Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2110138

Unveiling Music Development by Data Analysis

Summary

The way music evolves is just as complicated as living things. In order to better understand the role of music in the collective experience of human beings, we decided to quantify musical evolution, measuring the influence of previously produced music on new music and musical artists, so as to explore the evolution of music with the change of society.

For the first problem, we use the *influence_data.csv* to generate a directed network to show the relationship between artists and measure the importance of nodes by the Musical Influence(MI), which is a quantified result by **PageRank algorithm**. We also explore the importance of each genre in a sub-network, and verify our model according to the result generated from word cloud map.

For the second problem, we first use the **Entrophy Method** to calculate the weight of each music variable, and then calculate the similarity by **Standard Deviation** and conduct data analysis to measure the similarity within and between genres.

On the basis of the first two problems and for the third problem, we draw the thermodynamic diagram of correlation coefficient to get the variable with the largest **Correlation Coefficient** with YEAR, and compare the similarity and influence according to the variables of each genre. Also we visualise the way a genre changes over time and the relation between different genres.

For the fourth problem, we firstly inferred from **Principal Component Estimate**that **Energy, Danceability and Loudness** have a great positive impact on popularity. Then, we conducted **Cluster Analysis** and concluded that these three variables belong to the same type and the model is correct.

Based on the previous problems, we define revolution period by the increment of songs or musicans and take the 1960s as well as the 1980s as the revolutionary time. We also define musical revolutionists as those who are influential and whose genres differ from their influencers. They are listed and classified in later report.

For the sixth problem, we used **Time Series Analysis**, **Difference Operation** and **Partial Auto-correlation** to reveal the dynamism of the development of genres.

For the final seventh problem, we tried to identify the social, political and technological changes by the analysis of music data with a rapid change in the 1960s as an example.

By processing a huge amount of data and establishing models, we accurately explore the evolution of music over time.

Key Words: Reverse Pagerank; Principal Component Estimate(PCE); Correlation Analysis; Entropy Method; Time Series Analysis;

Contents

1 Introduction

1.1 Background

Music plays an irreplaceable role in the development of human society and human civilization. For an individual, there are many factors that influence his innovation, including his innate talent, his ability to perceive and present emotions, the changing situation of the society, the development of music technology, his personal background, life situation, and the influence of celebrity. For the whole society, the evolution of music is a spiraling process. The influences of predecessors on subsequent works, the influences of musicians on each other, and the response of the music industry to major social changes and a series of small changes accumulate step by step. The quantitative changes give rise to the qualitative changes, finally leading to the evolution of music.This kind of change will bring us new sounds or tempos. One of the manifestations is the emergence of new music genres, the other is the re-creation on the basis of existing genres.

1.2 Restatement of problems

Our purpose is to understand the role music plays in the collective human experience. Therefore, we need to quantify the development of music, that is, quantify the impact of previously produced music on new music and music artists. In this process, we can better examine the changing process of musicians and music genres, and explore how music develops with social change as time goes by.

On the whole, it's a question of evaluation. According to the given data, we set up two separate approaches from the perspective of musicians: (1) establish the network of influencers and followers; (2) select influence parameters; (3) establish influence model; From the perspective of music production: (1) select music characteristics; (2) establish music evaluation system; (3) establish music variables changing model

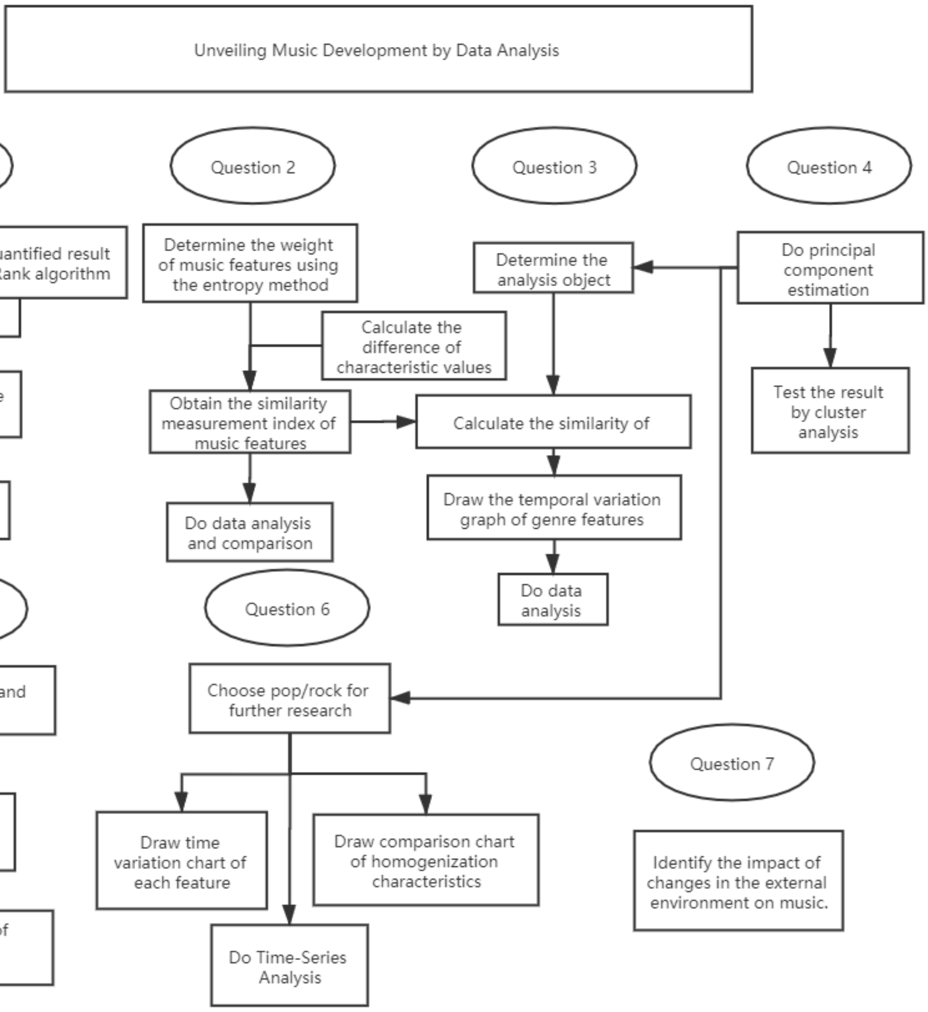
2 Analysis of the Problem

First of all, we use *data_influence.csv* to construct a directed network which points from follower to influencer and also establish a sub-network of genres according to genres which artists belong to. The PageRank algorithm is used to calculate the musical influence (MI) in the original network and sub-network, which represents its importance in the network. The word cloud map is generated according to the probability of genre occurrence to test the correctness of the model.

Second, in order to establish the measurement of music similarity, we use *data_by_artist.csv* to determine the weight of each variable when calculating similarity through entropy method. In the pair comparison of songs, the difference of their features' value after homogenization of the two songs is multiplied by the weight and then added to get the measurement index. When measuring the internal similarity of the genre, we use the standard deviation of each variable. And when it comes to different genres, we choose the standard deviation of the overall data . After comparative analysis, it is found that there is no absolute similarity among the artists within the genre.

Third, We combine the *full_music_data.csv* and the *influence_data.csv* data to obtain the music data of each genre, and take the mean value of its music variables as the characteristic value of the genre. The comparison process is similar to that described in the second problem. At the same time, in order to compare the characteristics of the genre more clearly, we homogenized the characteristic values of the selected genre and compared them by drawing graphs. Meanwhile, we visualize the changing process of both variables and genres.

Figure 1: The flow of our work


Unveiling Music Development by Data Analysis
Question 1
Question 2
Question 3
Question 4
Generate a directed network
Obtain quantified result by PageRank algorithm
Determine the weight of music features using the entropy method
Determine the analysis object
Do principal component estimation
Calculate the difference of characteristic values
Explore the importance of each genre
Obtain the similarity measurement index of music features
Calculate the similarity of
Test the result by cluster analysis
Generate word cloud map
Do data analysis and comparison
Draw the temporal variation graph of genre features
Do data analysis
Question 5
Question 6
Define revolution and revolutionist
Choose pop/rock for further research
Question 7
Carry out data statistics and analysis
Draw time variation chart of each feature
Draw comparison chart of homogenization characteristics
Identify the impact of changes in the external environment on music.
Get the period of revolution and revolutionists
Do Time-Series Analysis

Fourth, we believe that some variables have a greater impact than others on popularity and on musicians. Therefore, we use the principal component estimate method to obtain the regression equation between 12 variables and popularity. Among them, energy, danceability and loudness have greater positive effects, while duration_time has no effect on popularity and is irrelevant. Then, we calculate the correlation coefficient based on the *data_by_artist.csv* and visualized the model with the thermodynamic diagram. The maximum coefficient method is used for systematic clustering, and the 12 variables of atmosphere are classified into two categories. Energy, danceability and loudness are belonged to the same category. Thus, the model is verified as correct.

Fifth, we think that the great revolutionary periods are characterized by two things, a significant increase in the number of music or artists.Therefore, the period when the number of music increased rapidly was obtained through difference operation: the 1960-1970 are related to military, political, economic, cultural and other factors; The boom in musicians: 1990-2000s.

In our opinion, there are three characteristics of influencers: they are not the same genre as influencers; they have a decent popularity; and they influence many people. By analyzing the numbers, we've sorted out the reformers of major genres, which concludes Bob Dylan as the leader.

Sixth, we use the *influence_data.csv* to plot the change of popularity over time for some genres. We select pop/rock genres for further analysis and plot their genre development chart and characteristic changing by time change chart. Based on the previous correlation coefficient analysis, we selected energy for time series analysis and further evaluation.

3 Model Hypothesis and Symbol Description

3.1 Model Hypothesis

- The data set given in the question is sufficiently complete to reflect the status quo of music in the United States over the past hundred years.
- Consider 19 genres including Avant-Garde, Blues, Children's, Classical, Comedy/Spoken, Country, Easy Listening, Electronic, Folk, International, Jazz, Latin, New Age, Pop/Rock, R&B, Reggae, Religious, Stage Screen, Vocal. Others are excluded from consideration.
- The variables of songs and genre only consist of valence, tempo, loudness, mode, key, acousticness, instrumentalness, liveness, speechiness, duration_ms, regardless of lyrics and semantic analysis (not included in the data set).
- The development of music is not only influenced by the previous musicians, but also by military, political, economic, cultural and other factors.

3.2 Symbol Description

For convenience, some notations are first defined as follows:

Symbols	Definition
DAN	danceability
ENE	energy
VAL	valence
TEM	tempo
LOU	loudness
MOD	mode
KEY	key
ASO	acousticness
INS	instrumentalness
LIV	liveness
SPE	speechiness
MC	musical characteristics,a vector of 12 dimensions above
POP	popularity
DUR	duration(ms)
y	year
MI	musical influence, calculated by pagerank algorithm
β_i, c_i, d_i	unknown parameters to be calculated
R	the eigenvector in the pagerank algorithms

Symbols	Definition
DAT	the matrix from *data_by_artist.csv*
NM	number of musicians
NPC	number of principle component

4 Model Building and Solving

4.1 Problem 1

4.1.1 Model Building

Establish the artists directed network, and calculate the MI value of each node through the PageRank algorithm. Then establish the sub-network of the genre , and finally verify the model.

- Basic Hypothesis

  Quantity hypothesis: It is assumed that the more followers point to the influencer, the more important the influencer is.

  Quality hypothesis: If the importance of the follower is higher, the more important the influencer is, that is, the weight differ from follower to follower.

  Random listener: The listeners are exposed to music in their life randomly and have an incentive to search for relevant music.

- Formula Derivation

$$\mathrm{MI}\left(a_i\right)=\frac{1-d}{N}+d\sum_{a_j\in M(a_i)}\frac{\mathrm{MI}\left(a_j\right)}{L\left(a_j\right)} \tag{1}$$

  The $a_i\cdots a_n$ is the influencer; $M(a_i)$ is the collection of artists affected by a_i; $L(a_j)$ is the number of followers; N is the number of all artists in the collection; d is the damping coefficient, meaning the probability that an artist continues to influence new artists. The MI value of all artists in the collection can be represented by the feature vector of a special adjacency matrix.This eigenvector R is :

$$\mathbf{R}=\begin{bmatrix}\mathrm{MI}\left(a_1\right)\\ \mathrm{MI}\left(a_2\right)\\ \vdots\\ \mathrm{MI}\left(a_N\right)\end{bmatrix} \tag{2}$$

  R is also a solution to the following equations:

$$\mathbf{R}=\begin{bmatrix}(1-d)/N\\ (1-d)/N\\ \vdots\\ (1-d)/N\end{bmatrix}+d\begin{bmatrix}\ell\left(a_1,a_1\right) & \ell\left(a_1,a_2\right) & \cdots & \ell\left(a_1,a_N\right)\\ \ell\left(a_2,a_1\right) & \ddots & & \\ \vdots & & \ell\left(a_i,a_j\right) & \\ \ell\left(a_N,a_1\right) & & & \ell\left(a_N,a_N\right)\end{bmatrix}\mathbf{R} \tag{3}$$

The adjacency function L(ai, aj) represents the ratio of "number of artists affected by artist i from artist j "to "total number of artists affected by artist j". If a_j is not affected by a_i, then the previously mentioned "number of times from Artist j to Artist i" is zero.

To generalize the situation: for a particular j, there should be:

$$\sum_{i=1}^{N} \ell\left(a_i, a_j\right) = 1 \tag{4}$$

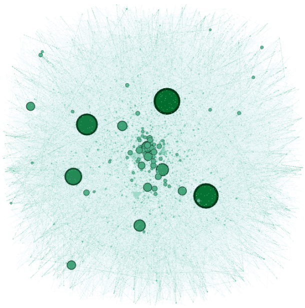
Set these formulas together, the value of PageRank eigenvector R can be estimated with high accuracy. Thus, MI values of musicians and music genres can be calculated.

4.1.2 Model Solving

- Data Processing

After MI value is calculated in this method, the directed network of each musician can be shown in the figure below.

Figure 2: The directed network of artists generated from influence data



In this figure, the larger the MI value, the larger the node size and the darker the node color. The artists with the highest MI value are:

Table 1: Top 10 most influential musicians

Artist Name	Musical Influence
Cab Calloway	0.020783
Billie Holiday	0.019663
Lester Young	0.016932
Louis Jordan	0.013613
T-Bone Walker	0.010007
Sister Rosetta Tharpe	0.009261
The Beatles	0.009
The Mills Brothers	0.007711
Mississippi Fred McDowell	0.007012
Mississippi Sheiks	0.006787

Sub-network is constructed according to the genre of artists:

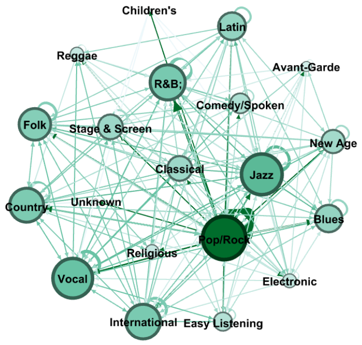


Figure 3: The directed network of genres generated from influence data

The genres with the highest MI value are:

Table 2: Top 10 most influential genres

Genre	Musical Influence
Pop/Rock	0.065622
Electronic	0.025617
Reggae	0.024988
Jazz	0.062194
Country	0.053017
Comedy/Spoken	0.034027
R&B;	0.054584
Classical	0.040535
Latin	0.044368
Vocal	0.060808

Thus we can come to a conclusion. The larger the MI value is, the more other nodes the node affects, and the more important the position of the node in the whole network is.

- Hypothesis Testing

  Analyze the data in *full_music_data.csv* and get the word cloud graph as follows:

  Again, PopRock took the top spot, which is consistent with the hypothesis.

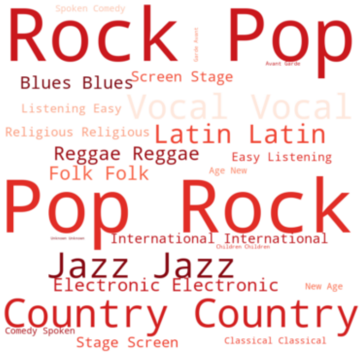


Figure 4: The wordcloud picture for verification

4.2 Problem 2

4.2.1 Model Building

- Basic Hypothesis

  Musical similarity is only related to its 11 characteristics. The similarity of music is objective. The subjective feelings of the audience are excluded from consideration.

- Basic Content

  The *data_by_artist.csv* is used to establish the music similarity metric. The selected features:

danceability	energy	valence	tempo	loudness	mode
key	acousticness	instrumentalness	liveness	spechiness	

  We get the weight of each feature using entropy method, and then it multiplied by the feature difference sum to get a unified index.

  Among them, for certain two songs, the variable difference is the absolute value after subtracting the homogenized characteristic value. For a genre, the characteristic difference is the standard deviation of its homogenized features' value. The larger the index value of similarity measurement is, the smaller the degree of similarity between two songs is. Vice versa .

- Formula Derivation(Entropy method)

  1. Assuming that the data has n rows of records and m variables, the data can be represented by an n*m matrix A (n rows and m columns, that is, the number of records in n rows and m characteristic columns).

$$A = \left[\begin{array}{c} x_1 \ldots x_m \end{array} \right] \tag{5}$$

  2. Normalization of data

  Xij represents the element in row i and column j of matrix A.

  Positive indicators:

$$x'_{ij} = \frac{x_{ij} - \min\left\{x_{ij}, \cdots, x_{nj}\right\}}{\max\left\{x_{1j}, \cdots, x_{nj}\right\} - \min\left\{x_{1j}, \cdots, x_{nj}\right\}} \tag{6}$$

  Negative indicators:

$$x'_{ij} = \frac{\max\left\{x_{1j}, \cdots, x_{nj}\right\} - x_{ij}}{\max\left\{x_{1j}, \cdots, x_{nj}\right\} - \min\left\{x_{ij}, \cdots, x_{nj}\right\}} \tag{7}$$

  3. Calculate the proportion of the i record in the j index

$$P_{ij} = \frac{x_{ij}}{\sum_1^n x_{ij}} (j = 1, 2, \ldots, m) \tag{8}$$

4. Calculate the entropy value of the j index

$$e_j = -k * \sum_{1}^{n} P_{ij} * \log\left(P_{ij}\right), k = 1/\ln(n) \tag{9}$$

5. Calculate the coefficient of difference of index j

$$g_j = 1 - e_j \tag{10}$$

6. Calculate the weight of index j

$$W_j = \frac{g_j}{\sum_1^m g_j} \tag{11}$$

Comparison model:

1. Calculate standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \tag{12}$$

2. Obtain the overall difference vector

3. Calculate similarity index

4.2.2 Model Solving

- Data Processing

  In order to unify the evaluation criteria, the data used to calculate the weight is *data_by_artist.csv* , which will not be changed in the subsequent calculation.The weight features is as follows:

Table 3: The weight of every parameter

Variable	Weight
danceability	0.015186
energy	0.025426
valence	0.02797
tempo	0.007553
loudness	0.003435
mode	0.06506
key	0.086929
acousticness	0.132906
instrumentalness	0.388894
liveness	0.052026
speechiness	0.194615

compare similarities between the two songs:

Example 1 :(Songs sung by the same singer) The obtained similarity index is:0.2217150172529077

Example 2 :(Songs sung by different singers) The obtained similarity index is 0.3966308818244329, which is less than that of Example 1.

Compare whether artists within genres are more similar than artists across genres:

Ideas:(1) Firstly, homogenize each music eigenvalue;

(2) Calculate the standard deviation of each musical feature among all artists, multiply it by the weight and add it to get the overall similarity index;

(3)Calculate the standard deviation of each musical characteristic of a genre artist, such as pop/rock genre, using the same method as (1);

(4) Make a comparison.

Table 4: The standard deviation of different parameters

	All	Electronic	Jazz	Pop&Rock
danceability	0.023	0.034	0.023	0.021
energy	0.047	0.036	0.035	0.035
valence	0.044	0.048	0.04	0.039
tempo	0.012	0.018	0.018	0.011
loudness	0.01	0.028	0.017	0.01
mode	0.026	0.085	0.069	0.016
key	0.073	0.088	0.056	0.075
acousticness	0.088	0.061	0.068	0.05
instrumentalness	0.056	0.115	0.079	0.04
liveness	0.015	0.035	0.017	0.014
speechiness	0.006	0.028	0.004	0.006

(1) all_artists:Index of similarity is 0.20236

(2) electronic:Index of similarity is 0.26760

(3) jazz:Index of similarity is 0.21480

(4) pop/rock:Index of similarity is 0.17498

- Results and Analysis

  We take Pop/Rock, Jazz, Electronic and compared them to all music, and find that songs within Pop/Rock are more similar than songs in all music, while the other two are less similar than all music. This indicates that there is no absolute similarity among artists within a genre under this model, which is in line with the diversified development of music of each genre.

  Jazz originated in the United States at the end of the 19th century and the beginning of the 20th century. Jazz emphasizes improvisation and is based on Shuffle, which is a combination of African black culture and European white culture, and its rhythm is extremely complex. Electronic music originated from the late 19th century to the early 20th century, including French specific music, Japanese electronic music, random music and other types of music, and it also under the influence of pop/rock music, jazz music and other music genres.There is no fixed equation for the composition of music, which rationalizes our modeling results to some extent.

4.3 Problem 3

4.3.1 Model Building

- Basic Hypothesis

$$r(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} \tag{13}$$

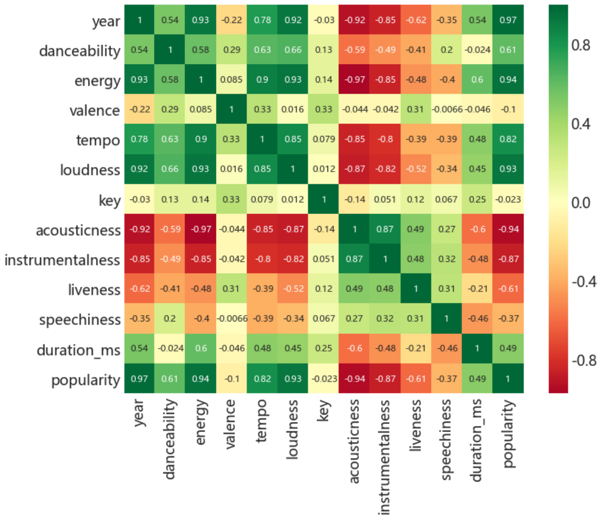


Figure 5: Correlation coefficient heat map generated from data by year

According to the graph above, music variables that are closely related to the year are selected for analysis.

For the correlation between a genre and the other genres, we preliminarily consider several factors: the correlation between genre variables and the influence between genres (refer to the directed network established in problem 1), the similarity of the various genres over time.

The considerations of factor 1 and factor 2 refer to the above, while for factor 3, the change of a variable of each genre with years is selected to be compared.

4.3.2 Model Solving

- Data Processing

  Compare similarities and influences between and within genres:

  Combining *full_music_data.csv* and *influence_data.csv*, we obtain the music data of each genre. The mean value of its music characteristics is taken as the characteristic value of the genre.

Table 5: The mean value of characteristics of different genres

	All	Pop&Rock	Electronic	Jazz	Country	R&B
danceability	0.527	0.493	0.626	0.603	0.515	0.577
energy	0.585	0.678	0.675	0.379	0.545	0.559
valence	0.543	0.526	0.485	0.512	0.577	0.611
tempo	0.509	0.531	0.502	0.344	0.445	0.375

Table 6: The mean value of characteristics of different genres

loudness	0.728	0.707	0.687	0.645	0.568	0.659
mode	0.821	0.862	0.635	0.682	0.988	0.713
key	0.501	0.506	0.559	0.452	0.534	0.515
acousticness	0.345	0.213	0.197	0.654	0.455	0.327
instrumentalness	0.136	0.112	0.369	0.42	0.044	0.047
liveness	0.192	0.187	0.267	0.17	0.217	0.175
speechiness	0.05	0.071	0.14	0.035	0.042	0.057

Compare the similarities between genres (the same calculation) :

(1) R&B and Jazz: 0.2083123672573685 (2) Country and Jazz: 0.21203612254563955

(3) Country and Eletronic: 0.21580027614343605

Images can show the similarity of each indicator more clearly:

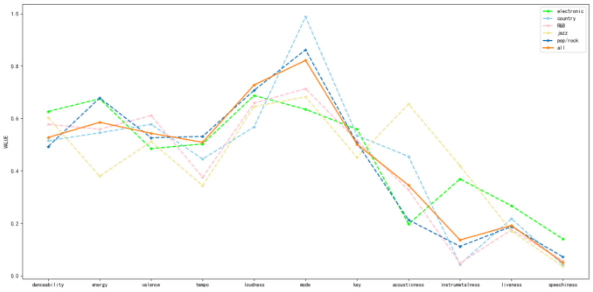


Figure 6: Comparison by variables

How genres change over time:

Taking all genres as an example, we analyze the change of their variables over time. The variables we selected are valence, energy, loudness, acousticness, tempo and instrumentalness. The images are as follows:

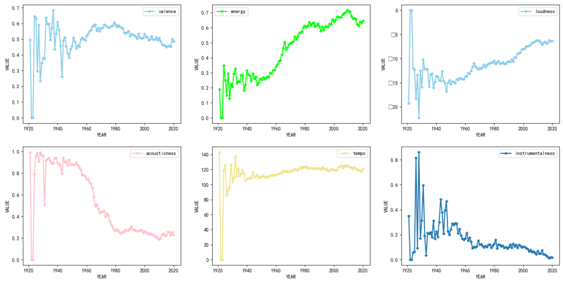


Figure 7: Variables changing by time

Whether a genre related to the others:

For each genre, as above, the change of a variable with years is compared.We select country, electronic, jazz, R&B to compare the change of energy over time.

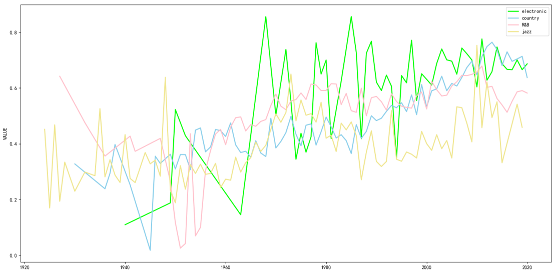


Figure 8: Variable "energy" changing by time

- Results and Analysis

  Analyzing the change of genre variables over time, it can be observed that both energy and loudness have a tendency to rise after 1950. And valence, acousticness, tempo with instrumentalness level off. The curves of energy of country, electronic, jazz, R&B are similar to the overall trend over time, which are fluctuating in roughly the same interval after 1950. This result verifies the close relationship between music, and the convergence of eigenvalues may indicates the inseparability and compatibility between different genres.The disorder of the data fluctuations before 1950 may result from lack of data.

  Music does not exist and develop in isolation. And there is no genre completely separated from all of other music. The beauty of music has a certain degree of similarity. Music genres maintain the similarity of the general trend while diversifying. R&B, according to the document, originated from African-American artists and is a blend of jazz, Gospel and electronic blues. Country music originated from Appalachian folk and bluegrass music, and has been continually adapted, innovated, and developed since then. Pop/rock music a mixed music in itself. From the history of different music genres, we believe that they are constantly exchanging and integrating in their development.

4.4 Problem 4

4.4.1 Model Building

We believe that popularity is a very important variable, because the influence of music is produced by the masses, and a higher popularity means that it can be spread more widely among the masses, thus influencing more musicians. In this part, a function will be established between various variables and popularity. It is concluded that the variables which have a greater influence on popularity are more significant.

- Basic Hypothesis

  POPULARITY is a function of 12 variables.

- Formula Derivation

$$DAT = \left(MC_1, MC_2, \cdots, MC_p\right) = \begin{pmatrix} VAL_1 & TEM_1 & \cdots & LOU_1 \\ VAL_2 & TEM_2 & \cdots & LOU_2 \\ \vdots & \vdots & & \vdots \\ VAL_n & TEM_n & \cdots & LOU_n \end{pmatrix} \tag{14}$$

$$POP = \beta_0 1 + DAT\beta + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2 I\right) \tag{15}$$

$$z = c_1 MC_1 + c_2 MC_2 + \cdots + c_p MC_p, \quad \sum_{j=1}^{p} c_j^2 = 1 \tag{16}$$

$$Z = \left(z_1, z_2, \cdots, z_p\right) = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix} \tag{17}$$

$$Y = \beta_0 1 + ZQ^T\beta + \varepsilon = \beta_0 1 + Z\alpha + \varepsilon \tag{18}$$

$$\hat{\beta} = (Q_1, Q_2)\begin{pmatrix} \hat{\alpha}_1 \\ 0 \end{pmatrix} = Q_1\hat{\alpha}_1 \tag{19}$$

- Calculating

Table 7: Parameters calculated by PCE

NPC	d_1	d_2	d_3	d_4	d_5	d_6
4	2.3427	9.9851	1.3858	0.0454	0.4432	-3.1236
5	2.5814	10.0634	1.4901	0.0450	0.4438	-3.0865
6	3.4558	10.9619	-1.9076	-0.0581	0.5602	-0.2684
7	0.9654	10.4680	-5.1621	-0.0847	0.6446	-3.5520
8	1.0523	8.7134	-7.1477	-0.0486	0.6369	-3.3861

$$\begin{aligned} POP = {} & d_1 DAN + d_2 ENE + d_3 VAL + d_4 TEM + d_5 LOU + d_6 MOD \\ & + d_7 KEY + d_8 ACO + d_9 INS + d_{10} LIV + d_{11} SPE + d_{12} DUR \end{aligned} \tag{20}$$

4.4.2 Model Solving

Among all the variables, energy, danceability and loudness have the most positive impact on popularity, while duration_time has no effect on popularity and is an independent variable.If energy, danceability and loudness belong to the same type of variables, the conclusion above is justified. Firstly, using the correlation coefficient between variables in the same way:

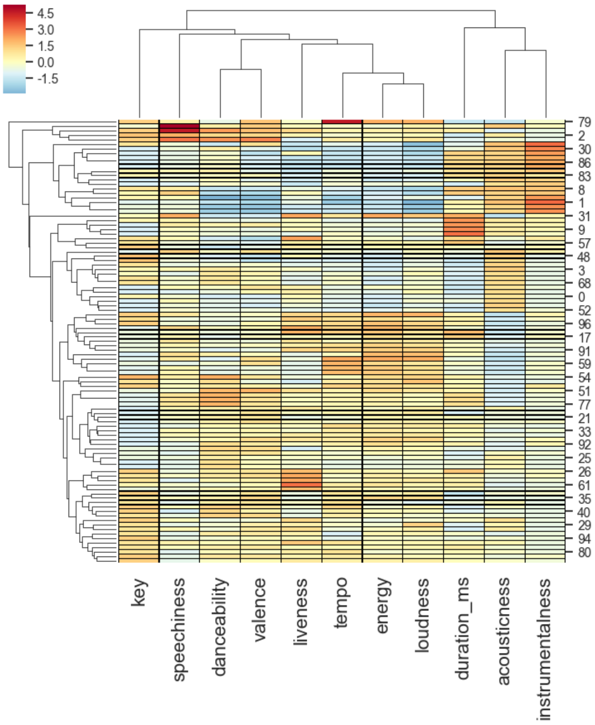


Figure 9: Cluster heatmap generated from data_by_artist

Then, according to their correlation, the 12 variables are clustered in R-type. Firstly, the data of each variable are standardized. UPGMA is used to measure the similarity between variables, and the cluster analysis is used to calculate the similarity between classes. Cluster analysis is performed on each variable. The quoted formula is:

$$R\left(G_1, G_2\right) = \max_{\substack{x_j \in G_1 \\ x_k \in G_2}} \left\{r_{jk}\right\} \tag{21}$$

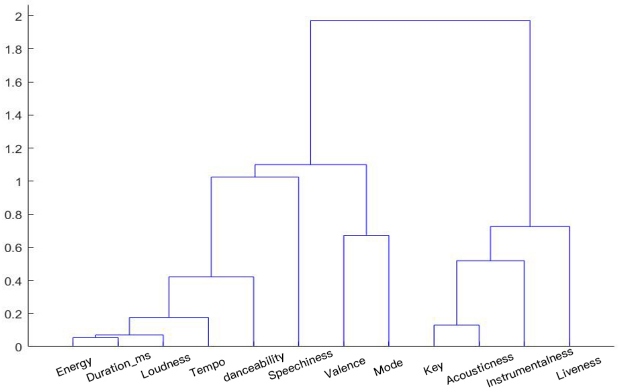


Figure 10: Cluster relationship

From this, the variables that determine the type of music can be divided into two broad categories. One is the nature of music, such as key, acousticness, instrumentalness and liveness; The other is emotional tendency, which can be distinguished as positive or negative music, based on factors such as energy, speechiness, loudness, tempo, and danceability. Energy, danceability and loudness, which have greater influence on popularity, are all variables affecting emotional tendency. Thus it verifies the validity of the model.

4.5 Problem 5

4.5.1 Model Building

- Basic Definition

  Great revolution: a significant change in the amount of music or in the number of musicians.

  Revolutionist: who has a certain popularity and influence on the entire music market, and influences many people as an influencer.

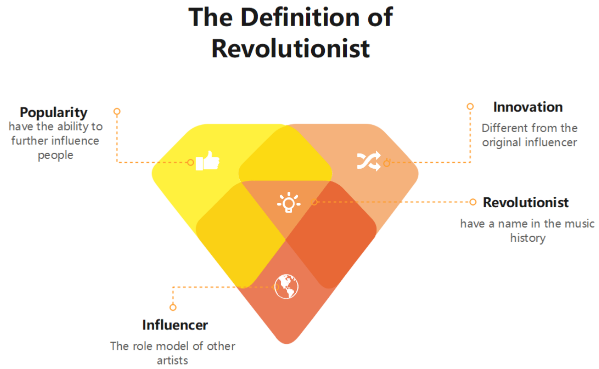

The Definition of Revolutionist
Popularity
have the ability to further influence people
Innovation
Different from the original influencer
Revolutionist
have a name in the music history
Influencer
The role model of other artists

Figure 11: Definition for clarification

4.5.2 Model Solving

According to the statistics of the number of songs each year, we get the image as shown in the figure:

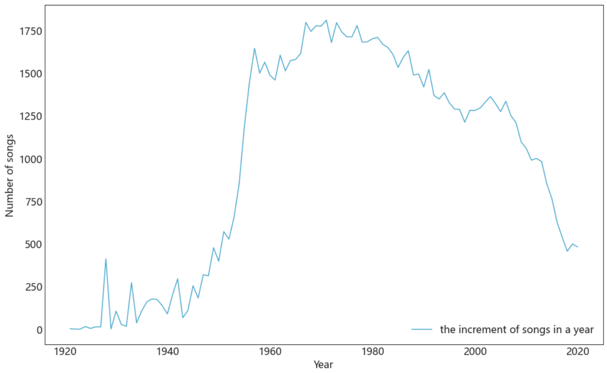


Figure 12: The number of songs changing by time

There was a significant increase in the number of songs from 1960 to 1970, so we think this period is revolutionary.

Politics: America's military and diplomatic tensions have made music, which is a widely recognized consumer product of mass culture, a spiritual comfort for the people.

Economy: With the development of industrialization, a large number of African Americans from the southern villages came to the cities, bringing their unique music and culture.

Culture: Rich Americans had more spiritual needs. Music represented the strongest sound of the time, making harmonious wishes.

Technology: The invention of vinyl records in 1948 greatly reduced the cost of recording production while radio, television, film and other technologies were booming.

The number of musicians in each year is statistically analyzed, and then the change of the increment of musicians over time is obtained by difference operation:

$$y_k = y_0 + ky, (k = 0, 1, \ldots, n) \tag{22}$$

$$\Delta NM\left(y_k\right) = NM\left(y_{k+1}\right) - NM\left(y_k\right) \tag{23}$$

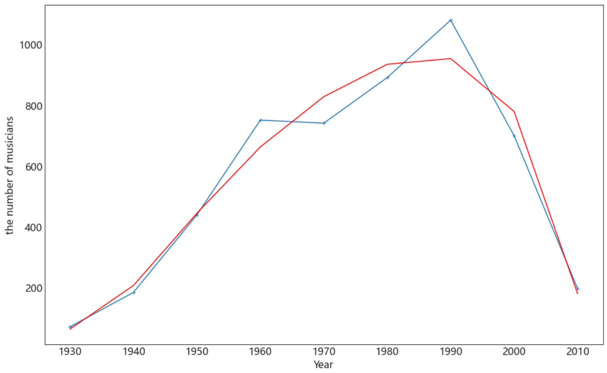


Figure 13: The increment of musicians changing by time

The period between 1980 and 1990 saw the greatest increase in the number of musicians, so we think it is also a revolutionary period. The digital age has been in full swing since the 1980s. Being a musician to release songs no longer required you to be a professional singer. Being a individual musician became possible. Thus, the number of musicians grew by leaps and bounds.

So we use the *influence_data.csv* for conditional filtering, with the condition "influencer schools!= followers genre" and then count function is used to calculate the number of people affected by each follower as an influencer, and vlookup function is used to link *data_ by_ artist.csv* to record the popularity. We use follower_number as the main keyword, popularity as the secondary keyword to get the most influential pioneer of each genre.

The conclusions are shown in the table below:

Table 8: Revolutionists from major genres

Genre	Revolutionist
Pop&Rock	Bob Dylan and Alice Cooper
R&B	Ray Charles and Teddy Pendergrass
Vocal	Billie Holiday and Tony Bennett
Electronic	Kraftwerk and Tangerine Dream
Blues	T-Bone Walker
Country	Bill Monroe
Stage&Screen	Andrew Lloyd Webber
Latin	Selena
Folk	Shirley Collins
Religious	Sister Rosetta Tharpe
Raggae	Ken Boothe

4.6 Problem 6

4.6.1 Model Building

- Fundamental Basis

Remove the duplicate items in the follower column of *influence_data.csv* and search for popularity in *data_by_artist.csv*. We sum the popularity of each year and genre and draw the following figure:

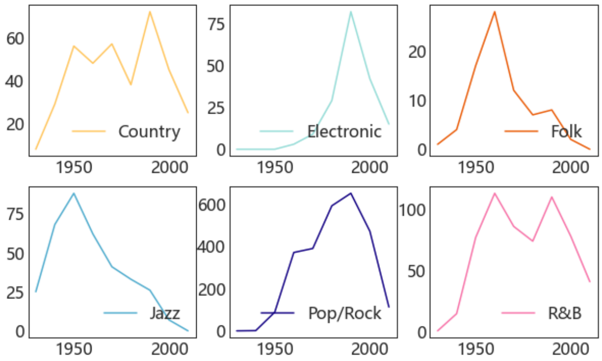


Figure 14: The popularity of genres changing by time

As can be seen from the figure, Pop&Rock is the most popular genre that has been booming in the last hundred years. Therefore, its development and historical evolution will be further analyzed in the following sections.

4.6.2 Model Solving

- Data Processing

In order to explore the influence of other genres on pop&rock, we classify the genres of influencers whose followers belong to pop&rock, and count the genres and active years of these influencers. After summarizing, we draw the following figure:

The development of pop rock school can be divided into three stages:

1.The early stage: it was influenced by traditional genres such as vocal, blues and jazz;

2.Golden middle period: pop&rock absorbed the essence of various genres and its style is a blend of diversified music.

3.Late modern times: modern music led by electronic also influenced the development of pop&rock

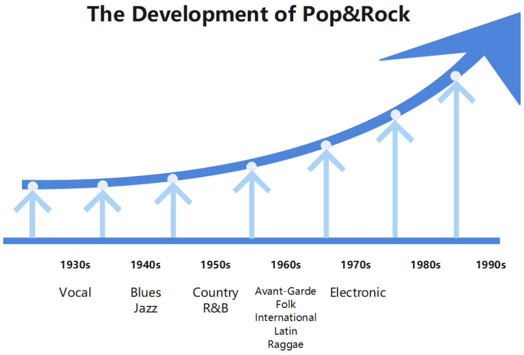


Figure 15: process description for clarification

Since pop&rock is more popular and widely influenced by other genres, this genre is chosen to analyse the changes of its musical characteristics over time.

In order to get a clearer general comparison diagram, we homogenize all characteristic indicators in this genre. The homogenization process is shown in the entropy method.

After homogenization, we get the general comparison diagram:

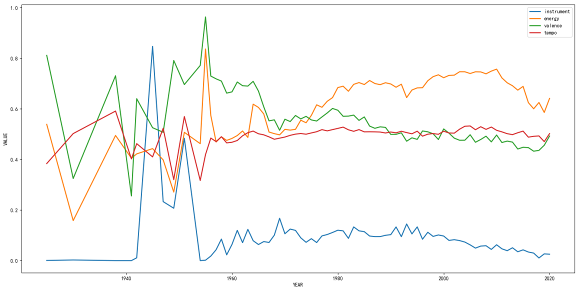


Figure 16: Characteristics of Pop&Rock changing by time

- Analysis

For pop/rock genre, the correlation table shows that danceability, one of its features, has a correlation with Year. Therefore, time series analysis is carried out for it.The time sequence diagram is shown below, with a monotonically increasing trend in the later period.

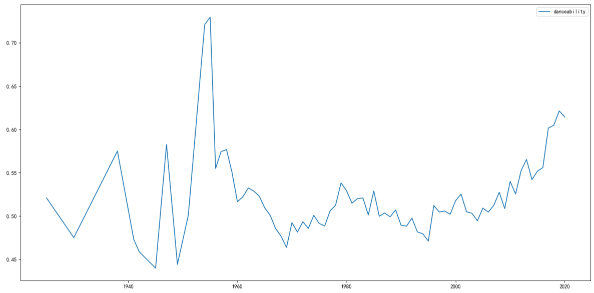


Figure 17: Time series of danceability of Pop&Rock

The autocorrelation graph is shown below:

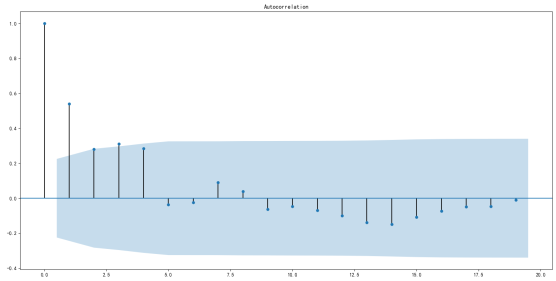


Figure 18: Autocorrelation of danceability of Pop&Rock

Because the time series of the image is not stable, the difference operation is carried out:

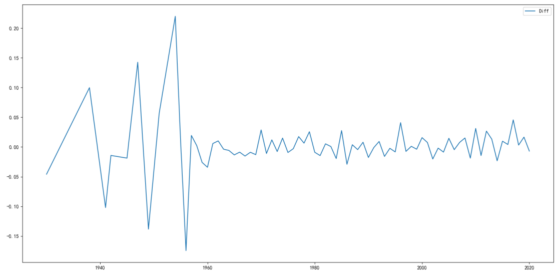


Figure 19: Time series of danceability of Pop&Rock after difference operation

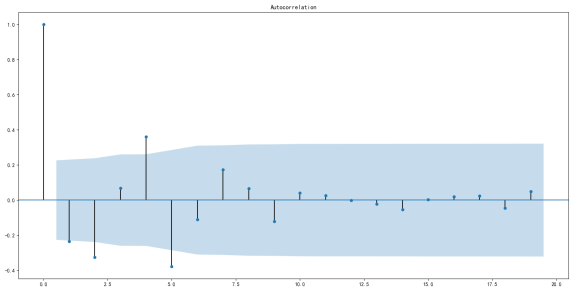


Figure 20: Autocorrelation of danceability of Pop&Rock after difference operation

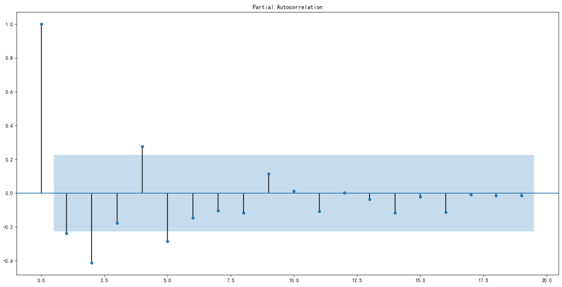


Figure 21: Partial correlation of danceability of Pop&Rock after difference operation

The sequence after difference is relatively stable.The results of white noise test : (array([4.28530905]), array([0.03844309])))

P value is less than significance level, so it can be considered as non-white noise. So we establish the ARIMA model, and predict its trend. The prediction results, standard error and confidence interval are as follows:

Prediction Results	0.639904	0.63824167	0.63657934	0.63491701	0.63325468
Standard Error	0.02362044	0.03340435	0.04091181	0.04724089	0.05281692
Confidence Interval	(0.593,0.686)	(0.572,0.703)	(0.556,0.716)	(0.542,0.727)	(0.529,0.736)

4.7 Problem 7

4.7.1 Model Building

Our networks can easily transcend the obstacle of a huge amount of data, and data analysis provides us with an ingenious mechanism to find out changes in musical parameters, thus attracting the attention of sociologists to explore the history and change of society.

Our network can analyze changes in genres, the emotional tone of the social music scene, and changes in the number of songs and musicians. In each case, there are profound social, political, and technological factors behind it.

The change of genre can reflect major social events. For example, in the 1960s, with the development of industrialization, the urban environment was seriously polluted. The middle class moved to the countryside one after another, while a large number of people from other ethnic groups came to the cities from the southern villages, bringing their unique music and culture. As we have analyzed in Problem 6, Avant-Garde, Folk, International, Latin, Reggae, and other genres were integrated to promote the development of PopRock at that time.

Genre changes can also reflect the emergence of new instruments. Traditional country music developed from the string music and traditional narrative songs of the 19th century. With a guitar or harmonica, the singer can freely express the happiness and sorrow in his heart. But in the 1950s, with the development of electroacoustic instruments, electronic instruments were gradually added to country music, and vocal choral music replaced the original high and lonely singing. At this time, country music gradually shows a more modern urban style.

The emotional tone of the social music atmosphere can reflect the background of the time. As solved in Problem 4, music with more energy, danceability and loudness was more popular. Generally speaking, in times of war, there are more grand and solemn songs and melodies courage and passion. In times of peace, songs that are lighthearted and romantic are more popular.

The increase in the number of music and musicians reflects the development in all aspects in both supply and demand. On the supply side, from recording technology in the 20th century to streaming media and the Internet in the 21st century, every advance has greatly reduced the cost of music innovation and gradually lowered the threshold for musicians. In the past, music was only for the rich, but now anyone with talent can become a pop star who gets attention. On the demand side, from expensive CDs to affordable MP3s and now apps on mobile phones, music is readily available and the market has expanded dramatically. The idol marketing of commercial record companies also increased the market demand for music. With the development of the economy, food and clothing is no longer a problem, people pay more and more attention to spiritual needs, music naturally has played an important role in it.

4.7.2 Model Solving

To be more specific, let's take the 1960-1970 period identified in Problem 5 as an example of social change. The 1960s was the era of the most prominent social contradictions in the United States and also brought the United States into the era of the "modern era" in an all-around way. Compared with the previous decades, this era did not have the horror of the world war, but American society experienced a huge era of turbulence. Looking at this period from a historical perspective, it is not hard to see that postwar young America accelerated its progress further. The deterioration of Sino-Soviet relations since the 1960s, the escalation of the Vietnam War, 32 countries have gained independence from colonial rule, African national liberation movement and Europe's construction of the Berlin Wall, etc. all led to profound social, political, and cultural changes in the United States. In this era, great changes have taken place in American society, the challenge for the three presidents, the social concept of "liberation or indulgence", the thoughts of "new left" surge, feminist social movement, the black civil rights movement, campus rebellion, and anti-war movement, the traditional socialism and the hippie movement, the U.S. mission to the moon, the "pop" revival of art... For almost a decade, American society has been racing along with the wheels of the "modern era" of social progress.

When looking at this period of history from a music research perspective, we can find that almost all of the American popular music took part in these social activities, more or less. Pop music is closely linked to social culture, from the American national ideology to the cultural discourse, from politics to families, from gender discrimination to the youth problems. The most important thing is that we understand the study of pop music should not simply stay in the music itself and simple analysis of works. The birth of any art is closely related to its specific era.

5 Model Evaluation and Improvement

5.1 Model Evaluation

Advantages:

- When calculating musical influence, we retroactively traces the source. The directed network points from followers to influencer, simulating the process of listeners searching for relevant music when listening to music, which is more practical.
- By synthesizing and simplifying the original variables, principal component estimate can

objectively determine the weight of each variable and avoid the arbitrariness of subjective judgment.

- Entropy method is to determine the variable weight according to the variation degree of the variable value, which is an objective weighting method, avoiding the deviation caused by human factors.

Disadvantages:

- The data set is not large enough to reflect the whole development of American music culture.
- Due to time constraints, the model can not be further optimized and correlated, and there are positive effects among some factors.

5.2 Model Improvement

- If more data were available, the model would be more accurate.
- The model can be further optimized to find out the positive relationship between some factors.

6 Conclusions

In order to better understand the role music has played in the collective human experience, we quantify musical evolution to measure the influence of previously produced music on new music and musical artists, so as to explore the development course of music with the change of society over time.

Through the PageRank algorithm, we quantify the influence of each node in the directed network. Through the entropy method, we calculate the weight of each music feature, so as to measure the music similarity and conduct data analysis. On this basis, we draw the heat map of correlation coefficient, analyzing the index with the largest correlation coefficient with YEAR, and compare the similarity and influence according to the characteristic values of each genre . At the same time, we also selected the index with the highest correlation through principal component analysis, and then conducted cluster analysis to find that these three variables belong to the same type, which verified the correctness of the model. Further, we not only explore the time periods and revolutionaries of major changes, but also reveal the dynamic nature of genre development, and finally try to identify the impact of environmental changes on music.

Through practice, it has been proved that the model used in this paper can explain the development of music accurately, which provides some help for future research.

7 One-page document to ICM society

It is our greatest honor to have this opportunity to share our insight into music study with you.

We genuinely believe that our approach to understanding the influence of music through networks has profound values to most of the people in society. In the status quo, people only care about the songs they listen to, and they often fall into the dilemma of no new interesting music, and friends around them are likely to have different preferences. But with our network, the recommendation algorithm can be optimized, which is more surprising and inclusive than the association analysis used by the general music APP. Although the genres may be different, they may still have similar thoughts, which can greatly enrich the spiritual life of the masses. Scholars of various genres are prone to be in conflict with or even attack each other for the so-called orthodoxy. For example, African American musicians attack white musicians for cultural appropriation and professionals attack pop music for being vulgar and boring, creating factionalism in a previously peaceful music world. With our network, people can understand that music is inherently interconnected, equal, and dynamic. They will be able to devote more energy to the innovation of music and thus creating more brilliant cultural crystals. Students usually only know some names but don't know the relationship between them, so they can't have a comprehensive grasp of the whole history of music.

Our network will also be helpful for them. Our model will improve a lot with richer data in three aspects. First, accuracy. With more data, as the sample size increases, our model will better reflect the changes and development of American music, and the random error in the sample will decrease. Second, diversity. If we have music data from different countries, we can infer the relationship between the two countries from their cultural exchanges. If detailed historical background data are available, changes in religious beliefs, social movements, and ideologies can also be inferred. Thirdly, complexity. We will use more complex mathematical models to accurately fit the actual situation.

After analyzing the data carefully, we are eager to give some suggestions for the study of music and its influence on culture. Pay attention to lyrics. Because of the widespread of songs, the influence of lyrics can not be underestimated, and the ideological content contained in them is a precious cultural treasure. Stress interdisciplinary study. We should not only focus on the music itself but also attach importance to taking advantage of other subjects. For example, music in times of war is sad, while music in times of peace is joyful. Through songs, a country's grand historical background and cultural heritage can be reflected. Also in combination with geography, folk songs reveal the local conditions and practices of native residents. With the development of computer science, we can use neural networks for natural language processing, which is much more efficient than human resources.

I would appreciate it if you carefully read our report.

8 References

References

[1] Yu Xiaofen, Fu Dai.Review of multi-index comprehensive evaluation methods [J]. Statistics and Decision,2004(11):119-121.

[2] Pan Yaxing. Research on the Generation of Ci Cloud Based on Python – Taking Chai Jing's Kanyan as an Example [J].Journal of Computer Science and Technology,2019,15(24):8-10.

[3] Wu JB, Ye LX, Jurke R. Application of ARIMA model in predicting the incidence of infectious diseases [J].Journal of Mathematical Medicine and Pharmacy,2007(01):90-92.

[4] BERGLUND JEFF, JOHNSON JAN, LEE KIMBERLI. Indigenous Pop:Native American Music from Jazz to Hip Hop. 2016.