Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2118261

Summary

Music has a history of thousands of years, and many music genres were born over time. Nowadays, music has become an important part of our daily life. In order to study the development of music and the relationship between artists, we need to develop an indicator to quantify the characteristics of music, and to reveal the changes in music development and the influence and similarity between artists.

The ICM Society requires our team to use the provided data to analyze the influence and similarity between artists, the influence and similarity among genres, how the genres change over time, and the relationship between music and historical background. In order to solve these problems, our team has established three models, music influence model, similarity analysis model and infectivity analysis model.

In the music influence model, we use the influence data between artists to map the influence network of artists and use the number of followers of influencer and followers to define the influencer's musical influence. Based on the definition of artist's influence, we analyze the influence network of artists and the meaning of musical influence in sub-networks.

In the similarity analysis model, we use Principal Component Analysis (PCA) to reduce the dimension of the data, then obtain the artist's data in terms of six principal components, and use the artist's Euclidean distance calculated with the principal components to define the artist's similarity measure. Based on the similarity measurement, we use a Clustering Algorithm to analyze the similarity of artists between genres and within genres and the similarity of music characteristics between genres. By obtaining a similarity matrix between genres, we conclude that the similarity of artists within genres is not significantly different from the similarity of artists between genres.

In the infectivity analysis model, we use the data of the above model combined with the data provided to define the level of influence of influencers on followers. We first employ an improved least squares (LS) regression to analyze the similarity and influence level between artists, and then use multiple regression analysis to draw the conclusions. We conclude that influencers do have an impact on followers, and some musical characteristics are more contagious with high confidence.

Finally, we combined the historical background to analyze the differences of genres, the changes in the popularity of genres over time, the changes in the musical characteristics of Pop/Rock genres over time, and the influence of music on culture.

Keywords:*Music Influence Model, Similarity Analysis Model, Infectivity Analysis Model, Robust Standard Error, Improved Least Squares, Regression*

Contents

1 Introduction

1.1 Background

Music is an important part in our daily life. In the past few decades, there have been several major musical revolutions, resulting in many different types of music. In order to understand how music develops, we hope to establish a way to quantify music development. Now ICM requires us to develop a model to measure the influence of music, analyze the connections between and within genres, and predict the development trend of the genres.

1.2 Our Goal

We were given four data sets, which included the characteristics of some music from the past few decades and the data of influencers and followers of some music genres.We need to build models to solve the following problems:

(1) Create one or more music influence networks to connect influencers and followers. Create a sub-network for contacting influencers and define a parameter to describe the music influence in this network.

(2) Use the provided data set to define a measure of musical similarity and analyze whether artists within genres are more similar than artists between genres.

(3) Analyze the similarities and influences between and within genres, the difference between genres and the difference levels between genres in different periods.

(4) Analyze whether the influencer actually affects the followers, and explain whether some characteristics of the music are "infectious" and thus affect the followers.

(5) Define the signs of major changes in music development, and analyze which artists are the influencers of these major changes.

(6) Analyze the influence of a type of music over time. Try to explain how some genres and artists changes over time.

(7) Represent the influence of society, politics and technology on music.

1.3 Our Work

In order to solve the above problems, we established three models, the music influence model, the similarity analysis model and the infectivity analysis model.

In the musical influence model, we define the artist's musical influence, map the artist's directed network of influence, and analyze the influence between and within genres.

In the similarity analysis model, we defined the similarity of artists, analyzed the similarity of artists between and within genres, and discussed the similarity between genres.

In the infectivity analysis model, we used the data of the above model to analyze whether the influencer really has an influence on the follower, and selected infectious music features.

Finally, we analyzed the data based on the historical background, and conducted a sensitivity analysis and evaluation of our model.

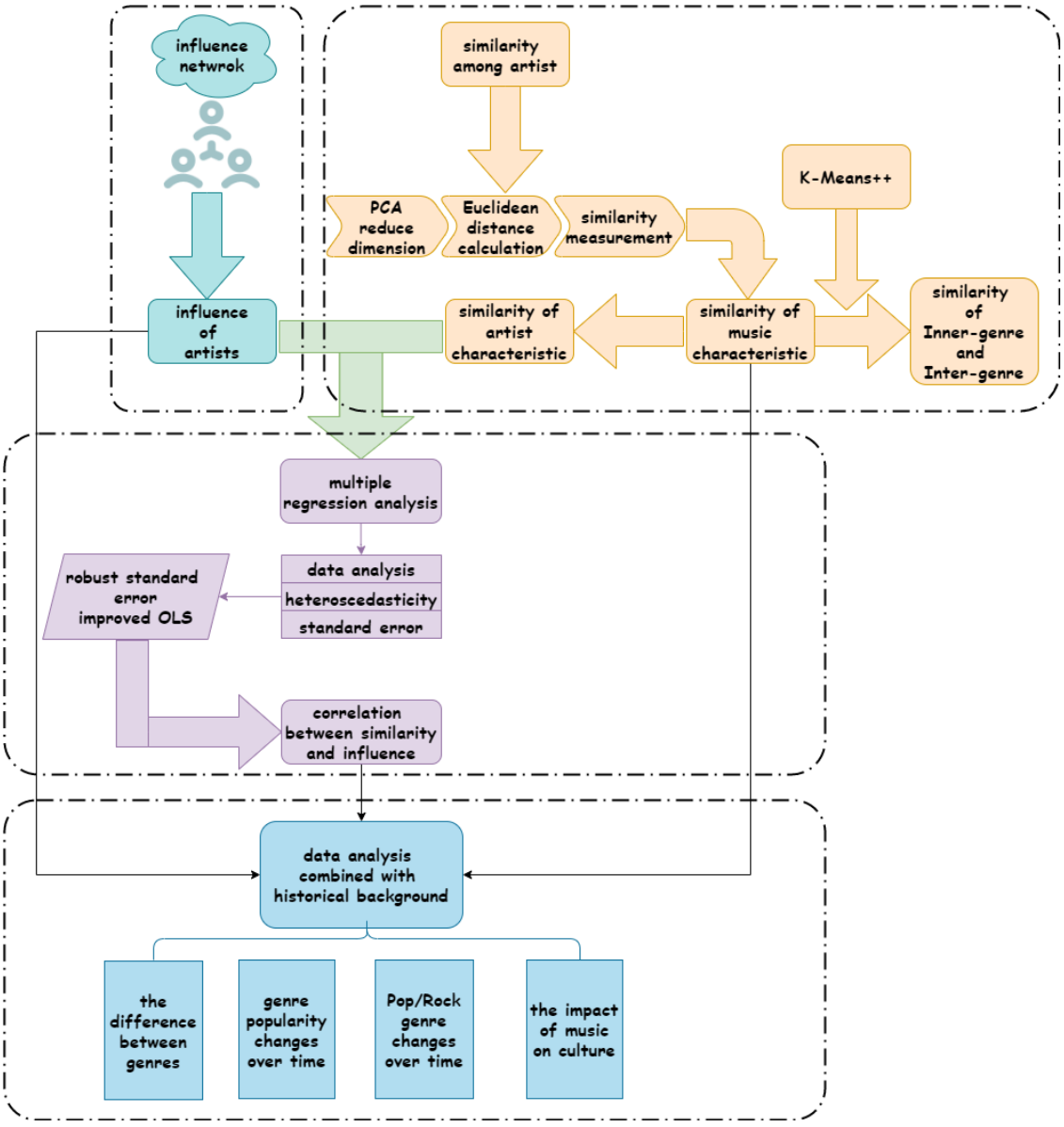


Figure 1: Overview of Our Work

2 Model Preparation

2.1 Assumptions

Since the problem involves big data, we need to make the following assumptions about the model, and these assumptions are reasonable:

(1) The influence of the indirect relationship does not exceed the two-level relationship.

Sometimes the indirect relationship cannot be ignored, but we think that the influence of the indirect relationship becomes smaller as the number of relationship levels increases, and the impact does not exceed the two-level relationship. For example, we think that a person's teacher has an influence on him, but his teacher's teacher has no influence on him.

(2) The artists in the genre are the same, and there is no weight. If we consider that each artist in the genre has different contributions to the genre, this will increase the difficulty of calculation. In order to simplify the calculation, we think that the contributions of artists within the genre are the same.

(3) Artists with great musical influence have a high level of influence on his followers. Our model thinks that an artist with greater influence has a higher level of influence on his followers, and an artist with no small musical influence has a particularly large influence on his followers.

2.2 Notations

Symbol	Description
NF_i	The number of followers of the artist i
α	Impact factors that balance the influence of indirect followers
MI_i	Artist i's musical influence
SDC_i	The standard deviation in cluster i
SDG_i	The standard deviation in genre i
NI_i	The number of influencers that influence follower i
IL_{ij}	The influence level of influencer i on follower j

3 Music Influence Model

In order to create a music influence network that connects influencers and followers, we build a music influence model. In this model, we first construct a directed graph connecting influencers and followers based on the influence_data data set. Then define the formula for calculating the musical influence of artists, calculate the musical influence of each artist according to the formula, and use this indicater to describe the influence of influencers in this network [3]. Finally, explain what this indicator measures in the influencer network sub-network.

3.1 Model Establishment

3.1.1 Directed Network Creation

We use python to draw pictures based on the influence_data data set.First, the name of the unique artist in the data set is used as a node, and then according to the relationship between the influencer and the follower in the data set, connect the node representing the influencer with the node representing the follower, and finally the arrow is used from the node representing the influencer points to the node representing the follower. We get a directional network connecting influencers and followers.

3.1.2 MI_i Parameters Development

In the second step, we need to determine the influence of each artist. Because we can only use the influence_data dataset, we can only determine the artist's musical influence through the network generated in the previous. We think that the musical influence of an artist has two parts.

One part is the number of followers of the artist(NF_i). In the network, we can easily count the followers of each influencer. The number of followers of an artist is equal to the outgoing degree of the node in the network. If the number of followers of an artist is larger, then the musical influence of the artist should be greater, which is consistent with our perception.

The other part is the influence of the followers of the artist's followers. Because the followers of an artist may also have followers, and this influence cannot be ignored sometimes. But this influence should be smaller as the number of layers increases, so we need to multiply this indirect relationship by a coefficient less than 1 to simulate its influence on the artist's musical influence. We call this coefficient the influence factor α.

In our daily lives, we usually remember the teacher of an outstanding person, at most we will hear of his teacher, and almost no one will remember his teacher's teacher. So here we think that the influence of this indirect relationship will last two levels at most, and after that, we think that the influence factor is zero [13]. We got the formula for calculating the artist's musical influence:

$$MI_i = NF_i + \alpha \sum NF_j \tag{1}$$

There may be a closed loop of relationships in the directed network (artist A is an influencer of artist B, and artist A is also a follower of artist B), which may lead to incalculable problems in the calculation process. Our formula solves this problem perfectly.The musical influence of the artist we get can measure the influence of an artist in this directed network, and the ranking of the artist among these artists can be obtained according to the order of influence.

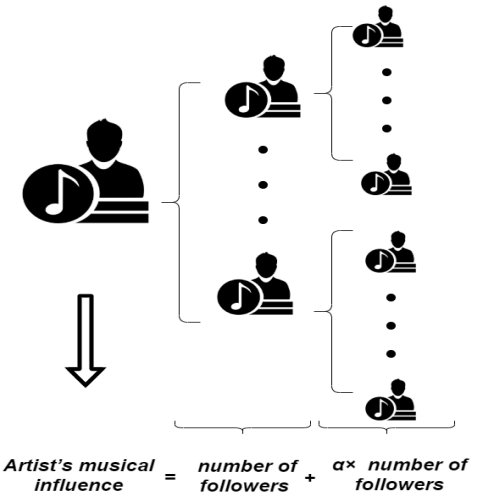


Figure 2: Formula Description Diagram

3.1.3 Directed Network of Genres and That of Time Periods Creation

In the third step, we take the artists of one of the genres and their influence relationship as a directed network. This directed network is a sub-network of the above directed network. Then take the influence relationship between these artists and calculate the influence according to the formula:$MI_{\hat{i}} = NF_{\hat{i}} + \alpha \sum NF_{\hat{j}}$, which is a subset of the above influence calculation formula(1).

Finally, the resulting directed network is the influence among artists within a genre, and the calculated influence($MI_{\hat{i}}$) is the influence of the artists in this genre within the genre. Because the point set of the directed network has shrunk, our musical influence in this sub-network represents the influence of an artist on other artists in this genre.

In the same way, we can also take the artists who debuted in a certain time period and their influence relationships to make a directed network, get the sub-network of this time period, and get similar conclusions.

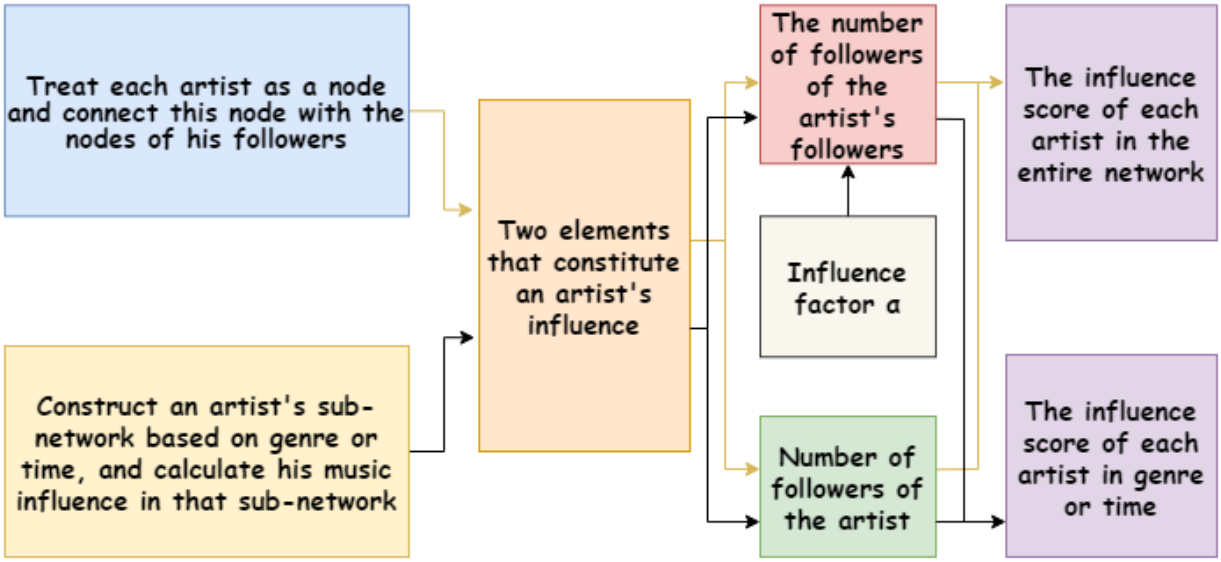


Figure 3: Overview of The Above Musical Influence Model

3.1.4 Influence Analysis between Genres

We choose two genres and analyze the influence between them. Because the musical influence defined above contains two parts, direct influence and indirect influence, we also need to consider these two parts when considering the influence between the two genres.

In the first part, we select two genres of artists that have mutual influence, and record the number of artists in one genre who have a direct influence on artists in the other genre, $NF_{\bar{i}}$.

In the second part, because we think that there are at most two levels of indirect influence, we only need to consider that artists of one genre indirectly influence artists of another genre

by influencing artists of other genres. Record the number of artists in the middle $NF_{\bar{j}}$. The influence between genres is obtained by the formula $MF_{\bar{i}} = NF_{\bar{i}} + \alpha NF_{\bar{j}}$.

3.2 Model Result

For the first step, we get a directed network, and then according to the second step, we get the music influence of each artist, and get the ranking of the artists according to the music influence.

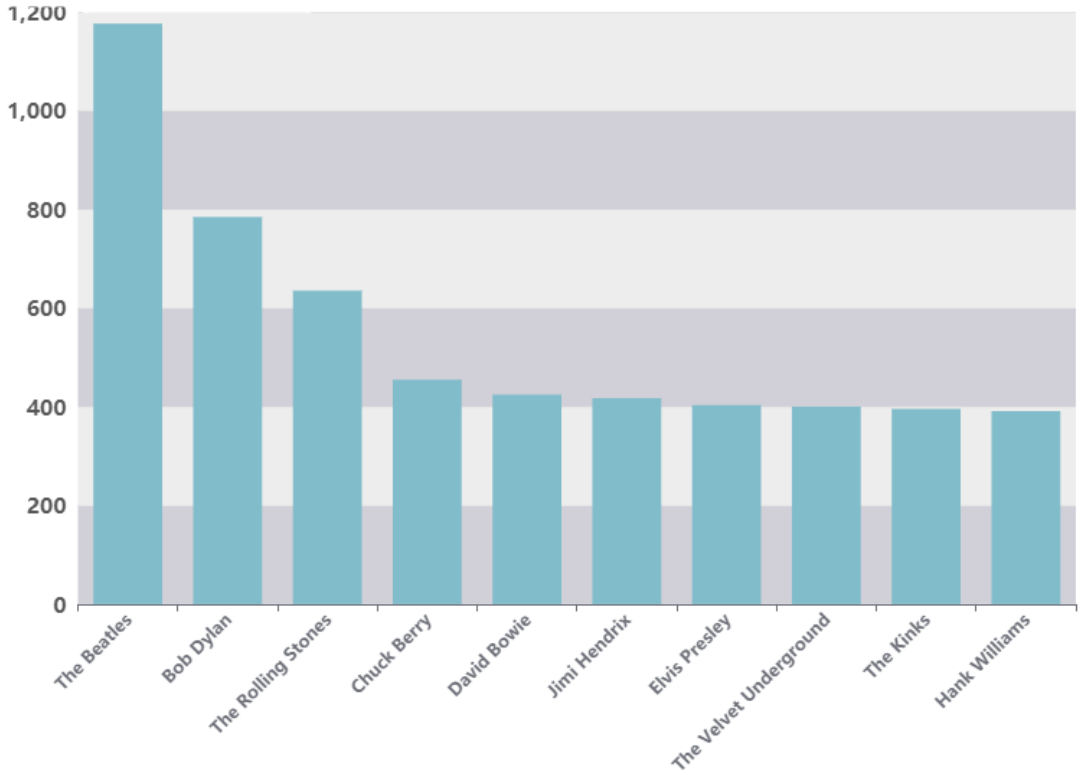


Figure 4: Music Influence Model's Result

From the figure, we can see that the Beatles ranked first in the rankings with more than 50% of the influence of the second place. Analyzing the data, we found that the Beatles have 615 followers. 158% of Bob Dylan, which shows that our calculation formula and result are correct.

For the third step, we select the Avant-Garde genre to analyze, and draw the directed network of this genre as shown in Figure 5 and the influence of artists in this genre in this genre as shown in Figure 6.(We have also produced influence diagrams of artists who debuted in 1940, please check in the appendix)

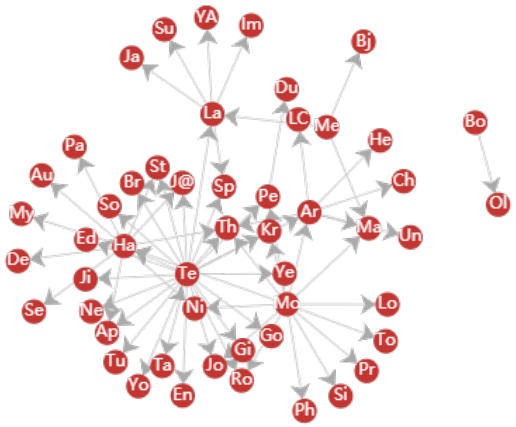


Figure 5: Avant-Garde Genre Directed Network

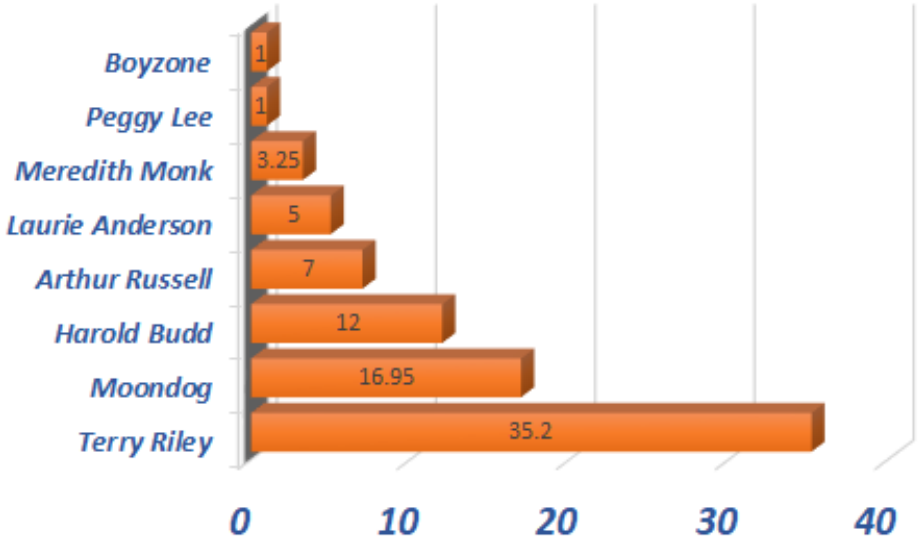


Figure 6: Artist Influence of The Avant-Garde Genre

It can be seen from Figure 6 that Terry Riley's influence is 35.2, ranking first in the Avant-Garde genre. Contact Figure 5 to find that Terry Riley (Te) is an influencer of many artists, so Terry Riley's influence score is higher.

4 Similarity Analysis Model

In this section, we define the measure of similarity between artists, analyze the similarity of artists between and within genres, and analyze the similarity between genres.

4.1 Model Establishment

4.1.1 Definition of Music Similarity Measure

In the data set we are given, music has characteristics such as danceability, energy, etc. We need to use these characteristics to determine the similarity of different music.

In order to cluster all artists according to their musical styles, so as to obtain the classification of multiple artists with similar musical styles, we need to establish a coordinate system based on the characteristics of music. Different musical characteristics are different coordinate axes, and each artist is a point of this coordinate system.

Due to the many features of music, there are many coordinate axes. In order to simplify the calculation, we first use principal component analysis (PCA) to reduce the dimensionality of the features of music [2].

Assuming that the data of all music characteristics of the artist constitutes a matrix x, we need to standardize each data in this matrix. Calculate the sum $\bar{x_j} = \sum_i^n x_{ij}$ and standard deviation σ_j of each column in x. According to the formula $Xij = \frac{x_{ij} - \bar{x_j}}{\sigma_j}$, we get the standardized matrix X.

$$x = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix} \qquad X = \begin{bmatrix} X_{11} & \dots & X_{1m} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{nm} \end{bmatrix} \qquad R = \begin{bmatrix} r_{11} & \dots & r_{1m} \\ \vdots & & \vdots \\ r_{n1} & \dots & r_{nm} \end{bmatrix}$$

Then according to the formula $r_{ij} = \frac{1}{n-1} \times \sum_{k=1}^{n} cov(X_{ki} - \bar{X_i}, X_{kj} - \bar{X_j})$, we get the covariance matrix R of X. Calculate the eigenvalue of R and the corresponding eigenvector, calculate the cumulative contribution rate of the eigenvalue $ContributionRate = \frac{\lambda_i}{\sum_i^p \lambda_i}$, take the eigenvector corresponding to the eigenvalue whose cumulative contribution rate reaches a certain value as the weight vector of the music indicator, and get different principal components.

We define the Euclidean distance between artists in these principal components as their musical similarity.

$$MusicalSimilarity = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} \tag{2}$$

4.1.2 Similarity Analysis of Inner-genre and Inter-genre by K-means++

Calculate the data of each artist in these principal components, and use the K-means++ algorithm to cluster the representative artists [9].

The K-means++ algorithm is a clustering algorithm [14]. After the user gives the classification number K, the system randomly selects a sample point as the initial center point of a cluster.

Calculate the distance from other sample points to the center point, and use the normalized value of the distance as the probability of selecting the center point of the next cluster (the greater the distance, the greater the probability), repeat until K center points are obtained until.

Then the system automatically calculates the Euclidean distance from each sample point to each center point, and classifies the sample points close to the center point into one cluster. Calculate the distance between sample points of the same cluster, find the position of the new center point and update the center point.

Repeat this step until the center point no longer changes, then stop and generate clustering results.

After obtaining the clustering results, we calculate the standard deviation in each cluster(SDC_i). At the same time, calculate the standard deviation within each genre(SDG_i) and compare them. If most of the SDCi are smaller than SDGi, it means that the similarity of artists within the genre is not obvious, and our clustering results show the similarity of artists more.

At the same time, SDG_i can also show the similarity of artists within the genre. If the SDG_i is smaller, the similarity of artists within the genre is higher.

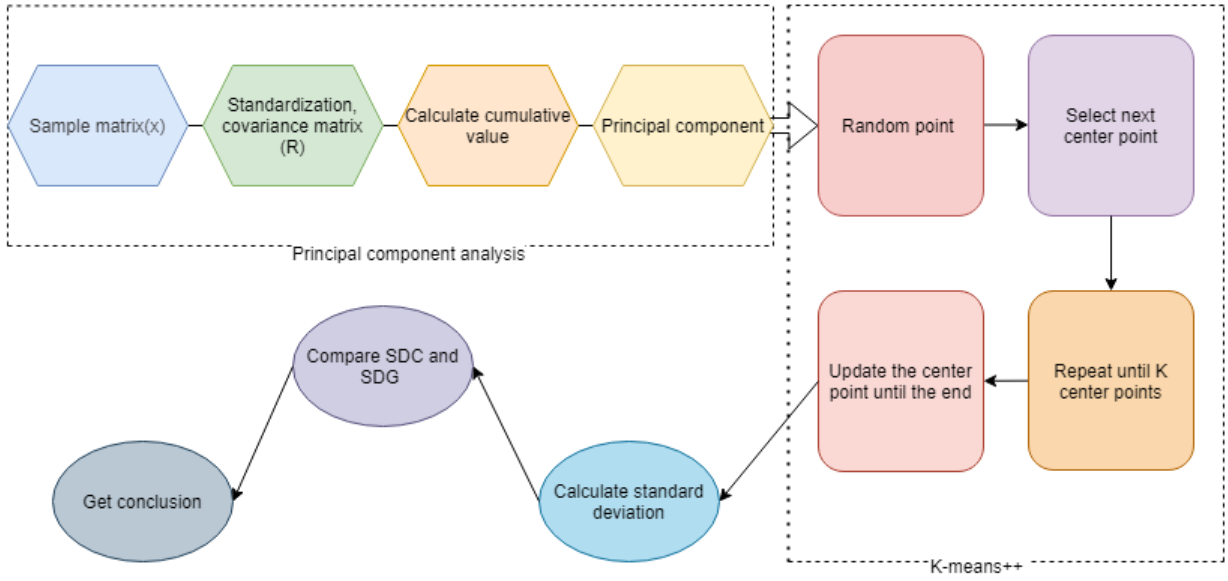


Figure 7: Overview of Above Steps

4.1.3 Similarity Analysis between Genres

Based on the music similarity defined above, we can calculate the similarity between genres. Because the genre is composed of many artists, the similarity of artists is measured by their European distance in the principal component. We think that the similarity of different genres is also measured by their European distance in the principal component [11].

Just like finding the center of an object in three-dimensional space, you need to average the coordinates of all points of the object to represent the coordinates of the center. Here, in order to determine the center of a genre, we need to average the values in the principal components of the artists in the genre, and use the average to represent the value of the genre in the principal components. Calculate the distance between the centers of different genres, and then we can get the similarity between the genres.

4.2 Model Result

4.2.1 PCA and Clustering Results

The results of PCA are as follows:

Table 1: Eigenvalue and Cumulative Contribution Rate

Eigenvalues	**2.9143**	**1.6881**	**1.3096**	**1.2299**	**0.9981**	**0.8830**	0.8296	0.7230
CCR	**0.2429**	**0.3835**	**0.4927**	**0.5951**	**0.6783**	**0.7519**	0.8210	0.8813

Because the cumulative contribution rate of the first 6 eigenvalues reached 75%, we decided to select these six eigenvectors as the coefficient vectors of the principal components.

In order to show the relationship between the clustering results and genres, we set the K value to be the number of large genres (the number of people exceeds 100). According to Is Intra-genre Correlation Higher than Inter-genre Analysis, we got the following clustering results and SDC.

Cluster	Euclidean_SD
1	0.619
2	0.646
3	0.632
4	0.635
5	0.878
6	1.114
7	0.484
8	1.913
9	0.531
10	0.690

Genre	Euclidean_SD
Avant-Garde	2.291
Blues	0.869
Children's	0.245
Classical	2.419
Comedy/Spoken	3.530
Country	0.565
Easy Listening	1.228
Electronic	1.051
Folk	1.009
International	1.286
Jazz	1.279
Latin	0.855
New Age	1.983
Pop/Rock	1.060
R&B;	0.829
Reggae	0.802
Religious	1.098
Stage & Screen	1.185
Unknown	1.567
Vocal	0.966

Figure 8: SDC and SDG

From the Figure 8, we find that the standard deviation of Children's genre is the smallest. After analyzing the data, we think that because there are only three artists in the Children's genre, and many of the audiences of this genre are children, the music characteristics of the artists in this genre are basically similar, resulting in a small standard deviation for this genre.

From the Figure 8 we can see that the maximum value of SDC is much smaller than the maximum value of SDG, and the overall SDC is also smaller than SDG. Therefore, we think

that the similarity of artists within genres is not greater than the similarity of artists between genres. In order to explore this issue in depth, we make a genre composition diagram of the sixth cluster.

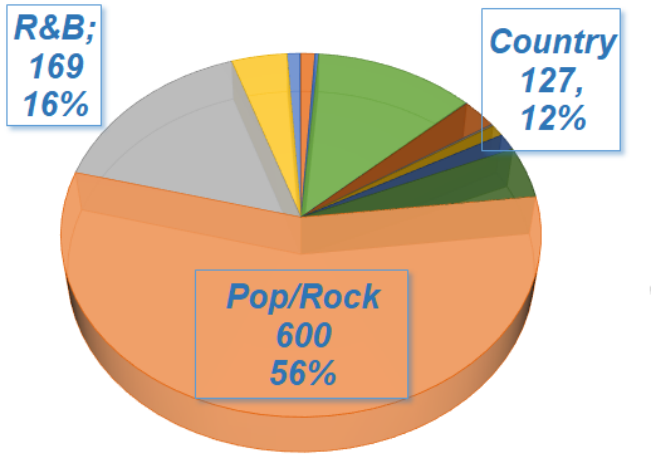


Figure 9: Genre Composition Diagram of The Sixth Cluster.

The sixth cluster consists of 1132 artists. In this cluster, there are more than half of Pop/Rock artists, 16% R&B artists and 12% Country artists. Explain that the sixth cluster is composed of more artists of different genres. Although the number of people is large, the standard deviation of the sixth cluster is not very large, only 1.114, which is acceptable. Therefore, we think that the similarity of artists within the genre is not greater than the similarity of artists between the genre.

4.2.2 Similarity Matrix between Genres

Calculating the similarity between multiple genres, we can get a similarity matrix as shown in the Figure 10.

	Avant-G	Classical	Country	Electronic	Folk	Jazz	Pop/Rock	R&B;
Avant-G	0.00	1.55	3.61	3.57	2.36	1.15	3.86	3.69
Classical	1.55	0.00	4.82	4.88	3.33	2.62	5.03	5.07
Country	3.61	4.82	0.00	2.23	1.86	2.77	1.40	1.20
Electronic	3.57	4.88	2.23	0.00	3.16	2.58	1.38	1.41
Folk	2.36	3.33	1.86	3.16	0.00	1.90	2.76	2.49
Jazz	1.15	2.62	2.77	2.58	1.90	0.00	2.99	2.65
Pop/Rock	3.86	5.03	1.40	1.38	2.76	2.99	0.00	1.35
R&B;	3.69	5.07	1.20	1.41	2.49	2.65	1.35	0.00

Figure 10: Similarity Matrix between Genres

We can see that the Euclidean distance between the Classical genre and other genres is larger, indicating that the percentage of similarity between the Classical genre and other genres is smaller.

5 Improved Regression Analysis with Robust Standard Error

In this section, we discuss whether influencers really influence followers, and analyze which musical characteristics are more "infectious".

5.1 Model Establishment

5.1.1 The Definition of Influence Level

In the above two models, we define the artist's musical influence MF_i by the number of followers of the artist, and define the artist's similarity measure by the Euclidean distance between the musical characteristics of different artists. In order to discuss whether influencers really influence followers, We need to study the influence level of influencers on followers

The influence level IL_{ij} of influencer i on follower j should include two aspects. On the one hand, it is the influencer's musical influence. We think that the greater the influencer's musical influence MF_i, the more the followers will receive the influence of this influencer. On the other hand, it is the number of influencers that influence a particular follower. We think that if the number of influencers of a followerNI_j is larger, the influence of each influencer on him is smaller.Therefore, we define the influence level of influencer a on follower b as follows.

$$IL_{ab} = \frac{MF_a}{NI_b} \tag{3}$$

5.1.2 Multiple Regression Analysis

In order to explore whether influencers really influence followers, we need to analyze the similarity of the musical characteristics of the two artists and the correlation of their degree of influence. We use the least squares method to perform multiple regression analysis on the musical characteristics. Due to the different dimensions of music characteristics, we first need to remove the influence of dimensions. Here, we use Z standardization method to standardize music characteristics.

$$\mu_i = \frac{\sum_j^n x_{ij}}{n} \tag{4}$$

$$\sigma_i = \sqrt{\frac{1}{n}\sum_j^n (x_i j - \mu_i)^2} \tag{5}$$

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{6}$$

Then we perform multiple regression analysis.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \tag{7}$$

5.1.3 Improved Least Square Regression with Robust Standard Error

The standard error is a statistic that describes the standard deviation of repeated sampling of the sample population. Each sampling of a sample population will generate a mean value, multiple sampling can get multiple mean values, and the standard error is the standard deviation of these means. The ordinary standard error is based on the same variance of the sample, but if the variance of the disturbance term is not constant in different observations, then there will be heteroscedasticity [15].

We know that cross-sectional data may have heteroscedasticity, and the results we need are obtained by testing the significance of the variables [4]. In the significance test of the variables, the constructed t-statistic is based on the correct estimation of the parameter standard deviation. At this time, if we still use the ordinary standard error without considering the existence of heteroscedasticity, the standard deviation of the parameter will be biased, which will cause the t-test to lose its meaning. Therefore, after using OLS regression, we need to test whether the model has heteroscedasticity [1].

$$nR^2 \backsim \chi^2(v), v = 77 \tag{8}$$

We make the assumption that there is no heteroscedasticity in the disturbance term.

$$H_0 : E(\epsilon_i^2 | x_2, \ldots, x_k) = \sigma^2 \tag{9}$$

Then we test the possibility of the disturbance term in the model.

Table 2: White Test Result

$\chi^2(77)$	177.03
Prob > χ^2	0.00

n(df)		1	2	.....	76	77	78
p	0.995	0.000	0.010	.....	47.997	48.788	49.582
	0.990	0.000	0.020	.....	50.286	51.097	51.910
	0.900	0.016	0.211	.....	60.690	61.586	62.483

Figure 11: Chi-square Distribution Table

From the Figure 11, the significance level of adjoint probability $Prob{\cdot}Chi-Square(77) > 48.788$, which means that we can reject the hypothesis (formula 9) at the 99.5% confidence level and believe that the disturbance term has heteroscedasticity [10].

Stock and Wastson (2011) recommend that the "robust standard error improved OLS" method should be used in most situations. So we use a robust standard error modified OLS to solve the problem of heteroscedasticity [12].

5.2 Model Result

First, we test the rationality of the model.

Table 3: Model Rationality Test

F(11,42317)	18.99
Prob>F	0.0018
R-squared	0.0058

It can be seen that the P value of F test of the model's joint significance test is <0.05, and the hypothesis 9 can be rejected at the 95% confidence level, which shows that our model is reasonable.

Then we use the model to analyze the significance test results of the independent variables.

$$t = \frac{\hat{\beta}_j - \beta_j}{Se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{(\sigma_\mu)^2 c_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{(\sigma_\mu)^2 (X^T X)^{-1}_{jj}}} (j = 0, 1, 2, ..., k) \tag{10}$$

Robust					
influence	Coef.	Std. Err.	t	P>t	Beta
danceability	-1.3318	0.7349	-1.8100	0.0700	-0.0095
energy	0.4019	0.6363	0.6300	0.5280	0.0043
valence	-1.0805	0.4857	-2.2200	0.0260	-0.0115
tempo	-0.0139	0.0042	-3.3300	0.0010	-0.0163
loudness	-0.1187	0.0246	-4.8300	0.0000	-0.0267
key	0.1726	0.0218	7.9100	0.0000	0.0468
acousticness	2.5247	0.3396	7.4300	0.0000	0.0373
instrumentalness	-0.4022	0.2826	-1.4200	0.1550	-0.0060
liveness	3.8523	0.8304	4.6400	0.0000	0.0292
speechiness	-5.8996	1.1435	-5.1600	0.0000	-0.0234
duration_ms	0.0000	0.0000	2.6400	0.0080	0.0148
_cons	4.5107	0.1394	32.3500	0.0000	.

Figure 12: Model Result

It can be seen that tempo, loudness, key, acousticness, liveness, speechiness, $duration_ms$, these indicators can reject the hypothesis at the 99% confidence level, and pass the t test. The valence and danceability can reject the hypothesis at 95% and 90% confidence levels respectively, and pass the t test. Therefore, we think that the similarity of these musical characteristics is significantly related to relative influence, and influencers do have an influence on followers.

6 Background-based Data Analysis

In this section, we will analyze the obtained data combined with historical background.

6.1 The Difference between Genres

In the second model, we analyzed the similarities between the genres and the artists within the genres, and got the conclusion that the similarities of the artists within the genres are no greater than the artists between the genres. Here, we discuss in depth what musical characteristics of artists within and between genres are similar, and what musical characteristics are different. We selected part of the music feature data of the major genres to make the following figure.

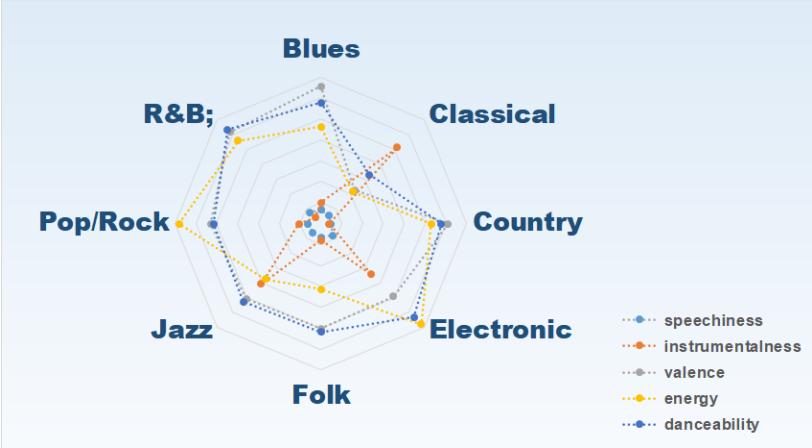


Figure 13: Comparison of Music Characteristics between Genres

We found that there are many different characteristics between different genres. The biggest difference is the instrumentalness indicator and energy indicator. The Instrumentalness indicator indicates whether the music contains human voices. If the value is closer to 1, the more likely it is to contain human voices. The Blues, R&B, Pop/Rock, and Folk genres basically contain no human voice, because most of them rely on musical instruments to express their emotions. Classical, Electronic, and Jazz genres basically contain human voice, because most of their music relies on the author to sing to express emotions.

The Energy indicator represents the intensity and activity of the music. If the value is closer to 1, the intensity and activity of the music are higher. The music of the Pop/Rock and Electronic genres has a strong energy, because most of their music is made with electronic devices and is full of rhythm. Classical genre music has the lowest energy because their music usually uses classical instruments.

6.2 Analysis of Genre Popularity over Time

We select the sum of the popularity of artists who debuted in a genre at a time as the popularity of the genre at that time, so that we can get the popularity data of different genres at different times. We use these data to get the following figure.

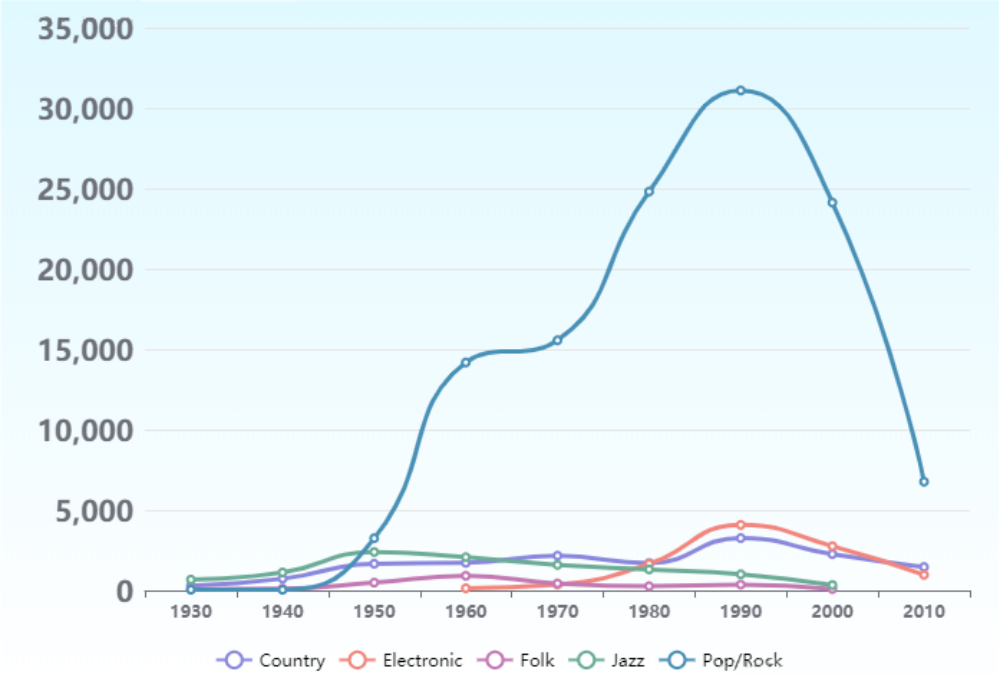


Figure 14: Genre Popularity and Time Relationship

We found that these genres were not very popular at the beginning of the 20th century, because World War I and World War II broke out during this period. People in this period are worried every day, so the popularity of music is not big.

By 1945, the Second World War ended, the world was peaceful, people were happy, and people needed music to express their inner joy, so the popularity of the genre rose. But Pop/Rock's growth rate is the largest, because Pop/Rock's style is more open and can better express people's emotions.

From the 1950s to the 1980s, many debut artists chose the Pop/Rock genre, so the popularity of the Pop/Rock genre rose sharply during this period. With the development of science and technology, electronic music was born in the 1960s and continues to develop. However, the overall popularity of the Pop/Rock genre continued to decline after the 1990s, which may be caused by the decomposition of this genre into many genres.

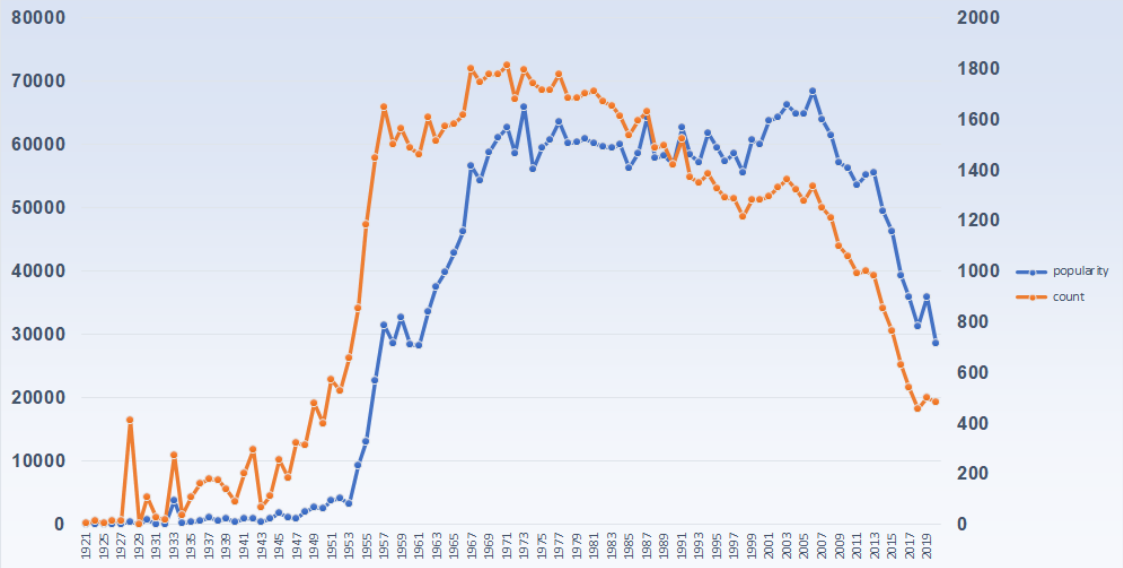


Figure 15: Pop/Rock's Popularity and Count

During the 1950s and 1960s, this period was a post-war recovery period. The relatively loose political environment after the war gave people ample space for the development of music after the war.

At that time the environment is fertile ground for the development of music, the fifties, there was born like Chuck Berry, James Brown famous musicians, their influence in our network were 454.65,340.35, the former is known as the "Rock of Shakespeare". With their efforts, Pop/Rock in this period had initial development. Then came the 1960s, and the development of Pop/Rock became more vigorous.

From our network, we found that in this era, there are bands or musicians like The Beatles, Bob Dylan, and The Rolling Stones. Their influence is huge, with 1176.1, 784.3 and 635.1 respectively. Their huge influence also allowed this genre to develop further.

6.3 Pop/Rock Genre Changes over Time

We choose Pop/Rock genre to analyze. We take the average of the musical characteristics of artists who debuted in a period within the genre as the musical characteristics of the genre in this period, and use these data to get the following figure.

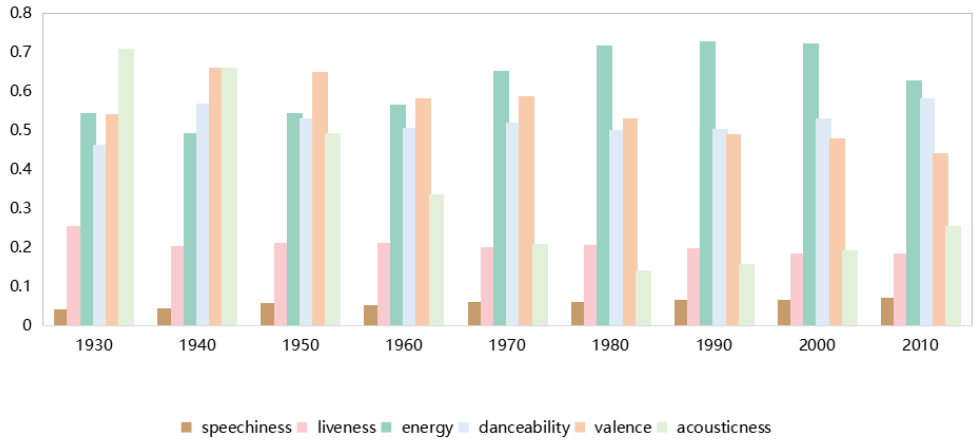


Figure 16: Pop/Rock Genre Changes over Time

We found that the energy indicator of the Pop/Rock genre rose before 1990, and the valence indicator rose before 1950, because the world war was going on before 1945, and people during the war needed music to encourage themselves. The higher the energy indicator and the valence indicator, the more encouraging music is. After the World War, people also needed encouraging music to express joy, so the energy indicator of the Pop/Rock genre was still rising, but the valence indicaor began to decline.

The Acousticness indicator was in a downward trend from 1930 to 2000. Because of the development of science and technology, people used computers and other electronic devices to process music, which led to a decline in the acoustics of music.

6.4 The Impact of Music on Culture

The influence of music on culture cannot be ignored. Culture is a series of common values, belief, knowledge, goals (goals), attitude, rules of conduct, etc. in society, all of which are reflected in music [7].

In the study of international relations, some Western scholars believe that Pop/Rock music became an effective weapon for American foreign cultural export during the Cold War. Because these two forms of music can embody the spirit of freedom and rebellion in the West, they have affected the attitude of young people in the Soviet Union and Eastern Europe towards the "totalitarian" regime [5].

From our data analysis, we can also find that the popularity of Pop/Rock, which flourished in the United States during that period, increased significantly, and it is likely that it had a certain impact on the political situation during the Cold War.

7 Sensitivity Analysis

In the network of the first model, we developed an indicator to measure the influence of a musician. This indicator relates to the influence coefficient of α. We change this coefficient,

that is, to adjust the way each musician uses indirect influence. The overall size of the influence gained. We let it fluctuate between 0.01 and 0.2 and get multiple results.

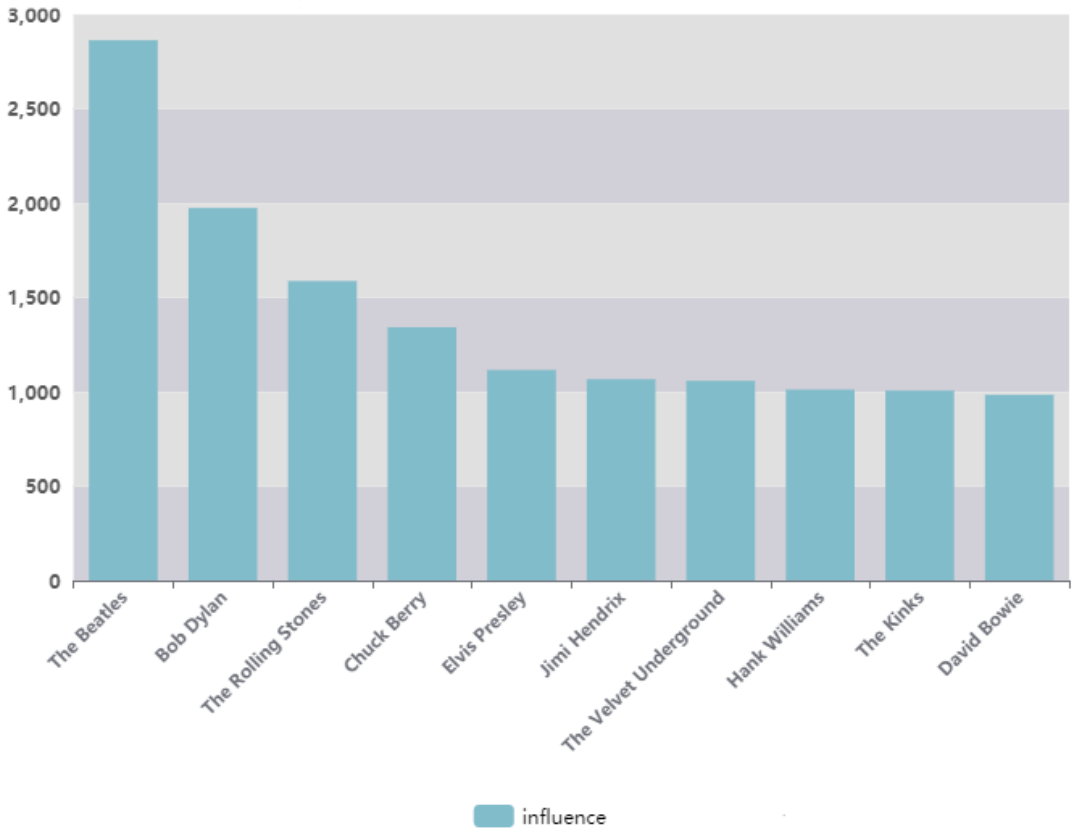


Figure 17: Rank When α=0.2

The ranking among musicians has not changed much. In the article, we show the top ten when α=0.05. From the results, we know that the adjustment of the coefficient will not have much impact on the ranking as long as it is within a reasonable range (1e-2 5e-1).

8 Model Evaluation

8.1 Strengthness

★ Our formula 1 for defining the influence of artists is reasonable. It not only considers the influence of indirect relationships, but also solves the closed-loop problem of the directed network in which artists interact with each other.

★ Our model analyzes artists, genres, and within genres from different angles, with rigorous structure and comprehensive conclusions.

★ Our model makes reasonable improvements to OLS to solve our problem.

8.2 Possible Improvements

♦ When we consider the contribution of artists within the genre to the genre, we think that each artist's contribution to the genre is the same, and there is no weight. If there is enough time, we can consider the weight of each artist.

♦ We have not considered that an artist whose small musical influence may have a great influence on his followers. If the data provided is rich enough, we can consider this situation.

9 A Document to ICM

To ICM Society:

Today, music has become an indispensable part of people's daily life. In order to explore the development of music and the impact of music on culture, we used the provided data to formulate an artist's music influence indicator and music similarity indicator. We analyzed the relationship between artists and genres, and studied the changes of genres and artists over time. The following is an overview of our work.

We have built three models to illustrate the definition and specific meaning of our indicators, and analyzed the relationship between artists and genres in the model. Through our defined music influence indicator, all artists can be ranked for influence, and the status of artists in this field can be obtained. And we initially defined the level of influence among artists, which allows us to analyze the influence among artists more completely. Analyzing these data, we found that influencers do have influence on followers, and the great artists do have greater influence compared with others. For example, the indicator of *The Beatles* is higher than that of anyone.

Through our defined similarity measure, we can calculate the music similarity between artists. Calculating these data, we find that the similarity between artists within genres is not necessarily greater than the similarity between artists between genres. This means that there are some artists who have a variety of music styles at the same time, which blurs the boundaries between genres. We can use similar methods to analyze the differences between genres by defining the average value of the musical characteristics of artists within the genre. The results tell us that there are big differences in instrument and energy indicators between genres, while the gap in other indicators is not obvious.

Using the data of the influence level and similarity between artists to perform regression analysis, we conclude that most of the musical features are infectious under high confidence. This shows that although the music characteristics of each genre are different, the music of each genre has a certain degree of appeal.

We analyzed the popularity of different genres at different times and found that the popularity of the genre is correlated with historical events. We also discussed how the music characteristics of the Pop/Rock genre change over time, and found that it is also related to historical events.

Through our model, you can have a clearer understanding of how music interacts and develops. Of course, there are some shortcomings in our model, but as more data is provided, our network can study the influence of music in society by improving the influence of artists on ordinary people.

Thanks for taking the time out of your busy schedule to read my document. Hope our model can help.

From Team #2118261

10 Appendices

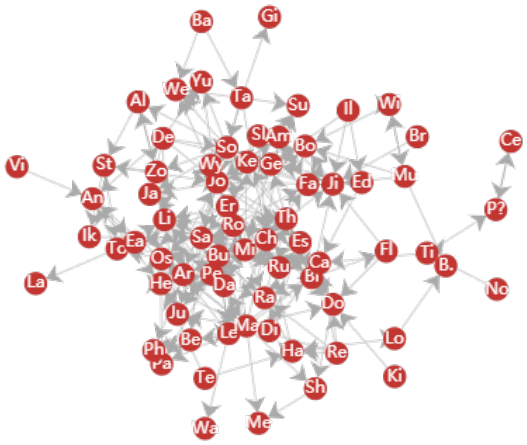
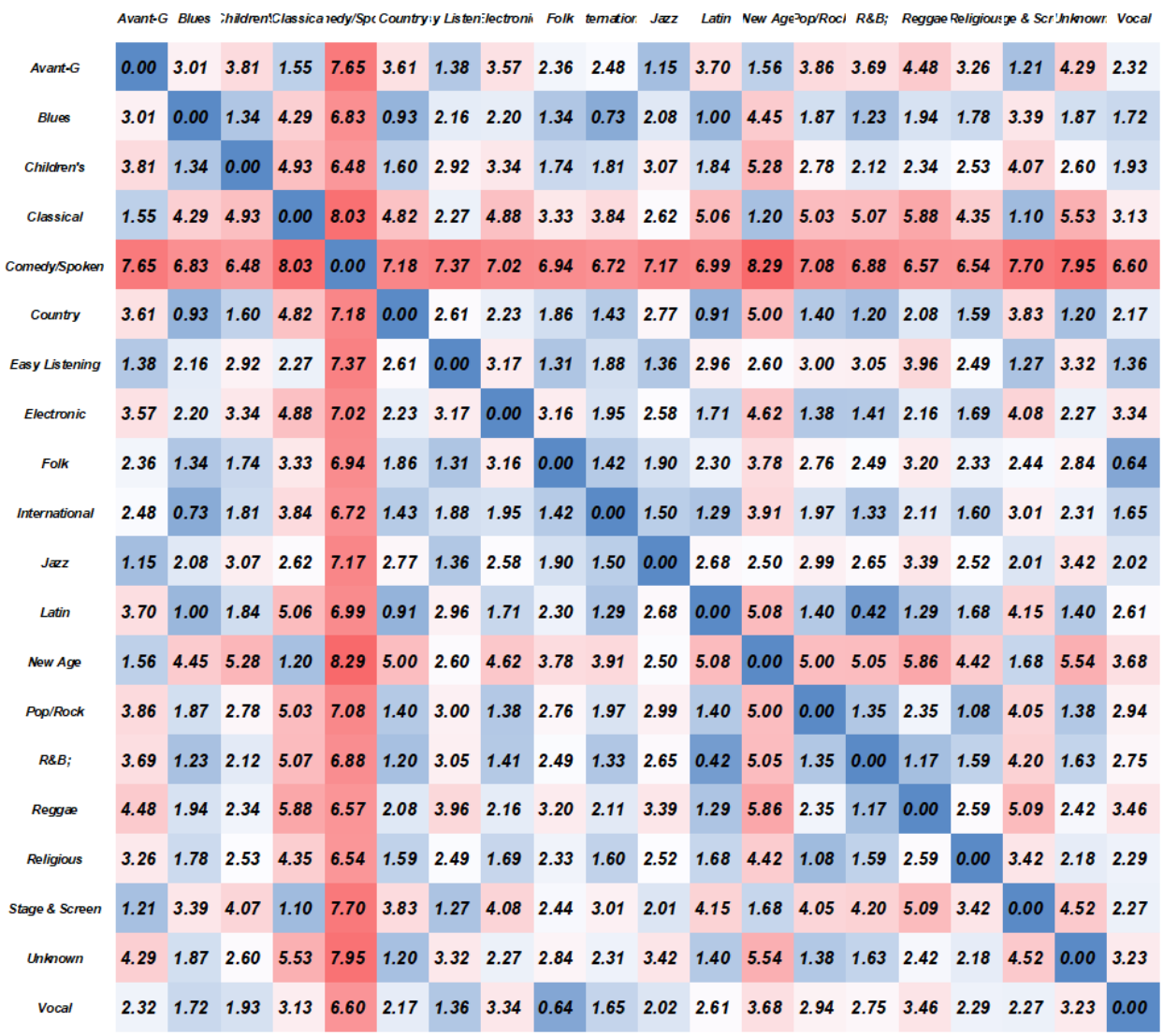


Figure 18: A Directed Network of Artists Who Debuted in 1940



	Avant-G	Blues	Children's	Classica	edy/Spc	Country	y Listen	lectroni	Folk	ternatior	Jazz	Latin	New Age	Pop/Roci	R&B;	Reggae	Religious	ge & Scr	Unknown	Vocal
Avant-G	0.00	3.01	3.81	1.55	7.65	3.61	1.38	3.57	2.36	2.48	1.15	3.70	1.56	3.86	3.69	4.48	3.26	1.21	4.29	2.32
Blues	3.01	0.00	1.34	4.29	6.83	0.93	2.16	2.20	1.34	0.73	2.08	1.00	4.45	1.87	1.23	1.94	1.78	3.39	1.87	1.72
Children's	3.81	1.34	0.00	4.93	6.48	1.60	2.92	3.34	1.74	1.81	3.07	1.84	5.28	2.78	2.12	2.34	2.53	4.07	2.60	1.93
Classical	1.55	4.29	4.93	0.00	8.03	4.82	2.27	4.88	3.33	3.84	2.62	5.06	1.20	5.03	5.07	5.88	4.35	1.10	5.53	3.13
Comedy/Spoken	7.65	6.83	6.48	8.03	0.00	7.18	7.37	7.02	6.94	6.72	7.17	6.99	8.29	7.08	6.88	6.57	6.54	7.70	7.95	6.60
Country	3.61	0.93	1.60	4.82	7.18	0.00	2.61	2.23	1.86	1.43	2.77	0.91	5.00	1.40	1.20	2.08	1.59	3.83	1.20	2.17
Easy Listening	1.38	2.16	2.92	2.27	7.37	2.61	0.00	3.17	1.31	1.88	1.36	2.96	2.60	3.00	3.05	3.96	2.49	1.27	3.32	1.36
Electronic	3.57	2.20	3.34	4.88	7.02	2.23	3.17	0.00	3.16	1.95	2.58	1.71	4.62	1.38	1.41	2.16	1.69	4.08	2.27	3.34
Folk	2.36	1.34	1.74	3.33	6.94	1.86	1.31	3.16	0.00	1.42	1.90	2.30	3.78	2.76	2.49	3.20	2.33	2.44	2.84	0.64
International	2.48	0.73	1.81	3.84	6.72	1.43	1.88	1.95	1.42	0.00	1.50	1.29	3.91	1.97	1.33	2.11	1.60	3.01	2.31	1.65
Jazz	1.15	2.08	3.07	2.62	7.17	2.77	1.36	2.58	1.90	1.50	0.00	2.68	2.50	2.99	2.65	3.39	2.52	2.01	3.42	2.02
Latin	3.70	1.00	1.84	5.06	6.99	0.91	2.96	1.71	2.30	1.29	2.68	0.00	5.08	1.40	0.42	1.29	1.68	4.15	1.40	2.61
New Age	1.56	4.45	5.28	1.20	8.29	5.00	2.60	4.62	3.78	3.91	2.50	5.08	0.00	5.00	5.05	5.86	4.42	1.68	5.54	3.68
Pop/Rock	3.86	1.87	2.78	5.03	7.08	1.40	3.00	1.38	2.76	1.97	2.99	1.40	5.00	0.00	1.35	2.35	1.08	4.05	1.38	2.94
R&B;	3.69	1.23	2.12	5.07	6.88	1.20	3.05	1.41	2.49	1.33	2.65	0.42	5.05	1.35	0.00	1.17	1.59	4.20	1.63	2.75
Reggae	4.48	1.94	2.34	5.88	6.57	2.08	3.96	2.16	3.20	2.11	3.39	1.29	5.86	2.35	1.17	0.00	2.59	5.09	2.42	3.46
Religious	3.26	1.78	2.53	4.35	6.54	1.59	2.49	1.69	2.33	1.60	2.52	1.68	4.42	1.08	1.59	2.59	0.00	3.42	2.18	2.29
Stage & Screen	1.21	3.39	4.07	1.10	7.70	3.83	1.27	4.08	2.44	3.01	2.01	4.15	1.68	4.05	4.20	5.09	3.42	0.00	4.52	2.27
Unknown	4.29	1.87	2.60	5.53	7.95	1.20	3.32	2.27	2.84	2.31	3.42	1.40	5.54	1.38	1.63	2.42	2.18	4.52	0.00	3.23
Vocal	2.32	1.72	1.93	3.13	6.60	2.17	1.36	3.34	0.64	1.65	2.02	2.61	3.68	2.94	2.75	3.46	2.29	2.27	3.23	0.00

Figure 19: Similarity Matrix Between Genres

References

[1] Oscar L Olvera Astivia and Bruno D Zumbo. Heteroskedasticity in multiple regression analysis: What it is, how to detect it and how to solve it with applications in r and spss. *Practical Assessment, Research, and Evaluation*, 24(1):1, 2019.

[2] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical methods*, 6(9):2812–2831, 2014.

[3] Pedro Cano and Markus Koppenberger. The emergence of complex network patterns in music artist networks. In *Proceedings of the 5th international symposium on music information retrieval (ISMIR 2004)*, pages 466–469. Citeseer, 2004.

[4] Matias D Cattaneo, Michael Jansson, and Whitney K Newey. Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361, 2018.

[5] Yudan Chen. Musical international relations: A cultural perspective of international relations studies. *China Foreign Affairs Review (Journal of China Foreign Affairs University)*, 3:120–134, 2011.

[6] Christophe Croux, Geert Dhaene, and Dirk Hoorelbeke. Robust standard errors for robust estimators, 2004.

[7] Donald Hodges. *Music in the human experience: An introduction to music psychology*. Routledge, 2019.

[8] B Huang. What kind of impact does our music really make on society, 2018.

[9] A comparative study of k-means, k-means++ and fuzzy c-means clustering algorithms. In *2017 3rd International Conference on Computational Intelligence& Communication Technology (CICT)*. IEEE, 2017.

[10] Junjie Liu and Yunzui Ye. Research on the influencing factors of urban residents' consumption: Based on the empirical study of guangxi. *"Journal of Guangxi Normal University" (Philosophy and Social Sciences Edition)*, 51(2):15–22, 2018.

[11] T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

[12] Giovanni Millo. Robust standard error estimators for panel models: A unifying approach. 2014.

[13] Judea Pearl. Direct and indirect effects. *arXiv preprint arXiv:1301.2300*, 2013.

[14] Michael Shindler, Alex Wong, and Adam Meyerson. Fast and accurate k-means for large datasets. In *nips*, volume 24, pages 2375–2383, 2011.

[15] Takashi Yamano. Lecture notes on advanced econometrics. *Lecture 9: Heteroskedasticity and robust estimators*, 2009.