Problem

D

Influence and similarity models in music networks

Summary

The relationship between music is like a complex network, intertwined with each other, weaving our life and art together, by exploring the changes in music, we can understand this wonderful art.

First, after addressing abnormal data in virtue of box plots, we establish an artist influence model through the typical sub data set of influencer texts. The model consists of three indicators. The index weight is calculated by the critic weight method to obtain the artist's musical influence. Based on the score of musical influence, a sub-graph of the influence direct network can be constructed through Gephi software to visually determine the manifestation of musical influence among artists. We use this model to calculate a series of data on the influence, such as average annual influence and personal influence.

In this modelling, in order to deal with the problems caused by high-dimensional data and calculate the similarity between individuals, the model we built combines two parts of knowledge, one is the algorithm required for data dimensionality reduction (Principal Component Analysis, Locality-Sensitivity Hashing), part of which is the algorithm (Cosine similarity, Euclidean distance) required to calculate similarity. We combine both parts of the algorithm to form our model algorithm PCA-CS, PCA-ED, LSH-ED, LSH-CS. Through these models, we can obtain the similarity of any two individuals (i.e., the similarity of the music of two artists). On this basis, we also used the stability test ADF test method and the time series ARMA, ARIMA, k-means clustering algorithms, and they assisted us to better solve some complex problems. These operations can be made in Python.

In addition, we also obtain some important similarity indicators artist_ratio (R) in music through data processing and mathematical calculations, which are used to evaluate whether an artist is more similar to artists in the genre, and the distance_judgement (dj), you can used to evaluate whether the music between two individuals is similar.

By using the model, we built, we solve the problems about music genres, artists, and music characteristics, and concluded that the music work of musicians in the same genre are more similar. The music of influencers can indeed affect the music of followers. And a series of questions about influence and music similarity.

Contents

1.Background

As we know, as a shared cultural heritage, music plays a significant role in human society and history. Music has many functions. For example, help people relieve stress, raise aesthetics accomplishment and exercise their thinking skills. In addition, music is beneficial to cross cultural exchanges and self-expression in communication, especially for individuals living in different races when language barriers existed in the old days.

Since its inception, music has been evolving with the development of the history so as to enrich it greatly. Music sometimes undergoes revolutionary changes, such as creating new genres, or making new sounds or rhythms. Sometimes these changes are due to one artist influencing another artist. Sometimes this is changed in response to external social factors, such as big international events or technological advances.

Therefore, if we can quantify some indicators of music, consider the musical characteristics of songs, and use this to capture the mutual influence between musicians. Then we can better understand the development of music in society in a macroscopic view over time.

2.Problem restatement

First, we need to filter the '*influence_data.csv* ' according to certain conditions, and visualize the filtered data in the direct network according to the musical influence. Then set up a music similarity measurement model, and use this model to explore the similarity of genres, the similarity of artists, and the similarity of music between influencers and followers.

After adding the indicator of time, discuss the changes of genres over time, the changes of music characteristics over time, and the changes of musical influence over time.

Finally, combining music characteristics, influence and other changes over time, find the "innovators" who can cause significant changes in music characteristics.

3.General Assumptions

1. **Artist's influence model (AIM):**

We believe that the 'Total number of people affected', 'The number of direct influences in the same genre (influence within the genre'), and 'The number of influences in the same era (the influence of the era)' are three objective factors for judging the influence of artists.

The active year of the influencer is later than the active year of the affected person, we remove it

2. Music similarity model (MSM):

Some genres with few artists, we eliminate them, and think it would not affect our other results Thinking that the infectiousness of music is closely related to its popularity.

Think that the musical characteristics of different genres in each era are the average of the musical characteristics of all artists of that genre in that era.

We divide the evolution of music over time into genre changes and music itself changes (genre does not change).

The number of people selected is 5000.

The notation table contains notations we use in this paper.

Symbol	Definition
X_{ij}	*The value of the j-th evaluation index of the i-th sample*
S_j	*Standard deviation of the j-th evaluation index*
r_{ij}	*Correlation coefficient between evaluation index i and j*
C_i	*The role of the i index in the entire evaluation system*
W_j	*The weight of the j index*
p	*The number of previous (or lagged) terms used within the model*
α_i	*Coefficient between 0 and 1*
w_t	*White noise term*
q	*The number of past error terms*
D	*The times of time series differenced*
dj	*The average distance between the genres*
$n\_clusters$	*The number of clusters we want to divide the data into*

Table 1: Notations

4.Data Wrangling and Exploring Analysis

There are four data sets in CSV format. After inspection, there are no missing data and format and content error data.

In order to view the outliers of the data, we can choose to sort the variables and compare statistics such as the maximum value or draw a scatter plot (visually showing the relationship between two variables is easy to detect graphically, but the disadvantage is that the essence is the relationship between the two variables) or draw a box plot. After considering the particular case of this problem, we finally choose to use descriptive analysis to do data exploration meanwhile draw box plots (see figure) to remove outliers. Elimination criteria: Calculate the minimum and maximum estimates of the data, data beyond this range may be an abnormal value. After weighing and analyzing, we choose to delete these items with outliers to ensure the precision of our data used for further analysis.

5.Model Construction

5.1 Critic Weight Method

(Deng Lijuan & Ma ailing, 2010) It is a comprehensive measure of the objective weight of indicators based on the contrast strength of evaluation indicators and the collision between indicators. Considering the variability of indicators while taking into account the correlation between indicators, the objective attributes of the data themselves are used for scientific evaluation.

Contrast strength refers to the size of the difference between the evaluation schemes of the same indicator, expressed in the form of standard deviation. The greater the standard deviation, the greater the fluctuation, that is, the greater the value gap between the various programs, the higher the weight.

The collision between the indicators is expressed by the correlation coefficient. If there is a strong positive correlation between the two indicators, the smaller the collision, the lower the weight.

$$\begin{cases} \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ S_j = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}{n-1}} \end{cases}$$

$$R_j = \sum_{i=1}^{p}\left(1-r_{ij}\right)$$

$$C_j = S_j \sum_{i=1}^{p}\left(1-r_{ij}\right) = S_j \times R_j$$

$$W_j = \frac{C_j}{\sum_{j=1}^{p} C_j}$$

5.2 Principal Component Analysis (PCA) to reduce dimensions

(Ringnér, 2008) This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

There are many ways to measure the similarities. (Qian, G., Sural, S., Gu, Y., & Pramanik, S., 2004)

5.2.1Cosine similarity

$$\cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2}\sqrt{\sum_{k=1}^{n} x_{2k}^2}}$$

5.2.2Euclidean distance

$$d(x,y) := \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

5.3 Locality-Sensitivity Hashing (LSH)

(Marçais, G., DeBlasio, D., Pandey, P., & Kingsford, C., 2019) For example, if we want to generate a 32-bit binary code, LSH randomly generate 32 projection vectors, $\{w_1, w_2, \ldots, w_{32}\}$, then each projection vector generates a binary code. The corresponding hash function can be: $b_k = \delta[w_k^\top x \geq 0]$. LSH initially was used in the inverted method to quickly search for the nearest neighbors and later was more used to generate random binary codes to approximate distances. We can use Hashing and quantization approximate nearest neighbor search method to deal with big data, thus simplifying distance calculation process.

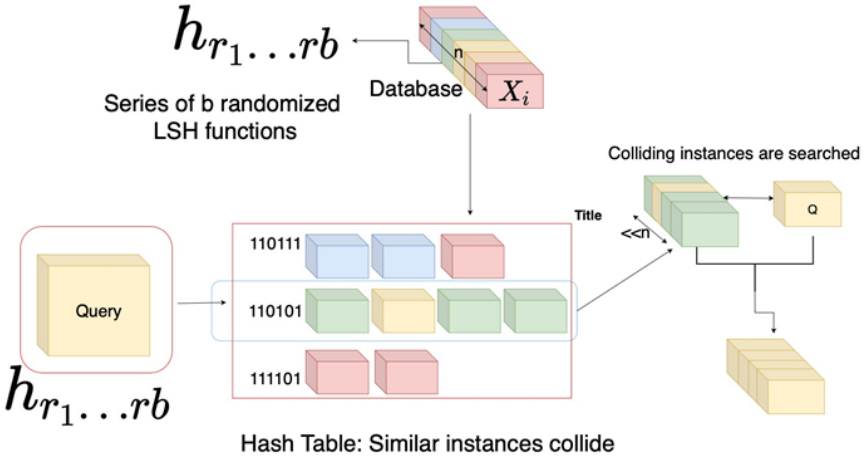


Figure 1

(Zheng, B., Zhao, X., Weng, L., Hung, N. Q. V., Liu, H., & Jensen, C. S., 2020) Points that are close in space have a higher probability of being assigned to the same bucket, while points that are far away have a high probability of being assigned to different buckets. In other words, for two points x and y, the probability that they are assigned to the same bucket by the hash function monotonically decreases as the distance increases.

For this complex music problem, due to the large amount and the high dimension of the data, when building the model, we choose to build four new algorithms by combining four classic and effective algorithms in pairs. In this paper, we call them PCA-CS, PCA-ED, LSH-CS, LSH-ED, and collectively referred to as Music similarity model. This model can effectively reduce the data dimensions, while avoiding dimensional disasters (data with too high dimension performs poorly on cosine similarity and euclidean metric algorithms).

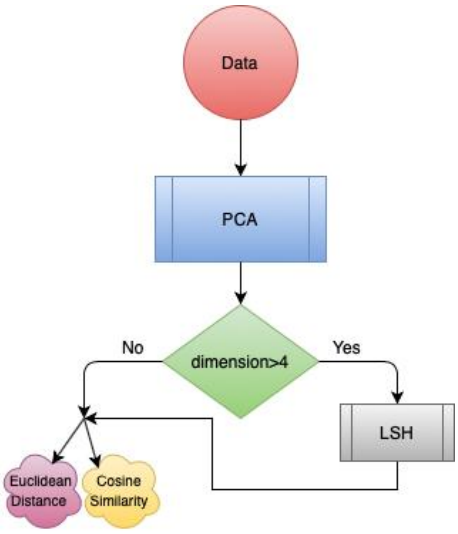


Figure 2

When addressing the problem, we choose to use one of these four algorithms according to the dimension reduction effect. We prefer to use the PCA algorithm for dimension reduction, because it retains high data accuracy (we require data accuracy to be more than 80%).

If the data after dimension reduction is found to be more than 4 dimensions, we will choose LSH to do dimension reduction, which can reduce the dimension to 3 or 4 dimensions. Then we use Cosine similarity or Euclidean distance to calculate the similarity between two objects.

5.4 ARMA

ARMA model is simply the merger between AR(p) and MA(q) models.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

5.5ARIMA

Differential autoregressive moving average model ARIMA(p,d,q), where d is the order of the data to be differentiated.

we combine the Autoregressive models and Moving Average models to produce more sophisticated models - Auto Regressive Moving Average (ARMA) and Auto Regressive Integrated Moving Average (ARIMA) models to do time series analysis.

6.Solve Problem

6.1 Problem 1:

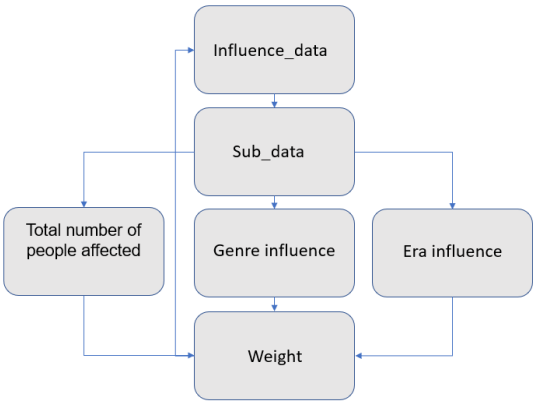


Figure 3

Indicator establishment part:
In this topic, we use the data in '*influence_data.csv*' to set measurement indicators, 'the total number of influencers', 'the number of direct influences in the same field (influence within the field)', and 'the number of influences in the same era (year Influence)'.

Data processing part:
Since the original data volume is too large and its performance in the directional network is poor, we randomly extract the original sample based on the influencer's age ratio to obtain a subset of the original sample. Then we analyze the sub-set through the indicators, and finally use the comprehensive evaluation model to determine the weight of the indicators.

Sampling:
We calculate the proportion of influence_data according to the year, and get the following year-to-person ratio. Assuming that the number of people selected is 5000, random sampling is performed according to the proportion of people in the year, and the following set of data (with map) can be obtained. Then the data is imported in Gephi, and we apply Fruchterman Reingold method is to layout. Finally, the nodes are sorted according to their indegree (the number of adjacent connecting lines), and the followers point to the influencers. By doing this, the direct network obtained.
Here we will not show the network's labels.

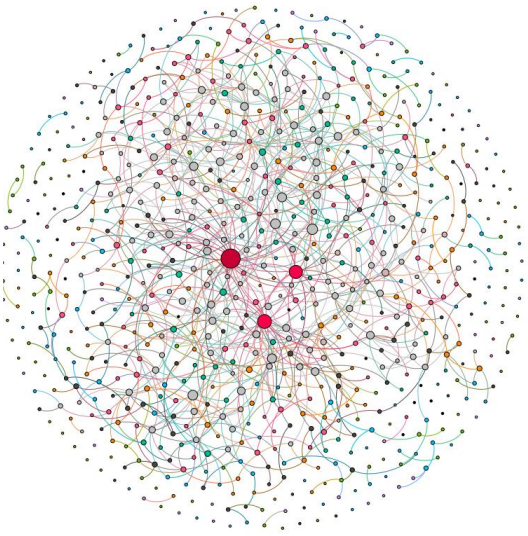


Figure 4

We sort the data chart of the network diagram by the indegree (the number of connected edges pointed by the arrow) to obtain the ranking table of the indegree. By comparing the indegree ranking table of the sub-data table with the indegree ranking table of the total data table, (Note: In the total data table, the average connection degree is 7.68; the maximum connection degree is 614; so only the top 100 rankings are used here) we find the accuracy of the sub_data table is 87.03% (the accuracy here means that the rate of top 100 people of the sub-data are also in the top 100 of the total data table is 87.03%), so it can be regarded as a typical sub-data of the total data set.

Perform the following index analysis on the subset:
1.Total number of people affected
2.Genre influence: rank of influencing people (compared with people in the same field)
3.Era influence: rank of influencing number (compared with people in the same age)

Table 2

name	ratio_year	ratio_genre	indegree
The Beatles	4.75543478	2.499107462	70
Bob Dylan	2.7173913	1.428061407	40
The Rolling Stones	2.64945652	1.392359871	39
Led Zeppelin	2.03804348	1.071046055	30
The Kinks	1.69836957	0.892538379	25
Jimi Hendrix	1.69836957	0.892538379	25
David Bowie	1.63043478	0.856836844	24
Marvin Gaye	3.0848329	3.744149766	24
Miles Davis	6.64739884	7.142857143	23

By applying the Critic weighting method, the artist influence model is obtained. Through this model, the weights of 'the total number of people affected', 'age influence', and, 'domain influence' can be obtained.

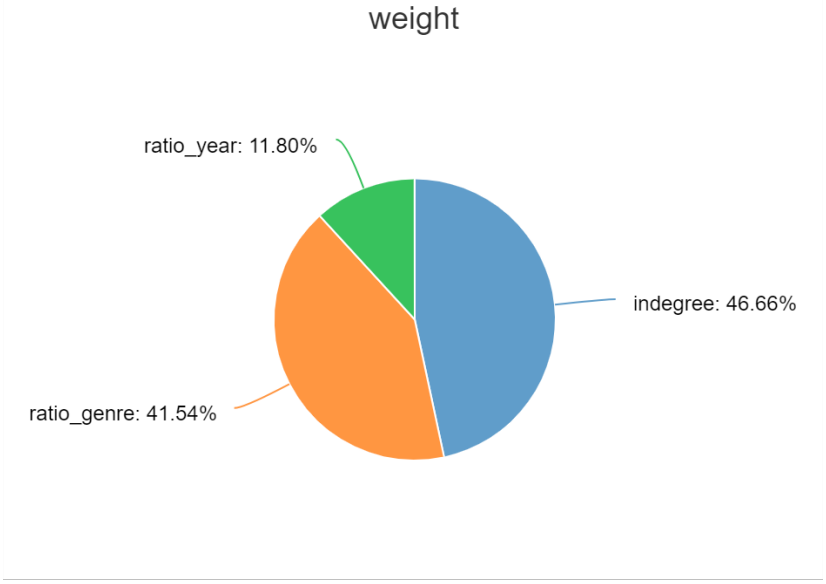


Figure 5

Taking musicians 'The Beatles', 'Bob Dylan', 'The Rolling Stones' as an example, the influence scores of them can be calculated by weights as, In the subset, by transforming the node index into their musical influence, the following figure is finally obtained, which can intuitively reflect the influence of music in the directional network.

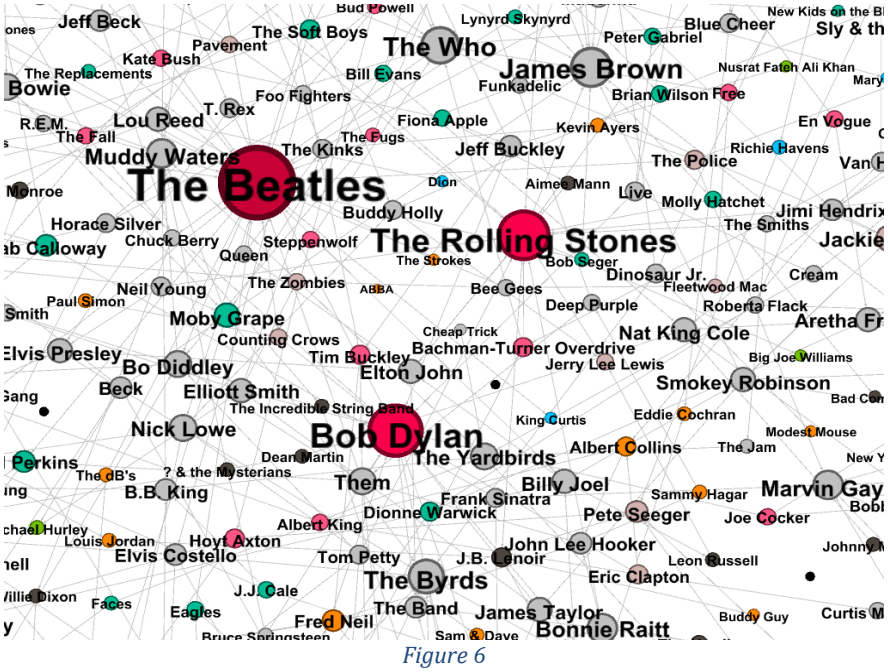


Figure 6

6.2 Problem 2:

We want to judge whether the artists within genre are more similar than artists between genres. Since the data_by_artist.csv data set can better reflect the differences between different artists compared to other data sets. We select this data set for analysis. In this problem, we use PCA dimension to reduce data and K-Means cluster analysis to evaluate the similarity of musicians' overall work. The following briefly explains the principle of the system and its process.

Since a musician has two features, music characteristics and vocal type, we hope to analyze from these two perspectives to see if musicians can be classified into the same category in both aspects. First, we use the PCA method to reduce the dimension of the data, which reduces the dimension of the music feature to 4 dimensions, the dimension of the vocal type into 3 dimensions and record them as fea_1, fea_2 respectively. Therefore, we can use k-means clustering analysis with confidence. According to calinski_harabasz_score, we find the number of clustering categories (n_clusters). fea_1 is clustered into two categories, and fea_2 is clustered into 4 clusters. We can find that the number of people is not much different in different categories, indicating that the clustering effect is relatively good.

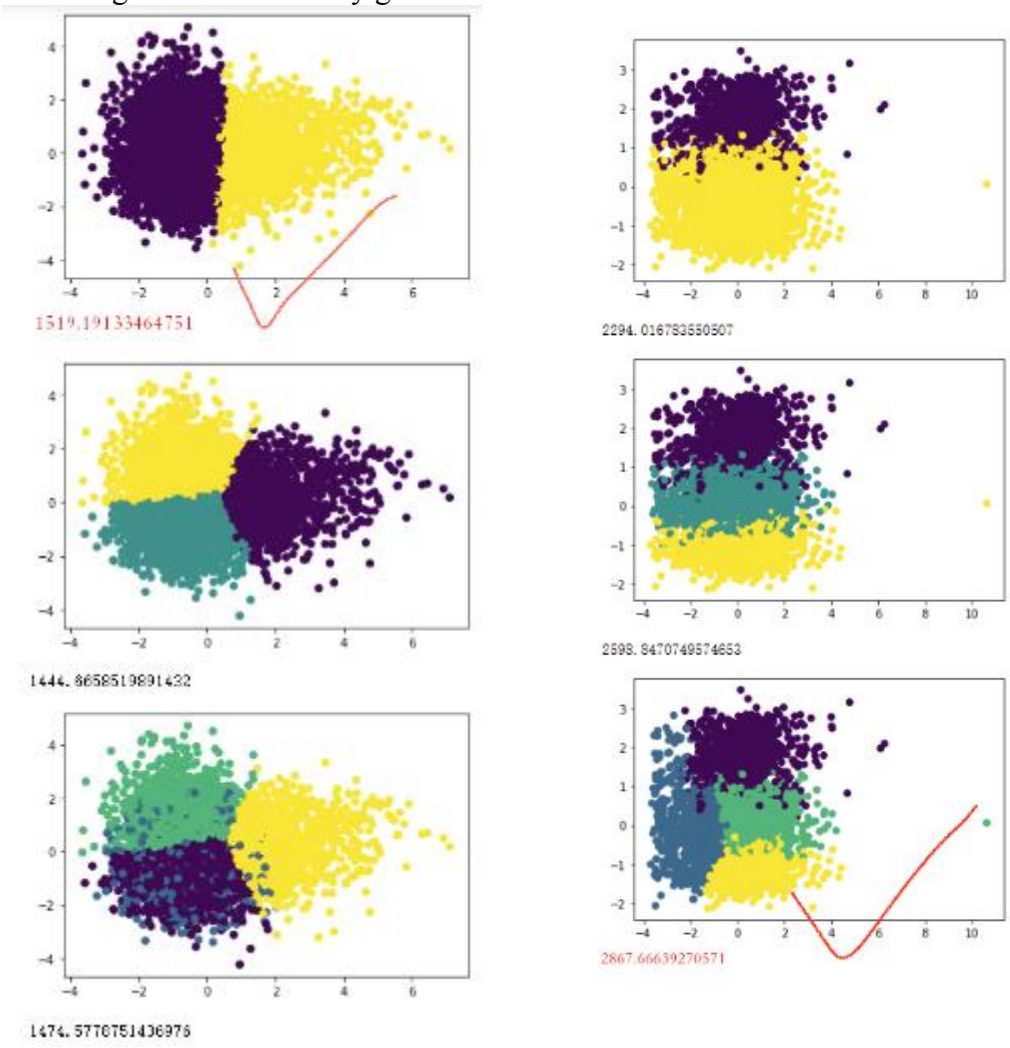


Figure 7

According to the completed clustering, we need to determine whether most artists of the same genre are now classified within the same category. Although the data set of data_by_artist does not indicate the genre of each artist, we can use the data set of influence_data to classify them correspond to each artist. Next, by traversing the number of artists of each corresponding genre in each cluster, we calculate the number of artists of each genre in different clusters and sort them. Now, we can get the ratio of the largest number to the total number of artists of the corresponding genre, and define it as s_rate. Finally, we can evaluate the similarity between artists. The corresponding s_rate is obtained by analyzing fea_1 and fea_2 respectively.

	Pop/Rock	Country	Classical	Electronic	Comedy/Spoken	Easy Listening	R&B;	Stage & Screen
fea_1 s_rate	0.762731	0.52381	1	0.746032	0.636363636	0.8125	0.741379	1
fea_2 s_rate	0.459429	0.472527	0.857143	0.301587	0.545454545	0.75	0.390086	0.551724138

	Reggae	Blues	New Age	Latin	Vocal	Jazz	International	Folk	Religious
fea_1 s_rate	0.83636364	0.50588235	1	0.76271186	0.9609375	0.84027778	0.60869565	0.8625	0.50980392
fea_2 s_rate	0.43636364	0.30588235	0.84210526	0.27966102	0.3125	0.58680556	0.32608696	0.5	0.43137255

Figure 8

We set that if artists of a genre account for more than the ratio R in the same cluster, we believe that the characteristics of the music works of artists in this genre are similar. According to the two features， we make separate judgments and we can conclude that most of the genres have similar features in the work of their artists. This is more in line with our normal judgment. The output work of artists in the same genre are usually relatively similar.

	similar	not similar
fea_1	13	4
fea_2	12	5

Figure 9

$$R = \frac{1}{n_{clusters}} + 0.1$$

It should be noted that due to the small number of people in some genres (i.e. 'Unknown', 'Children's','Avant-Grade'), we cannot judge whether the works of musicians of these genres are similar, so we can only choose to delete these three genres when judging.

6.3 Problem 3:

From the perspective of influence, we extract the three people with the highest influence, as shown in the following table. We divide the influence_data data set of the first three people in the table into two parts: the number of influencers in the same field and the number of influencers in different fields. Then we find that the number of influential people in the same field is far greater than that in different fields, as shown in the figure below.

name	same_genre	diff_genre
The Beatles	554	61
Bob Dylan	322	67
The Rolling stone	304	15

Table 3

In order to explore the similarities between genres, we choose to put the music characteristics and Vocal characteristics data of any two genres into our model to calculate the similarity between them.

First, we count the active years and genres of each artist and average the music characteristics of artists in the same genre and the same time period. Then we get the music characteristics of a genre in a specific year.

	genre	danceability	energy	valence	tempo	loudness	mode	key
0	R&B;	0.634267	0.561269	0.615420	116.730458	-9.341882	0.713860	5.658718
1	Pop/Rock	0.514568	0.680246	0.527397	124.326963	-8.592226	0.861882	5.556437
2	Religious	0.502250	0.561040	0.439797	117.043728	-9.070418	0.931034	5.344828
3	Blues	0.578501	0.463884	0.656524	120.046387	-11.758124	0.840000	5.390000
4	International	0.554290	0.461834	0.599220	115.038376	-12.177086	0.797468	5.481013
5	Country	0.578639	0.527955	0.610636	122.187117	-9.984130	0.987531	5.870324
6	Vocal	0.468220	0.294371	0.439201	112.091304	-13.353066	0.885350	5.178344
7	Jazz	0.529662	0.380152	0.511345	112.873902	-14.196530	0.683544	4.969620
8	New Age	0.381969	0.250221	0.244910	109.670616	-17.684459	0.694444	3.500000
9	Comedy/Spoken	0.571761	0.553670	0.581761	120.333646	-11.729833	0.777778	6.055556
10	Reggae	0.741573	0.562573	0.738345	116.422637	-9.416223	0.696296	6.155556
11	Latin	0.613488	0.581232	0.685083	117.697726	-9.260692	0.762115	5.396476
12	Folk	0.521888	0.305619	0.499050	120.238366	-14.160246	0.923913	5.467391
13	Stage & Screen	0.329756	0.281560	0.250811	105.899065	-15.951223	0.760000	4.300000
14	Classical	0.326512	0.203044	0.223425	106.414706	-19.581925	0.851852	5.259259
15	Easy Listening	0.445687	0.373131	0.404793	112.448004	-13.784661	0.913043	4.956522
16	Electronic	0.629574	0.672875	0.486272	120.182088	-9.174948	0.630542	6.157635

Figure 10

Next, we put it into our model and successfully calculated the similarity between every two genres. Here, due to the high dimension, the PCA reduction effect is poor, so we choose the LSH algorithm for dimension reduction and calculate their cosine similarity.

We believe that high cosine similarity indicates high similarity (See heat map on the right).
In addition, by using the LHS-ED and calculating the average value of the distance between the genres, we get an index called distance_judgement (referred to as dj) which equals to 2.1372. We think this indicator can be used to judge whether two types of music are relatively similar (i.e. if the Euclidean distance between two artists exceeds this indicator, we can think that their music works are not similar)

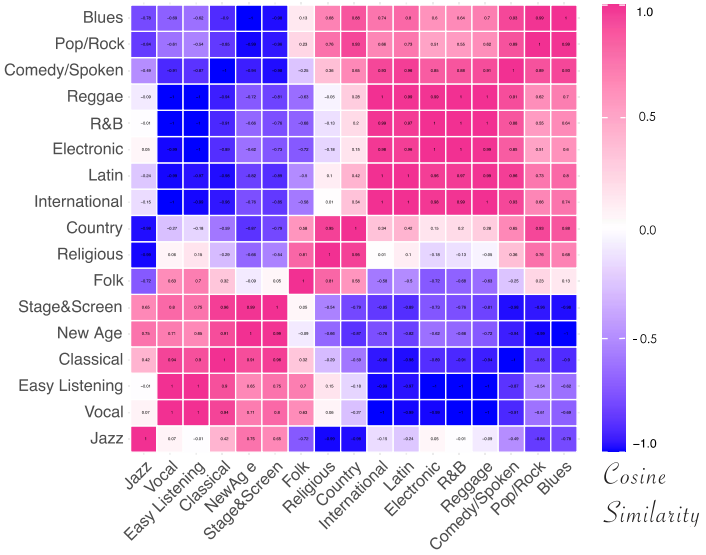


Figure 11

If we want to explore how a genre changes, we have to consider how the internal characteristics of a genre music change. Here we use the linear regression method to perform linear regression on each music feature, and get the result by observing the image.
After drawing each genre, we choose 'Pop/Rock' and ‘Vocal' as typical examples to answer this question. The reasons for choosing these two genres are as follows:

1. As larger music genres, these two genres have more musical works in each time period, which can better reflect the changes in each music characteristic.
2. The similarity between these two genres is very low (the square between them are blue on the heat map), so we can observe the changing music characteristics of the two genres with a large difference (to avoid ignoring many things if we only consider one genre)
3. Through observation, the images of these two genres have a better fit and can better reflect the changes in our music characteristics

We select images with obvious trends in music characteristics in the genre, and analyzed their changing feature. In two genres, we all select energy and valence for analysis.

Energy reflects the activity level of a song. The higher the value, the faster and louder the song is. We find that as time changes, the energy value of Pop/Rock is getting higher and higher, and the overall trend is very obvious. Although the energy value of Vocal has a relatively weak upward trend from a straight line, we can find that in fact, after the 1960s, the energy value of most eras is low. (Halevi-Katz, D., Yaakobi, E., & Putter-Katz, H., 2015)It does not pursue the intensity of music (It can be seen that the maximum value of Vocal is less than the minimum value of Pop/Rock)

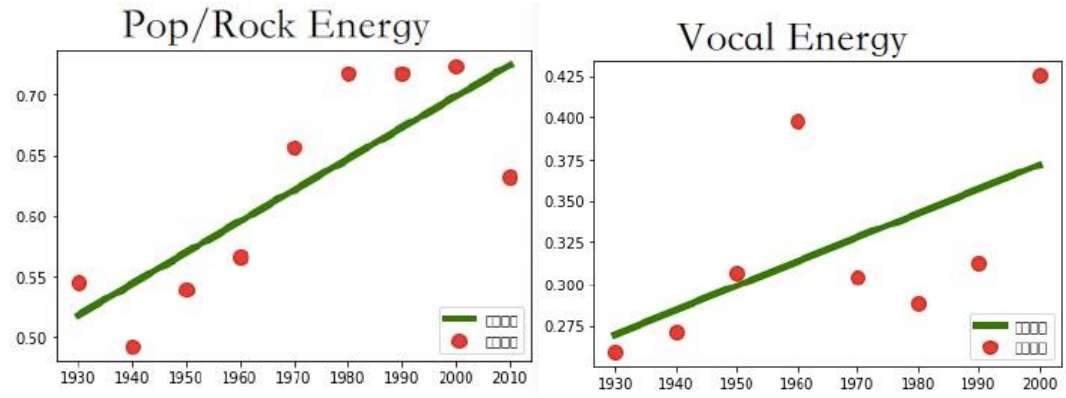


Figure 12

Valence expresses the positive degree of a song. The higher the value, the more positive the emotion of the song (i.e. happy). From these two pictures we can see that these two genre's Valence values are gradually declining, and the annual value of them is not much different. Our explanation is that as blacks gradually emerged in the music world (i.e. Chuck Berry), more and more black music works were spread, and the features of black music styles are mostly expressing dissatisfaction with society and racial discrimination. In addition, with the continuous changes of the times, many American singers have begun to describe some consumerism in their lyrics. These are all the negative aspects of society. Therefore, whether it is Pop/Rock music or Vocal music, the value of Valence is constantly declining.

6.4 Problem 4:

We want to judge whether influencers' music is more likely to affect the music of followers. By applying the influence model, we can find that The Beatles is one of the most influential artists, and the weak influencers is impossible to accurately determine whether the two are more similar because of the relatively small amount of data. Therefore, we select The Beatles as the analysis object. What we need to determine is whether the music characteristics of The Beatles' followers is more similar to the music characteristics of The Beatles. Since what we need to judge is whether the two musicians are similar, in this question, in addition to considering the characteristics of the music, we should also consider the vocal type of each artist.

We first select objects. Obviously, the followers of The Beatles need to be analyzed as one object. In addition, we also randomly select the same number of non-followers as their followers. We find that the data has 11 dimensions. When we use PCA to reduce dimension, the effect of dimension reduction is not ideal (we need to retain 7-dimensional data to ensure more than 80% of the data authenticity), so here, we apply the LSH-ED model to process the data and find that LSH can reduce the data to 3 dimensions. The similarity (Euclidean distance) between The Beatles and each musician is calculated reasonably in 3-D space. We find the two values by averaging the similarity between The Beatles and its followers and the similarity between The Beatles and its non-followers.

The distance between The Beatles and its followers	The distance between The Beatles and the other people
1.67	2.93

Figure 13

We find that although the Euclidean Distance of the people affected by The Beatles is relatively short, it is not very obvious. We speculate that it is because The Beatles is one of the most famous bands in the world. Despite of some people are not their followers, those people are followers of the people influenced by The Beatles. For example, the followers of The Beatles have an influential artist called Funkadelic, and he is an artist in the R&B category. He has influenced many R&B artists, so it will indirectly lead to other non-followers whose music is similar to The Beatles.

Besides, there may be another reason that although LSH can reduce the dimension better, due to the features of its algorithm, it will cause a loss of precision.

6.5 Problem 5:

In order to find musical evolution, we select the data_by_year.csv data set. Compare to other data sets, it clearly displays the time dimension, so that we can capture and analyze the changes over time, including changes in music characteristics and vocal characteristics. We decide to fit the time series with the data of music features and vocal features, and use image changes to determine when the revolution has occurred.

Since we need to observe the changes in the image, if we take too many pictures of all the music features, it will affect our observation, so we still use the PCA method to reduce the dimensionality of the music features and vocal features, and then record them as t1 and t2. We guess that because the amount of data is smaller and more continuous than data_by_artist.csv, the dimension reduction effect of t1 is better than the previous fea_1.

In order to make the image more accurately, we first performed ADF_test on the data to determine which data is more suitable for the ARMA model fitting. We find that among the 6 data, only the p-value corresponding to the third-dimension data of t1 and t2 are less than 0.05, which proves that only them are relatively stable and the effect of using the ARMA model for fitting is better, so we have made a difference on the other columns of data, which is the ARIMA fitting. Whether in the ARMA or ARIMA fitting, we determine the p, d, and q values by judging the AIC value. After confirming, we can draw the image. The image contains five images: 'Time Series Analysis Plot', 'Autocorrelation', 'Partial Autocorrelation', 'QQ Plot', and 'Probability Plot'.

According to the two images of 'QQ Plot' and 'Probability Plot', we find that some pictures which are fitting well of these six pictures. In these pictures, we can confidently observe the changes in the period of time based on the 'Time Series Analysis Plot' and use it to find the time when the revolution occurred. Through observation, we can clearly notice that the fluctuations in the 1920s and 1960s were very large.

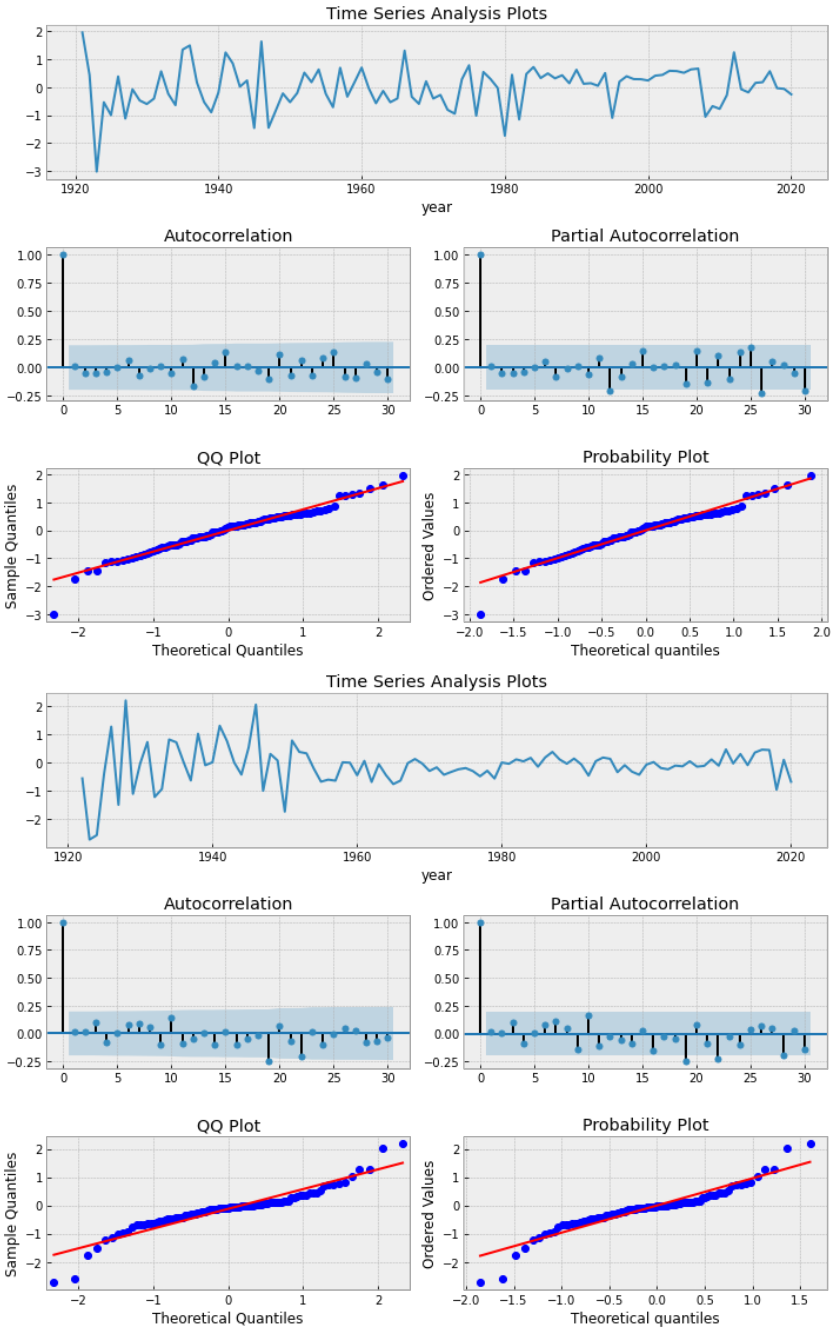


Figure 14

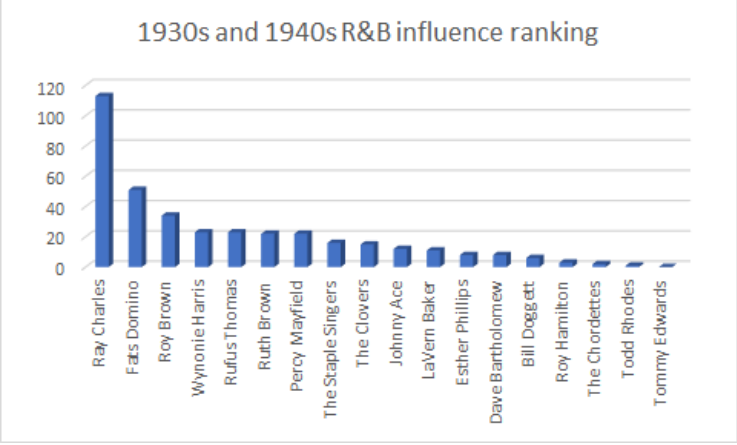
Next, we use simple linear regression to analyze the trend and use the weight coefficient to measure the strength of the tendency and understand the long-term comprehensive influence of a factor. After removing the factors of stable trend, combined with the previous time series chart, 20-50 volatility is large. We concentrate on the changes between 1920 and 1950. In view of the years we noted that there are changes, analyze the changes in musical characteristics in the years around the turning years, and find one or several characteristics that have changed the most. The key trend is stable, and we will discard it if it is not a fundamental factor affecting change. Danceability, energy, tempo, and loudness all show upward trends over time, while valence, acousticness,

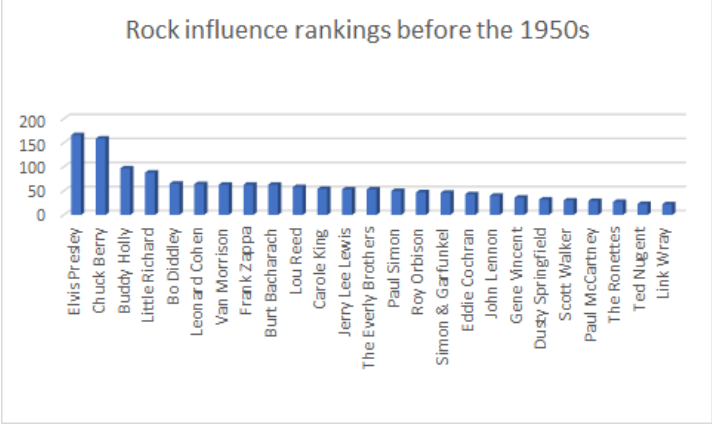
instrumentalness, livenss, and speechiness show downward trends. Combining the weight coefficient k and focusing on the distribution of the fitted trend line of the scattered points in the 1920s and 1950s, the analysis shows that the fluctuations of the danceability, valence, and instrumentalness on both sides of the line are large. So, these 3 features signify revolutions.

We divide the evolution of music over time into genre changes and music itself changes (genre does not change).

For the 20s to 50s, the fluctuations between the 20s and 50s are relatively large, which means that there is a big change in music characteristics in the 20s to 50s. For changes in this characteristic, the first thing that comes to mind is the more popular music genres in the years Because of the previous analysis, the similarity between certain genres is relatively low. Therefore, our primary goal is to identify the genre changes in music between the 20s and 50s, and get the most For popular music genres, take the 1950s as an example. If Rock was popular in the 1950s, first find the person who had the most influence on the Rock before the 1950s. It can be stated that this person played an important role in the change of this genre. You can call him a very prominent revolutionary (genre change revolutionary).

Jazz in the 1930s, R&B in the 1940s, Rock in the 50s





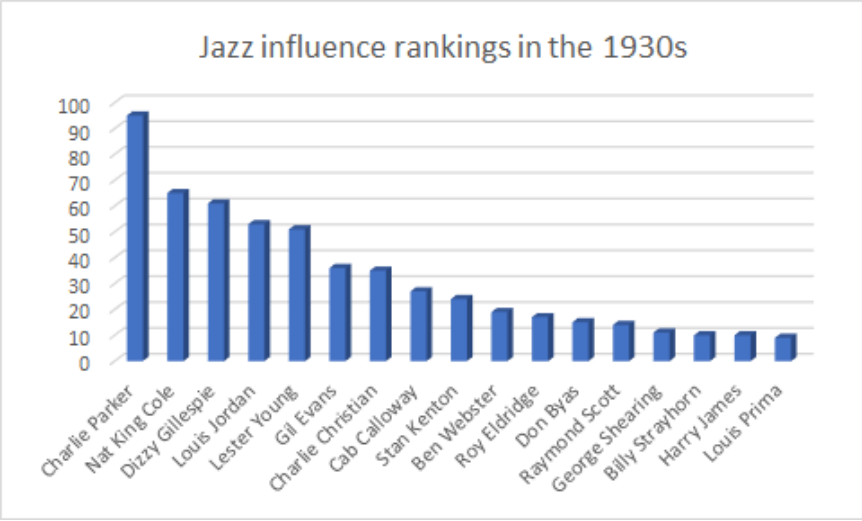


Figure 15

Through the data chart, you can get the ranking of Jazz, Rock, R&B influence, and extract the number one singer among R&B, Pop/Rock and Jazz.
R&B: Ray Charles: Ray Charles pioneered rhythm and blues music and was one of the first figures to be included in the Rock and Roll Hall of Fame.

Rock: Elvis Presley, he is regarded as one of the most significant cultural icons of the 20th century and is often referred to as the "King of Rock and Roll" or simply "the King".
Jazz： "Charlie" Parker Jr. (August 29, 1920 – March 12, 1955), nicknamed "Bird" and "Yardbird", was an American jazz saxophonist and composer.

Then there is a changed in the music itself. We think the judgment indicator is the influence in the same genre. If the impact in this genre is strong, then it will affect the creation of his followers in this field. If he has the greatest influence in music, then we have reason to suspect that he is a revolutionary in music.

In the network diagram, change the indicators of the directional network diagram to influence within the genre to visually see these revolutionaries.

6.6 Problem 6:

By applying the artist's influence model, we quantify the value of influence to each year, and draw a line chart of influence over time. This picture shows the overall influence and individual pop/rock influence. From this chart, we can observe the overall and pop/rock artists' influence values and their changes in each era.

After comparison, we find that the value of influence from the 1930s to the 1960s was relatively high, especially in the 1960s. After the above time series analysis, we also know that music has changed a lot from the 1920s to the 1960s, which shows that our analysis results can be mutually confirmed and the reliability is high.

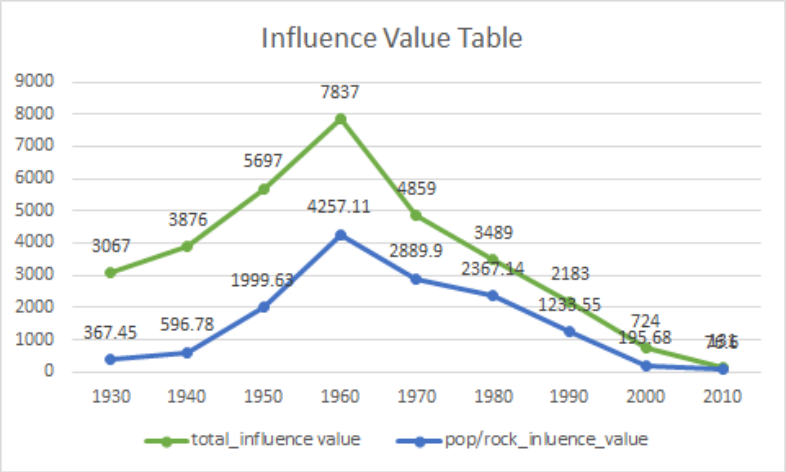


Figure 16

When we established the artist's influence model, we have already carried out a weight analysis of its influence indicators through the critic weight method. From the image, we can clearly see that the two indicators of indegree and ratio_genre can reveal whether the influencer is dynamic.

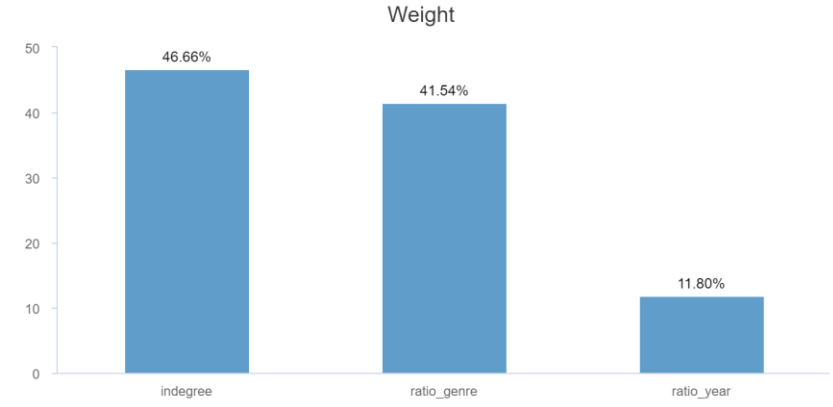


Figure 17

6.7 Problem 7:

Based on the Time Series Plot in problem 5 and basic historical knowledge, the great fluctuation between 1920s to 1950s corresponds to the emerge of blues actually. Blues singers were descendants of slaves and elements of their music reach back to African origins. The simple but expressive forms of the blues became by the 1960s one of the most important influences on the development of popular music throughout the United States.

(Zhou, 2015) The blues belongs to the black American ethnic group, and it is also a representation of the cultural characteristics of this ethnic group. The various musical characteristics of blues music are actually epitome of the past experience of American blacks, or in other words, epitome of the past experience of the human race: suffering, desolation, and helplessness. Blues music has come along with the African American ethnic group, and its style and subject matter selection has gradually changed from overall depression and melancholy to something joyful and pleasant, so that eventually blues music has penetrated into all aspects of popular and classical music. Among

them, it broke through its own cultural barriers and merged and sublimated with the cultures of other ethnic groups. But, even so, the unique imprint of blues music cannot be completely faded, just like b3 and b7 in the blues scale which are always reminding people not to forget the suffering years that the black American ethnic group has suffered. Blues' influence is not only limited to the category of popular music, but also penetrates into the field of classical music. For example, many classic blues music is adapted and processed in a symphony orchestra or the phenomenon of well-known blues singers and symphony orchestras has become the norm. Its influence is still spreading in the field of music.

The third technological revolution happens after the World War II and induces a musical revolution which can be revealed from our picture. The development of artificial intelligence and microelectronics technology, and the emergence of more new instruments, ranging from musical instruments to tools for recording music. Electronic becomes common. Since this, human beings have gradually entered the era of digital music. Pop was the mainstream in the 1920s. The American economic and political crisis in the 1930s and the development of the labor movement aroused American people's pursuit of freedom in the spiritual world of music. Rock emerged in the 1960s. (Inglis, 2017) A group of artists led by Bob Dylan devoted themselves to the revival of folk songs. In the 1970s, heavy metal bands set off an upsurge. At the same time, avant-garde rock and gorgeous rock entered the golden age, and New York punk and British punk emerged. In the 1980s, hip-hop, hard rock, and synthesizers became popular. All above are consistent with what we find about the processes of musical evolution.

7.Model Evaluation

7.1 Sensitivity Analysis

Our model allows us to change R (mentioned in the solution to Problem 2), which is a parameter used to compare the proportion of artists from a specific genre in the same category. We test the sensitivity of our model by changing the value of the parameter R to display reliability. The benchmark R value we selected earlier is 0.6, which divides the 17 genres into two categories (more similar within genres/with other genres). After calculation, the variations of the similarity

index within genre for characteristics and vocal types are 3.7307692 and 5.74358974. Therefore, our model is not sensitive to the value of the constant R set by us.

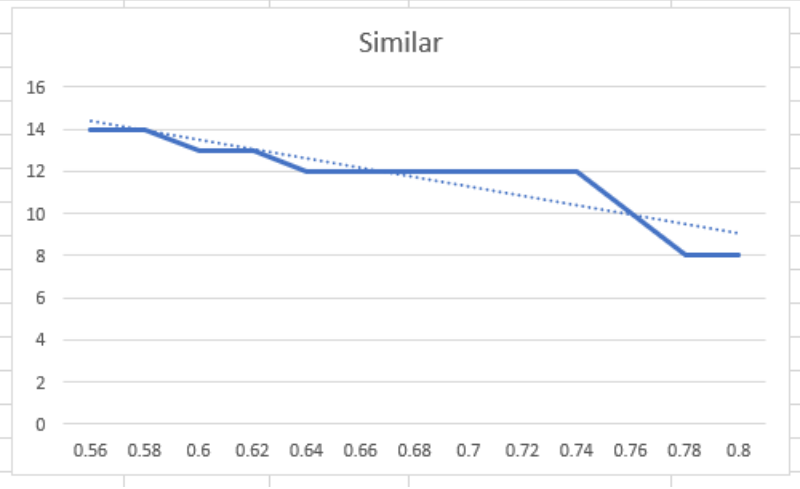


Figure 18

7.2 Strengths and Weaknesses

Strengths:

- By considering missing values and outliers, we reasonably clear the data.
- The use of LSH accelerates the calculation of similarity, and more ideally reduces high-dimensional data to low-dimensional data, which facilitates the calculation of Euclidean distance or cosine similarity
- For different dimensions, our music similarity model can choose a appropriate algorithm to calculate
- The critic weight method can help us accurately calculate the weight of the objective indicator of influence

Weaknesses:

- The influence model has certain limitations. The model only considers three more important indicators, but it does not necessarily mean that other indicators have no impact
- Whether it is PCA or LSH algorithm, in order to reduce the dimensionality, the data has sacrificed some accuracy, especially LSH
- Although our second question inferred that most artists of the genre are closer to the artists in this genre, due to the limitations of clustering, we can't explain well that artists of a certain genre are not similar to artists of other genres (i.e. in one category, although artists of genre A gather together, artists of genre B also gather in this category.)

7.3 Conclusions

This paper manages to develop a fully data-based model to build a network of music where it is feasible to figure out the similarities and influences between genres, artists, and explore the factors behind that influence the revolution.

For problem 1, we use the given dataset influence_data and manage to establish a model based on three indexes to do impact analysis. We select a subset of the dataset and visualize the results with a network graph.

For problem 2, we put forward an approach to deal with high dimension data and sum up an efficient method to compare similarities and influences between any two objects. In this paper,

we use this method to compare similarities and influences between and within genres, and also between artists. This method can be applied to compare genres and artists as well. Supplemented by clustering model, we set a parameter R to help us know whether an artist is more similar within genre or closer with styles of other genres. Alternatively, we can know whether artists within genre are more similar than artists between genres.
For problem 3, we put processed feature data of genres into the PCA-CS/PCA-model created by us and visualize the similarities between any two genres in a heat map. We come to the conclusion that influence within genres is larger by judging the similarity distance between two objects. The parameter distance_judgement(dj) is used to distinguish a genre. By researching on specific characteristics of a genre, we can know the changes of genres over time. Focusing on Pop/Rock and Vocal, we find the two genres have inconsistent trends in energy but share an descending trend when it comes to valence.

For problem 4, The Beatles has higher similarity rate with its followers than with non-followers, but the difference is not obvious. The similarity of the two is distributed near the distance index defined in the previous step. By judging the correlation between each music feature and the popularity, we can know which feature is the main feature. The time series tells us the great fluctuation from 1920s to 1960s. Among all these music features, the two characteristics danceability and energy have higher correlation with popularity, in other words, they are more "contagious".

For problem 5, the time series shows that the value of average music features fluctuates greatly between 1920s and 1960s. We find key characteristics are danceability and valence.
For problem 6, both the overall influence and the influence of Pop/Rock continue to increase. The weight analysis of the first question model (indegree, ratio genre) can help reveal whether an artist is a dynamic influencer.
For the last problem, we explain the fluctuation with the emerge of blues and the increasing popularity related genres like rock and jazz.

8.One-page document to the ICM Society

Dear Sir or Madam,

Our team proposed two models related to the music which amainly involves two aspects of influence and similarity respectively. They have the following significant values in the field of music and the other fields.

Firstly, it can be used in the evaluation system. There are various music award ceremonies in the current society(i.e.Grammy Award), so we need to know how to accurately evaluate one artist. If we can get some indicators what we need according to some of his status in the society, or his social review, we can put them into our artist influence model and judge his influence, which can be a basis for awards.

Secondly, it can be applied to the music recommendation system. When a user chooses a favorite music genre, We can make a prioritized recommendation based on the similarity between the music characteristics of each song and the music characteristics of the music genre. Even when a user listens to some songs for a long time and accumulates certain amount of data, we can accurately and intelligently recommend some songs that he likes with high probability according to the music characteristics of all the music he listens to.

In addition to art, in the market area, companies can count the characteristics of people who often buy their products, and use our similarity model to determine who can become their target customers. In turn, customers can easily find their favorite products based on their buying habits. This helps the company make a profit and save customers time spent on selecting products.

What's more, by processing the specific data, the influence of some social events can also be quantified by our influence model. After enough similar incidents have occurred, when some incidents occur again and these incidents are judged as similar incidents through the similarity model, we even can predict the subsequent impact of the incident. This can help the government or individuals better deal with problems.

When we consider more genres and more artists, the following changes may occur as a result.

Firstly, the influx of data from more genres and artists will have some impact on the weight indicators in the influence model, but it is still possible to quantify the impact of each person accurately.

When we calculate our dj index, its value may change to a certain extent, thereby changing our judgment of whether two individuals are similar.

More genres may cause the data to become more high-dimensional. When we judge whether artists within genre are more similar than artists between genres, the effect of clustering and classification may become worse, which affects the judgment and causes the sensitivity of R to increase.

As the amount of data and the dimension becomes higher, when we do dimension reduction (PCA or LSH), the accuracy of the data will be impaired, which results in bias in judgment.

In the future, if the model needs process more data or we use it to solve other problems, the model should establish a more appropriate dimension reduction method (i.e. 7 music characteristics are sorted into low-dimensional data according to their weights), rather than directly applying the PCA or LSH algorithm to reduce the dimension, which can ensure higher accuracy of the data. Of course, the solution of conventional problems can also be more accurate. Similarly, if we successfully achieve it, when analysing the changes in genre music characteristics, we don't need to pay attention to the time series of each characteristic and we can get its changing trend more simply and intuitively.

Besides, paying more attention to smaller music genres instead of simply thinking that big genres can represent most of the characteristics of the times. More and more accurate measurement methods can help artists learn more precisely the music they want. It can also help ordinary listeners better understand music (i.e. the connections and differences between two songs) Therefore, it can promote the development of the entire music industry.

Yours,
ICM team

9.References

1. Zheng, B., Zhao, X., Weng, L., Hung, N. Q. V., Liu, H., & Jensen, C. S. (2020). PM-LSH: A fast and accurate LSH framework for high-dimensional approximate NN search. Proceedings of the VLDB Endowment, 13(5), 643-655.
2. Marçais, G., DeBlasio, D., Pandey, P., & Kingsford, C. (2019). Locality-sensitive hashing for the edit distance. Bioinformatics, 35(14), i127-i135.

3. Inglis, I. (Ed.). (2017). Performance and popular music: history, place and time. Routledge.
4. Zhou, X. (2015). Cultural Representation and Cultural Studies. Shanghai People Press.
5. Halevi-Katz, D., Yaakobi, E., & Putter-Katz, H. (2015). Exposure to music and noise-induced hearing loss (NIHL) among professional pop/rock/jazz musicians. Noise & Health, 17(76), 158-164.
6. Ringnér, M. (2008). What is principal component analysis*?* Nature biotechnology*,* 26*(*3), 303-304.
7. Qian, G., Sural, S., Gu, Y., & Pramanik, S. (2004, March). Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In Proceedings of the 2004 ACM symposium on Applied computing (pp. 1232-1237).
8. Deng Lijuan & Ma ailing. (2010). Water saving irrigation scheme optimization based on critical weight and TOPSIS model. Water science and Engineering Technology (02), 10-12 doi:10.19733/j.cnki.1672-9900.2010.02.005.