Analyzing Spotify Music Data

Sai Jaswanth Kumar Kunku

AIT 580

Abstract

This research aims to investigate the evolution of music over the years as well as understand the relationship between the descriptive features of song (e.g. *Danceability*[1] and *Valence*[2] etc.) and its combined effect on the song popularity. This study uses data obtained from different Spotify playlists and applies statistical analytics, exploratory analytics, predictive analytics to uncover insights on how the demand for music is changing from 1957 to 2020 and identify the factors that support the song popularity. The results shows that these audio features have moderate predictive power in determining the song popularity and there are other factors associated with it. In addition to that, It also revealed that there has been a huge leap in the production of music from 21st century and is dominated by the pop music. In this study, we take a further look into the results obtained from the data and discuss limitations and outline the findings.

Introduction

Music plays an important role in the wellbeing of the human life and has become a daily driver of our life. With the intervention of technology as well as the Internet physical storages such as discs and tapes have now become things of past. As a result, people now can access music of their choice that may have been unexplored in the form of recommendations and this unprecedent change has opened a new page in the big data analytics. This study focuses on the data and audio features calculated by Spotify which is one of the fastest growing music streaming platforms with more than 70 million songs and 381 million monthly active users across the world [1]. By utilizing music data from the Spotify, this paper will explore the following research questions

a) *What audio features describe the genre?*
b) *How is the popularity varying among different Genres?*
c) *How are the audio features correlated with each other and to what extent it determines the popularity?*
d) *How is the music industry changing over the time?*

Literature Review

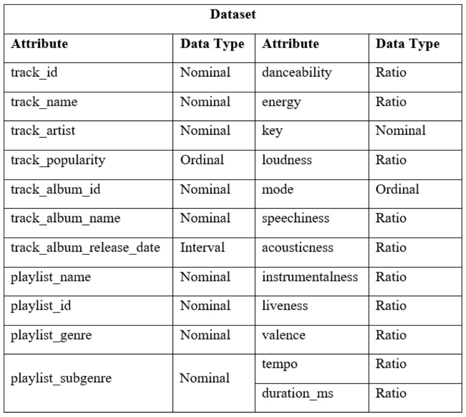
Although there hasn't been any extensive research on the Spotify data, there were few studies which focuses on various aspects of the subject and are relevant to my findings. For example, one study focused on the role of audio features on the overall popularity index, but they limited their research data to 19 albums of same artist which may introduce high bias and low variance (Sciandra et. al, 2020) [2]. Their findings showed that audio features such as *Speechiness*[3], *Instrumentalness*[4], *Liveness*[5] show a negative effect on the Popularity Index whereas *Tempo*[6] and Danceability charts do not show any effect at all.

In second study, (John, S. 2021) [3] analyzed the success of an albums by considering the similarities among popular albums, top albums of each year, audio features , the number of songs, duration, genre, and the popularity level of those albums using K-means clustering. The results showed that people like songs which have high danceability, *energy*[7], *loudness*[8] and dislike songs having *acousticness*[9] and speechiness. Although this study gives great insights on the music preferences of public the analysis is limited to only 21st century. It would be interesting to analyze the data before this period as well to find interesting patterns.

In the final study, (Bardhoshi, D. 2021) [4]. discovered the trends between subgroups of music by dividing on things like popularity or year it has been released. He classified the data into two groups. The first group contains data of top 100 songs which are popular, and second group contains the rest of the songs. While the analysis of relationship between audio features pretty much aligns with one of my research questions, the author in this study divided the data in groups i.e., top 100 popular songs into one group and 100 random songs from rest of the. In this study I am going to find the relationships by querying on the entire dataset which would give more accurate results.

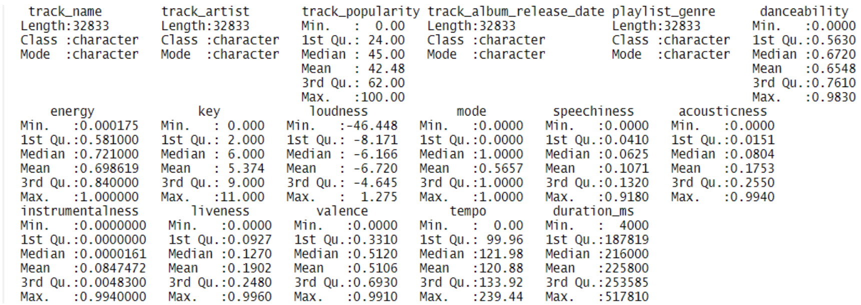
Tools and Methodologies

For the analyses presented in this work, the dataset was retrieved from a publicly available GitHub repository. The dataset [5] consists of 32,833 tracks from 473 playlists which were released between 1957 and 2020 with 23 attributes such as track popularity, playlist genre,duration,release date and other audio features etc. Table 1 shows the attributes present in the dataset and the data type they come under such as nominal, ordinal, interval, and ratio.



Dataset			
Attribute	**Data Type**	**Attribute**	**Data Type**
track_id	Nominal	danceability	Ratio
track_name	Nominal	energy	Ratio
track_artist	Nominal	key	Nominal
track_popularity	Ordinal	loudness	Ratio
track_album_id	Nominal	mode	Ordinal
track_album_name	Nominal	speechiness	Ratio
track_album_release_date	Interval	acousticness	Ratio
playlist_name	Nominal	instrumentalness	Ratio
playlist_id	Nominal	liveness	Ratio
playlist_genre	Nominal	valence	Ratio
playlist_subgenre	Nominal	tempo	Ratio
		duration_ms	Ratio

Table 1:Information about the Dataset



```
  track_name         track_artist        track_popularity track_album_release_date playlist_genre      danceability
 Length:32833       Length:32833       Min.   :  0.00   Length:32833             Length:32833       Min.   :0.0000
 Class :character   Class :character   1st Qu.: 24.00   Class :character         Class :character   1st Qu.:0.5630
 Mode  :character   Mode  :character   Median : 45.00   Mode  :character         Mode  :character   Median :0.6720
                                       Mean   : 42.48                                               Mean   :0.6548
                                       3rd Qu.: 62.00                                               3rd Qu.:0.7610
                                       Max.   :100.00                                               Max.   :0.9830
     energy              key            loudness            mode         speechiness      acousticness
 Min.   :0.000175   Min.   : 0.000   Min.   :-46.448   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.581000   1st Qu.: 2.000   1st Qu.: -8.171   1st Qu.:0.0000   1st Qu.:0.0410   1st Qu.:0.0151
 Median :0.721000   Median : 6.000   Median : -6.166   Median :1.0000   Median :0.0625   Median :0.0804
 Mean   :0.698619   Mean   : 5.374   Mean   : -6.720   Mean   :0.5657   Mean   :0.1071   Mean   :0.1753
 3rd Qu.:0.840000   3rd Qu.: 9.000   3rd Qu.: -4.645   3rd Qu.:1.0000   3rd Qu.:0.1320   3rd Qu.:0.2550
 Max.   :1.000000   Max.   :11.000   Max.   :  1.275   Max.   :1.0000   Max.   :0.9180   Max.   :0.9940
 instrumentalness       liveness          valence             tempo         duration_ms
 Min.   :0.0000000   Min.   :0.0000   Min.   :0.0000   Min.   :  0.00   Min.   :  4000
 1st Qu.:0.0000000   1st Qu.:0.0927   1st Qu.:0.3310   1st Qu.: 99.96   1st Qu.:187819
 Median :0.0000161   Median :0.1270   Median :0.5120   Median :121.98   Median :216000
 Mean   :0.0847472   Mean   :0.1902   Mean   :0.5106   Mean   :120.88   Mean   :225800
 3rd Qu.:0.0048300   3rd Qu.:0.2480   3rd Qu.:0.6930   3rd Qu.:133.92   3rd Qu.:253585
 Max.   :0.9940000   Max.   :0.9960   Max.   :0.9910   Max.   :239.44   Max.   :517810
```

Figure 1:Summary of the Dataset

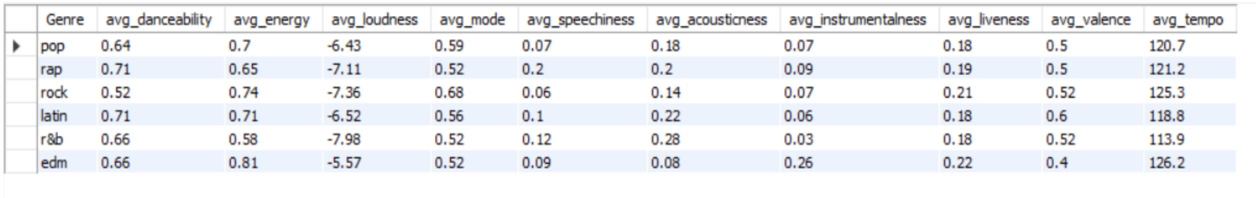
The dataset for this project was not readily downloadable as it is available in raw format. Hence it was retrieved from URL using R and written down as CSV file for preprocessing and analysis. The first step in this study's methodology was to clean the data, which was primarily done using R. There were quite many columns in the dataset which are not necessary for this research as they don't help in formulating the results and these were removed.

Next, the data was checked to find any missing data fields and it showed up 10 missing values in field track name. This was addressed by omitting all these rows. In addition to that, this data consists of same tracks with different versions such as remix, edit etc. that causes data duplicity and inaccurate results. This was taken care by eliminating the stop words present in the tracks using *stringr* library. Finally, to analyze time series data, a new data field named year has been appended to existing dataset after extracting it from the release date column using *lubridate* library. All these steps lead me bring down the number of data items from 32,833 to 24,671 which is quite significant difference considering the size of data.

After the data was cleaned, all the statistical summaries, correlations and the visualizations related to the research questions were done using R whereas Python was used to prepare the data for applying them on the visualizations supporting the research questions. In addition to that, MySQL was used to analyze fetch records from the dataset based on frequencies and percentages.

Results

i) Determining song characteristics



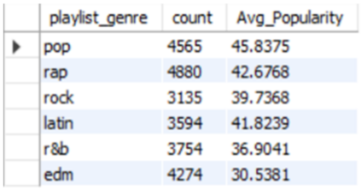
	Genre	avg_danceability	avg_energy	avg_loudness	avg_mode	avg_speechiness	avg_acousticness	avg_instrumentalness	avg_liveness	avg_valence	avg_tempo
▶	pop	0.64	0.7	-6.43	0.59	0.07	0.18	0.07	0.18	0.5	120.7
	rap	0.71	0.65	-7.11	0.52	0.2	0.2	0.09	0.19	0.5	121.2
	rock	0.52	0.74	-7.36	0.68	0.06	0.14	0.07	0.21	0.52	125.3
	latin	0.71	0.71	-6.52	0.56	0.1	0.22	0.06	0.18	0.6	118.8
	r&b	0.66	0.58	-7.98	0.52	0.12	0.28	0.03	0.18	0.52	113.9
	edm	0.66	0.81	-5.57	0.52	0.09	0.08	0.26	0.22	0.4	126.2

Table 2:Average song characteristic criteria of Genre

The above table explains that EDM songs has an average valence of 0.4, average loudness of -5.57 (lesser is more loud) ,average energy of 0.81 and instrumentals of 0.26 which is highest among other genres. In the next place, rap songs has an average speechiness,acousticness value of 0.2.While many values for Rap and Latin Genre overlap each other the Latin songs have less tempo

(118.8) and less instrumentalness (0.06) compared to Rap.R&b has an average accoustiness of 0.28 and tempo of 113.9 which explains that a track with accousticness and very low tempo can fall into this category.Comparing the values of pop with others Genre values it looks like pop orginated from all of these Genres.

ii) Statistical results between popularity and Genre



	playlist_genre	count	Avg_Popularity
▶	pop	4565	45.8375
	rap	4880	42.6768
	rock	3135	39.7368
	latin	3594	41.8239
	r&b	3754	36.9041
	edm	4274	30.5381

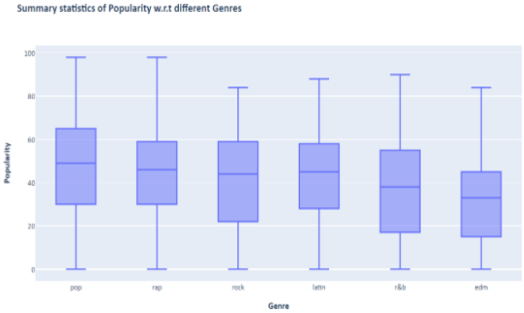
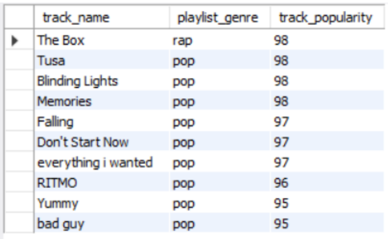


Table 3: Average song popularity for each genre

Figure 2:Box plot visualization of the song popularity



	track_name	playlist_genre	track_popularity
▶	The Box	rap	98
	Tusa	pop	98
	Blinding Lights	pop	98
	Memories	pop	98
	Falling	pop	97
	Don't Start Now	pop	97
	everything i wanted	pop	97
	RITMO	pop	96
	Yummy	pop	95
	bad guy	pop	95

Table 4:Top 10 trending songs

The results from table 3 and figure 2 explains that pop music has average popularity of 45.8% which is highest among. It can also be seen from the table 3 that all the top 10 trending songs is coming from Pop genre. In contrary. EDM music is least popular with an average popularity of 30.5% despite having more no of songs. It should also be noticed that Rap and Latin music are close in terms of average popularity although Latin music has less no of songs. Finally, rock and r&b music have an average popularity of 39.7% and 36% respectively which is decent.

iii) Popularity vs Genre

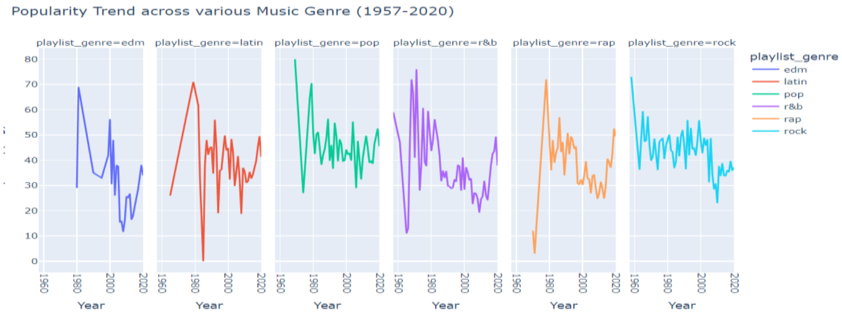


Figure 4:Line Graph visualization of Popularity Trend

Although the previous gave an overall estimate of popularity it would be clearly understood if we analyze it over the period. From figure 4. it can be seen that edm, rap and r&b music is losing its popularity from 1980 however they are gaining it ever since 2010.It is also important to notice that rock music which has been consistent over the years is not doing great in 21st century. Based on the above results it can be understood that music production companies are focusing on respective genres in the recent times based on the popularity obtained for the previous albums.

iv) Finding the relation between song Features

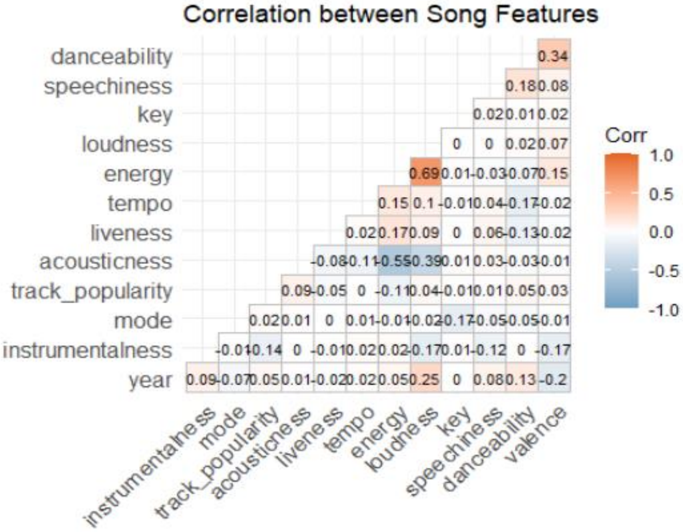
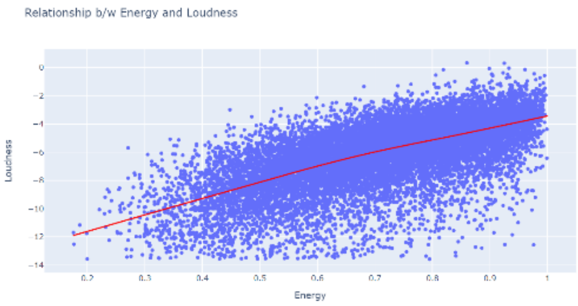


Figure 5: Correlation matrix of songs features



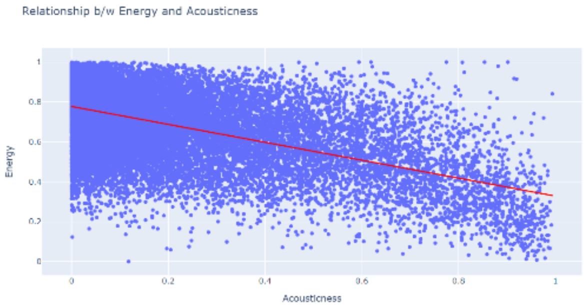
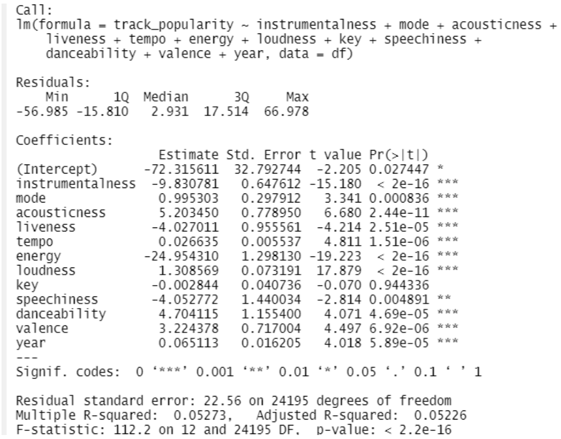


Figure 6: Scatterplot visualization of Energy & Loudness   Figure 7: Scatterplot visualization of Energy vs Acousticness

From the above correlation matrix it is understood that almost all the features expect danceability, energy, acoustincness , tempo and valance doesn’t show any relationship with each other. Since energy has a positive correlation of 0.69 with loudness which explains that increase in energy is causing more loudness. On the other hand there is a negative correlation of 0.55 between acousticness and energy which explains that the more acoustic the song it is less energetic .One interesting fact is that is that there is positive correlation of 0.25 between year and loudness. This explains that there is a significant increase loudness of music. In addition to that there is a negative correlation of 0.2 between year and valence which states that over the years sadder songs are being released.

v) Fitting a Linear model on the song features



```
Call:
lm(formula = track_popularity ~ instrumentalness + mode + acousticness +
    liveness + tempo + energy + loudness + key + speechiness +
    danceability + valence + year, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-56.985 -15.810   2.931  17.514  66.978

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -72.315611  32.792744  -2.205 0.027447 *
instrumentalness  -9.830781   0.647612 -15.180  < 2e-16 ***
mode               0.995303   0.297912   3.341 0.000836 ***
acousticness       5.203450   0.778950   6.680 2.44e-11 ***
liveness          -4.027011   0.955561  -4.214 2.51e-05 ***
tempo              0.026635   0.005537   4.811 1.51e-06 ***
energy           -24.954310   1.298130 -19.223  < 2e-16 ***
loudness           1.308569   0.073191  17.879  < 2e-16 ***
key               -0.002844   0.040736  -0.070 0.944336
speechiness       -4.052772   1.440034  -2.814 0.004891 **
danceability       4.704115   1.155400   4.071 4.69e-05 ***
valence            3.224378   0.717004   4.497 6.92e-06 ***
year               0.065113   0.016205   4.018 5.89e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.56 on 24195 degrees of freedom
Multiple R-squared:  0.05273,   Adjusted R-squared:  0.05226
F-statistic: 112.2 on 12 and 24195 DF,  p-value: < 2.2e-16
```

Figure 8: Fitting Linear model on audio features

To determine whether these features determine the popularity I have applied a linear model with popularity as predictor and the rest as my response variables .The linear model in figure 8 explains that *mode*[10] and speechiness are not explaining the track popularity as compared to other variables and can be ignored. It is also evident that *key*[11] is no predict power and it can understand from the fact that it is a nominal variable. After removing these three variables and fitting the linear model with 70% train data and testing on remaining 30% ,min max accuracy of 61.10% was obtained. Therefore it explains that there are additional factors in determining the popularity of a song besides these values.

vi) Song Duration vs Year

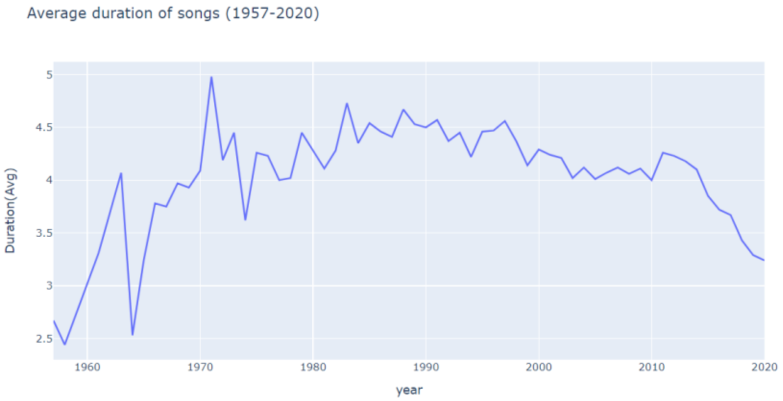


Figure 9: Line graph visualization of song duration (1957-2020)

To get an idea on the evolution of music with the time ,we will compare the differences in the values of song features, average song duration and total no of songs . The results from figure 9 shows that the average duration of songs has been increasing gradually from 1957 to 1971 and has been inconsistent. From 1988, the average duration of songs started to decrease gradually from 4.67 seconds to 4.29 before the beginning of 21st century. From 2000-2010 while the average song duration lasted around 4 minutes it came to 3.3 seconds in 2020.

vii) Differences in song features over the years

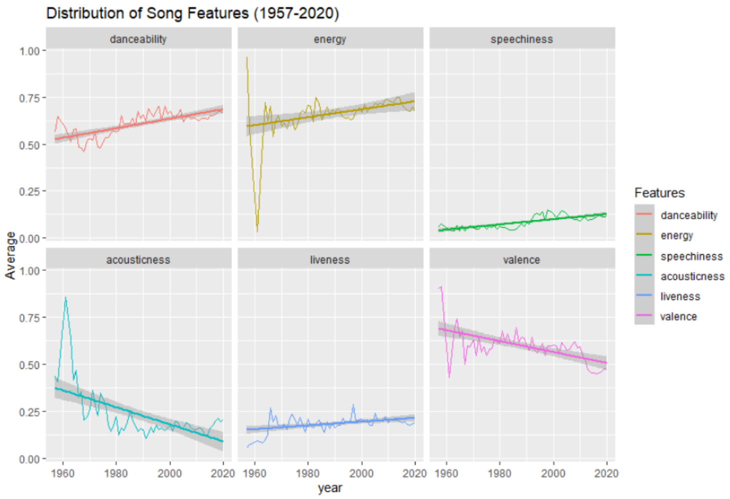
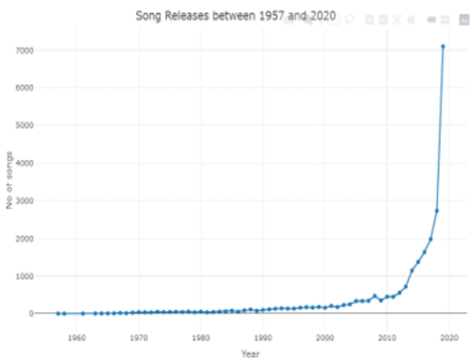


Figure 10: Line graph visualization of song features (1957-2020)

The above line graph illustrates that most of tracks that were composed in 21st century have almost similar values that of the previous years except Valance and Acousticness. The interesting fact is that the values of Danceability , Liveness and Speechiness tend to be constant over the years. On the other hand there is an increase in accousticness and decrease in valance in the music from 2010 and 2014 respectively. From the results we can conclude that more acoustic and sad musical tracks are being produced in the recent years.

viii) Determining the changes in song Distribution



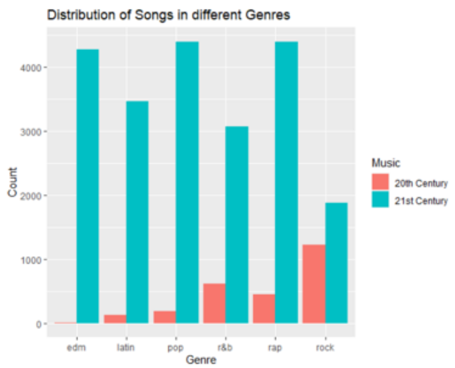


Figure 11: Line graph visualization of song features (1957-2020)

Figure 12: Bar graph visualization of song Distribution

The results from the line graph in figure 10 shows that number of songs have been significantly increasing over the years. In 21st Century alone, there has been a sharp increase in the no of songs reaching more than 7000. The bar graph distribution shows that pop, rap and EDM music are doing great in 21st century followed by Latin, music. Although rock and r&b music didn't do well in the past it outperformed other genres in 21st century.

Limitations

One of the main limitations I noticed when I started working on this research is the difficulty in finding a perfectly balanced dataset. All the datasets available online related to the study are mostly unbalanced as the data was obtained from various user playlists and not everyone can listen to all the available tracks.

In one of my findings, I was able to predict the song popularity using the song features with an accuracy of 61%.Although these features were able to explain popularity to a certain extent had there been few additional variables like the number of streams, revenue it is making based on the new subscriptions it would have to better accuracy. In addition to that if we had access to lyrics of the song, we could have determined the association of popularity with the lyrics.

Conclusion

The results obtained in this study through the analysis helped me preprocess large amounts of data, explore various and uncover various insights from the musical data. Through my findings the following research questions has been addressed.

i) What audio features describe the genre?

Music coming from EDM genre is making the audience happy, energetic, and engaged with its high instrumentalness, low valence and high energy values. While Rap and Latin music share similar characteristics in many aspects the key differences between these two are that Latin music lacks the tempo and uses fewer musical instruments whereas Rap music sounds more acoustic and contains more words. The Pop music which originated in the mid 1950's share all the features of other genres, and this is the reason why for increase in the popularity.

ii) How is the popularity varying among different Genres?

The Pop music which originated in the mid 1950's has been consistent among the years and is gaining popularity in the recent times and is trending in top. On the flipside, rock music which did well in the past is losing its popularity in 21st century. Latin and Rap has been inconsistent, but they are holding the top musical charts with its ever-increasing popularity. Although r&b and EDM music faced tough time in the early 21st century they started doing great in the market from the last decade.

iii) How are the audio features correlated with each other and to what extent it determines the popularity?

Most of the audio features doesn't show any correlation with other features expect danceability, energy, acousticness and tempo. The increase in the loudness is show is showing a positive effect on the energy. With the increase in the acousticness of music there will be a significant drop in the energy. It should be noticed that year by year the loudness is music is changing and music is more melancholic. Finally ,features such as mode, key and speechiness doesn't have any effect in determining the popularity.

iv) How is the music industry is changing over the time?

More number of tracks are being produced every year and judging by the no of songs in 21st century we can say that it is a good time for music artists. The average duration of songs decreased by a min from 4:30mins to 3:30 over the period. Features such as danceability ,energy, speechiness and liveness are showing a positive trend which is positive sign as it beneficial to the music lovers whereas valance is showing the negative trend which is not good to public as these songs may result in bad mood.

References

[1] Wikipedia contributors. (2021b, October 29). *Spotify*. Wikipedia. Retrieved December 4, 2021, from https://en.wikipedia.org/wiki/Spotify#Technical_information

[2] Sciandra, Mariangela & Spera, Irene. (2020). *A Model Based Approach to Spotify Data Analysis: A Beta GLMM*. Retrieved December 4,SSRN Electronic Journal. 10.2139/ssrn.3557124

[3] John, S. (2021, May 12). *Spotify Music Data Analysis - Web Mining [IS688, Spring 2021]*. Medium. Retrieved December 4, 2021, from https://medium.com/web-mining-is688-spring-2021/spotify-music-data-analysis-ed3d235023f2

[4] Bardhoshi, D. (2021, June 11). *Visualizing Spotify Songs with Python: an exploratory data analysis*. Medium. Retrieved December 4, 2021, from https://towardsdatascience.com/visualizing-spotify-songs-with-python-an-exploratory-data-analysis-fc3fae3c2c09

[5] Thomas Mock, & Kaylin Pavlik. (2020, January 20). *RforDataScience*. Github. [Data file - Raw Format], from https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-01-21

Glossary

1.**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. A value of 0.0 is least danceable and 1.0 is most danceable.

2. **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).

3. **Speechiness –** A measure that 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.

4. **Instrumentalness**-It is a measure which predicts whether a track contains no vocals. 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.

5. **Liveness -**It is a measure which 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 strong likelihood that the track is live.

6. **Tempo -**In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

7. **Energy -** It is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

8. **Loudness -**The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing 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.

9. **Acousticness**-It is defined as confidence measure from 0.0 to 1.0 of where 1.0 represents high confidence the track is acoustic.

10. **Mode -** It indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.

11. **Key -** Key is the main group of pitches, or notes, that form the harmonic foundation of a piece of music. In the dataset the key values ranges from 1-11. If no key was detected, the value is -1.