Do Popular Songs Endure?

Introduction:

Songs are an integral part of human life. There have been songs from centuries. Predicting whether a song catches the heart and soul of the audience, impacts the music industry. There are a lot of things that come into consideration that make a particular song popular on the music charts and amongst the music lovers across the globe. In the last century, the music industry has seen a huge transition from CDs to MP3 players and these days available on different streaming applications. But even with this technological advancements, the music industry is driven by users liking and disliking for a particular song.

Objective:

The main objective of this project is to develop a prediction model that applies the concepts of data science and predicts the popularity of a recording 'w' weeks from now, given that the current popularity is 'p' by analyzing the dataset of songs and artists starting from the last six to seven decades by taking into consideration a number of parameters to predict the current popularity of a particular song recording.

Data:

We obtained data from two primary sources, details of which are as follows:

1. **Billboard Data Top 100 Per Week**

Billboard publishes a list of Top 100 songs each week throughout the year. So we have dataset available for more than 60 years using the Billboard. We then used the Billboard API from the source(Link: https://github.com/guoguo12/billboard-charts) and we were able to get the list of songs for more than 60 years along with their artist name and current position in the billboard. Along with that, we got a number of other attributes as well. The dataset has the following attributes:

- title – The title of the track.
- artist – The name of the artist, as formatted on Billboard.com.
- peakPos – The track's peak position on the chart at any point in time, including future dates, as an int (or None if the chart does not include this information).
- lastPos – The track's position on the previous week's chart, as an int (or None if the chart does not include this information). This value is 0 if the track was not on the previous week's chart.
- weeks – The number of weeks the track has been or was on the chart, including future dates (up until the present time).
- rank – The track's current position on the chart.
- isNew – Whether the track is new to the chart.

2. **Spotipy Songs Popularity Data**

Spotipy is a lightweight Python library for the Spotify Web API. Using Spotipy Web API we are getting access to all of the music data provided by the Spotify platform. We then used the Spotipy API from the source(Link: https://github.com/plamere/spotipy) and got the popularity of the songs we got from the billboard. Then, we got the popularity of the songs from billboard by using the artist and track name and hitting them on the Spotipy API. Then we are using the popularity data to compare the actual result and predicted the result and we were able to depict using the Power Law function.

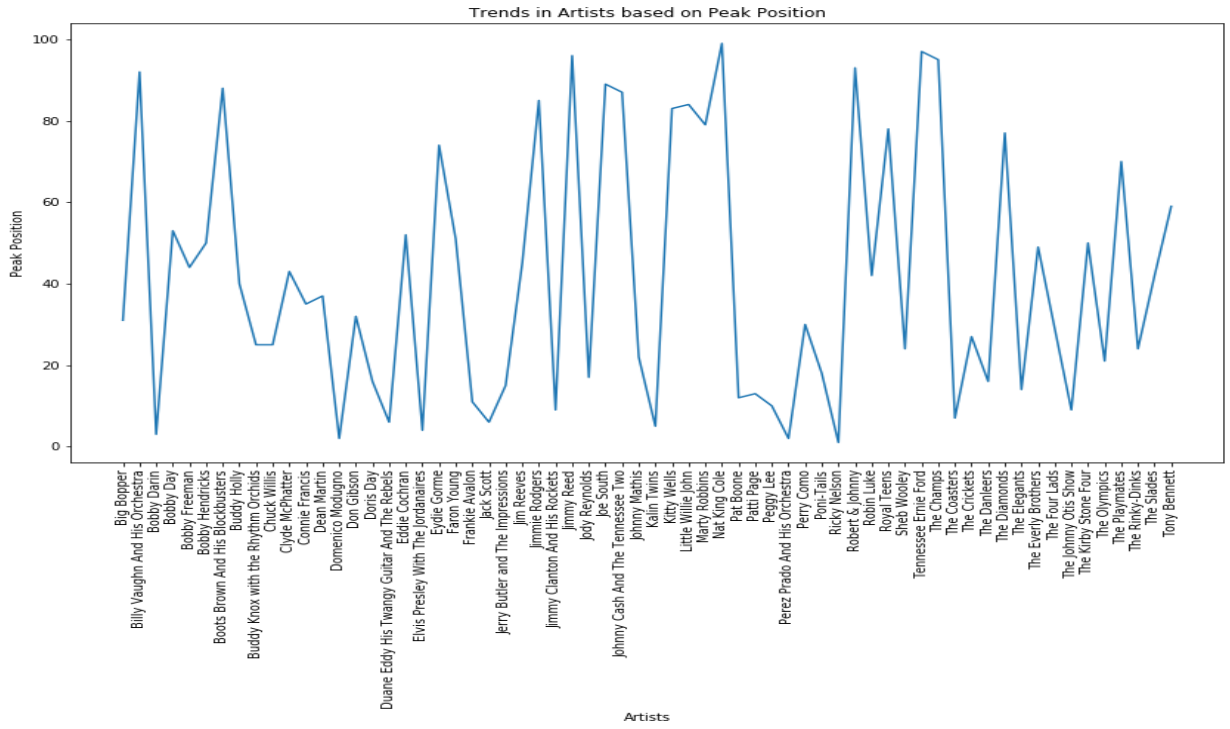
Data Preprocessing:

1. **Basic Cleaning:**

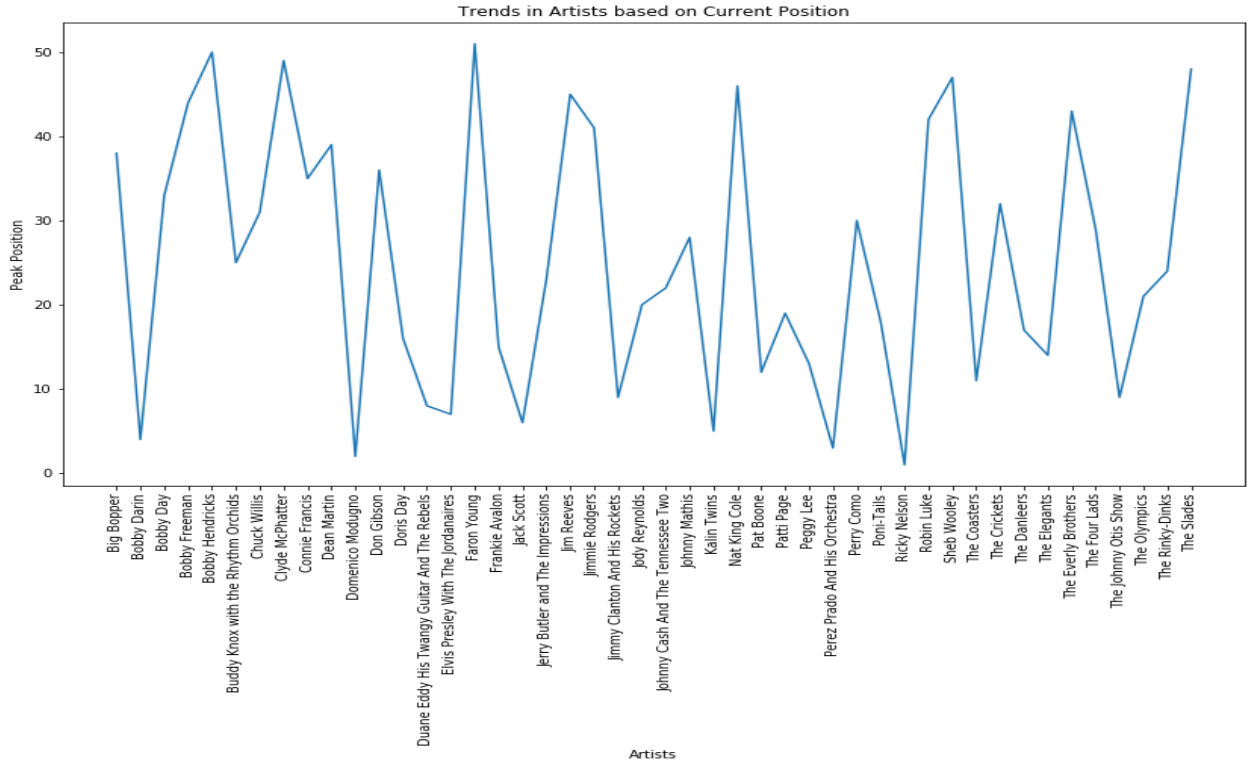
The dataset had various uncleaned features such as the weeks and current position fields in the billboard data had zero value. The reason the dataset was having zero values was that there were inconsistencies while calling the Billboard API for a few weeks in the billboard data. All such data items have been removed from the dataset. The dates in the dataset were in string format, so we parsed the dates and converted it to the standard date format. From the popularity data, we discarded any values which were zero, because there were a few songs whose popularity was returning a zero value because of the decreasing popularity of the song.

2. **Title-wise data aggregation:**

We initially had all songs in the dataset in the form of the week for a year and the billboard rank for that week from 1 to 100 along with the previous position of the week and the maximum rank achieved by the track. The idea was to get tracks which were recorded before the year 1961 and get the popularity of that particular track and then using this popularity we were predicting the popularity of that track after x number of years. In order to do this, we aggregated our data on the basis of title and selected a record with a year less than 1961. We did this for a few tracks and we generated a couple of interesting plots based on our observed data. Now for generating plots, we have taken about a hundred artists name and then using those artist names and the available data were about to get a few interesting observations from the given data



Here, we are finding trends in the artist names based on peak position. The plot denotes the maximum position attained by the artist during his lifetime. In this, we are grouping the data based on the artist and getting the data for artist rank on the value when he attained the maximum.



Here, we are getting the value for Artists in the current position and found out that the maximum value for the current position at the given time in the dataset for that artist.

Developing the Baseline Model (Monkey Model)

After we have gathered the data from billboard into the CSV file, we grouped the data by different titles. Later we added a new column for the year so that we can later extract the data based on year. After this we have grouped the data by titles, we have extracted the first appearance and last appearance of the song in the Billboard chart and made into another data frame. Then using the appearances computed the total number of years for which the song has actually endured. Then the data is sorted and we have picked the songs that are endured for a long time. We have then made the list of pairs consisting of a number of times the song has appeared in the billboard dataset and the name of the song. After preparing the list we have sorted the list based on the number of times it has appeared in the dataset. Now combining the above two results we have picked the songs in such a way that the songs are endured for the longest time and have appeared most of the time in the dataset to make the power law curve for the baseline model as accurate as possible.

From the list of songs that we filtered by the previous method explained, we grouped the data of each song by year and picked the minimum of the billboard's popularity. In this way we have collected a few sets of samples and have fitted the power law curve over the values and came up with the following power law formulas for each of the different songs:

1. Recording Name:- "**Call Me"**
   Formula **:**

$$Popularity\ of\ a\ recording = \frac{3 * Intial\_Popularity}{(Number\ of\ years)^{0.5}}$$

2. Recording Name:- "**Missing You"**
   Formula:-

$$Popularity\ of\ a\ recording = \frac{1.75 * Intial\_Popularity}{(Number\ of\ years)^{0.8}}$$

3. Recording Name:- "**Stay"**
   Formula **:**

$$Popularity\ of\ a\ recording = \frac{1.81 * Intial\_Popularity}{(Number\ of\ years)^{0.85}}$$

4. Recording Name:- "**Forever"**
   Formula **:**

$$Popularity\ of\ a\ recording = \frac{1.56 * Intial\_Popularity}{(Number\ of\ years)^{0.61}}$$

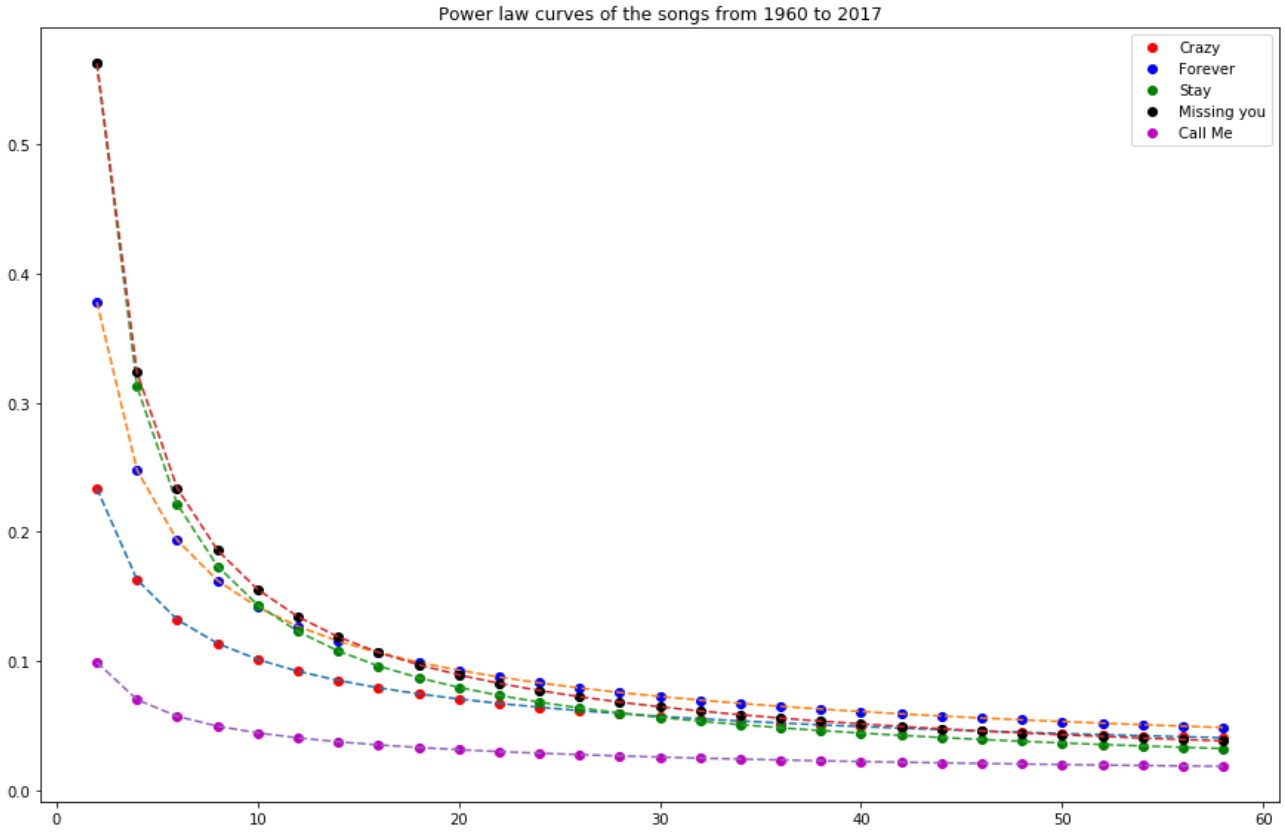
5. Recording Name:- **"Crazy"**
   Formula **:**

$$Popularity\ of\ a\ recording = \frac{1.5 * Intial\_Popularity}{(Number\ of\ years)^{0.52}}$$

In all the above we have assumed that popularity = 1/given_popularity, as we need the popularity to decrease after certain years.

We were able to successfully identify 5 recordings in 1960 that have endurance in 2018.

We generated the relevant plots for the songs and their comparison with the actual value

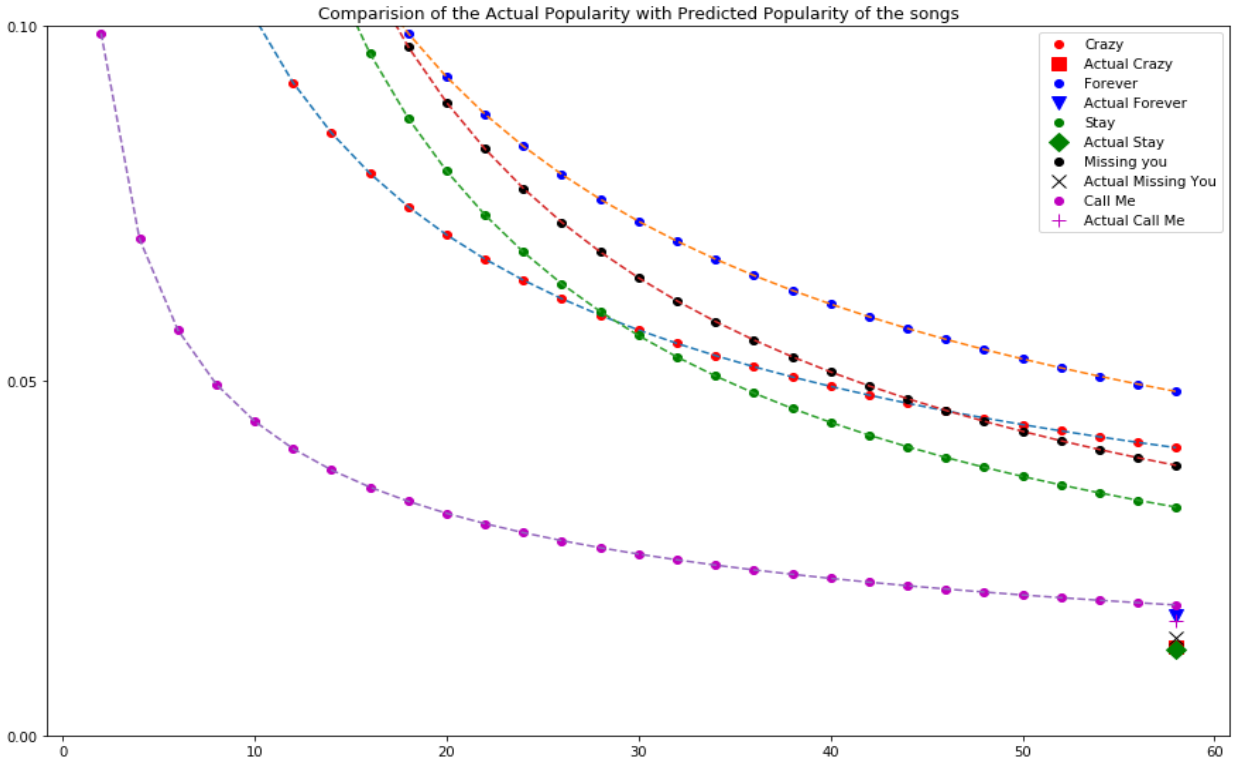


We have chosen only the above songs because we observed that some of the songs either don't follow the power law curve as their popularity tends to decrease and increase with increasing number of years or they don't have their appearance in the billboard charts for a significant period of time.

Comparing the Baseline Model Predictions with the Actual Values :

We have extracted the actual current popularity data of a recording from the Spotify API and the predicted popularity has been calculated based on the different Power Law functions for different songs. The difference between the actual values and the predicted values can be found in the plot below as well as a comparison between the values has been provided in the table below.

Name of the Recording	Actual Popularity	Predicted Popularity
Crazy	0.0125	0.041
Forever	0.017	0.0484
Stay	0.0122	0.032
Missing You	0.0137	0.038
Call Me	0.0161	0.0183



The difference between the predicted value and the actual value might be because of the decrease in the popularity of the artist or the other songs of the artist might have gained more popularity in the due course of time or there could be more other factors that can affect a popularity of a particular song and these factors will be analyzed in the future!

Future Plans:

1. We can consider the number of views of each track from Youtube and other song data sources to get a better prediction.
2. While our current model doesn't consider the aspect of selecting all the tracks as we are only interested in some of the tracks which are before 1970 to show a trend and we are predicting how that deviates from the actual predicted value.
3. We can plan to come up with an alternative error metric which considers the distance of the predicted probabilities to the actual value.
4. We will be applying Machine Learning models like SVM, LGBM, Deep Learning, Random Forest Classifier to analyze why there is a deviation from the actual value
5. We will be analyzing why the predicted popularity has deviated from current popularity.