What Makes A Best Selling Single?

Exeter Mathematics School

Introduction

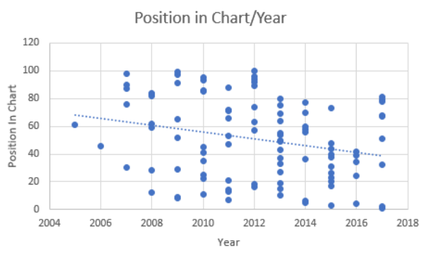
Ever since I was a young child I have had a love for music and this has inspired my project. I see a lot online about 'underrated' musicians and I wonder, well why aren't they famous? Why isn't their music best-selling like Taylor Swift and Ed Sheeran? Throughout this project I will aim to statistically analyse the top one hundred best selling singles of all time. From this, I hope to obtain a conclusion that is a list of factors that could increase the potential success of a song. My objectives are:

- To study the top selling songs and artists of the last century, to see if there is any common link between them
- To see if any lyrics or words are repeated in a number of these songs
- To see of the tempo (speed) of a song has any effect on its selling figures
- To see if the length of a song has any impact on its selling figures
- To analyse the dates of the songs, to see if population growth or another factor has caused songs produced more recently to have sold more copies
- To see if the release of apps such as Spotify has caused selling numbers to decrease, or if copies streamed is now part of the total figure
- To see if collaborations between two or more artists in effect means a better selling single
- To have a result that indicates what is involved in the composition of a best-selling hit

The Data and Initial Analysis

To find a definitive list of the top 100 best-selling singles, I chose to visit wikipedia[1]. However, I discovered that there are two types of singles - physical singles are those sold in Compact Disc (CD) or vinyl form, whereas digital singles are those sold online. I decided to look into digital singles, as I was more familiar with the data set and looking into the Spotify angle would be easier.

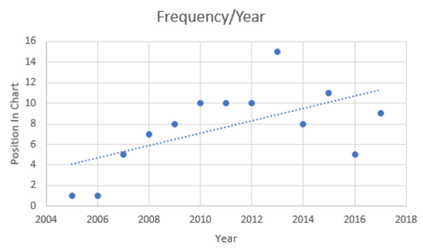
Firstly, I started analysing the songs by looking at the year in which they were released. To start with, I plotted the place in the chart (1-100) against the year they were released, as below.



I used Spearman's rank to find out the correlation between the two variables plotted in the graph.A value of 0 would mean no correlation, whereas a value of 1 would mean perfectly correlated. The formula for this is below:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

I only achieved a value of 0.27 - suggesting that there is a slight positive correlation indicating that a song charts higher overall if sold more recently. However, the small size of this number implies that this correlation isn't significant: year released doesn't necessarily impact a song's place in the top 100. I then decided to look at this from a different angle - I plotted the year released against the frequency of songs in the top 100.



The Spearman's rank for this graph was strong: 0.63. This indicates that the more recent the song, the more successful the song.

I then decided to analyse the artists that had sung these songs. To do this, I created a scoring system in which the person with the top place in the chart, Ed Sheeran, received 100 points, and the person in 100th place, both Flo Rida and Sia, received 1 point. I repeated this process three times - once with all of the songs selling over 10 million copies (the top 47), then the songs selling over 8 million (the top 77) and finally the complete data set.

Top 47:

- Ed Sheeran (89)
- Bruno Mars (73)
- Pharrell Williams (71)
- The Chainsmokers (66)
- Maroon 5 (65)

Top 77:

- Bruno Mars (190)
- The Chainsmokers (156)
- Pharrell Williams (154)
- Ed Sheeran (149)
- Maroon 5 (145)

Top 100:

- Bruno Mars (314)
- Rihanna (229)
- The Chainsmokers (225)
- Pharrell Williams (224)
- Maroon 5 (223)
- Ed Sheeran (217)

These are the top best selling artists - therefore songs released by these artists will be more successful. This is clear because of their consistency in the top 5 each time, bar Ed Sheeran's drop to 6th and Rihanna's climb to second when the data set expanded.

1 Frequency of Words

I used Python, with the help of two fellow students, Catherine Cronin and Oisin Wellesley-Miller, to create a code which took the lyrics from all of the songs in the data set (which I had already compiled from www.lyrster.com[2], some japanese ones being taken from another site[3]) and determined which words were of the highest frequency.

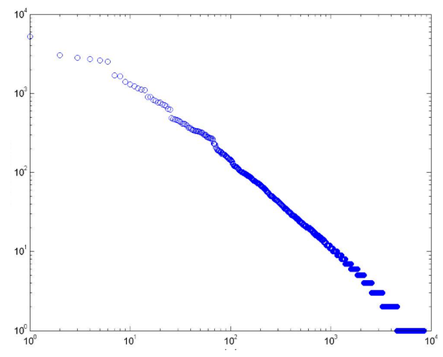
Word	Frequency	Word	Frequency
you	1731	oh	792
I	1603	to	777
the	1240	a	754
me	857	my	690
and	836	I'm	577

Although these seem like the normal most common words, which isn't a surprising result, it is worth noting that singing about someone else seems very popular - or the relationship between "you" and "I". This indicates that maybe love songs are the most popular. The whole collection of words that I compiled seemed to follow an exponential pattern, with the most words having lower frequencies and very few words having frequencies over 500. Looking further into this, I found that there is a law, Zipf's law, which states a similar conclusion but for the English Language.

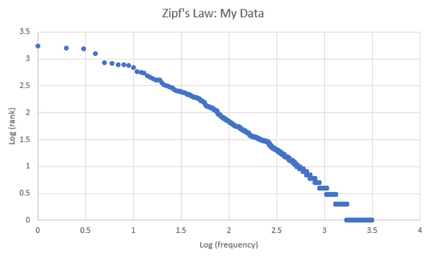
Wikipedia defines Zipf's law as stating that:

"given a large sample of words used, the frequency of any word is inversely proportional to its rank in the frequency table"

A graph of Zipf's law ranks the data and plots the logarithms of these ranks against the logarithms of their frequencies, like this:



My graph, which I plotted using the entirety of my song lyric data set, looks like this:



My data set clearly obeys Zipf's Law therefore - however to confirm this I worked out the equations for the lines of best fit for both graphs. For the whole English language, the equation was y = -1.116x + 7.9015, and for my data set it was y = -1.3318x + 4.5378. The difference in y-intercepts is because obviously words occur more in the English language than they do in song lyrics. But, the similar gradients confirm that song lyrics do follow a Zipf's Law pattern.

Therefore, when choosing which lyrics to use for a song:

- Love songs are more popular - especially in the second person, using the pronoun "you"
- Other than that, use the same words we use everyday in similar proportions because my graph obeys Zipf's Law with these words

2 Tempo

To investigate the tempo, I created a contingency table, having gotten the data online.[4]

Tempo (bpm)/ Place in Chart	Songs 1-20	Songs 21-40	Songs 41-60	Songs 61-80	Songs 81-100	Totals
$70 \leq x < 90$	2	1	1	2	1	7
$90 \leq x < 110$	5	6	8	7	7	33
$110 \leq x < 130$	9	7	5	7	11	39
$130 \leq x < 150$	2	2	2	3	1	10
$150 \leq x < 190$	2	3	3	1	2	11

I realised the best way to analyse this table was to use a Chi-squared test. However, in order to do this I needed to merge some rows together.

Tempo (bpm)/ Place in Chart	Songs 1-20	Songs 21-40	Songs 41-60	Songs 61-80	Songs 81-100	Totals
$70 \leq x < 110$	7	7	9	9	8	40
$110 \leq x < 190$	13	12	10	11	14	60

The formula for a chi squared statistic is:

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

This is where O is the observed value (in this case for example, it would be the numbers in the table) and E is the expected value (if there was no correlation between the two variables, all of the values in the table would be 10). My results give me a chi-squared statistic of 1.0467. With a 0.05 significant level, I obtain a p-value of 0.902642 which means the two variables have not got a significant correlation. In context, this means that as the songs become lower in the chart, this is not because they have an abnormal tempo. Basically, tempo has no impact on a song's place in the chart, although a large proportion of the songs have a tempo between 90 and 130 bpm.

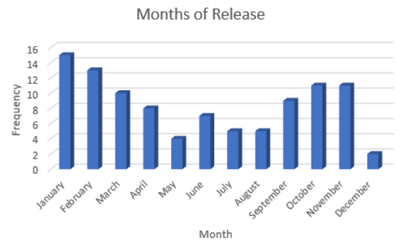
3 Length of Songs

Next, I decided to analyse the length of the songs - their duration in seconds. I didn't think these would be that varied in nature, so I calculated the average of this. But, after quickly scanning the data set of times, I spotted a huge anomaly in Greeen's "Kiseki", which was eight minutes and thirty-one seconds - the second longest song was much less at four minutes and fifty-eight seconds, so I decided to remove this song and create my average from the remaining ninety-nine. This was 229.6364 seconds, or approximately 3 minutes and 50 seconds. The fact that there was only one anomaly in the whole data set indicates to me that songs with an abnormal length do not chart high.

4 Dates Released

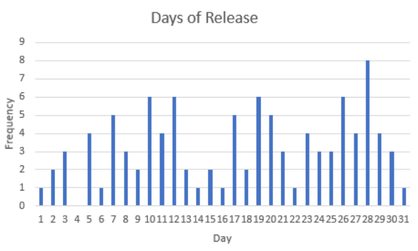
If a song was released in the winter, or summer, does it sell more than one sold in spring? I decided to analyse the dates all the songs were released in order to see if a pattern emerged.

I first looked at the months in which they were released, rather than the individual days. I ended up with this graph:



What this graph shows is that songs which are released in the earlier months of the year, and the first three months of the academic year, sell better than songs sold during the summer months and December. I think the reason that the last month is so low is due to it being the month of Christmas - people are more likely to listen to traditional songs on that subject.

I then analysed the days of the month in which the songs were released, to see if songs sold later in the month sold more than ones released earlier on. I didn't really expect to find a huge correlation here.



Surprisingly, there are days which could contradict my previous statement. The 28th, for example, seems to be a popular day, whilst the 4th has no songs at all. The later days in a month, bar the 31st (probably because there aren't thirty-one days in every month) seem to produce more best selling songs as a whole than earlier days and days that fall halfway between the two.

Overall, I can conclude from this that releasing a single on the 28th January will sell more copies than a song released on December 4th - so 'date released' does have an effect on how popular a song becomes.

5 Collaborations

Out of all 100 songs, 62 are from a single artist, 30 are from collaborations between two artists, 6 are from collaborations between 3 artists and 2 are from collaborations between 4 artists. No songs have any more people than this working together on a single.

Clearly, there is definitely a correlation between the number of artists a song is recorded by and its success in the charts. In order to produce a single that sells lots of copies, either working independently or with only one other is the best option.

6 One-Hit Wonders

I also calculated the number of one-hit wonders in this list - songs by artists that have no other successful singles. Only 9 of these were in this category. This could indicate that songs produced by artists who already have top singles, or will in the future, have a better chance of being in the top 100.

7 Spotify and Music Applications

A digital single is one that has been digitally downloaded - from applications such as iTunes, Spotify and Amazon Music, amongst others. This means that the songs in the data set I have been studying are bought online, rather than in CD or vinyl form. As this was only introduced in the early 2000s, it makes sense that most of the songs in the data set are from 2005. The easy access to songs both at home and on the go - carrying songs downloaded to a phone or another device is practical, and impossible to do with a CD player. This accounts for the vast rise in selling figures after these applications were introduced.

8 Conclusions

From this project, I believe that I have found a number of factors that increase the success of a song. These factors are:

- Time - songs sold more recently
- Artist - singers such as Bruno Mars and Ed Sheeran are more successful
- Genre - love songs, or songs about another person, are more popular
- Tempo - speeds between 90 and 130 bpm are popular
- Length - songs close to 3 minutes and 50 seconds in length are better selling
- Date - songs sold in the late periods of months earlier on in the year are better
- Collaborations - either working independently or with a single partner is popular
- One-Hit Wonders - having a hit single already/one coming up in the future is preferable
- Release - songs released digitally sell better than songs released physically

9 References

- 1. https://en.wikipedia.org/wiki/List_of_best-selling_singles#Best-selling_digital_singles
- 2. http://www.lyrster.com/
- 3. http:/www.animelyrics.com/
- 4. http://www.songbpm.com/