Project Final Report: Do Popular Songs Endure?

Objective

To analyze the performance of popular songs over time i.e. how popular or not-so-popular songs when originally released, are doing today?

Contributions of this paper

We propose a model which calculates the current popularity of an old song in terms of initial rank, time and artist popularity. We evaluate multiple models and show, through error distributions and root mean square error, that our proposed model performs quite well on test datasets. Using this model, we provide a list of 20 over and underperforming songs since 1958. We analyze the reasons for their outperformance and underperformance. While coming up with this model, we extensively study 4 datasets, namely Billboard Hot 100 (Yearly & Weekly), Spotify, WCBS and Last.fm. While Billboard and Spotify data is the primary dataset used for building our model, we use WCBS and Last.fm dataset to validate our model and do an analysis of songs popularity during interim years. We further analyze and justify the following interesting observations -

1. Social media impact, re-recording of a track, use of the track in a movie, death of the artist and occasion like Christmas or new year, are the primary reasons of over performance of songs.
2. Reasons for underperformance - Tragic events resulting in sales loss, Band disbanding, listeners loss, etc.
3. The popularity of a song decreases with the number of years elapsed since it's release. Specifically, songs recorded in recent years are more popular than older songs.
4. The current popularity of songs decreases as the rank in the chart increases (numerically). For example, songs ranked 100 in the initial billboard chart are lesser popular today than songs ranked 1.
5. For music groups with many charted songs: early songs endure more than their later works.
6. WCBS data-set suggests that popularity of song follows normal distribution.

Data Collection

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.
In particular, we have analyzed data from 4 different sources. Each of these datasets provides their unique advantages during our analysis.

Billboard Hot 100

Data

1. Yearly Dataset

We obtain 5200 songs using 'Billboard 100 Chart' for the years 1960 to 2012 from http://billboardtop100of.com/

2. Weekly Dataset

This includes songs from the year 1958 to 2017. In all our analysis, we have considered songs having the same name but different artists as 2 different songs.

We specifically obtain Track name, artist, rank, number of weeks it stayed on chart, date of appearance on chart.

Challenges
Both Billboard and Spotify have a difference in track and artists names for few songs. Ex. Billboard mentions artist as Akon while Spotify mentions Akon Featuring Lil Wayne & Young Jeezy. This sometimes led to a loss in data.

Why this dataset?
Billboard is the standard for music data. It is the only source which has data since 1958.

Spotify

Data
We used Spotify API to obtain the current popularity of any song and artists popularity in 2018

Challenges
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.

Why this dataset?
Spotify popularity score combines songs sales data and radio airplays, thus providing a more robust popularity measure as compared to WBCS.

WCBS

Data
We used WCBS yearly data for the validation of our model.

Challenges
This dataset has ranked 7” vinyls which has two songs on it. Hence it is difficult to know which song has contributed to what extent for the ranking of vinyl. And which song should be kept in creating the dataset? Also, the data available was only of 8 years. Hence it couldn’t be used for modeling.

Why this dataset?
This is the only authentic data set that provides a rating of old songs (1960’s). And as WCBS-FM is one of the most highest-rated stations in the United States, hence its longitudinal data is valuable.

Last.fm

Data
We have used Last.fm data to get track’s listeners count and play count

Challenges
Since, last.fm has officially released in 2002, the numbers are more biased toward recent years. Moreover, the play counts includes song being played by the same user again and again.

Why this dataset?
With Spotify popularity, many features apart from play count, sales data are included in order to get the popularity. However, Spotify doesn’t declare how it calculates these popularity score. Thus our assumption of exponential decay for popularity with time fails. Hence, we used actual listener count data from last.fm to do the analysis.

Baseline Model

<u>Dataset Used</u>: Billboard Hot 100 (Yearly)

We propose a baseline model which calculates the current popularity of an old song as a measure of time and initial popularity. We considered the following 2 different relationships between the parameters C, M, and y

where,
C - Current Popularity
M - Billboard Rank in the year of its appearance on Billboard
y - Number of years elapsed since its appearance

1st Model

C = beta0 * M + beta1 * y + beta2

We use linear regression to get the coefficients *beta0, beta1 and beta2.*

Results
Model: C = -0.14*M - 0.48*y + 73.16

Train MSE: 193.227
Test MSE: 179.80

2nd Model

C = M^beta0 * e^(beta1 * y + beta2)

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.

Results
Model: C = M^0.1 ∗ e ^ (−0.01∗y + 4.55)

Train MSE: 321.12
Test MSE: 348.06

Conclusion

Though we anticipate that the exponential model should perform well, we observe that the 1st model provides a lower MSE than the 2nd model. Thus, we consider the linear model for our analysis.

Improved Model

We included 4 new features in our model

1. Number of weeks the song stayed on the Billboard charts
2. Artist popularity
3. Re-use of the song in a movie, between release year and current year
4. Re-recording of the song, between release year and current year

We observed that except artist's popularity, none of the other features contributed significantly to the songs current popularity. Thus, we used linear regression to calculate the coefficients of the linear relationship. Using the results of linear regression, we propose the following linear model -

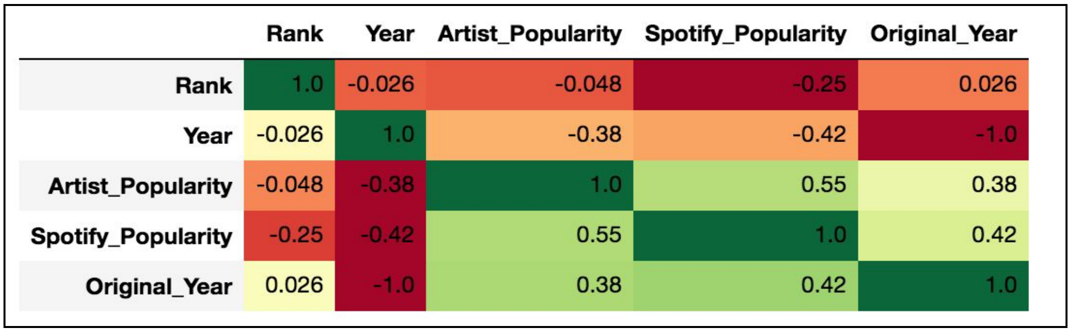
$$\mathbf{C = -0.13*M - 0.28*y + 0.46*A + 37.29}$$

where A is the artist popularity in the current year.

Train MSE: 145.95
Test MSE: 157.14

Evaluating Our Model

1. **Correlation Matrix**
   We calculated the following correlation matrix of "Spotify_Popularity" (Current Popularity) with all the other features and made the following observations.



	Rank	Year	Artist_Popularity	Spotify_Popularity	Original_Year
Rank	1.0	-0.026	-0.048	-0.25	0.026
Year	-0.026	1.0	-0.38	-0.42	-1.0
Artist_Popularity	-0.048	-0.38	1.0	0.55	0.38
Spotify_Popularity	-0.25	-0.42	0.55	1.0	0.42
Original_Year	0.026	-1.0	0.38	0.42	1.0

   1. Spotify popularity is inversely correlated to initial rank
      This is correct as songs which were quite popular initially (lower ranked numerically) should be relatively popular to other songs today as well

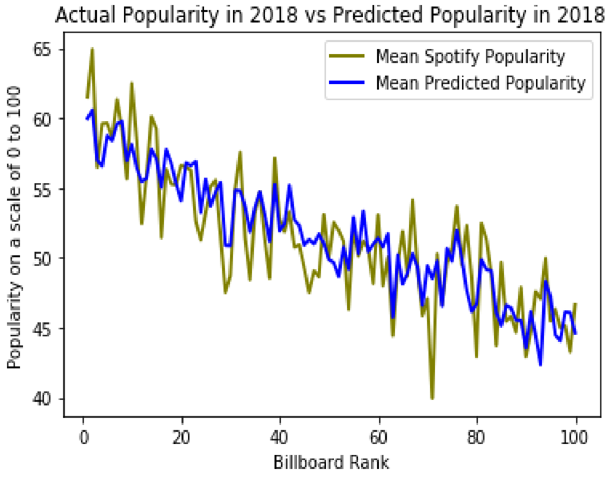
   2. Spotify popularity is inversely proportional to Y. This is expected as popularity should decrease with time.

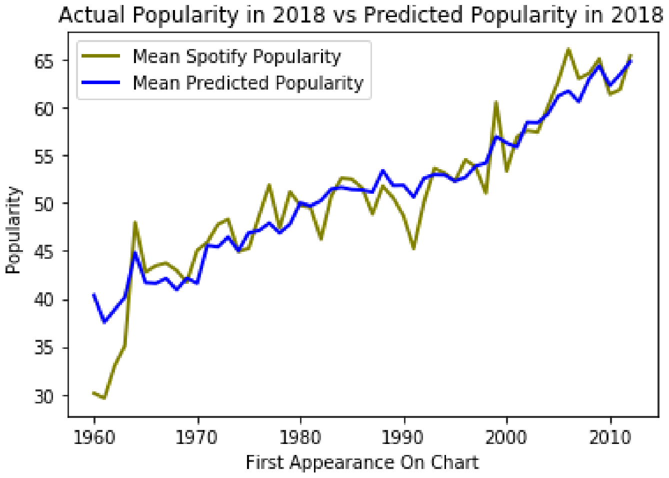
3. Spotify popularity is highly correlated to Artist_Popularity.

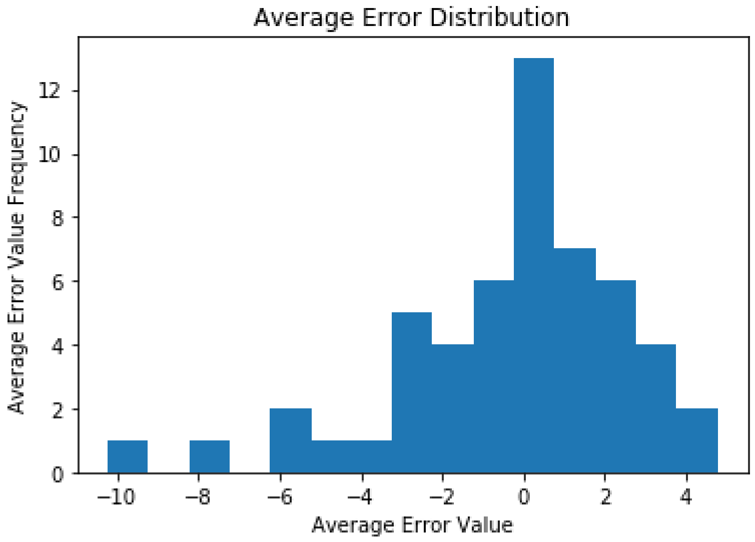
2. **Error Distribution**

   In order to verify that our model performs equally well for different billboard ranks, we plot the Mean Spotify Popularity and Mean Predicted Popularity vs Billboard Rank. The plot validates that our model's predictions follow Spotify popularity closely. It also verifies that song's popularity decreases as the rank increases (numerically).

   In order to verify that our model performs equally well for different years, we plot the Mean Spotify Popularity and Mean Predicted Popularity vs First Appearance on Chart. Except for the initial difference in popularity in 1960, our model predicts well for the other years.







The average error for our model varies from -3 to 3 primarily. As visible from the plots, our model's predicts fairly well.

Data Exploration and Analysis

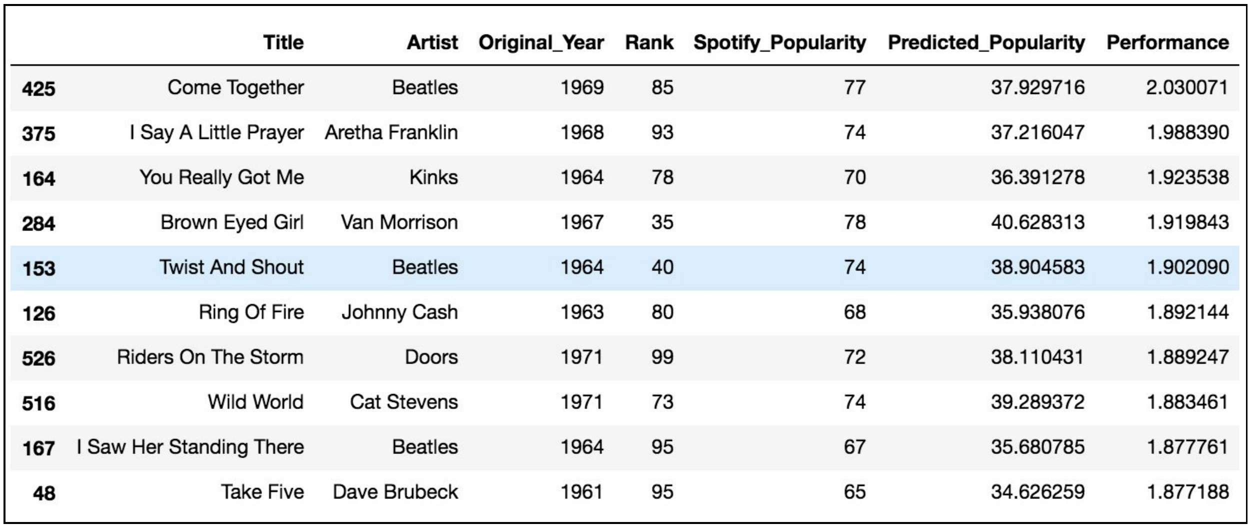
Experiment 1: Over and Under-Performing songs before 1980

Dataset Used: Billboard Hot 100 (Yearly)

Model Performance = Actual Popularity Today/Predicted Popularity Today

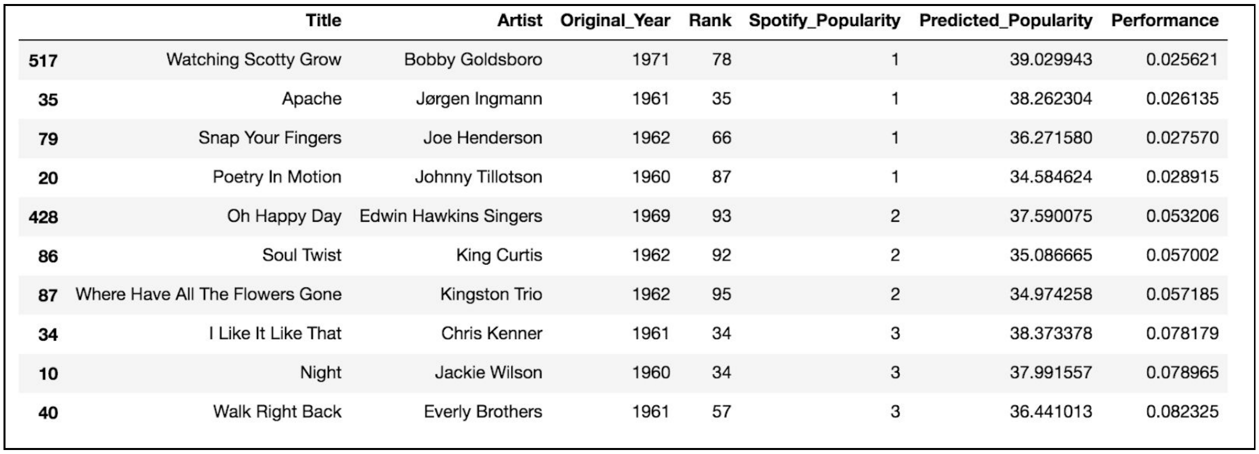
We used the above formula to obtain top 10 most over-performing songs between 1960 and 1980. The top over performer “Come Together” by Beatles has been ranked at #202 on the list of "The 500 Greatest Songs of All Time" by Rolling Stone. Many songs become hit several times over when they are "covered" by various artists. One such clear example is “Riders on the storm” which has been covered 23 times since its release.

List Of Over-Performing Songs before 1980



	Title	Artist	Original_Year	Rank	Spotify_Popularity	Predicted_Popularity	Performance
425	Come Together	Beatles	1969	85	77	37.929716	2.030071
375	I Say A Little Prayer	Aretha Franklin	1968	93	74	37.216047	1.988390
164	You Really Got Me	Kinks	1964	78	70	36.391278	1.923538
284	Brown Eyed Girl	Van Morrison	1967	35	78	40.628313	1.919843
153	Twist And Shout	Beatles	1964	40	74	38.904583	1.902090
126	Ring Of Fire	Johnny Cash	1963	80	68	35.938076	1.892144
526	Riders On The Storm	Doors	1971	99	72	38.110431	1.889247
516	Wild World	Cat Stevens	1971	73	74	39.289372	1.883461
167	I Saw Her Standing There	Beatles	1964	95	67	35.680785	1.877761
48	Take Five	Dave Brubeck	1961	95	65	34.626259	1.877188

List of Underperforming Songs before 1980



	Title	Artist	Original_Year	Rank	Spotify_Popularity	Predicted_Popularity	Performance
517	Watching Scotty Grow	Bobby Goldsboro	1971	78	1	39.029943	0.025621
35	Apache	Jørgen Ingmann	1961	35	1	38.262304	0.026135
79	Snap Your Fingers	Joe Henderson	1962	66	1	36.271580	0.027570
20	Poetry In Motion	Johnny Tillotson	1960	87	1	34.584624	0.028915
428	Oh Happy Day	Edwin Hawkins Singers	1969	93	2	37.590075	0.053206
86	Soul Twist	King Curtis	1962	92	2	35.086665	0.057002
87	Where Have All The Flowers Gone	Kingston Trio	1962	95	2	34.974258	0.057185
34	I Like It Like That	Chris Kenner	1961	34	3	38.373378	0.078179
10	Night	Jackie Wilson	1960	34	3	37.991557	0.078965
40	Walk Right Back	Everly Brothers	1961	57	3	36.441013	0.082325

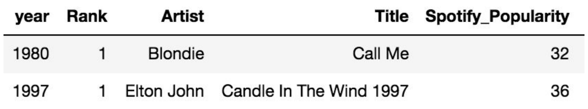
Experiment 2: Over and Under-Performing Song Detection using Outlier Detection

Dataset Used: Billboard Hot 100 (Yearly)

For each rank

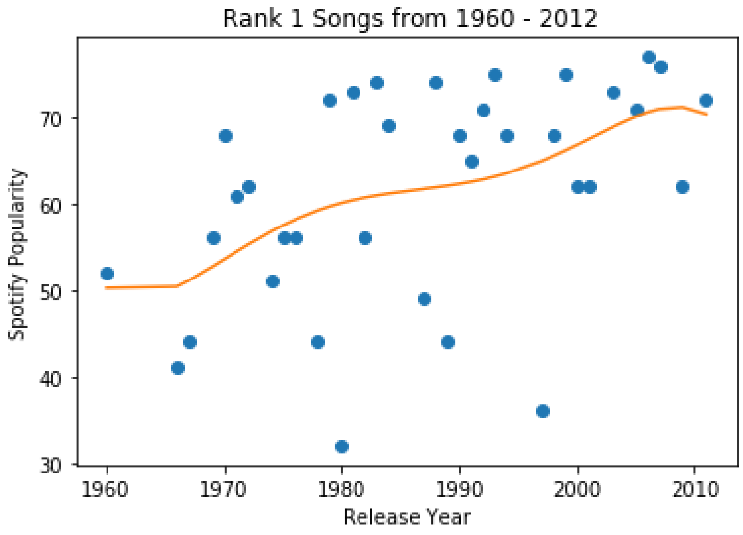
1. Created a scatter plot of all the songs which attained that rank.
2. Fit a curve which best represents the data in the plot
3. Detect the outliers in the plot using IQR detection method
4. Analyze if these outliers are under-performing or over-performing

Using the IQR(Inter-Quartile Range) method of outlier detection, we find out that 2 songs released in 1980 and 1997 have the lowest popularity and are thus outliers.

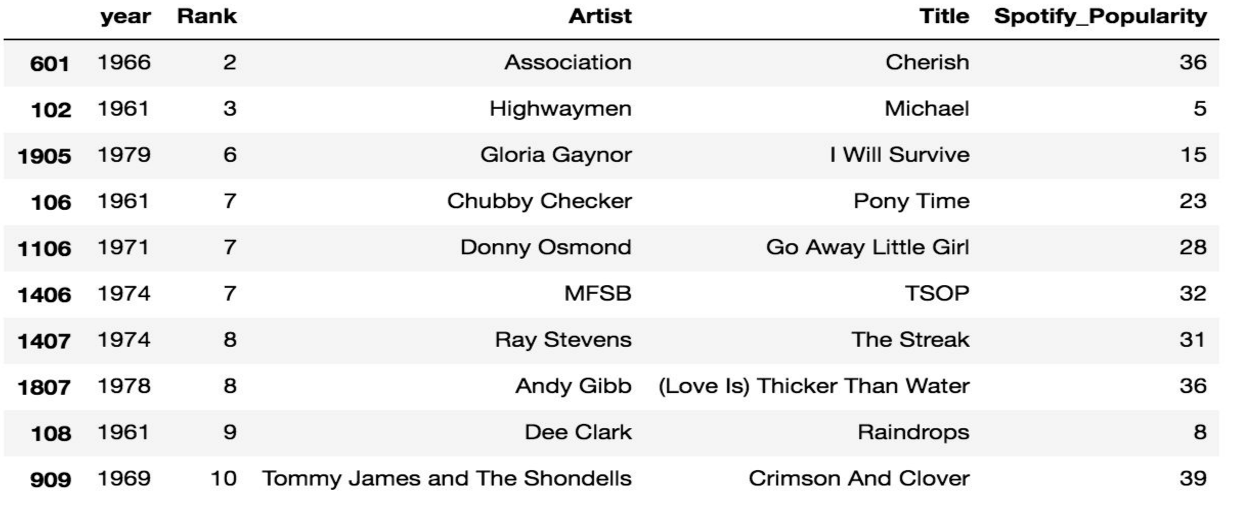


year	Rank	Artist	Title	Spotify_Popularity
1980	1	Blondie	Call Me	32
1997	1	Elton John	Candle In The Wind 1997	36

These songs have been ranked 1 in the past but are now under-performing. The band "Blondie" disbanded in 1982 and reformed in 1997. We suspect that this led to its decrease in popularity of the songs before the reformation.

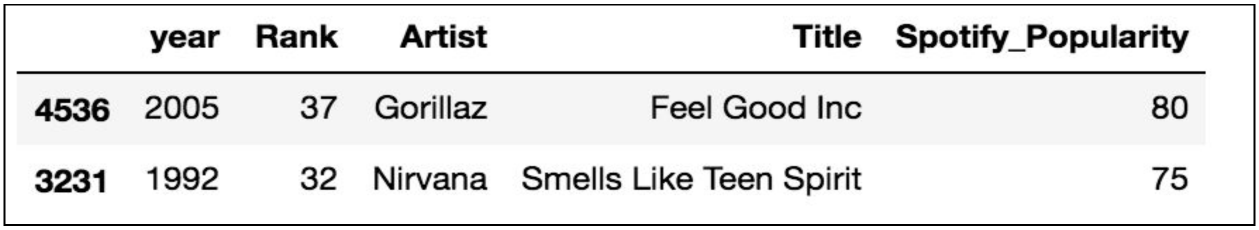


List of Under-Performing Songs using IQR outlier detection



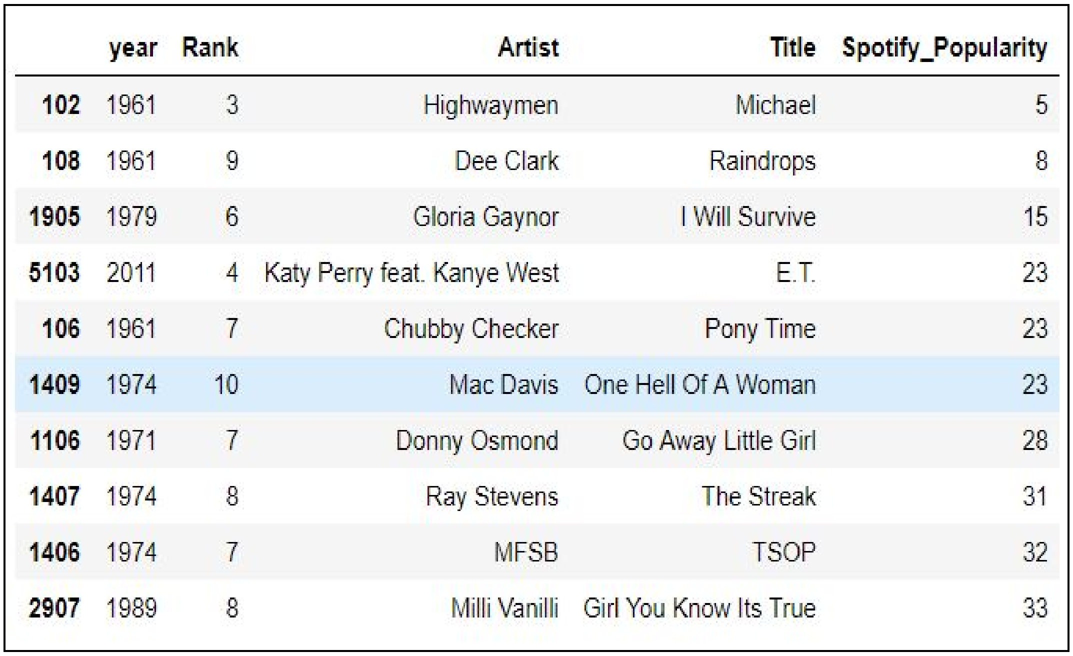
	year	Rank	Artist	Title	Spotify_Popularity
601	1966	2	Association	Cherish	36
102	1961	3	Highwaymen	Michael	5
1905	1979	6	Gloria Gaynor	I Will Survive	15
106	1961	7	Chubby Checker	Pony Time	23
1106	1971	7	Donny Osmond	Go Away Little Girl	28
1406	1974	7	MFSB	TSOP	32
1407	1974	8	Ray Stevens	The Streak	31
1807	1978	8	Andy Gibb	(Love Is) Thicker Than Water	36
108	1961	9	Dee Clark	Raindrops	8
909	1969	10	Tommy James and The Shondells	Crimson And Clover	39

List of Over Performing Songs using IQR outlier detection



	year	Rank	Artist	Title	Spotify_Popularity
4536	2005	37	Gorillaz	Feel Good Inc	80
3231	1992	32	Nirvana	Smells Like Teen Spirit	75

Dataset used: Billboard Hot 100 (Weekly)
Below is the list of top underperforming songs of all time, analysed on weekly data using IQR outlier detection.



	year	Rank	Artist	Title	Spotify_Popularity
102	1961	3	Highwaymen	Michael	5
108	1961	9	Dee Clark	Raindrops	8
1905	1979	6	Gloria Gaynor	I Will Survive	15
5103	2011	4	Katy Perry feat. Kanye West	E.T.	23
106	1961	7	Chubby Checker	Pony Time	23
1409	1974	10	Mac Davis	One Hell Of A Woman	23
1106	1971	7	Donny Osmond	Go Away Little Girl	28
1407	1974	8	Ray Stevens	The Streak	31
1406	1974	7	MFSB	TSOP	32
2907	1989	8	Milli Vanilli	Girl You Know Its True	33

Experiment 3: Do songs early in a group's career survive better or worse than those later in the career?

Dataset Used: Billboard Weekly Data with Song Popularity from Spotify

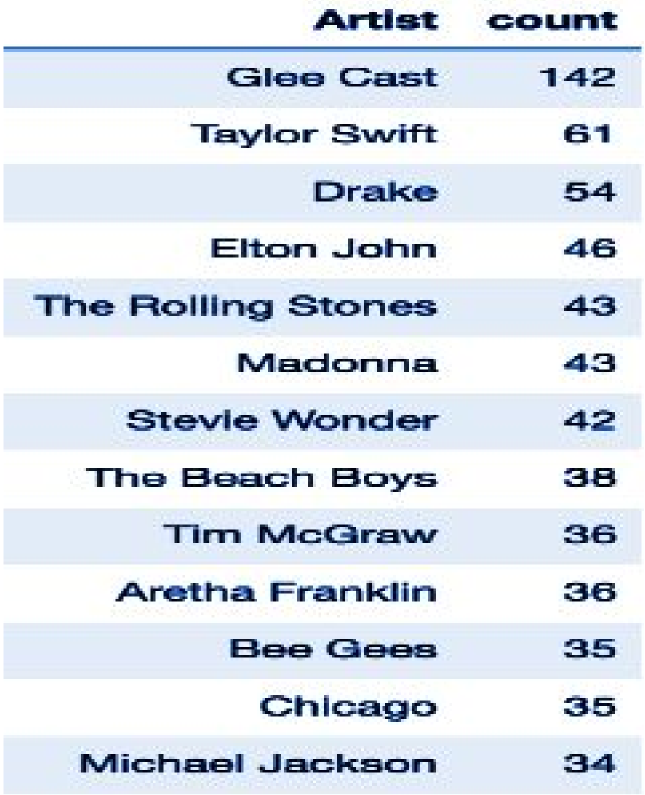
Work

We find top 15 music groups
having the most number of charted songs
Between 1958 to 2017.

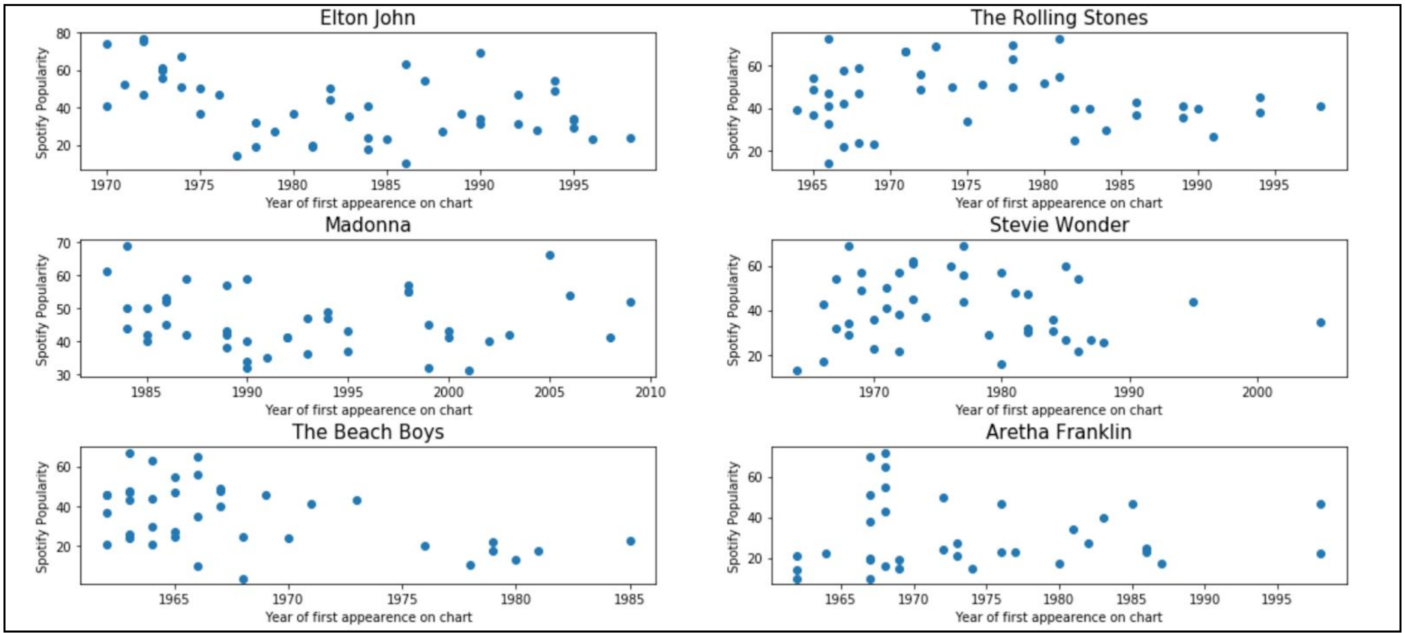
We create a scatter plot of the songs
recorded by each of these music groups
over the years.

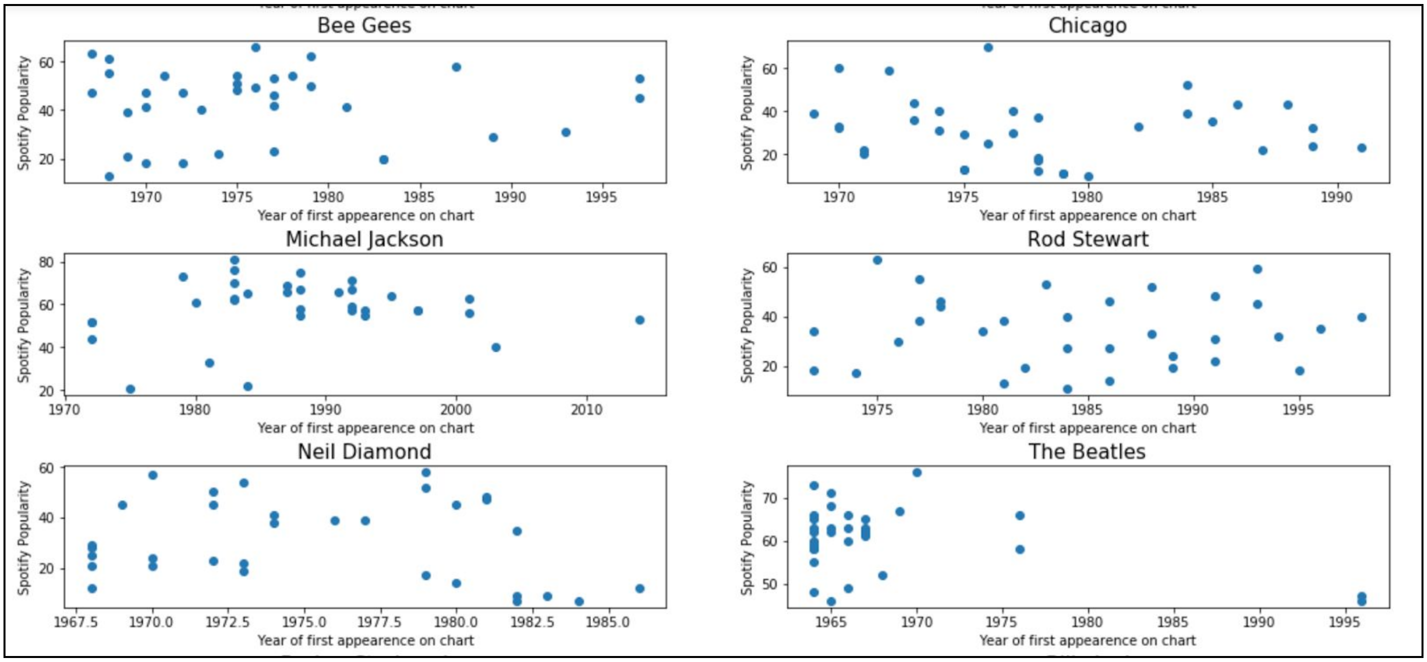
Observation

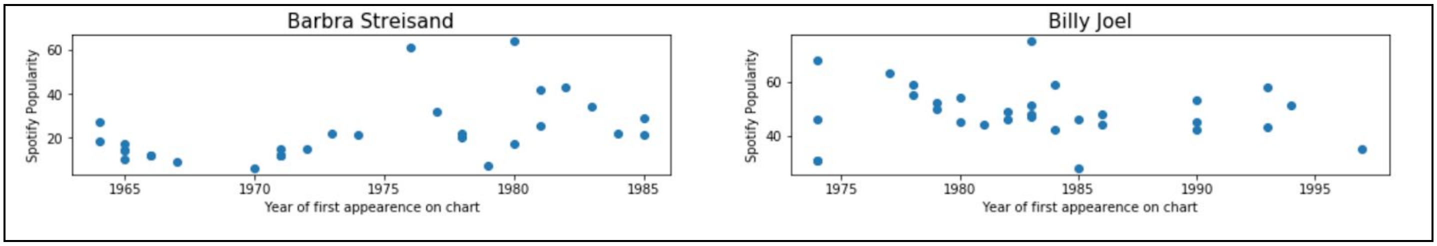
We observe that songs in the first half of
their lifetime have higher Spotif
popularity
than songs in the latter half of the
lifetime.



Artist	count
Glee Cast	142
Taylor Swift	61
Drake	54
Elton John	46
The Rolling Stones	43
Madonna	43
Stevie Wonder	42
The Beach Boys	38
Tim McGraw	36
Aretha Franklin	36
Bee Gees	35
Chicago	35
Michael Jackson	34







Experiment 4: Analysis from WCBS Dataset

Data Collection

We used data-set of WCBS FM radio station’s top 500 songs list which was provided by www.45cat.com starting from year 1973, but unfortunately, we could only find 8 years data. 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. But professor suggested that people usually used to buy vinyl for only A side songs. Hence in our dataset we would only consider A side songs.

Observation from Plots

During our initial analysis, we thought popularity to follow a power law decay but this analysis was based on spotify data. And it is true for old songs because when song appears on chart from that point it would follow power law decay for successive years (usual case) and this holds for spotify because spotify doesn't has popularity data at the time of release of song. This is somewhat overcome by WCBS dataset because in WCBS dataset for songs released around 1970, we have their popularity from when they were release. Hence in this case we have a plot in which popularity is observe as normal distribution for function of years.

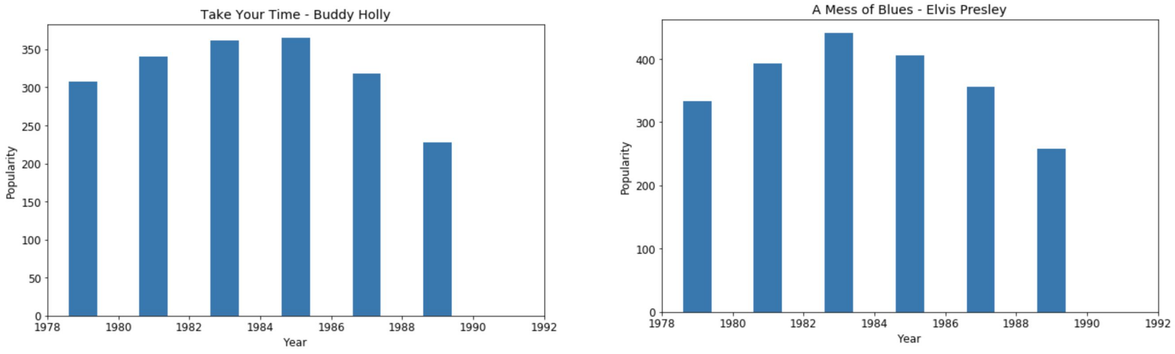
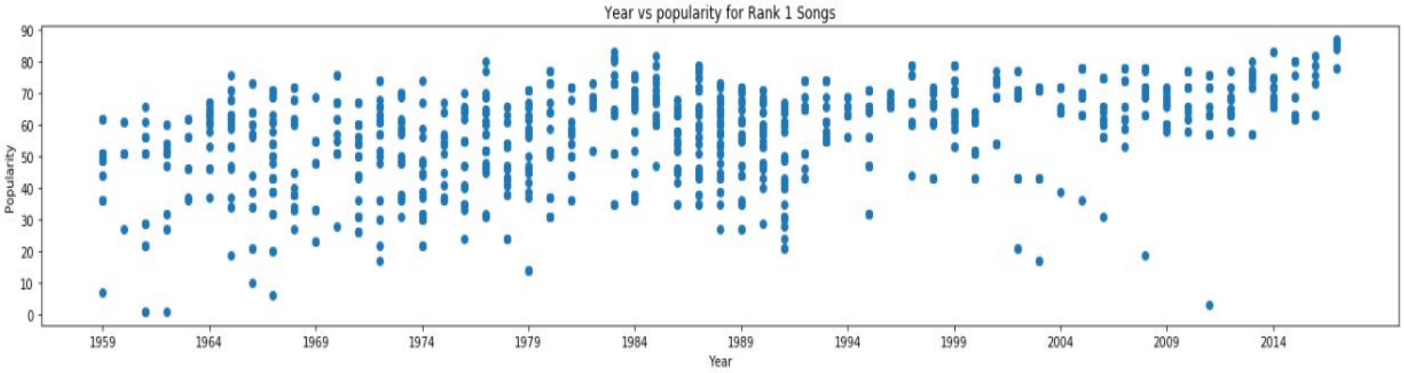
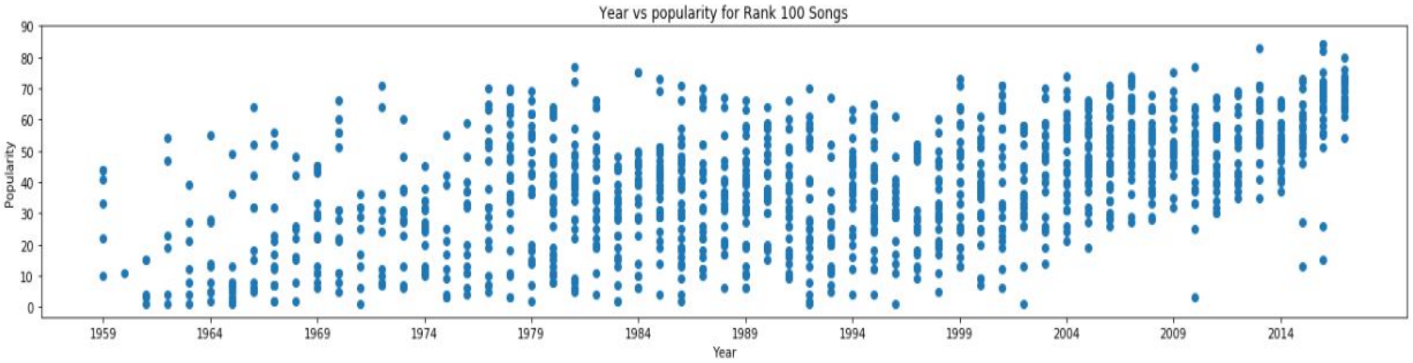


Fig Normal Distribution

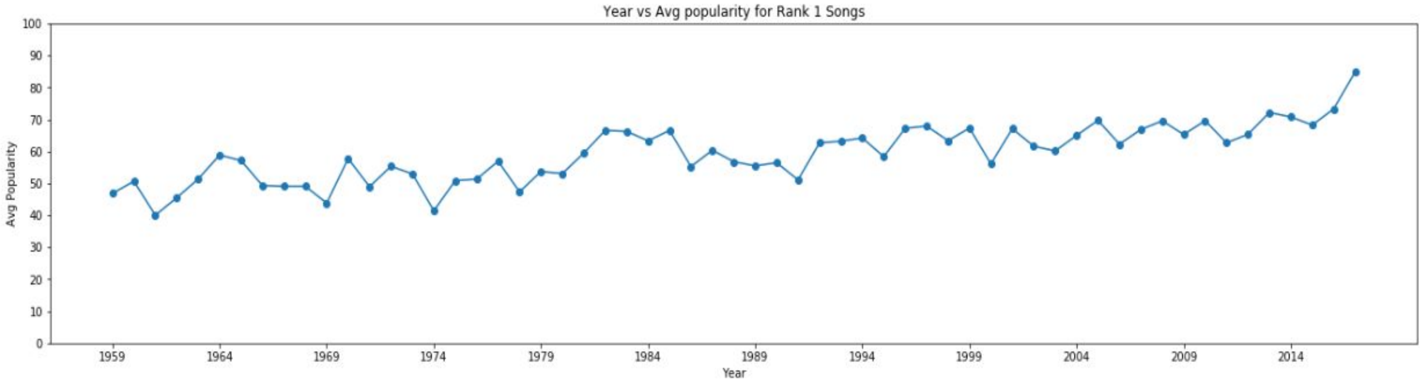
Experiment 5: Originally popular songs are currently more popular than not-so-popular songs?

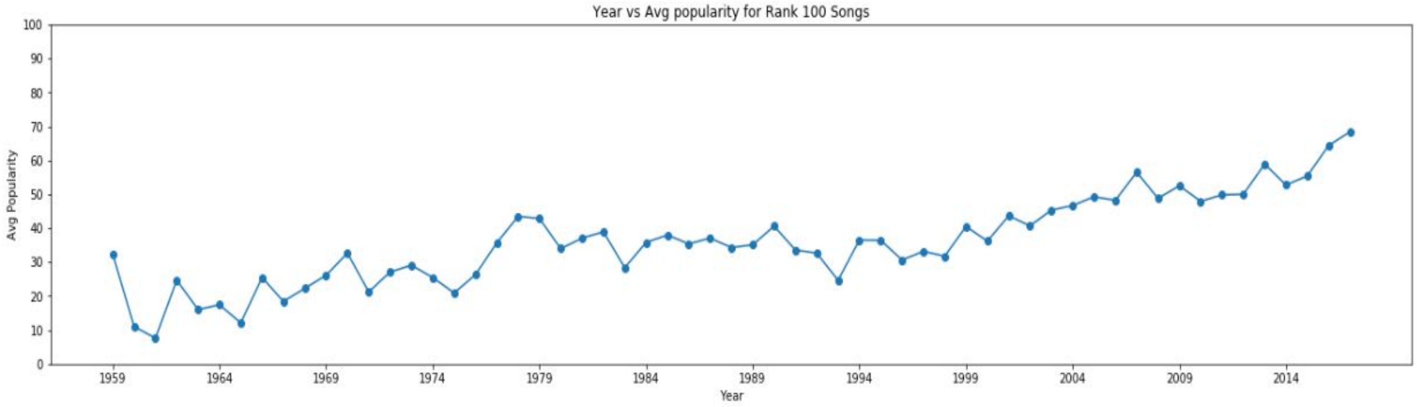
Below two scatter plot shows that, the songs which originally were popular are currently more popular when compared to songs which were not-so-popular. Rank 1 song scatter plot is at the upper end of the plot unlike points at the lower end of the scatter plot for Rank 100 songs. Additionally, recent songs have more Spotify popularity as compared to older songs. Songs which are deviating from expected behavior, for e.g. songs in year 2004-2014 which were expected to have higher current popularity but have less popularity (as from the below plot) are under-performing songs.



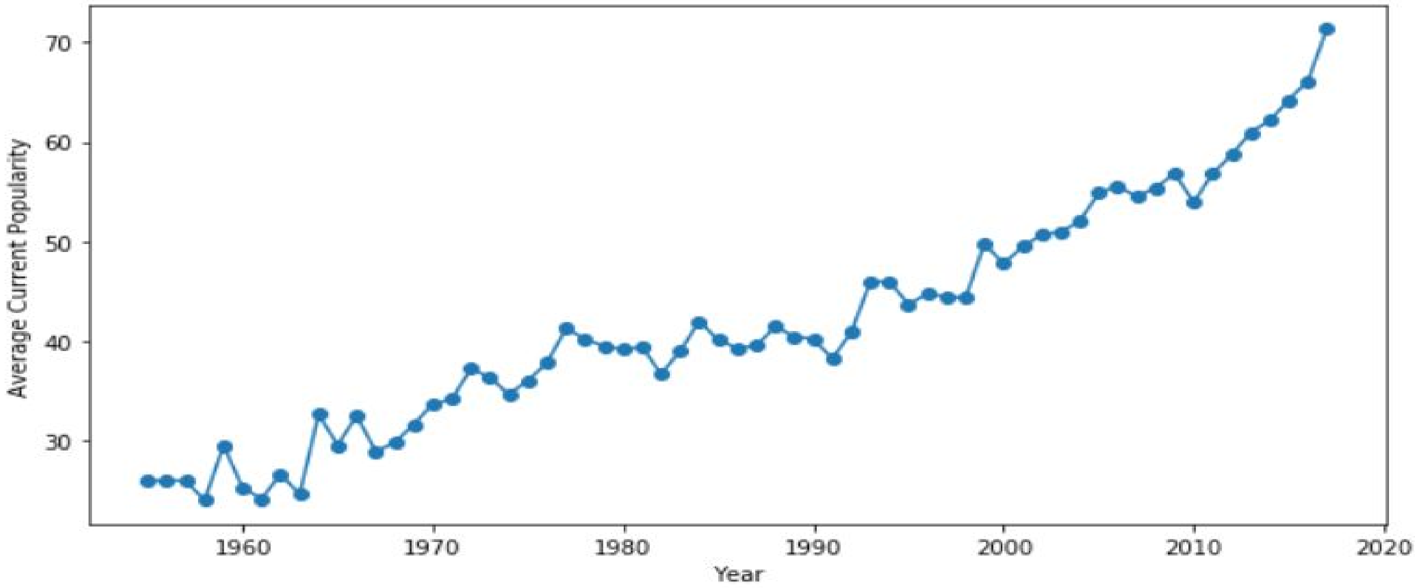


Averaging the current popularity of highest and lowest ranking songs from the Billboard weekly charts for each year, validates the analysis that recent and more popular songs are currently more popular when compared to older and not-so-popular songs.



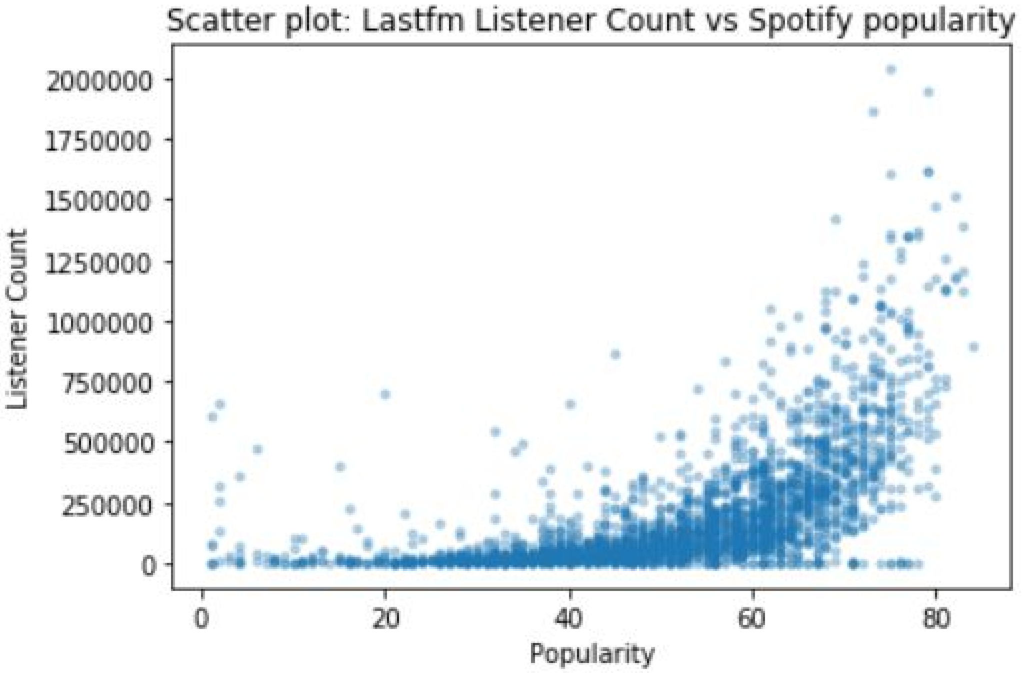


Below line plot depicts that the average popularity of songs year wise is more for with recent years.

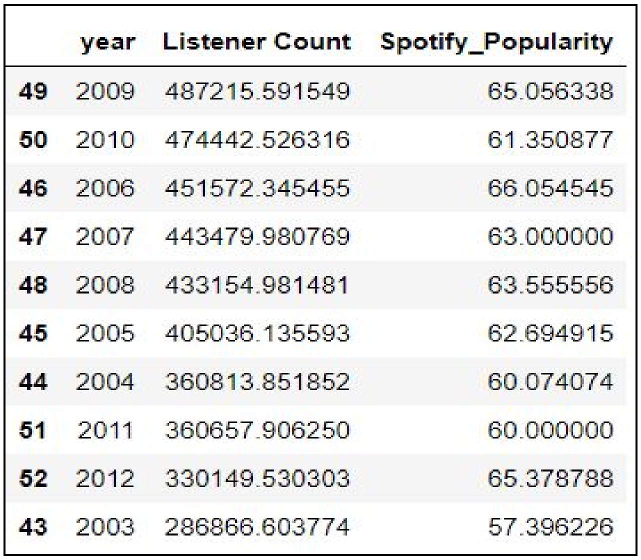


Experiment 6: Analysis on last.fm data

Correlation between *last.fm* ‘s listener count and Spotify Popularity: **0.596**
Correlation between *last.fm’s* play count and Spotify Popularity: **0.511**
The fact that listener count is more correlated than play count with current popularity of song is that, the play count on *last.fm*’s depicts the same song being played by same user multiple times, resulting in misleading numbers. So, for our analysis on *last.fm*’s data we have used listener count of any given track.



On finding the year since 2002, which has maximum number of average listeners to songs released particular year, we find that year 2009 has been most successful year.



	year	Listener Count	Spotify_Popularity
49	2009	487215.591549	65.056338
50	2010	474442.526316	61.350877
46	2006	451572.345455	66.054545
47	2007	443479.980769	63.000000
48	2008	433154.981481	63.555556
45	2005	405036.135593	62.694915
44	2004	360813.851852	60.074074
51	2011	360657.906250	60.000000
52	2012	330149.530303	65.378788
43	2003	286866.603774	57.396226

Experiment 7: Analysing over and under performing songs

Over performing Songs
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.

Underperforming songs

1. *Tragic Incidents*

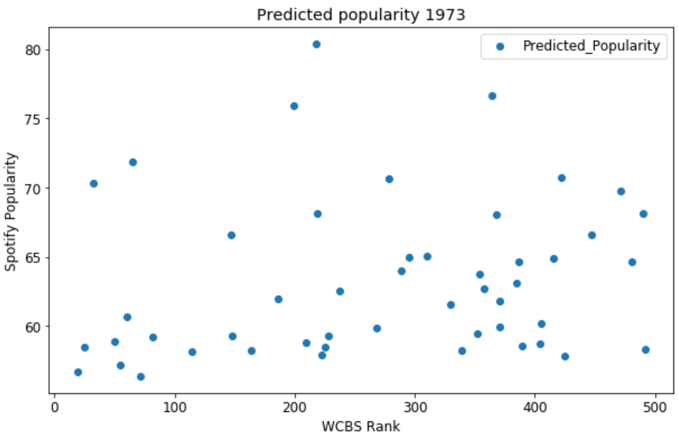
"Die Young" song by Ke$ha, originally released in 2012, debuted rank 13 on Billboard chart at the time of release. But later that year, because of Sandy Hook Elementary School shooting incident, the song was removed from many radio stations and lost around 19 Million listeners resulting in underperformance of song.

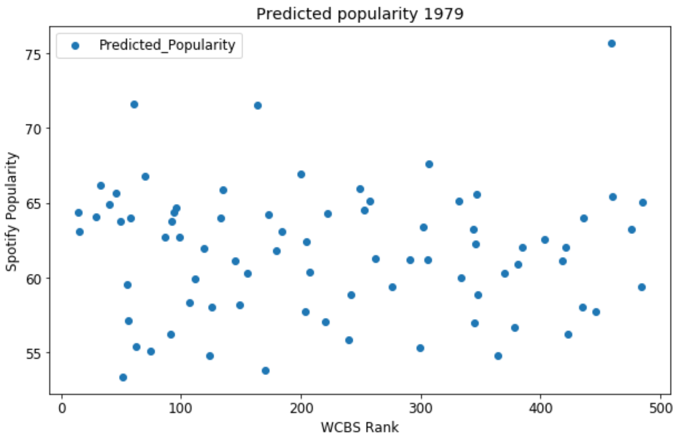
Experiment 8: Model Validation Using WCBS Data-set

To validate our model professor suggested us that we should use WCBS dataset. But there were some problems with this data. Earlier while cleaning WCBS data, we gave same rank to side B of the 7" vinyl. So first of all we needed to clean data again, considering only side A of the vinyl in our dataset.

After cleaning data, we needed to find songs that were present in our WCBS data set and should also be present in yearly billboard and spotify data-set. Hence we combined dataset based on song and artist. Artist was used because there were songs with same name but from different artists. After merging data sets, column names were updated to more meaningful names.

After cleaning data we used the formula generated from our model. In our formula we need to supply values of M and y where M is the initial billboard rank and y is the difference of predicted year and song release year. Having supplied the values we would get expected popularity for that song. We plotted the predicted popularity of a song against WCBS rank, and expected a graph where if rank increases then popularity should decrease. But on the contrary we got following plots:





At first we thought our model was giving inconsistent result but after analyzing closely we found that this was due to following, spotify ranking depends upon various factors such as sales, number of play counts, number of listeners, number of airplays etc. But WCBS depends upon only one factor i.e number of play counts. Hence the ranking supplied by WCBS won't be consistent with spotify dataset. Therefore our model which is created using spotify data-set would not perform as expected on WCBS dataset.

Conclusion

As seen from the plot "Year vs Avg popularity of all songs", we conclude that the average popularity of songs is higher for the recently released songs. However, popular older songs have a good average popularity (around 30 on a scale of 100) as well. Their popularity does not drop drastically over the years. Thus, most popular songs do endure over time.

We further analyze the reasons of over performance and underperformance of songs over time. We conclude that songs early in music group careers endure longer. We also present a decent model for prediction of the current popularity of an old song. It's difficult to get longitudinal data for songs over the years. We hope to present a better model if we are able to aggregate such data for a longer number of years.