Spotify Top Charts Analysis Final Report

An analysis of musical trends across several axes.

Liam McChesney
Computer Science
University of Colorado - Boulder
Boulder, CO USA
limc8676@colorado.edu

ABSTRACT

Throughout this project I data mined a dataset of Spotify Charts data. Using the af traits attached to each song I was able to cluster the songs into different groups. My main questions involved whether or not there are meaningful trends across region, chart, artist, rank, popularity, or trend. Using this analysis artists/producers may be able to tailor their songs in a specific way to reach a specific audience or achieve notoriety. Another possible application is that given a song that is already made you may be able to adjust your marketing strategy based on the results found in this report. A brief overview of the results includes that for the most part there are no meaningful trends across the above mentioned axes. Distributions remain relatively close to the overall cluster breakdown. However, there are some trends that were found that could be applied.

RELATED WORK

Angela Cottini did some work on the Spotify Top 200 chart from 2020-2021. She focused on the most frequent artists, song length, songs that stayed on the chart the longest, and release year. One interesting thing she found is that the number of times a song appeared on the Top 200 list was not correlated well with the number of followers the artist had. Another interesting find was that in a two-year period over half of the songs on the list are released in the first year of the period as opposed to before the period or in the second year. This helps us to see a little bit about how long it typically takes songs to become popular and how long they stay popular. (Cottini 2021)

Reika Fujimura did analysis on what musical elements are different on a yearly basis between the songs on the Top 200 list from 2018 to 2022. They found significant trends around Christmas time in the US region. Consistently the acousticness and valence (positivity of a song) would increase up till Christmas and then drop off immediately after. Other than that, valence seems like the trait the changed most from year to year where all other traits follow a similar pattern each year. Comparison of musical traits between regions was also done. The United States and Canada have relatively similar data whereas the United States and Japan have large differences in categories such as energy and duration. (Fujimura 2022)

Both of the above papers gave a good basis for my project. I took the type of work that Reika Fujimura did and applied it to different axes. For example, I used k-means clustering to cluster the data and analyzed the distribution of different clusters across multiple axes. As for the work that Angela Cottini did, I did some artist analysis to see if the top artists typically wrote the same type of song or if their songs follow the same distribution as the dataset.

DATASET

The dataset I used was the “Spotify Charts (All Audio Data)” pulled from Kaggle with over 26 million rows. This dataset contains the global “Top 200” and “Viral 50” Spotify charts from various countries/regions from January 2017 till March 2024. These charts are updated every two to three days. The dataset can be reached at the link below. Rows include title, rank, date, artist, URL, region, chart, trend, streams, track_id, album, popularity, duration_ms, explicit, release_date, available_regions, af_dancability, af_energy, af_key, af_loudness, af_mode, af_speechiness, af_acousticness, af_instrumentalness, af_liveness, af_valence, af_tempo, and af_time_signature.

https://www.kaggle.com/datasets/sunnykakar/spotify-charts-all-audio-data

TOOLS

The following tools were used for the project:

- **VSCode**: Environment
- **GitHub**: Submissions and Remote Project Storage
- **Jupyter Notebooks**: Format
- **Python**: Language
- **Pandas**: DML Library for analyzing data. Functions used include read_csv() and to_csv() to read and save the dataset respectively. Also, head() and info() for checking work and exploring the dataset. Lastly, group_by(), isin(), unique(), get_group(), count(), mean(), and sort_values() for data manipulation.
- **Scikit-learn**: Clustering, functions include the KMeans clustering setup as well as preprocessing for normalizing the af traits before clustering.
- **Seaborn and Matplotlib**: Graphical display libraries that included scatterplots, pie charts, and bar graphs.
- **Tabulate**: Tool for python to organized data frame printing

MAIN TECHNIQUES APPLIED

Data Cleaning: There were some errors in the release dates as some of them were after the date they appeared on the chart. Some songs have null values in the title, artist, available markets, popularity, and af columns. I did not use the release dates or available markets column. Since there were not very many null values between title (11 rows), artist (18 rows), and popularity and af traits (1% of data) I dropped the data points with null values. This resulted in a data size decrease of only about 1%. Streams had about one fifth of its data as null so this was not included in the analysis.

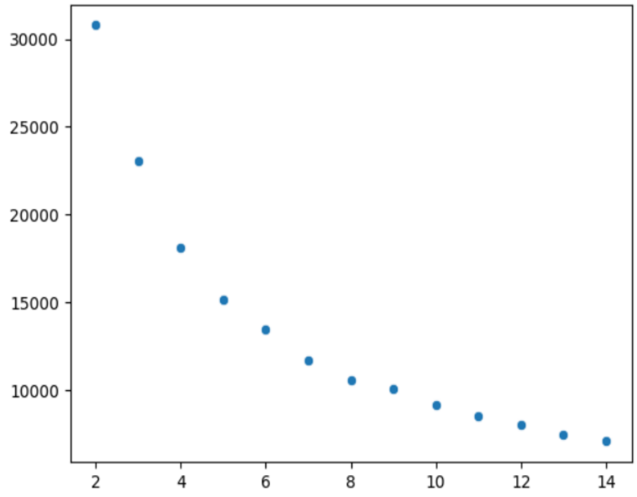
Data Preprocessing: I converted the trend field to a tertiary numerical value (-1 for down, 0 for same and 1 for up) which allowed some interesting calculations to be done on how the songs move up and down charts as well as saved memory. I also converted chart to a binary numerical value (2 for Top 200 and 5 for Viral 50). This helped in file size reduction. Several columns such as URL were unnecessary and were removed for ease of work.

Data cleaning/preprocessing was effective and reduced file size by 22 GB down to under 4 GB. Some preprocessing was completed including the transformation of string fields into integer fields (trend and chart). I did not need to break the data up into further chunks (split on region or chart) because the file reduction was so effective. Run times for any given block of code never took more than 10 minutes and rarely took more than 5. The following columns being dropped contributed to the drop in file size: url, track_id, album, streams, duration_ms, explicit, available_markets, and release_date.

Data Integration: This was not necessary as the data was all in one dataset to begin with.

Evaluation Methods:

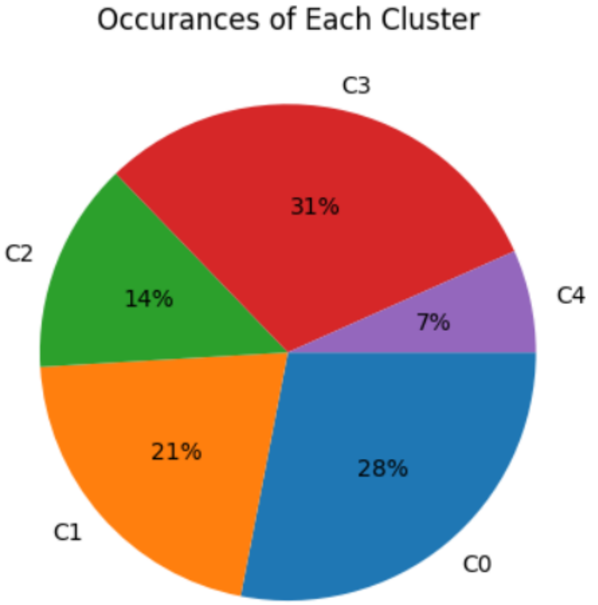
The bulk of my analysis came from using k-means clustering. I first performed normalization on the af traits. I used the scikit-learn function preprocessing(). Then I used the KMeans fit() function to make the clusters. I analyzed the inertias of clusters from 2-14 to see what the best k value would be for my analysis. Graph 1 below shows these inertias versus the cluster number. I settled on a k of 5 as it was close to the elbow of the graph and seemed like a good number of categories to break songs up into with not too few and not too many.



Graph 1

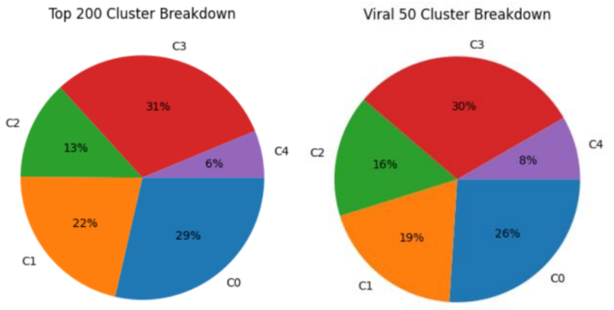
RESULTS

To begin with the results section, we take a moment to describe each cluster. Cluster 0 is the second most common cluster in the data set taking up 28%. It is characterized by a high average tempo as well as very low instrumentalness. C1 is the third most common cluster in the dataset taking up 21%. It is characterized by the highest average danceability, lowest average mode, and lowest average instrumentalness. C2 is the fourth most common cluster in the dataset taking up 14%. It is characterized by fairly low loudness and relatively high instrumentalness. C3 is the most common cluster in the data set taking up 31%. It is characterized by high average loudness and the highest average tempo. The distribution of the clusters can be seen in Pie Chart 1 below. Full characteristics as well as example songs for each cluster can be found below in the Appendix.



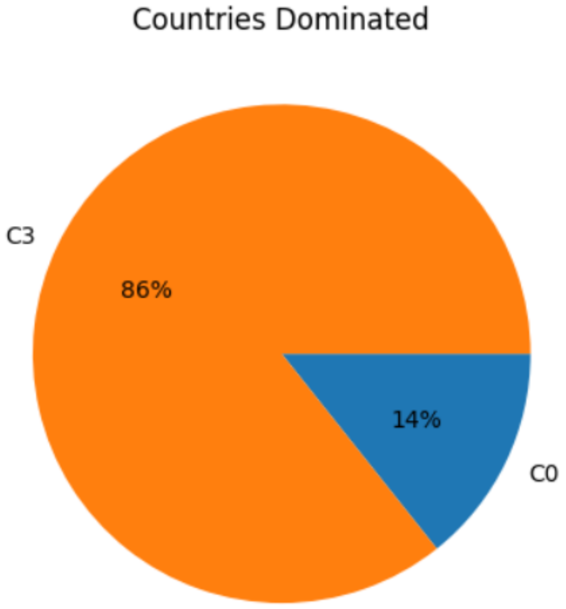
Pie Chart 1

Firstly, we look at the distribution as compared to the two different charts. The Top 200 chart is based on the total number of streams and is generally a good indicator of long-term popularity. The Viral 50 chart shows songs that are experiencing a sharp increase in streams which shows which songs are gaining popularity quickly. Appearance on this chart doesn’t necessarily indicate long term popularity or high stream count in the end. That said there was very little difference between the distribution of clusters across the two charts as can be seen in Pie Charts 2 and 3. The only noticeable difference between the two charts is that Cluster 2 takes 3% more spots on the Viral 50 chart. Perhaps songs in Cluster 2 may be slightly more likely to gain popularity quickly.



Pie Chart 2 and 3

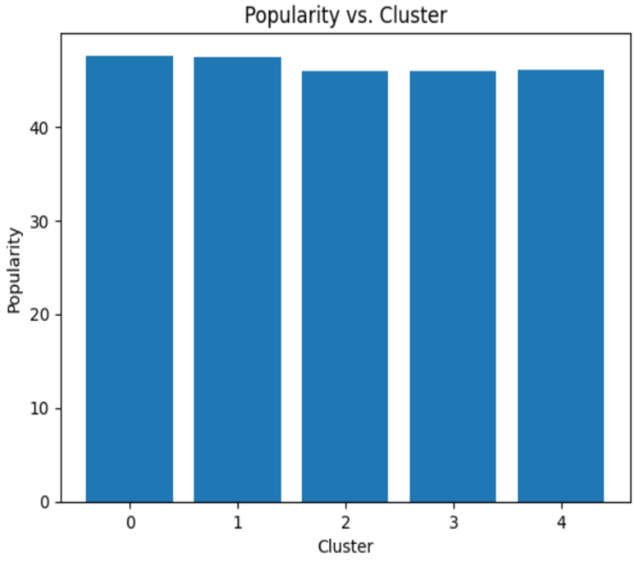
The next axis that we look at is the country/region axis. I took each country/region and looked at which cluster dominated it. One would expect that clusters 3 and 0 would dominate most countries if the distributions were similar. This is what I found as can be seen in Pie Chart 4.



Pie Chart 4

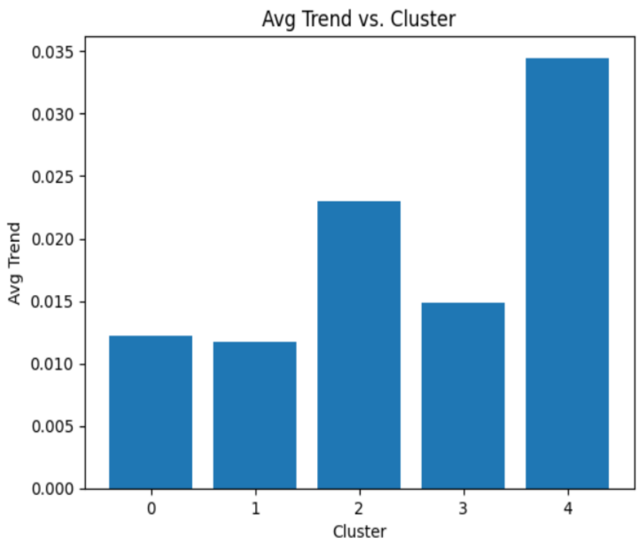
There are 70 different countries/regions in the data set so this split represents a 60-10 country/region split. At first glance this finding is relatively insignificant. However, upon further inspection of the 10 countries/regions where Cluster 0 dominates we find something interesting. The following countries/regions are the ones dominated by Cluster 0: Argentina, Colombia, Bolivia, Chile, Dominican Republic, Ecuador, El Salvador, Panama, Uruguay. All of these countries/regions are in Latin America! This is the most interesting thing I found in my analysis.

The popularity statistic is a measure that Spotify uses to analyze artists popularity. I looked at the average popularity of each cluster to see if more popular artists typically write one type of song. As can be seen in Bar Chart 1 there is practically no deviation between the average popularity of each cluster. This shows that artists generally write various songs in various clusters regardless of popularity.



Bar Chart 1

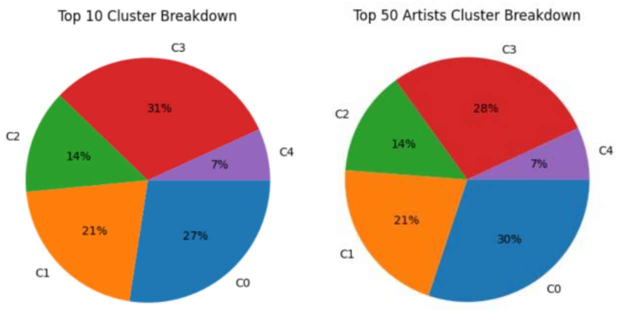
The next statistic that I analyzed was the average trend across each cluster. I converted the trend field to a tertiary numerical value to be able to do this calculation. 1 corresponds to an appearance or rise on the chart, 0 corresponds to maintaining rank, and -1 corresponds to a fall on the chart. Therefore, a positive average means that the song spent more days rising up the chart than falling off of it whereas a negative value would indicate the opposite. Interestingly every cluster had a positive average trend. That said, the average is quite low for all of them as can be seen in Bar Chart 2.



Bar Chart 2

Cluster 4 has the highest average trend and also is the most unique cluster of them all (as a reminder, cluster 4 is slower and more acoustic compared to all the other ones). This shows that slower more acoustic songs take longer to rise up the charts or fall off faster(or both) than faster paced more electric songs.

Lastly I looked at the top artists and the top songs of each chart to see if the top artists have achieved their success by writing a single song type or if the best of the best songs is dominated by a certain cluster. I looked at the top 50 artists based on appearances on the chart and every song that reached at least rank 10 on either chart. Both of these turned out not to have significantly different distributions as can be seen in Pie Charts 5 and 6.



Pie Chart 5 and 6

APPLICATIONS

Applications largely come from the general cluster distribution, country/region, chart, and trend analysis. The general cluster distribution tells us that the best song types to shoot for to achieve popularity would be Clusters 0 and 3 which make up 59% of the songs on the chart. This is 19% more than the two would take up if the distribution was even.

For the chart analysis the slight edge that Cluster 2 has on the Viral charts seems to indicate that they have a better ability to gain popularity quickly even if they don’t maintain popularity long enough to reach the Top 200 list. Knowing this, Cluster 2 songs may be beneficial to use in something like TikTok where virality is crucial. If you want your song to have more longevity perhaps avoid a Cluster 2 song. This also correlates with the general distribution trend as it is the second least frequent cluster in the data.

For the trend analysis we saw that Cluster 4 has a significantly higher trend average than the other categories. This indicates either a slower rise or a faster fall from these songs when compared to other clusters. Generally, neither of these would be a beneficial trait for a song to possess as a faster rise and slower fall means more time on top of the charts. Couple this with Cluster 4’s very small share on these charts in the first place and that indicates that Cluster 4 is a good song type to avoid when trying to create a popular song. Clusters to shoot for include 0,1 and 3 which all have low average trends. As well as average or above average general cluster distribution.

When it comes to the region/country dominance we found that most regions/countries were dominated by cluster 3 save for 10 Latin American Countries that were dominated by Cluster 0. Given this if you are trying to spread your music in Latin American countries/regions then Cluster 0 is a good cluster to emulate. Otherwise, if you have already written a Cluster 0 song and have advertising money to spend then Latin America may be the most effective place to spend money for that song. Couple this with Cluster 2’s general popularity across the world and it makes it a safe bet to try to emulate especially for Latin American artists.

REFERENCES

[1] Angela Cottini, 2021. Analysis of Spotify's Top 200 Chart (2020-2021) DOI: https://rpubs.com/arcottini/821526

[2] Reika Fujimura. 2022. Spotify Music Chart Trend Analysis. DOI: https://www.reikafujimura.com/post/spotify-music-analysis

APPENDIX

Cluster 0:

Title	**Artist(s)**
Black Beatles	Rae Sremmurd, Gucci Mane
One Dance	Drake, WizKid, Kyla
I Feel It Coming	The Weekend, Daft Punk
24K Magic	Bruno Mars
Bad Things (with Camila Cabello)	Machine Gun Kelly

Trait	**Average Value**
Danceability	0.686085
Energy	0.692992
Key	1.41752
Loudness	-5.21836
Mode	0.653873
Speechiness	0.107761
Acousticness	0.212659
Instrumentalness	0.00843946
Liveness	0.168638
Valence	0.54673
Tempo	133.688
Time Signature	3.97911

Cluster 1:

Title	**Artist(s)**
Bad and Boujee (feat. Lil Uzi Vert)	Migos
Fake Love	Drake
Closer	The Chainsmokers, Halsey
Don’t Wanna Know	Maroon 5, Kendrick Lamar
No Problem (feat. Lil Wayne & 2 Chainz)	Chance the Rapper

Trait	**Average Value**
Danceability	0.71484
Energy	0.671497
Key	9.36255
Loudness	-5.66475
Mode	0.490027
Speechiness	0.104119
Acousticness	0.220092
Instrumentalness	0.00685636
Liveness	0.16768
Valence	0.549332
Tempo	105.509
Time Signature	3.9814

Cluster 2:

Title	Artist(s)
Bounce Back	Big Sean
I Don't Wanna Live Forever	ZAYN, Taylor Swift
Chill Bill (feat. J. Davi$ & Spooks)	Rob $tone
Party Monster	The Weekend
No Heart	21 Savage, Metro Boomin

Trait	Average Value
Danceability	0.668556
Energy	0.501117
Key	2.26322
Loudness	-8.91202
Mode	0.626223
Speechiness	0.114764
Acousticness	0.381778
Instrumentalness	0.022206
Liveness	0.168072
Valence	0.437285
Tempo	105.559
Time Signature	3.94578

Cluster 3:

Title	Artist(s)
Starboy	The Weekend, Daft Punk
Broccoli(feat. Lil Yachty)	Shelley FKA DRAM
Let Me Love You	DJ Snake, Justin Bieber
Déjà vu	J. Cole
OOOUUU	Young M.A

Trait	Average Value
Danceability	0.686594
Energy	0.686516
Key	6.97107
Loudness	-5.49341
Mode	0.533315
Speechiness	0.110622
Acousticness	0.224421
Instrumentalness	0.0115122
Liveness	0.173424
Valence	0.549175
Tempo	134.237
Time Signature	3.96169

Cluster 4:

Title	**Artist(s)**
Caroline	Aminé
Controlla	Drake
Say You Won’t Let Go	James Arthur
iSpy (feat. Lil Yachty)	KYLE
I Want You Back	The Jackson 5

Trait	**Average Value**
Danceability	0.625392
Energy	0.447814
Key	7.9466
Loudness	-10.2482
Mode	0.567191
Speechiness	0.121153
Acousticness	0.45766
Instrumentalness	0.047308
Liveness	0.161588
Valence	0.39791
Tempo	90.1896
Time Signature	3.94632