CSC1109 - Spotify dataset analysis using Apache Pig and Hive

Ronan Kelly – 21374881

October 2024

https://github.com/Kellyronan03/CSC1109-Assignment1.git

1 Introduction

Hadoop is a distributed file system framework that allows for large scale data processing that is parallelised across multiple systems to increase efficiency. The framework employs the MapReduce technique which sends subsets of the overall datasets to different machines to be handled on a smaller scale. Pig and Hive are two languages that are implemented on top of the Hadoop system to allow for data processing and data querying on large datasets to take place.

I've decided to conduct my analysis on the Spotify dataset with metrics such as popularity and danceability. My initial hypothesis was that songs that have high danceability and energy scores would in turn have high popularity scores, my thought was also that these songs would be a favourite of consumers. So, this was why I decided to focus my analysis on songs with high energy and danceability.

2 Cleaning the data

Before I began analysing the data, I had to clean and process the data in Pig to correct any errors and allow for easier querying. When first cleaning I established which columns would be useful for my later analysis and dropped the rest, however the main cleaning that I conducted on the dataset was accounting for unacceptable values, for example where certain metrics such as popularity and energy were outside the previously specified range I would limit the value to 1. Another piece of cleaning I conducted was to remove any duplicate rows which were located in the dataset. I also standardised the duration value into seconds to make the value more readable. Finally the last piece of cleaning I conducted on the dataset was to replace any commas within the name of the track, album or artist were replaced with a space so as not to create problems when loading in the data for both Pig and Hive querying.

3 Queries

I began with the initially simple queries in both Hive and Pig, these queries won't require too much analysis and largely serve as a confirmation that the dataset is correctly formatted and cleaned.

As I said above the focus of all my analysis is on the energy and danceability of songs within the dataset and possibly if that correlates with popularity.

3.1 Most energetic songs

My first query was to discover which were the top 5 most energetic songs within the dataset.

```
White Noise Research    A Big Old Downpour      1.0
Akitsa  Affront Final   1.0
Jürgen Drews    Amigo Charly Brown      1.0
Scott Brown     Bass Be Louder - Edit   1.0
Ocean Sounds;BodyHI;Ocean Waves For Sleep       Calming Sea Waves       1.0
```

(a) Hive query results

```
(White Noise Research,A Big Old Downpour,1.0)
(Akitsa,Affront Final,1.0)
(Jürgen Drews,Amigo Charly Brown,1.0)
(Scott Brown,Bass Be Louder - Edit,1.0)
(Ocean Sounds;BodyHI;Ocean Waves For Sleep,Calming Sea Waves,1.0)
```

(b) Pig query results

As we can see from these results both Pig and Hive are the same this is as it proves that all the data was loaded correctly and that there was no issues from within the code. The 2 most energetic songs within the dataset 'A Big Old Downpour' and 'Affront Final' , however this does not tell us much as they all have the same scores for energy.

3.2 Most danceable songs

My next query was related to the most danceable, and finding out the top 5 for this category.

```
(Quantic,Sol Clap,0.985)
(That Girl Lay Lay,Medicaid Baby,0.984)
(Delano Smith,Inspiration,0.983)
(Oliver Schories,Daily Routines,0.982)
(dj funk,Bitches,0.981)
```

(a) Pig query results

```
Quantic Sol Clap        0.985
That Girl Lay Lay       Medicaid Baby   0.984
Delano Smith    Inspiration     0.983
Oliver Schories Daily Routines  0.982
dj funk Bitches 0.981
```

(b) Hive query results

Once again we can see that the results are both the same which is another good confirmation that all is well with the dataset. This time the danceability figure isn't the same across all the results so more can be taken from this result. I believe it is quite interesting that none of these songs would be notoriously popular which is strange because I initially believed that they would some of the most popular songs.

3.3 Songs with highest average energy, danceability, tempo and valence

For this query it was slightly more advanced with finding averages for a few different categories. I initially looked through the description of the dataset to discover which columns I believed would be the best matches for establishing a songs popularity. Through this research I settled on energy, danceability, tempo and valence. I would then compare these results with the most popular songs to see if there is any similarities verifying my hypothesis.

```
That That (prod. & feat. SUGA of BTS)
Percolator - Keep Movin' Mix
The Penguin Dance
Here We Go
Boom Boom Robot Da
```

(a) Hive query results

As we can see these are not necessarily the most popular songs ever, however, I would like to compare with what the data defines as the most popular songs to get a proper understanding.

```
Sam Smith;Kim Petras    Unholy (feat. Kim Petras)       100
David Guetta;Bebe Rexha I'm Good (Blue) 98
Manuel Turizo   La Bachata      98
Bad Bunny;Chencho Corleone      Me Porto Bonito 97
Bad Bunny       Tití Me Preguntó        97
```

(b) Most popular songs

As we can see there is no crossover between the songs that I initially found and what the dataset defines as the most popular songs. I find this interesting as I thought characteristics such as energy and danceability would be a good indicator of how well-liked and popular is.

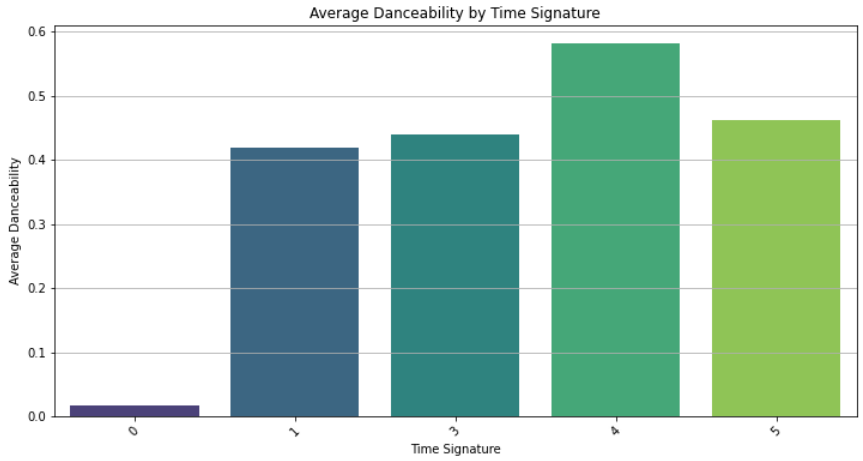
3.4 Time Signature with the best danceability

My next query involved finding out which time signature would have the highest average danceability as I thought it might be a good way of finding out which songs would be most popular.

```
4       0.5812857745524692
5       0.46153674475906925
3       0.440307478235801
1       0.41996413065404026
0       0.01676229521876476
```

(a) Hive Query Results

The most danceable time signature is 7/4 however I would say this is slightly skewed as majority of songs are produced at this time signature.



(b) Graph of average danceability of time signatures

3.5 Top 5 songs from the genre with the highest average energy and danceability

The next query I conducted on the Spotify dataset was an example of a join query where I initially create a subquery that calculates the genre with the highest average energy and danceability and joins that with the main query which returns the top 5 songs for danceability and energy in that specified genre.

```
Saturnus        Hate    death-metal
Stockholm Syndrome      Djin    death-metal
Resurgence      Hate    death-metal
Removed from Consciousness      PeelingFlesh;Jon Huber;Bludgeoned;Pathology;I Declare War       death-metal
Rid You of Your Flesh   LIK     death-metal
```

(a) Hive query results

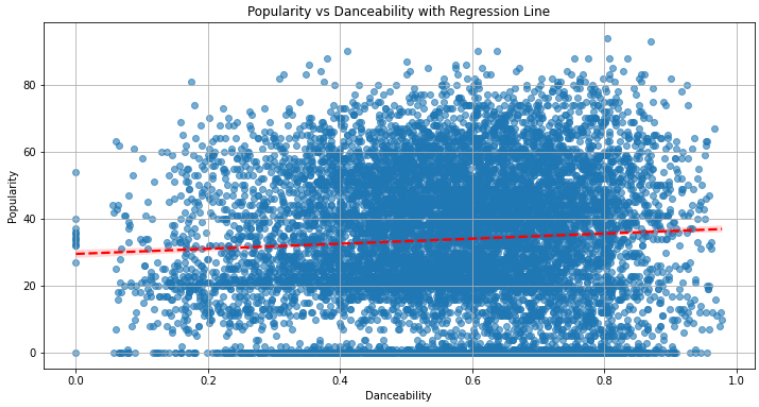
As we can see from this the genre with the highest average danceability and energy is Death Metal, although, this is a genre that you would associate with being very energetic and lots of dancing it wouldn't be the one I would expect to be top of that list which is interesting. As for the specific songs I'm not well aware of songs within the Death Metal genre so I'm not sure if these would even be considered as the best songs in this genre.

3.6 Correlation between various features

The final query I ran on my dataset was multiple different correlation experiments on a 20% sample of the entire dataset. I decided to sample the dataset to improve overall efficiency of the code. As I said at the start I stated that I believed that energy, danceability and maybe even tempo would be a good indicator of a songs popularity and so I decided to run a few correlation experiments to really test if any of them are actually correlated.

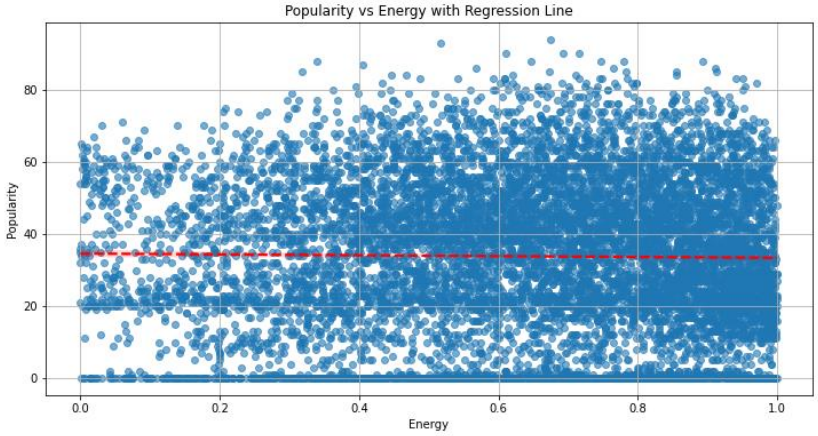
```
danceability vs popularity      0.04972553164689412
danceability vs energy  0.11503071861459581
danceability vs tempo   -0.04132458711875372
popularity vs energy    -0.018151049981364926
energy vs tempo 0.250491602596621 95
```

(a) Hive query results



(b) Graph of Danceability Vs Popularity

As we can a very slightly positive correlation between popularity and danceability, this isn't great for my initial hypothesis.



(c) Graph of Popularity Vs Energy

This shows a slight negative correlation which once again does not agree with my initial hypothesis. Overall, none of the correlation experiments agree with my initial statement showing that energy and danceability are not only not particularly correlated to each other but also have very little influence on a songs popularity which shocks me as I imagined a much greater relation between these features.

4 Conclusion

Overall this assignment has shown haw useful Pig and Hive have for querying and processing large datasets. The ability to be able to use efficient and readable scripting and SQL languages to query large datasets is hugely important. This has been a great opportunity for me to develop my SQL querying and scripting skills while also learning new skills in processing and managing large datasets has been very enjoyable.

Through my analysis on the Spotify dataset and my initial hypothesis on how the features are related, I discovered that these features do in fact have very little influence on each other. I also learned a lot about danceability and energy and which genres would be considered the most energetic and danceable while also discovering the most popular time signatures and tempos. This has been a great experience working with and getting to know this dataset.