Analysing Spotify Tracks Dataset Using Apache Pig and Hive

Name: Tanuli Liyanage
Student Id: 22100555
Module: CSC1109 Data at Speed & Scale

1. Introduction

For this project, I selected the Spotify Tracks dataset from Kaggle which contains data of over 100,000 Spotify songs of different artists, genres, and albums with a variety of audio features. The dataset is considered quite large and requires tools and technologies that are more advanced and powerful. For processing and analysing the above dataset, I used the Hadoop ecosystem. Firstly, the dataset was cleaned using Apache Pig, then Apache Hive was used for simple and complex queries. Throughout the process I have analysed musical trends and patterns to derive interesting insights that could help explain what makes a track popular.

All scripts, Pig and Hive queries and output files used in this project are available in the Git repository here.

2. Data Cleaning

The data cleaning was performed using Apache Pig. The CSVExcelStorage function from Piggybank was used to handle the commas inside quoted strings and using tabs as the delimiter proved to be more accurate than using a comma.

1. **Removed the header row** - The first row contained the column names which was considered as a data row, hence it was removed.

2. **Dropped the index column** - The first column was unnamed and had no analytical importance, hence it was removed.

3. **Filtered out missing and empty values** - Null and missing values were removed to ensure the data is consistent and accurate for analysis.

4. **Trimmed whitespaces** - Unnecessary whitespaces were removed to maintain the quality of data.

5. **Removed duplicate rows** - Applied the distinct operator to remove 450 duplicated rows to ensure data integrity.

```
Input(s):
Successfully read 114001 records from: "file:///lab/data/spotify_tracks.csv"

Output(s):
Successfully stored 113549 records in: "file:///lab/outputs/cleaned_data"
```

Figure 1: Record counts of the original and the cleaned datasets

3. Simple Query Analysis

After cleaning the dataset, simple queries were performed using both Apache Pig and Hive to analyse the data. The same two queries were repeated in both languages and given below are the results of each query and analysis of each result.

Query 1: Most popular Country song by Zach Bryan that has a danceability score > 0.5 and is the most popular.

track_name	artists	album_name	popularity	danceability
Late July	Zach Bryan	Open the Gate	42	0.532
Oklahoma Smokeshow	Zach Bryan	Hideaway - New Boots	5	0.544
Oklahoma Smokeshow	Zach Bryan	Country Car Hits	3	0.544
If She Wants a Cowboy	Zach Bryan	Good Times Country	3	0.553
Oklahoma Smokeshow	Zach Bryan	Never Leave - Country Magic	1	0.544

Figure 2(a): Query 1 result from Hive

```
(Late July,Zach Bryan,Open the Gate,42,0.532)
(Oklahoma Smokeshow,Zach Bryan,Hideaway - New Boots,5,0.544)
(Oklahoma Smokeshow,Zach Bryan,Country Car Hits,3,0.544)
(If She Wants a Cowboy,Zach Bryan,Good Times Country,3,0.553)
(Oklahoma Smokeshow,Zach Bryan,Never Leave - Country Magic,1,0.544)
```

Figure 2(b): Query 1 result from Pig

The results obtained are arranged in descending order by popularity and it's clear that the results are the same in both Hive and Pig. 'Late July' in the 'Open the Gate' album has been the most successful song by Zach Bryan with a moderate danceable score of 0.532. When comparing the tracks, it's interesting to see that 'Oklahoma Smokeshow' with a slightly higher danceable score of 0.544 was a lot less popular than 'Late July'.

Query 2: The Genres of music that has a high valence score

track_genre	avg_valence
salsa	0.8145360721442889
forro	0.7604993987975944
rockabilly	0.7267367367367376
afrobeat	0.6984746746746758
ska	0.6967199999999996
children	0.6937879759519042
samba	0.6934091000000006
pagode	0.6879780000000002
kids	0.6814791120080723
party	0.6811789738430577

Figure 3(a) : Result from Hive

```
(salsa,0.8145360721442875)
(forro,0.7604993987975959)
(rockabilly,0.7267367367367367)
(afrobeat,0.6984746746746753)
(ska,0.6967200000000001)
(children,0.6937879759519041)
(samba,0.6934090999999999)
(pagode,0.6879780000000006)
(kids,0.6814791120080722)
(party,0.6811789738430579)
```

Figure 3(b) : Result from Pig

The results above show that genres such as salsa, forro, rockabilly, afrobeat have the highest average valence scores. Valence means the musical positiveness in a track and the above genres typically represent energetic, lively and joyful musical styles. Children and kids' music are always designed to be happy and playful which explains their high valence scores. Therefore this query results confirms that valence effectively captures the positivity and brightness of music.

4. Complex Query Analysis

Moving on from simple to more complex queries to analyse the dataset further using Apache Hive.

Query 1: Average Speechiness and Popularity by Genre

track_genre	avg_speechiness	avg_popularity	total_tracks
comedy	0.756	24.64	996
j-dance	0.22	26.69	993
dancehall	0.187	33.47	999
kids	0.153	14.94	991
funk	0.152	32.32	1000
hardcore	0.149	36.11	999
grindcore	0.138	14.63	998
sad	0.13	52.38	1000
hip-hop	0.13	38.08	991
french	0.127	41.08	999
hardstyle	0.122	26.6	998
reggaeton	0.121	23.86	1000
happy	0.12	21.72	999
latino	0.115	25.83	993
reggae	0.115	20.63	1000
death-metal	0.111	32.18	999
metalcore	0.11	43.48	1000
emo	0.109	48.13	1000
show-tunes	0.108	31.26	999
turkish	0.105	40.7	999

Figure 4(a): Complex Query 1 result

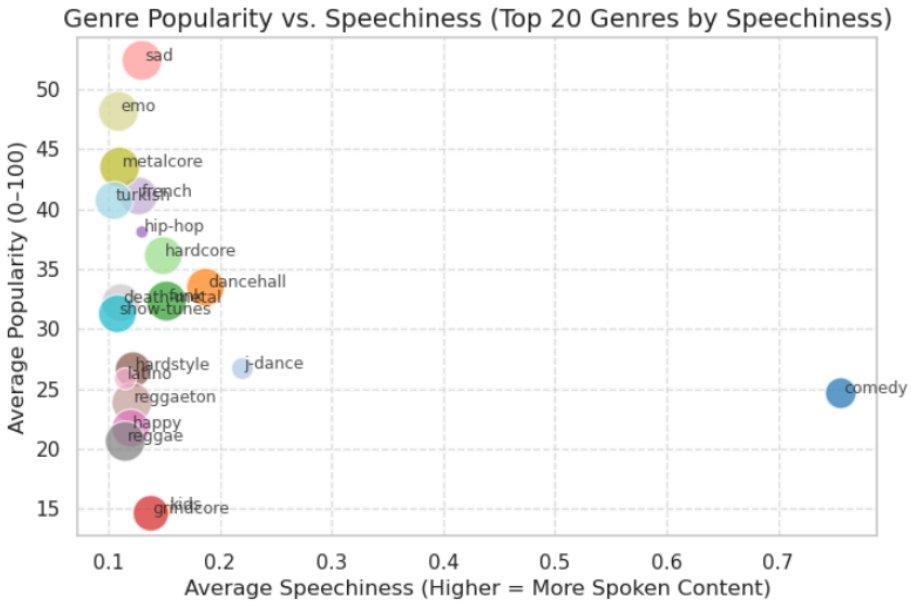

Figure 4(b): Correlation between Popularity and Speechiness

The goal of this query was to understand if songs with high spoken content are popular or not. According to the scatter plot above, it's very clear that as the speechiness increases, popularity decreases. Despite the large difference of average speechiness between comedy and j-dance genres, the popularity scores are very similar. Songs under the genre sad are noticed to be very popular and it is understandable to have less words in sad music as they tend to have more instrumental effects. From the scatter plot it can be seen that larger bubbles indicate genres with more available tracks which suggests those tracks are mainstream genres such as funk and sad.

Overall, there's no direct linear relationship between speechiness and popularity, but the pattern show that genres with high speechiness are less mainstream (eg: Comedy) and genres with moderate speechiness have more popularity (eg: hiphop, french, hardcore)

Query 2: Valence vs Danceability by Genre

v.track_genre	avg_valence	avg_danceability
salsa	0.815	0.668
forro	0.76	0.65
rockabilly	0.727	0.561
afrobeat	0.698	0.669
ska	0.697	0.581
children	0.694	0.716
samba	0.693	0.575
pagode	0.688	0.578
kids	0.681	0.779
party	0.681	0.667
disco	0.671	0.677
rock-n-roll	0.67	0.55
reggae	0.648	0.745
reggaeton	0.643	0.759

Figure 5(a): Complex Query 2 result

It can be seen that salsa, forro, rockabilly, afrobeat, children, samba are the genres that have tracks that make one feel the happiest. Songs from these genres are naturally upbeat and cheerful which explains the reason for high valence scores.

Genres with high valence also have high danceability. The scatter plot below in Figure 5(b) clearly shows the correlation between valence and danceability. The top genres such as salsa, samba, afrobeat, reggaeton are highly dance focused genres and those data points scattered towards the right corner of the scatter plots confirms that happier tracks are more danceable.

Tracks from reggaeton genre are very highly danceable however it has a lower happiness score compared to other dance focused genres like salsa and sumba. This could be due to reggaeton music often containing melodies that use minor keys that evoke somber feelings. This indicates that high danceability doesn't always mean happiness in music.

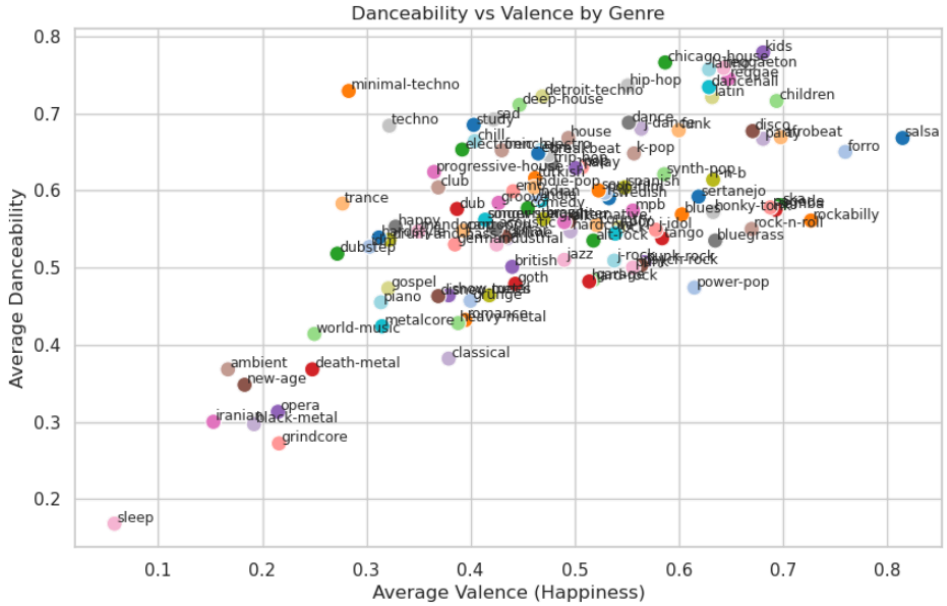

Figure 5(b): Correlation between Valence and Danceability

Query 3: Top Energetic Artists by Genre (10% sample)

artists	track_genre	avg_energy	avg_popularity
Willi Herren	disco	1.0	30.0
Markus Becker;Die Mallorca Cowboys	party	1.0	0.0
Zorra	heavy-metal	1.0	19.0
Watain	black-metal	1.0	18.0
disrupt	grindcore	1.0	12.0
Banner Pilot	power-pop	1.0	24.0
dj funk;TJR	chicago-house	1.0	30.0
Ruido Blanco Hart;Ondas Alfa;Música de concentración profunda	world-music	1.0	23.0
Rotterdam Terror Corps	techno	1.0	29.0
Jürgen Drews	party	1.0	27.33

Figure 6(a): Complex Query 3 result

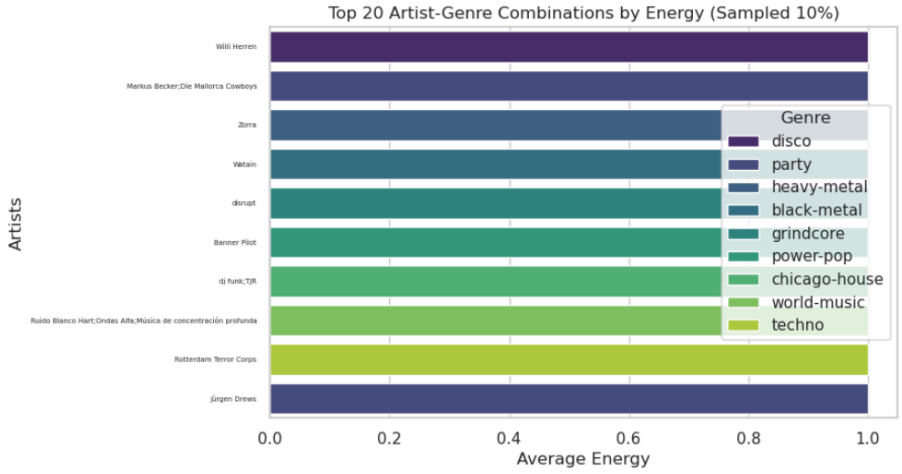

Figure 6(b): Top 20 Artist-Genre Combinations by Energy (Sampled 10%)

I have used TABLESAMPLE to get a 10% sample of the dataset to focus on energy per artist and their genres. According to the above results and the horizontal bar chart, the sample seems to be biased towards high energy tracks and that makes it quite hard to draw meaningful conclusions about energy per artist as the whole dataset is not represented by this sample.

However, based on this 10% bucket sample, the above artists and genres appear to have the highest energy tracks of score 1.0. And it makes sense to have genres such as disco, party, heavy-metal and techno among the track genres as these genres are usually considered very loud and intense.

Although the tracks are extremely energetic, popularity scores are varied. Some highly energetic tracks are not popular at all. For example, the party song by Markus Becker and Die Mallorca Cowboys. That shows how energy alone doesn't determine how successful or popular a song is.

5. Conclusions

In conclusion, this project has helped me understand how a large dataset can be processed, cleaned and analysed using the Hadoop ecosystem. Through a range of simple and complex queries, I have revealed some interesting musical trends. Genres such as salsa, samba and afrobeat were found to have valence and danceability. The speechiness and popularity analysis highlighted that songs with more spoken content are less popular than melodic genres. Finally, the sampling query illustrated how sampling bias can affect analytical accuracy.