Blueprints for a hit song

Darian Tomašević

Mentor: prof. dr. Aleksandar Jurišić

18th of February, 2018

1. Introduction

Today's music industry consists of companies and individuals that earn a substantial amount of money for creating and performing new songs. Imagine the effect on such an industry if we could make a blueprint for a hit song. If song generation could be automated at least to some degree, it would be a game changer. The idea might seem far-fetched, but with enough data and the use of statistical analysis anything is possible.

Considering that the early texts of music theory date back to the Mesopotamians (1500 BCE) we can safely assume that debates about song writing have been going on for generations. Even the topic of music information retrieval is not new, as such APIs have already been built and used in other projects. Most research I have come across focuses more on analysing a given song such as determining a song's emotion [Jamdar, Abraham, Khanna & Dubey, 2015], or identifying a particular song via minimal input, such as a whistling [Arentz, Hetland & Olstad, 2007] and last but not least classifying a song's genre [Neumayer & Rauber, 2007]. Seeing the past research, I have decided to concentrate on finding trends in popular songs by analysing the pure numbers, which represent various properties of these songs.

The aim of this project is to identify a few common attributes of hit songs and see if and how they have changed during the past decades. This will help artists, by giving them a few tips regarding song making, which might increase their chances of writing and recording a number one hit.

After the Introduction, we will go through the data collection process in Section 2. Having familiarised ourselves with the utilised APIs, we will continue with the analysis and dissection of the gathered data in Section 3, such as the song's key (Section 3.1), tempo (Section 3.2), duration (Section 3.3) and loudness (Section 3.4). In the ensuing Section 4 we will quickly summarize our results and end with some closing thoughts.

2. Data collection

As our resource for measuring success we will use the Billboard Hot 100 charts from 1950 onward [2] and my first plan was to extract information about all the songs using the Billboard's API. However, this was easier said than done, as the site's API is rather slow and unreliable, which is understandable since it is still unofficial. Thus we will focus on a smaller set of songs, as we set out to find every number one song in the past 67 years.

Having compiled the names, we turn to the Echo Nest API, which is now integrated into a streaming service called Spotify [9]. Luckily the Echo Nest API is easier to work with than the Billboard's and we are able to obtain a myriad of fascinating data from it, such as the song's key, tempo, and even some more abstract information. More details regarding all of the measurements will be explained later in the report.

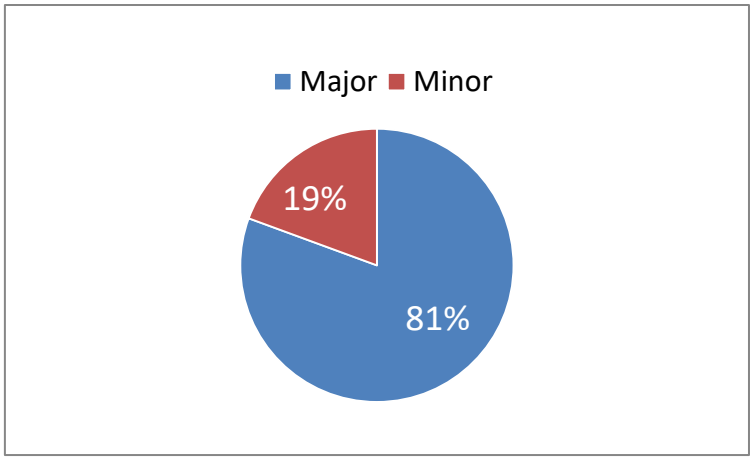
3. Data analysis

After arranging the gathered data into a neat spreadsheet, I started the search for interesting attributes that would be suitable for the project. Without success, I tried to figure out how abstract information such as danceability was computed, hence I chose to rather focus on the quantitative properties that seemed most significant to the song's identity, such as the key of the song, it's tempo, duration and loudness. Even though most of them are self explanatory, I will still go include a theoretical summary of specific musical terms.

3.1. The Eternal Question: Major or Minor

The main way to distinguish major keys from minor is to check whether a half step difference occurs between the 3rd and 4th or between the 2nd and 3rd note of the scale. This greatly alters the mood of the song, for example music based on major keys is considered "happy" while the minor keys are seen as "melancholy".

Taking into account what we just learned in combination with our results we see that about 80% of the number one songs were written in a major key, meaning that most of them are "happy" sounding. The aforementioned results are depicted in the following graph, see Graph 1.



Graph 1: Pie chart of songs written in major or minor key

To figure out if our sample size is "big enough" we can use the about to be mentioned formula with our data:

$$p = 0.19, \qquad q = 1 - p, \qquad n = 67, \qquad \alpha = 0.05,$$

$$p \pm z_{\alpha/2}\sqrt{\frac{pq}{n}} = 0.19 \pm 0.001$$

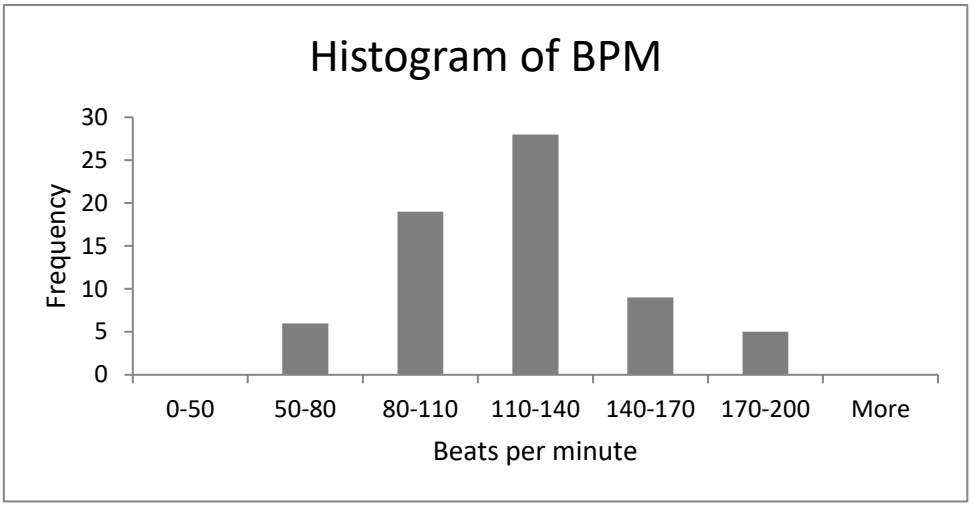
When interpreting said numbers we follow the rule of thumb [7, page 300] that the size of a sample is acceptable when $np \geq 4$ and $nq \geq 4$. The results might seem remarkable as we could conclude, with 95% probability, that 19% of all popular songs are in a minor key. However, the main problem is that we had not chosen our sample completely randomly, so we cannot generalise our results.

3.2. Not quite my tempo

Tempo, in the music world, is the pace or speed of a given song. It is typically measured in beats per minute (BPM). As an example 60 BPM signifies one beat per second.

To see if there exists a particular tempo that is preferred in popular music, we check how the tempos of our number one songs are distributed. Due to the sample size being larger than 30, we can assume that the distribution of tempos will bear resemblance to a bell curve, due to the central limit theorem.

For the sake of easier data representation we divide the songs into groups with a range of 30 beats per minute, as seen in Graph 2. Despite analysing only 67 songs, we can already notice that the histogram begins to resemble a normal distribution.



Graph 2: Distribution of tempo values

We can calculate the average tempo and the standard deviation:

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i \ , \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$\bar{x} = 119.94 \text{ BPM}, \qquad s = 27.31 \text{ BPM}$$

The average of around 120, which is considered standard in popular music, has many explanations such as its similarities to our walking pace (likewise used in military marches). Some theories also mention an evolutionary explanation, which states that at some point in time a songwriter had success with that specific tempo and other artists started to copy it to the point where even the default setting for today's audio recording software is set to 120. There exist many more theories, but one other particularly intriguing thought is that 120 can neatly be divided with our time measurements into 2 beats per second.

Next, we will use

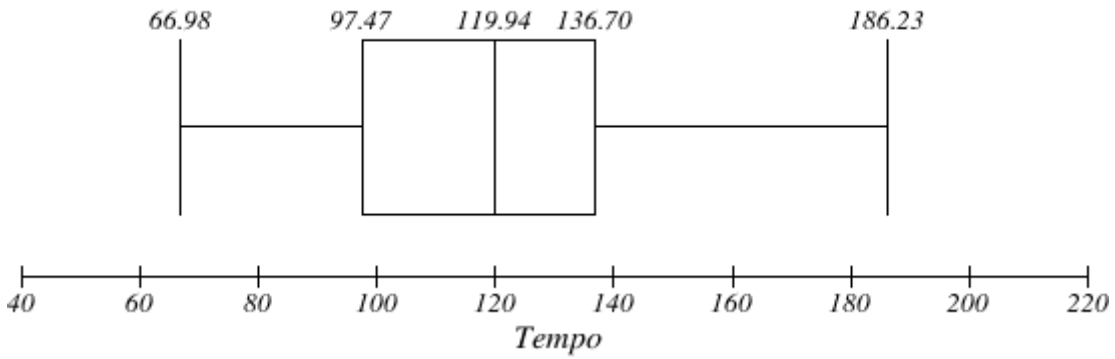
$$P\left(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

to calculate the 95% confidence interval for our mean:

$$P(114.451 \leq \mu \leq 125.428) = 0.95$$

The results allow us to decide, with 95% probability, on an area in which the average tempo of the population will lie in.

We can also check how our data is represented by a box plot, see Graph 3.



Graph 3: Box plot for tempo values

With a quick glance at the box we see the maximum at 186.23 BPM and the minimum at 66.98 BPM, the first quartile at 97.47 BPM and the third at 136.7 BPM, while the mean is at 119.94 BPM.

Furthermore we can test a few hypotheses regarding our mean. Let us say our null hypothesis states that the mean is 120 while the alternative is that the mean is not 120. Using the following formulas we can accept one of the previously mentioned hypotheses.

$$\mu_0 = 120, \qquad H_0: \mu = \mu_0, \qquad H_1: \mu \neq \mu_0,$$
$$\alpha = 0.05,$$
$$\mathrm{TS} = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n},$$
$$\mathrm{TS} = -0.01781, \qquad t = \pm 1.96, \qquad P = 0.985,$$
$$K_\alpha = (-\infty, -1.96) \cup (1.96, \infty)$$

Interpreting the results, we see that the critical values are ± 1.96 and that the test statistic is -0.0178. Therefore, in this classical frequency framework, we cannot reject the alternative hypothesis H_1, but we also cannot reject the null hypothesis H_0 because the P-value is larger than α and furthermore because the test statistic does not fall into the rejection region K_α.

3.3. Pressed for time

Viewing the durations of all the songs we can suspect there to be a slight increase in duration over the years, thus our goal is to find the correlation between song duration and the year that song charted in.

The Pearson correlation coefficient is a measure of linear correlation between two variables, X and Y. Its value lies between -1 and 1, where the two values represent a total linear relationship, negative and positive respectively, while 0 expresses no linear correlation.

Next, we will calculate the average song duration and the before mentioned coefficient.

$$\bar{x} = 227 \text{ seconds}, \qquad \bar{y} = 1984 \text{ years},$$
$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{k_{X,Y}}{s_X s_Y}$$

By inputting our data into the formula, we obtain the following correlation coefficient

$$r_{X,Y} = 0.338\,.$$

To understand the obtained value, we follow the rule of thumb [3] which states that our coefficient describes a low correlation, as it lies close to the negligible correlation range. Despite the description, we still see an increase in song duration from around 200 seconds in the 1950s to around 250 seconds in the 2010s. This means that songs today are almost a whole minute longer, which is substantial with regard to the length of the whole song.

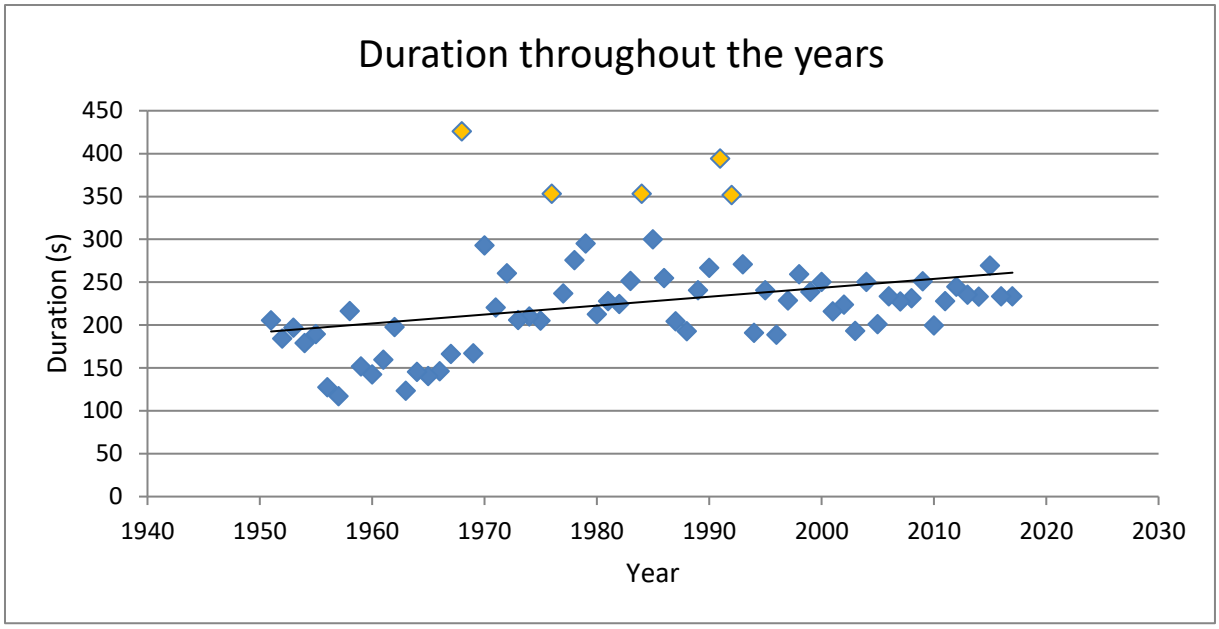
There exists a logical explanations for the phenomena: music was first recorded on phonographs and later on disks which both could hold about 3 minutes of information, but after the disks were replaced by CD's, which could store more data, song duration was not considered a limit anymore.

Even though the Pearson correlation coefficient is rather low, we can still calculate the regression line with the ensuing formulas, where Y represents song duration and X the year it charted in:

$$Y = \alpha + \beta X, \qquad a = y - \beta x, \qquad b = \frac{k_{X,Y}}{s_X^2}, \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$
$$s_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}, \qquad k_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

By inserting our data into the formulas we acquire the regression line Y, which is also depicted in the following graph, see Graph 4.

$$Y = -1835 + 1.039X$$



Graph 4: Graph of song durations throughout the years with the corresponding linear regression line. The orange squares in our case represent the outliers.

3.4. It might get loud

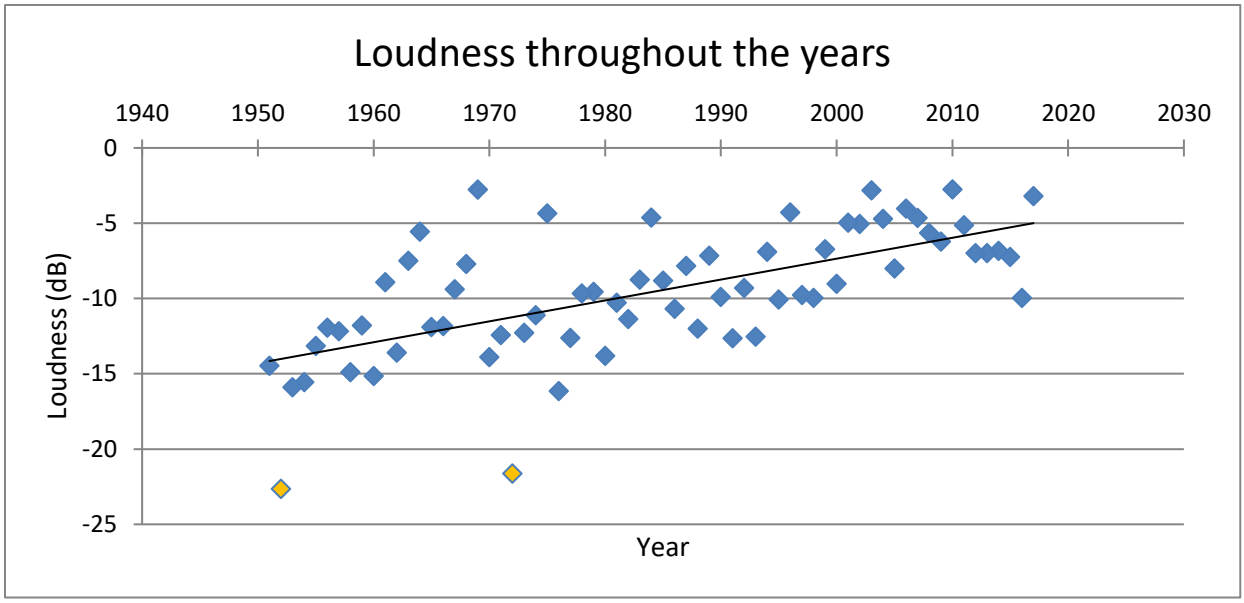
From the Echo Nest API we were able to obtain a Loudness value, measured in decibels (dB). Despite its name, this does not represent the actual volume, which you can adjust with a volume knob. Instead, it's a measure of loudness between the loudest and the quietest part of a song. Throughout the years the quietest parts have become louder whilst the loudest parts have gotten closer to the maximum, that being 0 dB. Note that when recording the decibels are measured between -60 and 0.

After gathering the data it was obvious that the loudness increased tremendously over the years. Once again we check the Pearson correlation coefficient between the song's Loudness and the Year it charted in, utilizing the before mentioned formulas.

$$r_{X,Y} = 0.641$$

Based on the result and the rule mentioned in the previous section we see that the correlation between our variables, is not the strongest, but still represents a moderate uphill linear relationship. Our next step is to calculate the regression line Y with the formulas used previously and make a graph depicting the line.

$$Y = 0.138X - 284.8$$



Graph 5: Graph of song loudness throughout the years with the corresponding linear regression line. The orange squares represent outliers with high leverage.

We can also test the null hypothesis H_0 that the variables are not linearly dependant. The formula for the test statistic is as follows:

$$\mathrm{TS} = \frac{r\sqrt{n-1}}{\sqrt{1-r^2}} \sim \mathrm{Student}(n-2)$$

If we insert $n = 67$ and the already calculated $r_{X,Y} = 0.641$ into the formula we reach the about to be mentioned results, where TS is the test statistic, P is the P-value and the Student distribution defines the rejection region K_α.

$$H_0\colon r_{X,Y} = 0,$$
$$\alpha = 0.05,$$
$$\mathrm{TS} = 6.731, \qquad P = 5.17 * 10^{-9}, \qquad t = \pm 2.00,$$
$$K_\alpha = (-\infty, -2.00) \cup (2.00, \infty)$$

Because the P-value is smaller than our α we can reject the null hypothesis H_0 that states the variables are not linearly correlated. This means that even though the correlation coefficient is neither 1 nor -1 it is still statistically important and that there is indeed a linear dependency between the two variables.

4. Conclusion

Figuring out a blueprint for song making is a rather difficult task, since emotions play such a fundamental part in music, thus it cannot just be expressed with numbers. Even though, we have figured out a few general tips to increase one's chances of having a Billboard hit, such as write in a major key, keep the tempo at about 120 BPM, keep the length of the track around 3 to 4 minutes and last but not least do not be afraid to record as loud as possible.

I was rather pleased with how the project turned out, but if I were to re-do it, I would probably try to gather a bigger data set, for example the top 100 songs of each year, which might prove difficult due to the Billboard API being unreliable. I was also going to include some more abstract attributes such as danceability, but sadly I did not find a good explanation of how the numbers were computed in addition to some of the Echo Nest API's data being rather sketchy.

Who knows, maybe in the future we will be able to make a generator of new hit songs, but then again, should we really strive to achieve such a goal? From a technological standpoint it might seem incredible, but it might also represent the surrendering of one of the few creative processes we have left in today's day and age.

5. Literature and references

[1] W. A. Arentz, M. L. Hetland, B. Olstad. (2007) Retrieving musical information based on rhythm and pitch correlations.

[2] Billboard.com (https://www.billboard.com/charts/hot-100/) (27.12.2017).

[3] D. E. Hinkle, W. Wiersma, S.G. Jurs, Applied Statistics for the Behavioral Sciences (5th edition), Boston: Houghton Mifflin, 2003.

[4] M. Hladnik. Verjetnostni račun in statistika, Zapiski predavanj, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2002.

[5] A. Jamdar, J. Abraham, K. Khanna, R. Dubey (2015). Emotion analysis of songs based on lyrical and audio features.

[6] A. Jurišić. Verjetnostni račun in statistika, zapiski, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2017.

[7] W. Mendenhall, T. Sincich. Statistics for Engineering and the Sciences (5th edition), Prentice-Hall, Inc. Upper Saddle River, USA, 2017.

[8] R. Neumayer, A. Rauber. (2007) Integration of Text and Audio Features for Genre Classification in Music Information Retrieval.

[9] Spotify.com (https://beta.developer.spotify.com/console/get-audio-features-track/) (28.12.2017).

[10] Spotify.com (https://developer.spotify.com/web-api/console/get-search-item/) (28.12.2017).