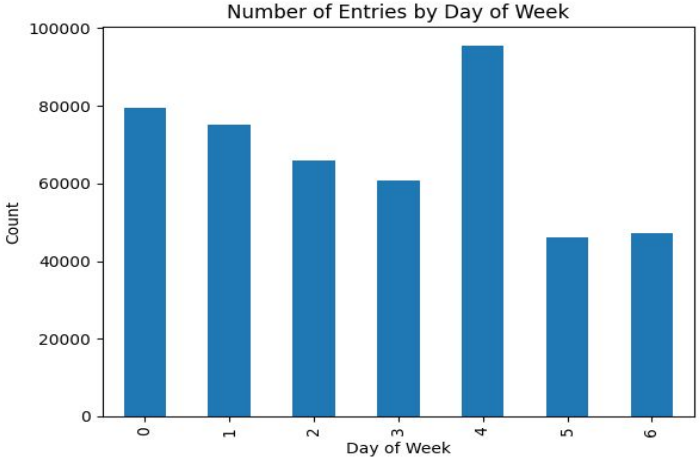
With Log transform: 40850 outliers

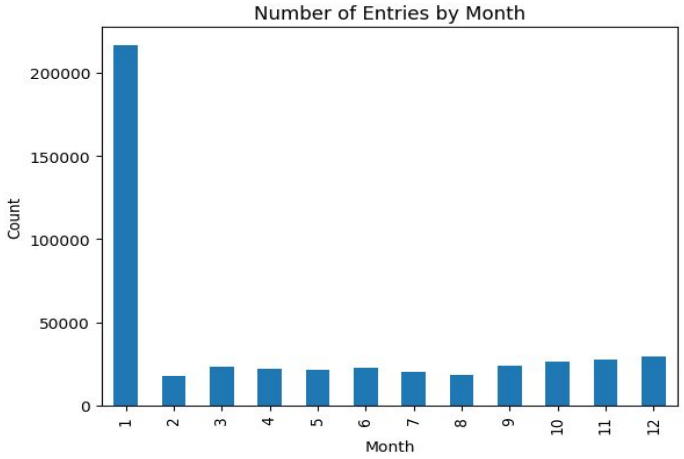


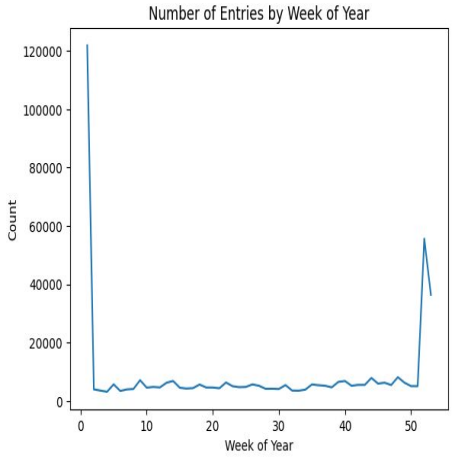


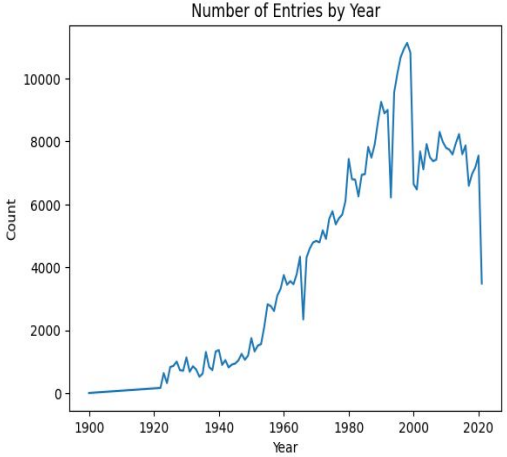
With Log transform: 296 outliers

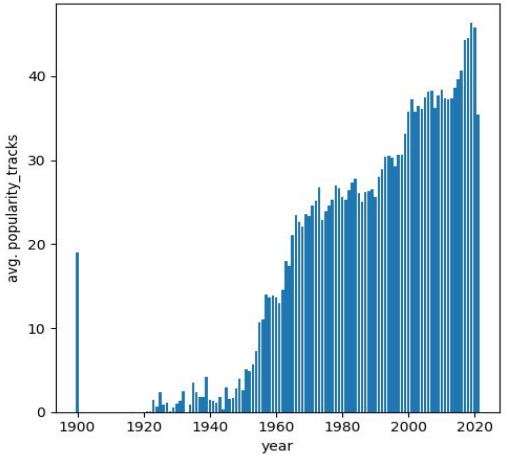
Tracks & Time







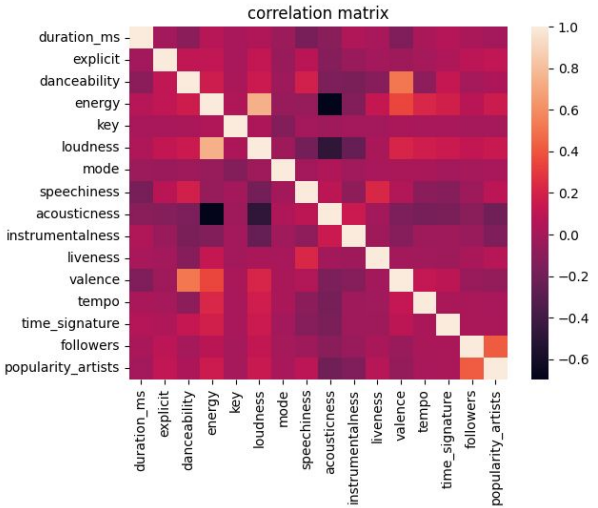


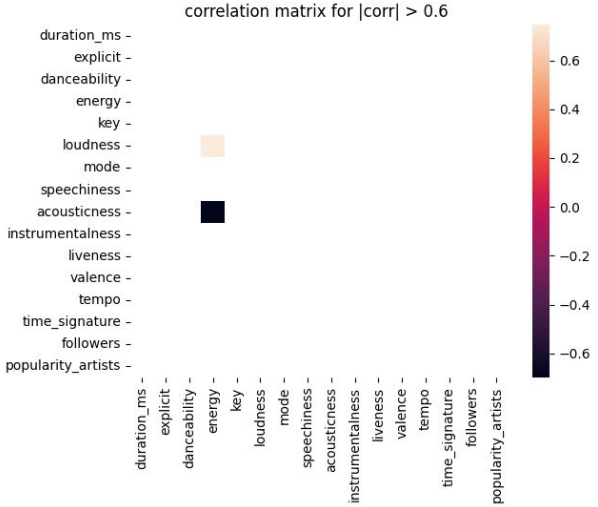


Conclusions

1. More tracks in the dataset were released in January than any other month
2. Tracks were more often released at the beginning of the year or at the end.
3. The distribution of tracks within the week is fairly balanced.
4. Many tracks in the dataset were released in the 1990s, 2000s, and 2010s
5. The average popularity score has increased year over year, probably due to the increase in access to mainstream media consumption tools such as radios, TVs, internet, mobile phones, Spotify, etc.

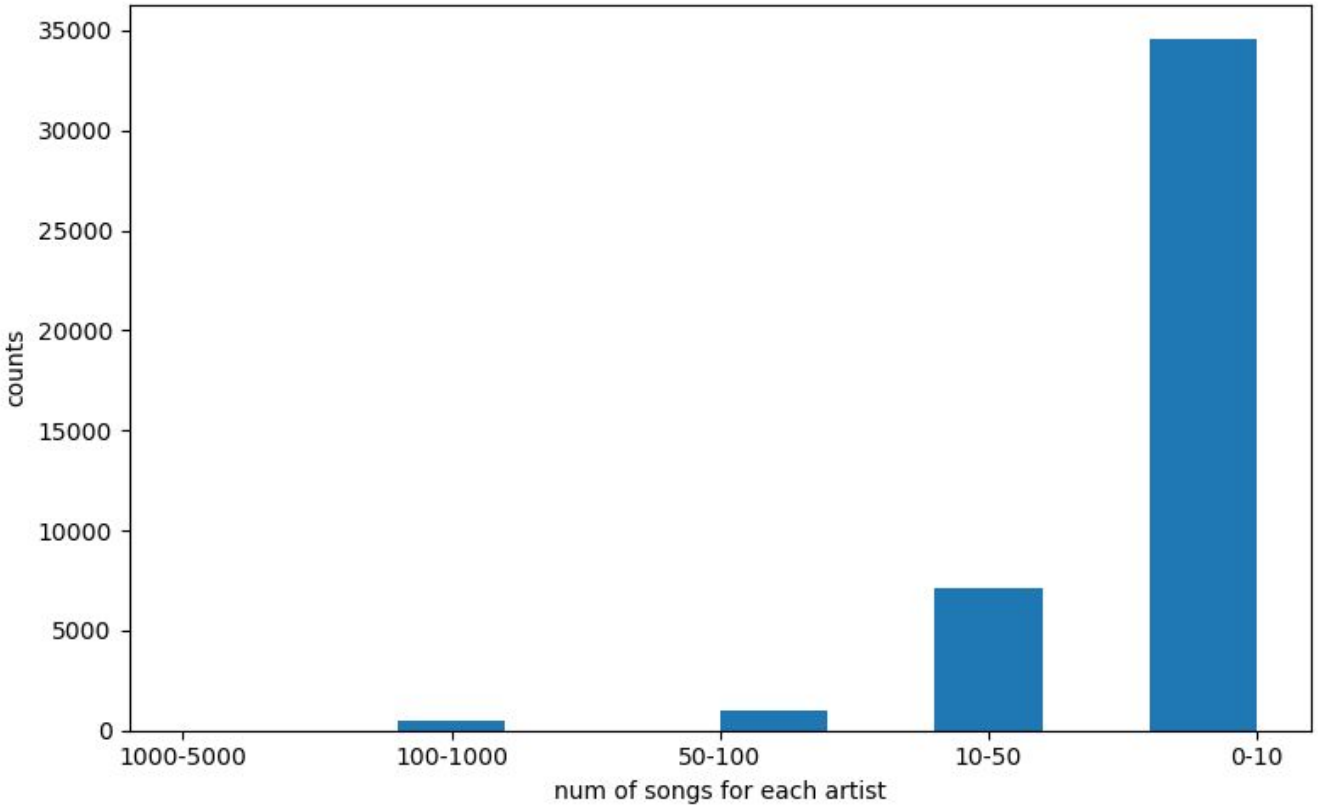
Correlation Matrix Between Numerical Features





With a threshold of 0.6, **loudness and energy** and **acousticness and energy** are highly correlated. Might be useful to eliminate **energy** from the numerical feature list

Artists Frequency



This plot shows the frequency of each artist show up in the dataset.

Most artist have 0 to 10 tracks in the dataset. Only a small number of artist have tracks larger than 50.

We further separate all artists with total tracks larger than 100, and plot all the feature histograms same as slide 3. It shows the distribution are very similar with overall dataset.

Cleaning and Sampling

- In tracks.csv, the column artists is reformatted. “[“ and “]” are stripped away. Same preprocesses are applied on columns genres, id_artists.
- Two datasets are merged on the id of artist. After the merge, the dataset contains 470k rows and 55k unique artists. On average, each artist has 8.5 tracks.
- About 37k rows in the merged dataset have empty genre, we decided to drop those rows. After dropping those rows, on average, each artist has 10 tracks.
- The feature instrumentalness measures whether the track contains no vocals. Confidence is higher as the value approaches to 1.0. Values above 0.5 are interpreted as instrumental tracks. There are 100k outliers, consider this feature might correlate to popularity, we decide to keep the outlier.
- Liveness detects the presence of audience. If the value is greater than 0.8, very likely the track is live. We decided to remove all the row have liveness greater than 0.8. There are 30k outliers. If a song is performed live by an artist, very likely there is a identical song that is not from live. Therefore, we decide to delete all the outliers.
- For speechiness, below 0.33 means the track only have music, between 0.33 and 0.66 means the track may contain both music and speech. Above 0.66 means the track is made entirely of spoken words. There are about 40k rows have speechiness above 0.66. Very likely those tracks are not songs. Since this project focus on identifying popular songs, we decided to remove all outlier rows.

ML techniques

1. One approach is to define popularity score bins, then we have a smaller classification problem. Potential classification methods include:
    - Random forest, XGBoost for classification
    - Logistic regression
    - Neural network
    - Support vector machine
2. An unsupervised approach: first cluster the dataset, then interpret each cluster based on given genres. For example, suppose we found a cluster that contains a lot rock songs with high popularity, then we can define the cluster as popular rock music. Potential unsupervised methods include:
    - PCA
    - t-SNE
    - K-means clustering