Université Jean Monnet - Saint-Etienne

Spotify Data Analysis

Beggari Mohamed Islem
Moudjahed Mohamed

6 janvier 2023

Table des matières

1 Introduction **2**

2 Overall Description **3**

2.1 Description of the features of the data-set 3

2.2 Some statistics about the data-set and the variables 3

2.3 Analysis goal 4

3 Analysis **5**

3.1 the behavior of the feature and its influence on the popularity 5

3.1.1 What attributes make a song popular in general ? 5

3.1.2 feature Engineering 6

3.1.3 Does this "hitfactor" control absolutely the popularity ? 7

3.2 Hypothesis 1 7

3.3 Conclusion 1 8

3.4 Hypothesis 2 8

3.5 Training a Classifier on "genre" 9

3.6 Understanding the "genre" column 10

3.7 Conclusion 2 10

3.7.1 what characteristics have every genre ? 10

3.7.2 which "genre" was popular in the thirties ? 10

4 Conclusion **11**

5 Bibliography **12**

1 Introduction

This report summarizes all of the primary statistical modeling and analysis results that we could extract from the data.

The purpose of this report is to document the different methods used during this analyses and to clarify the goal of this study which is understanding the evolution of the music and it's popularity throughout time and the features that could influence it.
The project utilized a case-study approach using prior built-in music playback data collected from Spotify's API given by the Professor.
Additionally, this report is designed to serve as a template for describing our interpretations about the result found during the analysis.

The remainder of this report is organized as follows :
Section 2 gives an overall description about the data and the different feature we deal with along the study.
Section 3 presents the Analysis that will be based on the popularity and the different features that could have a (positive/negative) influence on it and we will try to answer some questions like :

- how the popularity of the songs change with years ?
- how could we make a popular song ?
- what kind of music people liked to hear in a certain period ?
- and more other questions ...

Section 4 presents the conclusions we have come to after making this study and we will be recalling all the necessary steps to build and implement the analysis

2 Overall Description

2.1 Description of the features of the data-set

the data-set is composed of 19 different features of different types (int/float/string/object), and a total of 169909 rows(songs) the meaning of each feature :

feature	Description
Year	the release year of the recording.
Acousticness	The higher the value the more acoustic the song is.
energy	The energy of a song, the higher the value, the more energetic. song
Liveness	The higher the value, the more likely the song is a live recording.
Loudness	The higher the value, the louder the song.it varies between -16 and -4
Valence	The higher the value, the more positive mood for the song.
Danceability	The higher the value, the easier it is to dance to this song.
Duration	The length of the song
Speechiness	The higher the value the more spoken word the song contains.
Popularity	The higher the value the more popular the song is.
Artists	the name(s) of the artist(s) that made the song.
Tempo	tempo is the speed or pace of a given piece and derives directly from the average beat duration.
Instrumentalness	The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
release date	the day/moth/year of the release of the song
Genre	the genre of the track

TABLE 1 – Description of the features

2.2 Some statistics about the data-set and the variables

we calculated some measures about the data (mean/standard deviation/min/max/...) and how many null values there are

— **count :** is the total of the quantitative non-null variables
— **mean :**is the average value of each feature
— **std :**is the standard deviation of each feature
— **min/max :** are the minimal/maximal values that could have every feature
— **25% - /50% - /75% :** are the quantiles of this data (values that divide the data into four equal parts)

	count	mean	std	min	25%	50%	75%	max
acousticness	169909.0	0.493214	0.376627	0.0	0.0945	0.492000	0.8880	0.996
danceability	169909.0	0.538150	0.175346	0.0	0.4170	0.548000	0.6670	0.988
duration_ms	169909.0	231406.158973	121321.923219	5108.0	171040.0000	208600.000000	262960.0000	5403500.000
energy	169909.0	0.488593	0.267390	0.0	0.2630	0.481000	0.7100	1.000
instrumentalness	169909.0	0.161937	0.309329	0.0	0.0000	0.000204	0.0868	1.000
key	169909.0	5.200519	3.515257	0.0	2.0000	5.000000	8.0000	11.000
liveness	169909.0	0.206690	0.176796	0.0	0.0984	0.135000	0.2630	1.000
loudness	169909.0	-11.370289	5.666765	-60.0	-14.4700	-10.474000	-7.1180	3.855
mode	169909.0	0.708556	0.454429	0.0	0.0000	1.000000	1.0000	1.000
speechiness	169909.0	0.094058	0.149937	0.0	0.0349	0.045000	0.0754	0.969
tempo	169909.0	116.948017	30.726937	0.0	93.5160	114.778000	135.7120	244.091
valence	169909.0	0.532095	0.262408	0.0	0.3220	0.544000	0.7490	1.000
popularity	169909.0	31.556610	21.582614	0.0	12.0000	33.000000	48.0000	100.000
year	169909.0	1977.223231	25.593168	1921.0	1957.0000	1978.000000	1999.0000	2020.000

FIGURE 1 – Information about the features

2.3 Analysis goal

As we can see there is a Strong correlation between popularity and [year/energy/loudness/acousticness] and low correlation with the other features
So our study will be about the popularity and try to find out what feature can influence it and how.

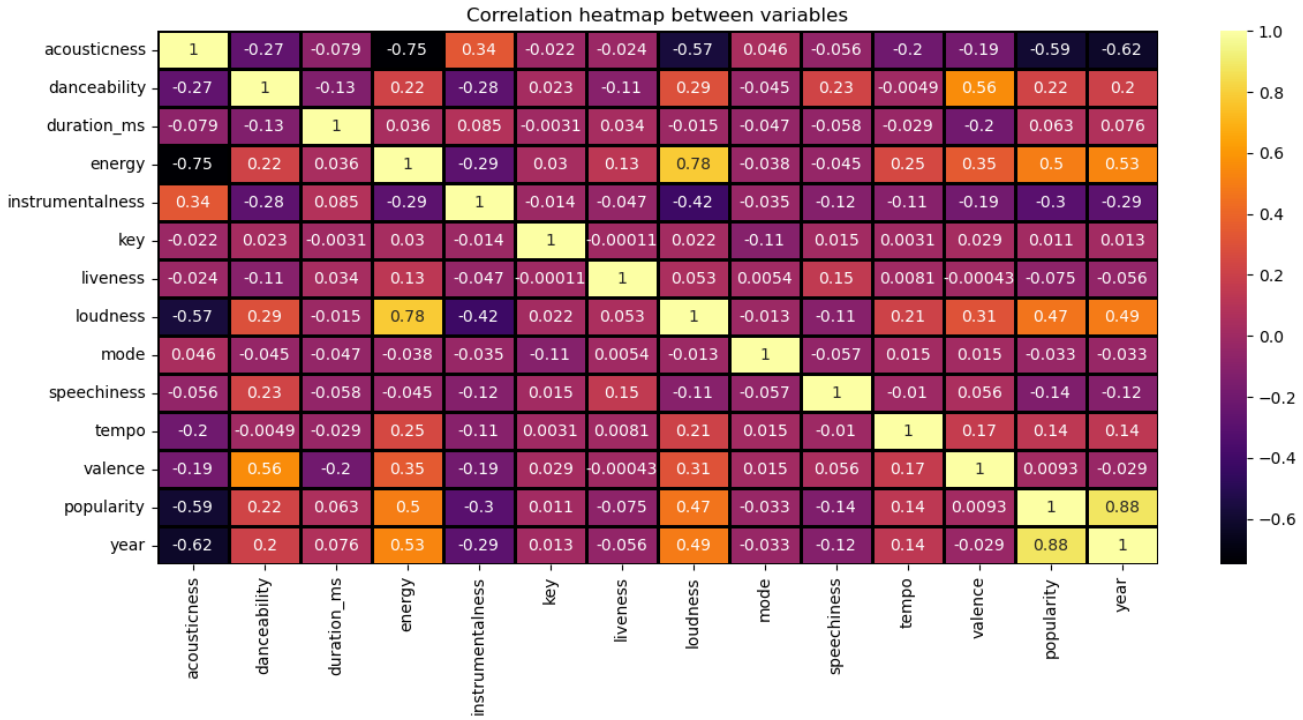

FIGURE 2 – correlation coefficients table

3 Analysis

3.1 the behavior of the feature and its influence on the popularity

3.1.1 What attributes make a song popular in general ?

based on the correlation table popularity have a strong relation with [year / loudness / energy / accousticness] so we will try to take a closer look and analyse the behavior of every feature of these :

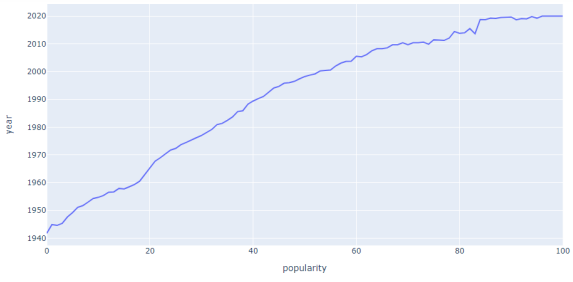

Figure : the behavior of year with respect to popularity

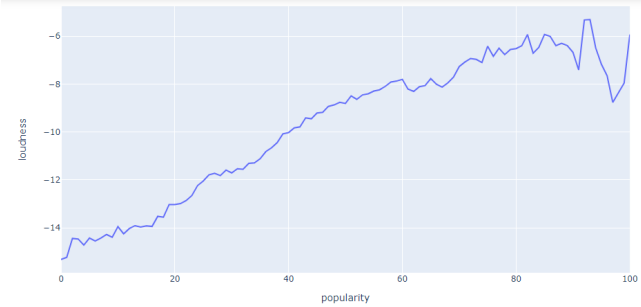

Figure : the behavior of loudness with respect to popularity

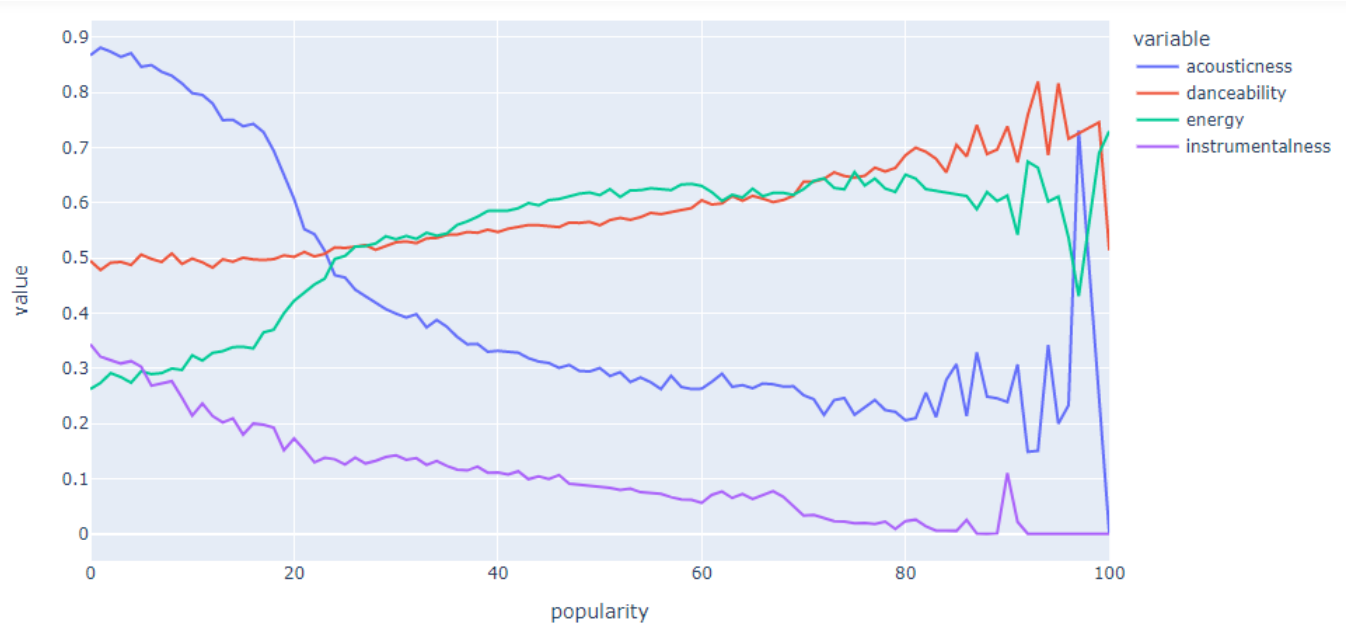

FIGURE 3 – the behavior of the different features with respect to popularity

3.1.2 feature Engineering

As we can see all of the previous features have a general behavior but always there is an exception where the factor was not good but the song was popular though we think that the reason of this is that when some feature is down. the other features compensate, that's why we end up with a big popularity even though a feature was not optimal) so here we're going to make a new feature called (hitfactor)which is a combination between all these features

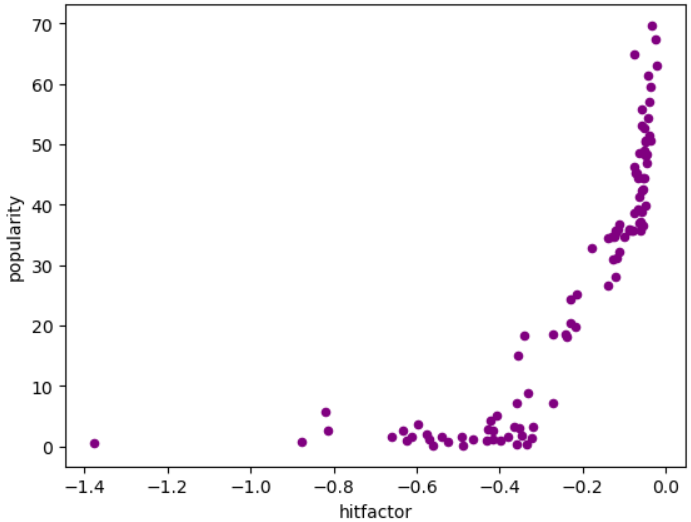

FIGURE 4 – the behavior of the hitfactor with respect to popularity

we can see that the nearest the hitfactor to the 0 value the more popularity we could have (popularity increases when hitfactor get close to 0)

3.1.3 Does this "hitfactor" control absolutely the popularity ?

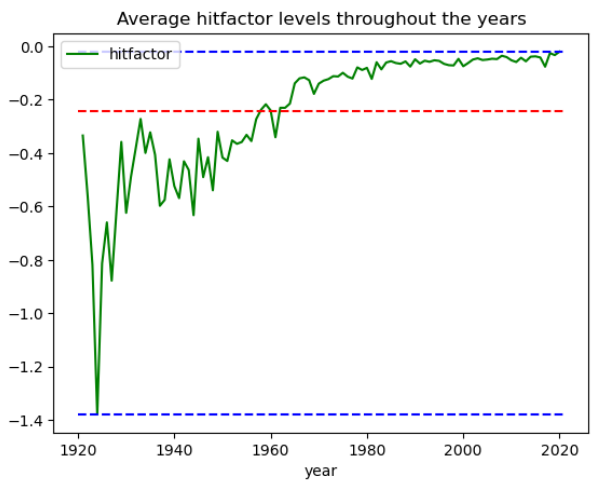

Figure : how hitfactor changes with the time

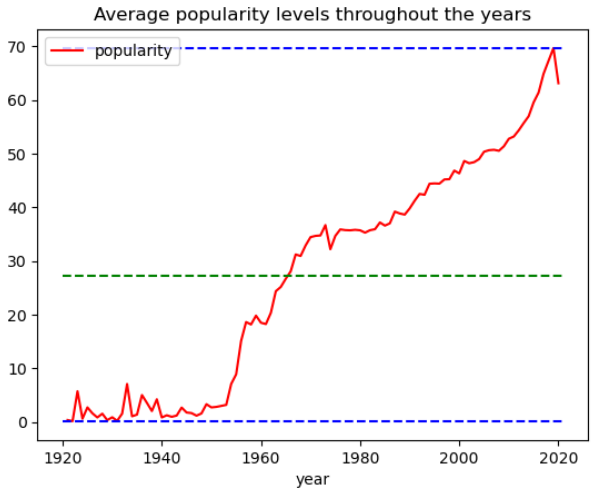

Figure : how popularity changes with the time

The hitfactor have nearly the same behavior over time as popularity but in the period between **[1920-1940]** the popularity of the songs is quite constant While the hitfactor is having ups and downs, So **NO** the hitfactor is not totally controlling the popularity

The question now is : why the popularity did not follow the hitfactor perfectly ? what are the factors that control the popularity with the hitfactor ? why does the graph of hitfactor go so much down in the 30's ?

3.2 Hypothesis 1

we assume that the popularity did not go down with the hitfactor because other features were compensating the difference To see if this is true of false we will try to see the behavior of the other features [speechiness / tempo /liveness / valence] while the thirties (1920-1930) - if their behavior is different than usual, that explain everything otherwise the hypothesis is wrong

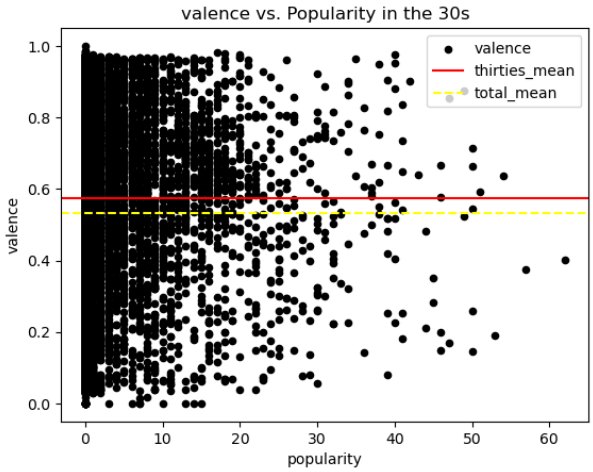

Figure : Valence behavior

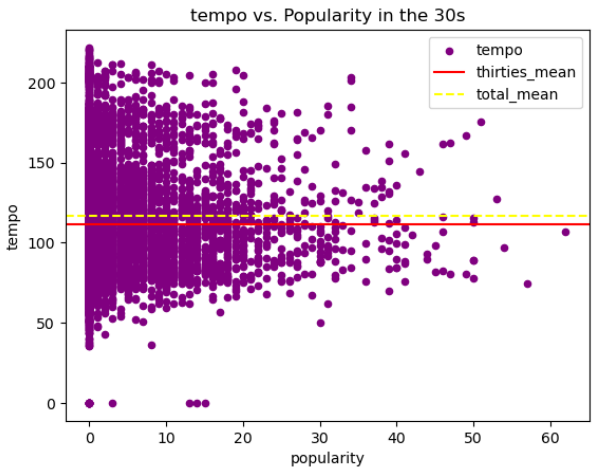

Figure : Tempo behavior

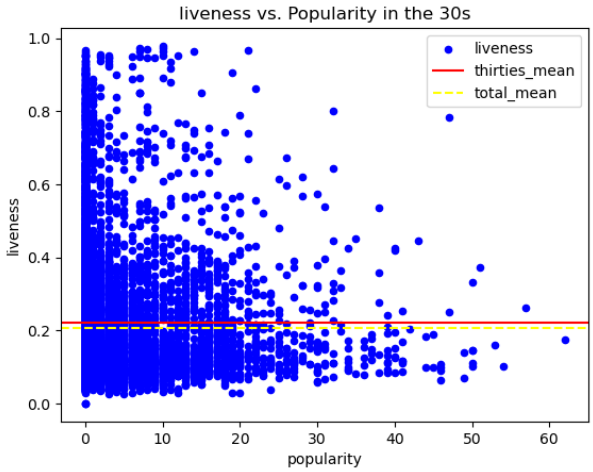

Figure : liveness behavior

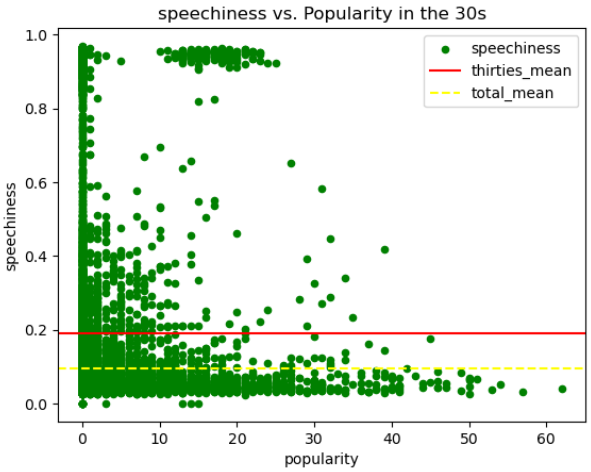

Figure : Speechiness behavior

3.3 Conclusion 1

Hypothesis 1 was not right because [speechiness / tempo /liveness / valence] behavior is nearly the same in the thirties with the general behavior

3.4 Hypothesis 2

here we assume that the reason why the hitfactor is going down in the thirties unlike the popularity is that the musical tastes in the 30's were a little different so the "genre" of music which was popular in the 30's was maybe (not very loud / based on acousticness/ ...) So to see what type of music was popular in the 30's we have to add a "genre" column to our data-set and finish the study

3.5 Training a Classifier on "genre"

because we don't have a "genre" column on the original data-set we decided to search for an external data-set that have the "genre" column and train a classifier on it so that we can apply that model on our original data-set to predict the "genre" of the songs we have.
the data-set we found have **42305 rows × 23 columns** but to make the prediction possible afterwards we kept only the columns that we have in our original data-set and we added a new column to convert the 'genre' feature to integers so that we can learn it and predict it after that.

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	genre	genre_int
0	0.831	0.814	2	-7.364	1	0.4200	0.059800	0.013400	0.0556	0.3890	156.985	124539	4	Dark Trap	0
1	0.719	0.493	8	-7.230	1	0.0794	0.401000	0.000000	0.1180	0.1240	115.080	224427	4	Dark Trap	0
2	0.850	0.893	5	-4.783	1	0.0623	0.013800	0.000004	0.3720	0.0391	218.050	98821	4	Dark Trap	0
3	0.476	0.781	0	-4.710	1	0.1030	0.023700	0.000000	0.1140	0.1750	186.948	123661	3	Dark Trap	0
4	0.798	0.624	2	-7.668	1	0.2930	0.217000	0.000000	0.1660	0.5910	147.988	123298	4	Dark Trap	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
42300	0.528	0.693	4	-5.148	1	0.0304	0.031500	0.000345	0.1210	0.3940	150.013	269208	4	hardstyle	9
42301	0.517	0.768	0	-7.922	0	0.0479	0.022500	0.000018	0.2050	0.3830	149.928	210112	4	hardstyle	9

FIGURE 5 – new data information

we used different machine learning algorithms to train the model and we took the one making the best score in term of precision

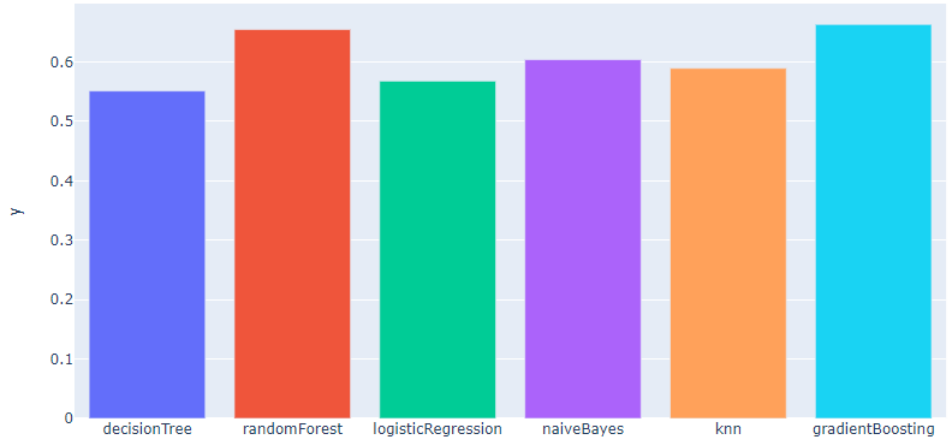

FIGURE 6 – precision of the different ml-algorithms on the external data-set

We decided to go with the **Gradient-Boosting** with **66% precision** and apply it on the original data

3.6 Understanding the "genre" column

After applying the model on the original data we got a "genre" column, so our goal now is to understand how does it devises the data and try to use it to answer out question of the **hypothesis 2**

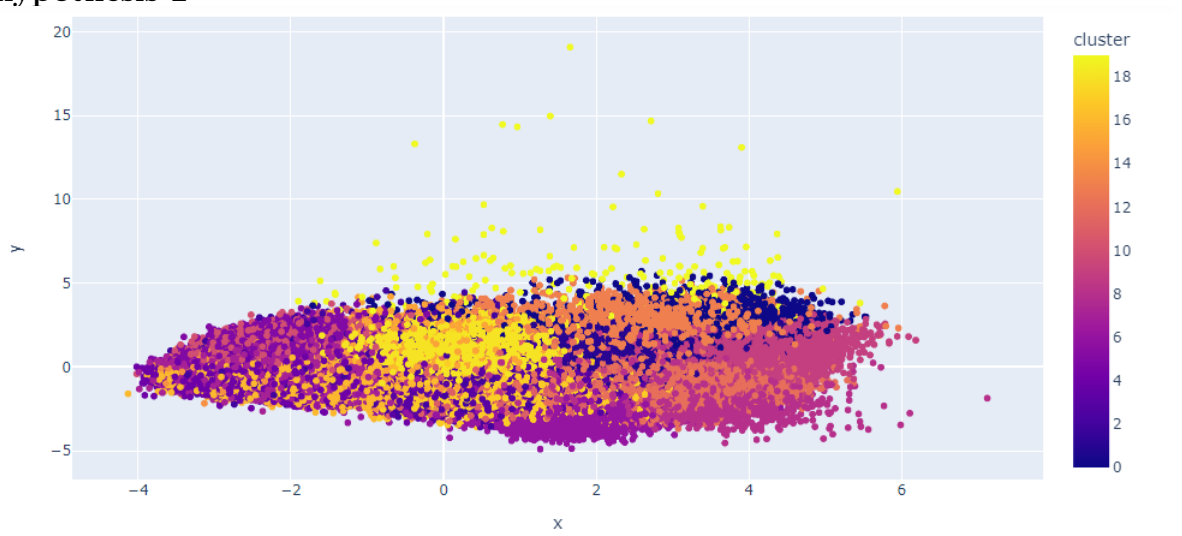

FIGURE 7 – clustering of the original data-set based on the "genre"

3.7 Conclusion 2

to verify if hypothesis 2 is true or not we have to answer these 2 questions first :

3.7.1 what characteristics have every genre ?

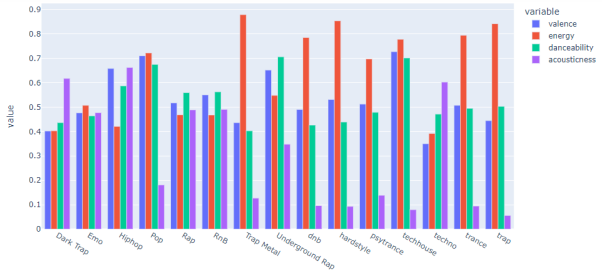

Figure : different features of every genre

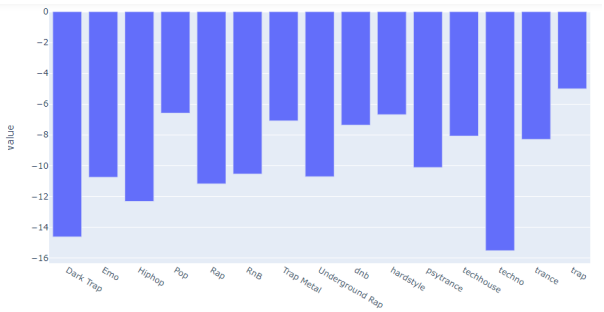

Figure : loudness of every genre

3.7.2 which "genre" was popular in the thirties ?

the most popular "genre" of music in the thirties was "hip-hop/rap" and as we can see in the previous plots the "energy" and "loudness" in the "Hip-hop/rap" type is very low and we did see in the table of correlations that there is a strong positive correlation between

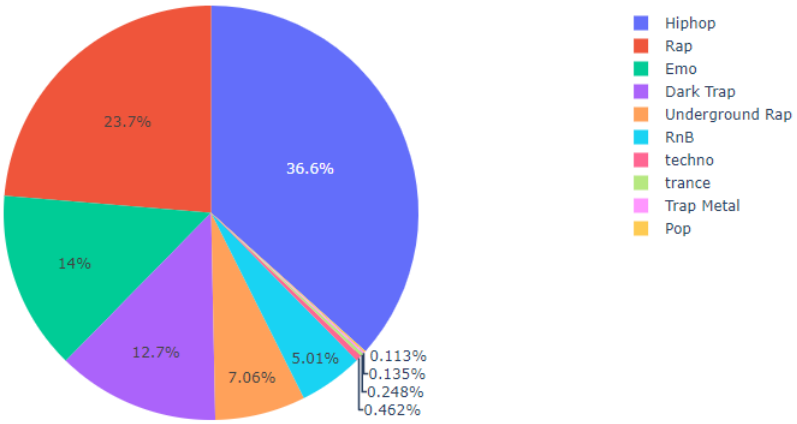

FIGURE 8 – Top genres in the 30's

"energy/popularity" and "loudness/popularity" that explain why the hitfactor was going so much down : (because people were listening to "hip-hop / rap " so much and nearly nothing else. and because of the characteristics of "hip-hop / rap" (very low energy and loudness) the hitfactor goes so much down (and not the popularity)

4 Conclusion

In this Study we tried to define the factor that can make a popular song and at the end we found out that we might have models to predict what people could like but always there are some exceptions that make our predictions not very accurate .
In our case it was the people's taste that changed in a certain time **(30's)** or it might be also the artists and the type of art they are making have changed.
we must also point out that our "genre" column was only **66% correct** so we are not certain of our results.
We tried to make some **hyper-parameter tuning** and also tried to **detect the outliers** in the external data and delete them to increase the precision of the **Gradient-Boosting** algorithm but it didn't increase it unfortunately

5 Bibliography

— https ://www.kaggle.com/code/swapnalvarma/spotify-data-analysis/data
(the new data trained)

— https ://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features **(spotify API)**

— https ://www.python-graph-gallery.com/ **(the plots ideas)**

— https ://plotly.com/python/plotly-express/
(plotly-express for the groupped bar plots)

— https ://www.kaggle.com/code/jaminjamin/comparing-ml-models-for-predicting-song-popularity/notebook
(inspiration for the classifier implementation)

— https ://numpy.org/doc/ **(numpy documentation)**

— https ://pandas.pydata.org/docs/ **(pandas documentation)**