University of Pisa

Department of Informatics

Master’s Degree in Data Science and Business Informatics

**DM1 - project**

**Data Understanding, Preparation**

**& Clustering**

Authors:

Mohamed Arafaath Sathik Basha

Vincenzo Rocchi

Cristian Ferrara

Sommario

[Index 1](#_Toc120488103)

[Data Understanding 2](#_Toc120488104)

[Data semantics 3](#_Toc120488105)

[Distribution of variables and statistics 4](#_Toc120488106)

[Dataset preliminary analysis 5](#_Toc120488107)

[Data quality 6](#_Toc120488108)

[Missing values: 6](#_Toc120488109)

[Variable transformation: 7](#_Toc120488110)

[Correlation matrix and eliminated variables 7](#_Toc120488111)

[Outliers 8](#_Toc120488112)

# Data Understanding

“The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats.” We are looking at part of the RAVDESS original dataset, more precisely at the voice-only one. The idea that lays behind the construction of this dataset is due to the changes occurred in the emotion studies during the last decade, a lot has changed in how we see and treat emotions and so are the studies associated with the topic. The importance of a reliable and validated record of expression of emotions is crucial to the integrity of the studies focused on vocal recognition or more in general sound classification and recognition.

# Data semantics

Analyzing it from up close we can clearly distinguish two parts, one containing the details of the recordings, the details of the actor that recorded them and the technical audio data, the other containing the extracted statistical data used for: zero-crossing rate, Mel-Frequency Cepstral Coefficients, spectral centroid, and the stft chromagram. The first half represents the 40% of the total data and the other the remaining 60%. Starting the analysis from the non-statistical part the “*modality*” represents the type of file and in this case, there are audio only ones, but they can be further distinguished by the “*vocal\_channel*” attribute that specifies if the recording was spoken or singed. The phrases utilized by the 24 actors (identified by being assigned a number from 1-24 under the “*actor*” attribute and further divided by “*sex*” in F or M) are “Kids are talking by the door” and “Dogs are sitting by the door” the one used is specified under the “*statement*” attribute. The phrases are repeated 2 times for each actor the order can be seen under the “repetition” attribute. The most interesting part comes when looking at the “emotion” and “emotional\_intenisty” attributes that indicate respectively the emotion that is interpreted by the actor and the intensity at which that emotion is represented. Those attributes alone though are not useful enough, but combining them with the statistical data describing the audio waves, make us seek for a correlation path between the emotion and the small changes in the spectrum of the audio recording. Such features are already used for various audio self-recognition techniques, the “MFCC” or Mel-Frequency Cepstral Coefficients is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency and is used in music information retrieval applications such as genre classification, audio similarity measures, etc. The “SC” or spectral centroid has many applications in audio classification and music classification and it’s used in processing of audio signals, especially music genre classification. The “STFT” or Chroma STFT value of an audio basically represent the intensity of the twelve distinctive pitch classes that are used to study music. They can be employed in the differentiation of the pitch class profiles between audio signals.

A quick recap of all the attributes and their definitions:

**Descriptive part**:

1. **modality** (audio-only)
2. **vocal\_channel** (speech, song)
3. **emotion** (neutral, calm, happy, sad, angry, fearful, disgust, surprised)
4. **emotional\_intensity** (normal, strong). NOTE: no strong intensity for the 'neutral' emotion
5. **statement** ("Kids are talking by the door", "Dogs are sitting by the door")
6. **repetition** (1st repetition, 2nd repetition)
7. **actor** (01 to 24)
8. **sex** (M, F)

**Technical part:**

1. **channels** (number of channels; 1 for mono, 2 for stereo audio)
2. **sample\_width** (number of bytes per sample; 1 means 8-bit, 2 means 16-bit)
3. **frame\_rate** (frequency of samples used (in Hertz))
4. **frame\_width** (Number of bytes for each frame. One frame contains a sample for each channel.)
5. **length\_ms** (audio file length (in milliseconds))
6. **frame\_count** (the number of frames from the sample)
7. **intensity** (loudness in dBFS (dB relative to the maximum possible loudness))

**Statistical part:**

1. **zero\_crossings\_sum** (sum of the zero-crossing rate)
2. **'mean', 'std', 'min', 'max', 'kur', 'skew'** (statistics of the original audio signal)
3. **mfcc\_ 'mean', 'std', 'min', 'max'** (statistics of the Mel-Frequency Cepstral Coefficients)
4. **sc\_ 'mean', 'std', 'min', 'max', 'kur', 'skew'** (statistics of the spectral centroid)
5. **stft\_ 'mean', 'std', 'min', 'max', 'kur', 'skew'** (statistics of the stft chromagram)#my initial idea of correlating the stat part is probably wrong or I did it too quickly im gonna work on it better tomorrow morning already sent you a message.

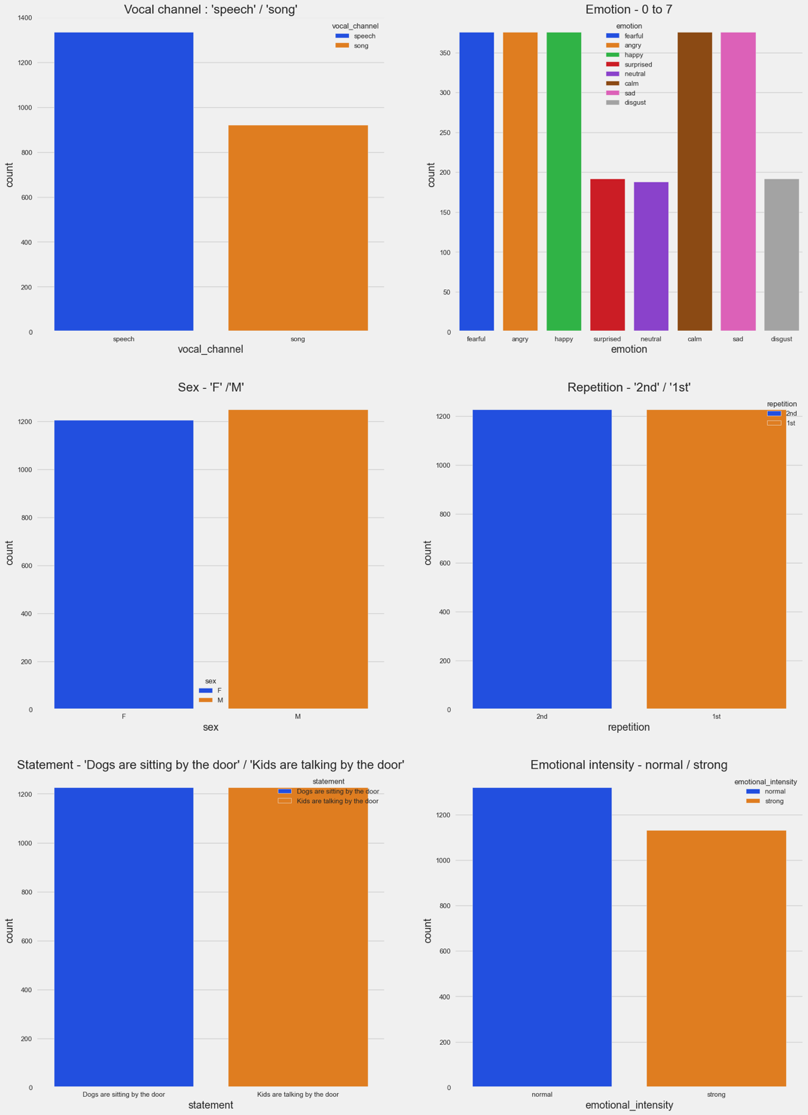
# Distribution of variables and statistics

To better understand the dataset, we decided to visualize the distributions of all variables through histograms and boxplots, that are presented below, no clear pattern emerges from this first visualization except for the vocal channel and the emotion attribute that show less presence of different characteristics throughout the test.

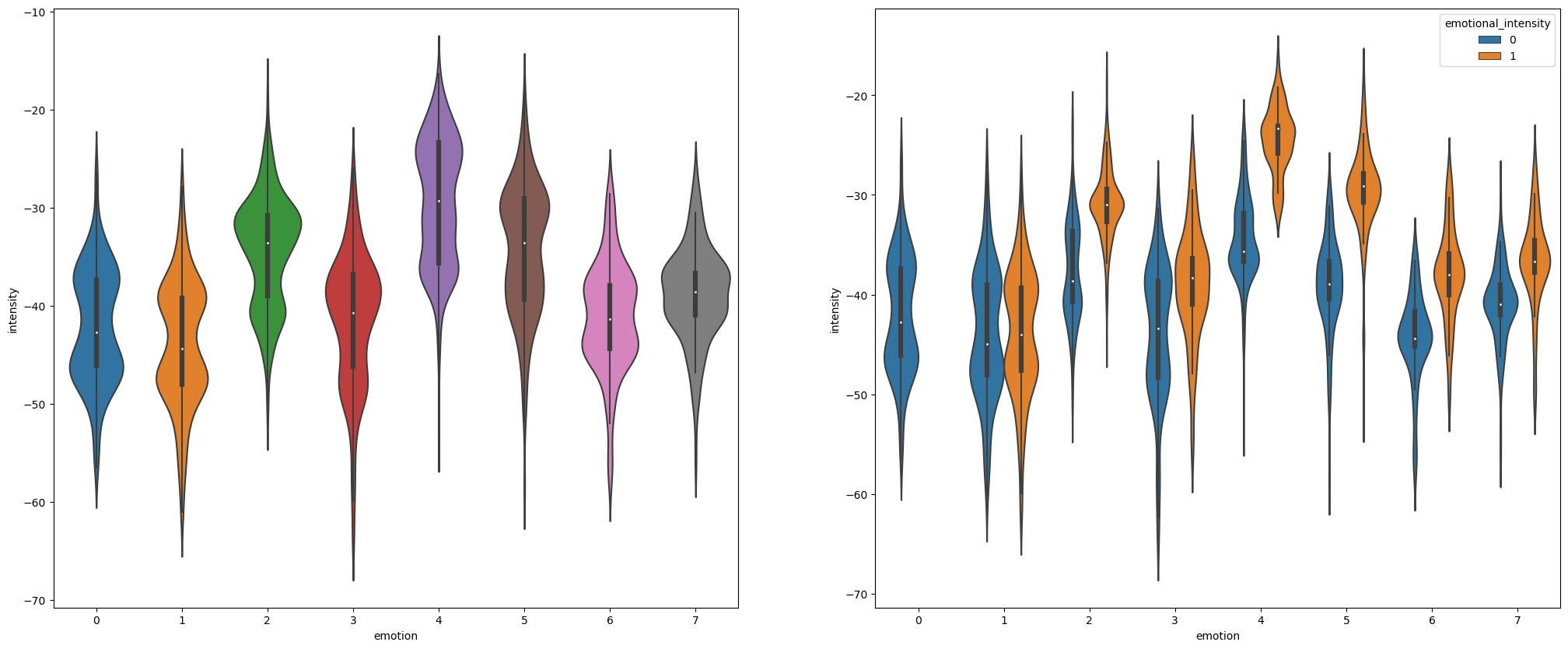
**FIGURES 1. HISTOGRAMS OF THE DISTRIBUTION OF**

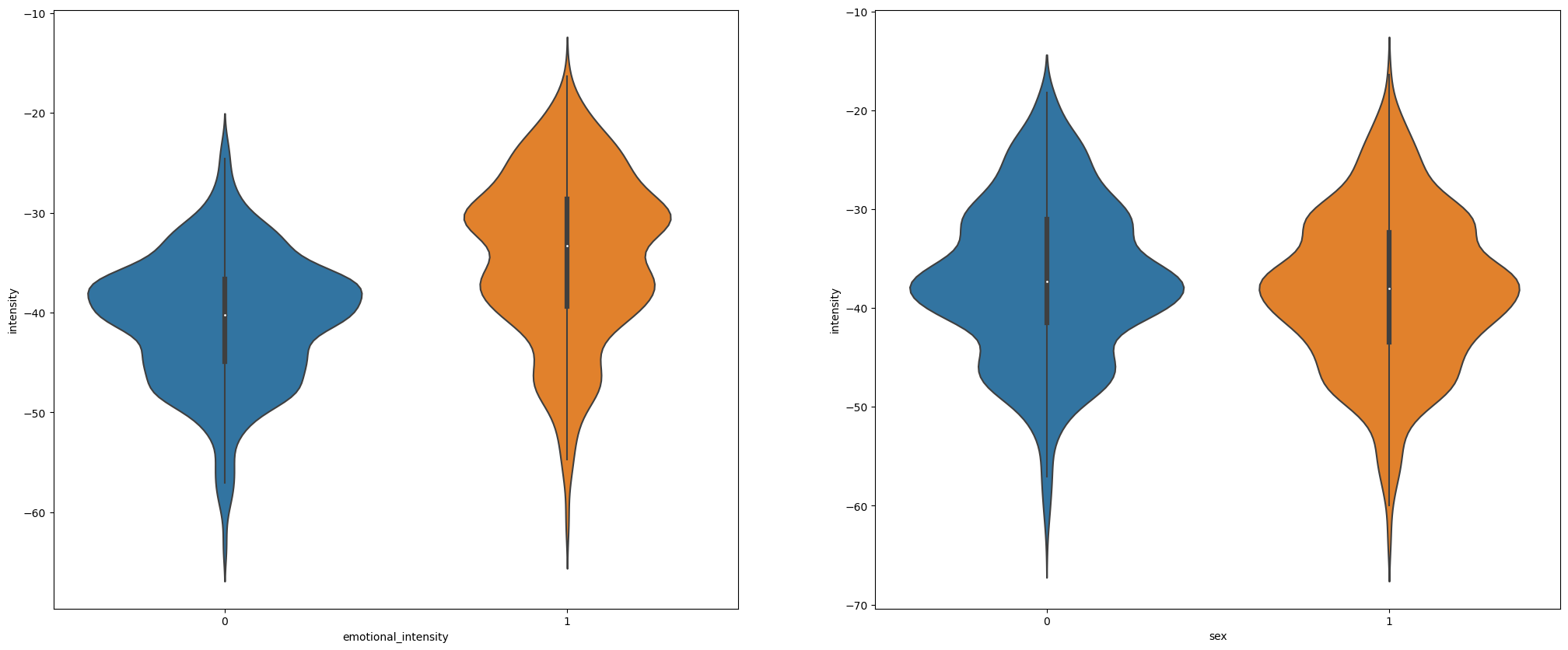
**1. NUMERICAL VARIABLES 2. CATEGORICAL VARIABLES**

A picture containing text, dog

Description automatically generated

Since we concluded that the most interesting attributes to plot mainly against are the intensity and the emotion one, we created some violin plot that helps us visualizing rapidly the trends between those 2 attributes and the other chosen: for the first one (emotion) we decided to plot against emotion, emotional intensity and sex. The emotion is mapped as follows (0 = neutral, 1 = calm, 2 = happy, 3 = sad, 4 = angry, 5 = fearful, 6 = disgust, 7 = surprised), the intensity is represented negatively relative to the maximum possible level of loudness in negative dB, it works in exactly the same way as the positive representation, but with every negative 10 dB simply indicating a factor of 10 LESS. So, -10 dB is 1/10th that of 0 dB, -20 dB is 1/100th of 0 dB, -30 dB is 1/1000th of 0 dB and so on again. We can conclude that when people are angry, fearful and happy the intensity can be much higher than the other feelings and the probability of a more intense sound is much higher; in the disgusted and surprised state the intensity is generally on pair with the neutral and calm feelings although the mean is higher, it presents a shallower 1.5x interquartile range and for so a higher probability of it being at a higher level than the neutral and calm state, making it more distinguishable and less situational; when they present a sad expression the intensity can be much lower than the other emotions but the probability of it being at a normal level is still higher than it being less intense and this makes the sadness the less distinguishable emotion of the bunch. Comparing the intensity and the emotional intensity (0 = normal, 1 = strong) we can see that the normal intensity has a broader range, a lower mean and a shallower inter quartile range, the probability of it being not so loud is much higher than the strong emotional intensity and so is less variable and more distinguishable. Next to the first plot is the one that merges the first two and let us comprehend how the emotional intensity also plays a big role in the definition of the loudness and the ability to recognize a felling by the loudness of the actor. The last plot is the sex against the intensity and we can clearly see that there are only small variation between male and female and the correlation with the intensity of the statement.

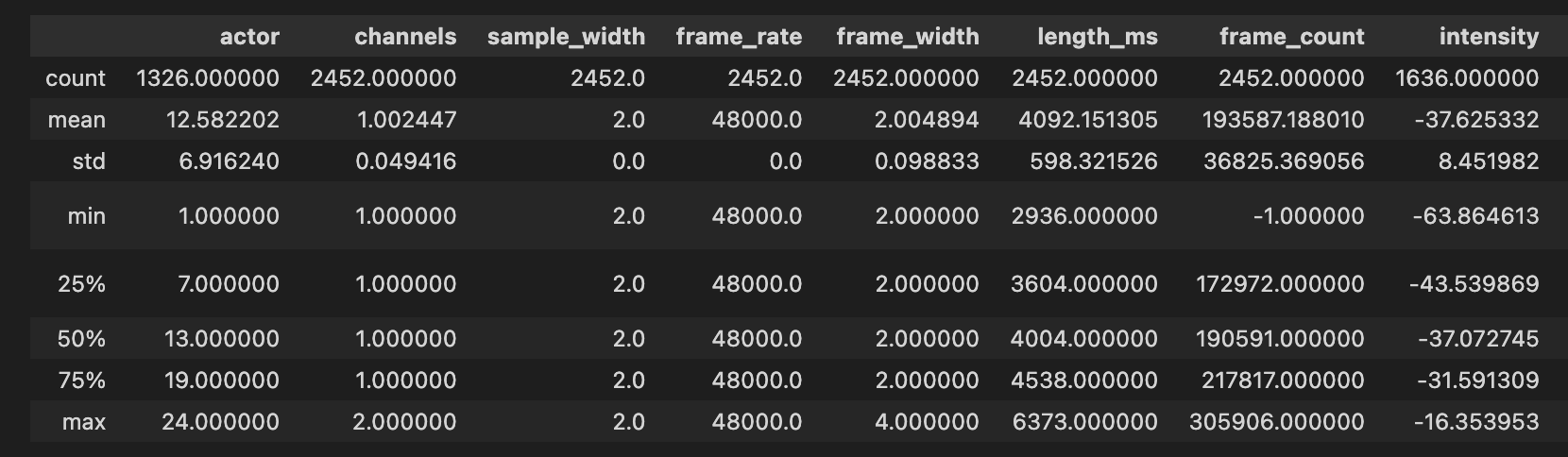




# Dataset preliminary analysis

The “RAVDESS” data set we are working on is composed by 2452 rows and 38 columns, the first 14, as said, are the one that not make an in depth statistical analysis on the more technical audio part, those are the ones that we’ll be analyzing as in the first part of the report and these are also the ones that require an in depth data quality assessment. Doing so we’ll address the data quality part of the essay and make our lives much easier finding the purpose of the analysis itself. Starting with a quick look on the statistical means of the numerical data:

Table 1 - statistical means of the dataset

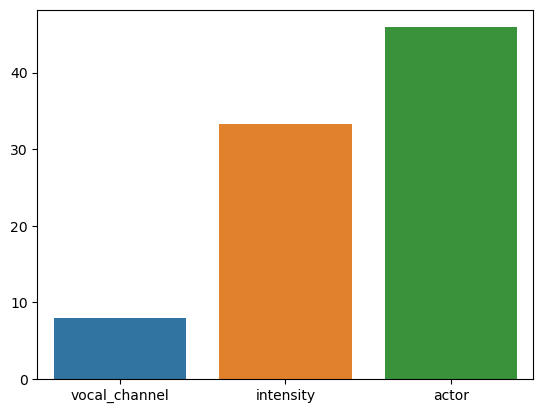


The values extracted, as at this stage, are: firstly the count of the values itself and this first data already tells us that some of the attributes lack a lot of records and that we’ll be dealing with a massive amount of missing values later on, not all the

attributes are full of data and some are missing as much as 50% of all the records done. The min/max let us take a quick snap on the length and the extremities of the values in the records itself. The std deviation and the mean help us further identify mismatches in the distribution of our data, by comparing the std value with the mean we can see that most of the data does not vary by a great amount and that the distribution rests accordingly. Utilizing exclusively the std deviation is not quite enough to take a full picture of the data deviation in the dataset though. Lastly, accounting for the interquartile range, looking at the difference between the first and the third quartile (50% on the table), we can say that our dataset does not sit in a broader than normal deviation range. The analysis helped us also identifying some of the problems with the data quality in our dataset and so we can conclude that it clearly presents some errors probably due to the registration of the records themselves.

# Data quality

## Missing values:

We can currently see that all the values indicating something that is logically obvious are in fact useless for the analysis and so do the attributes that account only for one value. The first obvious column we account for is the modality one, as we already said, the modality in this dataset will be voice-only and not include the facial video recording part. We are done for now and we’ll be looking again at unique values later on to get the job computationally easier. Assessing the missing values is the priority now in getting a better understanding of the distribution and the correlation of all the variables. The picture is much more complete and we can confirm the exact percentages of the missing values already discovered in the first preliminary analysis. In fact the vocal channel column has 7,99%, the intensity one has 33,28% and the actor has 45,92% of the total values missing. The first one to asses is the vocal channel and we are filling it with the mode of the values recorded, due to the relatively small percentage of missing values. Regarding the intensity we grouped the data by emotional intensity, vocal channel and emotion (which are the most correlated ones) and then filled with the mean of the resulted grouped dataset. This procedure assures us, by comparing a distribution graph of the before and after, that filling the missing values hasn’t moved the weight of the distribution, biasing the results.

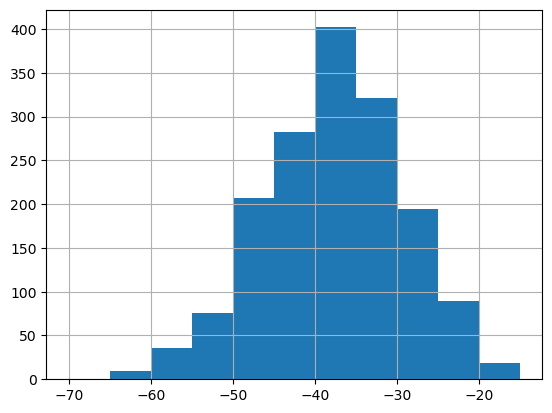
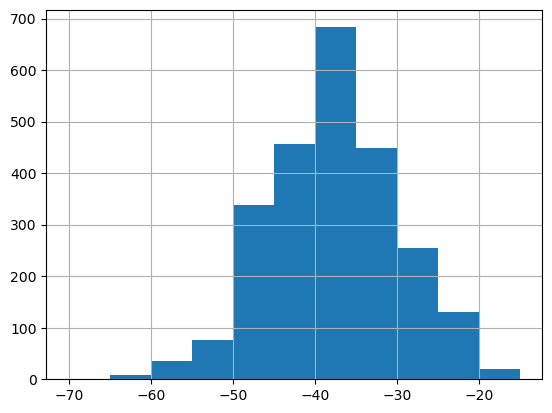


Table 4 - after

Table 3 - before

The last attribute that has missing values is the actor one, they account for as much as 50% of the total data, and no visible correlation appears between this and other attributes, not having enough information to fill the data resulted in us deciding that is better to entirely drop the column.

## Variable transformation:

Chart, histogram

Description automatically generatedLooking at the categorical values we have in our dataset we can see how most of them are composed by only 2 different values, so we used the binary encoding to transform them over one hot encoding as its composed of just binary values. This transformation accounts for the lack of the computational possibility to process text values and get information as the correlation between the different attributes. All the categorical attributes except the emotion one are expressed this way, for this last one instead of using a binary encoding technique (0,1), as the values can be ordered in a range that goes from the less intense emotion, it being the “neutral” one, to the highest intense emotion, it being “surprised”, we have encoded the values with an ordinal range of 0 – 7 based on its intensity. The mapping schema will be useful as we go along in our analisys to make plotting and working with the data easier in a data manipulation perspective. Except the mapping part, regarding variable transformation, we did not see the need to go with log transformed data since most of them, like you can see in the figure “” below the skewness of the dataset is only 0.5 (falling in the normal zone), therefore the dataset is marginally left-skewed but a log-transformation is not useful enough to justify the use of it. Furthermore, like already said above, the variability of the data (obtained by comparing the mean and the STD) is marginal, therefore even in this possible case for log-transformation, we did not see the need to apply it.

## Correlation matrix and eliminated variables

After dealing with missing values and transforming the variables above in numerical ones we can create a correlation matrix, also related to the first part of the dataset, that gives us a quick look on the relations between the various records. Before proceeding further with it is appropriate to reduce the dataset by removing other not useful records. The two attributes that are representing only one unique value are: sample width and frame rate. Those 2 have the same value across all the records, as said before, are useless for our analysis and therefore will be dropped entirely. Managed those two we can go on with the correlation matrix.

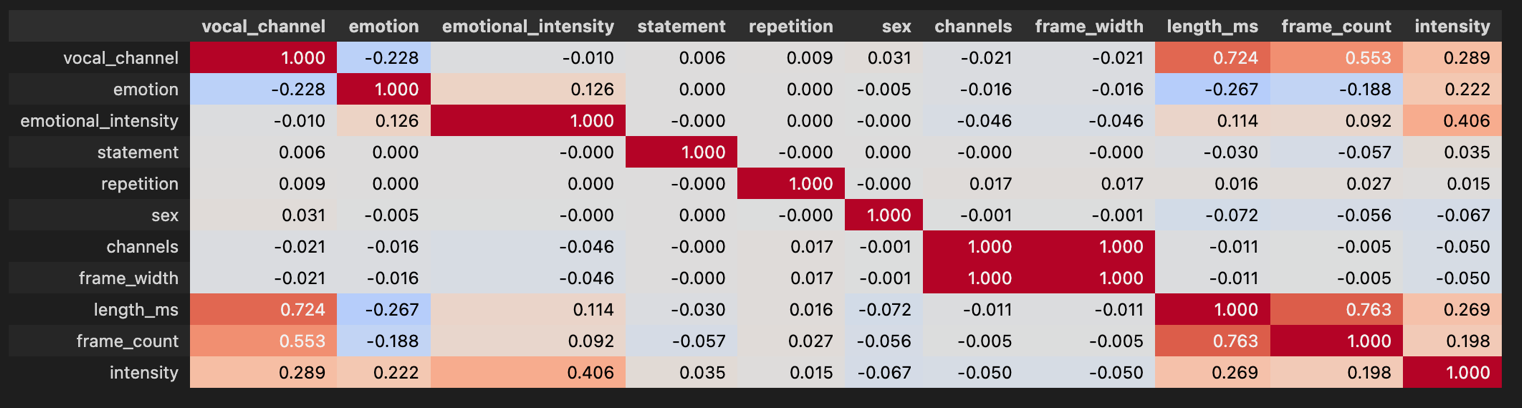
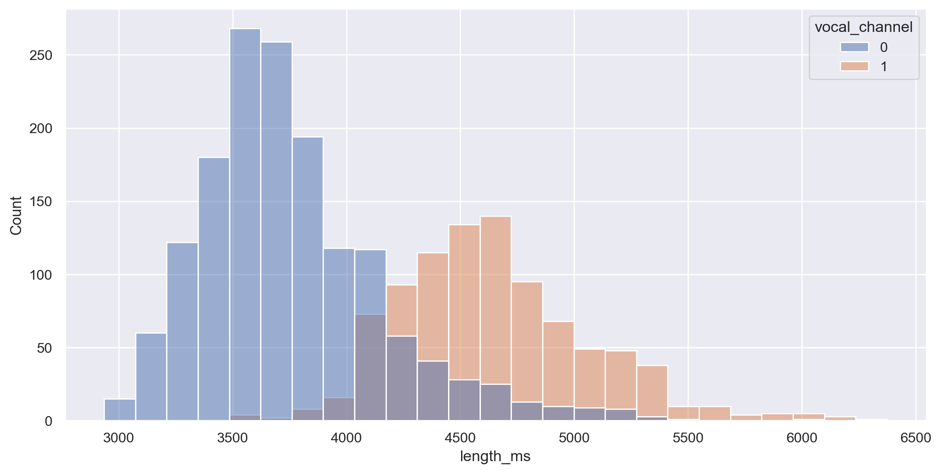
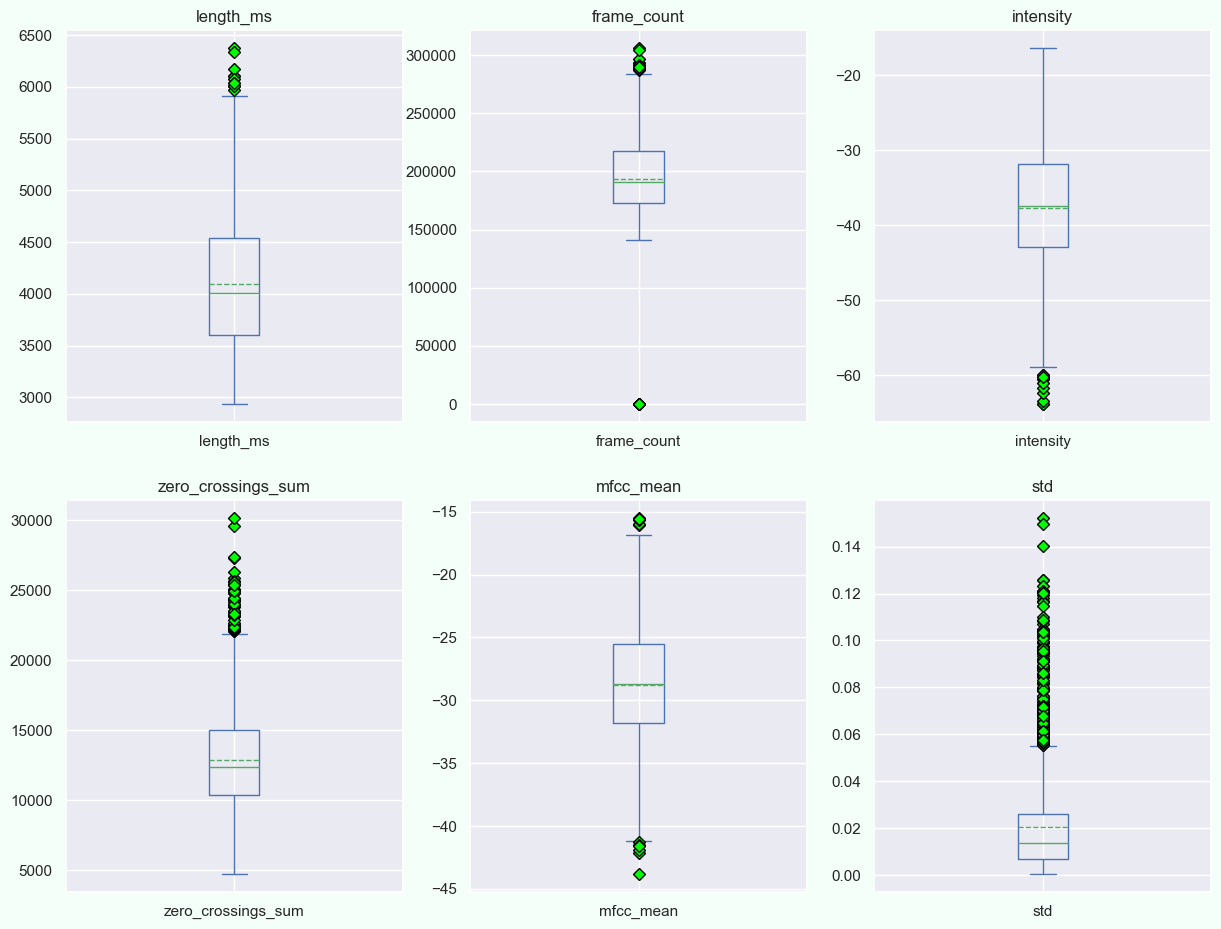


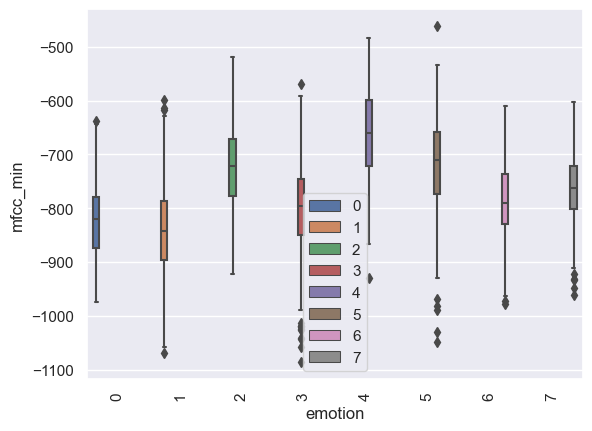
Table 5- non statistical part correlation matrix (Pearson method)

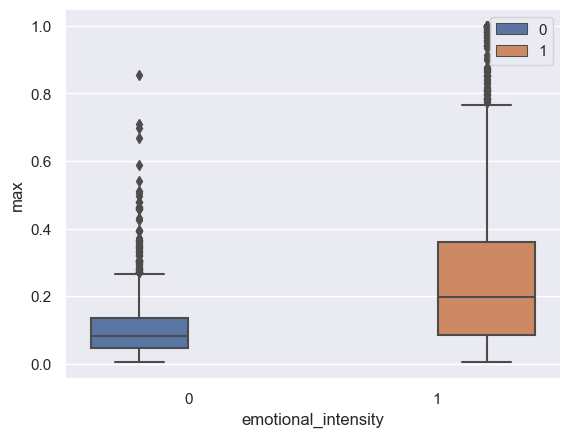
To create the correlation graph we used the Pearson corr. method together with the Kendall one to better understand the type of correlation between the two variables. Briefly there are no highly correlate values except for the ones we already used to fill the missing values and the length in ms and the frame count. Those last two are positively correlated (0.73P - 0.97K) as the length of the recording increases so does the frame count due to the recording been physically bigger and requiring more frames to be stored, the two methods used help us understand that there is a stronger monotonic relation (kendall) and so the variables move for 97% of the time in the same direction but not at a constant rate (only 73%). Another strong positive correlation is between the length and the vocal channel, as the length increases the vocal channel does to (0 = speech, 1 = song), so we can conclude here that the lengthier audio recordings are the one in which the actors sing (a visual representation of the output on the right)

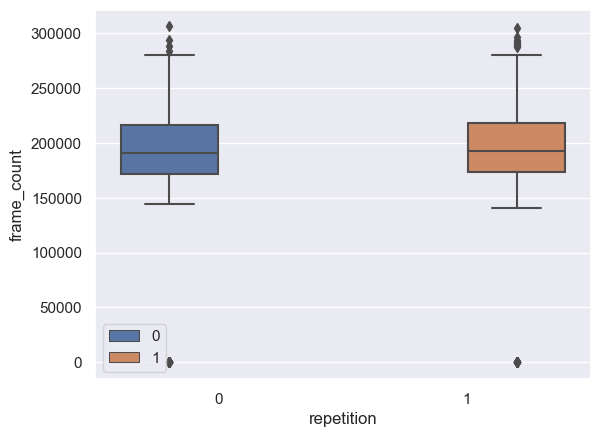
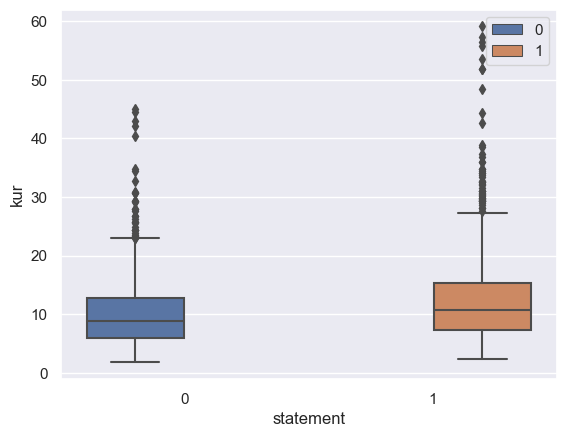
## Outliers

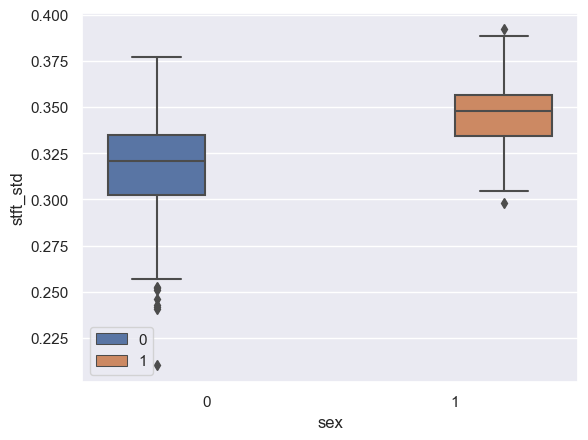
Checking if there are outliers in the dataset is as important as accounting for missing values, it makes the representation of the data itself clearer and not biased by far data points (unusual values that vary from one attribute to another). To identify the outliers there is no strict mathematical rule but we can use boxplots to get a visual representation of the values outside the normal range and decide from there. In the case of the RAVDESS dataset there is no excessive variability in the data regarding the non-statistical part, however the statistical means part is highly variable and presents a lot of outliers and removing all of them resulted in a dataset more than halved. Therefore, we defined a method to get the outliers of all the categorical attributes compared to the most correlated continuous one. When we got all the outliers, following the method explained above, we intersected those that we found and removed the one that are common between the analyzed one. Using this approach guarantees us continuity and gives a less biased dataset from a logical-mathematical perspective (even though the statistical correlation between the records is nothing to be impressed with, this approach implies a more conscious understanding of the dependencies of the records).



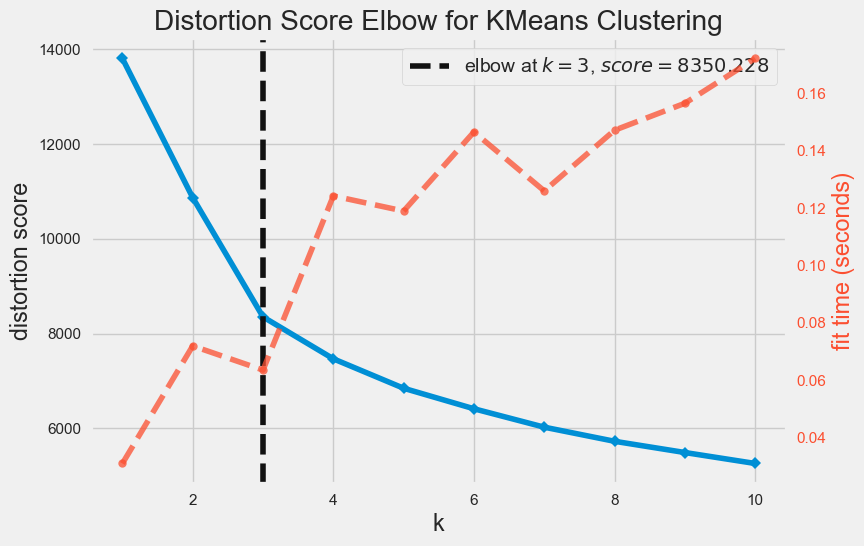




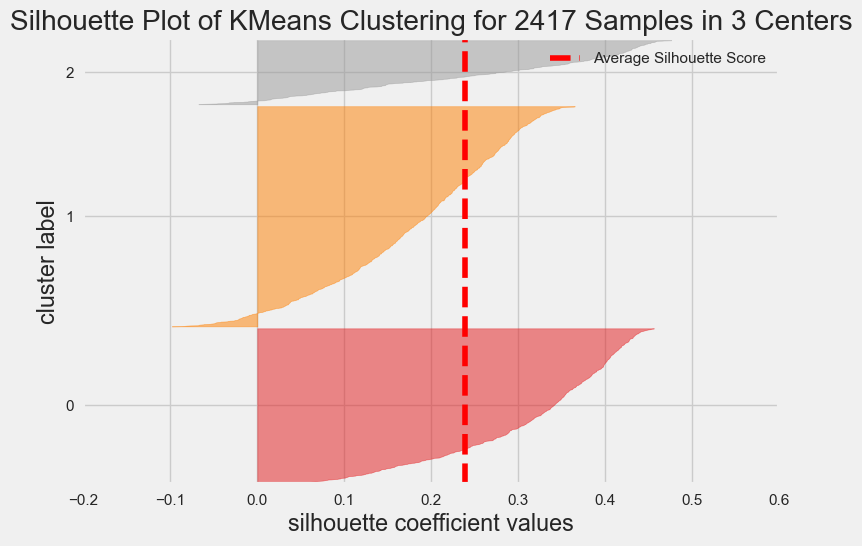




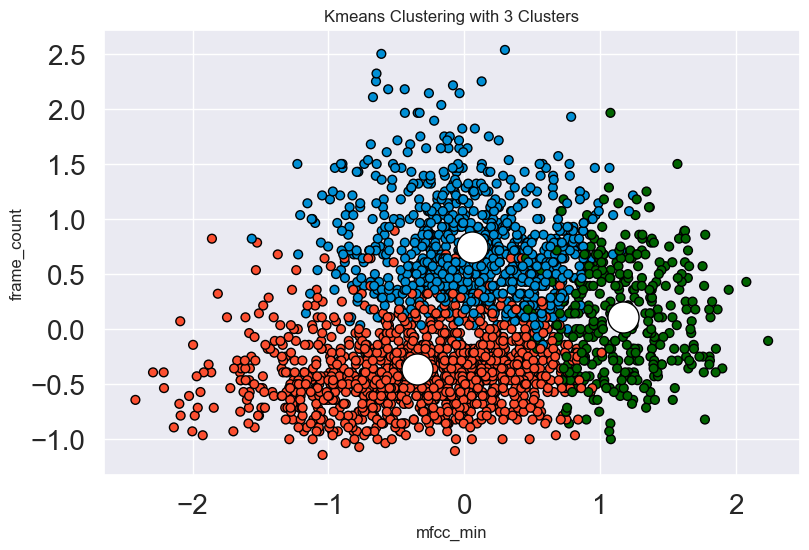
# Clustering

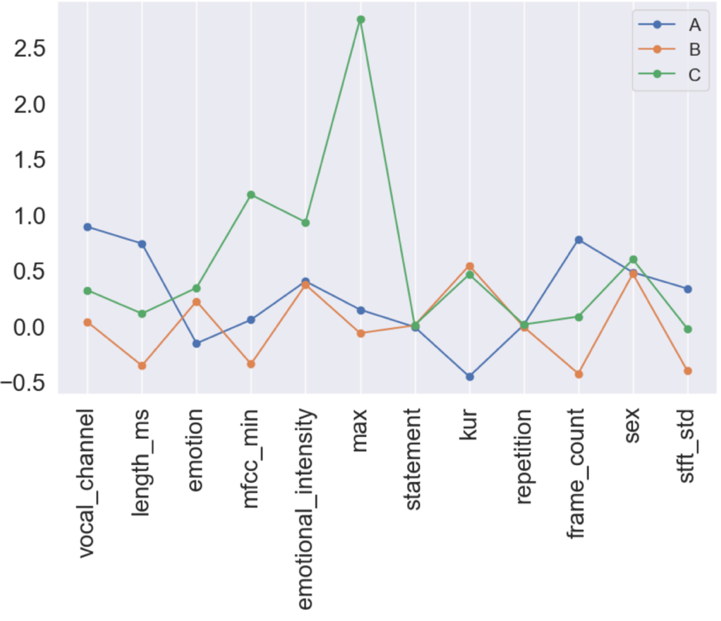
In this section we will deal with clustering the units present in the dataset, in order to obtain groups that have common characteristics. What is expected is to find any correlations that, in the previous phase of Data Understanding had not emerged. The chosen attributes to carry out the cluster analysis they will all be quantitative and will be specified from time to time for each technique used. First, the K-means clustering technique will be applied. After that, they will be shown also the results related to the DBSCAN and Hierarchical techniques. In order to do clustering, after our analysis, we chose the following attributes mentioned below for best results which are as follows: "vocal\_channel", "length\_ms", "emotion", "mfcc\_min", "emotional\_intensity", "max", "statement", "kur", "repetition", "frame\_count", "sex", "stft\_std”. We have used Min\_Max Scaler to scale the values in dataset to start with clustering.  Our goal was to identify any correlations with the attributes considered.

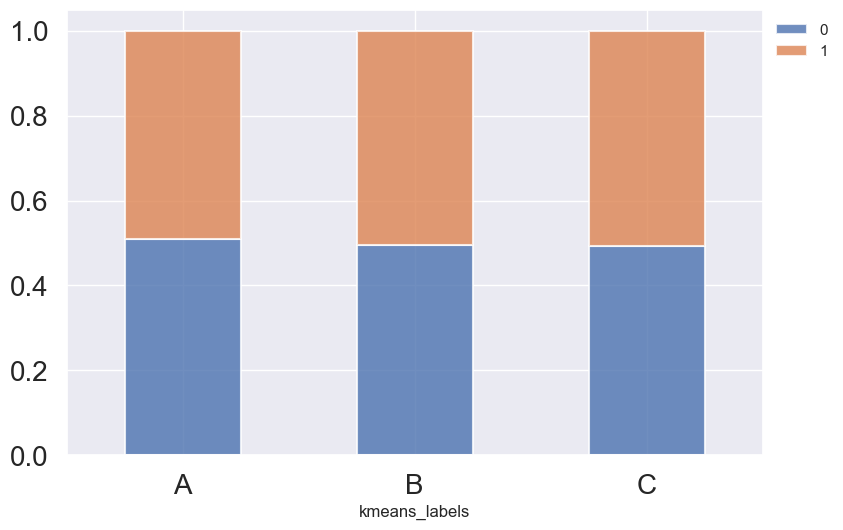
## K-means

The **K- means** is a partitioned group analysis algorithm that allows to subdivide a set of objects into k groups based on their attributes.It is a variant of the expectation-maximization (EM) algorithm whose goal is to determine the k groups of data generated by Gaussian distributions. The initial value of the centroids was defined using the 'k- means++' method. This method initializes the centroids so that they are far apart, thus allowing better results than the random initialization method. The latter, in fact, could lead to a non-optimal but local solution. However, with the chosen method, the centroids will be more precise and therefore fewer iterations will be required. To identify the optimal number of clusters (k), it was decided to calculate the value of the SSE (Sum of Squared Errors for a number of clusters ranging from 2 to 11 and also plot them to find distortion score elbow with which we found that at k = 3 we had an elbow score of 8350.23 after which k-means algorithm tries to minimize distortion (SSE) between each observation vector and its dominating centroid. We also calculated and the Silhouette Coefficient and also plotted to view the average Silhouette Coefficient which in our case was 0.24. The value of the Silhouette coefficient in the best case is close to 1, therefore the value of k is sought such that this is the maximum and in our case k=3 gave us the best coefficient.

From the figure shown, it can be seen that the clusters are much more influenced by the variables mfcc\_min,  frame\_count, length\_ms and stft\_std.. Indeed, the centroids of Clusters B and C have similar values for the emotion, statement, kur, repetition and sex attributes, while the centroids of Clusters A and B have a similar values for the emotionl\_intensity, statement, repetition and sex attributes. The clusters are well seperated at vocal\_channel, length\_ms, mfcc\_min, frame\_count and stft\_std. By our analysis, we found that, the 3 different clusters are perfectly separated at mfcc\_min and frame\_count which is plotted below.We should also note that, inspite of having few outliers, our K Means algorithms has perfectly distinguished one cluster from another.







When we tried to compare the clusters with all the categorical attributes, the attribute “Statement” had the highest percentage(42 %) of values which were similar to the clusters formed by our K-Means Algorithm.

## DBSCAN

The dbscan is based on density because it connects regions of points with sufficiently high density. DBSCAN is one of the most used clustering algorithms and is also the most cited in the scientific literature. DBSCAN estimates the density around each point (item) by counting the number of points in a neighborhoodee (o epseps) specified by the user, and applies call thresholdsminPtsminPtsto identify the “core”, “border” and “noise” points. In a second step, the core points are gathered in a cluster, if they are "density-reachable" ("reachable by density", i.e. if there is a chain of core points in which each point falls within the eps-around the following). Finally the edge points are assigned to the clusters. The algorithm requires only the parametersepseps e minPtsminPts.

Shows the distribution of the cluster and noise points with respect to the target variable in the scatter plot below:

Chart, line chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

The best value for the Epsilon parameter is found near the 'elbow' of the function represented. This is because, above that part of the curve, increasing Epsilon does not further increase the maximum distance between points. Having determined the parameters, the DBscan technique was applied to the entire dataset.

The predicted labels {{-1: 79, 0: 2338, 1: 0} out of which, label “-1” which indicates the outliers, label “0” is the only cluster which is been found and label “1” has no instances at all. And, the calculated Silhouette coefficient is equal to 0.33.

Good results are obtained considering MinPoints=3 and Epsilon=1.25.. It can be seen that the 79 noise points are effectively distributed in the upper part of the graph, i.e., for high values of the variable relating to the high “kur” and can be considered outliers which are viewed by white color except the points which are black.

When we tried to compare the clusters with all the categorical attributes, the attribute “Statement” had the highest percentage(49 %) of values which were similar to the clusters formed by our Dbscan Algorithm.

2.3 Clustering with agglomerative hierarchical technique

For the application of the agglomerative hierarchical technique we considered the same 12 variables used for the K means and the DBscan. The distance function used will be the Euclidean distance.

For problems related to computational complexity, a clusterization with the K Means technique was first carried out, 4 different aggregation methods were considered : full bond, single bond, average bond and Ward's method. The results are shown using the dendrogram in Figure 2.7

For each of the methods considered, the height at which the cut was chosen was mainly based on the possibility of obtaining 3 clusters. Table 2.4 shows the results obtained. At first glance, it can be seen that, using the single bond method, the merging distance is very low up to the penultimate iteration.

Chart, histogram

Description automatically generated Chart

Description automatically generated

Chart

Description automatically generated Chart

Description automatically generated

Figure 2.7: Dendogram for different aggregation methods

For each of the methods considered, the height at which the cut was chosen was mainly based on the possibility of obtaining 3 clusters. Table 2.4 shows the results obtained. At first glance, it can be seen that, using the single bond method, the merging distance is very low up to the penultimate iteration.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cutting height | Silhouette Coefficient | Predicted Labels  0 1. 2 | | |
| Complete | 9 | 0.3316340087583796 | 2234 | 15 | 168 |
| Single | 2.15 | 0.3535557732337175 | 2414 | 1 | 2 |
| Group Average | 5.2 | 0.32842667898157857 | 2339 | 77 | 1 |
| Ward‘s Method | 62 | 0.18732693703351677 | 755 | 1099 | 563 |

The table shows that, using the single link method, all but two vehicles belong to a single cluster. This can be explained by the fact that the single bond joins the points based on the smallest distance between all pairs. The averaging method allows to obtain highly unbalanced clusters. In fact, clusters 1 and 3 have a very small size compared to cluster 2.

Although the Silhouette score is very high using “single” method, the choice is oriented towards a more balanced clustering albeit with a lower Silhouette score. For this reason, Ward's method was considered. Hence, we prefer the silhouette coefficient 0.18 , with the predicted labels {0: 755, 1: 1099, 2: 563}.

When we tried to compare the clusters with all the categorical attributes, the attribute “Statement” had the highest percentage(38 %) of values which were similar to the clusters formed by our Dbscan Algorithm.

Fig: Scatter plot of the target variable in the three clusters obtained with the single link method

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

2.4 Final evaluation of the best clustering algorithm

Table 2.7 shows the Silhouette score for the considered techniques. The K-Means technique creates 3 clusters of balanced dimensions but with a low Silhouette score, as well as the by Heirarichal Clustering methods which created highly unbalanced clusters with a high silhouette score. Although the Silhouette score suggests that the best method is that of the single link, we believe that strongly unbalanced clusters as in our case are not a good result. For this reason, our choice was oriented towards the DBscan technique.

Table 2.7: Silhouette score for the different techniques considered

|  |  |  |
| --- | --- | --- |
| Clustering Algorithms | No. Of Clusters | Silhoutte Score |
| K-Means | 3 | 0.24 |
| DBscan | 1 | 0.33 |
| Heirarichal | 3 | 0.18 |

Ward's method allows to obtain good results but less defined clusters than the K-Means technique.  
The Silhouette score for the two techniques considered confirms our hypothesis.

In conclusion, the DBscan technique, as well as the K-Means Clustering, are to be considered good techniques to identify clusters. But, by our analysis, the best clustering technique is DBscan because it has the highest Silhouette score, idenitifying balanced clusters and outliers as well.