Combining Data-Science and Music

Using a data-approach to creating a “Popular Song”

Davis Li

March 9th, 2021

1 Executive Summary

Ever since I was six, my parents put me into piano lessons. For the first couple of years, I was pushed into playing and performing classical music. Although the technical aspects of classic pieces intrigued me, I have always had a showman side of me and wanted to play pop songs. Back in elementary school when I played in the music room at lunch, I always attracted an audience playing the newest hits from Justin Bieber, Taylor Swift and Maroon 5. This was when I decided to pivot into playing and arranging pop song pieces. This hobby of mine eventually turned into my first ever part-time job as a wedding pianist, where I played hours and hours of my own arranged pop songs. Through arranging different pop songs, I have noticed trends here and there, but I have always wondered what makes a pop song “popular.”

Fast forward to today, with my limited-yet-functional skills in data science, I want to use data to explore what makes a popular song.

2 Technical Exposition

2.1 Data Pre-Processing

I obtained the data from a Kaggle dataset based on the Spotify API. This dataset included all the Spotify songs (175k+ songs) from 1920 to 2020. Given that the data originated from the Spotify API, there was not much data cleaning that had to be done. All work was done through Python, and you can find the full Jupyter Notebook on my Github.

2.12 Data Attributes Exploration

The attributes of this data set were:

- **Acousticness**: Confidence measure from 0.0 to 1.0 of how acoustic it is
- **Danceability:** Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity, measured from 0.0 to 1.0
- **Duration_ms:** Duration of track in milliseconds
- **Energy:** The perceptual measure of intensity and activity. Energetic tracks feel fast, loud and noisy. Death metal has high energy, while a Bach prelude scores low
- **Instrumentalness:** Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal

content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

- **Key:** The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.
- **Liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides a strong likelihood that the track is live.
- **Loudness:** The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing the relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 dB.
- **Mode:** Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- **Speechiness:** Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- **Tempo:** The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, the tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **Time_signature:** An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
- **Valence:** A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

When looking at these attributes, there is one common problem. For attributes like acousticness, danceability, or instrumentalness, we do not know what goes into these variables. Therefore, it's hard to understand the difference between 2 similar values on the same attribute. The good news is, we can easily work with attributes such as key, mode and tempo.

2.13 Data Time Range

Although there is data dating back to 1920, I was skeptical whether it would be good enough data for us to analyze. From plotting a histogram (Figure 1), we can see that there are significantly fewer data before 1950.

After plotting a time series on acousticness to see how data prior to 1950 looked (Figure 2), we can see that it messed up our data quite a bit. Therefore, time-series data will mostly be conducted from 1950 to 2020.

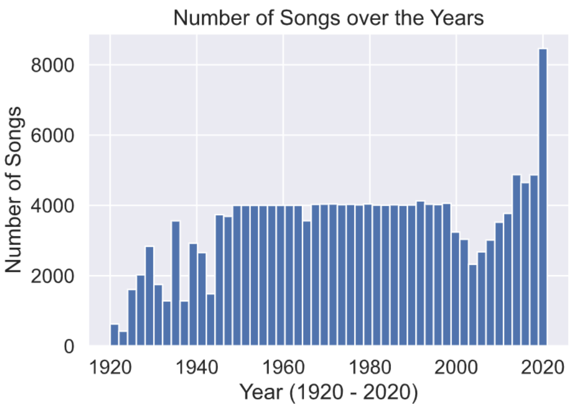


Figure 1: Histogram of Songs

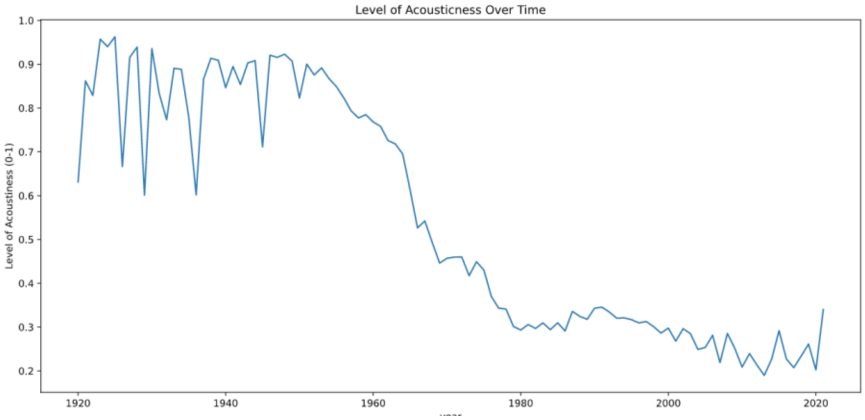


Figure 2: Level of Acousticness Over Time

3 Exploratory Data Analysis

3.1 Trends Over Time:

3.1.2 Popular Song and Artists

Using the data set we have, I want to find the most popular songs as of the moment, and the most popular artists of all time. Given that my Spotify has the Billboard Top 100's constantly on repeat, I recognized the most popular songs immediately. What surprised me was the top artists. I thought it would've featured more recent artists like Travis Scott or Ed Sheeran, but the top artists were actually from the '90s, with Beatles leading and Frank Sinatra coming in a close second. Who knew there were so many Frankie fans! However, I was amused to find a good hoard of Swifties (Taylor Swift fans) there. Fun fact, I still have the sheet music for "You Belong to Me" and "Back to December" that I arranged years ago!

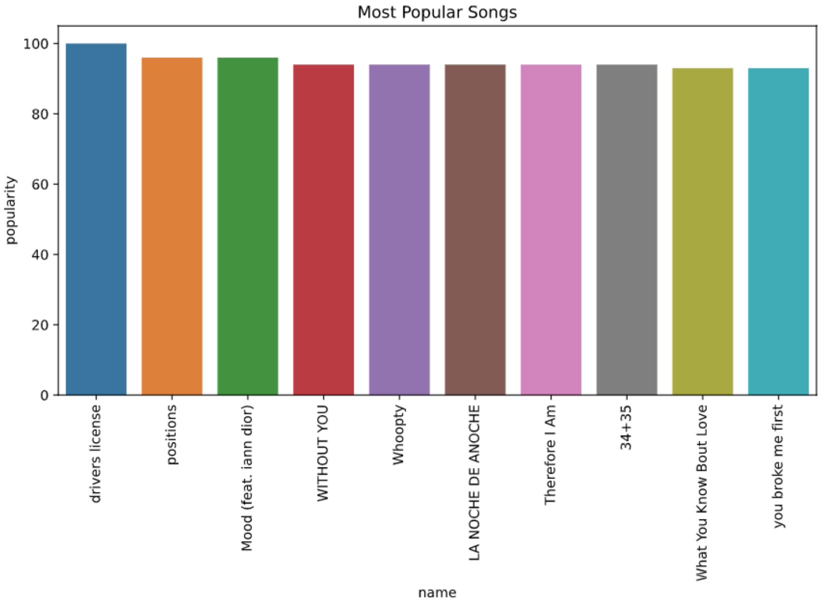


Figure 4: Most Popular Songs

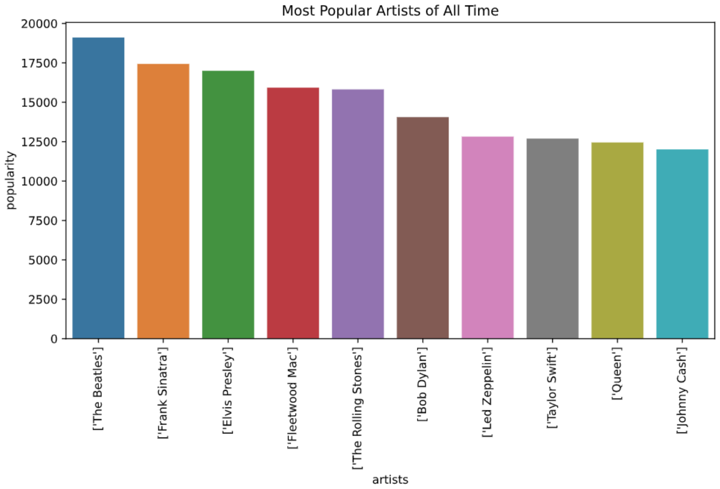


Figure 5: Most Popular Artists of All Time

3.1.3 Attributes Over Time

To first explore the data, I wanted to look at how song styles were changing over time. To describe the style of songs, I chose to look at the attributes of acousticness, danceability, energy, loudness and tempo. There are a couple of conclusions we can get to.

- **Acousticness:** It seems that acousticness has been on a downwards trend since the late 1960s, which could be due to the rise of disco (popularity of disco accelerated from '70s throughout '80s) to replace the once-popular country/classical music.
- **Danceability/Energy/Loudness/Tempo:** It appears that with the fall of acousticness and the rise of disco/EDM/hip hop, newer songs with electronic/computer-generated beats are more energetic, thus having higher danceable, loudness and tempo.

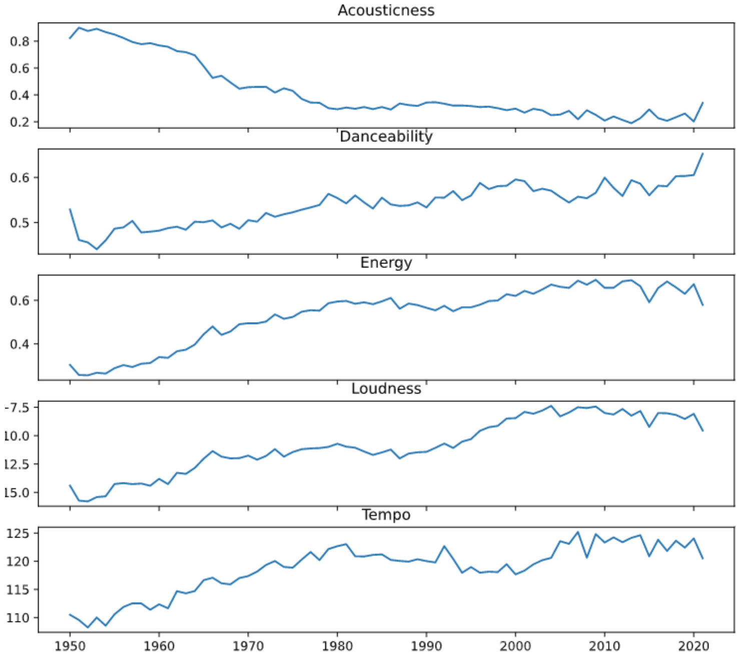


Figure 6: Selective Song Attributes Over Time

I also wanted to look at the duration of songs over time, to see whether things have changed. It seems like songs drastically got longer from 1950 to 1980 to about 230000ms (3.83 minutes). However, since 2010, songs have gotten shorter to about 210000 ms (3.5 minutes).

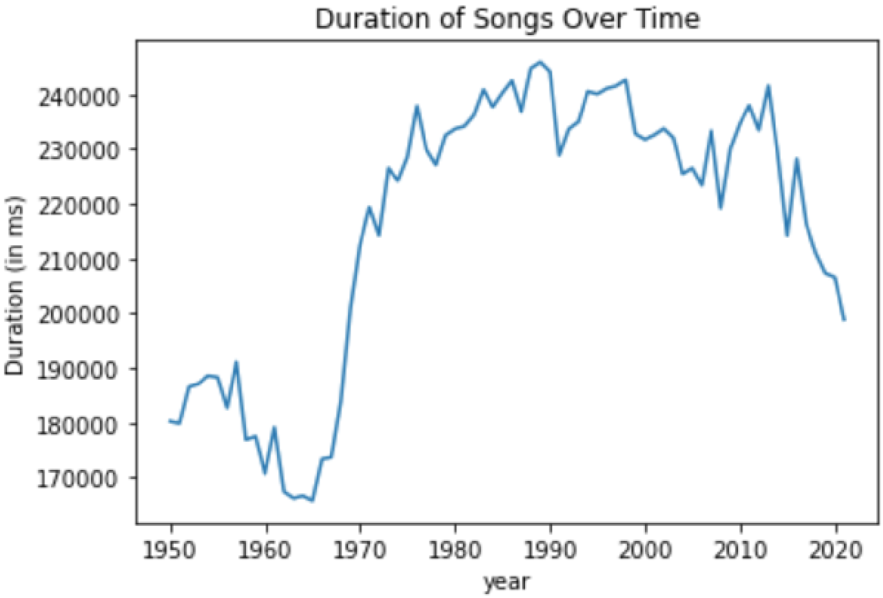


Figure 7: Duration of Songs Over Time

3.2 Attributes vs Popularity

3.2.1 Key & Mode vs Popularity

Let's first take a look at how modes affect song popularity. Before running the analysis, I would've thought that major keys are much more popular than minor keys. I was proven wrong - there seems to be not much difference (almost exactly the same).

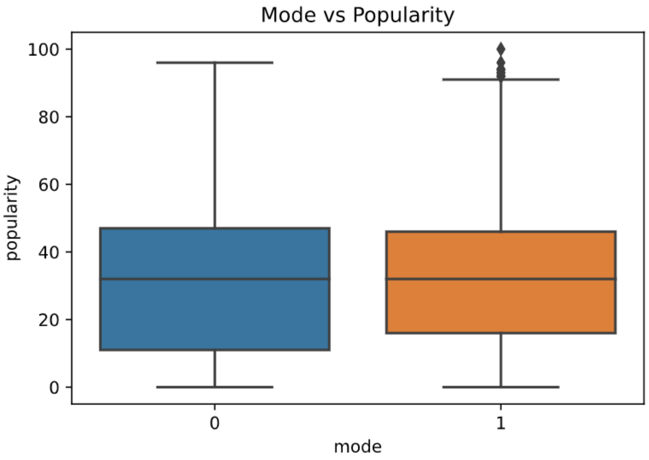


Figure 8: Mode vs Popularity

Just looking at the modes isn't enough, let's take a look at the different keys then come back to the modes. My initial hypothesis was that songs in a white key would be much more popular than songs in a black key. From Figure 9, you can tell that white keys have a higher median of popularity than black keys. Keep in mind that 0 is C, and every +1 is +1 semitone (so 1 is C#/Db, 2 is D.etc). We can gather that Keys E (3), A (9) and D (2) are the most popular, while D# (3) is the least.

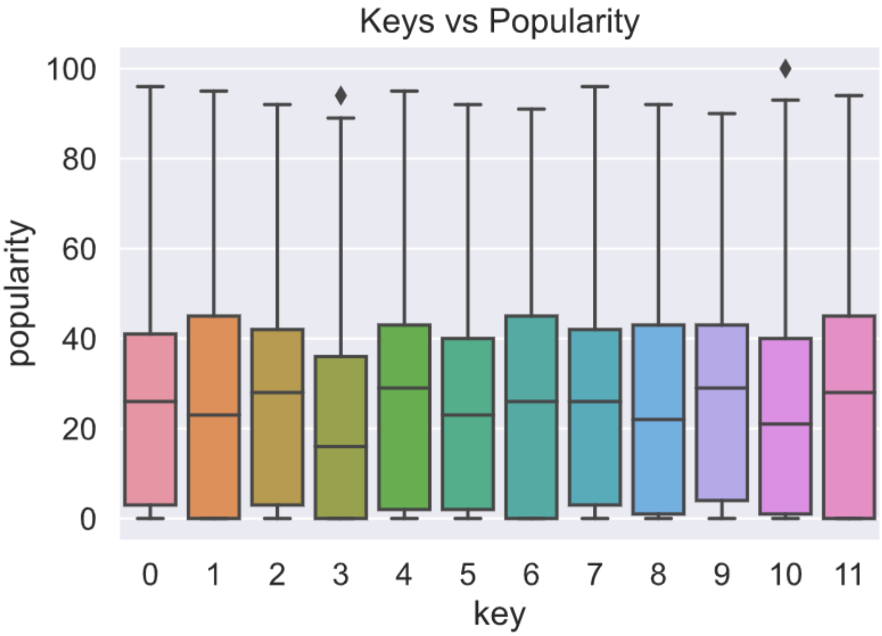


Figure 9: Key vs Popularity

Now, I want to find out the relationship of keys and mode on popularity. I hope that by doing a double plot, I would be able to see a trend that I wasn't able to see before on the singular plots.

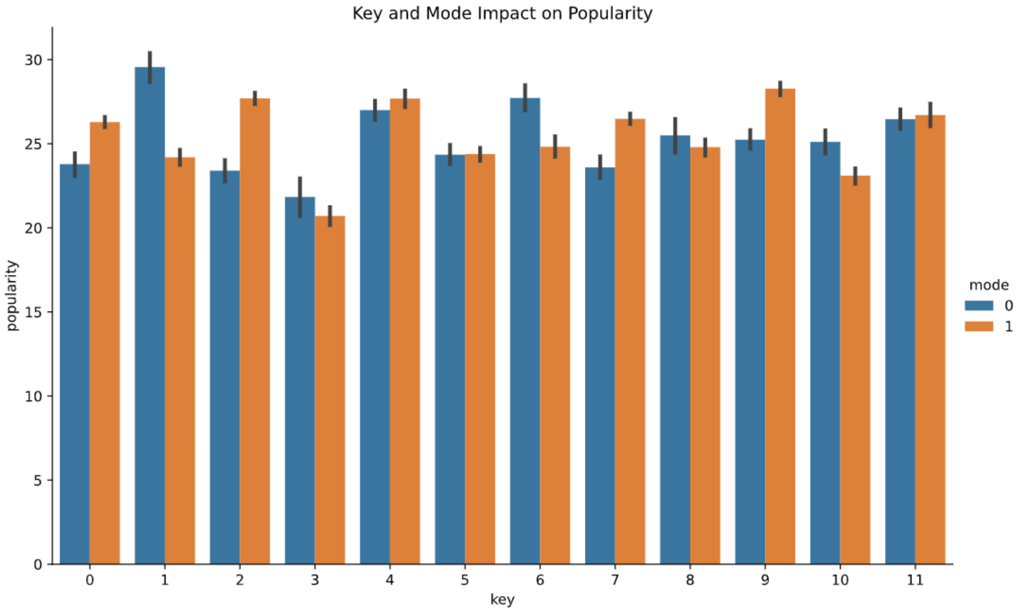


Figure 10: Key and Mode vs Popularity

As we can see, more popular songs seem to be in a major mode when it comes to white keys (C, D, E, A.etc), while minor modes are more popular amongst black keys (D#, Bb.etc).

From this analysis, we can conclude that the most popular songs are written using white major keys.

3.2.2 Energy vs Popularity

I walked into this plot thinking that higher energy led to higher popularity, but I walked out semi disappointed. When I first plotted this, it was just a huge solid box. I realized I had too many data points, so I turned the alpha to 0.02. We can see that increase in energy seems to slightly increase popularity, but there wasn't much of a relationship overall.

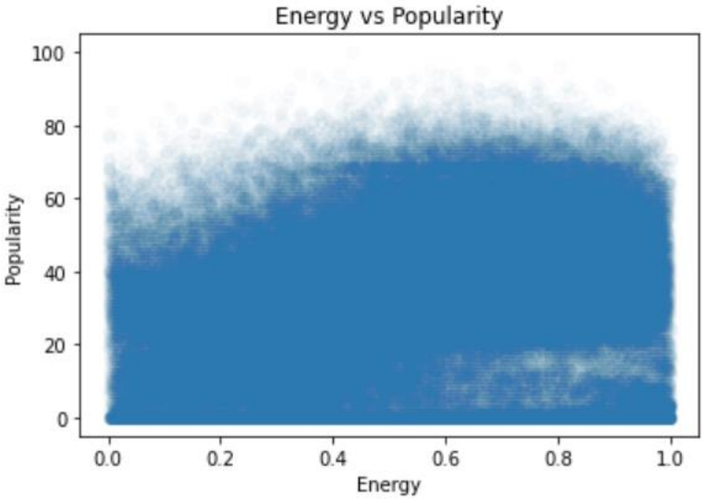


Figure 11: Energy vs Popularity

3.2.3 Loudness vs Popularity

My hypothesis was similar to energy vs popularity where I thought increased loudness led to increased popularity. Expecting to be disappointed, this time I was pleasantly surprised. We can see a relatively linear slope between the relationship of loudness v popularity. So I guess to make a popular song, you have to crank your volume up a couple of notches?

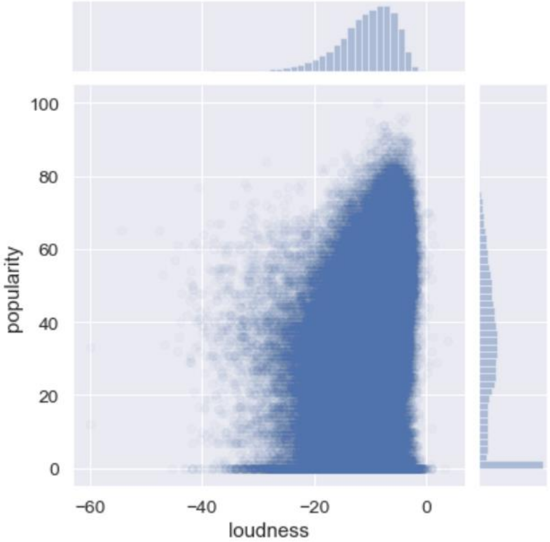


Figure 12: Loudness vs Popularity

3.2.4 Attribute Correlation Heat Map

To view the relationship between the rest of the variables, I plotted a correlation heatmap. Here are some observations:

- As pointed out above on the time series plots, acousticness has a strong negative correlation with energy, loudness, and tempo
- There was a 0.35 correlation between explicitness and speechiness. I'm guessing that the songs with higher speechiness are rap songs, which would be more explicit.
- In terms of popularity, we can see that acousticness has a -0.4 correlation, while attributes like loudness and energy have a positive correlation (0.34 and 0.33 respectively)
- Instrumentalness has a -0.3 correlation with popularity, I am guessing this also has to do with the rise of electronic/computer-generated music genres.

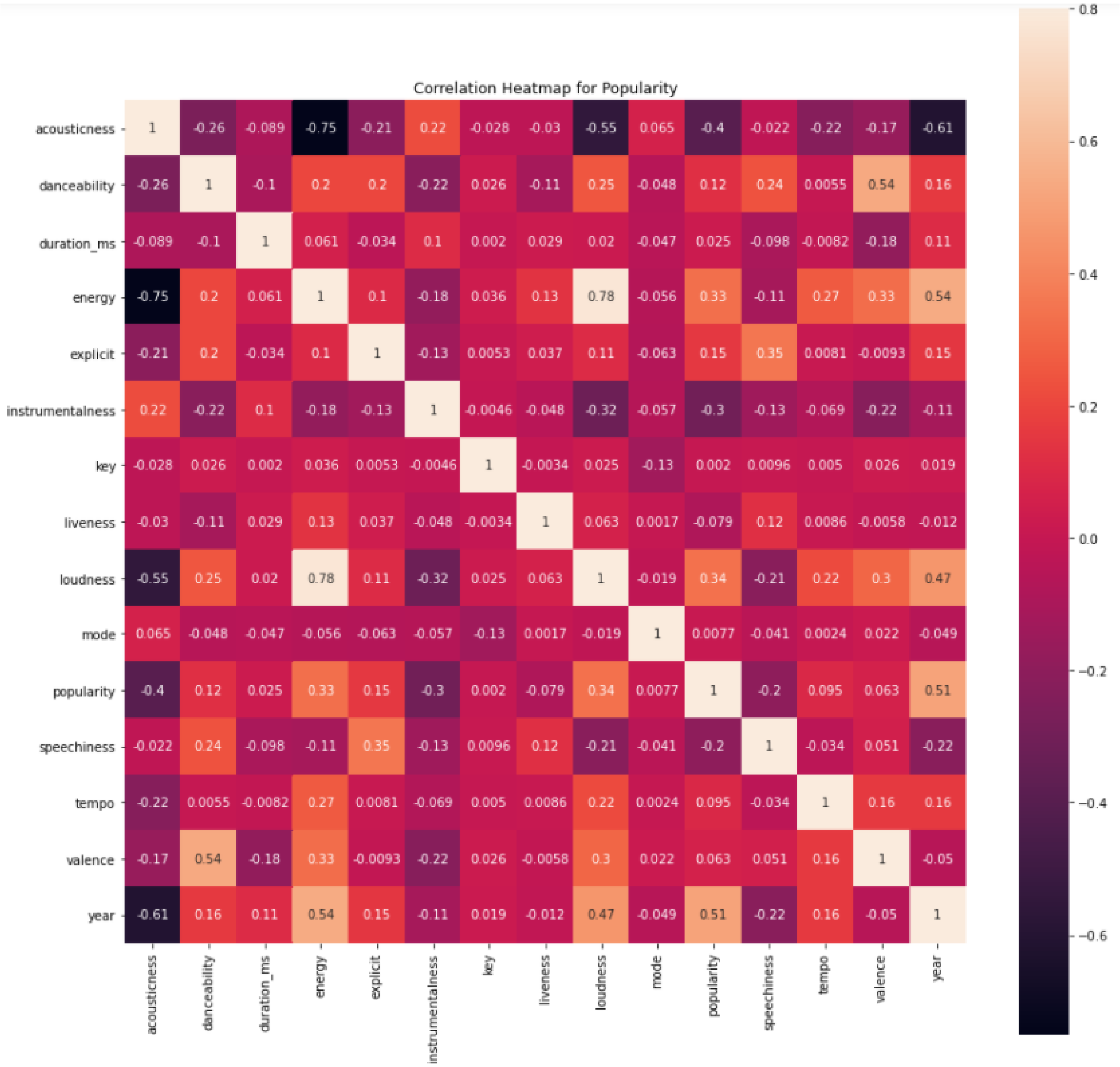


Figure 13: Correlation Heatmap

4 Song Structure Research

Drifting away from the data side of things, I wanted to use this opportunity to also display some of my anecdotal findings with creating popular songs. In this section, I'll walk through the common I-IV-V-VI chord progression, as well as the typical song structure of ABABCBB. This information will help accompany the information above to creating a popular song.

4.1 The I-IV-V-VI

The I-IV-V-VI, also known as the 1-4-5-6 chord progression, is commonly used in popular songs. I like to call this the lazy man's chord progression because it will literally sound good with anything as long as you keep the melody in the key. If you want to explore a more complicated but similar concept, check out the chord progression used for Pachelbel's Canon in D.

Here is a list of songs that all use the I-IV-V-VI chord progression:

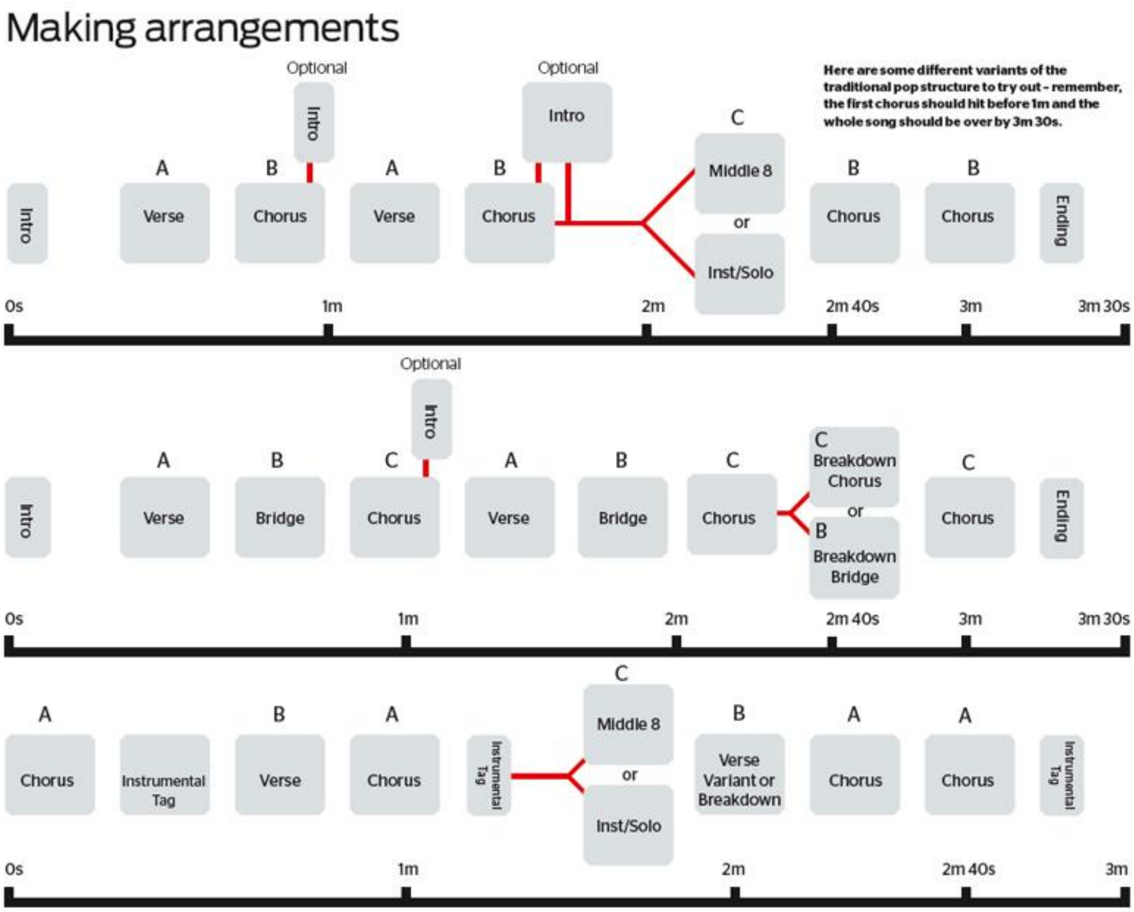
- "I will always love you" – Whitney Houston
- "Purple Rain" – Prince
- "All My Loving". "A Day in the Life", "A Hard Day's Night", "Real Love", and 10 + other Beatles songs!
- "Every Breath You Take" – The Police
- "Africa" – Toto
- "Perfect", "Castle over the Hill" – Ed Sheeran
- "Back to December" – Taylor Swift
- "All of Me" – John Legend
- "Cruise" – Florida Georgia Line
- "Don't Let Me Down" – The Chainsmokers
- "Drive By", "Hey Soul Sister" – Train
- "Hello" – Adele
- "No Woman, No Cry" – Bob Marley
- "Wedding Dress" – Taeyang

I think I've made my point here. All the songs listed above ranging from genres like rock, country, and EDM all use a variation of the I-IV-V-VI chord progression.

My advice? Figure out what key you want the song to be in (using the data we derived above) and use the I-IV-V-VI – it has worked for most of these popular songs!

4.2 The Breakdown of Pop Song Structure: ABABCBB

From arranging a ton of pop pieces, I have also noticed a common structure between them. After doing some digging, this structure is well known as the ABABCBB – and the many variations that follow. For a lack of better words, I have attached a pictorial guide on how this looks. As shown on the pictorial, this structure allows songs to end at around 3.5 minutes. If you recall back to our song duration time series plot, the average duration of songs in the past 10 years has been exactly around 3.5 minutes. Once again, making connections between data and general music trends, and using it to prove our point!


Making arrangements

Here are some different variants of the traditional pop structure to try out – remember, the first chorus should hit before 1m and the whole song should be over by 3m 30s.

Figure 14: Pop Song Structure (Source: www.musicradar.com/)

5 Conclusion: How to create a "popular song"

5.1 Conclusions

As shown from the data, most popular songs today tend to be louder, more energetic, and have a higher tempo. In contrast, songs with more instrumentals and acousticness do not do as well. We can also see that more popular songs are constructed on white major keys, with the most popular being E major, A major, and D major. In addition, most songs end at around 3.5 minutes as a general rule of thumb. Although not backed up by data, the I-IV-V-VI chord progression and ABABCBB structure are usually used in popular songs well.

5.2 Other Influencing Factors

If you do everything I mentioned above, you will meet the requirements of popular songs, but you will probably not become popular. As shown in Figure 13 (Correlation Heatmap), most attributes do not have a strong correlation with popularity. This is because there are so many other reasons that contribute to the success of a song. I will finish off by explaining a couple of these factors.

- **Artist:** This one should come as no surprise. It is much easier for an already well-known artist to release a new song that becomes popular.
- **Strong Team of Producers, Publishers, and Promoters:** The production quality and distribution power of publishers are extremely important when it comes to getting attention to a new song. Also, a good promotion and marketing team will make or break the sales.
- **Theme and Lyrics:** Are you singing about your lovely pet fish or a breakup that left you in the ruins? Choosing the right theme for a song will influence the lyrics and how well it sits with the audience. If you take a look at the Billboards, most songs are about love.
- **Luck:** Sometimes artists are one-hit wonders, where they release a hit and never see that type of success again. It is hard to predict the industry, and thus hard to become a hit overnight. Upon the base layer of skill, there is an added variable of luck – don’t underestimate this.

In conclusion, I hope this data analysis of music gave you a good understanding of what makes a popular song, and how you can try to create one yourself!