

**TASK**

**Exploratory Data Analysis on the Spotify Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

This Exploratory Data Analysis will be done on a dataset which was retrieved from Spotify. The Spotify.csv dataset is constructed of 17 columns and 2017 rows. The columns are categorical variables, each category represents an attribute of a song. Some columns contain continuous and discrete data while other columns contain string and object types.

Some columns are:

* Artist
* Song\_title
* Loudness
* Tempo

Each of the 2017 rows lists a song as well as its attributes. In general the dataset is in very good condition so would be fairly simple to produce some insightful visualizations.

**DATA CLEANING**

* The column headers do not contain any hyphenated word but they do contain words with underscores, these will be perfect for constructing visualizations as there will not be ant errors.
* All datatypes were logged correctly so there was no need to convert any of the datatypes.

MISSING DATA

* This data was taken from the Spotify database so there were no Nan or Null values.
* This data was pre-processed and structured to construct a machine learning algorithm to predict if a song would be liked by a user.
* In other words, the algorithm is used to generate daily playlists on Spotify for the user.

DATA STORIES AND VISUALIZATIONS

1. Song Tempo Like/ Dislike Distribution:

As we can see the graph is more or less normally distributed, we can see that listeners prefer songs with a moderate tempo.

1. Song Danceability Like Distributions:

This graph is skewed slightly to the left we can see that users prefer songs with a higher danceability score.

1. Song Duration Like Distribution:

These values are skewed right, and we can infer that listeners do not prefer songs with a lower duration but instead prefer songs with a larger duration.

1. Song Loudness Like Distribution:

The data for this is skewed right. The graph indicates that users do prefer loud music but not at extreme volumes.

1. Song Speechiness Like Distribution:

The data is greatly skewed to the right. We can see that user do prefer more speech in the songs that they listen to.

1. Song Valence Like Distribution:

The is data is roughly normally distributed, users prefer songs with a higher valence.

1. Song Energy Like Distribution:

This data is skewed to the left. We can infer that users enjoy songs at roughly 0.8 units of energy.

1. Song Accousticness Like Distribution:

This data is skewed to the right and we can clearly deduce that users prefer songs with lower accousticness.

1. Song Instrumentalness Like Distribution:

This data is highly skewed to the right and we can infer from the graph that users dislike songs that contain low instrumentals.

**THIS REPORT WAS WRITTEN BY : Kamerin Moodley**

