# Project Progress Report: Do Popular Songs Endure?

## Objective

The objective of this project is to analyze the performance of popular songs over time i.e. how popular or not-so-popular songs when originally released, are doing today? We aim to develop a baseline model which calculates the current popularity of an old song as a measure of time and initial popularity. We will further enhance this model by adding features which cause over-performance and under-performance.

# Approach

## Analysis of Billboard Hot 100 and Spotify Popularity Score

Commercial Sales data and Radio airplay are primary sources which reflect the popularity of a song. After the arrival of the internet, several other parameters like the number of song audio downloads, streaming activity on apps like Pandora, song video views on platforms like YouTube and song name search on search engines like Google also depict the performance of the song. However, data from such platforms is available only for the last few years. For example, Google trend data is available only since 2004. YouTube view data is available only since 2011. While these platforms are excellent sources to get current popularity, they fail to provide long-term popularity of songs released in the 60s, 70s etc. Thus, we decided to use chart rankings and FM airplay as our primary source for longitudinal data.

### Data Collection and Cleaning

Our preliminary data includes songs from the year 1955 to 2017. In all our analysis, we have considered songs having the same name but different artists as 2 different songs.

1. The initial popularity of songs
   We used Billboard Hot 100 Rank Yearly to obtain the initial popularity of 5200 songs from the year 1960 to 2012.

2. Current Popularity of all songs in 2018
   We used Spotify API to obtain the current popularity of any song. Spotify does not provide an exact string search. Searching a track name gives many results. We, thus, used track title, artist name and year of release to filter the exact song in the results. After applying these filters, we could get the current popularity of 3161 songs. Thus, we could not consider 100 songs from every year. For example, we have 25 songs for the year 1964.

### Baseline Model

Considering the data of initial and current popularity of 3161 songs, we came up with the following baseline model.

$$C = \mathrm{M}^{2.82} * e^{-0.01*Y-8.54}$$

where
C is current popularity (2018) on a scale of 1-100 (100 is best)
M is initial popularity on a scale of 1-100
Y = 2018 - Year in which the song appeared on the Billboard chart

Rank to Popularity Conversion
Initial Popularity(M) = 101 - Billboard Hot 100 Rank
As we considered top 100 songs of each year on the billboard, we scaled their popularity between 90 to 100 scale. This was done to match with Spotify's popularity scale for top performing songs.

Model Intuition
We expected to find an exponential decrease in songs popularity over the years.
Thus, we used log-level regression to find out the relationship between different C, M, Y parameters.

Regression Model
Equation - $log(C) = beta0 * log(M) + beta1 * Y + beta2$
Model Input - $log(M)\ and\ Y$
Model Output - $beta0, beta1, beta2, log(C)$
We trained on 80% data (2500 points) and tested on 20% (600 data points).

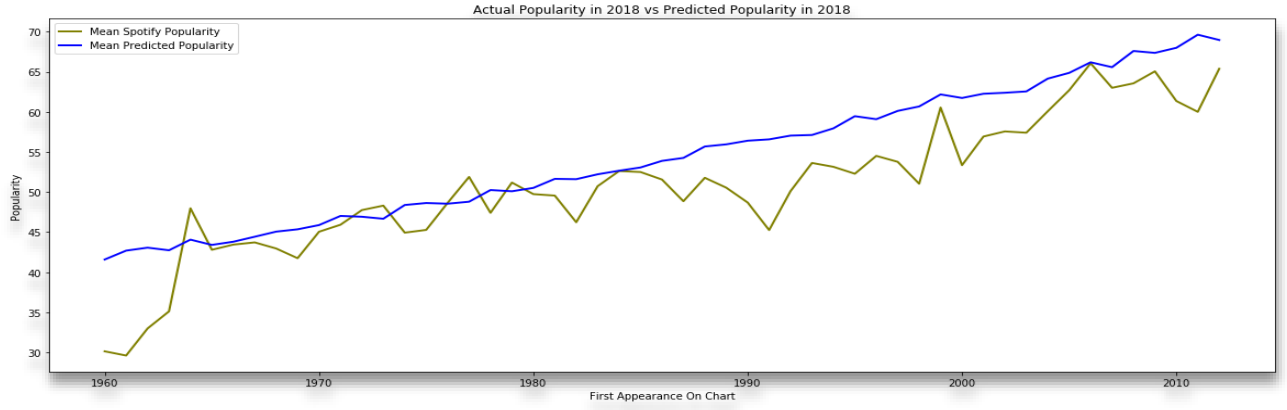
## Analyzing Model's Performance

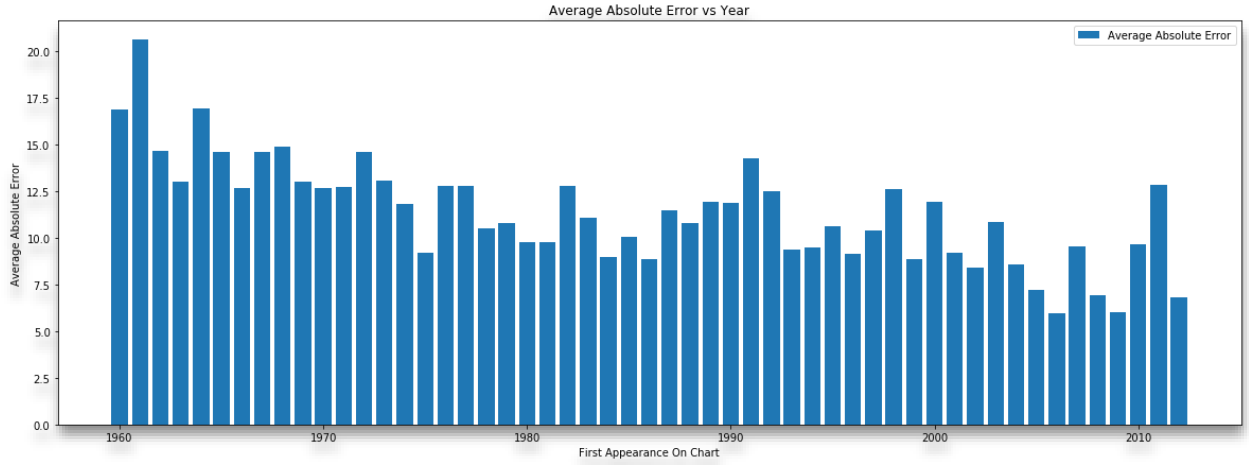
Results
Train MSE - 0.18
Test MSE - 0.21

**What we think is correct about our formula?**

1. C is directly proportional to M - This is correct as songs which were quite popular initially should be relatively popular to other songs today as well.
2. C is inversely proportional to Y - This is correct as popularity should decrease with time

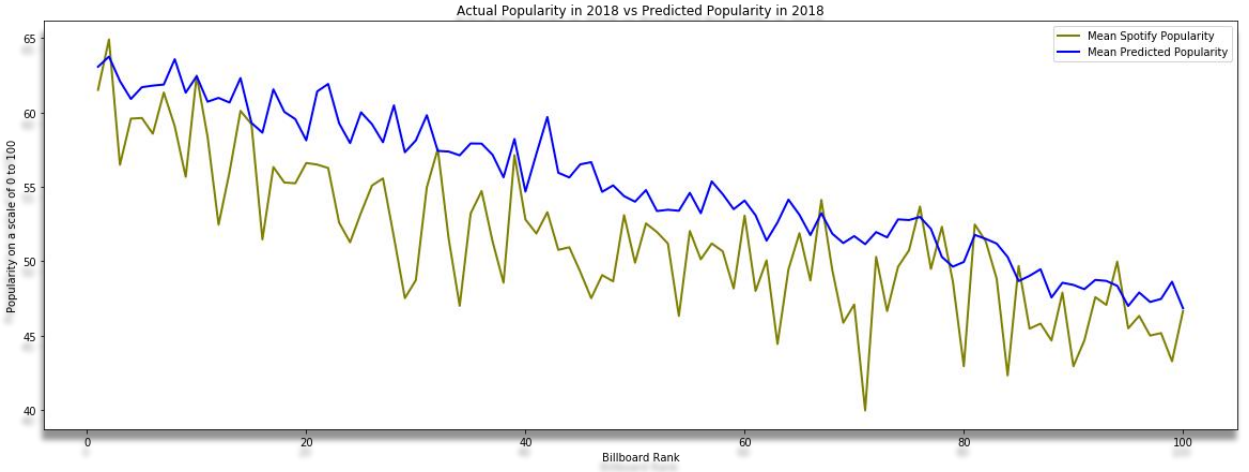
As visible from the chart below, our model performs well over the years.

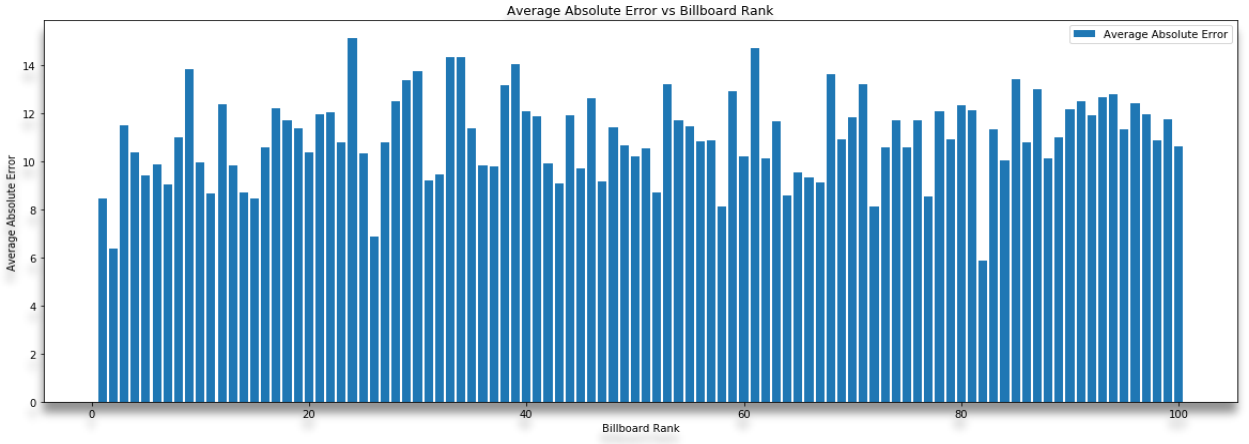





The average absolute error for our model is 11.04. Except for the large absolute error in 1960, our model gives an accurate current popularity score to a song.

Further, our model does not have anomalies when plotted over different billboard rank.



**What we think could go wrong in our analysis?**
As our model has been obtained by training a regression model on 80% of the dataset, it's bound to perform well. It would be interesting to see how this model performs on the larger test dataset. The rankings from WCBS-FM dataset would be a good way to test our model on longitudinal data of a song.

**Over-Performing Songs**

$$Model\ Performance\ =\ Actual\ Popularity\ Today/Predicted\ Popularity\ Today$$

We used the above formula to obtain top 20 most over-performing songs since 1960.
The top overperformer "Come Together" by Beatles has been ranked at #202 on the list of "The 500 Greatest Songs of All Time" by Rolling Stone. Many songs become hit several times over when they are "covered" by various artists. One such clear example is "Riders on the storm" which has been covered 23 times since its release.

| | Title | Artist | Original_Year | Initial_Rank | Billiboard_Popularity | Spotify_Popularity | Predicted_Popularity |
|---|---|---|---|---|---|---|---|
| **425** | Come Together | Beatles | 1969 | 85 | 91.515152 | 77 | 40.712812 |
| **375** | I Say A Little Prayer | Aretha Franklin | 1968 | 93 | 90.707071 | 74 | 39.312069 |
| **526** | Riders On The Storm | Doors | 1971 | 99 | 90.101010 | 72 | 39.750661 |
| **167** | I Saw Her Standing There | Beatles | 1964 | 95 | 90.505051 | 67 | 37.533878 |
| **48** | Take Five | Dave Brubeck | 1961 | 95 | 90.505051 | 65 | 36.424585 |
| **164** | You Really Got Me | Kinks | 1964 | 78 | 92.222222 | 70 | 39.576961 |
| **126** | Ring Of Fire | Johnny Cash | 1963 | 80 | 92.020202 | 68 | 38.941595 |
| **639** | Space Oddity | David Bowie | 1973 | 97 | 90.303030 | 71 | 40.810617 |
| **634** | Money | Pink Floyd | 1973 | 92 | 90.808081 | 72 | 41.457555 |
| **825** | Rock And Roll All Nite | Kiss | 1976 | 95 | 90.505051 | 73 | 42.319330 |

**Under-Performing Songs**

| | Title | Artist | Original_Year | Initial_Rank | Billiboard_Popularity | Spotify_Popularity | Predicted_Popularity |
|---|---|---|---|---|---|---|---|
| **3057** | Lighters | Bad Meets Evil feat. Bruno Mars | 2011 | 34 | 96.666667 | 1 | 72.311093 |
| **2413** | Get It On Tonite | Montell Jordan | 2000 | 24 | 97.676768 | 1 | 66.705789 |
| **3005** | Take It Off | Ke$ha | 2010 | 59 | 94.141414 | 1 | 66.442075 |
| **35** | Apache | Jørgen Ingmann | 1961 | 35 | 96.565657 | 1 | 43.729778 |
| **517** | Watching Scotty Grow | Bobby Goldsboro | 1971 | 78 | 92.222222 | 1 | 42.446615 |
| **79** | Snap Your Fingers | Joe Henderson | 1962 | 66 | 93.434343 | 1 | 40.248403 |
| **20** | Poetry In Motion | Johnny Tillotson | 1960 | 87 | 91.313131 | 1 | 36.977543 |
| **2988** | Your Love Is My Drug | Ke$ha | 2010 | 28 | 97.272727 | 2 | 72.864577 |
| **3056** | Blow | Ke$Ha | 2011 | 33 | 96.767677 | 2 | 72.524375 |
| **2446** | Party Up (Up In Here) | DMX | 2000 | 71 | 92.929293 | 2 | 57.961891 |

Sometimes the popularity of artist also affects the popularity of song. For example, Kesha, a very popular singer in 2010-11, released only 1 song between 2013-2017 due to her involvement in legal matters. We suspect that her decrease in popularity caused several of her songs (3 songs in our top 10 list) to underperform.

# Analysis of WCBS Dataset

### Data Collection

WCBS-FM is one of the highest-rated classic hits stations in the United States. It ranks the top 500 songs of all time, as voted by the station's listeners. We used this dataset which was provided by www.45cat.com starting from year 1973, but unfortunately, we could only find 8 years data(1973, 1979, 1981, 1983, 1985, 1987, 1989, 1993, 1995). Hence the analysis was done using only 8 years dataset. One thing to note is, WCBS does not rank individual songs but ranks 7" vinyl, and each vinyl has two songs in it. Hence, we got rating of 1000 songs each year.

### Data Cleaning

The data-set had some unnecessary information in columns like label, vinyl cover, format and cat#. First, we removed these columns. Data-set contained a column which had information of artist, song names and rank in a single cell. So, we split this column into different columns of artist, song name and rank. As vinyl has two songs in it, we gave two songs (of the same vinyl) same rank. In Year column some values had month, date and year and some had only year. To maintain uniformity in the column and to facilitate the use of the column we kept only the year.

| | Rank | Song | Year |
|---|---|---|---|
| 1 | 251.0 | The Lonnie Donegan Skiffle GroupA: Rock Island... | Mar 1956 |
| 2 | 252.0 | Isley BrothersA: This Old Heart Of Mine (Is We... | 28 Jan 1966 |
| 3 | 253.0 | The Righteous BrothersA: Ebb TideB: (I Love Yo... | Nov 1965 |
| 4 | 254.0 | The AngelsA: My Boyfriend's BackB: (Love Me) N... | Jul 1963 |
| 5 | 255.0 | Johnnie And JoeA: Over The Mountain Across The... | Apr 1957 |

Fig 1. Clean Part 1

| | Rank | Year | Artist | Song1 |
|---|---|---|---|---|
| 0 | 252 | 1966 | Isley Brothers | This Old Heart Of Mine (Is Weak For You) |
| 1 | 253 | 1965 | The Righteous Brothers | Ebb Tide |
| 2 | 254 | 1963 | The Angels | My Boyfriend's Back |
| 3 | 255 | 1957 | Johnnie And Joe | Over The Mountain Across The Sea |
| 4 | 256 | 1965 | The 4 Seasons | Let's Hang On! |

Fig 2. Final Clean Data.

**Observation from Plots**

After cleaning the data-set we wanted to gain some insights, hence we made following plots: 1. Rank vs Year 2. Popularity vs Year for each song. This was done so that we could find out how the rank of the song changes with time. And as we expected rank increased i.e popularity of song decreased as time passed. Below are some of the plots which support our expectation. Not all the plots were this clear in showing relationship.

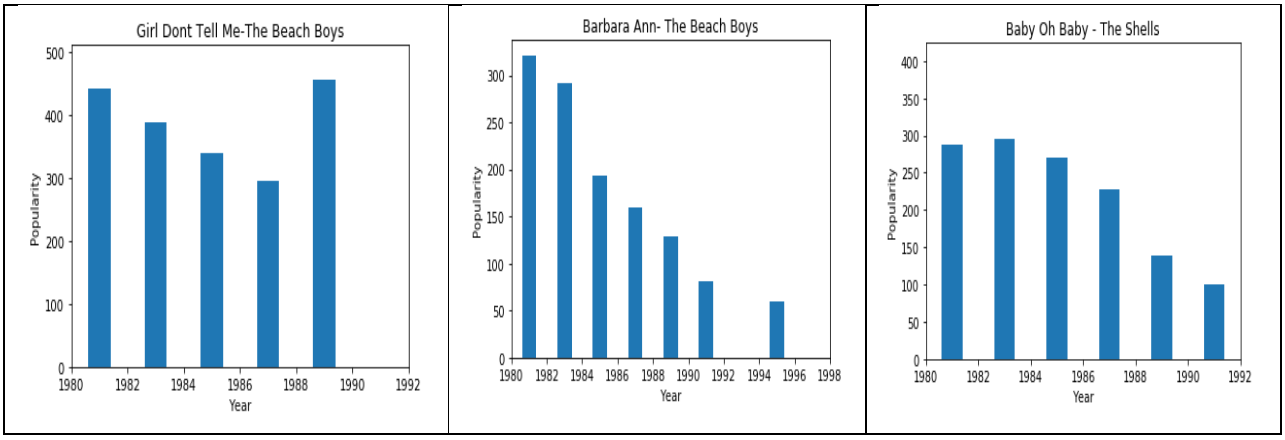



Fig 3. Popularity vs Year

There were also some anomalies observed in the plots. The first type of anomaly (Type 1) was rank of song increased till certain year but then started decreasing (i.e. popularity decreased at first then popularity started increasing). This anomaly is depicted in fig 4. The possible cause of this could be band got famous again, there was cover that was sung, song got incorporated into a movie, etc.

The other type of anomaly (Type 2) that was observed was rank of the song improved in subsequent years (popularity increased). This might be since song was not appreciated at first, but as more and more people started listening it got famous. This anomaly is depicted in fig 5. Here My Prayer is a famous song which was originally written in 1926. Since then many recordings of this song has been done and among them most famous one was the recording by the 'The Platters' in 1956 since then it has been incorporated into various movies

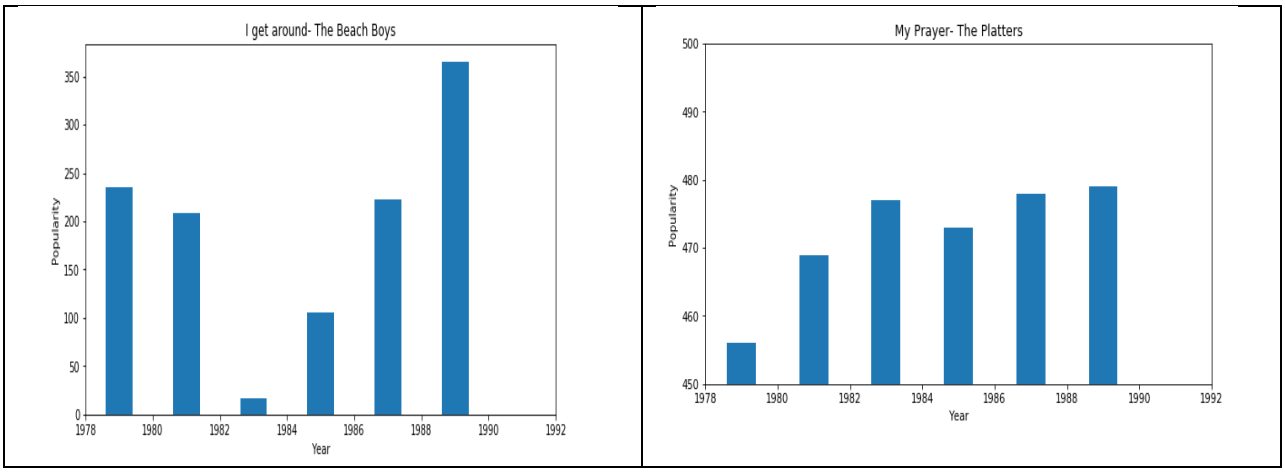



Fig 4 Anomalies

**Challenges in Dataset**

The dataset is of WCBS FM which ranked Vinyl Albums, these vinyls contains two sides A and B having one song per side. Hence for each rank there were two songs and while cleaning we gave same rank to both the songs. Doing this might not be an accurate measure of song rank. Also, we are not sure that the rank given to the vinyl was due to the performance of both the songs or best song. If rank was given based on both the songs, then technique used by us might not be an accurate measure to predict the trends. The dataset also contained multiple entries for the same vinyl, these multiple entries was since it was first released in Canada then in U.S.A by different distributing company. Hence WCBS considered them different (weird). So, offset this we chose the best rank for the vinyl, hence the songs.

As discussed above, the popularity of any song gradually decreases with the time. But this doesn't always hold true and we have analyzed various reasons for any song to not follow the gradual decay curve.

## Analysis of Weekly Billboard Charts

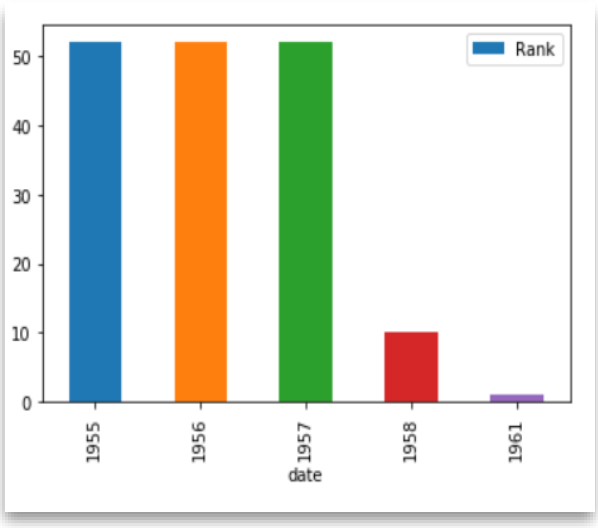
The reason for doing analysis on weekly charts instead of year-end chart is to get more inclusive idea of how popularity of songs changes over time. There are many songs which might get into Billboard hot 100 weekly list but won't appear in year-end Hot 100 charts, so analyzing weekly charts will give larger dataset and hence deeper insights.
We have scraped the data from Billboard website for hot-100 songs of each week from 1955-2017. The data has columns: **[**date, Rank, Title, Artist**]**

The dataset contains all the recordings by any artist for given year, that appeared on the billboard hot-100 chart weekly. Below are the steps followed for analysis:

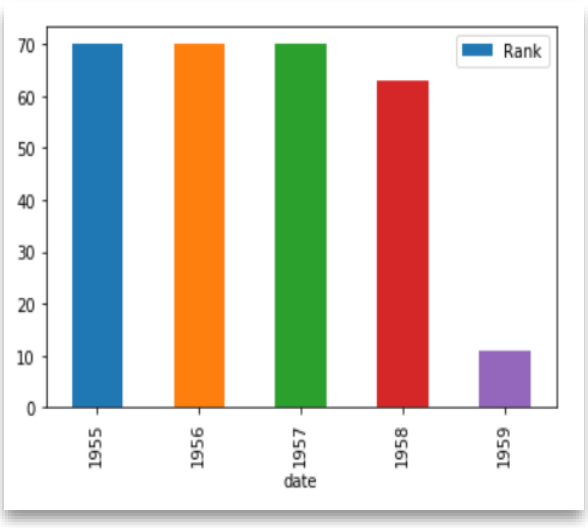
1. Pick particular year, say 1960, and get all the weekly hot-100 recordings from that year. Since different recordings are present on the chart at different ranks in different weeks, we have considered the peak position of recording on the chart for analysis.
2. Now, check if any song which were present on chart in 1960, re-appeared ever after that.

3. If the song, re-appeared on the charts, what are the different reasons for the song to be over performing.

Expected behavior for any recording is shown below for two popular songs of year 1955. Here rank stands for popularity of the song.




'*All I have to do Is Dream*' by The Everly Brothers, remained in the Billboard weekly hot-100 chart for consecutive 4 years. As depicted popularity of the song decreased over the years.

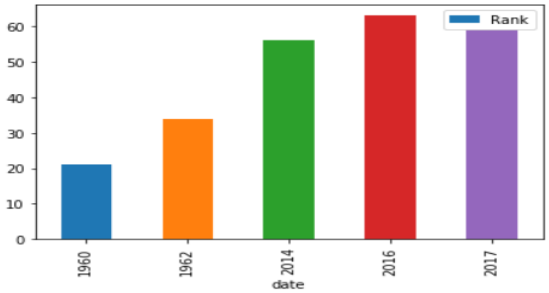



'*Chantilly Lace*' by Big Boppers, remained popular during initial year of release but decreased with the time.

However, upon analysis, we have found different recording which re-appeared on the chart even after many years of its release, is basically because of some events affecting the popularity of the song. Some of these events are

1. **Use in Movie**
   Do you Love Me? by The Contours originally released in 1962. This song appeared o chart in 1988 again because the song was used in 1988 movie, *Alvin and the Chipmunks*.

2. **Occasion related Songs**
   The Christmas Song by Nat King Cole, originally released in 1960, appeared on charts over the time, as the data from billboard hot-100 reveals. If we plot the popularity of this song over the years when it has appeared on the weekly charts, the popularity has increased over the years. The noticeable thing here is that the song has appeared on charts only during the Christmas time. The bar plot below depicts the same



3. **Cover of the song**
   *The Twist* by Chubby Checker released in 1960 came on top of billboard list in 1960 and was present on the Billboard chart till 1962. The reason is because song gave birth to the Twist dance craze, making it popular among celebrities and their followers. The recording re-appeared on the chart again in 1988, as it was recorded by The Fat Boys and not the original artist.

4. **Social Media Impact**
   Billie Jean by Michael Jackson originally released in 1983, #1 on Billboard weekly hot-100 chart, never appeared on charts again until May 2014. In May 2014, a viral video of a high school-aged teenager imitating Jackson's Motown 25 performance of the song helped the song re-enter the Billboard Hot 100 at number 14. Similarly, Livin' on a Prayer by Bon Jovi released in 1986, #1 on Billboard list, re-charted again in 2013 after a video went viral.

5. **Death of Artist**
   I would Die 4 U by Prince and The Revolution, originally released in 1984, reached #8 on Billboard charts but never appeared on the charts again until 2016, where the song re-charted on Billboard Hot 100 at #39 after Prince's death. Similarly, Under Pressure by David Bowie originally released in 1981, re-charted again in 2016 after Bowie's Death.

# Future Work

Currently, our analysis has been done on 3 separate datasets i.e.

1. Billboard Hot 100 Year Dataset and Spotify Popularity Score
2. WCBS-FM Top 500 Dataset
3. Billboard Weekly Data

Next, we plan to merge these 3 datasets and generate popularity score for the same songs over a larger span of years. This would help us to better test our baseline model over longitudinal data. We also plan to analyze more features which affect the over-performing/under-performing songs. Incorporating such features in our regression model would result in better accuracy.

We intend to analyze the decrease in song popularity as a power-law function. This can be achieved only if we could find songs which have been popular over multiple years. We have found *https://top40weekly.com/, https://singleschronology.wordpress.com/2014-2/* as 2 promising datasets which provide us the popularity of songs in 2013. Further, we plan to search for WCBS Dataset for more years as it provides popularity of a lot of old songs.

We have handled the challenges in a certain way and in future, if we have a larger dataset, we could improvise on the techniques used while handling challenges in the dataset. With the weekly billboard hot-100 dataset, songs which were present on the charts for longest and for consecutive weeks, can be analyzed to get the reasons for popularity of top songs. And this analysis can be used to predict the popularity of songs in future.