mini-project-1

Akash Mahajan (akashmjn@stanford.edu), Raunaq Rewari (raunaq@stanford.edu)

Dataset - Spotify's Worldwide Daily Song Ranking

Introduction

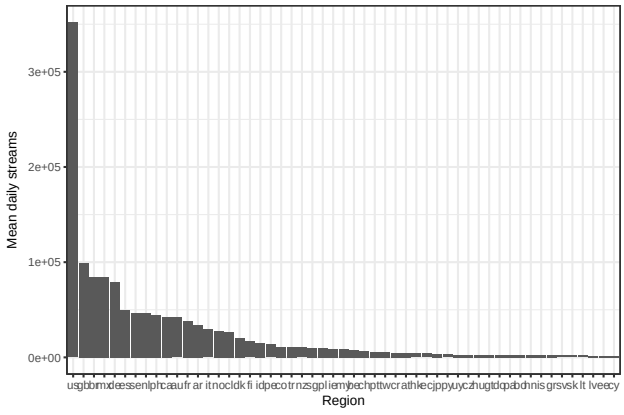
Whether we like it or not, we have all been inflicted by 'Despacito' at some point. Released early this year, the song quickly amassed 4.6 billion streams by July 2017, making it the most streamed song in history[1], and the quickest to reach 3 billion views on Youtube, despite being in a completely foreign language.

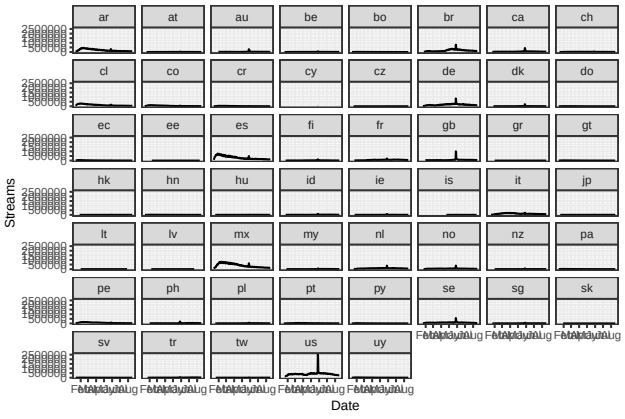
Motivated by our pain, admittedly awe as well, and a general curiosity about how such phenomenon spread, we set out to explore a dataset from Spotify[2]. What are the signs that a song will make the charts? What influences how long it will stay? Are there any interesting global correlations (or lack of) in the spread of popularity?

Dataset overview and validation

Our dataset comprises of the daily top 200 most listened songs (by stream count) on Spotify in 54 regions over a period of 1st January 2017 to 17th August 2017 (source: [3], gathered via an automated crawler accessing the Spotify Web API). The regions comprise 53 countries, and a 'global' region contining overall top songs. For each day and region, we have a list of the top 200 songs, with their position, a unique identifier, basic information such as track name, artist, and the number of streams. In total, this amounts to about 2 million rows including 4682 unique artists, and 11932 unique songs tracked over this period.

This data has no missing values. However some initial exploration (below) shows that for simplicty of analysis, some of the regions can be filtered out due to a low number of mean daily streams. Also, looking at the region-wise trends for 'Despacito', we notice a massive spike that probably needs to be re-validated via the API. If these are indeed outliers, they could be interpolated using a sliding median window.





[1] https://www.theverge.com/2017/7/19/15997816/despacito-song-luis-fonsi-daddy-yankee-justin-bieber-most-streamed
[2] A global music streaming service present in 53 countries https://www.spotify.com/us/
[3] https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking

Analysis

Possible Objectives

For a concrete analysis of our data, we formulate our problem in two possible ways.

Binary reponse variable: given historical data for songs in the top 200, can we predict in advance if a song will make it to the top 20?

Continuous reponse variable: how long will a song stay in the top 20, or forecasting number of streams for a song in the next month.

Transformation

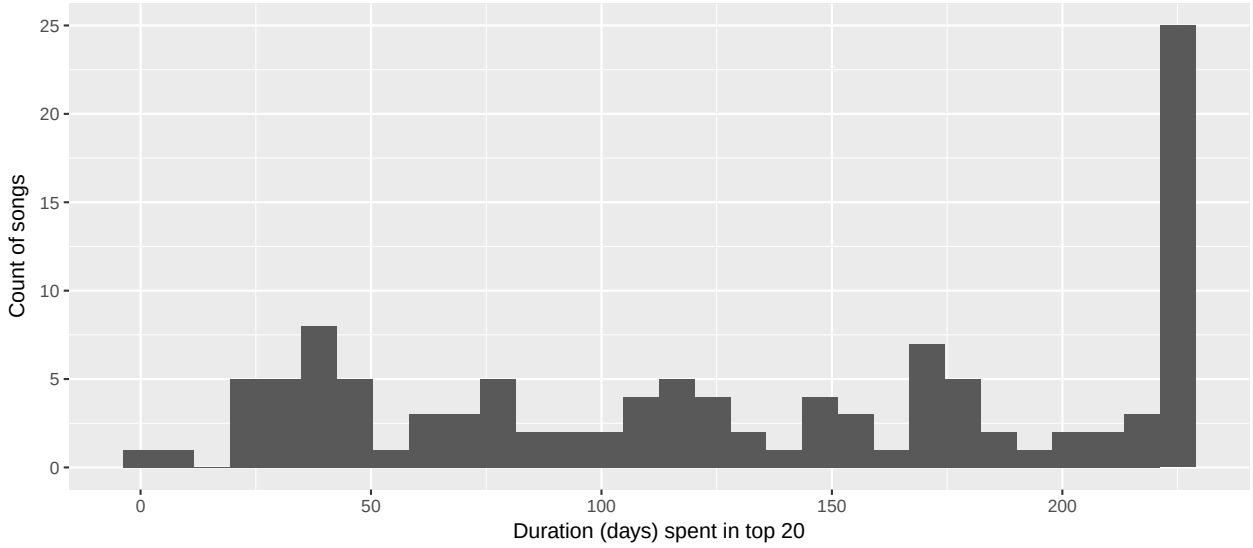
For the objectives outlined, we need to transform our data. While we our data is fundamentally in a time-series form, for our binary classification task we will aggregate data for a period of time by song. This will require extraction of some features related to the dynamics, popularity in different regions, and potentially additional information about the song from the API [^4] such as genre, audio analysis etc. A hold-out set can be created by subsetting out a time period / or particular songs, from our dataset.

[^4: https://spotipy.readthedocs.io/en/latest/]

Preliminary Exploration

So far we have done an initial exploration to get an understanding of dynamics of song popularity. For simplicity, we look at just the global region below where for our time period of interest, there have been 114 songs that have entered that charts at some point in their life.

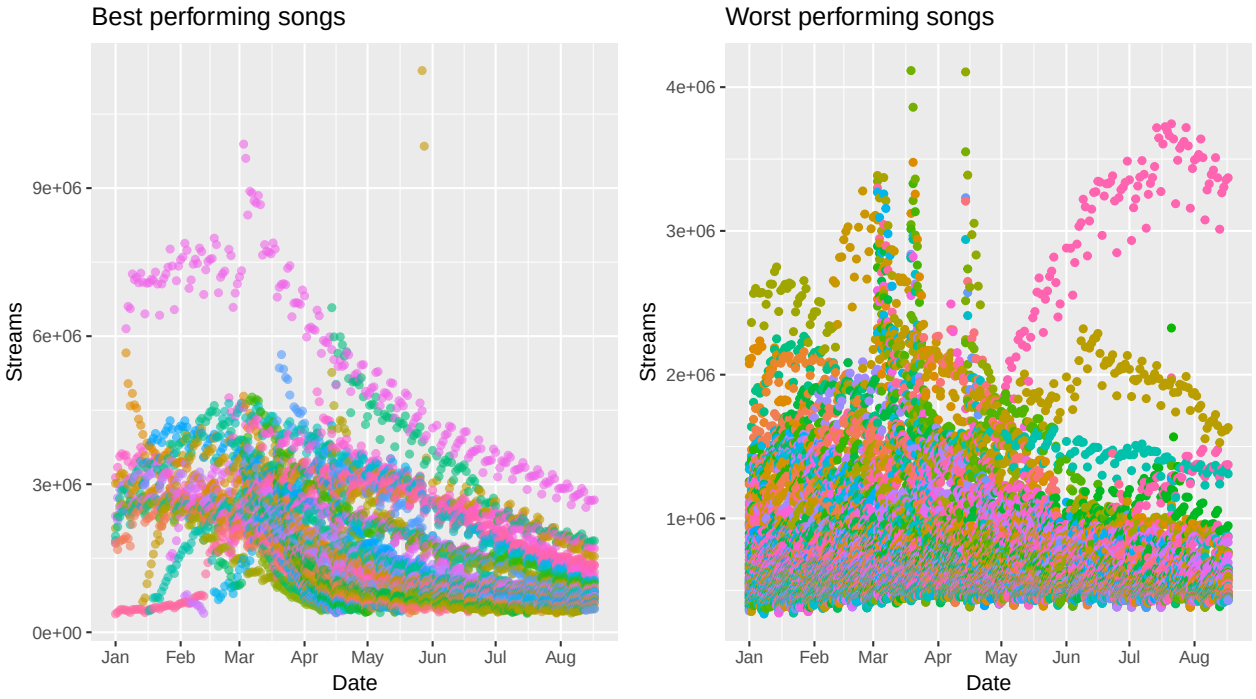
Given this peaking dynamics of song popularity, for these 114 songs we want to look at the duration spent in the top 20 and check if this is well distributed enough.



We see that it is fairly evenly distributed, except for a large number of songs that might be extending over our entire time range. Reasons for this will need to be investigated further, and might influence our final choice of continuous variable.

The following two plots show the streaming trends for the best and worst performing songs respectively. It can be seen that there is seems to be a much more predictible pattern for the top songs. Each song seems to

rise in popularity pretty early in its life and slowly die out. The worst performing songs on the other hand seem to not have a discernible pattern.



Open Questions

Going forward from this point to the next stage in our analysis, i.e. modelling the song popularities there are a few questions that need to be addressed.

1. What kind of features do we build from the temporal history of songs? Should we use time-series models?
2. How do we detect / measure the correlation between popularity in regions? The dataset has thousands of time series for each song across multiple regions. Coming up with a metric to combine and detect this will be hard.
3. Are there any additional correlations with genre, artist, lyrics etc. worth exploring?
4. Deciding our continous response variable between either a forecasting approach or duration in top 20.