Spotify Track Analysis :

1. Introduction

The music industry has undergone a significant transformation over the past decade, primarily due to the emergence of streaming services. One such service that has had a profound impact on the music industry is Spotify, which has changed the way people listen to music. With a vast collection of songs spanning multiple genres and over 356 million active users, Spotify has become a music streaming powerhouse.

This data analysis project aims to examine the Spotify Tracks Dataset from Kaggle to answer the following critical business questions:

1. What are the duration times of different music genres?
2. How can we describe the sound of each music genre?
3. Are there any significant variations in the characteristics of popular songs across different music genres?

To accomplish our objectives, we will begin with an overview of the dataset, followed by an analysis of the three business questions listed above. We will then conclude our report with a summary of our findings and recommendations for further analysis.

2. Data Cleaning

2.1 Dataset

The original dataset has 114 000 rows and 21 columns. The dataset had a considerable number of duplicate values. Moreover, we observed numerous inconsistent values, such as inconsistent naming conventions for track_id, artists, album_name, and track_name .

```
SELECT * FROM Spotify_tracks ;
```

track_id	artists	album_name	track_name	popularity	duration_ms	explicit	danceability	energy	key_	loudness	mode	speechiness
5SuOikwiRyPMVoIQDJUgSV	Gen Hoshino	Comedy	Comedy	73	230666	FALSE	0.676	0.461	1	-6.746	0	0.143
4qPNDBW1i3p13qLCt0Ki3A	Ben Woodward	Ghost (Acoustic)	Ghost - Acoustic	55	149610	FALSE	0.42	0.166	1	-17.235	1	0.0763
1iJBSr7s7jYXzM8EGcbK5b	Ingrid Michaelson;ZAYN	To Begin Again	To Begin Again	57	210826	FALSE	0.438	0.359	0	-9.734	1	0.0557
6lfxq3CG4xtTiEg7opyCyx	Kina Grannis	Crazy Rich Asians (Original Motion Picture Soun...	Can't Help Falling In Love	71	201933	FALSE	0.266	0.0596	0	-18.515	1	0.0363
5vjLSffimiIP26QG5WcN2K	Chord Overstreet	Hold On	Hold On	82	198853	FALSE	0.618	0.443	2	-9.681	1	0.0526
01MVOl9KtVTNfFiBU9I7dc	Tyrone Wells	Days I Will Remember	Days I Will Remember	58	214240	FALSE	0.688	0.481	6	-8.807	1	0.105
6Vc5wAMmXdKIAM7WUoEb7N	A Great Big World;Christina Aguilera	Is There Anybody Out There?	Say Something	74	229400	FALSE	0.407	0.147	2	-8.822	1	0.0355
1EzrEOXmMH3G43AXT1y7pA	Jason Mraz	We Sing. We Dance. We Steal Things.	I'm Yours	80	242946	FALSE	0.703	0.444	11	-9.331	1	0.0417
0IktbUcnAGrvD03AWnz3Q8	Jason Mraz;Colbie Caillat	We Sing. We Dance. We Steal Things.	Lucky	74	189613	FALSE	0.625	0.414	0	-8.7	1	0.0369
7k9GuJYLp2AzqokyEdwEw2	Ross Copperman	Hunger	Hunger	56	205594	FALSE	0.442	0.632	1	-6.77	1	0.0295
4mzP5mHkRvGxdhdGdAH7EJ	Zack Tabudlo	Episode	Give Me Your Forever	74	244800	FALSE	0.627	0.363	8	-8.127	1	0.0291
5ivF4eQBqJiVLSIAE9jRyl	Jason Mraz	Love Is a Four Letter Word	I Won't Give Up	69	240165	FALSE	0.483	0.303	4	-10.058	1	0.0429
4ptDJbJl35d7gQfeNteBwp	Dan Berk	Solo	Solo	52	198712	FALSE	0.489	0.314	7	-9.245	0	0.0331
0X9MxHR1rTkEHDjp95F2OO	Anna Hamilton	Bad Liar	Bad Liar	62	248448	FALSE	0.691	0.234	3	-6.441	1	0.0285
4LbWtBkN82ZRhz9jqzgrb3	Chord Overstreet;Deepend	Hold On (Remix)	Hold On - Remix	56	188133	FALSE	0.755	0.78	2	-6.084	1	0.0327
1KHdq8NK9QxnGjdXb55NiG	Landon Pigg	The Boy Who Never	Falling in Love at a Coff...	58	244986	FALSE	0.489	0.561	4	-7.933	1	0.0274
6xKeQgzfjixSUld14qUezm	Andrew Foy;Renee Foy	ily (i love you baby)	ily (i love you baby)	56	129750	FALSE	0.706	0.112	2	-18.098	1	0.0391

2.2 Removing inconsistent character (Data Validation)

We noticed that the data contained some inconsistent characters in the track_id ,artists , album_name and track_name columns . The possible reason for this is that the Unicode characters in the original data were not valid UTF-8 characters, even though they were converted to the appropriate CSV format before being loaded into the MySQL database. To deal with this issue, we used a different character set or collation that is compatible with the data.

The following query first converts the column to the 'latin1' character set to handle any encoding inconsistencies. It then converts the result to a binary format and converts it again to the 'utf8mb4' character set, specifying the 'utf8mb4_general_ci' collation. This should replace any Unicode characters with their correct Unicode equivalents

```sql
SELECT track_id , CONVERT(CAST(CONVERT(track_id USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci
from spotify_tracks ;
SELECT artists , CONVERT(CAST(CONVERT(artists USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci AS artists_correction
from spotify_tracks ;
SELECT album_name , CONVERT(CAST(CONVERT(album_name USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci
from spotify_tracks ;
SELECT track_name , CONVERT(CAST(CONVERT(track_name USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci
from spotify_tracks ;

-- update table --
UPDATE spotify_tracks
SET track_id = CONVERT(CAST(CONVERT(track_id USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci;
UPDATE spotify_tracks
SET artists = CONVERT(CAST(CONVERT(artists USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci;
UPDATE spotify_tracks
SET album_name = CONVERT(CAST(CONVERT(album_name USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci;
UPDATE spotify_tracks
SET  track_name = CONVERT(CAST(CONVERT( track_name USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci;
```

artists	artists_correction
Shirley Carvalhaes	Shirley Carvalhaes
Jhonas Serra	Jhonas Serra
Kiara VitÃ³ria;Kellen Byanca	Kiara Vitória;Kellen Byanca
Samuel Messias;Midian Lima	Samuel Messias;Midian Lima
Dazaranha	Dazaranha
Um Barril de Rap	Um Barril de Rap
Matanza	Matanza
Strike	Strike
AndrÃ© ValadÃ£o	André Valadão
Sarah Farias	Sarah Farias
Catedral	Catedral
BarÃ£o Vermelho	Barão Vermelho
ZÃ© Ramalho	Zé Ramalho
MinistÃ©rio Vineyard	Ministério Vineyard
Renato Russo; 14 Bis	Renato Russo; 14 Bis
Frei Gilson	Frei Gilson
Lauriete	Lauriete
VOLAC;illusionize;Andre Longo	VOLAC;illusionize;Andre Longo

2.3 Populating missing columns

We populated missing columns in the dataset, at first look it seems that, country, cast and director are the most affected columns, so we start by populating the Table using the description column,

```
SELECT t1.show_id, t1.type, t1.title, t1.director, t1.cast, t1.country, t1.date_added, t1.release_year, t1.rating, t1.duration, t1.listed_in, t1.description
FROM netflix AS t1
INNER JOIN (
    SELECT MIN(show_id) AS min_show_id, description, director, cast, country
    FROM netflix
    WHERE director IS NOT NULL AND cast IS NOT NULL AND country IS NOT NULL
    GROUP BY description, director, cast, country
) AS t2 ON t1.description = t2.description
WHERE t1.director IS NULL AND t1.cast IS NULL AND t1.country IS NULL;

UPDATE netflix AS t1
INNER JOIN (
    SELECT MIN(show_id) AS min_show_id, description, director, cast, country
    FROM netflix
    WHERE director IS NOT NULL AND cast IS NOT NULL AND country IS NOT NULL
    GROUP BY description, director, cast, country
) AS t2 ON t1.description = t2.description
SET t1.director = t2.director, t1.cast = t2.cast, t1.country = t2.country
WHERE t1.director IS NULL AND t1.cast IS NULL AND t1.country IS NULL;
```

then we populate the country using the director column.

```
SELECT nt2.title ,COALESCE(nt.country,nt2.country) AS new_country
FROM netflix  AS nt
JOIN netflix AS nt2
ON nt.director = nt2.director
AND nt.show_id <> nt2.show_id
WHERE nt.country IS NULL ;

update netflix AS nt
JOIN netflix  nt2
ON nt.director = nt2.director
AND nt.show_id <> nt2.show_id
SET nt.country = coalesce(nt.country,nt2.country)
WHERE nt.country IS NULL ;
```

director	cast	country
Antoine Fuqua	Denzel Washington, Ethan Hawke, Scott Glenn,...	United States
Unknown director	Richard Mofe-Damijo, Dakore Akande, Bimbo M...	Nigeria
Unknown director	Kashmira Irani, Chandan Anand, Dinesh Mehta,...	India
Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Mieko Harad...	Japan
Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Koji Tsujitani...	Japan
Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Koji Tsujitani...	Japan
Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Koji Tsujitani...	Japan
Unknown director	Nicole Byer, Jacques Torres	United States
Masahiko Murata	Junko Takeuchi, Chie Nakamura, Rikiya Koyama...	Japan
Hajime Kamegaki	Junko Takeuchi, Chie Nakamura, Noriaki Sugiya...	Japan

2.4 Treating the Duplicates

We conducted a count of duplicate records in the dataset and discovered that 422 rows had identical values for track_id, artists, album_name, track_name, and track_genre. As a result, we removed these duplicate rows from the dataset.

```
SELECT track_id, artists, album_name, track_name, track_genre  ,count(*)
FROM spotify_tracks
GROUP BY track_id, artists, album_name, track_name, track_genre
HAVING COUNT(*) > 1;
```

track_id	artists	album_name	track_name	track_genre	count(*)
73tHDm0hBREbIJU4NrbGKy	Wolfgang Amadeus Mozart...	Mozart: A Night of Classics	Sonata for Piano and Violin in E flat, K.481: 3f. ...	classical	2
1NoDZgAwhDmLXBdfxaFPVM	Wolfgang Amadeus Mozart...	Mozart: A Night of Classics	Sonata for Piano and Violin in A, K.305: 2e. Vari...	classical	2
7zo05whm8LBnnt1dG979ov	Wolfgang Amadeus Mozart...	Mozart - A Classical Dawn	Six German Dances, K.571: No. 5 in B Flat Major	classical	2
4GdCiwOXIpmQGUWl5XBqyr	Wolfgang Amadeus Mozart...	Mozart - Inspiring Classics	6 Variations in G minor for Piano & Violin on "Hél...	classical	2
2bNeq7dW6z9umgGfHGLdcL	Wolfgang Amadeus Mozart...	Mozart - Inspiring Classics	Symphony No. 33 in B-Flat Major, K. 319: 2b. A...	classical	2
5SBm6jz3bNhnbgerw5sJ9g	Wolfgang Amadeus Mozart...	Mozart - Inspiring Classics	8 Variations on 'Dieu d'amour' from 'Les mariage...	classical	2
040CvqW0wo3b77EBRfgC14	Wolfgang Amadeus Mozart...	Mozart - A Classical Dawn	Andante and allegretto for Piano and Violin in C,...	classical	2
2WvKORgp0fzCJAAvG8cN3K	Wolfgang Amadeus Mozart...	Mozart - A Classical Dawn	Symphony No. 25 in G minor, K.183: 1. Allegro ...	classical	2
1ghv8smG2wTI2Rhs2eghih	Wolfgang Amadeus Mozart...	Mozart: A Night of Classics	Piano Sonata No.14 in C Minor, K. 457: 2c. Ada...	classical	2
1SZp7sIqzHHh1YMaMu8FL2	Wolfgang Amadeus Mozart...	Mozart - A Classical Dawn	12 Variations on an Allegretto in B Flat, K.500: ...	classical	2
227o6yibgqhCnZ4i0w77CE	Wolfgang Amadeus Mozart...	Mozart - A Classical Dawn	Fantasia in C minor, K.475: Allegro	classical	2

3. Exploratory Data Analysis

3.1 Analyzing the duration time: Horizontal Bar Chart

The mean duration time of tracks in each genre ranges from 133,854.66 seconds for grindcore to 371,933.56 seconds for Detroit techno. The standard deviation is highest for breakbeat at 214,310.03 seconds, while the lowest standard deviation is for honky-tonk at 33,044.72 seconds.

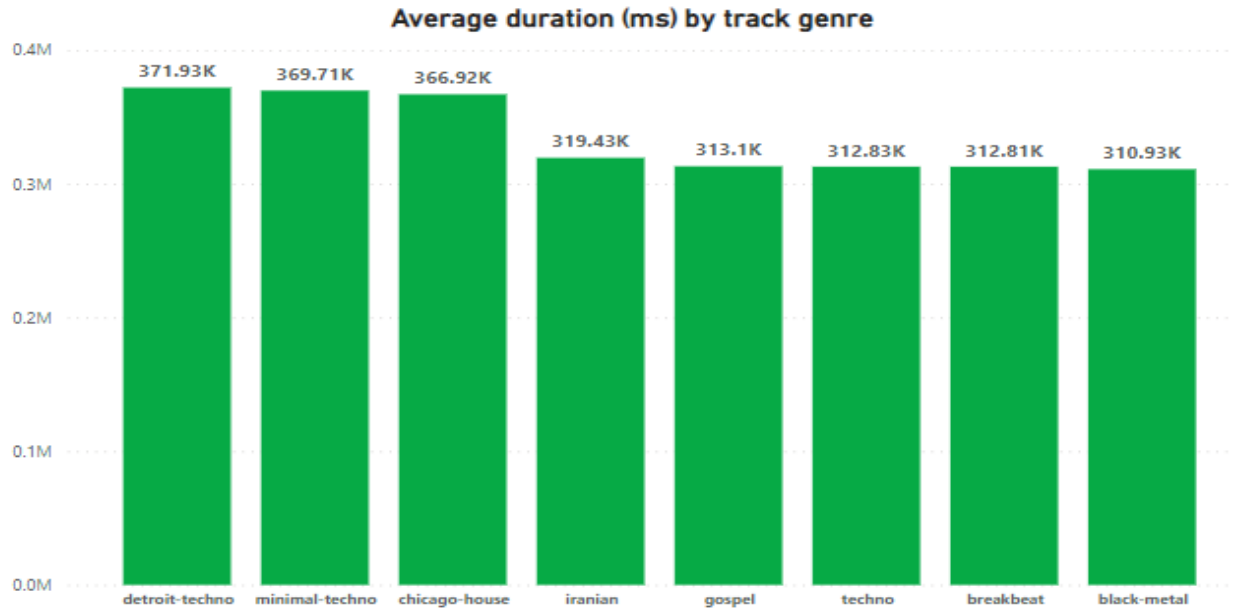
The median duration time for tracks ranges from 111,072.50 seconds for grindcore to 362,027.30 seconds for Detroit techno. The median duration time is an indicator of the typical duration for a track, and it may be less influenced by outliers than the mean duration time.

There are some interesting insights we can draw from this data. Genres that are often associated with dance, such as house, electronic, and techno, tend to have a lower mean and median duration time, indicating that they are often composed with the intention of being played in a continuous mix. On the other hand, genres such as black metal and industrial tend to have longer mean and median duration times, suggesting that these tracks may be more complex or layered than other genres.

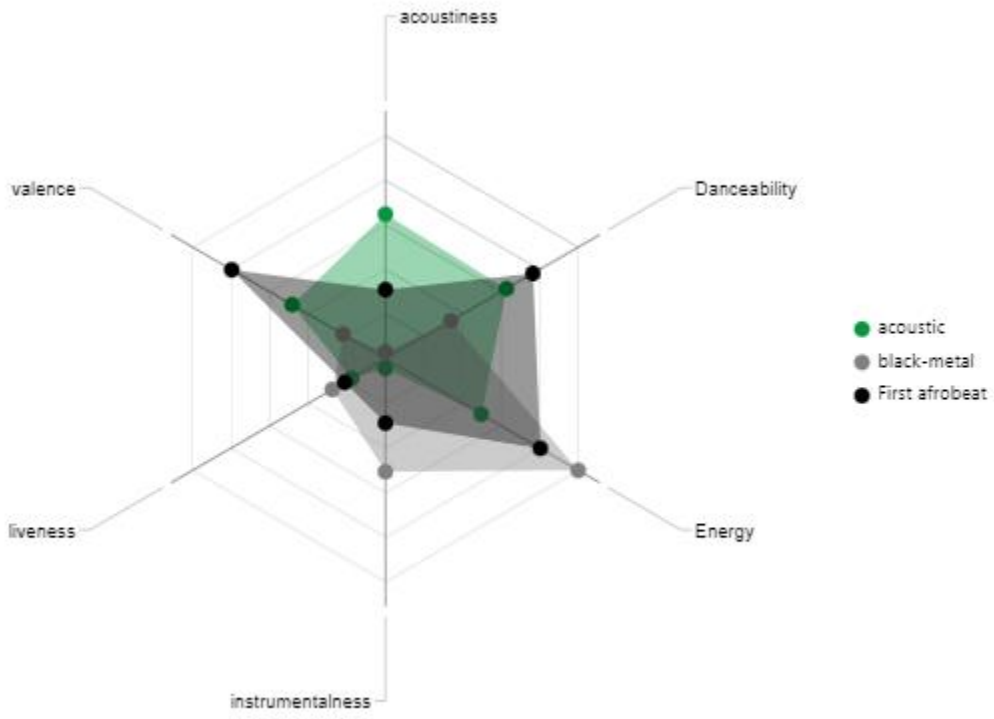
Additionally, the standard deviation of duration times can be an indicator of how consistent the genre is in terms of track length. For example, honky-tonk has the lowest standard deviation, suggesting that tracks in this genre tend to be a similar length, while breakbeat has the highest standard deviation, indicating that tracks in this genre can vary widely in length.

It's also worth noting that there are some genres with a relatively small sample size, which may impact the accuracy of the data. For example, there are only 6 tracks in the dataset labeled as grindcore, which may not be enough to draw meaningful conclusions about the genre as a whole.

Overall, the duration time of tracks can provide insight into how the music is intended to be listened to, and can help identify patterns in the composition of different genres.



3.2 How does a genre typically sound? (Radar Chart or Bar chart)



Based on the information presented, it appears that the acoustic genre is generally characterized by music that is not very danceable, with low energy and speechiness, high acousticness, low instrumentalness, moderate liveness, and moderate valence. In contrast, other genres have their own typical sound characteristics. For example, afrobeat is typically more danceable and has higher energy and valence compared to acoustic. Black metal typically has low danceability and high instrumentalness, while classical music typically has high acousticness and low energy.

It's important to note that these averages are not representative of every single song in the genre, but rather an overall trend of the genre's sound. There can be variations within each genre, and some songs may deviate from the typical characteristics of their respective genres.

3.3 do characteristics of popular songs vary across genres?

The question of whether the characteristics of popular songs vary across genres is a topic of great interest to both music enthusiasts and industry professionals. We decided to focus on popular songs that are categorized within genres with an average popularity index value exceeding 45. This led us to analyze six distinct genres: pop-film, k-pop, chill, sad, grunge, indian, and anime. then we utilized a statistical method called ANOVA (Analysis of Variance) to examine the audio features of popular songs in the Spotify dataset . The audio features that were analyzed include danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo.

ANOVA, or Analysis of Variance, is a statistical technique used to determine if there are significant differences between groups. By using ANOVA, we were able to identify which audio features varied significantly across different music genres.

The following tables display the significant results of the ANOVA analysis conducted on the characteristics of popular music genres in the Spotify dataset. Notably, the tables illustrate an overview of average values of danceability, energy, speechiness, and instrumentalness, as well as the ANOVA results such as P-value, F, and F critical with significant level of 5%:

Groups	Average danceability	Average Energy	Average instrumentalness	Average speechness
pop-film	0.6004	0.6790	0.0040	0.0620
k-pop	0.6458	0.4243	0.0095	0.0825
chill	0.6739	0.4551	0.1200	0.1092
sad	0.7037	0.7999	0.0990	0.1395
grunge	0.4619	0.5924	0.0335	0.0614
indian	0.5914	0.7013	0.0266	0.0675
anime	0.5326	0.6156	0.1988	0.0762

ANOVA	F	P-value	F crit
danceability	179.020	8.9627E-195	2.102
energy	198.243	8.0814E-213	2.102
loudness	93.659	5.4955E-108	2.102
speechiness	52.430	1.25492E-61	2.102
acousticness	181.714	2.4704E-197	2.102
instrumentalness	51.319	2.46322E-60	2.102
liveness	10.147	3.90994E-11	2.102
valence	31.442	7.8374E-37	2.102
tempo	15.533	1.26914E-17	2.102

These results show that there are significant differences in audio features across different music genres as indicated by P-values<0.05 and F> Fcritic.

The ANOVA results show that danceability varies significantly across different genres, with chill having the highest average danceability score of 0.69, followed by sad and k-pop, while grunge has the lowest average danceability score of 0.46. In terms of energy levels, grunge has the highest mean energy level of 0.79, while chill has the lowest mean energy level of 0.41. Loudness levels also vary significantly across genres, with chill having the lowest mean loudness level of -14.36, while k-pop has the highest mean loudness level of -5.54.

Speechiness and acoustiness levels also vary significantly across genres, with sad having the highest mean speechiness level of 0.11, while grunge has the lowest mean speechiness level of 0.04, and chill having the highest mean acoustiness level of 0.58, while grunge has the lowest mean acoustiness level of 0.15. Finally, instrumentalness levels also vary significantly across genres, with chill having the highest mean instrumentalness level of 0.55, while pop-film has the lowest mean instrumentalness level of 0.03.

The insights derived from the analysis suggest that each genre has distinct characteristics that set it apart from other genres. These insights can be useful for Spotify in improving their recommendation system and for users in discovering new songs and genres that align with their preferences. Additionally, music producers and artists can use this information to create songs that fit the desired features of a particular genre and engage their target audience more effectively.

Overall, the ANOVA analysis provides a comprehensive understanding of how different characteristics vary across genres and how this information can be utilized to enhance the music experience of both Spotify and its users.

4. Recommendations:

- Spotify can use this analysis to improve its recommendation system and provide users with a more personalized music experience by taking into account the characteristics of the music they enjoy listening to.
- Spotify can provide users with genre-specific playlists that match their listening preferences.
- Spotify can work with artists and producers to create music that aligns with the characteristics of the most popular songs on the platform.
- Spotify can also use this analysis to target its marketing and advertising efforts towards users who prefer certain genres or characteristics of music.

References:

Kaggle. (n.d.). Spotify Tracks: https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

Appendices

GitHub Repository: Link

- Spotify_track_cleaning.sql
- Spotify_track_analysis.sql