Analysis on Spotify Dataset

Christiane Manuela AYO NDAZO'O
Gloria ISEDU

January 2023

1 Introduction

Data analysis aims to inspect, describe, transform, visualize, and model data in order to discover useful information, suggest conclusions, and support decision making.

In this project, we wanted to have this analytical approach in order to extract useful information from the spotify dataset. With the data provided by the data set(we will describe the data set later in the report) and the algorithms learned in class, the questions that came to mind were (1) what makes an artist/song popular, (2) can we guess from quantitative data about a song in which year it was released, (3) can we perform a classification, (4) are there clusters in this data?

It is the answers to these questions that will shape this report.

2 Overall Description

1. **Shape**

   Spotify dataset has **169909 rows** and **19 columns**. Also, there is no missing value in the data set.

2. **Features and data types**

   The features of this dataset are : **acousticness(float)**, confidence measure between 0 and 1 of whether the track is acoustic; **artists(object)**, artists who performed the track; **danceability(float)**, between 0 and 1, it describes how suitable a track is for dancing; **duration_ms(int)**, duration of the track in milliseconds; **energy(int)**, measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity; **explicit(int)**, whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown); **id(string)**, Spotify ID for the track; **instrumentalness(float)**, between 0 and 1, predicts whether a track contains no vocals; **key(int)**, between -1 and 11, key the track is in; **liveness(int)**, between 0 and 1, detects the presence of an audience in the recording; **loudness(int)**, between -60 and 0, overall loudness of a track in decibels (dB); **mode(int)**, modality (major or minor) of a track; **name(object)**, name of the album; **popularity(int)**, popularity of the artist[1]; **release_date(object)**, date the album was first released; **speechiness(float)**, between 0 and 1, detects the presence of spoken words in a track; **tempo(float)**, overall estimated tempo of a track in beats per minute (BPM); **valence(float)**, measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track; **year(int)**, year the song has been released;

3. **Quantitative variables analysis**

   According to the information provided by spotify[2], of these 19 columns, 13 were quantitative variables. The table below shows the mean, standard deviation, minimum, maximum and percentiles (0.25, 0.50) of these variables.

Variable	mean	std	min	max	25%	50%
acousticness	0.493214	0.376627	0.0	0.996	0.0945	0.492
danceability	0.53815	0.175346	0.0	0.988	0.417	0.548
duration_ms	2.314062e+05	1.213219e+05	5.108e+03	5.4035e+06	5.108e+03	1.7104e+05
energy	0.488593	0.267390	0.0	1.0	0.263	0.481
instrumentalness	0.161937	0.309329	0.0	1.0	0.0	0.000204
key	5.200519	3.515257	0	11	2	5
liveness	0.206690	0.176796	0.0	1.0	0.0984	0.135
loudness	-11.370289	5.666765	-60	3.85	-14.47	-10.47
popularity	31.556610	21.582614	0.0	100	12	33
speechiness	0.094058	0.149937	0.0	0.969	0.0349	0.045
tempo	116.948017	30.726937	0.0	244.091	93.516	114.778
valence	0.532095	0.262408	0.0	1.0	0.322	0.544
year	1977.223231	25.593168	1921	2020	1957	1978

Table 1: Mean, std, min, max and percentiles of quantitative variables

4. **Qualitative variable** We have 1 qualitative variable which is the **mode**

3 Analysis

1. **Principal Component Analysis(PCA)** Given the dimension of our dataset, we wanted to reduce its number of variables to 2 or 3 using PCA. This reduction would be valid if and only if these components kept a good part of the initial variance of the data. From the scree plot in figure 1, we see that for this data set, the first 2 principal components represent the information in the data well enough.

[1] In this dataset, the popularity of an artist is confused with the popularity of his song

Analysis First, we did an analysis on the second component(PC2) and observed that the PC2 was capturing songs that had danceability, tempo, valence and speechiness equal to 0 from 1991. After this first analysis on the PC2, we did another analysis on the first and second principal components together(that is, PC1 and PC2 together). In the second analysis we deduced that first component captures information about the year.

2. **Popularity**

Our first analysis was on popularity. As said before, popularity describes how popular is an artist/a song. Thus, the series of experiments we conducted aimed to determine the criteria that make an artist popular. The model used is a linear regression model because popularity is a continuous variable.

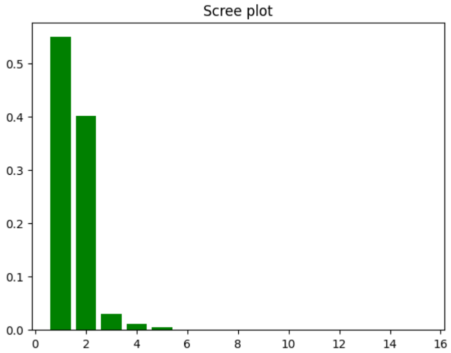

Figure 1: Scree plot

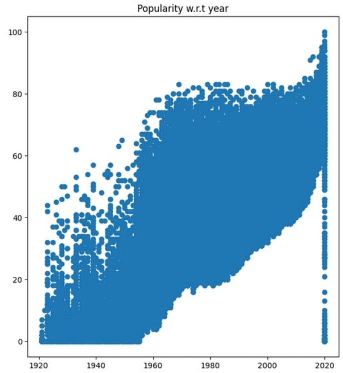

Figure 2: Popularity as function of year

(a) **Popularity and year**

i. **A simple regression model**

According to the graph in figure 2, popularity increases with the year. So there is a linear relationship between popularity and year. Based on this observation, we hypothesized a simple linear regression of popularity by year.

The model constructed from this hypothesis yielded a result of 77.56%, meaning that 77.56% of the variation in popularity is explained by the year.

ii. **A polynomial model**

Seeing the look of the graph in figure 3, we decided to try to express popularity as a polynomial of the year. Aware of the impact of the degree of the polynomial on the quality of the model, we decided which degree to choose for the polynomial by grid search cross-validation. From that operation, the best degree was three(3) and we got an average score of 77.75 %.

Using year as a polynomial increases a little bit(0.19) the score of the model. This can be explained by the shape of the model, which more closely matches the appearance of the graph in the figure2

(b) **Popularity and other variables**

Although popularity is more correlated with the year, we may have missed some information by looking only at the year. Therefore, the purpose of this section is to examine other features that might influence popularity.

Since we obtained 12 other numerical variables (apart from popularity), we first performed a feature extraction using sklearn's Recursive Feature Extraction (RFE) [1] to determine which features had an influence on popularity. The features that emerged were: acousticness, danceability, instrumentalness and speechiness.

The figure 3 shows the overall comparison score of all the popularity regressors on training and test data.

On this plot, we see that the multiple linear regressor with the features of acousticness, danceability, instrumentalness and speechiness features has the best score. Adding those parameters added 1% percent of accuracy.

Conclusions :

- For an artist/a song to be popular, five(05) criteria should be taken into account : the year he releases the song, the acousticness of the song, the danceability of the song its instrumentalness and its speechiness.
- The models are as good on training data as they are on test data, which means that there is no underfitting or overfitting. All these models are therefore reliable

(c) **Analyze of the data of popular songs over years**

Knowing that popularity is a function of year, acousticsness, danceability, instrumentalness, and speechiness, we wanted to take a detailed look at the values that have increased popularity over the years.

i. **A first attempt** In our first attempt, we tried to examine the data by time intervals to see if we could retrieve information about the predominant criteria for popularity. We wanted to see linear correlations of popularity over time. But unfortunately, we did not find any interesting results.

ii. **Another attempt** Since we are interested in the popularity of the songs, the other view we had was to look directly at the most popular songs over the years and analyze their data directly.

The table below shows the general conditions for a good popularity. But there are songs that have been popular without strictly respecting these conditions.

Period	Acousticness	danceability	instrumentalness	speechiness
1921-1941	High(>0.7)	Medium(0.4-0.7)	Low(<0.1)	Low(<0.1)
1942-1965	High(>0.6)	Medium(0.3-0.5)	Low(<0.1)	Low(<0.1)
1966-1998	Low(<0.4)	Medium(0.3-0.6)	Low(<0.1)	Low(<0.1)
1999-2020	Low(<0.3)	High(>0.5)	Low(<0.1)	Low(<0.1)

Table 2: Range of values of acousticness, danceability, instrumentalness and speechiness of popular songs over years

Over the years, the criteria for popularity have changed, mainly in terms of acousticness and danceability. We find that songs are becoming more versatile, joining singing and dancing performances.We also observe that most of the popular songs were songs with a high valence(knowing that a high valence corresponds to a happy or positive song)

3. **Year** Another question was to know if we could predict the year from the variables.

   (a) **A simple linear regression model**

   We started our experiment using the popularity. We saw above that popularity was strongly correlated to the year. So we started building a simple linear regression model. We got a score of 77.32% of accuracy.

   (b) **A polynomial regressor model**

   Using the same approach as with popularity, we expressed the year as a polynomial model of popularity. The score obtained was 78.39%.

   (c) **A multiple linear regression model**

   As we did with popularity, we looked for the features that had an effect on the year. These features are : acousticness, danceability, energy, loudness, popularity and valence.

   The figure 4 shows the overall comparison of linear regression models.

   Conclusions :

   - The multiple linear regression model is more accurate than all the other models.
   - The year is determined by 06 features: popularity, acousticness, danceability, energy, loudness and valence.
   - The models are as good on training data as they are on test data, which means that there is no underfitting or overfitting. All these models are therefore reliable and we don't need to make a regularization.

4. **Mode**

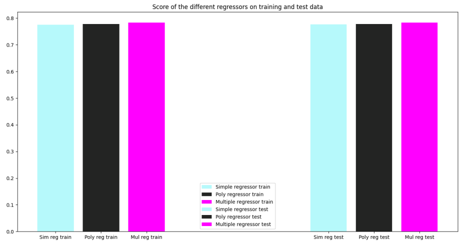

Figure 3: Overall comparison of regression models of population on score

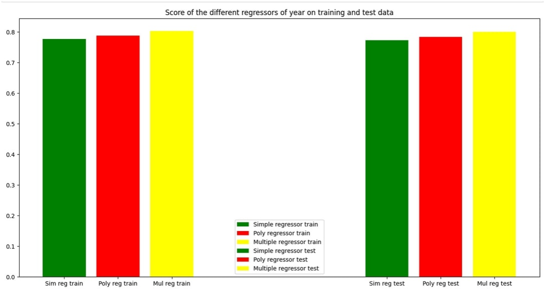

Figure 4: Overall comparison of regression models of year on scores

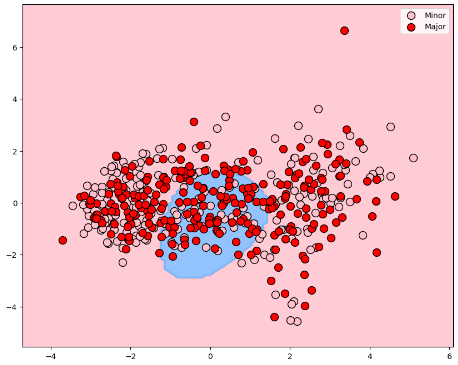

Figure 5: SVM Classifier for the mode of a song

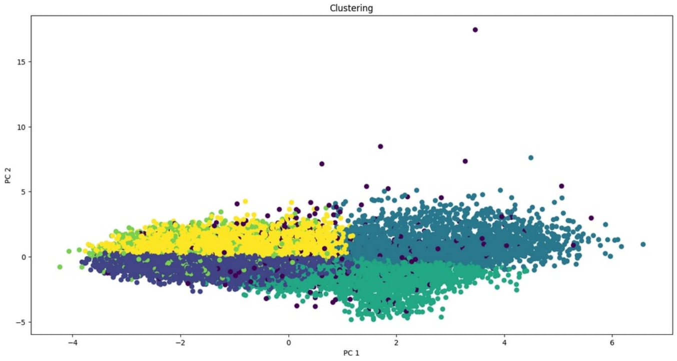

Figure 6: Clustering

Mode (major or minor) is a categorical variable. Given the songs, we wanted to know if we could have a good **binary classifier** of the mode. In the figure 5, we have the decision region of the support vector machine classifier (SVM) where the blue outline represents the region where the songs will be classified as major and the pink represents the region where the songs will be classified as minor.

Our classifier is quite poor (53.4% accuracy) and this could be explained by the fact that the mode is determined by the third note of a song. And with the characteristics of our dataset, it may have been difficult to clearly establish a similarity measure.

5. **Clustering**

We finished our analysis by clustering. The goal was to see if there are clusters in our dataset and if so, what information they represent. To be able to understand these clusters, we added the gender column to our dataset. Unfortunately, we did not have the genres for all songs. So we focused on the songs for which we had genres. The results of this clustering are shown on the figure 6.

We have 6 clusters with a lot of overlapping :

The yellow cluster which actually overlaps the light green cluster. If we make a link with the results we got from the criteria that make a song popular, we will notice that all those songs represent songs with small acousticness. These clusters represent the emergence of hip-hop and rock genres in all their forms.

The dark blue cluster is overlapping the light blue cluster. We remain on the era of development of rock, but we also have r&b. It seems like american and british cultures were influencing the rest of the worlds because we also see genres like japanese chillhop which is japanese music influenced by hip hop.

The green cluster describes emergence of afro, asiatic and south american genres of songs.

The light blue cluster overlaps the dark blue cluster. In addition to british and american genres, we have the emergence of traditional afro genres and some latino genres.

4 Conclusion

Spotify dataset allowed us to make use of different machine learning algorithms such as linear regression(for popularity and year), classification(for mode) and clustering.

These analysis helped us understanding parameters that make a song popular. Also, we have been able to determine in a certain mesure the mode of the songs and discover how different genres were mixing over the years.

At the end of this analysis, the question that remains is: which binary classification algorithm could perform better than the Support Vector Machine we performed?

From this question, improvements to this project could be made by testing different clustering and classification algorithms to compare them based on their score.

References

[1] Rfe sklearn.

[2] Spotify api.