Determinants of Song Longevity: A Study of Spotify Weekly Charts

Aya Bekhtiar, Arthur Morvan, Aliénor Sabourdin

December 3, 2025

Github Repository : `https://github.com/Ayabekhtiar/spotify_charts_project`

This report revolves around studying Spotify tracks data, more precisely the weekly global charts spanning from 2017 to 2020. The aim of our work is to attempt answering the following question: *Are there specific song characteristics that make a hit have lasting success rather than just short-term virality?* We've went beyond an analysis of 'what makes a hit' and focused on sustainability of popularity.

Dataset Documentation

The main collected datasets come from `spotifycharts.com` and `kaggle.com`. These datasets contain information about tracks, their rankings, and their audio features.

Spotify Charts Webscraping

Using the Selenium library, we web scraped global ranking CSV files from the week of December 29th, 2016 to the week of December 31st, 2020. Incidentally, the weekly charts renew each Thursday, so we decided to include the extra days of 2016 and 2021 because the first week of 2017 officially starts in the last days of 2016, and the last week of 2020 extends into early 2021.

Specifically, there is a file **regional-global-weekly-YYYY-MM-DD.csv** for each week containing the top 200 songs of that week's Spotify chart. For each song, the dataset indicates: the artists' names (`artist_names`), the title under which it is published (`track_name`), its producers (`source`), the highest rank it has ever reached (`peak_rank`), its previous rank if any, and -1 if none (`previous_rank`), the number of weeks it spent on the charts (`weeks_on_chart`), and its number of streams (`streams`).

Spotify Web API and How Kaggle Saved the Day

Unfortunately, Spotify Web API blocked access to its audio features section in May 2025 for individuals, reserving its usage for organizations that submit a request. Our alternatives were to pay third-party data providers to obtain this data, which was not feasible, or to find Spotify audio features datasets on Kaggle. We found three datasets on Kaggle that provided audio features but only for songs released between 1921 and 2020, so the latest year we could work on was 2020. Besides that, the choice of 2017 was simply because it has the earliest accessible week on `spotifycharts.com`.

We use 3 kaggle datasets for data enrichment:

- **tracks.csv**: contains for each song its id (`id`), title (`name`), length in ms (`duration_ms`), artists' names (`artists`), artists' id (`id_artists`), release date (`release_date`), and the following Spotify audio features: `danceability`, `energy`, `popularity`, `key`, `explicit`, `loudness`, `mode`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence`, `tempo`, `time_signature`.
- **Kaggle enrichment 2**: contains a CSV that displays for each song its id (`id`), title (`track_name`), artists' names (`artist_names`), length in ms (`duration_ms`), producers (`source`), number of weeks spent on the charts (`weeks_on_chart`), number of streams (`streams`), and the following Spotify audio features: `key`, `mode`, `time_signature`, `danceability`, `energy`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence`, `loudness`, `tempo`.
- **Kaggle enrichment 3**: a folder containing two CSV files, each providing for every song the artists' names (`artist_name`), ID (`track_id`), name (`track_name`), length in ms (`duration_ms`), and the following Spotify audio features: `acousticness`, `danceability`, `energy`, `instrumentalness`, `key`, `liveness`, `loudness`, `mode`, `speechiness`, `tempo`, `time_signature` `valence`, `popularity`.

Final Datasets We Work with

Our cleaning, analysis and visualizations are based on the following data storage pipeline:

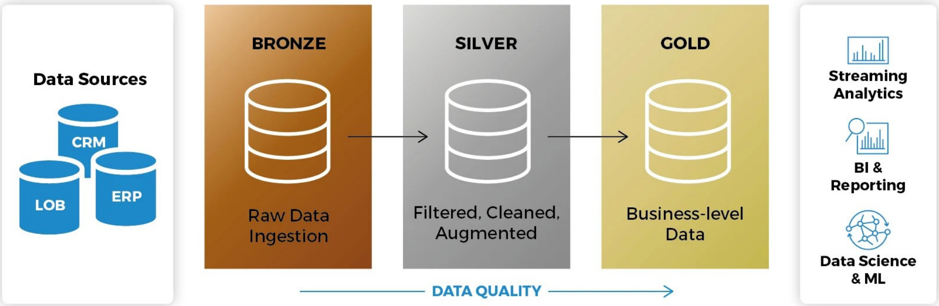

Figure 1: Medaillon Architecture (*Databricks.com*)

- **Bronze Layer** (`data/bronze/`) – raw, unprocessed input data:
  - weekly Spotify charts
  - tracks metadata and audio features found on kaggle, which had been retrieved via the Spotify API
- **Silver Layer** (`data/silver/`) – data processing steps between raw input files (bronze) and clean final files (gold):
  - consolidated weekly charts
  - merged datasets combining chart entries with track features
  - stored as Parquet files for improved performance and compression
- **Gold Layer** (`data/gold/`) – final analytics-ready data:
  - fully processed and enriched dataset

  - missing values handled and feature types normalized
  - stored as Parquet files

The reason we chose this architecture is mainly because having the bronze layer allow us to keep all the raw data used in the study for documentation, security and verification purposes.

Results: Findings and Limitations

Findings

Before answering our research question, we first needed to define what it means for a song to "stay a long time" in the charts. Based on our data, we found that a track remains on the charts for an average of 17 weeks. However, the median is only 8 weeks, which shows that our data has a highly skewed distribution: while most songs disappear quickly, a few stay for a remarkably long time, creating important outliers.

Another important preliminary question is: *what makes a song enter the Spotify charts?* According to Spotify's support website, the main element for a song to enter the chart is its weekly number of streams. When examining weekly streams for the most frequent songs in the charts, we noticed three recurring patterns: (1) the number of streams peaks around the song's release and then slowly fades over time; (2) the song experiences two or three smaller peaks in streams; (3) it maintains steady, lower but still significant streaming levels. We also noticed that the most consistent song on the chart (appearing in 207 out of 210 weeks) actually has lower streams than most of the other highly consistent tracks, indicating that lasting a long time does not necessarily equal peak popularity.

With this context in place, we explored whether audio features themselves could help explain why some songs last and others do not. To do so, we created two categories of songs: "short-lived hits," songs that stayed for less than 8 weeks (the median time a song spends on the charts), and "long-lasting hits," songs that stayed longer. By comparing the audio features of these groups, we found that long-lasting hits tend to be slightly more danceable, with higher valence (more positive sounds) and slightly higher energy levels. They are also less acoustic, less speechy, and a bit slower. Lastly, they tend to be slightly longer, by around 2 seconds.

We also tested the correlation of all these audio features with the number of weeks a song spent on the charts. This analysis does not output significant findings: all correlations remain below an absolute value of 0.06, showing that there is no audio feature that can predict chart longevity linearly. However, considering the correlations they have between themselves (for instance, danceability having a -0.2 correlation with acousticness), there seems to be a coherent pattern: long-lasting hits typically have a polished, upbeat, danceable pop profile, while short-lived songs are slightly more extreme (more acoustic, faster, or more speechy).

Since there were no linear relationships but still a visible trend, we wanted to push our analysis further and test whether there are more complex, non-linear interactions between audio features and longevity. In other words, 'can a specific *combination* of audio features determine whether a song lasts on the charts?' Although this analysis showed a more significant influence of audio features (especially a song's duration, its tempo, and its danceability), it also showed that while they do have an influence, it is relatively weak and not sufficient to explain why a song would stay on the charts.

Thus, while audio features may contribute to longevity, they do not determine it. External factors such as promotion, timing, cultural moments, or artist fame likely play a more important role.

Limitations

The main limitation is the reliance on Spotify's audio features. While they offer useful metrics, they still simplify many complex musical attributes. Some feature computations (e.g., danceability) are not transparent and can be subjective. They also omit important elements such as melody, harmony, arrangement, and production quality.

Furthermore, we cover charts only between 2017 and 2020. While Spotify's official website states that the main driver for entering the top charts is weekly streams, we know that other factors, such as social virality or algorithmic playlist exposure, may play a role and are impossible for us to take into account. Additionally, the chart methodology described on Spotify's website in 2025 may not be identical to the one used during our study period, meaning that some chart presence drivers may be unknown today.

Another limitation is that listening behaviour changes with cultural moments, seasons, or personal trends. These evolving listening habits directly affect streaming patterns, yet they cannot be captured with our data. As a result, part of a song's chart longevity may simply reflect changes in audience behaviour rather than any intrinsic audio characteristic.

Despite these limitations, the patterns revealed across distributions, trajectories, correlations, and predictive attempts all point to the same conclusion: the sound of a song alone does not determine how long it will last.

Reflection on the data lifecycle

Through this project, we directly applied the data lifecycle. Each stage came with its own challenges, even the very first one. Theoretically, our data collection was straightforward: everything was openly accessible, no private data involved. However, we encountered many issues: some pages were inconsistent, scraping had to be automated carefully, and one of us even had their Spotify account temporarily removed due to "suspicious activity", even if the charts are stated as entirely public data.

The real challenge, however, was data cleaning. It was by far the densest, challenging, and important phase of the lifecycle. One difficulty was that it was our first encounter with the data, so what was useful or irrelevant was not yet fully defined. For instance, our first plots looked strange, and after digging in our cleaning steps, we realised that one track could appear under multiple IDs, when appearing as an EP or in an album for example. We had to decide what made a track unique and implemented a unique ID system for songs. Working with multiple datasets also made us question on how to merge entries, and when to trust or override the dataset. Handling missing values was equally complex. The explicit column, was incomplete, but as we thought it was an interesting feature, we tried to fill in its missing instances. But we found no external of sufficient quality, and filling it with its distribution features in the data (e.g. median or mean) would have introduced bias. Our only option was to infer on the artists and title. It worked really well, but required using the Gemini API. Thus, we decided that dropping the rows for which it was empty was the most rigorous approach, as they represented a small part of our dataset. All these choices were difficult to make, especially since they determined what our analysis could legitimately claim.

Storage was comparatively simple. We organised the data into Bronze–Silver–Gold layers for reproducibility, documentation and security purposes, and as the dataset was not massive, we stored it in Parquet format, allowing better compression, speed and schema enforcement without unnecessary complexity.

Data analysis brought an unexpected issue: although we knew what question we wanted to answer, we did not know which metric would actually resolve it. At first, we generated many graphs and statistics and after a through exploration, we could identify some convincing patterns, not through a single graph, but by analysing them together. At first, we expected to find a linear combination, but our findings made us reconsider, and we pushed our analysis deeper, ending up looking for more complex feature patterns.

The communication phase, so writing this report, made us look back on our entire study. It helped us identify an important merging error in our third dataset, which we corrected, improving our results. Besides, it made clear what our results meant: some insights we thought were important turned out unclear or unsupported when trying to explain them, so we bettered our analysis again, and ended up with what we consider being solid findings.

Overall, we saw how much each step is interdependent, and how the quality of the final output ultimately depends on the compounded success of all preceding steps.

Key Takeaways

This project allowed us to apply the concrete data science and engineering skills we developed throughout the course and in our prior experiences to a challenging and interesting problem

(1) Even if a project is well defined and its steps clear,it might not unfold as expected. Public data turned out to be less easy to handle than we assumed, especially the discrepancy between how accessible it says it is and how it really is.

(2) We had mentionned in class that cleaning the data usually represents 80% of the time spent on a project. This was indeed the case for us, and thorough cleaning became the intellectual core of the project. As stated above, we've made deliberate, transparent choices. And the challenge came both from identifying the cells that needed processing and from finding coherent ways to enrich or clean the data without compromising quality or introducing bias.

(3) We also learned to addapt, comming prepared is important but we've had to iteratively adapt question the different techiques used at each step of the data lifecycle. For that visualising is key and we relied on extensive prints and visuals to check if the steps where behaving as espected. Furthermore, we learnt not to be afraid to question our early assumptions to better our analysis. When we saw how weak our first correlations were (visualisation 5) and adapted our model to a non linear one.

Let's not forget that the main lesson we take from this project is that there is no straightforward answer to how to make a song that will be popular for years. This means two things: there are other parameters than just audio features that make a hit last, such as marketing or seasonality; and maybe also a song's musicality and "greatness" cannot really be measured through it's levels of "danceability" and "accousticness", and could come from an undefineable artistic genius.

That's good news for artists, as they may still have a few years before having to compete with an AI beatmaker!