Music Influencer Model

Name

Elaine He, Teng Liang, Richard Yang

Abstract

Music has been and still is a very important aspect of society. From the earliest records of civilization, music has perpetuated throughout culture and tradition. However, throughout all these thousands of years, one must ponder how music has evolved over the years. After centuries, what is the result of this evolution? What are the trends that led to this? What makes specific genres and music popular? What kind of music does the majority of people like? These are the questions to which our research attempts to answer. The previous research on this topic either excludes the musical trend throughout the years or is too vague on the description of the songs. This is significant because music takes years to demonstrate any notable patterns and extreme detail is needed to examine these patterns thoroughly. We aim to answer these inquiries with detailed evidence and still be broad enough to include yearly trends and patterns. Our methods and approach toward this target are by utilizing quality data descriptive methods to derive a clear trend in the musical data. Methods such as heat maps and regression graphs will aid in determining trends. Furthermore, we will use models such as OLS and TensorFlow to determine which variables and features affect the popularity of music the most. Categorizing the data by dividing each genre into a percentage of the whole music industry, and comparing the changes year by year, we were able to identify important trends and patterns throughout the music industry in the last decade. Important findings include the overall dominance of pop/rock and the recent rise of R&B, parallel with the decline of country music. Finally, we ran XGboosting on the features and found that acousticness has the most impact on popularity, followed by loudness. What is significant about our findings is that we answered our two main questions regarding the evolution of music and the features which impact its popularity, driving that evolution.

Music Influencer Model

In a CNN article, it was stated that “music is present in every part of our lives. Our spiritual rituals are framed with songs, children learn the alphabet through song and the malls and cafes we visit during our leisure time are rarely silent”. With music surrounding peoples’ everyday lives, this study will investigate what influences the popularity and genre of music throughout the years. We pose the question of whether the mood of the music affects the mood of the society, comparing trends throughout the years. We also want to explore the influence of musicians as influencers with the thought of social media.

There has been an abundance of previous research regarding the influence of music, however, these studies mostly take psychological and sociological perspectives. Therefore, we want to delve into this topic using analysis of data. There are some intriguing points regarding this topic we would like to examine. Recent popular music artworks appearing on the market seem to be labeled as “negative”, “sad”, “dark” more in comparison to the old musical works that usually would be considered as more positive. Is there a trend that can be reflected on the data that people are more inclined to listen to more melancholy music? In the dataset we use, there is a characteristic in the data called “danceability”. We expect it to keep growing as social media like Tik Tok or Instagram have become viral.

Hypothesis

We hypothesize that loudness and energy impact popularity positively while accousticness and instrumental impacts popularity negatively.

Method

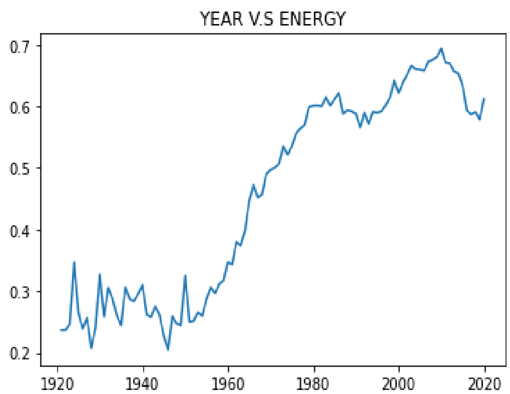
Dataset

For our research, we are utilizing a music dataset containing four subsets which are full music data, data by artist, data by year, and influence data. These four subsets incorporate the variables artist name, influencer name, active start, year, genre, popularity, danceability, energy, valence, tempo, mode, key, acousticness, instrumentalness, liveness, speechiness, duration. Each subset incorporates these variables in a different manner. All four subsets will be displayed in the appendices. As an example, Figure 1 contains the first 5 rows of the subset of data by year.

	year	danceability	energy	valence	tempo	loudness	mode	key	acousticness	instrumentalness	liveness	speechiness	duration_ms	popularity
29	1950	0.491433	0.324979	0.551834	111.768489	-13.506982	1	7	0.866723	0.262045	0.212791	0.118183	222606.0595	2.703500
30	1951	0.459766	0.250260	0.431762	109.183712	-15.948836	1	0	0.904933	0.306456	0.215993	0.107834	214660.0035	2.799000
31	1952	0.462659	0.251581	0.436885	107.591127	-16.055367	1	5	0.861886	0.277559	0.227615	0.164197	223649.8685	2.986500
32	1953	0.436234	0.265310	0.423354	109.014178	-15.551768	1	0	0.891139	0.290270	0.225679	0.096054	217921.8221	3.177436
33	1954	0.466630	0.259361	0.446452	108.532237	-15.594470	1	5	0.867919	0.278372	0.219946	0.115995	222630.2310	7.083500

Figure 1: Data by year

As a note, we've removed data from 1920 to 1950. Our reasoning is that this was a period of turbulence, as it encompasses the period from which World War I ended to when World War II ended. We hypothesized that not a lot of musical records were recorded well during this time, and the few that were recorded, would not represent the whole musical atmosphere. Therefore, by cutting out data from this period of time, the noise in this dataset is significantly reduced. For example, Figure 2 and Figure 3 display the changes that occurred as a result of removing the period in question.



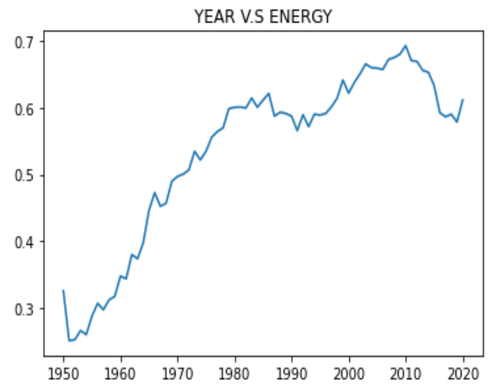


Figure 2: Year vs. Energy including 1920-1950

Figure 3: Year vs. Energy excluding 1920-1950

Pretest

Before conducting the research, we ran some preliminary tests on the dataset using regression and correlation models. These models reveal some trends and developments of music throughout the years.

One of the first tests that we ran was a comparison of the appearances of different genres throughout the years. We did this by comparing each genre's percentage makeup of the whole music industry year by year. To be more specific, we group by the musician genres based on year and then count it by percentage. As a result, we came up with the following graph.

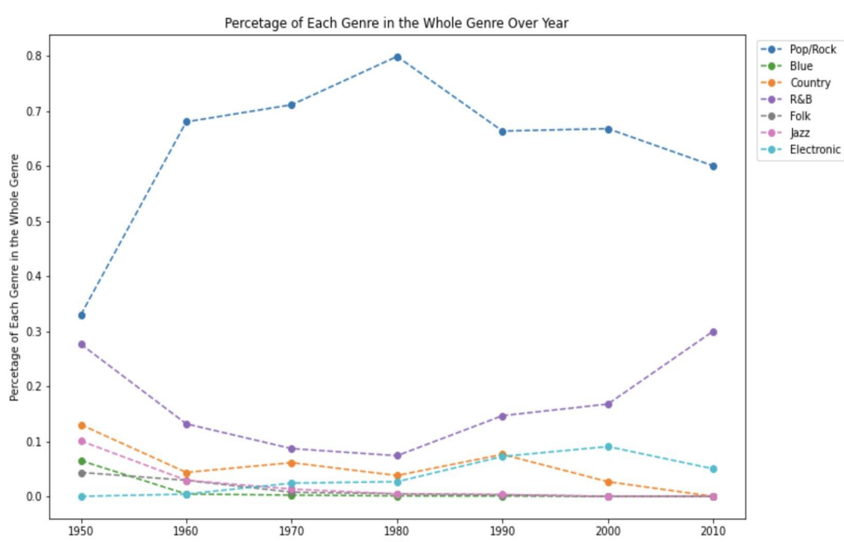


Figure 4: Percentage of each genre in the whole music industry over year

Figure 5 reveals the trend of musical genres throughout the years. It is clear that op/rock has been the most popular for pretty much all the time. Some other clear trends include the rise of R&B since 1980, the constant decline in jazz throughout the whole decade, and the recent decline in the country for the last 20 years. This particular graph is important because it shows the rise and fall of genres throughout the years, also bringing up a potential hypothesis that the music today is less diversified than before, as Pop/Rock, R&B, and Electronic music account for over 95% of all the musician genres today. Furthermore, it is evident to note that this graph doesn't measure popularity, instead of the number of music in each genre that is released every year. In a way, it could be interpreted as popularity, but we will take a closer look into what affects popularity.

The correlation heatmap reveals the relationship between the different features in the dataset. As seen in Figure 1, some notable correlations that affect popularity are as follows: loudness, energy, acousticness, and instrumentalness. It is clear that loudness and energy have a strong and positive relationship with popularity, while acousticness and instrumentalness have a negative relationship with popularity.

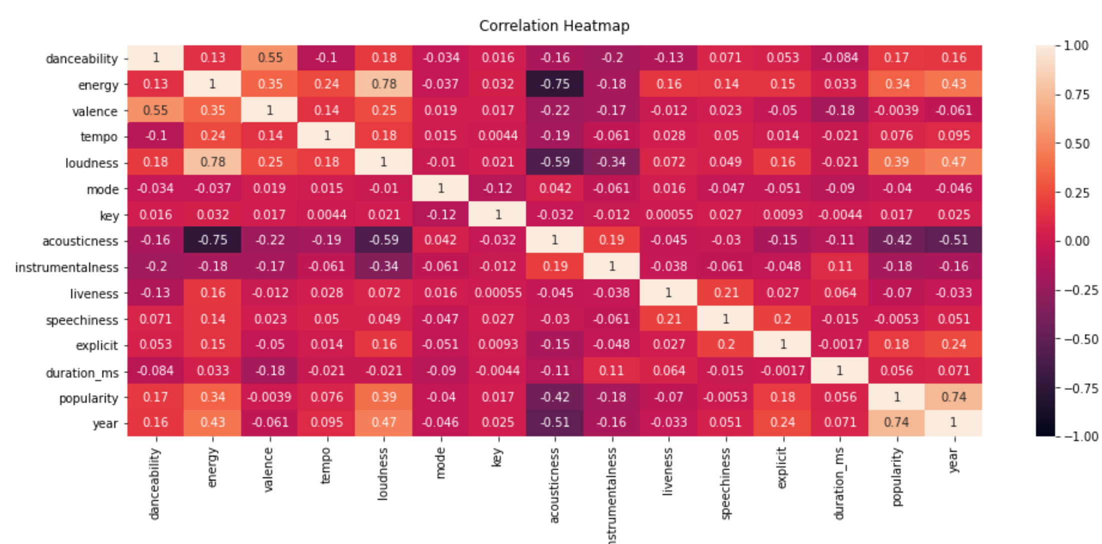


Figure 5: Correlation Heatmap between the features

We also ran simple regression with the year being the independent variable, and the other musical features in the dataset being the dependent variable. The simple regression graphs demonstrate some clear trends in music throughout the years. All the simple regression graphs will be included in the appendices. Notably, popularity demonstrates an evident upwards trend throughout the years with the coefficient being 0.93540. This could be because music has become more accessible in recent years, and this could be an indicator that music is also gaining more influence. The graphs for danceability, loudness, and energy also demonstrate a clear upwards trend throughout the years with the coefficients being 0.75499, 0.90251, and 0.52014 respectively. The graphs for acousticness and instrumentalness display a downward trend throughout the years, with the coefficients being 0.41263 and 0.73870. For other dependent variables, the trend was less clear. And this could be an indication that those musical features are not affected in the long term throughout the years.

It should be noted that although these variables do not display a clear rising or falling trend throughout the time period from 1950 to 2020, they do show trends when the time period is shortened. For valence, it demonstrated a rising trend from 1950 to 1980 and then demonstrated a decreasing trend from 1980 to 2020. This results in a very low coefficient of 0.03494. This indicates that regression models may not fit all the variables in the dataset. And for certain variables, the trend may be variable, including both positive and negative growth, from the time period of 1950-2020.

Design

In our procedure, we are running an experimental research design. Our dependent variable is popularity while our independent variables are the 15 independent features ('danceability', 'energy', 'valence', 'tempo', 'loudness','key', 'acousticness', 'instrumentalness', 'liveness', 'speechiness', 'explicit','popularity', 'year', 'release_date', 'song_title (censored)') . Our goal is to test which of the 15 features has the most notable and significant impact on popularity.

Procedure

We used the correlation heatmap and simple regression to preliminary examine the relationship between variables in the dataset. We continued with running an OLS multiple regression to test the hypothesis.

Results

OLS Regression Results

Dep. Variable:	popularity	**R-squared:**	0.977
Model:	OLS	**Adj. R-squared:**	0.973
Method:	Least Squares	**F-statistic:**	231.5
Date:	Fri, 13 Aug 2021	**Prob (F-statistic):**	3.15e-44
Time:	06:07:55	**Log-Likelihood:**	-161.88
No. Observations:	71	**AIC:**	347.8
Df Residuals:	59	**BIC:**	374.9
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	170.7117	39.130	4.363	0.000	92.412	249.011
danceability	87.2597	24.775	3.522	0.001	37.685	136.835
energy	-80.8659	30.635	-2.640	0.011	-142.167	-19.565
key	-0.1823	0.122	-1.495	0.140	-0.426	0.062
valence	-54.1954	12.395	-4.372	0.000	-78.998	-29.392
tempo	-0.2322	0.271	-0.857	0.395	-0.775	0.310
loudness	2.6294	0.723	3.636	0.001	1.183	4.076
acousticness	-68.1380	14.455	-4.714	0.000	-97.062	-39.214
instrumentalness	-41.6523	18.467	-2.255	0.028	-78.605	-4.700
liveness	-4.7708	34.778	-0.137	0.891	-74.362	64.820
speechiness	-95.6493	30.444	-3.142	0.003	-156.568	-34.731
duration_ms	-4.31e-05	2.87e-05	-1.501	0.139	-0.000	1.44e-05

Omnibus:	6.456	**Durbin-Watson:**	1.536
Prob(Omnibus):	0.040	**Jarque-Bera (JB):**	6.265
Skew:	0.490	**Prob(JB):**	0.0436
Kurtosis:	4.077	**Cond. No.**	3.71e+07

Figure 6: OLS regression with Popularity as the dependent variable

We ran an OLS multiple regression model with popularity as the dependent variable, and all other variables in the data by year subset. The r-squared value for this regression is rather strong, being 0.977. This value indicates that the independent variables in this model explain 97.7% of the variance of the dependent variable. The intercept for this regression is 170.7117, which indicates when the coefficients of the independent variables are 0, popularity will be 190.7117. For examining the coefficient of the dependent variable, a one-unit increase in the independent variables is associated with a coefficient increase in popularity. For example, a one-unit increase in danceability is associated with an 87.2597 increase in popularity. The p-values in the OLS output indicate the probability of getting this coefficient if the true coefficient is 0. We will be using 0.05 as our significance level. For our regression output, the p-values for intercept, danceability, energy, valence, loudness, acousticness, instrumentalness, and speechiness, the p-values are under the significance level of 0.05. And the p-values for key, tempo, liveliness, and duration are larger than the significance level 0.05.

To further test the importance of features that affect popularity, we apply XGBoosting to evaluate the feature importance of the independent variables. Three methods (weight, cover, gain) are used to illustrate the most important features of popularity. The graphs are shown below. As we can see, the two methods show three different outcomes. “Loudness” is the most important feature under method “Weight”, while “Accousticness” is the one under both methods “Weight” and “Gain”. It’s reasonable to deduce that “Accousticness” can be the most important variable affecting popularity. Next time we will examine this deduction from a more rigorous and mathematical perspective.

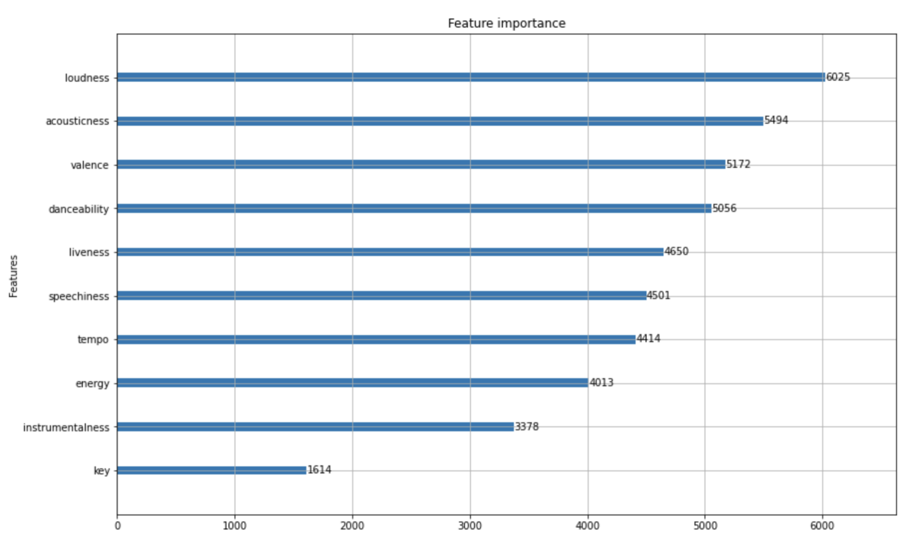


Figure 7: XGBoosting weight and features

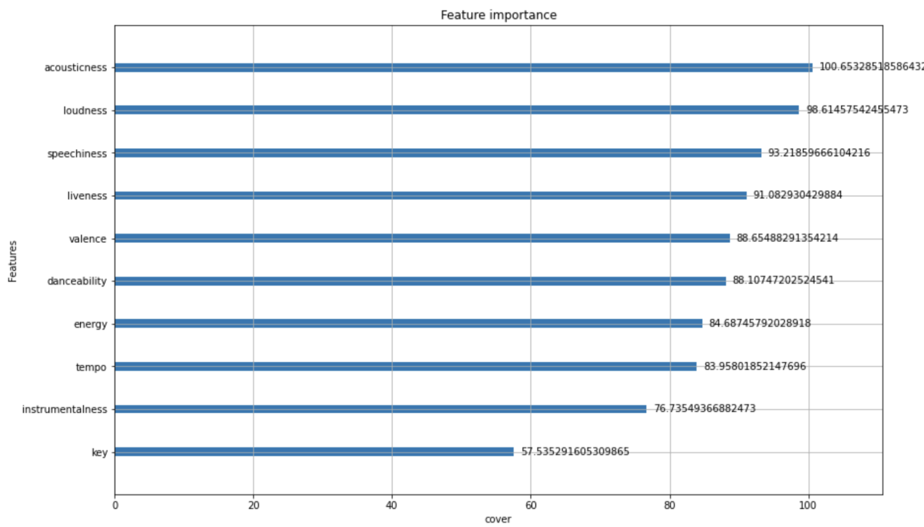


Figure 8: XGBoosting cover and features

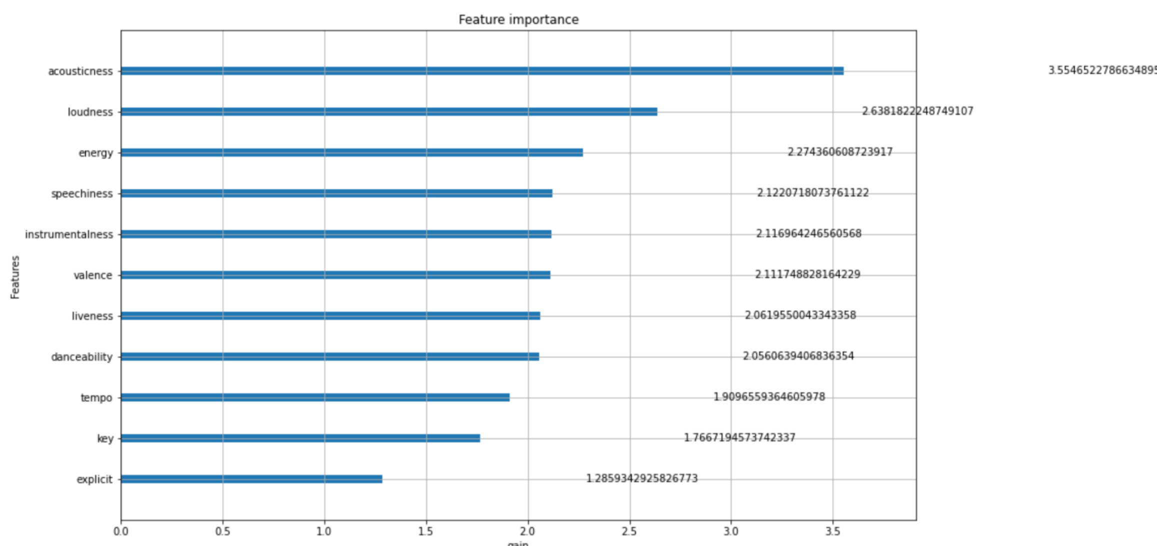


Figure 9: XGBoosting gain and features

Discussion

Limitations

One limitation of our research lies in the foundation of the dataset. We do not know how this dataset was obtained or the methods used to collect it. There could be certain biases in the collection of the data. Another limitation is the number of models that we have done. We have done graphs such as linear regression OLS, and other analyses of the data, but we would also like to do some Tensorflow. If time allows it, that is what we are planning to do in the future. Furthermore, we would like to use clustering to reveal subgroups within each genre. For instance, within pop/rock, there would be music that has both high energy and low energy. We categorize these two groups into subgroups of the pop/rock genre. Doing this will allow us to see the many different subgroups within each genre, and further analysis will lead to the finding of their impact on each respective genre and the music industry as a whole.

Future Implications

The following results that we've obtained through our tests have significant future implications. First, knowing which specific features impact popularity the most may assist future artists to make music that more people find enjoyable and pleasant. Artists may use specific tempos or tones to connect more with their audience and try out new approaches that they may not have tried before knowing these data points. This may enrich the musical genre greatly and lead to music being more popular and having a more significant impact on a person's life. As stated in the beginning, music has the capacity to affect mood and livelihood. If more enjoyable music is made, then more people would have an increase in mood and overall happiness.

To further our research, we would like to utilize additional subsets and utilize new models. We would like to use neural networks to run multiple regressions with popularity being the dependent variable and year and other variables being the independent variables, acquiring more sophisticated results. Combine social media data of these musicians with their own characteristics and analyze their influence. Using the data by artist and influencer data to build a model to analyze the following trend of different musical genres. Build a model categorizing the danceability, valence, tempo, and other technical aspects of music into different musical genres.

Appendices

Full Music Dataset

	artist_names	artists_id	danceability	energy	valence	tempo	loudness	mode	key	acousticness	instrumentalness	liveness	speechiness	explicit	duration_ms	popularity	year	release_date	song_title (censored)
0	["Fat Freddy's Drop"]	[178301]	0.600	0.365	0.131	130.046	-13.083	0	9	0.06720	0.585000	0.0921	0.0498	0	437200	54	2005	2005	Ernie
1	["Fat Freddy's Drop"]	[178301]	0.874	0.326	0.179	119.620	-13.302	0	11	0.01360	0.148000	0.0993	0.1310	0	581008	53	2005	2005	Wandering Eye
2	["Fat Freddy's Drop"]	[178301]	0.670	0.531	0.336	139.385	-8.267	0	9	0.01560	0.345000	0.3060	0.0377	0	431293	55	2009	8/7/2009	The Raft
3	["Alexander O'Neal"]	[625201]	0.761	0.702	0.850	104.773	-8.523	1	7	0.10800	0.000031	0.0935	0.0389	0	304427	34	1991	1/1/1991	All True Man
4	["Alexander O'Neal"]	[625201]	0.661	0.828	0.902	115.078	-12.673	0	11	0.27600	0.000001	0.2870	0.0390	0	264933	37	1987	7/29/1987	(What Can I Say) To Make You Love Me
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98335	['ZZ Top']	[690254]	0.276	0.892	0.715	80.475	-7.035	1	11	0.40900	0.000000	0.7140	0.0893	0	115973	33	1975	4/18/1975	Jailhouse Rock - **** Remaster
98336	['ZZ Top']	[690254]	0.700	0.592	0.906	109.847	-10.434	1	11	0.10800	0.018400	0.1180	0.0681	0	263627	32	1976	11/29/1976	It's Only Love
98337	['ZZ Top']	[690254]	0.709	0.709	0.863	111.544	-12.023	1	2	0.11800	0.000011	0.1250	0.0370	0	158400	31	1987	1987	Balinese
98338	['ZZ Top']	[690254]	0.552	0.651	0.533	161.548	-10.624	0	4	0.00494	0.034900	0.1320	0.2290	0	232533	43	1992	4/13/1992	La Grange - **** Remaster
98339	['ZZ Top']	[690254]	0.546	0.864	0.863	145.652	-7.632	1	7	0.09080	0.008140	0.2770	0.0393	0	137240	43	2005	7/19/2005	****

98340 rows × 19 columns

Influence Data

	influencer_id	influencer_name	influencer_main_genre	influencer_active_start	follower_id	follower_name	follower_main_genre	follower_active_start
0	759491	The Exploited	Pop/Rock	1980	74	Special Duties	Pop/Rock	1980
1	25462	Tricky	Electronic	1990	335	PJ Harvey	Pop/Rock	1990
2	66915	Bob Dylan	Pop/Rock	1960	335	PJ Harvey	Pop/Rock	1990
3	71209	Leonard Cohen	Pop/Rock	1950	335	PJ Harvey	Pop/Rock	1990
4	91438	The Gun Club	Pop/Rock	1980	335	PJ Harvey	Pop/Rock	1990
...	...	...	...	...	...	...	...	...
42765	580300	Sufjan Stevens	Pop/Rock	1990	3661738	Rosemary & Garlic	Pop/Rock	2010
42766	261309	Vybz Kartel	Reggae	2000	3670556	Trinidad Cardona	R&B;	2010
42767	467203	Michael Jackson	R&B;	1960	3670556	Trinidad Cardona	R&B;	2010
42768	2518003	Popcaan	Reggae	2000	3670556	Trinidad Cardona	R&B;	2010
42769	2896351	Tommy Lee	Reggae	2000	3670556	Trinidad Cardona	R&B;	2010

42770 rows × 8 columns

Data by Artist

	artist_name	artist_id	danceability	energy	valence	tempo	loudness	mode	key	acousticness	instrumentalness	liveness	speechiness	duration_ms	popularity	count
0	Frank Sinatra	792507	0.384478	0.238017	0.364288	110.181698	-14.271141	1	5	0.735648	0.020855	0.232106	0.049614	189179.9255	26.004383	1369
1	Vladimir Horowitz	119107	0.343210	0.118844	0.225951	94.900679	-23.193418	1	1	0.990070	0.879508	0.183812	0.043360	266541.1251	3.592378	1207
2	Johnny Cash	816890	0.619803	0.449381	0.680662	115.037747	-11.593104	1	10	0.685637	0.022647	0.242243	0.098216	162279.2672	26.614130	1104
3	Billie Holiday	79016	0.572637	0.201368	0.498934	109.912172	-13.225966	1	5	0.908499	0.013064	0.217727	0.062432	185131.4530	15.621005	1095
4	Bob Dylan	66915	0.512598	0.477932	0.551934	126.160149	-11.184330	1	7	0.562567	0.034211	0.308978	0.064535	256713.4203	30.860806	1092
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5849	Natalie La Rose	3359519	0.830000	0.520000	0.735000	104.990000	-8.714000	1	0	0.000792	0.000013	0.065600	0.037600	189907.0000	64.000000	1
5850	Sarah Ross	3381566	0.721000	0.944000	0.626000	85.002000	-5.982000	1	8	0.013000	0.000000	0.320000	0.159000	262760.0000	52.000000	1
5851	Rotimi	3410250	0.637000	0.501000	0.431000	103.993000	-6.148000	0	0	0.229000	0.000059	0.099000	0.187000	185461.0000	71.000000	1
5852	Jillian Jacqueline	3455945	0.547000	0.672000	0.283000	155.791000	-5.023000	1	11	0.304000	0.000000	0.099600	0.049600	213133.0000	58.000000	1
5853	Jaira Burns	3639618	0.566000	0.769000	0.385000	170.036000	-4.342000	1	7	0.018300	0.000000	0.108000	0.087200	191100.0000	74.000000	1

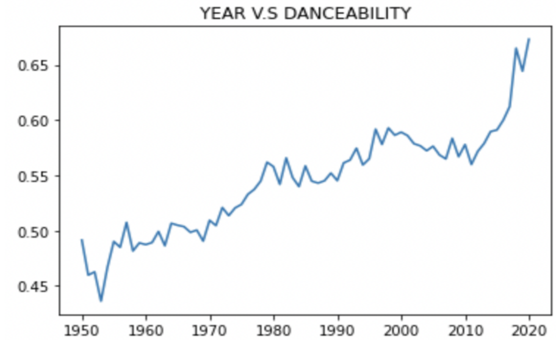
5854 rows × 16 columns

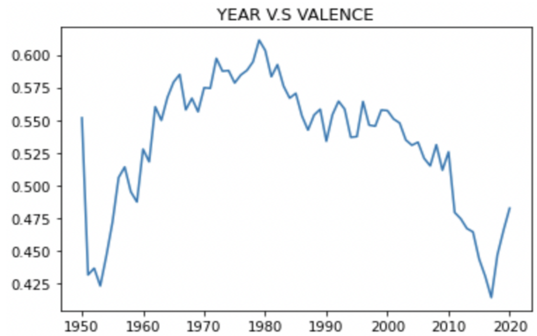
Data by Year

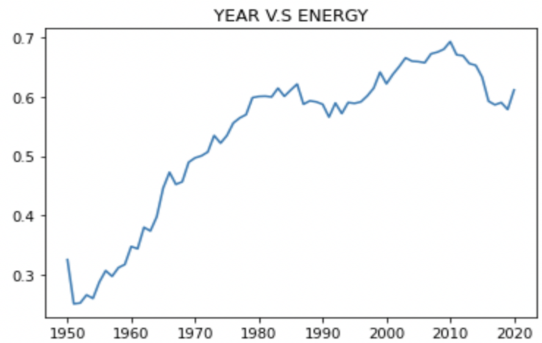
	year	danceability	energy	valence	tempo	loudness	mode	key	acousticness	instrumentalness	liveness	speechiness	duration_ms	popularity
0	1921	0.425661	0.236784	0.425495	100.397758	-17.095437	1	7	0.895823	0.322330	0.215814	0.077258	229911.9141	0.351562
1	1922	0.480000	0.237026	0.534056	101.376139	-19.179958	1	10	0.939236	0.440470	0.238647	0.115419	167904.5417	0.138889
2	1923	0.568462	0.246936	0.624788	112.456598	-14.373882	1	0	0.976329	0.401932	0.236656	0.098619	178356.3018	5.727811
3	1924	0.548654	0.347033	0.668574	120.653359	-14.202304	1	10	0.935575	0.583955	0.237875	0.090210	188461.6498	0.603376
4	1925	0.571890	0.264373	0.616430	115.671715	-14.516707	1	5	0.965422	0.408893	0.243094	0.115457	184130.6996	2.707224
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	2016	0.599976	0.592877	0.430769	119.070344	-7.949913	1	0	0.280290	0.074646	0.180198	0.107298	219400.7638	61.371254
96	2017	0.612286	0.586739	0.414465	116.840277	-8.422697	1	1	0.289916	0.098209	0.194218	0.111752	209343.6130	64.861500
97	2018	0.664930	0.590591	0.447141	122.004325	-7.253666	1	1	0.271941	0.035948	0.171781	0.128140	200919.1190	67.276000
98	2019	0.644215	0.578796	0.465856	118.868163	-8.041738	1	1	0.289298	0.076518	0.167161	0.124799	197733.1330	69.655500
99	2020	0.673077	0.611914	0.482755	121.228704	-7.204024	1	1	0.247374	0.039052	0.177048	0.143505	197114.6623	63.111048

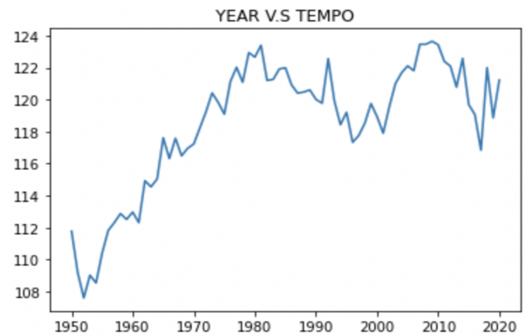
100 rows × 14 columns

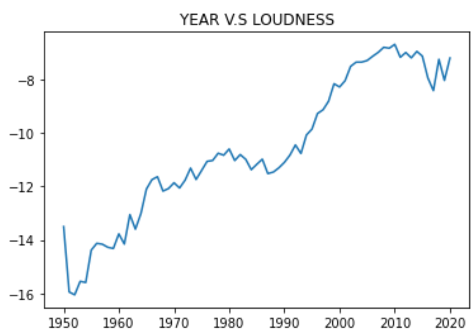
Simple Regression Graphs

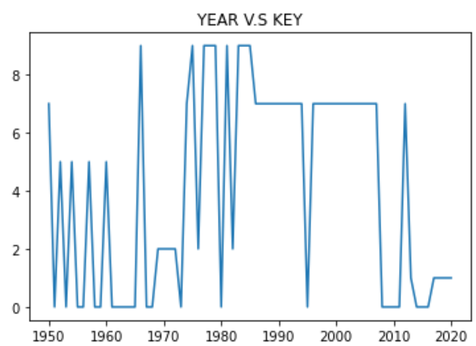


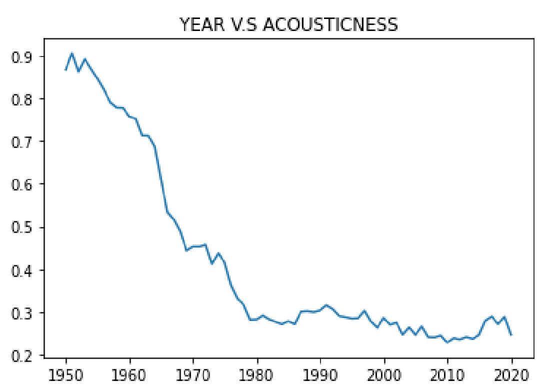


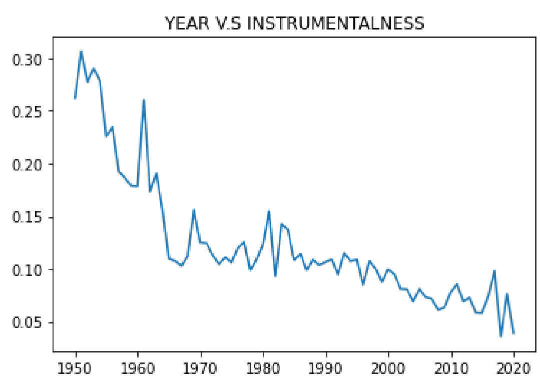


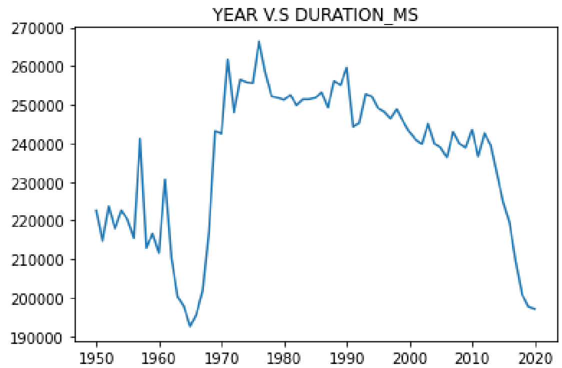


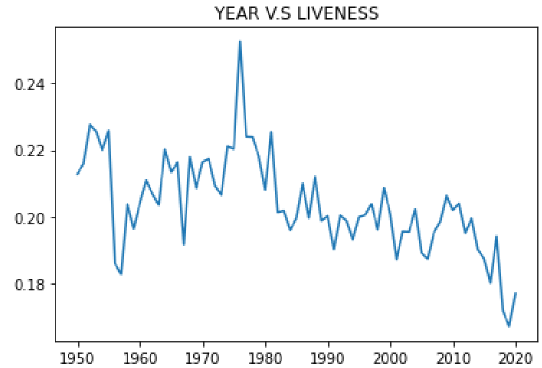


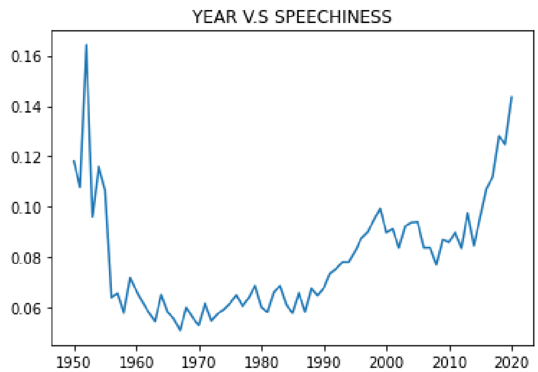


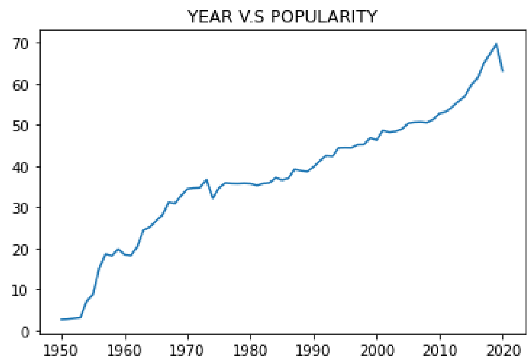


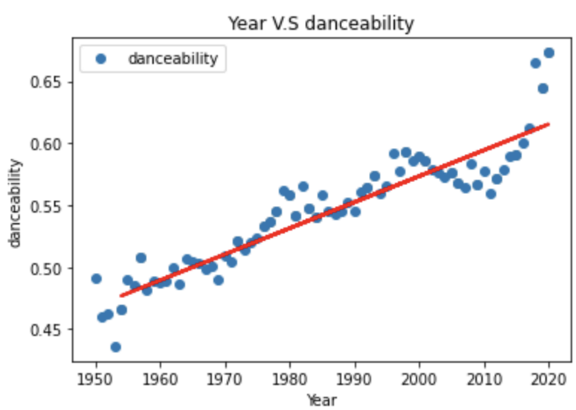




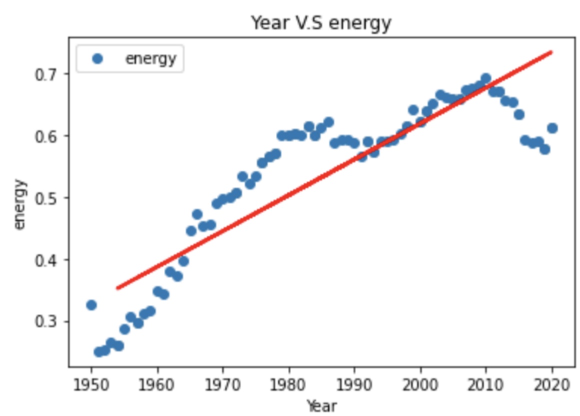




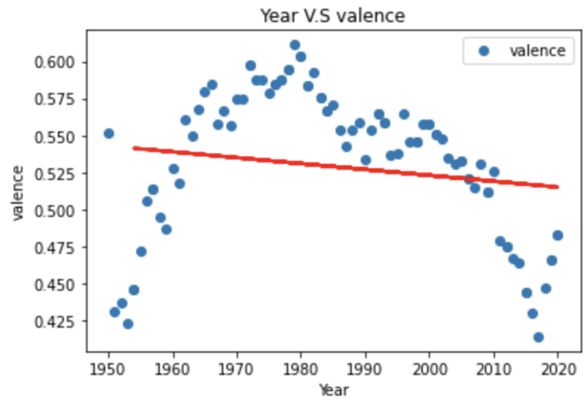




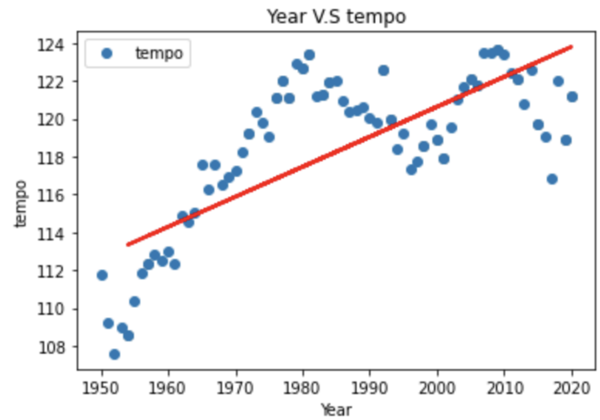
0.7549867651410209



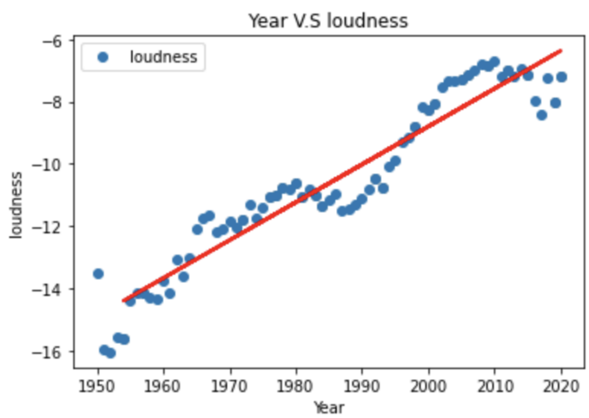
0.5201384978977612



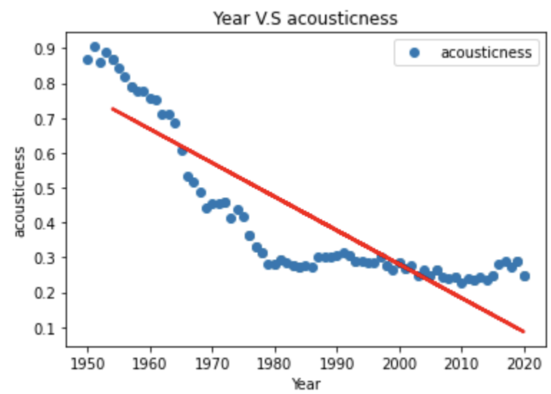

```
0.034940691107105404
```



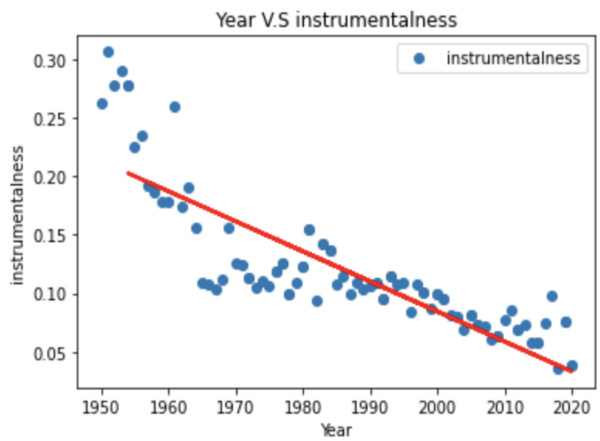

```
0.057353369529476894
```



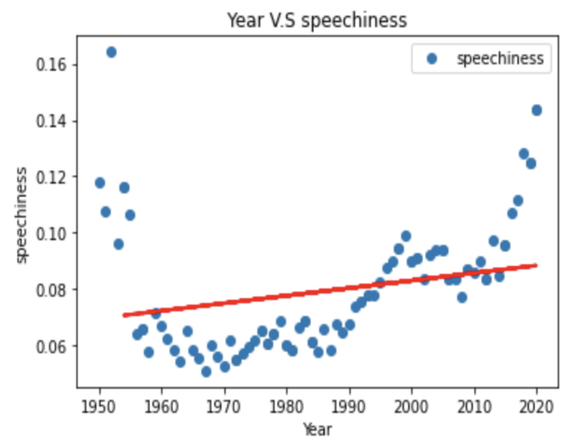

```
0.9025124054971821
```



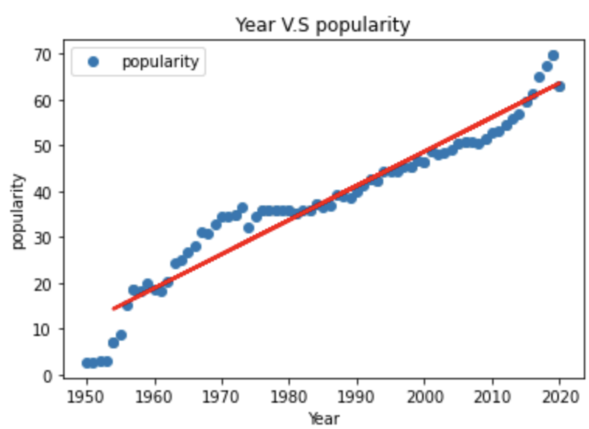

```
0.42162509549220484
```



0.7387038991509387



0.19407180529255663



0.9353958844494671