Spotify Data Analysis

By Isabella Lindgren

Background

- Founded in 2006
- Digital audio streaming platform
- 248 Million MAU’s
- 113 Million Subscribers

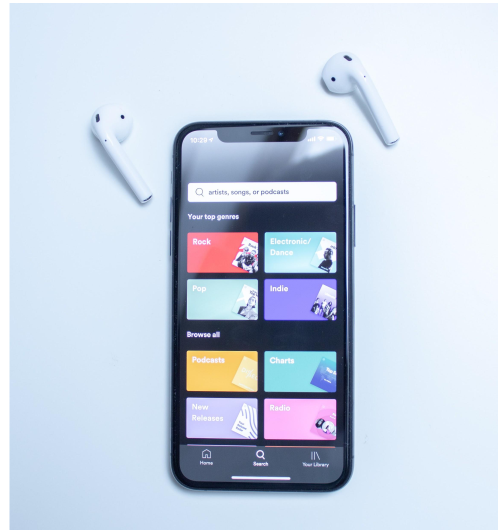


Photo by Patrik Michalicka on Unsplash

For those who aren’t familiar with Spotify, it is a digital audio streaming platform that originated in Stockholm, Sweden. It was founded in 2006 by Daniel Ek and has rapidly grown in popularity worldwide since then. Spotify provides access to music, podcasts and video content from record labels and media companies to consumers around the world! Users are able to browse an incredible variety of tracks by artist, album, or genre and can create, edit and share playlists. Spotify currently has about 248 million monthly active users and 113 million subscribers, with its paid user base growing around 31% each year.

Business Understanding

How can we provide a unique and exceptional listening experience to our users?

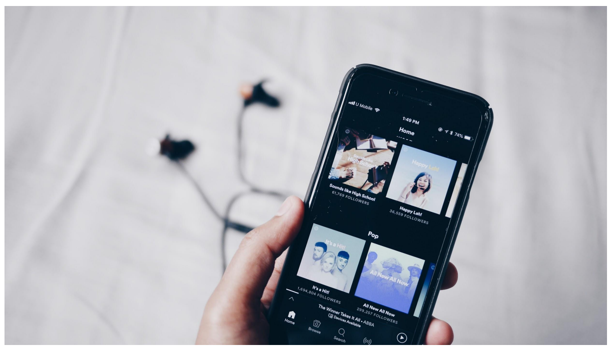


Photo by Fixelgraphy on Unsplash

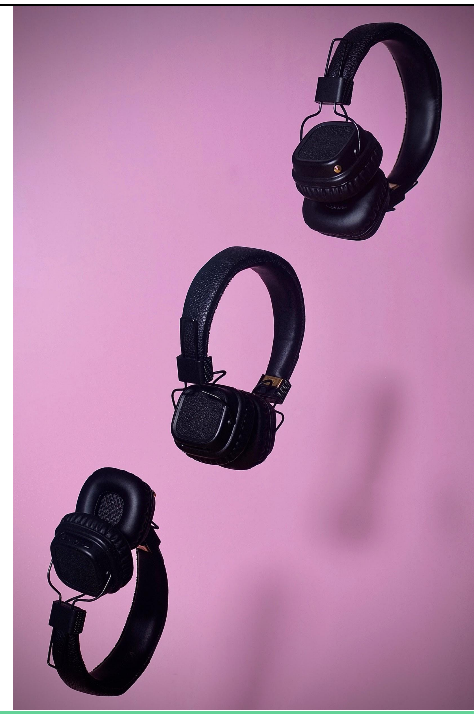
So, I approached this project as a Spotify Data Scientist and our main objective as such is to figure out – **How can we provide a unique and exceptional listening experience to our users?**

In order to answer that question, we need to gain a better understanding of the music choices of our current users so we can optimize our platform to not only expose our users to new content, but content they would actually be interested in. We are looking to gain Premium (paying) users and retain them for the long term.

Methodology

- Data Gathering
- Exploratory Data Analysis
- Hypothesis Testing
- Clustering
- Machine Learning

Photo by insung yoon on Unsplash



Our first step was to gather our data to analyze. We gathered data from two sources - the ‘Top Popular Songs from 2010-2019’ dataset from Kaggle and my personal Spotify data obtained using the Spotify API. These datasets included song features such as the track name and artist, beats per minute, energy, valence or mood, popularity, danceability, loudness, acousticness, speechiness etc.

Using that data we explored the relationships between these song features and performed some hypothesis testing. We then used various clustering algorithms to create ready-made playlists and lastly performed some machine learning classifications on our clustered data.

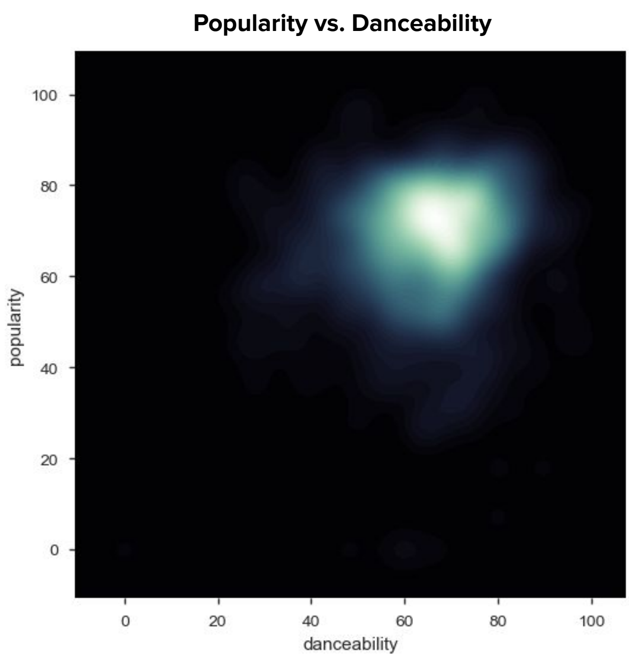
One of the main questions we were looking to investigate is – What makes a song popular? From analyzing the most popular songs charting, can we determine what song features influence popularity? Do these songs follow some sort of pattern?

1. **Does valence have a statistically significant effect on the popularity of a song?**

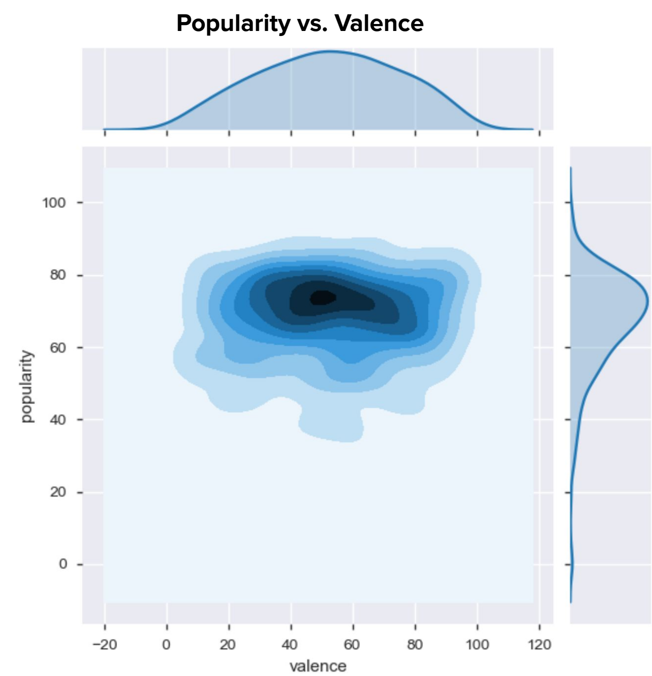
2. **Does danceability have a statistically significant effect on the popularity of a track? At what levels of danceability?**

We visualized the relationships between the different song factors and focused on two questions in particular for our hypothesis testing:

1. Does valence have a statistically significant effect on the popularity of a song?
2. Does danceability have a statistically significant effect on the popularity of a track? At what levels?



From our Kaggle popularity dataset, we could visualize that the most popular songs tended to be around a danceability score of 60-80, which is pretty high! So, we could see that in the past decade, people really liked a song they could bop along to. But what about the mood?



We could see here that charting songs tended to have a valence within the 50 – 60 range – this is slightly on the higher side, meaning the charting songs of the last decade were more happy in nature.

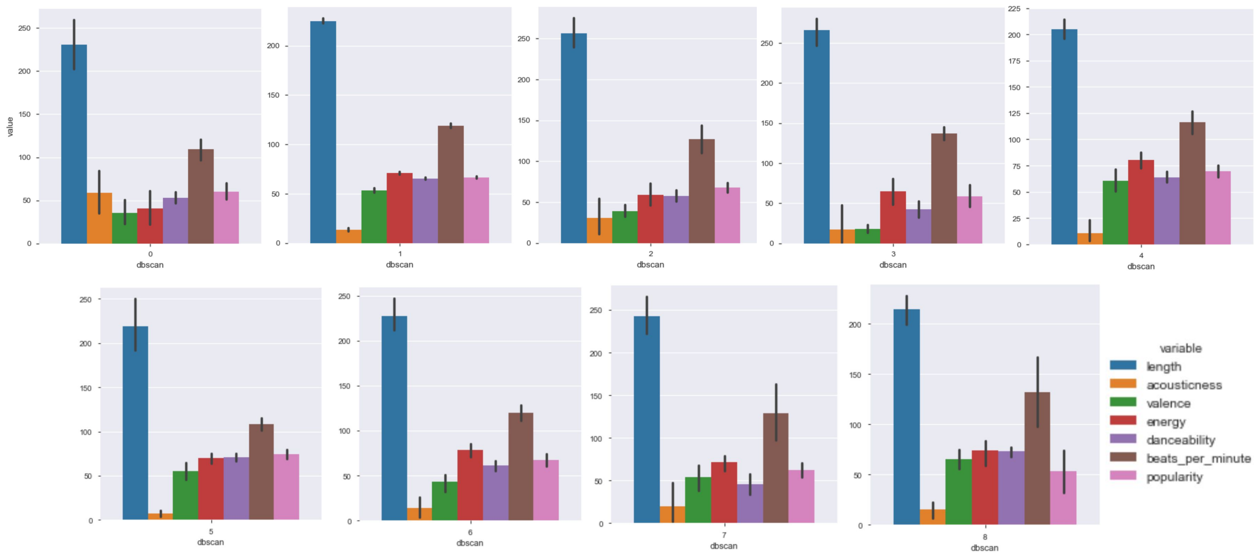
Hypothesis Testing Results

- Neither valence or energy seemed to influence the popularity of a song ($p > 0.05$)
- Outside features
  - Artist popularity
  - Artist social media presence
  - Big events
  - Record Company

We found that neither valence or energy seemed to influence the popularity of a song.

This may have to do with outside features that are not included in our data - such as the artist's popularity or their social media presence, if the artist appeared in any major events (coachella, super bowl, grammy's, etc), the influence of the record company of the artist, what age group is the majority of our listeners? More information is needed to determine an answer to our original question – What makes a song popular? These are things we may want to consider in order to better cater to our user base and also expand our user base.

Song Recommendation by Attributes



On the other hand, using data that we do have, we used different clustering algorithms to group songs with similar attributes together. Our dbscan method performed the best and we ended up with 9 distinct clusters shown here. Each color represents a song feature (as we see in the legend).

We can see that cluster 0 is characterized by high acousticness and low energy compared to the other clusters. Cluster 4 has very high danceability and energy.

Ready-Made Playlists

Click on the icons to see the playlists of some of our clusters!

Using these cluster labels, we created ready-made playlists so users can groove to songs that have a particular vibe they are looking for in the moment. This makes it easy for the user to stay engaged and improves their overall experience. By understanding the user's song selections, we can recommend other popular songs that have similar attributes that they would have a greater probability of liking.

Now let's have a look at some of our clusters in playlist form!

Predicting My Mood through Spotify

- 3 Clusters/Moods
  - Energetic, Chill, Cheerful
- 96.6% Accuracy

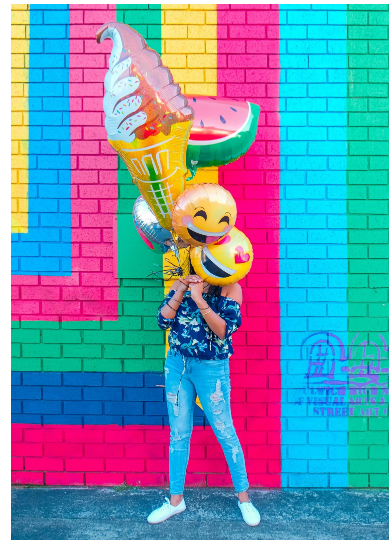


Photo by Lidya Nada on Unsplash

Now to get a different perspective, I had a look at my own personal Spotify data. Music is a medium that conveys emotion or a mood. People usually listen to songs that align with the mood that they are feeling. Using Kmeans clustering, I classified my tracks into 3 distinct clusters based on the key emotions that I associate with the majority of songs in a particular cluster: Energetic, Chill, and Happy. Using various machine learning models, our best model classified tracks with 96.6% accuracy! This could be extremely beneficial in the song recommendation process because users may be more receptive to new content that conveys the mood that they are feeling.

Business Recommendations

- Ready-Made playlists based on listener preference
- Song Recommendations on similar features
- Playlists for Mood

Photo by Mohammad Metri on Unsplash

Future Work

- Time series to see what genres are growing in popularity
- Include popularity ranking of artist
- Neural Network for song recommendation

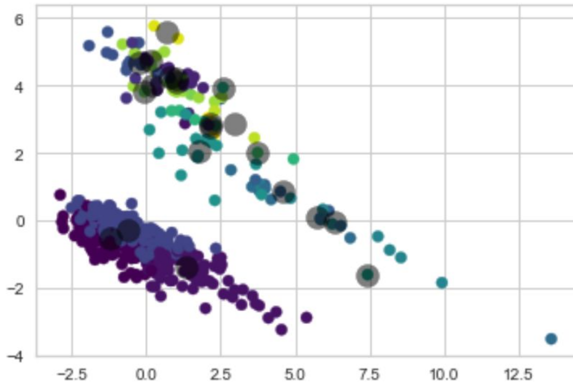
Future work – I would like to perform a time series analysis to see what genres are growing in popularity. I would also like to include popularity ranking and social media presence of the artists to our data. I would also like to create a neural network that would recommend songs and continually recommend relevant content as the users' taste changes over time.

Thank you

Thank you and happy listening!

Appendix:

K Means Clustering with all subgenres




```
Kmeans silhouette score:  0.214579002236143
Agglomerative Clustering silhouette score:  0.17505349846513404
DBScan silhouette score:  0.5384261496477445
```

The **silhouette** value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The **silhouette** ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

1. Does valence have a statistically significant effect on the popularity of a song?

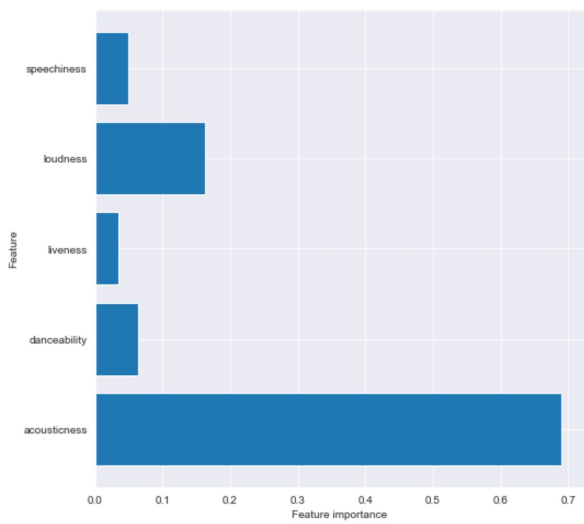
```
p-value:  0.3048121571837561
Failed to reject Null Hypothesis
```

2. Does danceability have a statistically significant effect on the popularity of a track? At what levels of danceability?

```
p-value:  0.19869512082605628
Failed to reject Null Hypothesis
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
3.0	4.0	2.9	0.9	-13.3767	19.1767	False
3.0	5.0	8.5978	0.6266	-5.8513	23.0469	False
3.0	6.0	7.175	0.7458	-6.3209	20.6709	False
3.0	7.0	8.5842	0.5295	-4.6901	21.8585	False
3.0	8.0	9.2401	0.4419	-4.1293	22.6094	False
3.0	9.0	12.1646	0.1933	-2.4656	26.7949	False
3.0	10.0	4.8611	0.9	-14.7948	24.517	False
3.0	nan	-58.25	0.0033	-104.6456	-11.8544	True
4.0	5.0	5.6978	0.8482	-6.2413	17.637	False
4.0	6.0	4.275	0.9	-6.491	15.041	False
4.0	7.0	5.6842	0.7263	-4.8026	16.1711	False
4.0	8.0	6.3401	0.6216	-4.2669	16.947	False
4.0	9.0	9.2646	0.3013	-2.8931	21.4224	False
4.0	10.0	1.9611	0.9	-15.9309	19.8531	False
4.0	nan	-61.15	0.0012	-106.8263	-15.4737	True
5.0	6.0	-1.4228	0.9	-9.1528	6.3072	False
5.0	7.0	-0.0136	0.9	-7.3499	7.3226	False
5.0	8.0	0.6422	0.9	-6.8647	8.1492	False
5.0	9.0	3.5668	0.9	-6.007	13.1406	False
5.0	10.0	-3.7367	0.9	-19.9839	12.5104	False
5.0	nan	-66.8478	0.001	-111.9052	-21.7904	True
6.0	7.0	1.4092	0.9	-3.8046	6.623	False
6.0	8.0	2.0651	0.9	-3.3863	7.5164	False
6.0	9.0	4.9896	0.582	-3.0739	13.0532	False
6.0	10.0	-2.3139	0.9	-17.7195	13.0917	False
6.0	nan	-65.425	0.001	-110.1858	-20.6642	True
7.0	8.0	0.6558	0.9	-4.2211	5.5328	False
7.0	9.0	3.5804	0.8697	-4.1065	11.2673	False
7.0	10.0	-3.7231	0.9	-18.935	11.4887	False
7.0	nan	-66.8342	0.001	-111.5287	-22.1397	True
8.0	9.0	2.9246	0.9	-4.9254	10.7745	False
8.0	10.0	-4.379	0.9	-19.6738	10.9159	False
8.0	nan	-67.4901	0.001	-112.2129	-22.7672	True
9.0	10.0	-7.3035	0.9	-23.712	9.1049	False
9.0	nan	-70.4146	0.001	-115.5304	-25.2988	True
10.0	nan	-63.1111	0.0011	-110.0978	-16.1244	True




```
{'Random Forest': 95.55555555555556,
 'KNN': 94.44444444444444,
 'Decision Tree': 90.0,
 'XGBoost': 93.33333333333333,
 'SVM': 96.66666666666667,
 'Naive Bayes': 93.33333333333333}
```

Model Accuracy Predicting My Spotify Mood