STUDY OF HOW GENRES INFLUENCES EACH OTHER IN LISTENING CHOICE:

THE SPOTIFY CASE

TABLE OF CONTENTS

1. INTRODUCTION AND RESEARCH QUESTION

Is it possible to know how people choose the music they want to listen? This choice is influenced by several variables, most of them are difficult or impossible to measure. However, looking at the listeners behavior it is possible to detect if there are some elements in their choices thar influence more or less the choice of another elements. To carry on such analysis data are fundamental. In the past this information would have been difficult to collect, but nowadays thanks to music streaming platforms are every day more available and accurate.

According with the 2023 report “Music Streaming worldwide” of Statista, the number of subscribers to music streaming services has grown from 304 million in 2019 to 713 million in 2023 (+134%) (in revenues, from 1.3 billion to 19.3 billion) (Statista, 2023). In particular in this sector on of the biggest company is Spotify. According with Statista’s data in Q3 of 2023 it has been ranked as the first music streaming platform for the number of subscribers, 223 million (Apple Music has just 89.9 million) (Statista, 2023).

Going more in detail, according with the 2024 Spotify report from Statista, in the Q4 of 2023 the monthly average of active users was 602 million (Statista, 2024), meaning that the amount of data it can collect is enormous.

In addition to it, Spotify provides some useful APIs that permit access to user’s information about behaviors. The combination of all these elements makes this provider the perfect choice for our research.

We have said that of course not all the variables that make a user decide which song or genre it prefers to listen are available. However, from Spotify what we can get are the information about artists, genres and in general songs that users like to listen to more.

In our project we decided to focus on the genres area. And in particular on how they are connected to each other: can we detect some genres that are more likely to influence the choice of a user in the selection of other genres?

The best way of doing it is through a Network Analysis, a technique that permits to study how the nodes of a network interact with each other.

As it will be possible to see in the next pages, in order to set it we use as nodes the artists extracted from a list of 36 genres decided by the researchers from the beginning and connect to each other through a feature that Spotify provides, the so called “related

artists". These features connect each artist to other ones according to the users' listening behaviors.

What we expected is that yes there is a correlation between different genres, and we expected that this correlation is stronger for genres that are quite popular (the popularity is calculated using a Spotify variable that takes in consideration not only the number of listeners of a particular artist and the number of its followers but also the listeners behaviors).

Taking in consideration all these elements we set the analysis as follows: a first part of data gathering and variable exploration, using python and R. And a second part in which the network analysis is carried out using R.

In the following paragraphs, we are going to each step of the research from the data gathering part using python to setting the network analysis on R.

2. DATA COLLECTION AND ANALYSIS

The first part of the project consists of the data collection and analysis. In this section the dataset is constructed, and a descriptive and exploratory analysis conducted.

2.1. DATA GATHERING AND DATASET CONSTRUCTION

The first thing to do in order to be able to carry out the research has been the data gathering.

The first step of this process has been the selection of the genres. Spotify provides a list of genres it uses to classify the songs. However, the full list was too long and too detailed, since some of the most popular genres are divided into sub genres. For this reason, we decided to create a list by our own selecting the items from the upper list using the criteria of: the most common genres. Then 36 items have been picked up:
Pop, Rock, Rap, Jazz, Blues, Folk, Metal, Country, Classical, Reggae, Punk, Techno, Trance, EDM, Dubstep, Roots, R&B, Indie, Trap, Instrumental, Hip-hop, House, Salsa, Flamenco, Goa, Gospel, Tango, K-pop, Swing, Dark, Funky, Piano, Grime, Aggrotech, Fusion, Industrial.

As soon as the list have been created, it has been used to set the first Spotify's API: "https://api.spotify.com/v1/search ", that permits to collect a list of Spotify's artists for each one (1000 max). The only limitation that has been applied is related the market of reference, the Italian one. The number of followers and the popularity (a Spotify measure that using several features creates a ranking of the artists) have been also gathered. Finally, a first dataset has been created with each row a different singer, in total of 26.418.

The next step was to connect each artist in the dataset with another using the Spotify's concept of the "related artists". An artist is related to another when it is similar to it. Similarity is based on analysis of the Spotify community's listening history. In order to achieve it Spotify provides a useful API: https://api.spotify.com/v1/artists/{id}/related-artists, this permits to obtain for each artist given as input a list of singers. To simplify the data collection a limitation in the gathered data has been put: only the related artists that were already in the dataset has been saved in the datasets. In this way, each artist could present a related singer and there is the necessary information for each related singer.

The final dataset is then composed by: 26.418 rows and 10 columns, some of them have been used just as a check. The ones that provide actual useful information are: *name, genre_search, genres, popularity, followers, related_to.*

ID	NAME	GENRE_SEARCH	GENRES	POPULARITY	FOLLOWERS	RELATED_TO
06HL4z0CvFAxyc27GXpf02	Taylor Swift	Pop	Pop	100	104.000.000	0C8ZW7ezQVs4URX5aX7Kqx,1McMsnEElThX1knmY4oliG,6jJ0s89eD6GaHleKKya26X, […]
3TVXtAsR1Inumwj472S9r4	Drake	Pop,Rap,R&B, Hip-hop	canadian hip hop,canadian pop,hip,hop,pop rap,rap	96	84.922.890	1RyvyyTE3xzB2ZywiAwp0i,1URnnhqYAYcrqrcwql10ft,6l3HvQ5sa6mXTsMTB19rO5, […]

Table 1: Dataset example

The final dataset than has been converted in a .csv file.

2.2. EXPLORATORY AND DESCRIPTIVE ANALYSIS

As soon as the dataset has been created an exploratory/descriptive analysis has been carried out.

Let's start by looking at the artists dataset.

As already mentioned, it is composed by 26.418 artists. Each artist can be referred to more than one genre:

Number of genres	1	2	3	4	5	6
Number of Artists	20.918	3.270	775	262	90	4

Table 2: Genres for artist

It is possible to see that most of the artists (83%) present only one genre. The other 17% instead can have from 2 to 6 genre combinations. For this reason, each genre has been analyzed separately in order to normalize the data having so a better understanding of them.

The distribution of the genres is then:

Genre	Number of Artists
Trap	1000
Trance	1000
Techno	1000
Tango	502
Swing	971
Salsa	819
Roots	986
Rock	1000

Fusion	1000
Funky	367
Folk	1000
Flamenco	761
EDM	1000
Dubstep	1000
Dark	1000
Country	1000
Classical	1000

Reggae	1000
Rap	1000
R&B	1000
Punk	1000
Pop	1000
Piano	1000
Metal	1000
K-pop	1000
Jazz	1000
Instrumental	1000
Industrial	1000
Indie	1000
House	1000
Hip-hop	1000
Grime	609
Gospel	1000
Goa	255

Blues	1000
Aggrotech	216

Table 3: Distribution of genres

it is possible to see that most of the genres presents several artists equal to 1.000 (75%), the maximum limit of Spotify's API. The others are almost all close to 1000, just three genres present a low number of artists. These genres are: Goa, Funky and Aggrotech. It is not surprising that these genres are not the ones we would call "popular".

Each artist then is characterized by the number of followers it has on Spotify and a metric called popularity than, taking also in consideration the followers (), it uses also other parameters to define a rank, a position is then given to each artist.

Let's look at both:

According with the first feature, the number of followers, we can see that the 5 artists with the highest value are:

NAME	FOLLOWERS
Ed Sheeran	113.101.989
Taylor Swift	103.747.466
Ariana Grande	94.844.417
Billie Eilish	92.107.642
Drake	84.922.890

Table 4: Top 5 artists for n° followers

Instead, the most popular ones are:

NAME	POPULARITY
Taylor Swift	100
Drake	96
Bad Bunny	95
Kanye West	95
The Weeknd	95

Table 5: Top 5 artists for popularity

Looking the two features on a higher level the popularity and the number of followers features, the averages of them are respectively: 37 (the maximum is 100) and 472.786. Both these values are quite low comparing with the values of the five top singers.

Let's check the distributions:

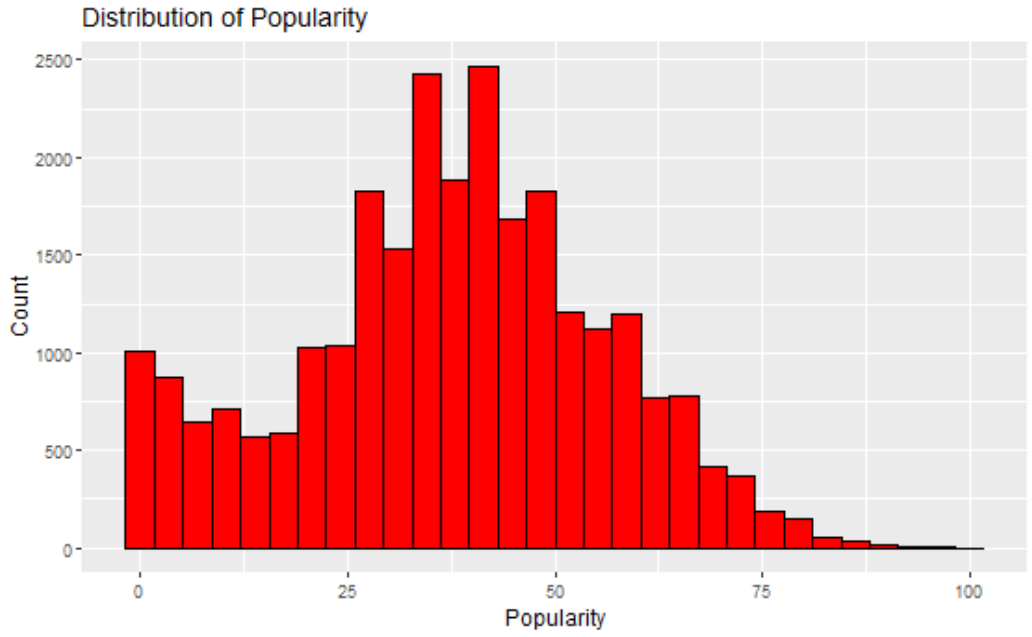

Figure 1: Distribution of Popularity

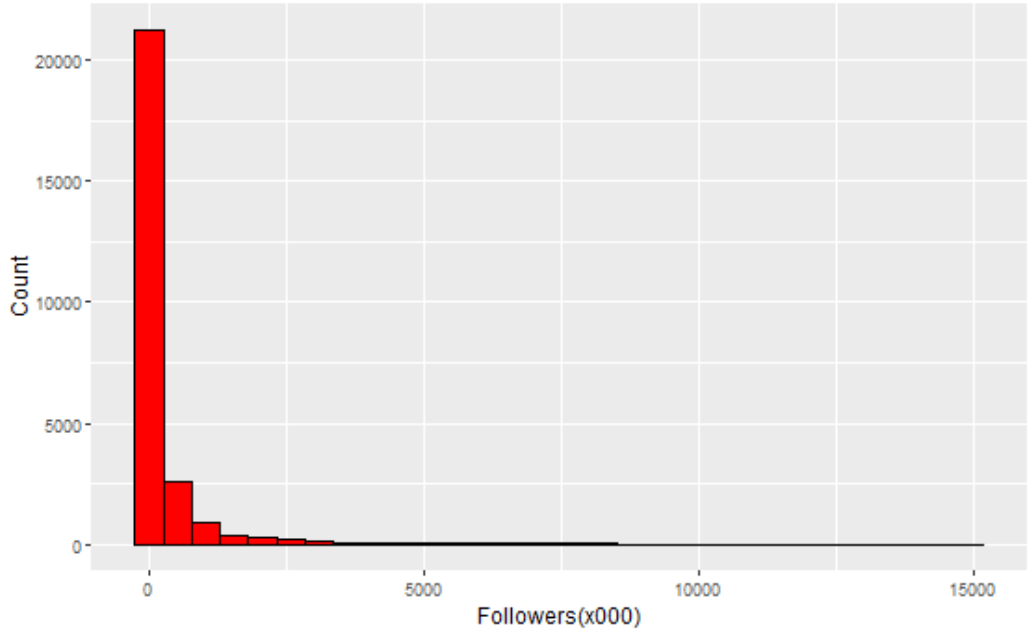

Figure 2: Distribution of Followers(x000)[1]

[1] The artists with a number of followers higher than 15.000.000 have been removed from the plot in order to make easier the reading

From the plots above it is possible to see how most of the artists presents a popularity rank lower than 50 (the 50% of the total artist present a popularity lower than 38 and only the 1% higher than 80).

Talking instead about the number of followers (the real amount has been divided by one thousand for a better reading of the data) also here it is possible to see a concentration over a set range of values. Considering the overall range of values that goes from 0 to 113.101.989, the 50 % of the singers present several followers lower than 26.000 and just the 32% higher than 100.000 (8% higher than 1.000.000).

Taking in consideration all this information, it is possible to analyze the degree of each metric for each genres. The two features are calculated using the average of the artists' values.

Looking at the Popularity, the five top genres are:

GENRE	POPULARITY
Pop	69.33
R&B	69.05
Hip-hop	62.27
Rock	61.77
Rap	61.61

Table 6: Top 5 Genres for Popularity

The five lowest ones instead are:

GENRE	POPULARITY
Goa	5.36
Tango	13.89
Aggrotech	15.45
Industrial	16.98
Grime	19.16

Table 7: Bottom 5 Genres for Popularity

Looking at the general distribution of the poularity between generes, we have that the majority of the genres presents an average popularity between 30 and 50.

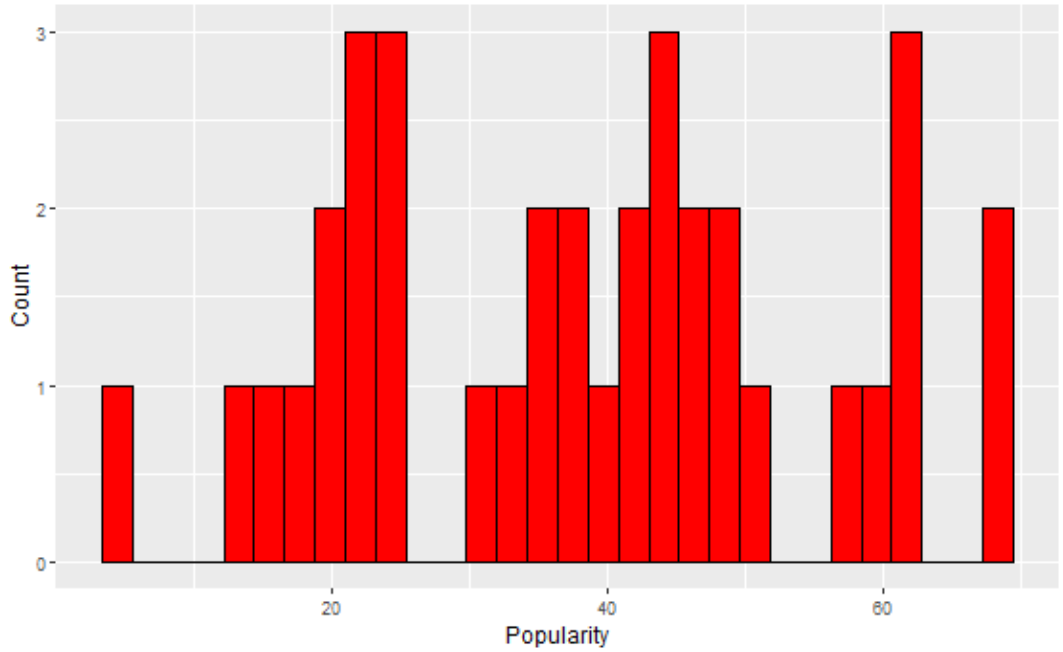

Figure 3: Popularity distribution for Genres

Talking about the followers, we can see that the top five genres are:

GENRE	FOLLOWERS
Pop	5.134.590.47
R&B	4.542.701.53
Hip-hop	2.300.264.77
Rap	2.256.985.47
Rock	2.199.871.73

Table 8: Top 5 Genres for Followers

Instead, the five lowest ones are:

GENRE	FOLLOWERS
Goa	1.883.71
Tango	5.773.69
Aggrotech	12.432.16
Dubstep	23.189.92
Funky	28.840.67

Table 9: Bottom 5 Genres for Followers

The distribution of the average followers for the genres reflects what is visible for the artists. Most of the genres are ranked in the lower part of the plot.

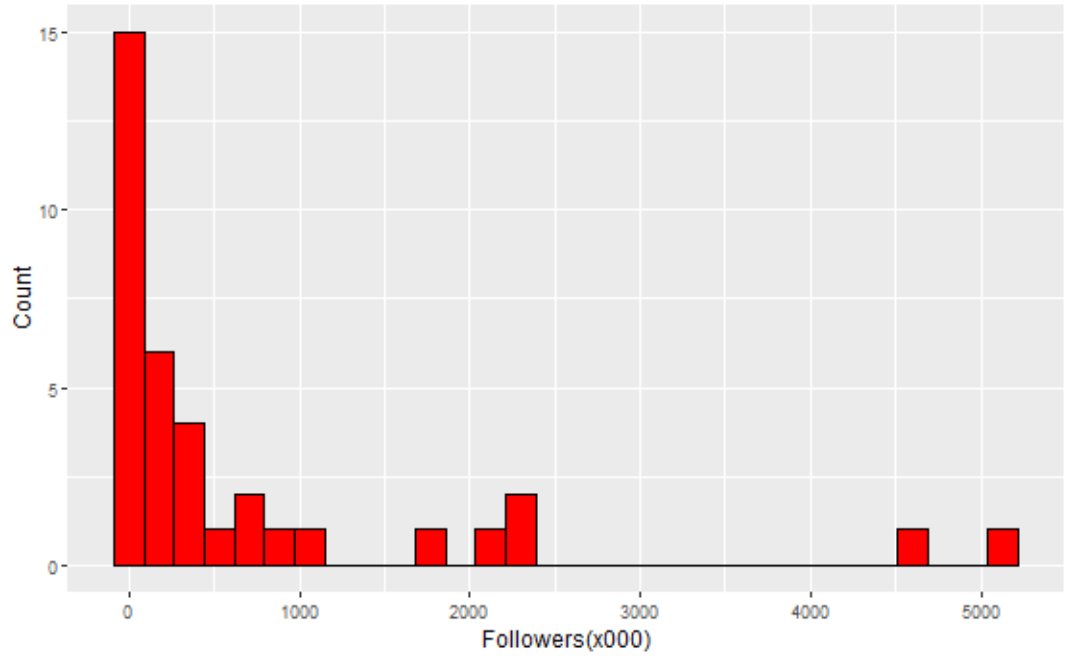

Figure 4: Followers(x000) distribution for the Genres

The two measures are both interesting to study, however they are also strongly correlated with each other.

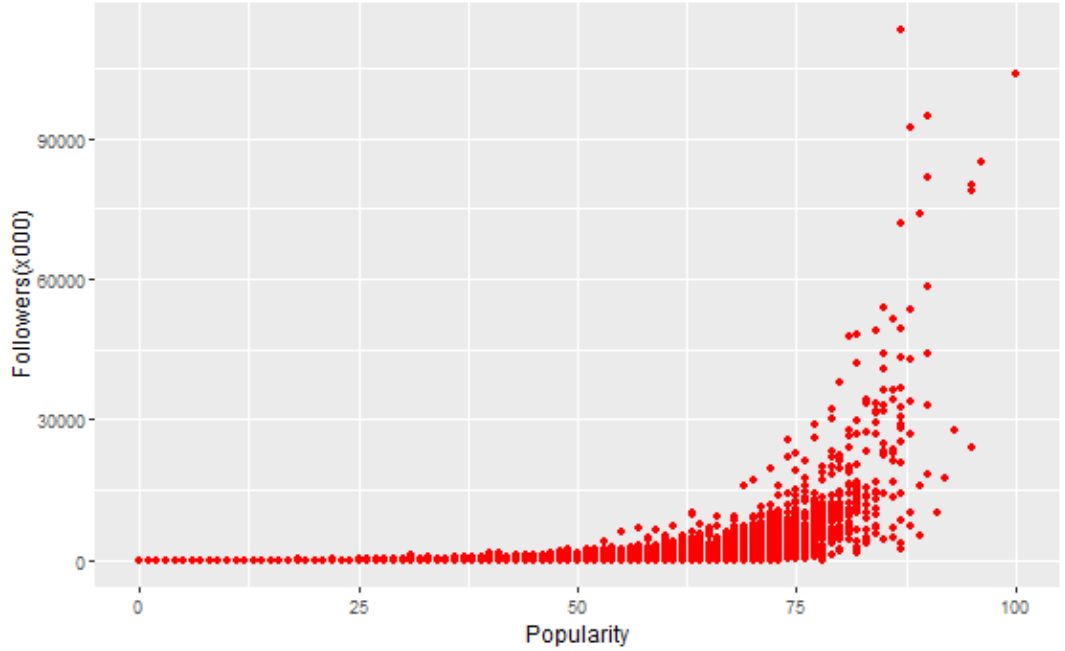

Figure 5: Correlation between the number of followers and the popularity

For this reason, to carry out the analysis we decided to focus just on the popularity measure. This choice come from two principles:

- The popularity measure is normalized. The values in fact- are part of a ranking meaning that they are easier and more accurate to elaborate.

- Since the popularity measure is a calculated measure not only takes in consideration the number of followers an artist has but also other measures like how much time a song has been listed to.

3. NETWORK ANALYSIS

Network analysis is a method used to study the structure and dynamics of networks, which are collections of interconnected entities. It examines the relationships (edges) between the artists (nodes). The data provided by Spotify are inherent just to the artists. As a result, we used this information to gain insights to the genre, particularly which genres are the most influential. So, we performed network analysis on the artists, compute metrics (such as eigen centrality, closeness centrality, betweenness centrality and degree centrality) to answer the research question.

Before performing network analysis, we had to face different problems. Firstly, each artist has similar artists (we have just the ID of these ones), all contained in a single string. So, we had to split each ID of a similar artist. After that, we had another issue: artists have a different number of similar artists, and this makes it impossible to create a network. In order to solve this problem, we used the unnest function, and duplicated the artists for each ID similar artist.

ID	NAME	GENRES	RELATED_TO
06HL4z0CvFAxyc27GXpf02	Taylor Swift	pop	0C8ZW7ezQVs4URX5aX7Kqx
06HL4z0CvFAxyc27GXpf02	Taylor Swift	pop	1McMsnEElThX1knmY4oliG
06HL4z0CvFAxyc27GXpf02	Taylor Swift	pop	6jJ0s89eD6GaHleKKya26X

Table 10: Exploded dataset by related artist and genre

With these transformations, we passed from a dataset of about 26.418 observations to more than 414.504 rows. After that, we used the “igraph” library to create the network, and then compute the metrics written above to gain insights from the genres.

3.1. METRICS

We computed four different metrics in order to gain interesting information about genres.

3.1.1. DEGREE

First, degree centrality. Degree centrality determines the influence within a network based on the number of direct connections (edges) it has to other nodes (and thus this helps us to define which artists are the most directly connected to the other ones).

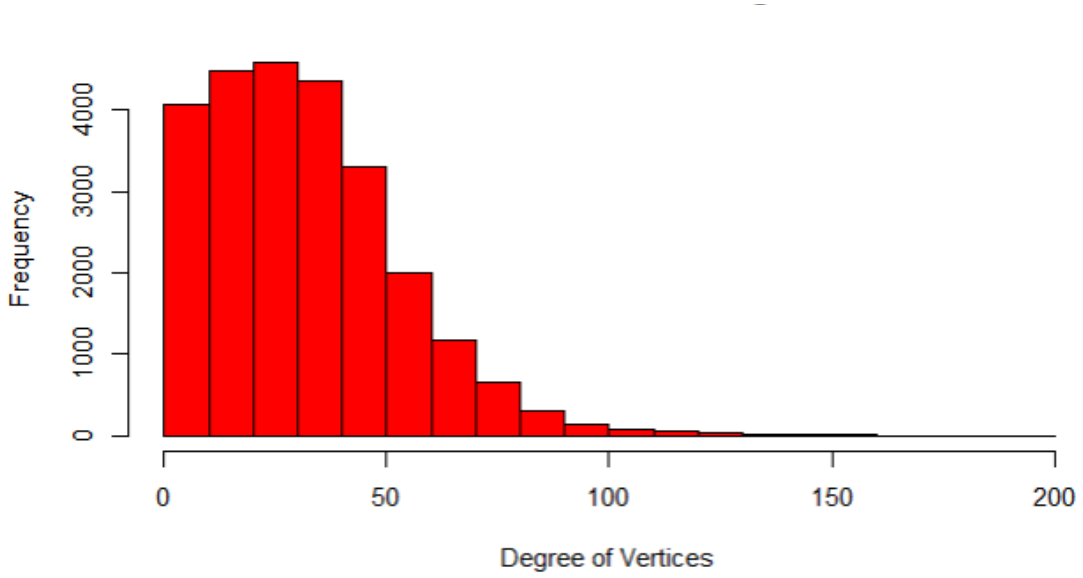

Figure 6: Distribution of Node Degree

As we can see from the plot, a major part of the observations reaches a few degrees between 10 and 40. Higher is the number of degrees, lower is the frequency of the genres. This suggests that just a few genres reach a great number of degrees. In addition, the probability that genres reach 30 degrees is 0.5 (so, the 50% of the genres have a maximum of 30 degrees), while the probability that genres have more than 38 degrees is just 0.2 (so, just 20% of the genres have a number of degrees major than 38).

3.1.2. CLOSENESS

Then, we used closeness centrality which helps us to find influential artists based on how quickly they can connect with other artists in the network. This approach helps highlight the importance of network position and connectivity speed in understanding influence within the network.

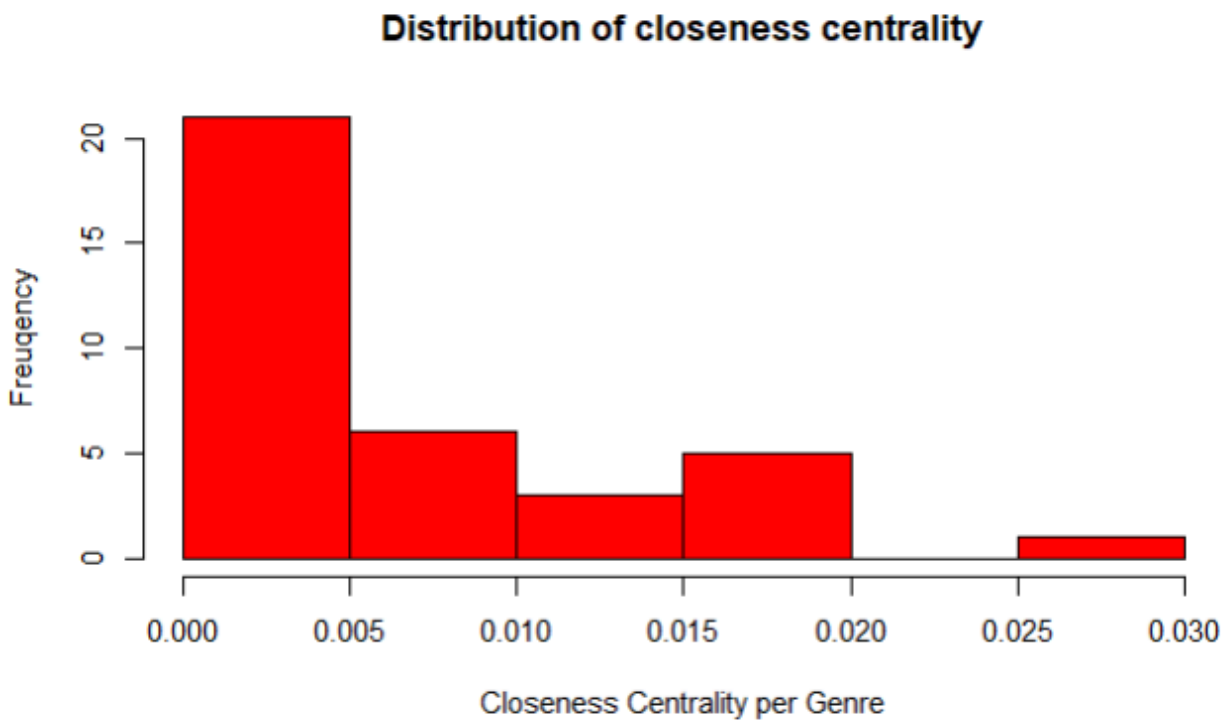

Figure 7: Distribution of Closeness Centrality

As we can see from this plot, less than 1% of the genres achieve a closeness at least equal to 0.02. Many of them have a closeness overall between 0.000 and 0.005.

3.1.3. BETWEENNESS

Another important metric is the betweenness centrality. It is a measure of a node's importance based on the number of times it acts as a bridge along the shortest path between two other nodes. It reflects how often a node lies on the shortest paths between other nodes in the network, indicating its role in facilitating communication or influence between different parts of the network.

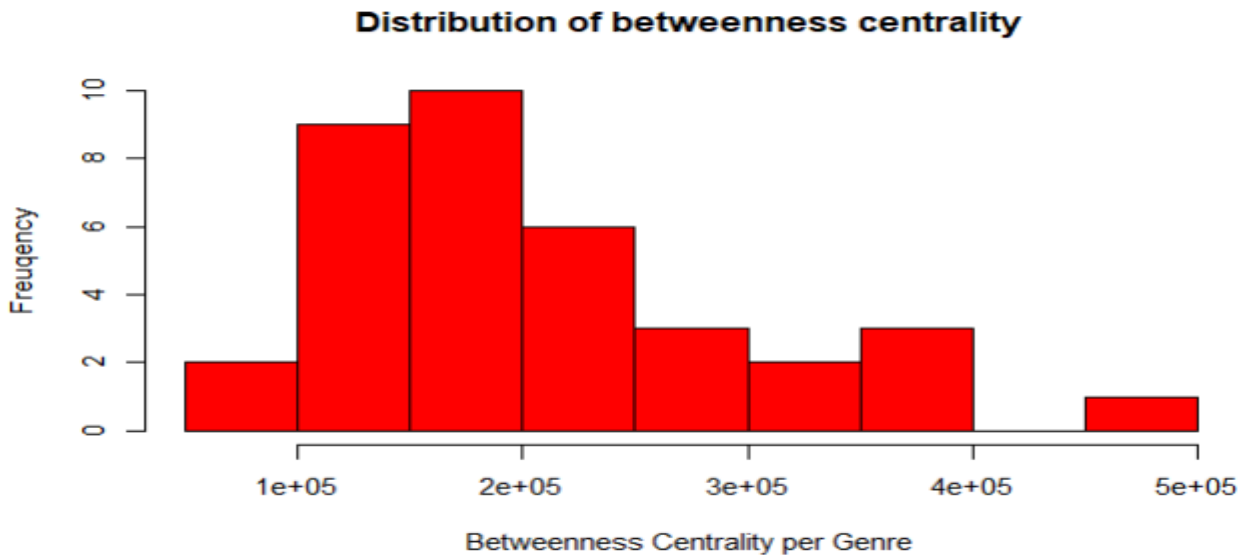

Figure 8: Distribution of Betweenness Centrality

From this plot, it is possible to see that major part of the genres have a betweenness centrality between 100.000 and 250.000.

3.1.4. EIGEN

The last important measure that we used in our research is the eigenvector centrality, which is a measure of a node's influence in a network, considering not just the number of connections (edges) it has, but also the quality and influence of those connections. In other words, a node with high eigenvector centrality is connected to many nodes that are themselves influential.

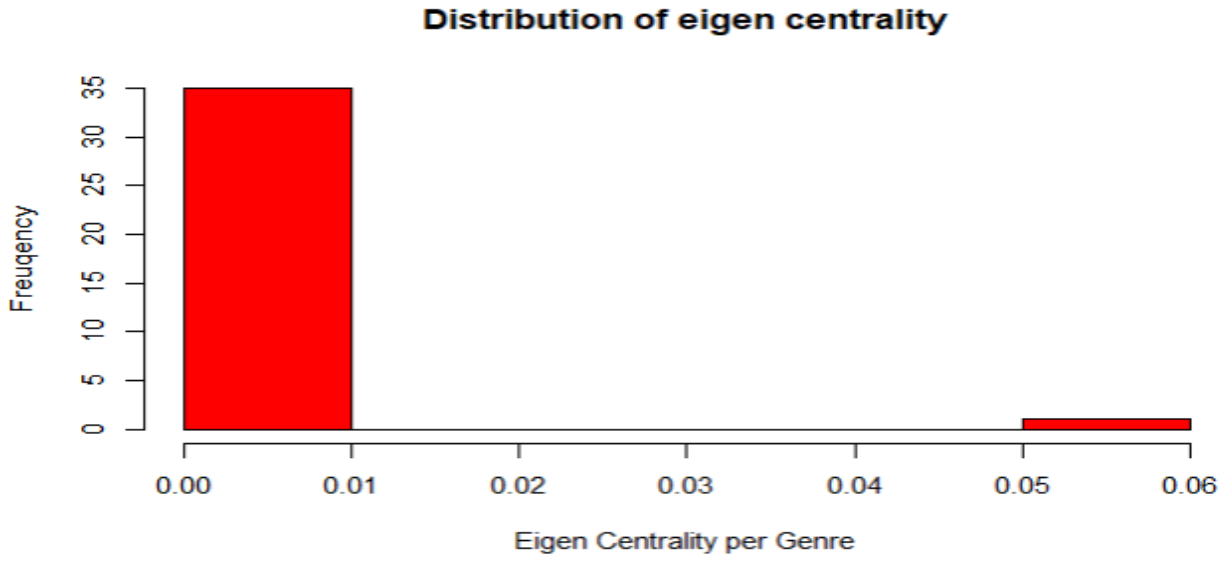

Figure 9: Distribution of Eigen Centrality

Here we can notice that the most observations are between 0.00 and 0.01, suggesting that the most of observations have a very low eigen centrality overall.

3.2. RELATIONSHIPS BETWEEN METRICS

After that we computed these metrics, we inserted them into a new dataset, containing just the relevant fields for our analysis. We created a new dataset just with the unique artists, and then we computed a mean for each genre, weighted for popularity. The important fields we decided to consider are: artist's id, name of the artist, genres, followers, weights (based on popularity), and the metrics written above.

Then, because the research question is about genres, and not artists, we computed the mean (we decided that this metric is the most suitable because genres differ in terms of number of artists, and it leaves the results unaffected by those). After that, to gain more insights, we took just numerical variables and created a correlation matrix for these ones (it is a table showing correlation coefficients between variables, indicating the strength and direction of their linear relationships. Each cell in the matrix represents the correlation between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

	TOT_DEGREE	TOT_POPULARITY	TOT_CLOS_CENTRALITY	TOT_BETW_CENTRALITY	TOT_FOLLOWERS	TOT_EIGEN_CENTRALITY
TOT_DEGREE	1.0000000	0.619601999	-0.05559069	0.57537413	0.47831230	0.090912107
TOT_POPULARITY	0.61960200	1.0000000	0.12592119	0.78042203	0.75373532	0.001260639
TOT_CLOS_CENTRALITY	-0.05559069	0.125921187	1.000000	0.03009621	-0.13906229	0.336611113
TOT_BETW_CENTRALITY	0.57537413	0.780422033	0.03009621	1.0000000	0.61451335	0.129717584
TOT_FOLLOWERS	0.47831230	0.753735324	-0.13906229	0.61451335	1.0000000	-0.086432249
TOT_EIGEN_CENTRALITY	0.09091211	0.001260639	0.33661111	0.12971758	-0.08643225	1.000000

Table 11: Correlation Matrix between variables

As we can see from this table, popularity has a strong positive relationship with degree, betweenness and followers (same for followers, that is strong positive relationship with

degree, popularity and betweenness). Betweenness centrality has a strong positive relationship also with degree centrality. About closeness centrality, we can see that it has a positive relationship with eigen centrality.

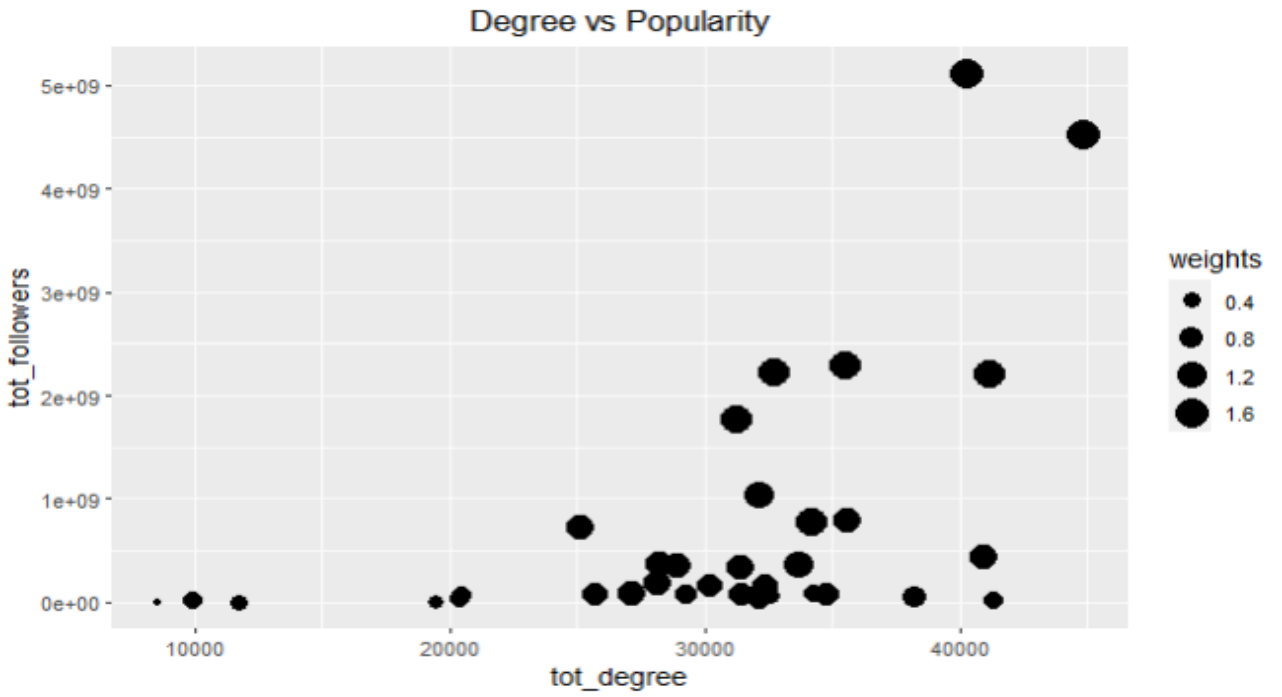

Figure 10: Degree vs Popularity

As we can see from the plot, there is a strong positive relationship between degree centrality of the genre and the followers with popularity (the weights).

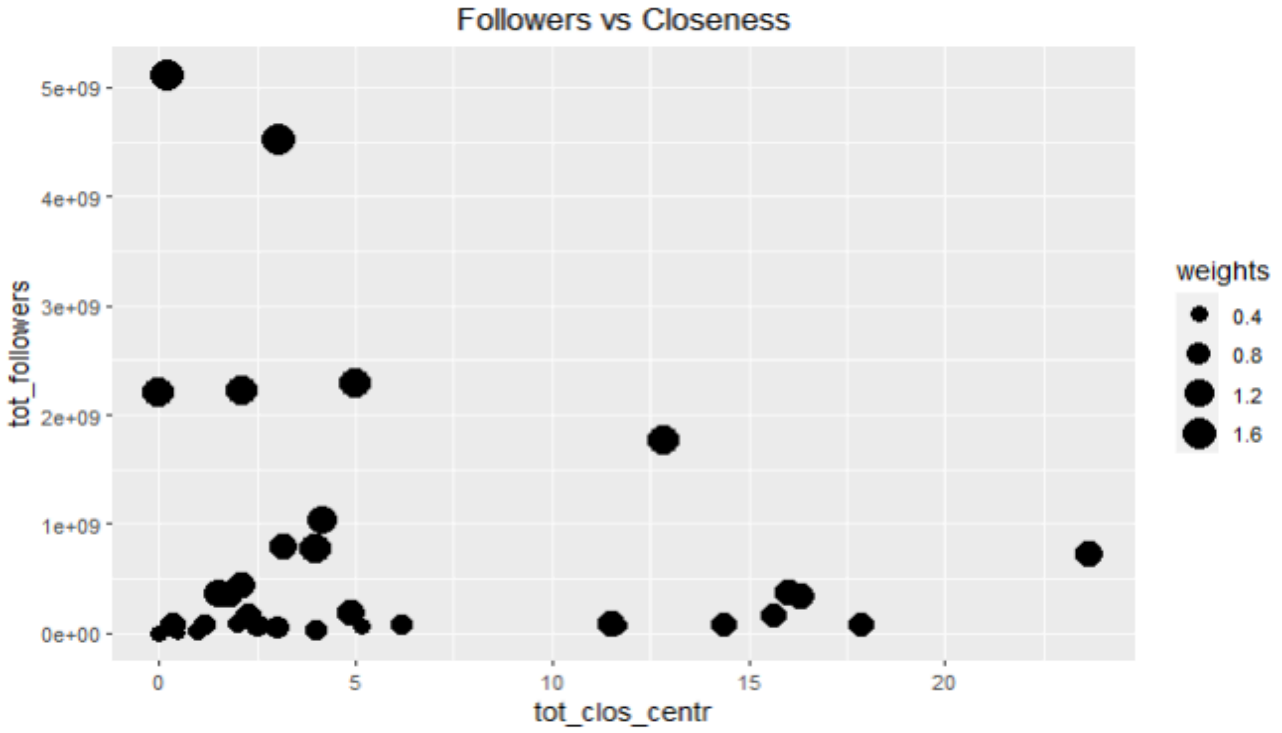

Figure 11: Followers vs Closeness

As we can see from this plot, genres with high levels of closeness don't imply that they have a large number of followers, and in addition, genres with high levels of closeness are the ones with small values of weights (and so, of popularity).

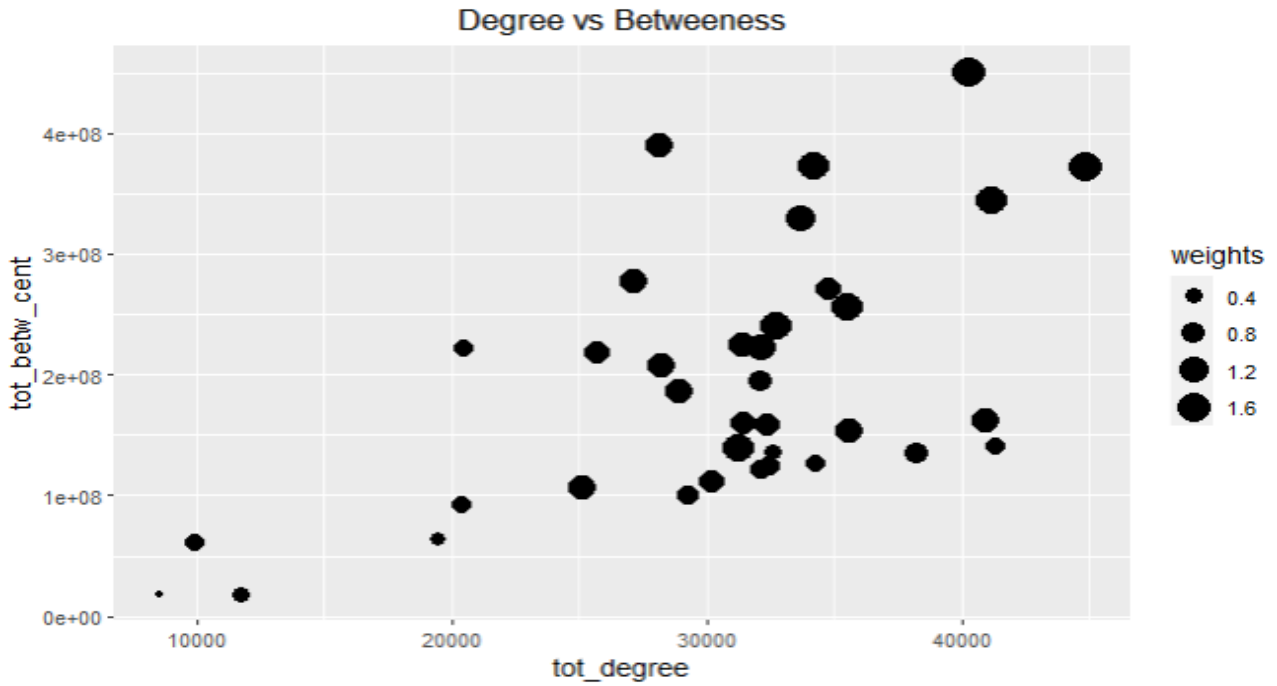

Figure 12: Degree vs Betweenness

Here we can notice the positive relationship between degree and betweenness centrality, and looking at the patterns into the weights, we can see that as popularity increases betweenness and degree increases.

3.3. COMPARISON BETWEEN GENRES

After looking insights into those metrics, we generate a new dataset with all the metrics specified before and the relevant fields about artists (id of the artist, its name and the genre). In this phase we had to solve another issue: how to determine which are the most influential genres? We have a lot of metrics. We normalize degree, closeness and betweenness and then we create a new variable (named influential score) which is the result of the sum of these three variables. The higher the influential score, the higher is the influence of the genre. Below the results.

GENRE	INFLUENTIAL SCORE
Instrumental	3.1335008
R&B	1.9534282

Pop	1.9062755
EDM	1.8895510
K-pop	1.7978064
Folk	1.7694802
Indie	1.7564505
Classical	1.7083999
Jazz	1.6996985
Rock	1.6830114
Piano	1.6726935
Gospel	1.5820069
Hip-hop	1.5723793
House	1.5491288
Trap	1.5480998
Fusion	1.4428250
Reggae	1.3872877
Country	1.3619037
Rap	1.3527310
Trance	1.2781052
Dubstep	1.2743657
Metal	1.2686705
Roots	1.2625575
Techno	1.2545264
Industrial	1.2462182
Blues	1.1719565
Flamenco	1.1553842
Punk	1.1346883
Swing	1.1291450
Dark	1.0706765
Salsa	0.9261379
Grime	0.8294132
Tango	0.5965250
Funky	0.4001010
Aggrotech	0.3026354
Goa	0.2318889

Table 12: Influential score for Genres

1. **Top Influential Genres**:
    - **Instrumental music** emerges as the most influential genre with an Influential Score of 3.1335. This high score suggests that Instrumental music holds a central position in the music network, potentially due to its versatility and

widespread use across different contexts, such as in movies, video games, and background settings.

- **R&B (1.9534)**, **Pop (1.9063)**, and **EDM (1.8896)** follow closely behind. These genres are known for their broad appeal and significant cultural impact, especially in mainstream media and global music charts.
- **K-pop (1.7978)** and **Folk (1.7695)** also score high, reflecting their growing influence globally. K-pop's rise can be attributed to its international fanbase and online presence, while Folk music's enduring popularity may be due to its deep cultural roots and storytelling traditions.

2. **Moderately Influential Genres**:
   - Genres like **Indie (1.7565)**, **Classical (1.7084)**, **Jazz (1.6997)**, and **Rock (1.6830)** maintain a solid presence, indicating their consistent relevance in the music landscape. These genres, while not as mainstream as Pop or EDM, continue to be influential due to their dedicated followings and contributions to music innovation.
3. **Less Influential Genres**:
   - **Salsa (0.9261)**, **Grime (0.8294)**, **Tango (0.5965)**, **Funky (0.4001)**, **Aggrotech (0.3026)**, and **Goa (0.2319)** are among the least influential genres in the study. These genres may be niche or region-specific, which could explain their lower centrality and popularity scores within the broader music network.

Implications of the results::

- **Cultural and Market Influence**: The dominance of Instrumental, R&B, and Pop genres highlights their pervasive impact on both global and local scales. These genres' high scores suggest they are not only popular but also occupy key positions in the network of musical influence, potentially driving trends and shaping the development of other genres.
- **Diversity of influence**: The wide range of Influential Scores across genres suggests a diverse and multi-faceted music ecosystem where different genres influence various segments of the market. Even though some genres like Funky,

Aggrotech, and Goa have lower influence scores, they cater to specific audiences and contribute to the overall richness of the music landscape.

- **Genre Evolution and Adaptation**: The moderate to high scores of genres such as K-pop and Indie indicate how genres can evolve and increase their influence over time. K-pop's rise, in particular, reflects the power of digital platforms and globalized music consumption patterns. This suggests that genres can enhance their influence by adapting to new media and audience trends.
- **Future trends:** :As the music industry continues to evolve with technological advancements and shifting consumer preferences, genres that can leverage these changes may see an increase in their Influential Scores. For instance, genres with strong online communities or those that are highly adaptable to different media formats may become more central in the future.

4. CONCLUSION

The initial idea of the result was that the genre most popular was also the most influential one. The idea was that due to these high values it would have been largely connected with other genres.

However, what comes out from the research is that the popularity element is not the element with the strongest influence over the ability of influencing the user choice. The result of the analysis suggests that the Instrumental, which presents a popularity below the average, is at the top of the rank.

Even if the result is not what we have expected it can be explained by considering that the Instrumental genre is quite transversal around the different genres.

4.1. LIMITATIONS

The results of the research need to be read keeping in mind the limitations of it.

We can summarize them in three main points:

- the source of the data. We decided to use Spotify as a source of the data for the research, for the reasons presented in the introduction. However, Spotify is not the only music service in the market, meaning that the results are biased.
- The used dataset. Even if Spotify provides server useful APIs that permits it to gather a lot of data, it puts some limitations on the amount of data that can be collected. And in addition, we don’t have the possibility to choose which data we want. An example is the number of artists it is possible to collect. Even if the amount is not small, 1000 artists is the maximum, we cannot control the way Spotify provides us the data. Finally, the same artist could belong to more than one genre. Meaning that the amount of artists that we could collect each time was less than 1000.

CONTRIBUTE OF EACH STUDENT TO THE PROJECT

Data collection: Federico

Exploratory Data Analysis: Federico

Network Analysis and Conclusion: Alberto

Presentation: Alberto

Report: both

BIBLIOGRAPHY

Spotify. (s.d.). *Get Artist*. Tratto da Spotify for Developers: https://developer.spotify.com/documentation/web-api/reference/get-an-artist

Spotify. (s.d.). *Search for Item*. Tratto da Spotify for Developers: https://developer.spotify.com/documentation/web-api/reference/search

Statista. (2023). *Music Streaming worldwide.*

Statista. (2024). *Spotify.*