Michael Adu-Gyamfi

EXPLORATORY DATA ANALYSIS

This is a dataset of Spotify tracks over a range of 125 different genres. Each track has some attributes related to songs associated with it.

Column Description

track_id: The Spotify ID for the track

artists: The artists' names who performed the track. If there is more than one artist, they are separated by a semi-colon.

album_name: The album name in which the track appears.

track_name: Name of the track.

popularity: The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity.

duration_ms: The track length in milliseconds.

explicit: Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown).

danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.

key: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.

loudness: The overall loudness of a track in decibels (dB).

mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66

describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.

liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

time_signature: An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of 3/4, to 7/4.

track_genre: The genre in which the track belongs.

INFORMATION ABOUT DATA

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 114000 entries, 0 to 113999
Data columns (total 21 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        114000 non-null  int64
 1   track_id          114000 non-null  object
 2   artists           113999 non-null  object
 3   album_name        113999 non-null  object
 4   track_name        113999 non-null  object
 5   popularity        114000 non-null  int64
 6   duration_ms       114000 non-null  int64
 7   explicit          114000 non-null  bool
 8   danceability      114000 non-null  float64
 9   energy            114000 non-null  float64
 10  key               114000 non-null  int64
 11  loudness          114000 non-null  float64
 12  mode              114000 non-null  int64
 13  speechiness       114000 non-null  float64
 14  acousticness      114000 non-null  float64
 15  instrumentalness  114000 non-null  float64
 16  liveness          114000 non-null  float64
```

```
 17  valence           114000 non-null  float64
 18  tempo             114000 non-null  float64
 19  time_signature    114000 non-null  int64
 20  track_genre       114000 non-null  object
dtypes: bool(1), float64(9), int64(6), object(5)
memory usage: 17.5+ MB
```

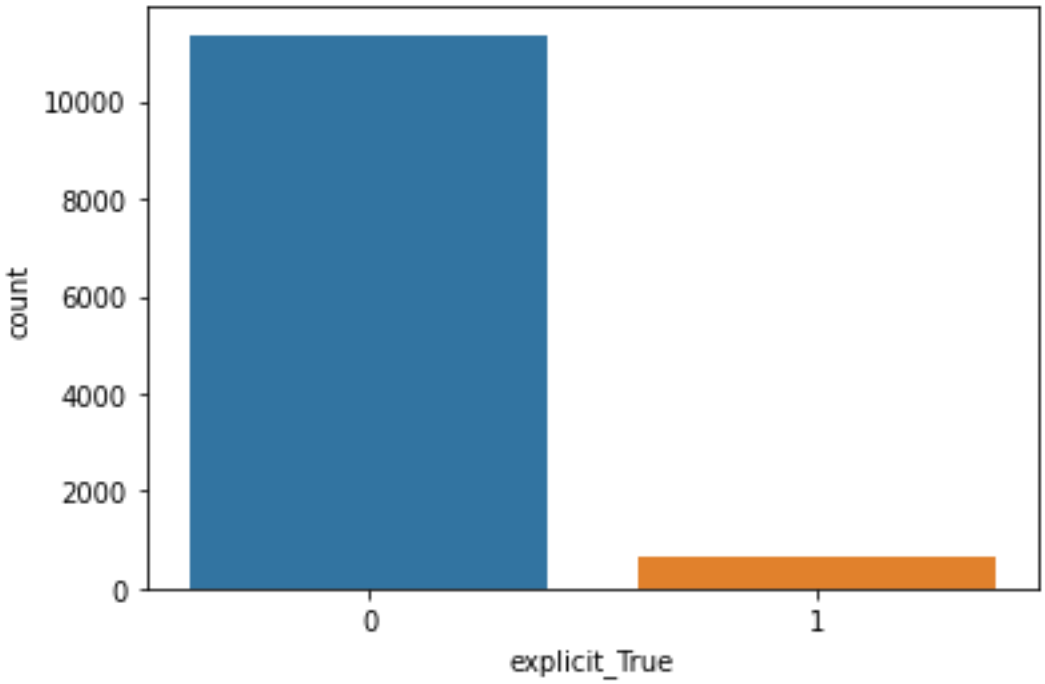
INITIAL PLAN FOR DATA EXPLORATION.

1. Data Cleaning
    - Reading and understanding the data
    - Handle duplicates if any
    - Handling missing values if any
    - Handling Outliers and removing them
2. Visualize some elements in the data.
    - Before log transformation and after log transformation
    - Top 20 albums
    - Histogram of song tempo
    - Bar chart of explicit and non-explicit songs
3. Query and answer the following questions.
    - Who is the most popular artists.
    - What is the most popular song.
    - What is the most popular genre.
    - Does the explicit nature of a song affects it popularity.
    - How does the tempo of a song affect its popularity.
4. Hypothesis testing

DATA CLEANING AND FEATURE ENGINEERING

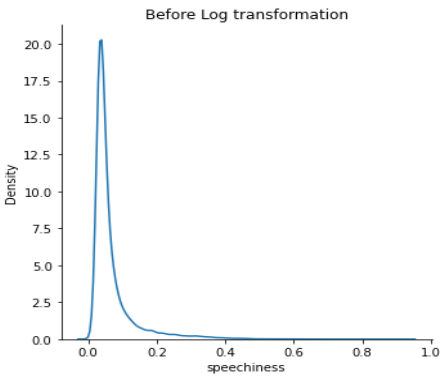
- The dataset has 11400 rows and 21 columns.
- The dataset has no duplicates.
- One row in the data had NaN values which were removed.
- Using z-score analysis it was realized that `duration_ms` and `speechiness` had outliers.Hence a log transformation was performed on them. Additionally, log transformation was performed on all attributes with skew variables greater than 0.75 .
- For `explicit` column, one-hot encoding was performed to change the True and False values to zero to represent False and one to represent True.
- The `duration_ms` columns was converted to minutes.

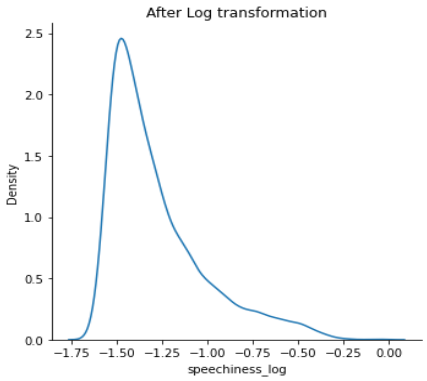
DATA VISUALIZATION



- From the bar chart above, it could be seen that songs with explicit content was lesser than songs without explicit content.

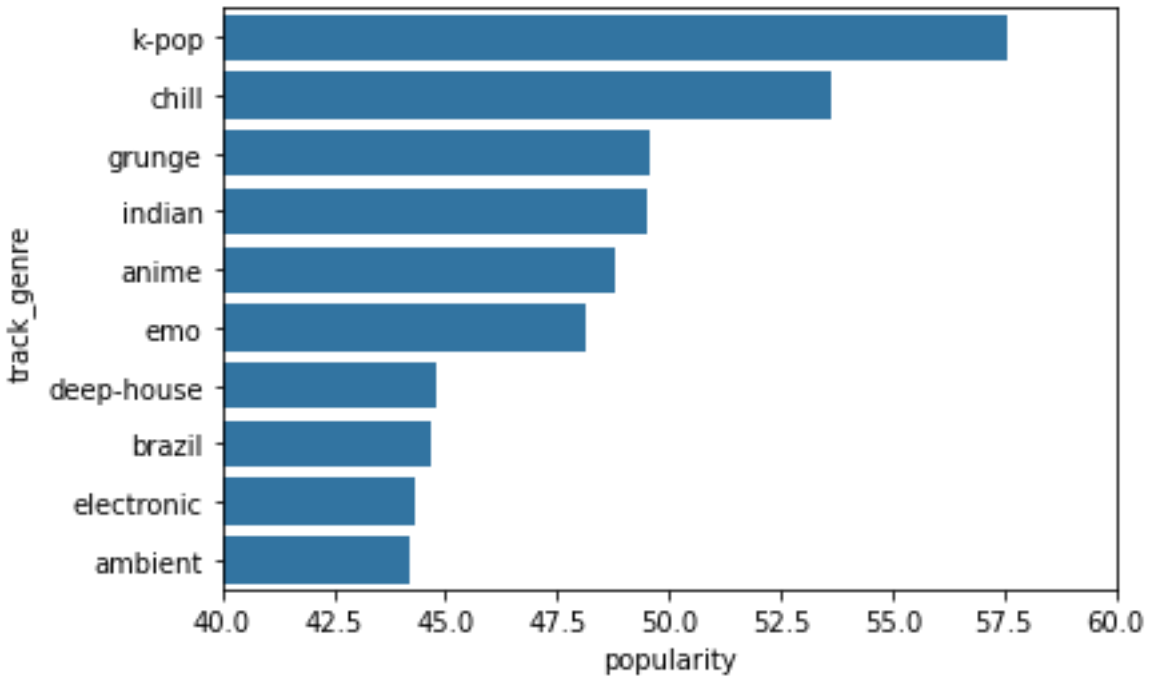
LOG TRANSFORMATION OF SPEECHINESS



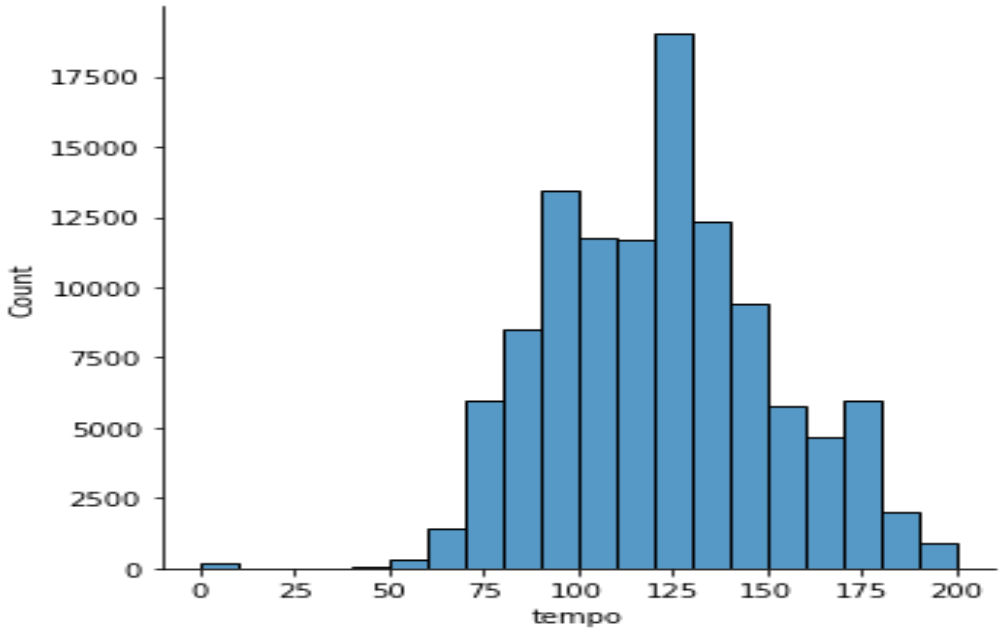


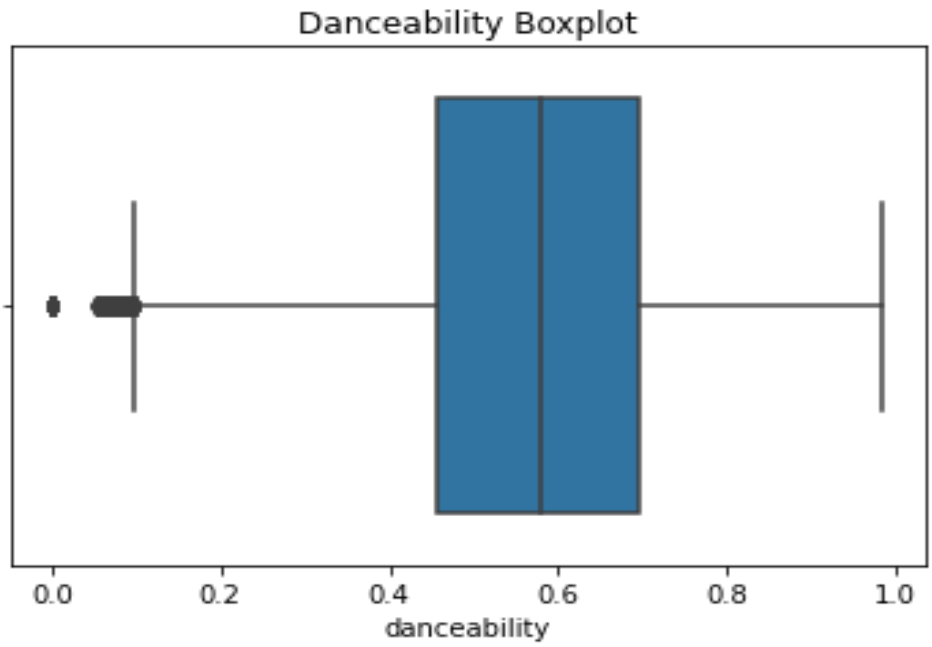
- `Speechiness` was one of the variables that were skewed. After log transformation, It can be seen that the data is now centered.

Top 10 genres



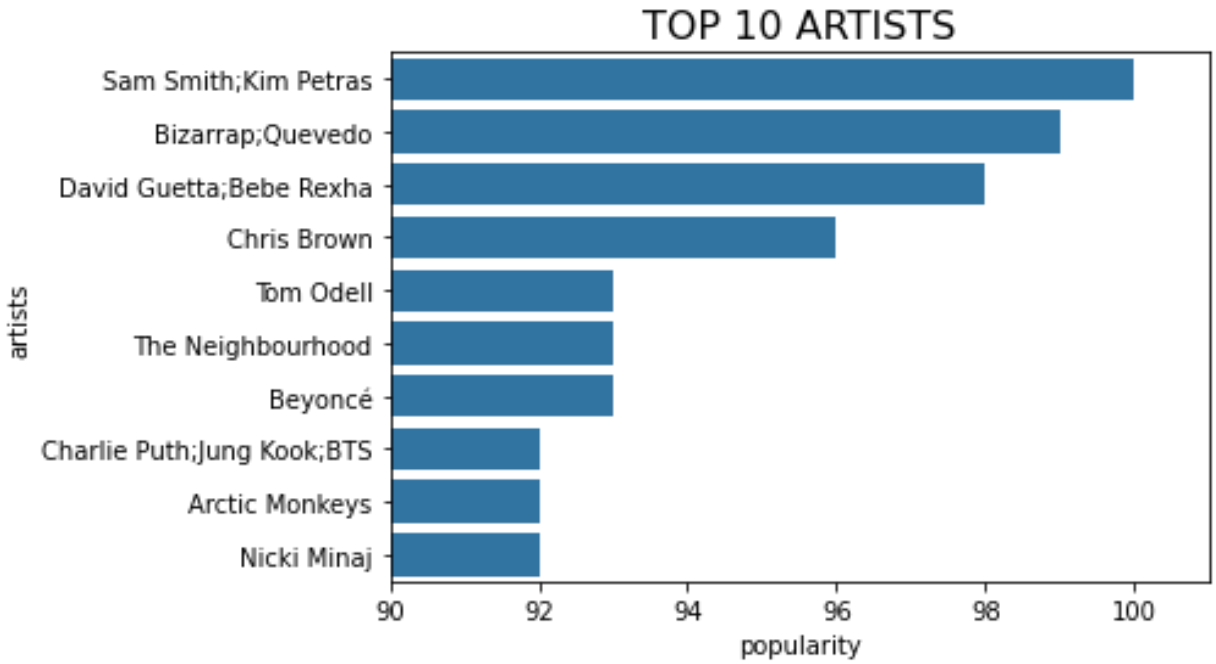
DISTRIBUTION FOR SONG TEMPO



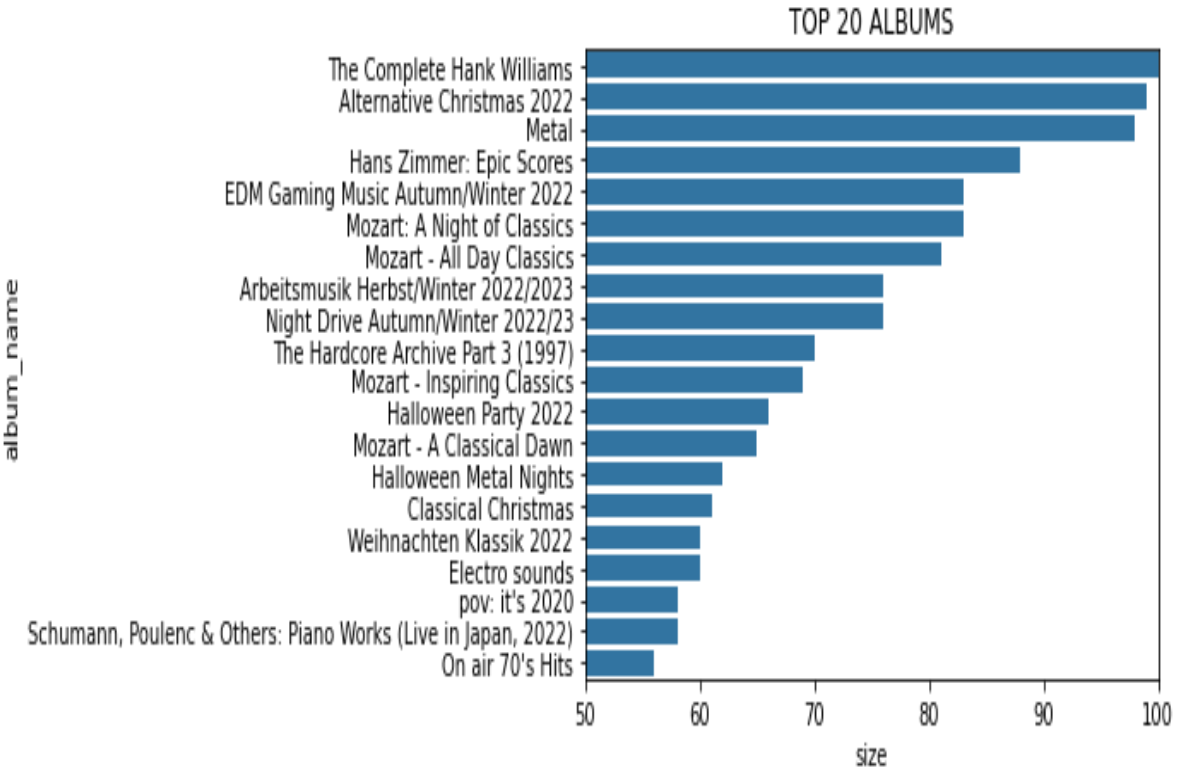


TOP 5 TRACKS

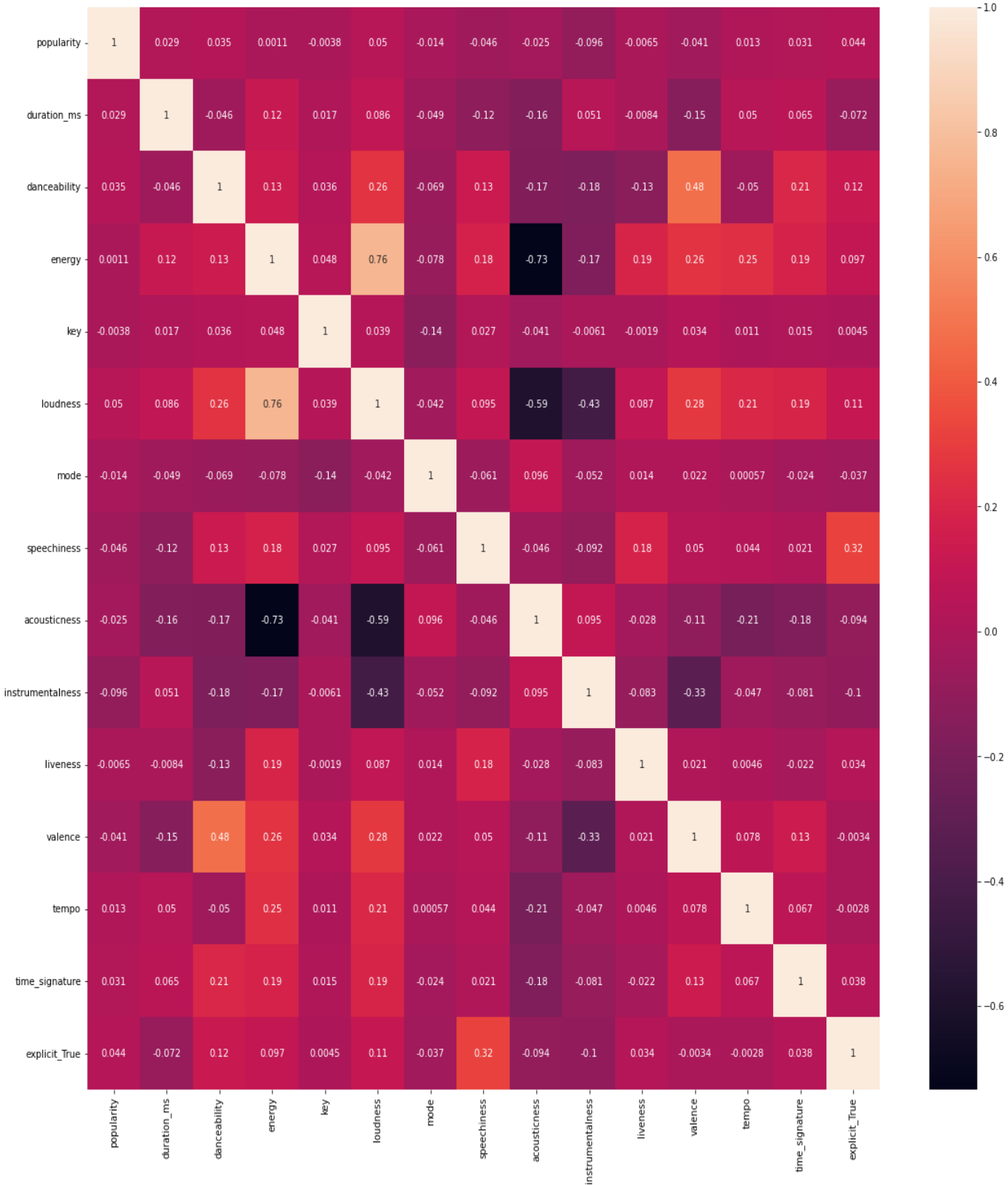
	track_name	artists	popularity
20001	Unholy (feat. Kim Petras)	Sam Smith;Kim Petras	100
81051	Unholy (feat. Kim Petras)	Sam Smith;Kim Petras	100
51664	Quevedo: Bzrp Music Sessions, Vol. 52	Bizarrap;Quevedo	99
89411	La Bachata	Manuel Turizo	98
81210	I'm Good (Blue)	David Guetta;Bebe Rexha	98



TOP 20 ALBUMS



CORRELATION MATRIX



INSIGHT FROM CORRELATION MATRIX

- Song with higher acousticness had lots of energy which is expected.
- Songs perceived as being positive or upbeat(valence) had a high correlation with the danceability of the song.
- `loudness` had a very strong correlation with the energy of a song.
- The explicitness of a song does not affect its popularity.

Hypothesis Testing

Hypothesis

1. Most modern songs have a duration of more than 3000 milliseconds.
2. The probability of selecting an explicit song is 0.5. That means there are equal chances of selecting or not selecting an explicit song.
3. The longer the duration of a song , the higher its popularity.

CONDUCTING FORMAL SIGNIFICANCE TEST FOR HYPOTHESIS 1.

This is to investigate whether the duration of a song is random or if they are intentionally made to have a duration of more than 3000 milliseconds.

Null Hypothesis: The duration of most songs are less than 180000 milliseconds (3 minutes). The probability of selecting a song that is more than 180000 milliseconds is 0.5.

Alternative Hypothesis: The probability of selecting a song that is less than 180000 milliseconds (3 minutes) is greater than 0.5. This means that most songs have duration of above 3000 milliseconds.

After 100 observations were sampled, 71 had a duration greater than 180000(3 minutes). The test statistic was therefore 0.71. Binomial distribution was used for the hypothesis testing.

The p-value was found to be 0.0063%. There is way less than 5%. Hence we can reject our null hypothesis and accept the alternative hypothesis.

Inference

The alternative hypothesis is accepted. The probability of selecting a song whose duration is longer than 3 minutes is high. It can be concluded that most of the songs being realized are more than 180000 milliseconds(3 minutes) longer.

SUGGESTIONS FOR THE NEXT STEPS IN ANALYZING THE DATA.

- Boxplots can be used to visualize the statistical measures of columns of interest.
- Using Machine Learning to identify patterns in data.
- Adopting Machine Learning methods to predict the popularity of some songs based on their attributes.

QUALITY OF DATA

The data is ideal for data analysis. As a result, no tiring data cleaning and feature engineering methods were required.

In addition, the number of observations was satisfactory. 11400 observations were provided. This is good for machine learning algorithms although a lot more observations would lead to improved accuracy.