Final Team Project

Puerta de Toledo

Master in Big Data Analytics

Statistical Learning

Ion Bueno Ulacia NIA 100364530

Daniel Gil Santiuste NIA 100364564

Daniel Martín Cruz NIA 100384121

Ion Bueno Ulacia NIA 100364530
Daniel Gil Santiuste NIA 100364564
Daniel Martín Cruz NIA 100384121

uc3m | Universidad Carlos III de Madrid

Contents

Ion Bueno Ulacia NIA 100364530
Daniel Gil Santiuste NIA 100364564
Daniel Martín Cruz NIA 100384121

List of Figures

List of Tables

Before starting, it is important to highlight that we have modified our dataset reducing the number of possible genres of the categorical variable to only four of them: classical, electronic, hip-hop and jazz. The goal of this simplification is to get clearer distinctions between these genres.

Also mention that some original variables which do not provide any useful information for the project have been removed, as author or id of the song. From the 18 original variables, only 11 quantitative variables and the qualitative variable of interest are used.

1. Missing data

There are missing values in the `tempo` and in `duration` variables. In total there are **3.608** samples with one missing value and **183** with two from a total of **20.000** instances.

1.1. Inputting to variable tempo

Before inputting the missing data, a preprocessing was necessary since these values were represented in the dataset as `"?"`. They were replace by `NA` and the whole data was converted to numeric.

We employed the **predictive mean matching (PMM)** to input the missing values. In figure 1 it is shown how they fit in the scatter plot between `tempo` and `liveness`.

Figure 1: Scatter plot of tempo and liveness after imputing new values with PMM

1.2. Inputting to variable duration

As in previous section, it was required a preprocessing to change the null values, since they were represented as `"-1"`. The method PMM was also employed for this variable. In figure 2 it is shown the result in a scatter plot between `duration` and `liveness`.

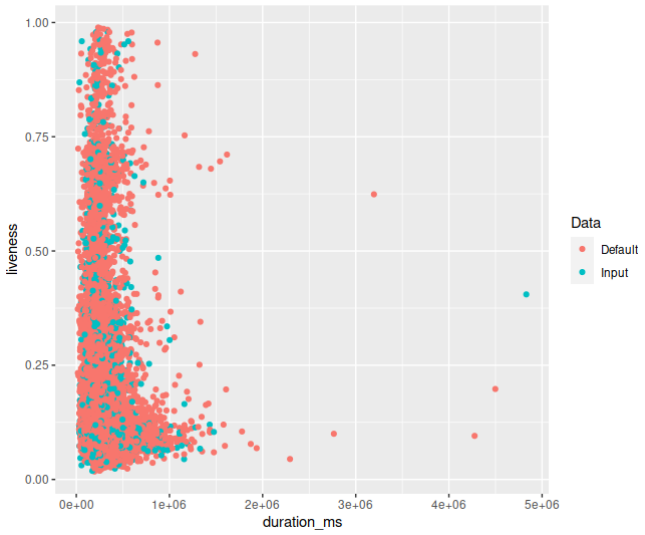


Figure 2: Scatter plot of duration and liveness after imputing new values with PMM

2. Graphical analysis of the variables

In this section there will be shown different graphical representations of the data in order to derive some useful and representative information of it that permits to have some insights of its behaviour. There will also be a special consideration with the qualitative variable to see which variables are the most informative to distinguish the groups formed by such variable.

2.1. Boxplots

The first proposed plot used to get visual information of the data are boxplots conditioned by the categorical variable, differentiating the samples by its class (classical, electronic, hip-hop and jazz).

For the sake of concreteness, there will be shown only those quantitative variables that permit a better differentiation over the categorical variable:

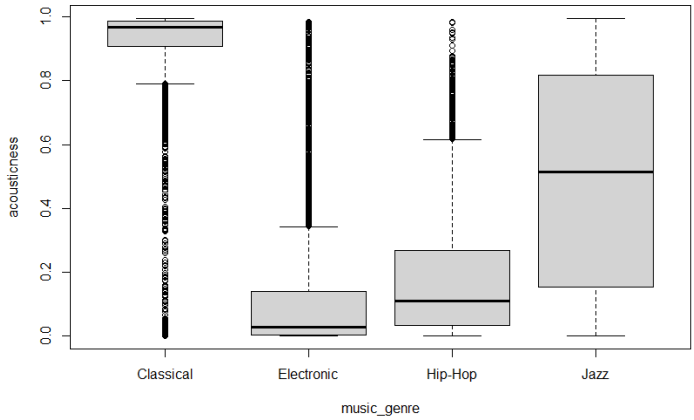


Figure 3: Boxplot of acousticness depending the music genre.

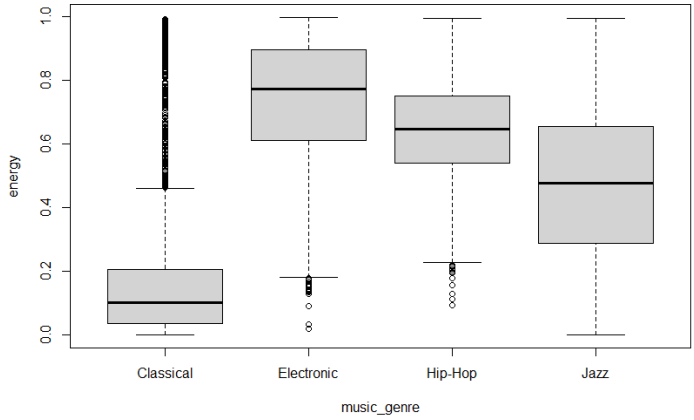


Figure 4: Boxplot of energy depending the music genre

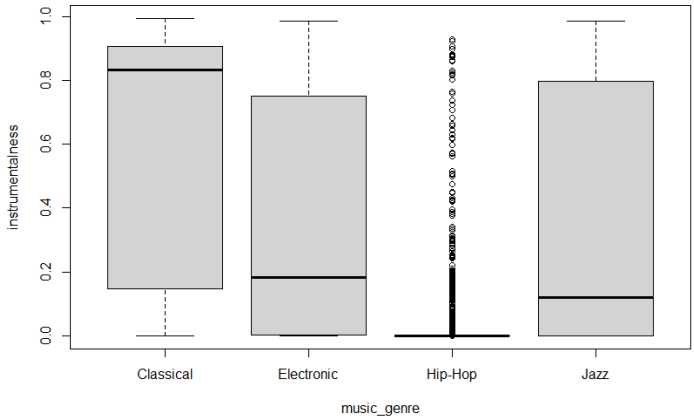


Figure 5: Boxplot of instrumentalness depending the music genre

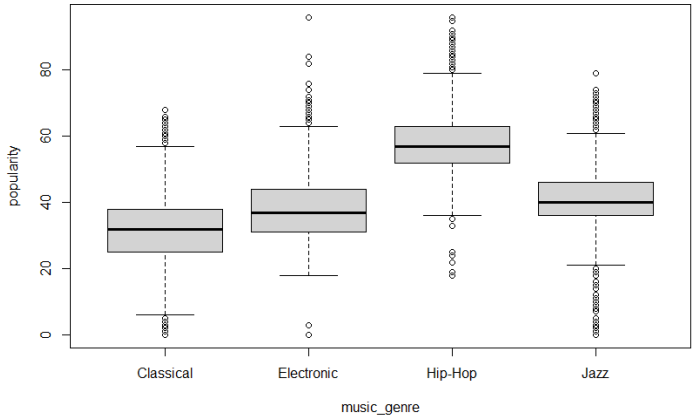


Figure 6: Boxplot of popularity depending the music genre

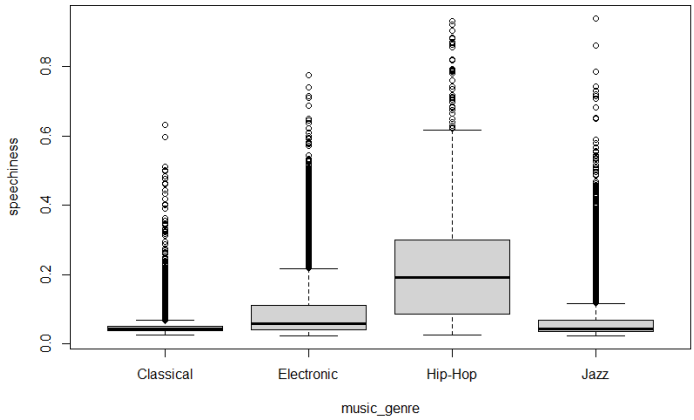


Figure 7: Boxplot of speechiness depending the music genre

The first conclusion you can get with these plots is that the classical music will be the genre that will be easier to detect as it can be seen in the previous plots.

There are other genres more similar such as electronic and jazz music (according to the characteristics studied here) but it can be seen that variables like *acousticness* can be useful in this differentiation.

2.2. Pairs plot matrix

If the goal to accomplish is to distinguish groups using variables, a graphical representation that can help is the pairs plot: a matrix with scatter plots involving all the possible pairs of variables.

Here is the pairs plot without differentiating between categories:

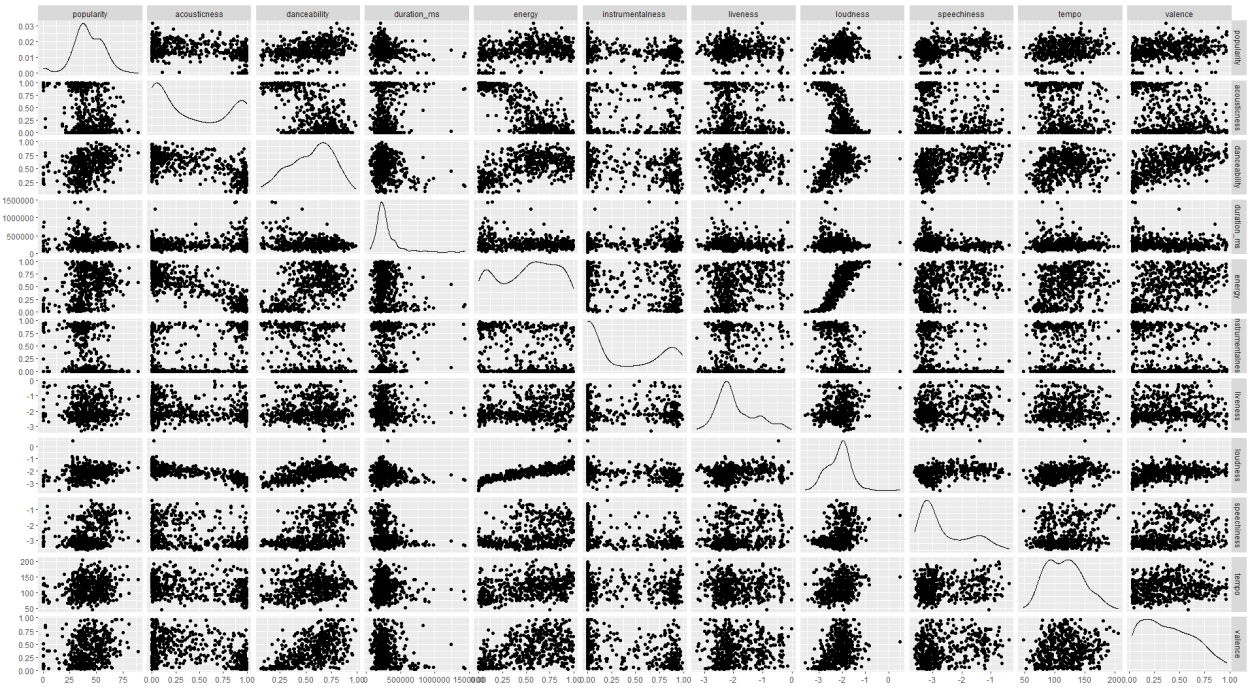


Figure 8: Pairs plot of the dataset

It is not possible to see a clear differentiation of the four groups in any of the plots. What can be seen is here is that it may be possible to differentiate some group from the others as it was concluded from the previous subsection with the boxplots analysis. In this way, it can be possible to differentiate between all the groups, not only with a pair of variables, but with all the variables in the dataset.

Now it will be shown the same plot but using different colors for each of the categories:

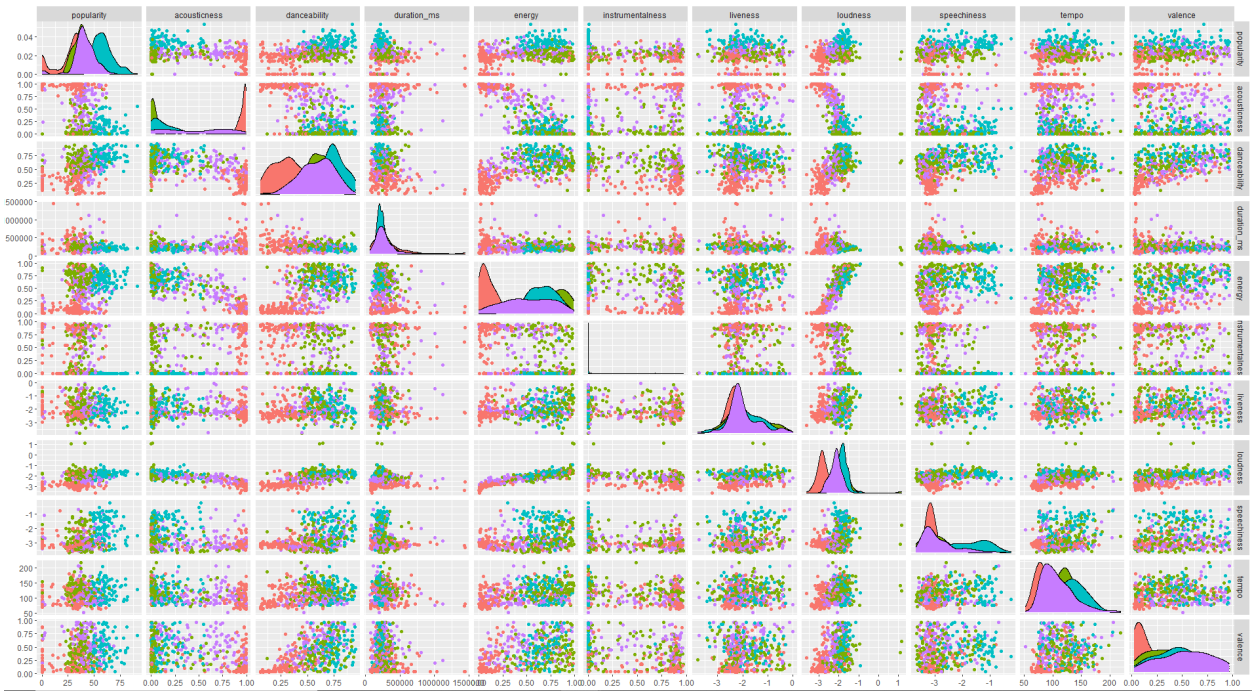


Figure 9: Pairs plot of the dataset distinguishing by classes

As it was mentioned with the previous plot, there is no pair of variables with a clear differentiation of groups but the goal will be to use all the variables to classify registers properly.

2.3. PCP

The Parallel Coordinates Plot is a very useful tool in order to see differences in the behaviour of the four classes present in this dataset. This plot has been created with the function `geom_pcp`. Where each color corresponds to a different musical genre.

In order to improve interpretability and performance, there will be used a random subsample of the dataset of 500 registers.

It is possible to see this PCP in figure 10.

Figure 10: PCP

There are some interesting insights here. It can be seen again that the category most "different" with respect to the other ones is Classical music, showing a low popularity, danceability, energy and loudness and high acousticness and instrumentalness. It is also seen that Hip-Hop music tends to have always the same value of instrumentalness.

This plot can be useful too in order to study outliers, but this will be further explained in section 4

2.4. Data transformations

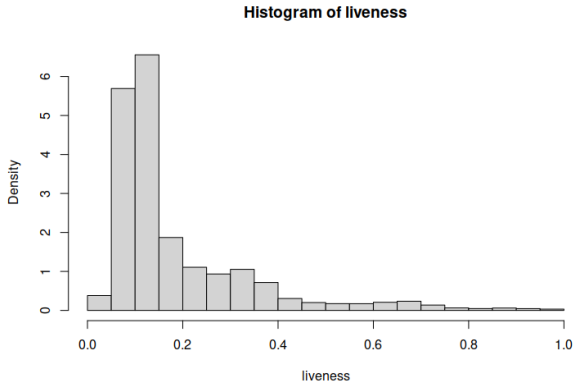
The next plot to investigate were histograms, that allow to have an insight of the distribution of the data.

The main goal of this subsection will be to detect the more skewed variables and to decide of it was convenient to transform the data in such a way that this skewness is reduced. The procedure was looking the histogram of each variable as well as calculating the skewness coefficient to decide if it was suitable to perform a transformation.

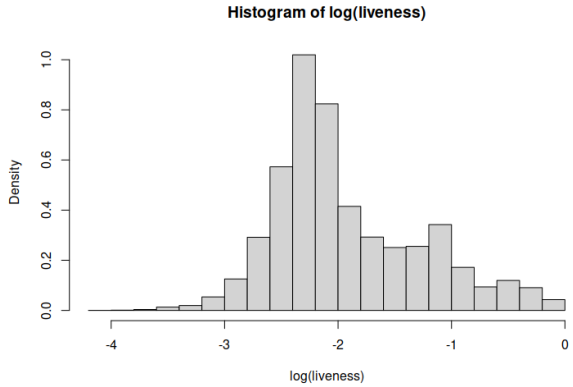
As there are in total 11 quantitative variables, we are going to mention only the ones where it was necessary the transformation.

2.4.1. Liveness

In this case the distribution is right skewed as it can be seen in figure 11a. For this reason a **logarithmic** transformation is performed. The skewness coefficient was **2.226**, and was reduced to **0.719**.



(a) Histogram of liveness

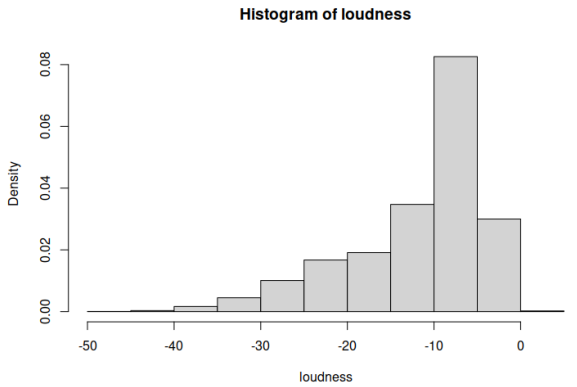


(b) Histogram of transformed liveness

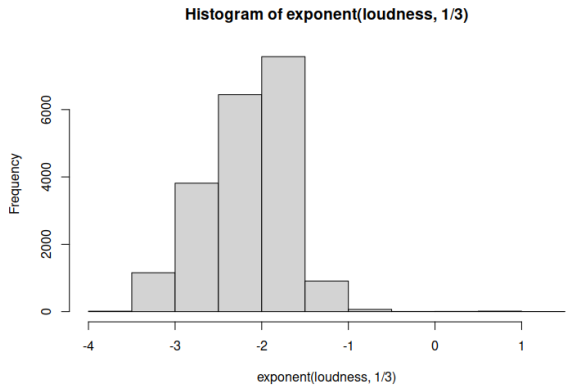
Figure 11: Histogram of liveness and its transformation

2.4.2. Loudness

As it is shown in figure 12a, the distribution is left skewed with negative values. For this reason the logarithmic transformation cannot be applied without a shifting. In this case we applied a **cube root** transformation, which can handle negative values and provides a more symmetric distribution than a logarithmic with a shifting. The original distributions had a skewness of **-1.216** and was reduced to **-0.133**.



(a) Histogram of loudness

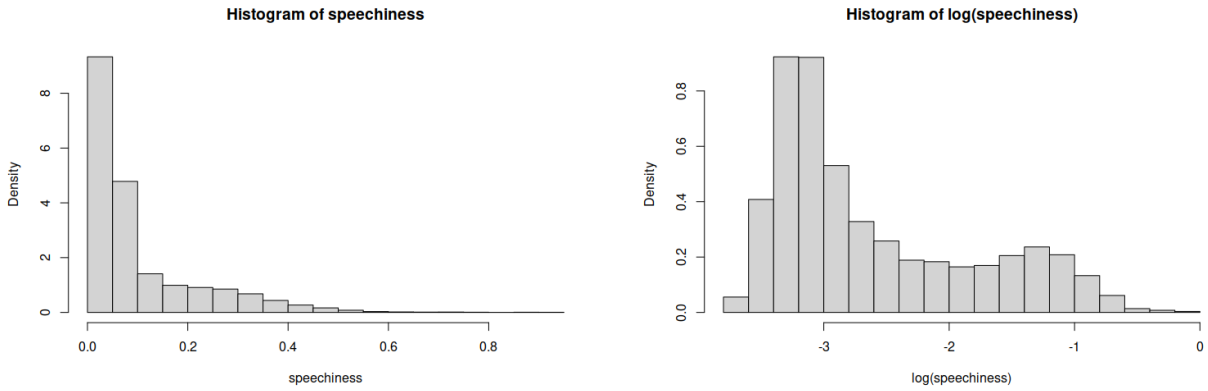


(b) Histogram of transformed loudness

Figure 12: Histogram of loudness and its transformation

2.4.3. Speechiness

The last one corresponds with a right skewed distribution where it was a applied again a **logarithmic** distribution. The skewed coefficient was decreased from **2.109** to **0.886**.



(a) Histogram of speechiness

(b) Histogram of transformed speechiness

Figure 13: Histogram of speechiness and its transformation

3. Main characteristics of quantitative variables

Once the graphical analysis of the dataset was carried out and the most informative quantitative variables were discovered, a study on the main characteristics of each quantitative variable is made.

They are considered as main characteristics the mean vector, the covariance matrix and the correlation matrix. We will go through a first analysis with all the observations in the dataset. Afterwards, a similar study will be carried out on each of the groups defined by the qualitative variable. Finally, a comparative stage would give us important insights about the data we are using for the project.

An important detail that should be taken into account from the comparison conclusions is that they will be taken only on a subset of music genres, not on the whole music universe. So they should be treated like that, because if new music genres were introduced, the extracted conclusions could, logically, change. A remarkable aspect is considering the transformations of variables liveness, loudness and speechiness instead of original variables. To ease the reading of the conclusions, it is omitted from now on, but should be taken into account.

3.1. Study with all the variables in the dataset

The mean vector for the whole dataset can tell us about the general musical characteristics of the genres under study. For each genre, it will be seen which of the characteristics are clearly above or below the mean and conclusions will be made from it.

Covariance and correlation matrix can give us some insights and conclusions, in this case, about the characteristics' linear relations for the music sample considered in the project. For the sake of simplicity and to ease the interpretability of the results, a graphical representation of the correlation matrix is shown and studied. By using this measurement, the linear dependence among the different characteristics can be compared as the correlation is independent from the scaling of the variables.

As it can be seen in figure 14, there exists a strong positive linear relationship between energy and loudness. It can also be said that, in general, danceability of a music piece increases with the popularity, energy and loudness of the song (all of them have a positive linear relation with the former characteristic). Another important

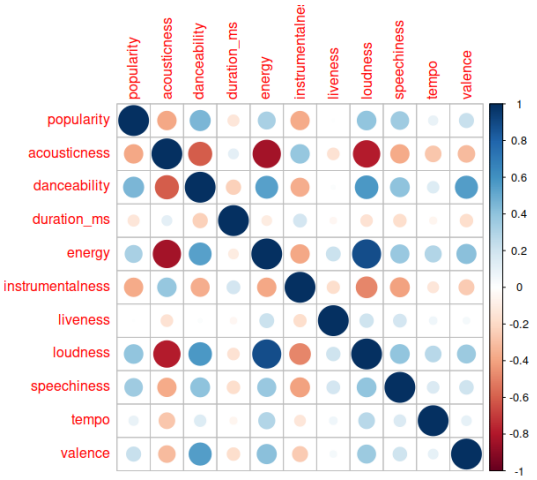


Figure 14: Correlation plot for all the genres in the dataset

linear relation is the one between the positiveness of a song (valence) and its danceability. On the other hand, when studying the strongest negative linear relationships, it can be established that acousticness is inversely proportional to danceability, energy and loudness.

In fact, all these statements could be known in advance, without carrying out the present study. They are part of the music culture. Now, we will see whether these "music laws" stand for all the music genres under study or not. Additionally, it will be checked if new particular relations could arise in any of the genres.

3.2. Study per music genre in the dataset

The following table contains the mean vector of the characteristics corresponding to all the observations and for each of the genres:

Feature	All genres	Classical	Electronic	Hip-hop	Jazz
Popularity	41.69	29.32	38.11	58.40	40.93
Acousticness	0.42	0.87	0.12	0.18	0.49
Danceability	0.56	0.31	0.62	0.71	0.58
Energy	0.51	0.18	0.74	0.64	0.47
Instrumentalness	0.33	0.60	0.35	0.01	0.35
Liveness	-1.93	-2.05	-1.84	-1.83	-1.99
Loudness	-2.16	-2.73	-1.85	-1.87	-2.20
Speechiness	-2.61	-3.06	-2.64	-1.83	-2.9
Valence	0.40	0.21	0.39	0.47	0.51
Duration (ms)	239764	278014	244553	198396	238092
Tempo	115.33	103.95	125.54	119.98	111.86

Table 1: Mean vector for each variable in the dataset.

From the mean vectors table, it can be seen that hip-hop, which is the most popular one, has danceability, speechiness and loudness clearly above the mean of the studied genres. In contrast, its instrumentalness is much less than the others (in fact, this is the genre lowering the overall average). Then it could be thought

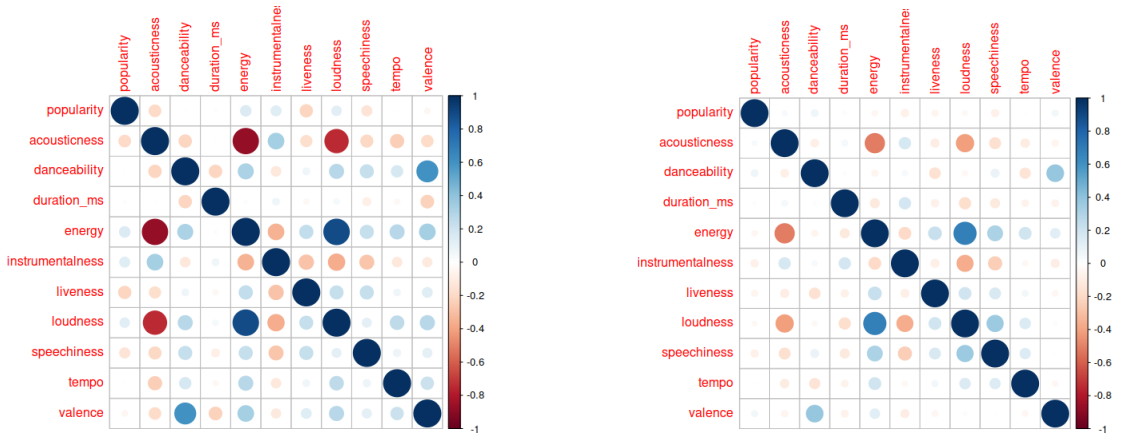
that the current musical taste in terms of popularity is having less instrumental and more speaking parts in the songs. The opposite situation is reflected in classical music, which has a very low popularity joint with the highest instrulmentalness and the lowest speechiness.

Another conclusion would be that people likes dancing loud songs. This can be seen not only on the hip-hop sample (higher danceability and second higher loudness means) but also for the electronic, which presents the second higher danceability and the highest loudness means. This was something stated when analysing the total correlation matrix, so it can be seen that, in fact, holds.

When studying the valence of the genres (musical positiveness conveyed by a track), the fact that the most popular genre is not the one sounding more positive could catch our attention (although hip-hop which is the most popular is the second in the valence classification). Jazz is the most positive within the genres under study. This shows that a genre does not need to transmit positive feelings to be socially accepted and preferred, which is an important statement about people behaviour.

Sideways, analysing people attitude, most energic music is not the most preferred one neither. Maybe because people uses music to relax or distract from other things, which could explain that energy is not the most preferable feature in a song. Another interesting aspect is that popularity order is the same as less duration order, meaning that most popular genres in mean are those with the shortest songs, in mean.

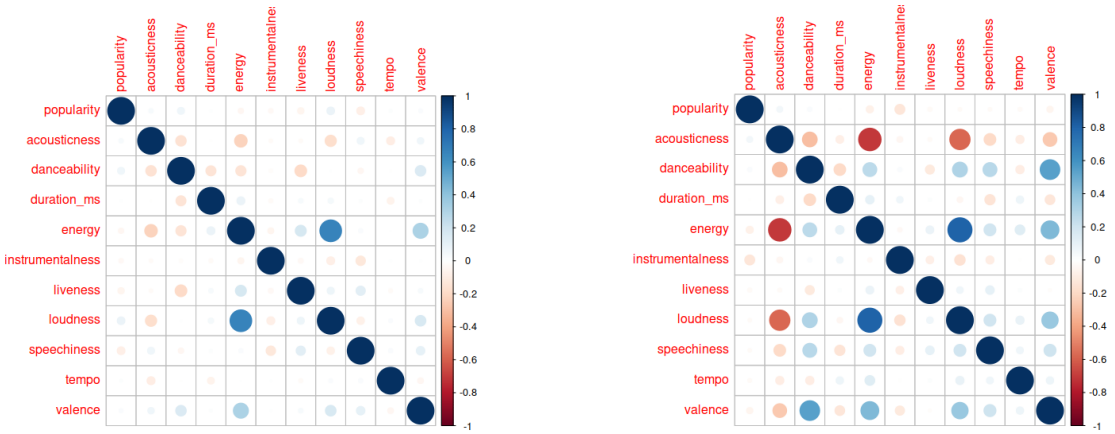
After this first analysis, the relationships between variables for each genre are studied by a similar plot to the one used for all the observations.



(a) Correlation plot of classical music

(b) Correlation plot of electronic music

Figure 15: Correlation plot per genre (I)



(a) Correlation plot of hip-hop music

(b) Correlation plot of jazz music

Figure 16: Correlation plot per genre (II)

From the correlation plots for each genre, it can be stated that all the chosen genres follow, on their own, a similar trend to the one of the aggregated group when analysing the features relationships. But there are some exceptions.

In the case of electronic music, more energic songs do not mean a higher danceability anymore. For this genre, there exists a slightly negative linear relationship, not positive as before. This can be explained by the very nature of the music genre (which is the most energic one among the chosen), whose most energic songs are so energic that receive a low danceability score. Then one of the “music laws” which were established at the beginning does not hold for this genre as it represents one of the extremes of that law. Something similar happens for hip-hop (the second most energic one), for which this same explanation stands as well.

This study of the extremes of each variable presents also interesting effects on other song features such as the danceability. The genre with a higher score on it was hip-hop. Previously, a strong linear relation between danceability and valence was seen, but when analysing for this particular genre (the most danceable one), this relation gets much weaker, almost inexistent in comparison with the other studied genres.

In this way, the proposed analysis was carried out. Some interesting conclusions were extracted from it that give us a deeper knowledge on the project dataset and how the different music genres interact among them.

4. Outliers

Due to the high distortion that outliers can generate in classification, it is required to look for them. However, it is important to mention that most of the variables are bounded between 0 and 1, so it is more likely having samples which belong to the tails of the distribution rather than outliers.

In order to have an intuition of the data, in figure 8 it can be seen how the variables are related in pairs. Several variables are between 0 and 1 as `popularity`, `acousticness` or `danceability`. Taking into account the high number of samples, it makes sense that the whole space is filled in many pairs as `acousticness` and `danceability` or `liveness` and `instrumentalness`.

However, other pairs with variables as `duration` and `loudness` show some points which can be classified as outliers. Mention these two variables are not bounded.

As the goal of the project is the classification of four classes (Classical, Electronic, Hip-hop and Jazz), it makes more sense inspecting the variables per group, in the pairs plot of figure 9. The detection of outliers is going to be carried per class, since the samples of other groups are only going to introduce noise during this process.

The method employed corresponds with the **minimum covariance determinant (MCD) estimator**. This algorithm is mainly appropriate for approximately symmetric datasets. For this reason, the first approach was transforming the variables per group, since in this step there is not any prediction, the unique goal is cleaning the data. However, the distinction between points as possible outliers was less clear than using the transformations for the whole dataset, explained in section 2.4.

The followed steps are the same in all groups.

- Calculate the robust mean vector, covariance matrix and correlation matrix of the data set employing the `CovMcd` function of the library `rrcov`. Mention different values for the parameter h, which corresponds with the percentage of observations included to minimize the determinant of the covariance matrix, have been tried. The value of h which provides a more clear difference between outliers and non-outliers has been selected.
- Compare the Mahalanobis distances of all samples to evaluate which points could be considered outliers.
- Perfom a PCP plot to visualize the behaviour of possible outliers.

4.1. Outliers in Classical group

After calculating the robust mean vector, covariance matrix and correlation matrix of the split dataset, in figure 17 it can be seen how the total variability is reduced since the first eigenvalue is decreased, which is the one providing more weight. It has been used a value of $h = 0.95$.

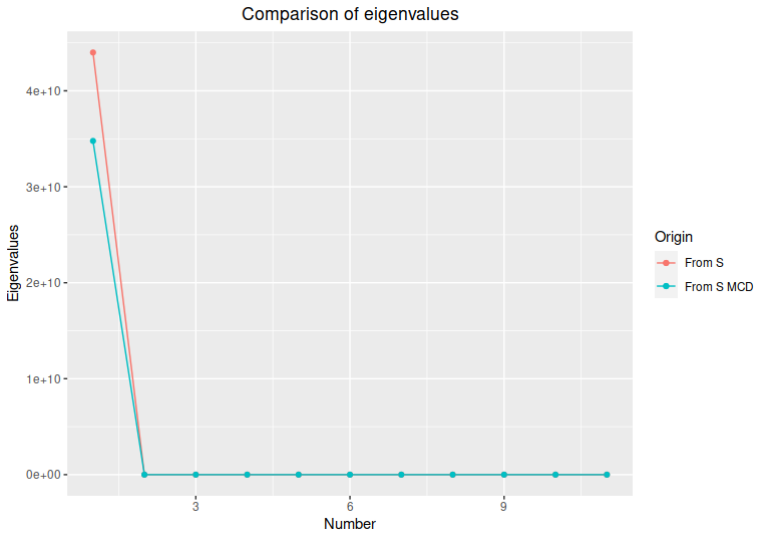
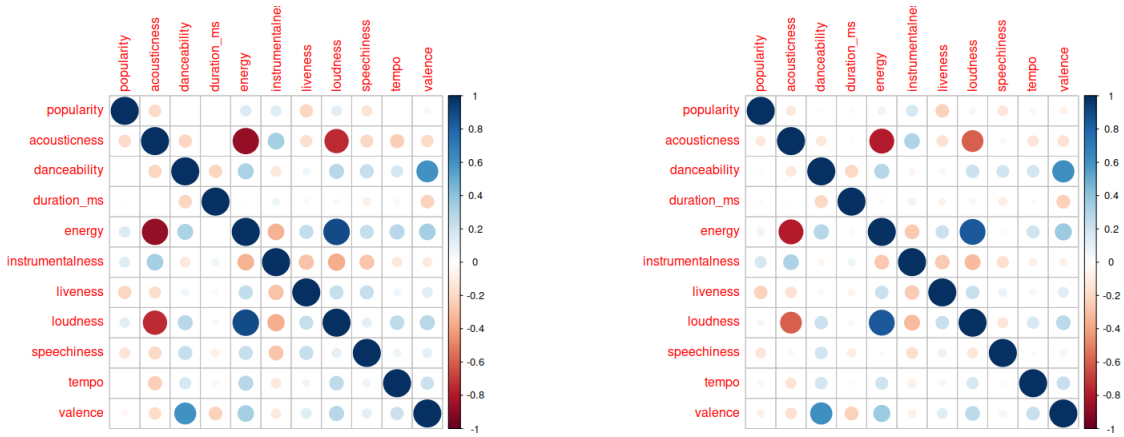


Figure 17: Comparison of eigenvalues in the Classical group

However, there is no great difference and consequently the correlations between variables do not variate a lot, which can be checked in figures 18a and 18b.



(a) Correlation matrix

(b) MCD correlation matrix

Figure 18: Correlation matrices comparison of Classical group

Taking into account the criterion of the Mahalanobis distances, many points could be considered outliers as it is reflected in figure 19a. In spite of that, it is a huge amount of points compared with the total number of samples. Looking into the logarithmic distances, figure 19b, it can be seen how only three points are in the upper range. For this reason, for the Classical group only these **three points** are considered outliers and consequently removed.

(a) Squared Mahalanobis distances

(b) Log of squared Mahalanobis distances

Figure 19: Mahalanobis distances of Classical group

In order to ensure the selection of outliers, a PCP plot is shown in figure 20 where the behaviour of the rest of points is plotted respect the ones chosen. In this case it can be seen how `duration` and `loudness` are slightly different than the others.

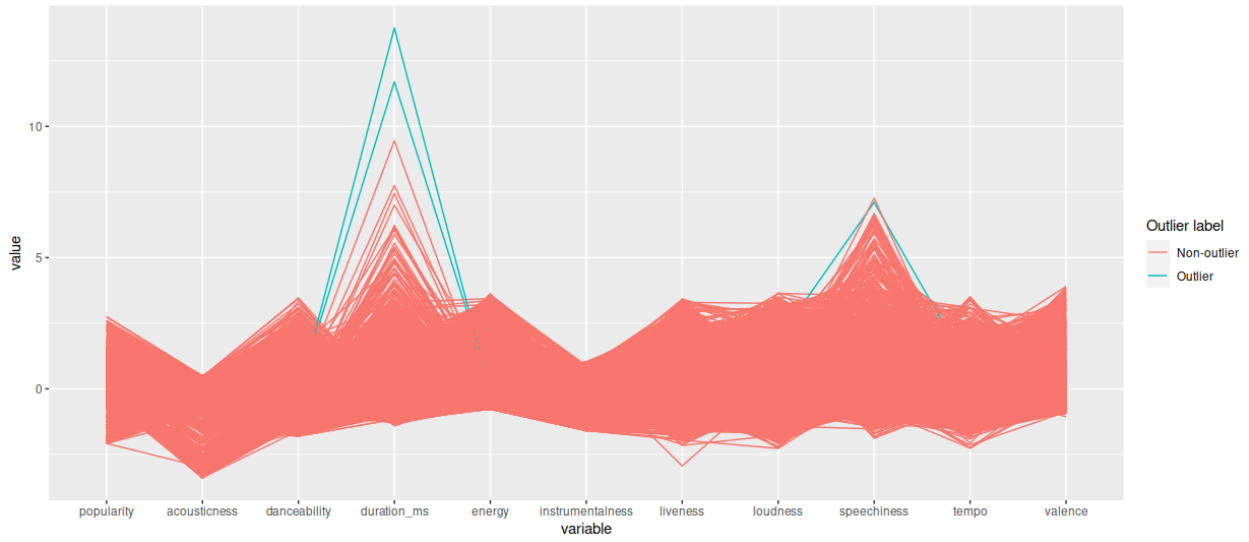


Figure 20: PCP plot with possible outliers in the Classical group

4.2. Outliers in Electronic group

In this group the MCD estimator is more effective since the variance of the covariance matrix is reduced significantly, plotted in figure 21. A value of $h = 0.85$ has been used.

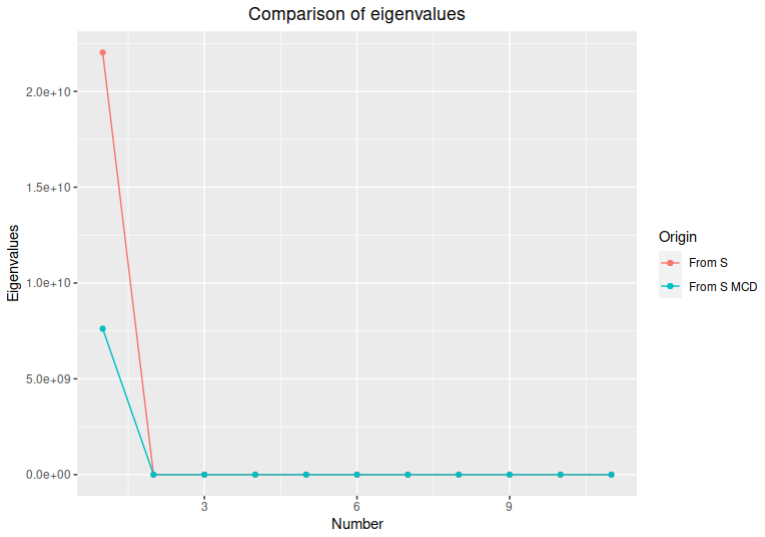
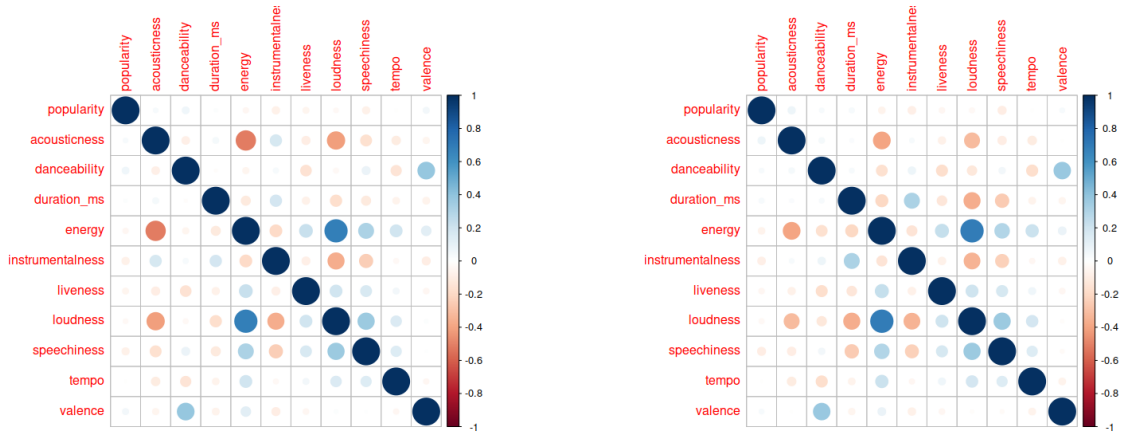


Figure 21: Comparison of eigenvalues in the Electronic group

Consequently, the differences between the original correlation matrix (figure 22a) and the MCD matrix (figure 22b) are more significant. Some variables are less correlated as `energy` and `acousticness` while others more as `loudness` and `duration`.

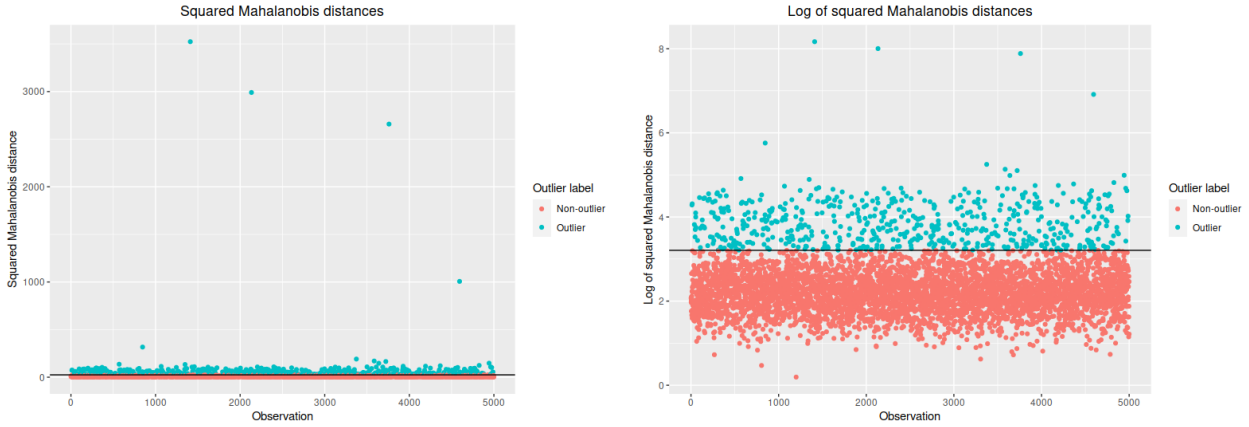


(a) Correlation matrix

(b) MCD correlation matrix

Figure 22: Correlation matrices comparison of Electronic group

The differences in the Mahalanobis distances (figure 23a) is huge, where it can be clearly seen how four points are different than the rest. For this reason these **four point**s are considered outliers.



(a) Squared Mahalanobis distances

(b) Log of squared Mahalanobis distances

Figure 23: Mahalanobis distances of Electronic group

As in the previous group, the patterns are compared with a PCP plot in figure 24. However, in this case this plot does not provide clear information, since the outliers are not distinguished. The points are not differentiated for their values in `duration` and `loudness`, then, other variables should have this role.

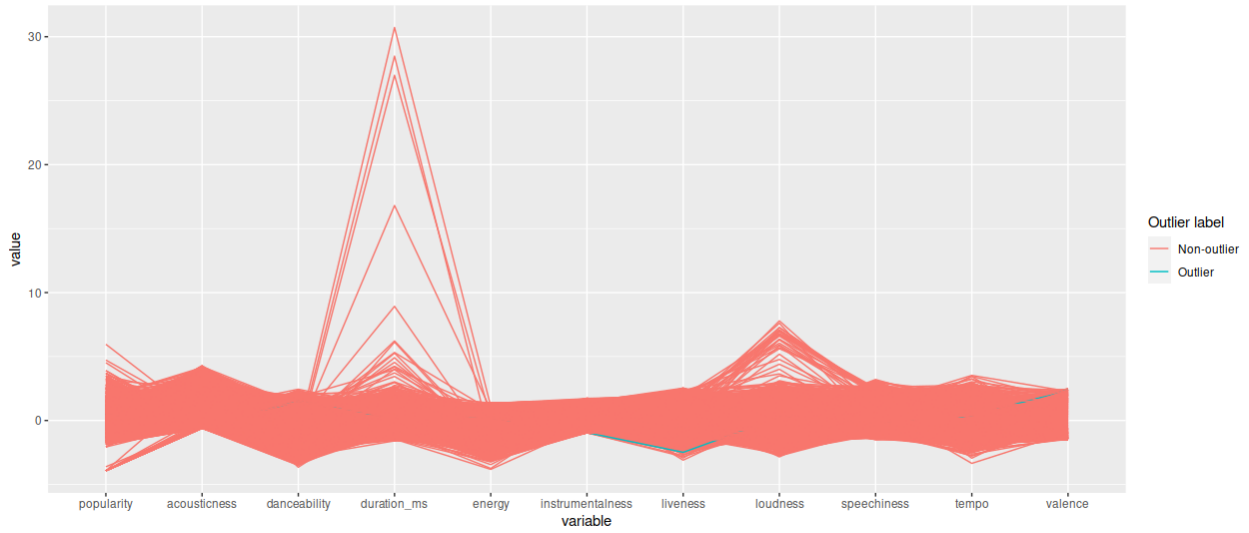


Figure 24: PCP plot with possible outliers in the Electronic group

4.3. Outliers in Hip-hop group

This group is the most controversial. Firstly, the variance cannot be well reduced as it is shown in figure 25. In this case $h = 0.9$.

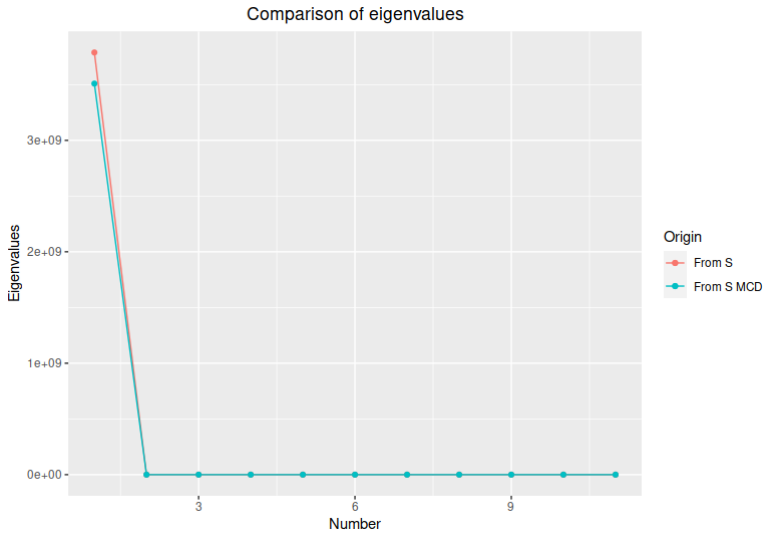
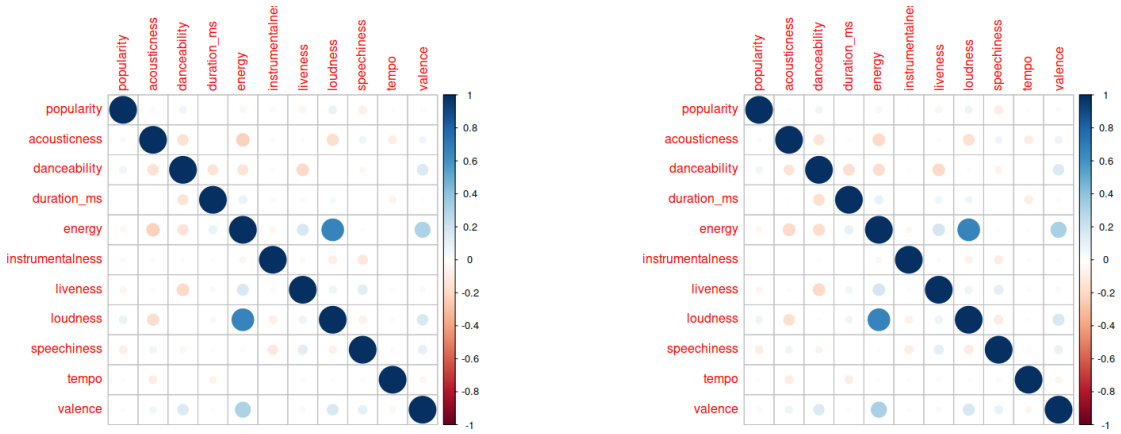


Figure 25: Comparison of eigenvalues in the Hip-Hop group

Due to that, there are almost not difference between the relationships of the variables. The correlation matrix (figure 26a) and the MCD matrix (figure 26b) are practically identical.

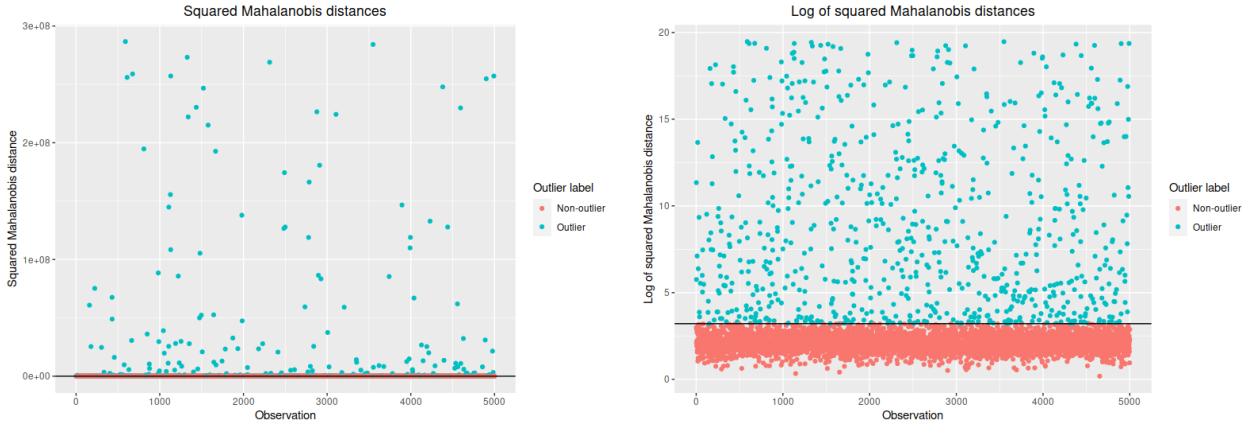


(a) Correlation matrix

(b) MCD correlation matrix

Figure 26: Correlation matrices comparison of Hip-Hop group

The critical point comes with the Mahalanobis distances, where a lot of samples (figure 27a) can be outliers. If the logarithmic distances are considered (figure 27b), there is not a number of points whose distances are so different than the mean. For this reason, **no points** are removed in this group.



(a) Squared Mahalanobis distances

(b) Log of squared Mahalanobis distances

Figure 27: Mahalanobis distances of Hip-Hop group

In this case, the PCP plot (figure 24) shows the patterns of all the points which could be considered outliers. As it can be seen, the point is in the `instrumentalness` variable, where most of samples have a small value near zero, in contrast with the outliers.

For this reason is not a good option removing these points, since these samples could correspond with hip-hop songs where some instruments appear, which is not very common. They are data from the tails of the distributions, rather than outliers.

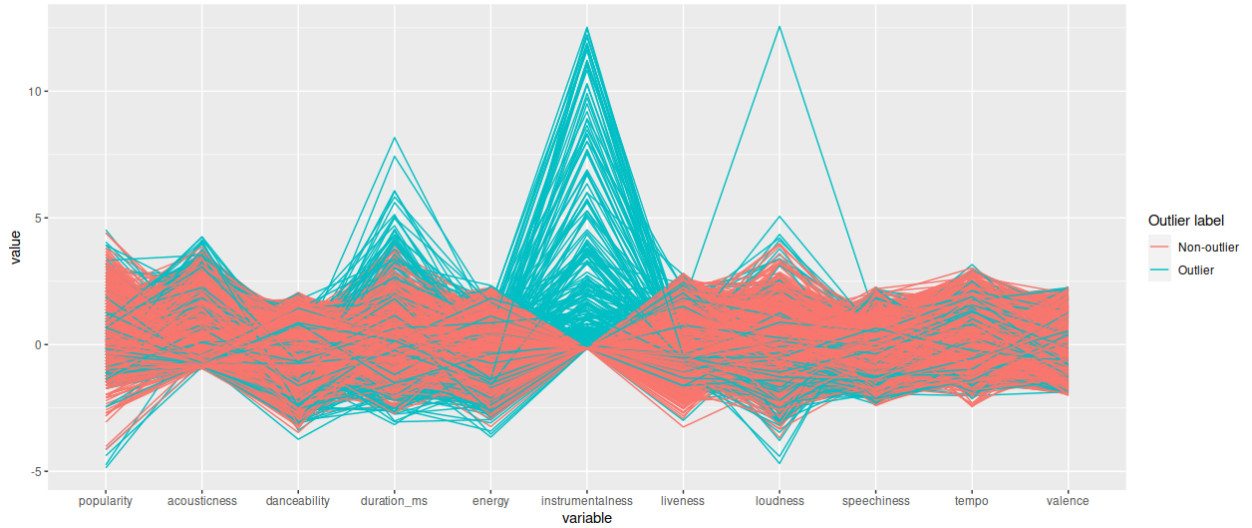


Figure 28: PCP plot with possible outliers in the Hip-Hop group

4.4. Outliers in Jazz group

In the last group the difference in the eigenvalues of the covariance matrices is not very significant, figure 29. It has been used $h = 0.95$.

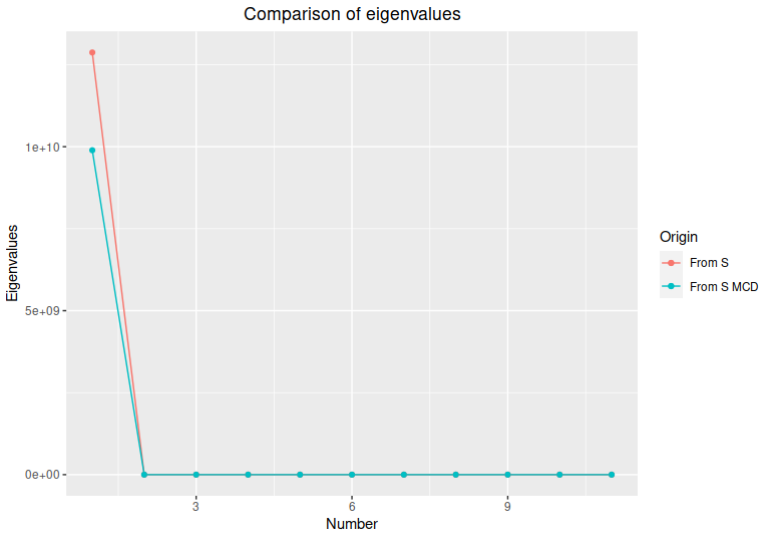
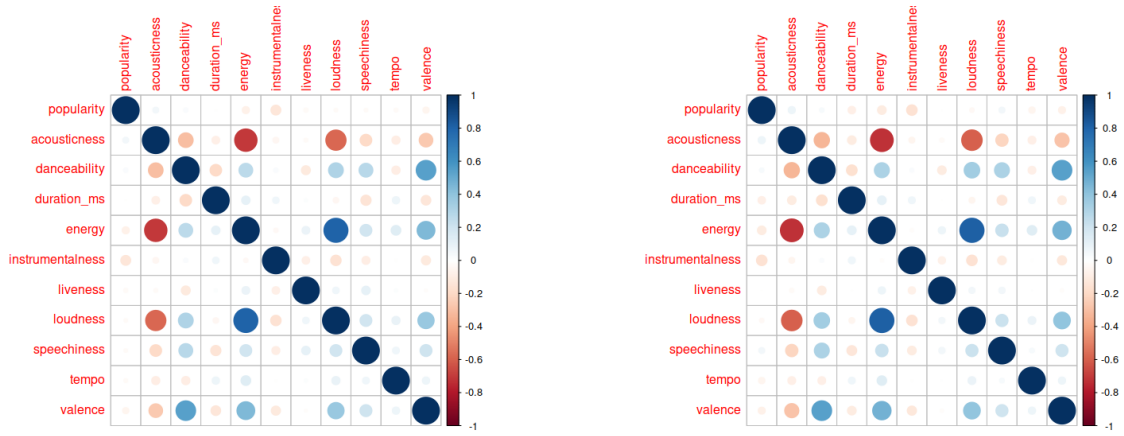


Figure 29: Comparison of eigenvalues in the Jazz group

For this reason, the relationships between variables plotted in the correlation matrices are almost the same, figures 30a and 30b.

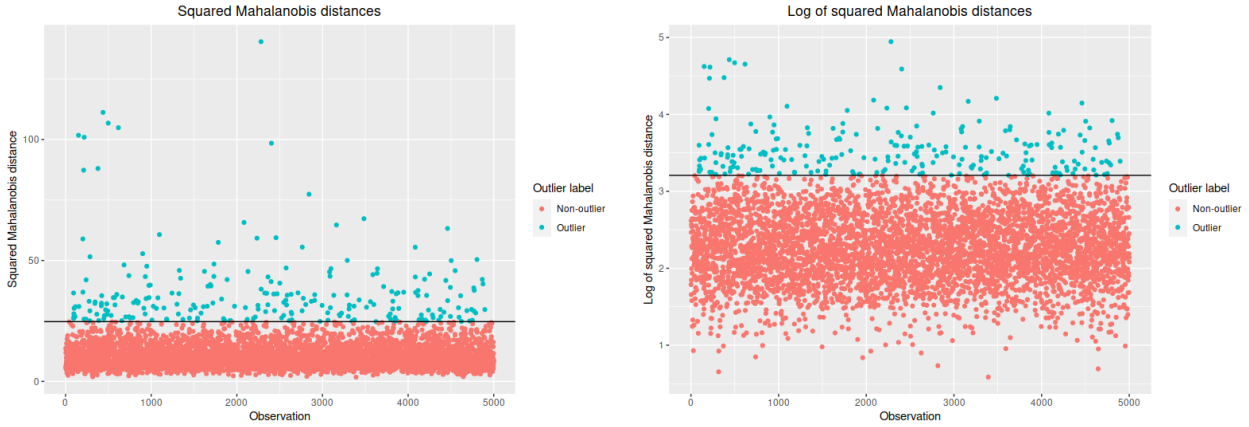


(a) Correlation matrix

(b) MCD correlation matrix

Figure 30: Correlation matrices comparison of Jazz group

In contrast, using the Mahalanobis distance criterion some points can be clearly differentiated from non-outliers. In this case there are nine points which are widely separated from the rest. For this reason, these **nine points** are considered outliers and removed from the dataset.



(a) Squared Mahalanobis distances

(b) Log of squared Mahalanobis distances

Figure 31: Mahalanobis distances of Jazz group

In figure 32 it is shown the patter of the selected points. However, it cannot be appreciated any clear difference respect the value of the variables using this plot.

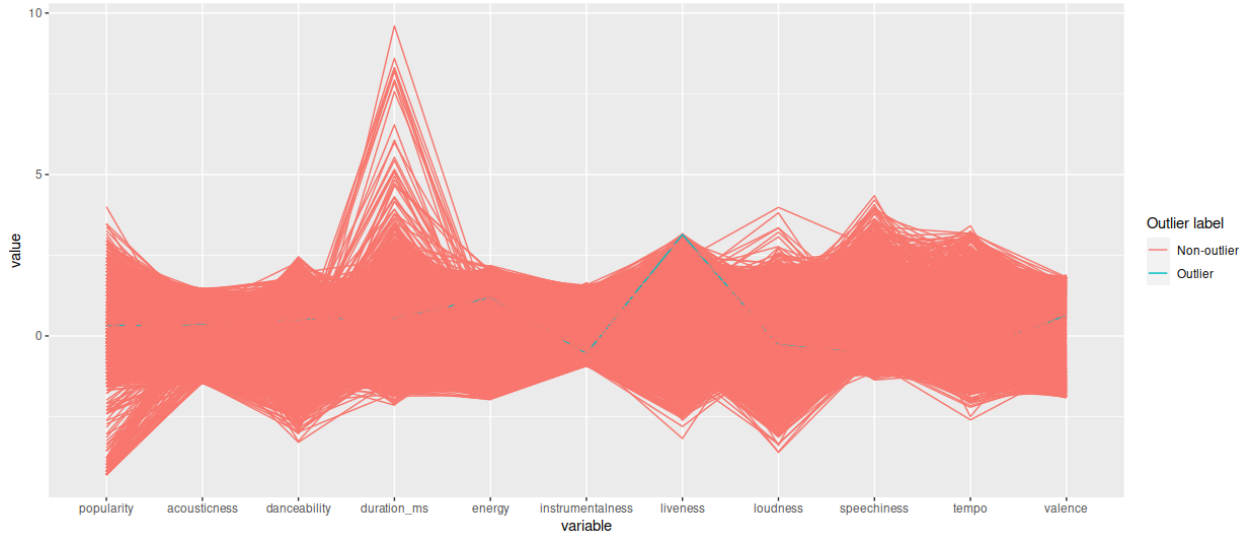


Figure 32: PCP plot with possible outliers in the Jazz group

5. Dimension Reduction Techniques

The idea is to obtain the main characteristics of the data based on the new variables obtained through transformations. In particular, it is interesting for knowing the presence of groups and/or outliers. In addition to, these new variables can be used in classification, since sometimes they improve the results respect the given predictors.

Mention it is used 11 quantitative variables, so there is going to be 11 transformed components.

5.1. Principal component analysis (PCA)

The idea is to find the components which represent better the variability of the data. The first point was looking for the variance explained by each component. In figure 33, it can be seen how the first component explains more than **40% of variance** in the dataset. Then there is a huge jump into the second one which only adds 10.6% more. Then the variance is slightly decreased along the rest of components.

With only **5 components**, it is possible to explain **76.77%** of the variance.

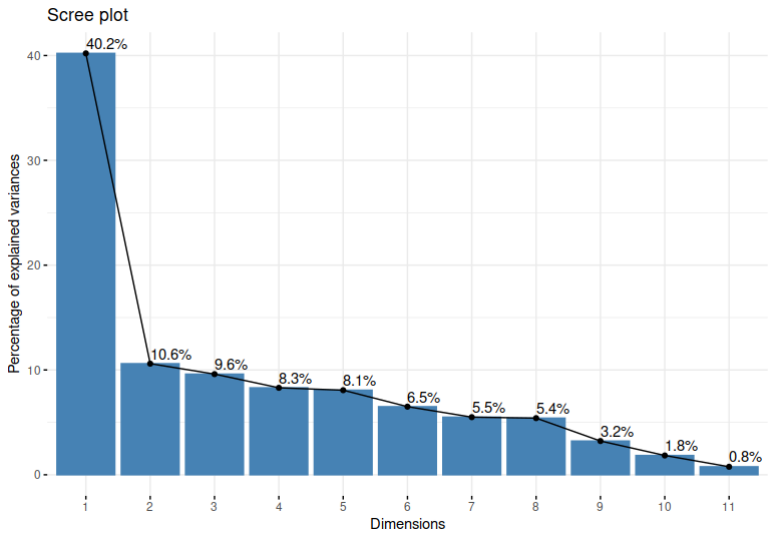


Figure 33: Percentage of explained variance by each principal component

However, looking into the data representation using the first and second principal components (a total of 50.6% percentage of variance), there is not a clear pattern (figure 34). The four groups are almost gathered in the same area. Nevertheless, it could be useful to distinguish some genres. For example, classical and hip-hop are separated, but respect the other three several points are mixed.

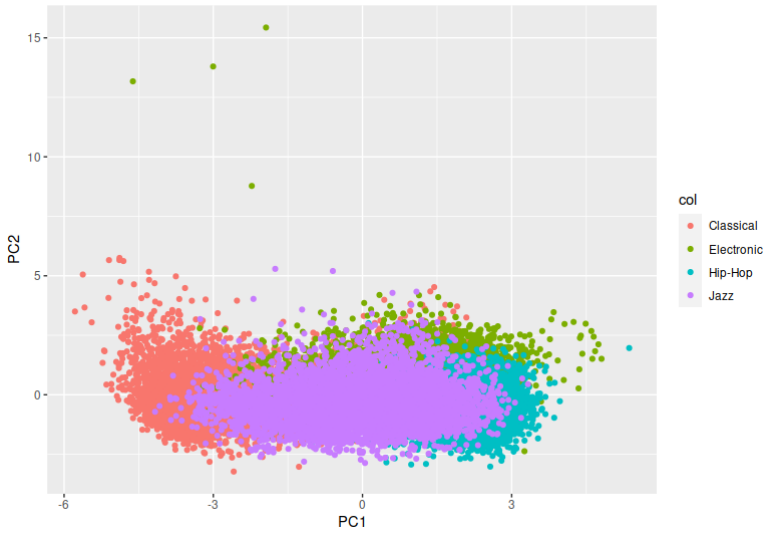


Figure 34: Data representation using first and second principal components

Mention four points from electronic group which are very separated from the rest. Selecting the first 5 components (76.77% of variance), it is shown a pairs plot in figure 35 where it can be seen how these points differ from the rest in all components. Due to that, they could be outliers of the electronic group.

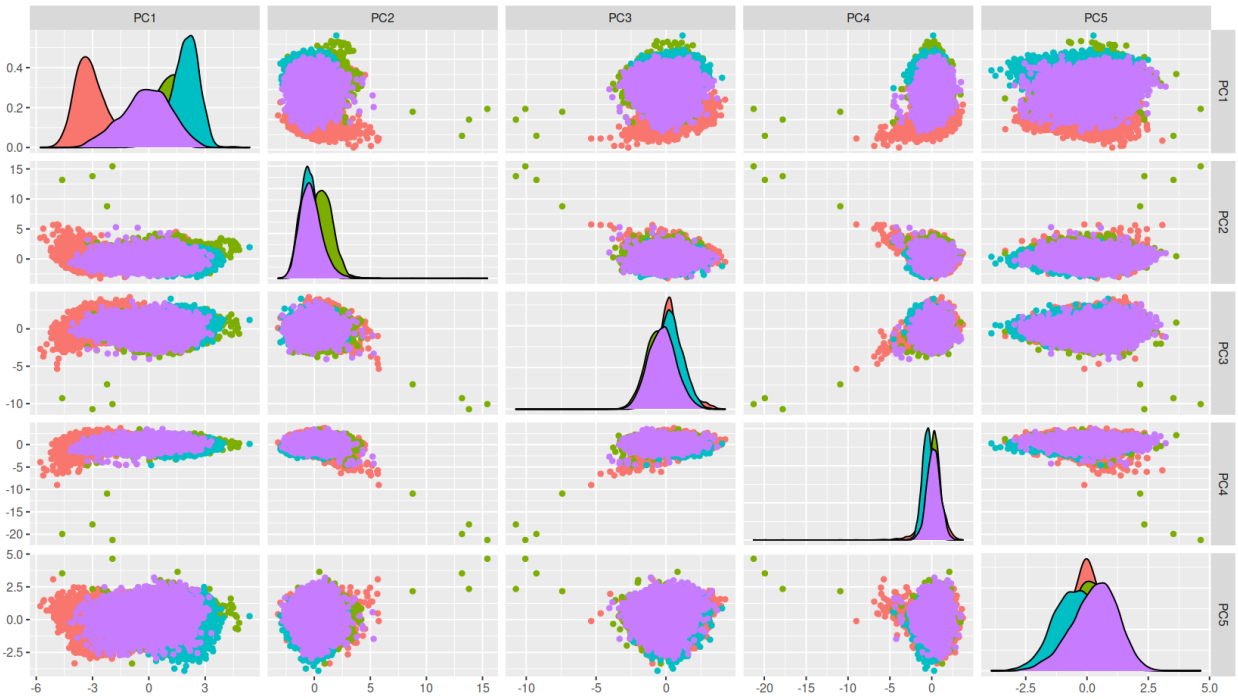


Figure 35: Pairs plot using the first five principal components (76.77% of explained variance)

It is important also notice how the groups are not well distinguished in any combination of PCs. The first one provides the best result, but as it has been commented, only differentiating classical group from the rest.

5.1.1. Interpretation of the main principal components

A critical step into dimension reduction is interpretating and knowing what the PC mean or represent, at least the most representative ones.

In this case, the **first PC** stands out from the rest differentiating the four groups. That is why this it implies this high percentage of variance explained.

In figure 36 the variables' weights of first principal component appear. It can be seen how it provides a more positive weight to songs more danceable and loud, and negative to presence of instruments or acousticness. That makes sense with figure 34, where hip-hop and electronics have positive PC1 values and classical negative. The first component could be a variable indicating energetic songs without presence of acoustic instruments.

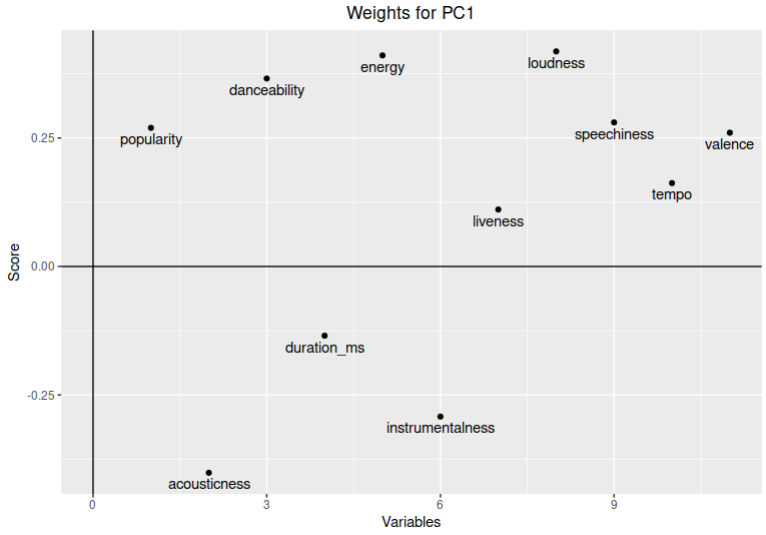


Figure 36: Variables' weights of first principal component

The correlation matrix between the first five principal components and the original variables is presented in figure 37. The first one represents with more weight (positive or negative) the variables in the sense explained before.

To highlight how the rest of components provide more or different weights to combinations not explored before. For example how the PC2 mainly classify by the variables with less weight in PC1 (`duration`, `liveness` and `tempo`) or PC3 providing a high positive score to `liveness`.

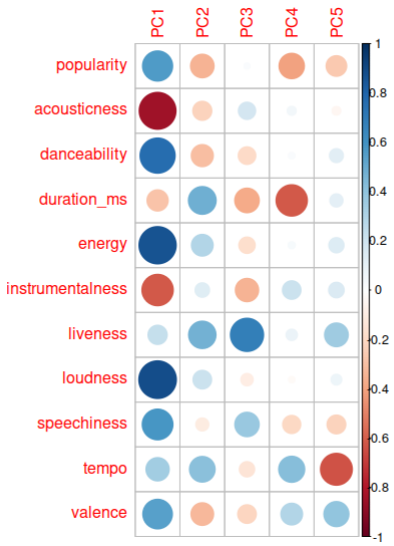


Figure 37: Correlation matrix between the variables and the first five principal components

5.2. Independent component analysis (ICA)

This technique is more aggressive since it employs components which are directly independent, instead of highly uncorrelated. Mention it usually provides a different result or combination than PCA, so that is the reason why it is useful to employ it.

Once the components are obtained, they are sorted by the negative entropy which represent. Mention in this case the first IC components does not correspond with the highest negative entropy, that is why in figure 38 it is included the number of the original index component. As expected, the first one implies the highest amount with a significant difference from the rest.

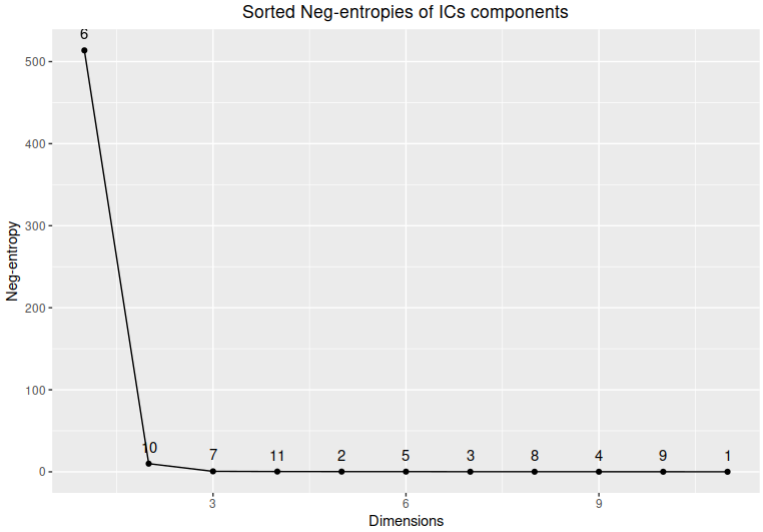


Figure 38: Independent components sorted by negative entropy

However, as with PCA, in figure 39 it can be seen how the groups are not well distinguished using the first two

IC with highest negative entropy (6 and 10 respectively). Mention the dispersion of points in the electronic group. As before, let's examine the pairs plot in order to decide if these ones could be outliers.

Figure 39: Data representation using first and second independent components

In the pairs plot of figure 40 only appear the **first five independent components** sorted by the negative entropy (figure 38). There have only been plotted five because as more were included, the scatter plots were being less informative, mixing most of the points from different groups.

Mention no IC separates well the points, but several strange samples of electronic group appear. Nevertheless, there are many, and the process of outliers have been already carried out, so this information was not considered.

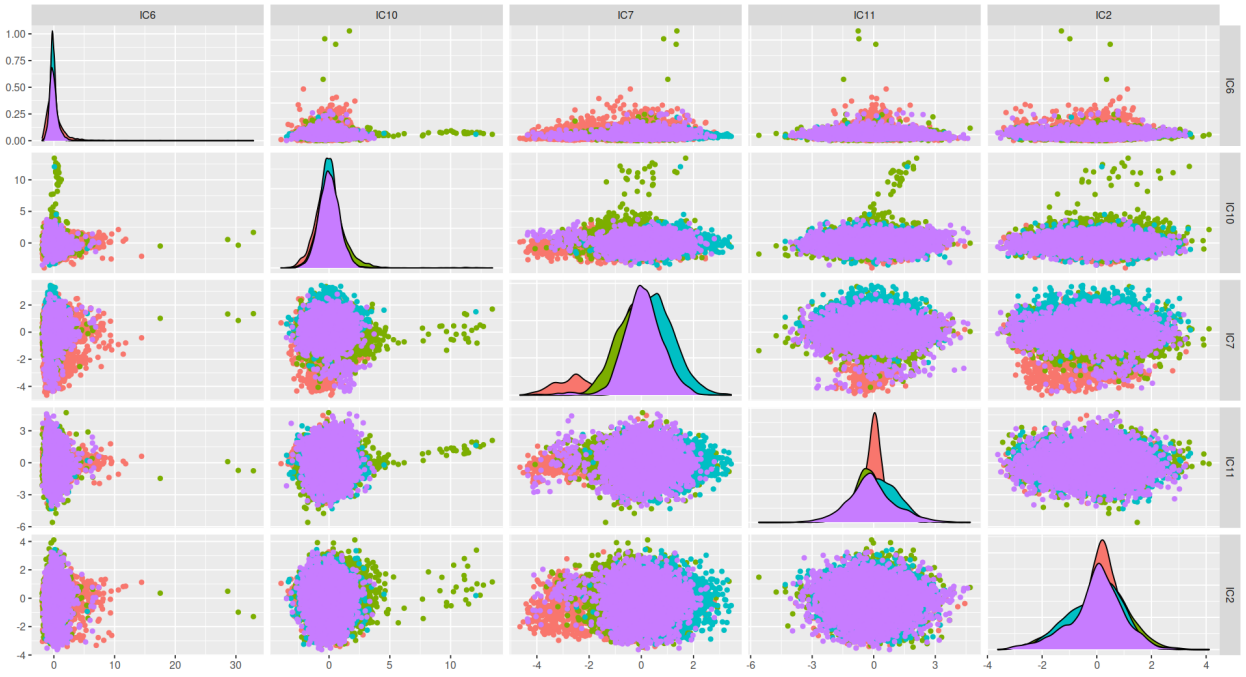


Figure 40: Pairs plot using the first five independent components

5.2.1. Interpretation of the main independent components

Unlike with PCA, here it is plotted the correlation matrix of all components in figure 41. It can be seen how first components are clearly related with a variable, while the last one represents more a combination. However, there are always a main one variable represented.

The **first IC** classifies by `duration`, but, as it can be seen in figure 39 is not a good parameter. Mention other IC which are closely related with a variable as IC7 with `popularity` or IC2 with `liveness`, and do not provide a good result distinguishing the musical genres.

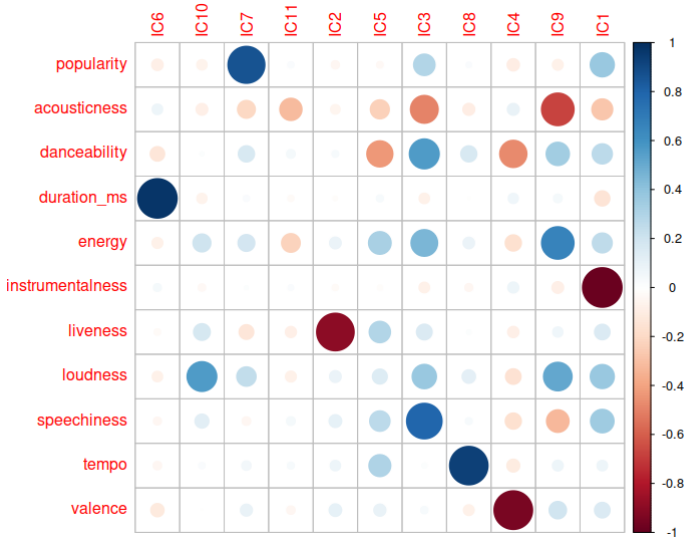


Figure 41: Correlation matrix between the variables and the independent components

Last picture, 42, provides an insight of the different results given by PCA and ICA. As it can be seen, there is not a clear relationship of one-by-one between principal and independent components. That is the reason each strategy could lead to different representations of the data, altough in this case no one was optimal.

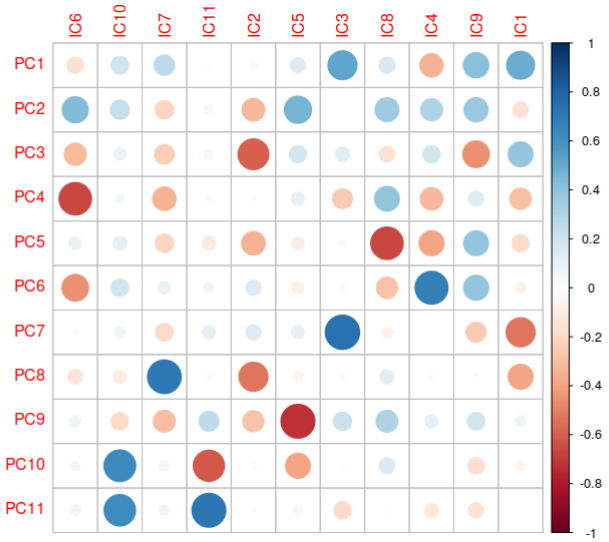


Figure 42: Correlation matrix between the principal and independent components

Nevertheless, mention similarities as the one given by IC6 and PC4. In the correlation matrix of the ICs, figure 41, it can be seen how IC6 is positive correlated with `duration`. In contrast, PC4 in the correlation matrix in figure 37 is negative correlated with this variable. That is the reason these two components have a high negative correlation.

6. Unsupervised classification

In this section, there will be performed different Unsupervised Learning methods in order to group the data samples according its category. The difference with respect to the following section in where there will be done Supervised Learning, is that the target variable *music_genre* will not be used in this case.

6.1. Estimation of K

The first question that must be asked in every unsupervised learning problem is: how many different groups there exist? Or in other words, what is the value of K? In the problem proposed in this project, it is well known that the there are four different categories corresponding to four different music genres: classical, jazz, electronic, and hip-hop music; but, as it has been previously stated, all the analysis performed in this section is done without taking into account the target variable. This means that the first step of this unsupervised learning section will be to estimate the correct number of groups.

In order to achieve this goal, there will be used two classical methods in cluster analysis: the Elbow method and the Silhouette method.

6.1.1. Elbow method

In the first case the method to estimate the optimal number of K will be the widely used Elbow method. This option consists on running k-means using the data for a range of K values and extract the total within-cluster sum of squares value from each mode. After this, the sum of squares is plotted and the goal then is to find the "elbow" of the function, that value of K will correspond to the optimal K. Given that it is already know that K is 4, the range used for this plot will be from 1 to 10.

Another aspect important to be mentioned is the fact that the dataframe used for this plot is scaled. This is very important because differences in the scale are very sensitive when it comes to the k-means method. This is specially important in the problem addressed here because almost all the variables have values between 0 and 1 while some of them are unbounded, having values of even hundreds of thousands as in the case of the duration in miliseconds.

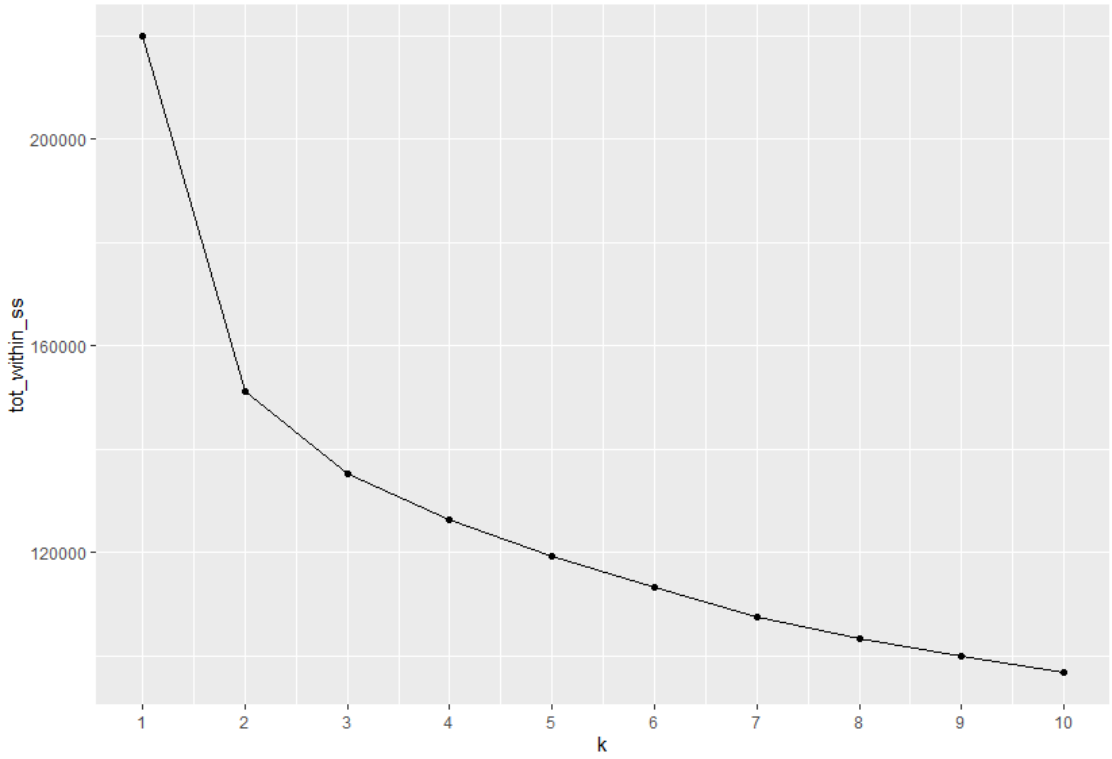


Figure 43: Elbow method plot

According to the previous plot, the optimal number of groups that should be used in this cluster analysis would be 2 because it is the value of K where the elbow is located. As it is already known, this value of an optimal K of 2 is not corresponded with the real value of 4. This is a direct of consequence of something that have been seen all along the project: the categories are not very well differentiated. This fact will make the formation of groups more difficult, but this problem will be addressed deeply in the next and last section.

6.1.2. Silhouette method

Another method dominant when it comes to the determination of the number of clusters is the Silhouette analysis. It consists on calculating how similar each observation is with the cluster it is assigned relative to other clusters. The silhouette of a point is given by:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}}$$

And, as it can be seen it ranges from 1 to -1, in where a positive value means that x_i is well matched to its own cluster. In the Silhouette analysis plot, it can be seen the average of this silhouette.

Here is the plot:

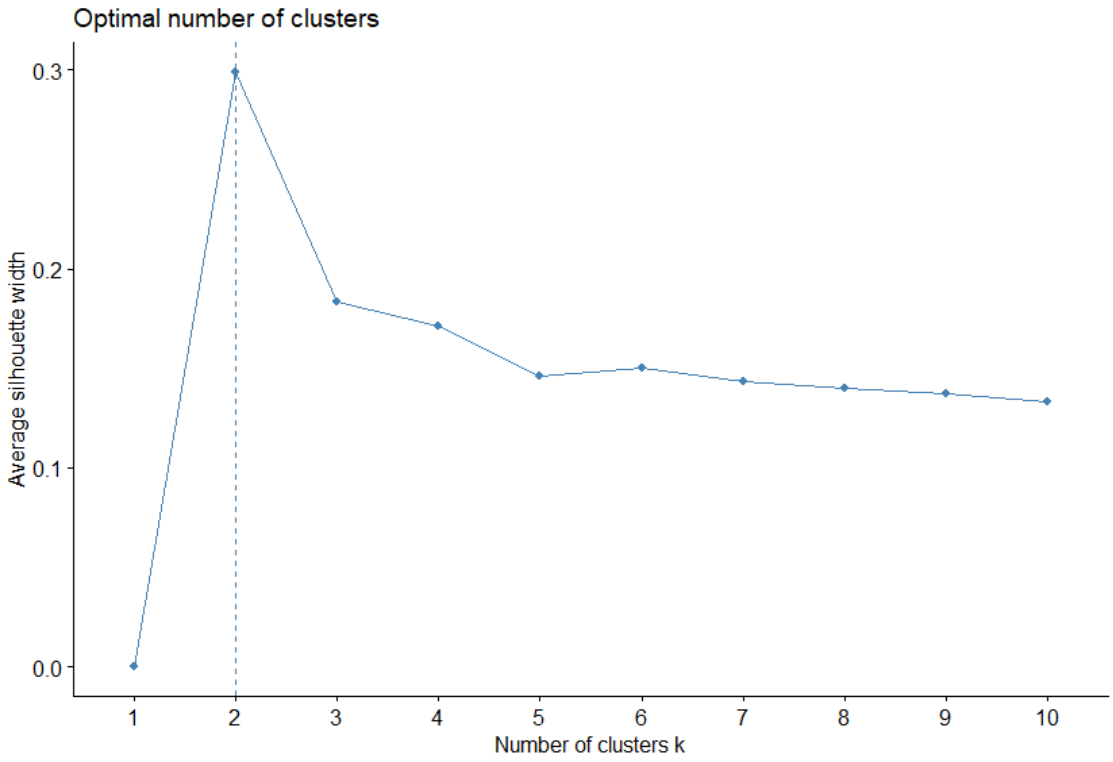
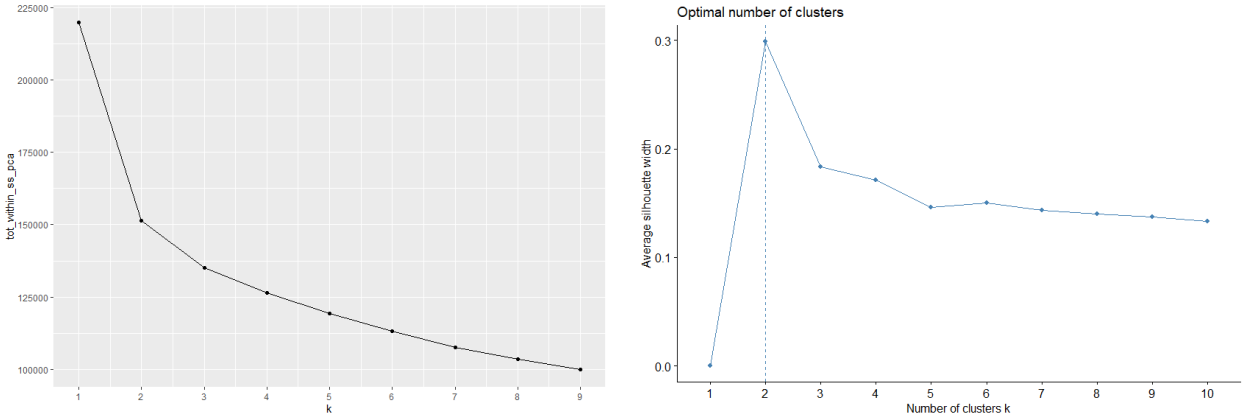


Figure 44: Silhouette method plot

As the Elbow method already indicated, the optimal value of K estimated by the Silhouette method is 2. As it was previously said, this is a consequence of having categories that are difficult to differentiate.

Here is also the repeated plots but using the principal components calculated in the previous section.



(a) Elbow plot using PCA

(b) Silhouette plot using PCA

Figure 45: Repeated plots using principal components

As expected, using the principal components of the data instead of the original data scaled, the results obtained are the same, with an optimal K of 2.

Despite the fact that these two methods "recommend" to use a K value of 2, the unsupervised learning methods that will be performed in the following subsections will be executed with the true value of K that is 4.

6.2. Partitional clustering

Having already estimated the number of clusters present in our data, it is the moment of proceeding with the grouping. The first option that will be presented in this project will be the use of the most popular clustering algorithm: **the K-Means algorithm**. This algorithm is characterized for being very efficient when clustering large data matrices. On the other hand, one of the main disadvantages that it has is that it does not find clusters of arbitrary shapes.

In order to use K-Means in R, it will be used the `kmeans()` function contained in the `stats` library.

One way to see how well the classification went is through the confusion matrix, the problem is that, given it is an unsupervised learning problem, it is not possible to know the corresponding genre to each of the clusters. It will be assumed that each of the groups will correspond to the genre most typical in that group. This is the confusion matrix generated:

Music genre	Classical	Electronic	Hip-hop	Jazz
Classical	4418	337	7	235
Electronic	164	2575	585	1672
Hip-hop	7	365	4247	381
Jazz	991	583	487	2930

Table 2: Confusion matrix using K-Means clustering.

The most remarkable aspect that can be highlighted seeing this confusion matrix is that the genres of *Classical music* and *Hip-hop music* are the ones better identified when using K-Means, classifying correctly more than

4000 samples of the ∼5000 existing samples for every music genre. This could have been expected because, as it has been seen all along the project, these are the genres better differentiated from the rest classes present in the dataset.

6.2.1. Visualization

The library `factoextra` provides a very useful function: `fviz_cluster()` that generates the following plot:

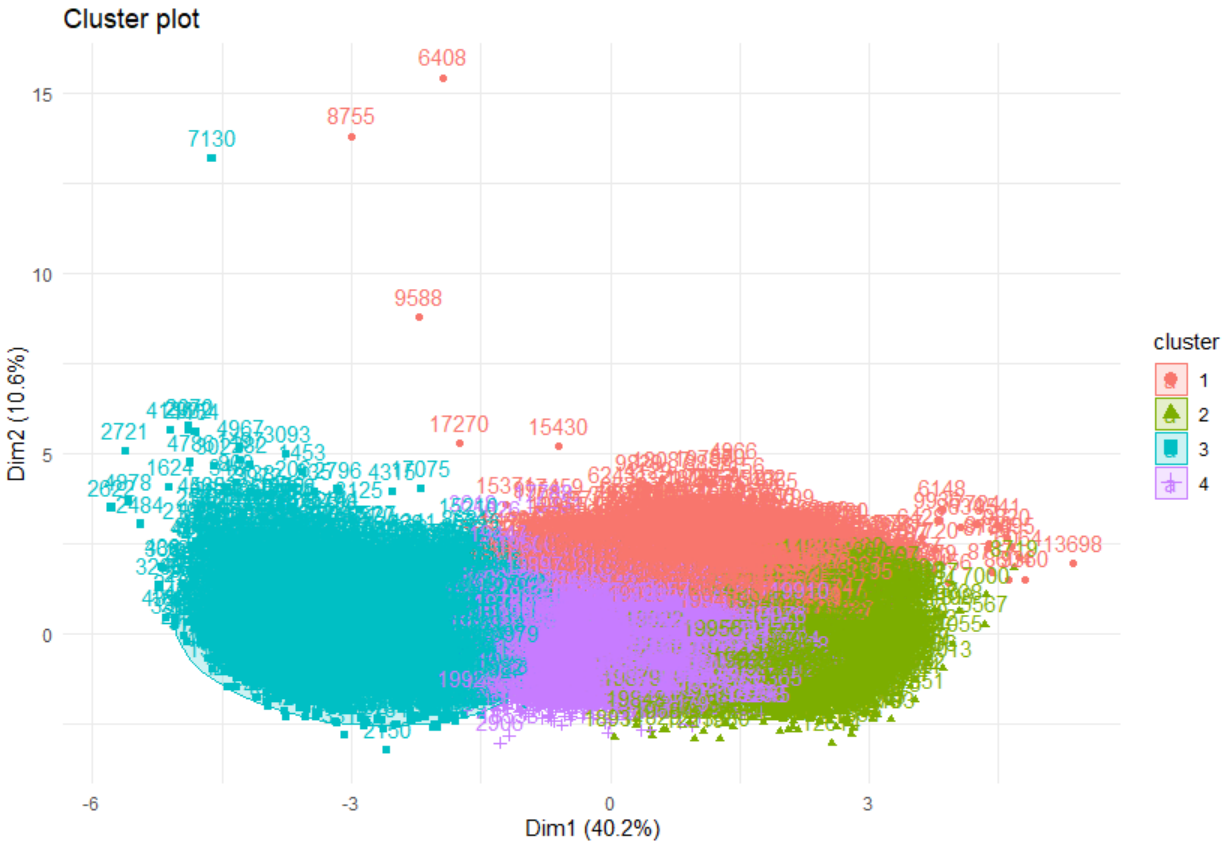


Figure 46: Confusion matrix of K-Means clusters

Here it can be seen a scatter plot where the x and y axis are the two first principal components of the data that, together, they explain more than the 50% of the variance of the data. In that scatter plot are all the 5000 samples of the dataset coloured according the group assigned by the K-Means algorithm.

6.3. Hierarchical clustering

The idea of Hierarchical clustering approach the differentiation of groups by dividing or agglomerating groups of samples until it is obtained the desired number of clusters. This division or agglomeration is done trying to minimize the distance between samples of the same group. The most popular distances are the following:

- Single linkage
- Complete linkage
- Average linkage

- Ward linkage

This method can be implemented in R thanks to the `hclust()` function provided by the `stats` library, selecting the type of distance with the `method` attribute.

The problem encountered is that with almost all the methods we encountered the similar problem, a very imbalanced clustering like the one that can be seen in the following table using the *Complete* linkage:

Cluster	1	2	3	4
Number of samples	19638	337	6	3

Table 3: Number of samples per class using Complete Linkage

However, a method that procuded a more accurate result was the *Ward* linkage. The reason of this is that this method provides with solutions close to the ones given by K-Means. This is the resulting number of samples for each cluster:

Cluster	1	2	3	4
Number of samples	5548	7556	2682	4198

Table 4: Number of samples per class using Ward Linkage

This is the resulting confusion matrix:

Music genre	Classical	Electronic	Hip-hop	Jazz
Classical	4391	475	4	127
Electronic	105	3432	302	1157
Hip-hop	18	1365	3576	41
Jazz	1034	2284	316	1357

Table 5: Confusion matrix using Hierarchical clustering with Ward linkage.

In this table, it is possible to observe that Classical music is still properly classified, correctly identifying more of the 80% of the samples corresponding to that category. Observing the clustering for the groups that correspond to Electronic and Jazz music, the classification is not too bad, correctly classifying around 3500 song that belong to these methods. However, when it comes to Jazz music, the classification is not good at all, as it is classifying incorrectly more than 3500 samples, grouping them as Classical or Electronic in most of the cases.

6.3.1. Visualization

As it has been already depicted, hierarchical clustering starts from a unique group containing all the samples of the dataset and divides this group in order to minimize some distance metric. This consequent division produces the dendogram. Using the library `dendextend`, the following plot of the dendogram can be generated, obtaining the following result using Ward linkage metric:

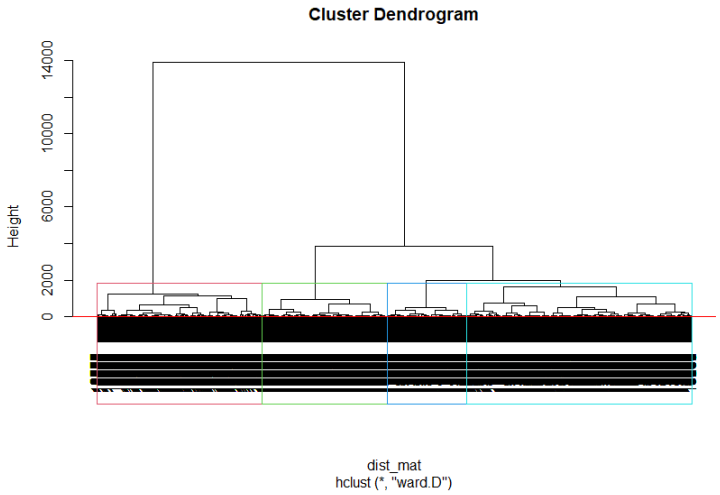


Figure 47: Cluster dendogram and groups classified

The goal of this plot is to see how the dendogram is divided creating the clusters and observing which branches of the plot belong to each of the groups, the problem is that due to the elevated number of samples, this is not clear at all. In any case, it is good to have an idea about how this hierarchical clustering works.

7. Supervised classification

As the initial descriptive analysis showed, it is not easy to distinguish among the different groups. Although some characteristic features could be extracted for each music genre, they are not enough to classify the music pieces just by their analysis.

In this way, some methods will be tried for solving the problem of predicting the music genre from the song features. As according to the “No Free Lunch” theorem, no method dominates all others in every possible dataset, different alternatives on supervised classification should be tried. Their performances will be compared and then the best one according to this process will be chosen and used for new predictions. For this comparison, the whole dataset is divided into two partitions, determining the training and test datasets.

These methods of supervised classification will use the Bayes rule as it is proved that, under a probabilistic framework, a method using it minimizes the error rate. Then different classifiers that estimate the necessary conditional probabilities are proposed.

7.1. K-Nearest Neighbours (KNN)

This classification method uses a hyper-parameter, k (number of neighbours), which needs to be estimated. For this purpose, a value range between 1 and 30 is explored. By performing leave-one-out cross-validation with the training dataset partition, this parameter is chosen so that the minimum error rate is obtained.

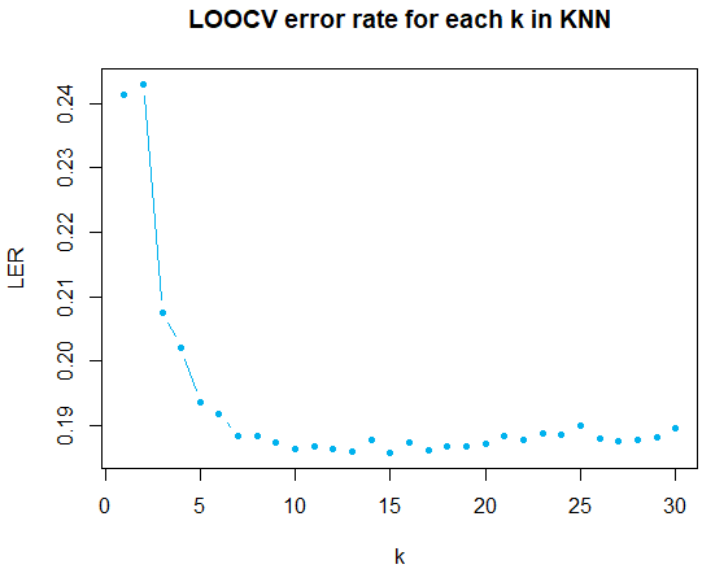


Figure 48: Selection of hyper-parameter K for KNN.

In the end, this parameter is set equal to 15 and used in the final model to classify the test dataset. The obtained test error rate is 0.1931.

7.2. Methods based on the Bayes Theorem

Afterwards, this supervised classification task is performed with different methods using the Bayes Theorem. These are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Naive Bayes (NB).

In each case, the parameters of the model are estimated with the training sample. Afterwards, the models are applied on the test dataset, obtaining a test error rate equal to 0.2168 for LDA, 0.2181 for QDA and 0.6966 for NB. Then, none of them is performing better than KNN.

7.3. Logistic regression

In the end, the logistic regression method is tried on the project problem. There is also a process of estimating the model parameters with the training dataset.

In this case, the backward selection method is applied on the model too. In this way, it is tried to improve the test error rate by deleting predictors without discriminatory power and not having much influence in the classification. Nevertheless, a similar performance as without it is obtained, so it is discarded.

The test error rate for this method is 0.2075.

7.4. Performance comparison

Now, it is time to select the best way to solve the proposed problem. As it can be seen from the results comparison, the chosen method for solving the problem of supervised classification will be KNN, as it is the one obtaining

the lowest error rate in the test performed.

Some other alternatives were tried, too. No improvement was appreciated by discarding predictors that were highly correlated (as it is the case of energy and loudness) or could be considered as non-relevant after the descriptive analysis. The new variables obtained by PCA and ICA were also introduced into the classifiers, but the previous KNN performance was not even approached (although the error rate was improved a bit for the rest of classifiers). Some of the most promising results from these last tries are collected in the following table along with the ones obtained with the original predictors:

Classifier	KNN	LDA	QDA	Logistic reg.
Original variables	**0.1931**	0.2168	0.2181	0.2075
PCA	0.2025	0.2110	0.2158	0.2061
ICA	0.2015	0.2110	0.2157	0.2059

Table 6: Classifiers' error rate comparison.

After the best classification method has been chosen, it could be interesting to analyze the corresponding confusion matrix for the given problem. As it can be seen in the table below, Classical and Hip-hop music are the best differentiable music genres. This is a trend that has been spotted all along the project and could result evident just be listening to one song from each genre. In this case, logic reasoning coincides with statistical behaviour, which is something that does not always happen.

Music genre	Classical	Electronic	Hip-hop	Jazz
Classical	1305	86	2	111
Electronic	34	1077	135	221
Hip-hop	0	45	1475	35
Jazz	126	258	100	986

Table 7: Confusion matrix using KNN.

It can be concluded that KNN is the best method for supervised classification. One of the biggest disadvantages of this method arises when dealing with unbalanced datasets. Nevertheless, as this project deals with a balanced dataset, the above-mentioned disadvantage is overcome. Then we can take advantage of having to adjust just one hyper-parameter and the resulting computational cost reduction for solving the supervised classification problem.