Abstract:

In this project, I got dataset that contains more than 160.000 songs collected from Spotify Web API.

In the first section, I explored the database. First, I import the data from a CSV file to Python and I explored the data in statistical aspects using Pandas library which helped me find interesting information about the dataset. Then I proceeded to more advanced analysis and used more advanced functions to learn more meaningful information about my data, do advanced analytics, and present it with visualization so I could see the results more clearly and see the data behavior. In order to use the machine learning and deep learning algorithm, I performed pre-processing to convert the raw data to a sustainable format for analysis. Then, I perform a re-analysis of the data in order to draw new conclusions about the relevant information.

In the last section, I used the Dataset after pre-processing and built a 'Kmeans' model for the unsupervised learning algorithm and a Decision Tree model for the supervised learning algorithm. I used my previous knowledge to arrange the data to better fit the model.

Data details:

Primary:

- id (Id of track generated by Spotify)

Numerical:

- acousticness (Ranges from 0 to 1)
- danceability (Ranges from 0 to 1)
- energy (Ranges from 0 to 1)
- duration_ms (Integer typically ranging from 200k to 300k)
- instrumentalness (Ranges from 0 to 1)
- valence (Ranges from 0 to 1)
- popularity (Ranges from 0 to 100)
- tempo (Float typically ranging from 50 to 150)
- liveness (Ranges from 0 to 1)

- loudness (Float typically ranging from -60 to 0)

- speechiness (Ranges from 0 to 1)

- year (Ranges from 1921 to 2020)

Dummy:

- mode (0 = Minor, 1 = Major)

- explicit (0 = No explicit content, 1 = Explicit content)

Categorical:

- key (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on…)

- artists (List of artists mentioned)

- release_date (Date of release mostly in yyyy-mm-dd format, however precision of date may vary)

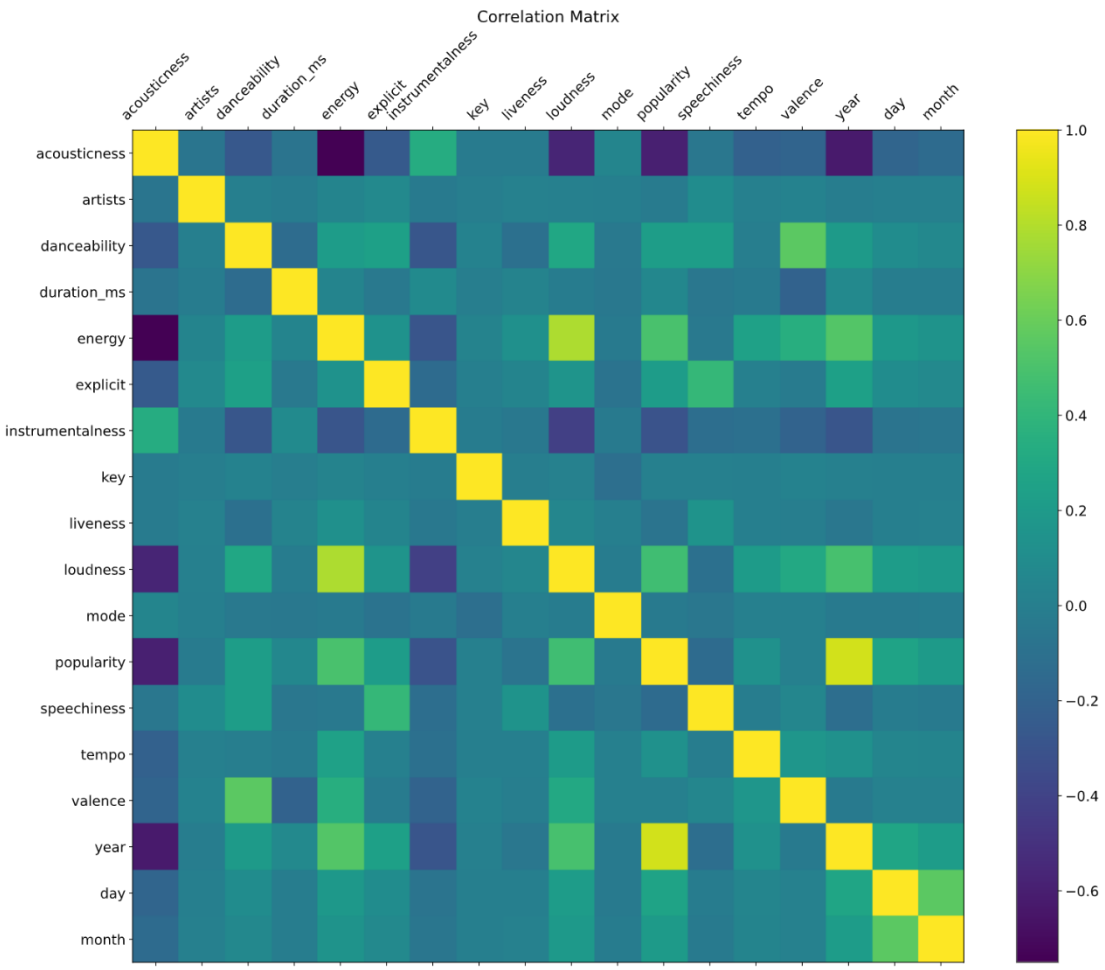
- name (Name of the song)

descriptive details:

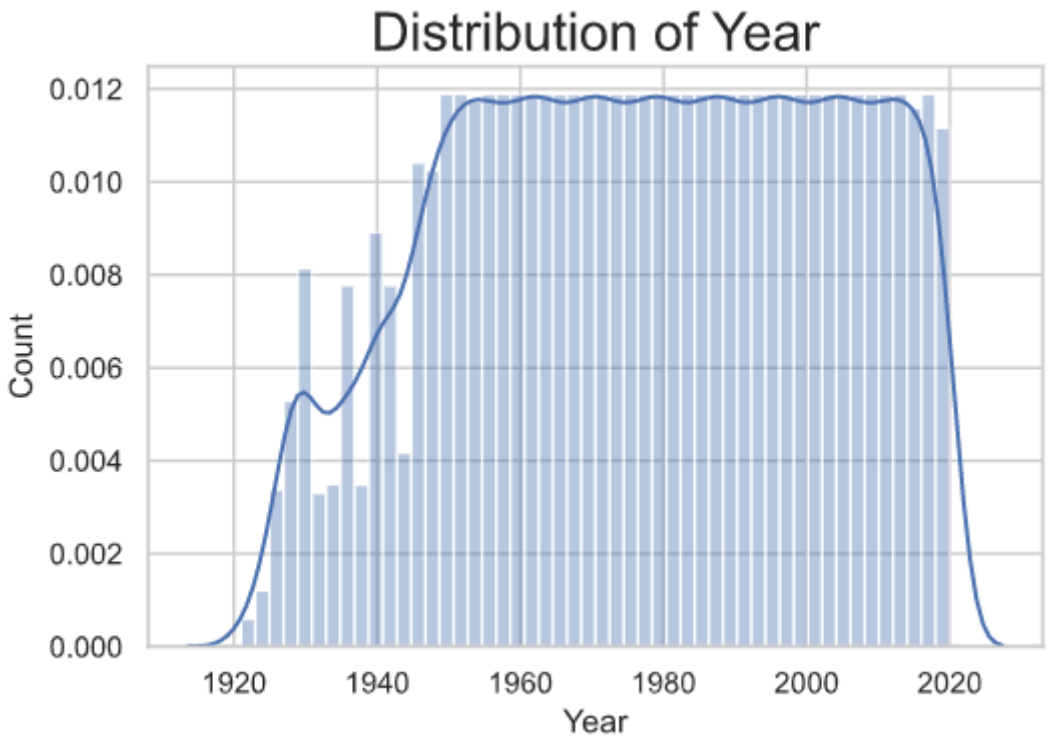
	acousticness	danceability	duration_ms	energy	explicit	instrumentalness	key	liveness	loudness
count	169909.000000	169909.000000	1.699090e+05	169909.000000	169909.000000	169909.000000	169909.000000	169909.000000	169909.000000
mean	0.493214	0.538150	2.314062e+05	0.488593	0.084863	0.161937	5.200519	0.206690	-11.370289
std	0.376627	0.175346	1.213219e+05	0.267390	0.278679	0.309329	3.515257	0.176796	5.666765
min	0.000000	0.000000	5.108000e+03	0.000000	0.000000	0.000000	0.000000	0.000000	-60.000000
25%	0.094500	0.417000	1.710400e+05	0.263000	0.000000	0.000000	2.000000	0.098400	-14.470000
50%	0.492000	0.548000	2.086000e+05	0.481000	0.000000	0.000204	5.000000	0.135000	-10.474000
75%	0.888000	0.667000	2.629600e+05	0.710000	0.000000	0.086800	8.000000	0.263000	-7.118000
max	0.996000	0.988000	5.403500e+06	1.000000	1.000000	1.000000	11.000000	1.000000	3.855000

	mode	popularity	speechiness	tempo	valence	year
count	169909.000000	169909.000000	169909.000000	169909.000000	169909.000000	169909.000000
mean	0.708556	31.556610	0.094058	116.948017	0.532095	1977.223231
std	0.454429	21.582614	0.149937	30.726937	0.262408	25.593168
min	0.000000	0.000000	0.000000	0.000000	0.000000	1921.000000
25%	0.000000	12.000000	0.034900	93.516000	0.322000	1957.000000
50%	1.000000	33.000000	0.045000	114.778000	0.544000	1978.000000
75%	1.000000	48.000000	0.075400	135.712000	0.749000	1999.000000
max	1.000000	100.000000	0.969000	244.091000	1.000000	2020.000000

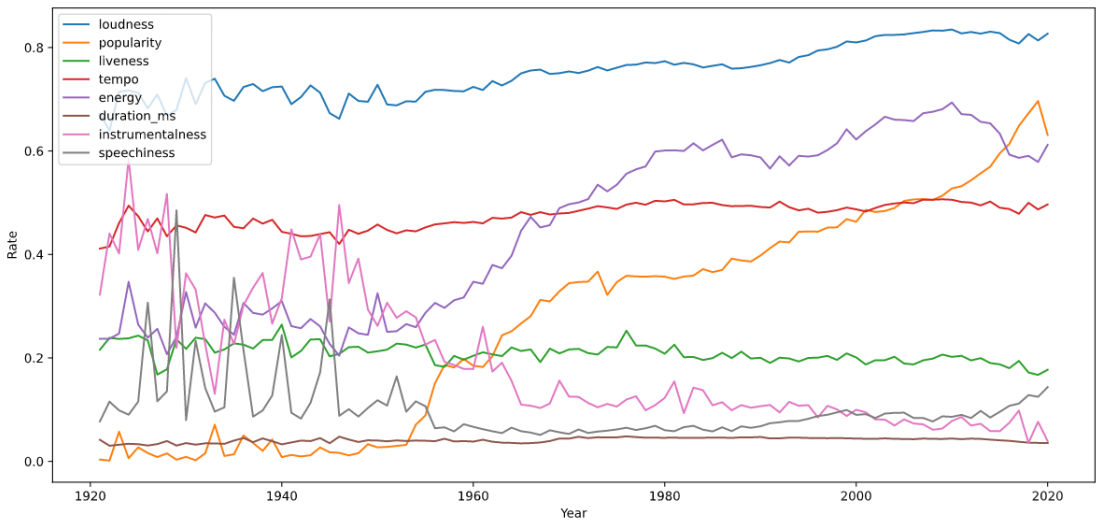
The graph below shows the correlation between each two columns. There is strong correlation between year and popularity and between loudness and energy.



The graph below shows the release date's distribution over the years. Most of the songs in the dataset are between 1950 to 2020.



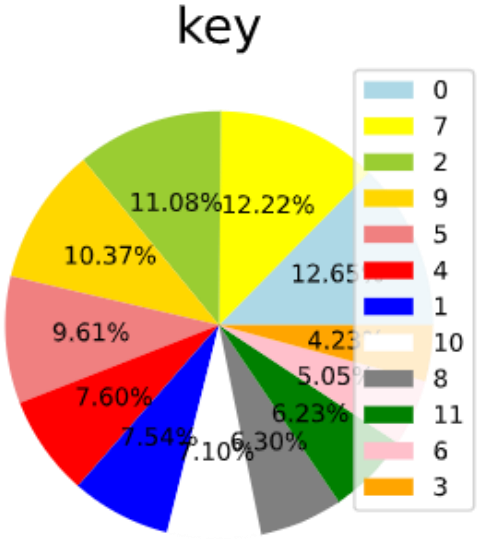
The graph below shows the regression of factors over the years. The loudness, energy, popularity and tempo are increase, and the instrumentalness, liveness and speechiness decrease over the years. The duration remains constant. From this graph and according to the correlation graph , we can learn that when the loudness is increase the energy is increase (and the opposite), and the energy and popularity both increase drastically over the years.



The output below shows the songs popularity average of each key. The most popular (38) key is 1 in 0 (minor) mode and the most unpopular (23) is 3 in 1 mode (major).

```
mode  key
0     0      29.121667
      1      38.175475
      2      27.521559
      3      28.761965
      4      34.174586
      5      30.420311
      6      37.694821
      7      28.770787
      8      34.490385
      9      32.915916
      10     33.600941
      11     35.456159
1     0      31.225350
      1      33.580017
      2      33.117899
      3      23.314990
      4      32.409233
      5      27.650991
      6      32.641608
      7      31.906789
      8      29.713910
      9      33.277764
      10     26.893159
      11     33.924113
Name: popularity, dtype: float64
```

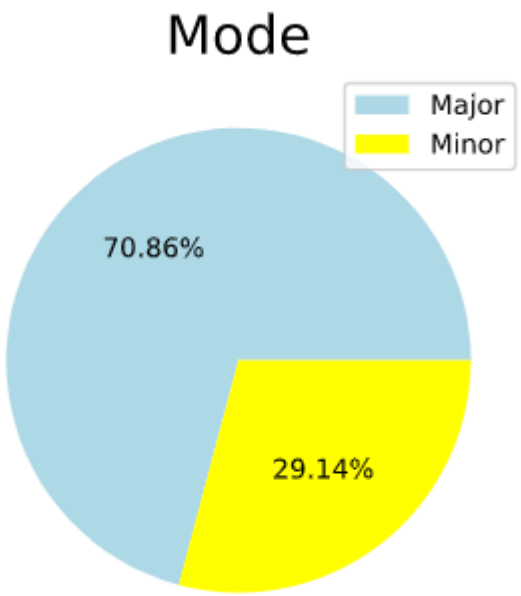
The graph below shows the musical key's distribution of songs. The most common key is 0 and the most uncommon is 3.



The output below shows the tracks popularity average of each month. The most popular month is 5 (May) and the most unpopular is 12 (December). From this output, we can learn that the unpopular songs release around the end of the year and the popular around the middle of the year.

```
month
1      0.249088
2      0.410828
3      0.409145
4      0.424634
5      0.429527
6      0.404876
7      0.383070
8      0.414713
9      0.419573
10     0.419076
11     0.396811
12     0.212638
```

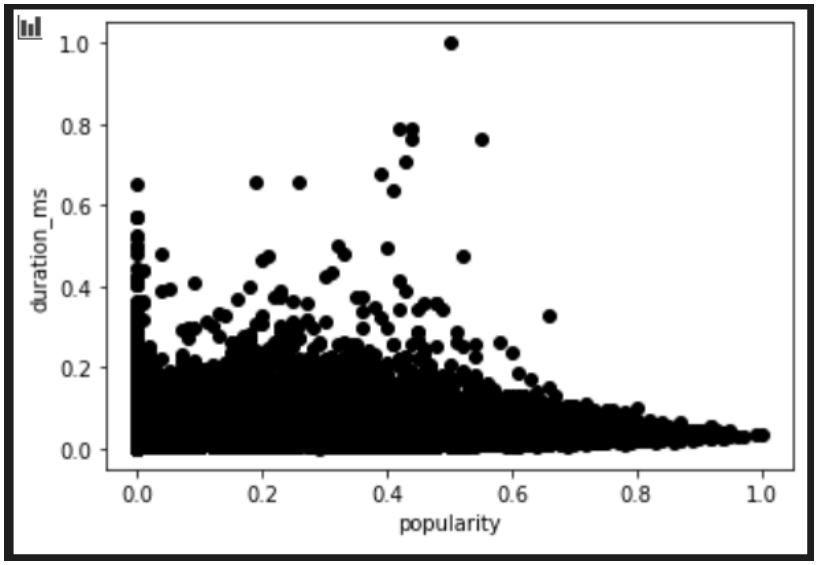
The graph below shows the song mode's distribution – major or minor. 70.86 percent of the songs are in major mode and 29.14 percent are minor mode. From this graph and according to the key-popularity output, we can learn that the most of the songs wrote in major key but the minor songs are much more popular.



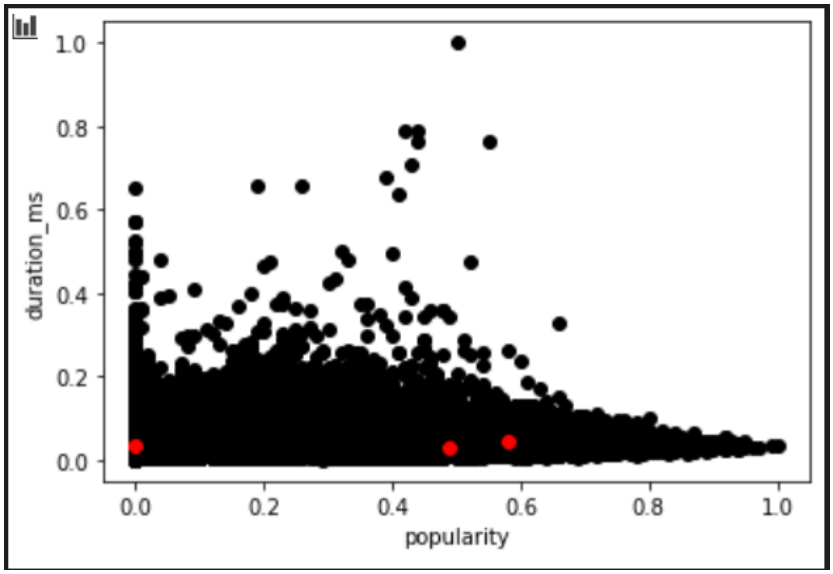
Unsupervised Learning – K-mean

The question for the unsupervised learning is: Can I split my data to 3 clusters such that every cluster is including popularity rate with the same song duration?

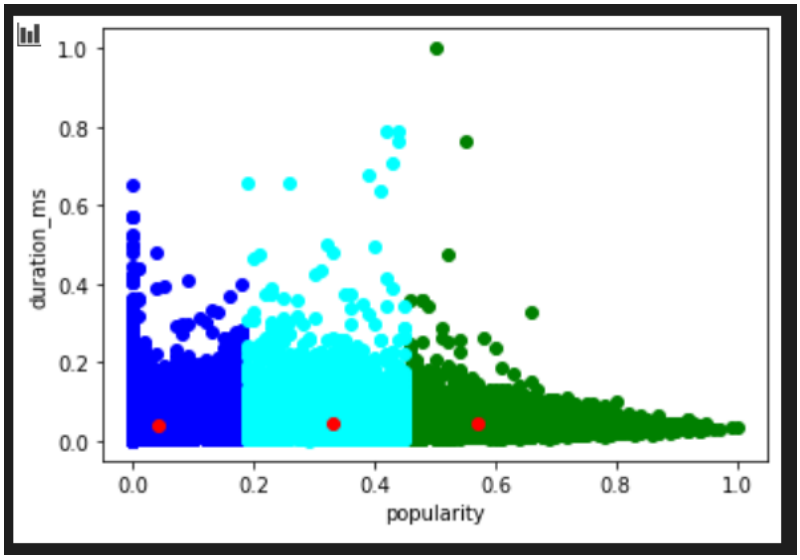
I took two variables from the data – “duration_ms” and “popularity”. I picked these two variables and visualize the data points:



Steps 1 and 2 of K-Means were about choosing the number of clusters (3) and selecting random centroids for each cluster.



Next, visualize the clusters i have got:



The red dots represent the centroid of each cluster.

Supervised Learning – Decision Tree

The question for the unsupervised learning is: Can i predict the release year of track by the rate of Danceability/Duration/Energy/Loudness/Tempo?

1. Decision Tree for Classification

In this section i predicted the release year of song depending upon different attributes of the song.

Evaluating the Algorithm

At this point I have trained my algorithm and made some predictions. Now I see how accurate my algorithm is by used of confusion matrix, precision, recall, and F1 score.

	precision	recall	f1-score	support
29	0.24	0.21	0.27	1921
17	0.11	0.12	0.11	1922
42	0.28	0.26	0.31	1923
47	0.23	0.23	0.22	1924
54	0.27	0.28	0.25	1925
186	0.50	0.48	0.51	1926
111	0.38	0.40	0.37	1927
228	0.50	0.49	0.51	1928
199	0.53	0.51	0.56	1929
accuracy			0.16	33982
macro avg	0.18	0.18	0.18	33982
weighted avg	0.16	0.16	0.16	33982

From the confusion matrix, you can see that out of 33,982 test instances, my algorithm misclassified 28,544. This is 16 % accuracy.

2. Decision Tree for Regression

For regression I use Decision Tree Regressor class of the tree library.

To evaluate performance of the regression algorithm, the commonly used metrics are mean absolute error, mean squared error, and root mean squared error.

	Actual	Predicted
34530	1963	1958.0
54303	2007	1997.0
119544	1954	1957.0
132124	1986	1982.0
165993	1981	1969.0
...	...	...
27979	1978	1980.0
108428	1937	1943.0
83427	1974	1977.0
67859	1992	1936.0

Mean Absolute Error: 8.999151511584564

Mean Squared Error: 172.4670765241729

Root Mean Squared Error: 13.132672101448849

The mean absolute error for our algorithm is 8.9, which is less than 10 percent of the mean of all the values in the 'Year' column. This means that my algorithm did a fine prediction job.

Conclusions

From the results of the graphs , it can be concluded that there is connection between the popularity to the features of the song. We can see the feature affect each other and specially on the popularity. According to the rate / year regression graph, it can be seen from the data that the popularity rank of the songs is rising rapidly starting around 1950 and the other features starting change as well. The energy, loudness and tempo increase compared to instrumentalness, liveness and speechiness that decrease, and it is make sense according to the music that has become more electronic over the years. From the key / mode graphs, it can be concluded that the most of the songs wrote in major scale that consider to happy scale compared to the minor that consider to sad scale that it makes sense but, the popular rank of the minor scale is higher. The reason could be that minor keys are sometimes said to have a more interesting, possibly darker sound than plain major scales.