# 2.1 Centroid-based methods

## 2.1.1 Pre-processing and partitioning

For centroid-based methods we define a function that draw three plots: the SSE/K plot, which can help us to estimate the number of clusters and to better understanding the choices made by us, the KMeans plot and the Bisecting KMeans plot. Then, the function can be invoked to draw these plots between two attributes. We could have divided the choice of the number of clusters into two variables, but the goal is to compare the graphs produced using the two methods, so we decided to assume only one number of clusters for both.

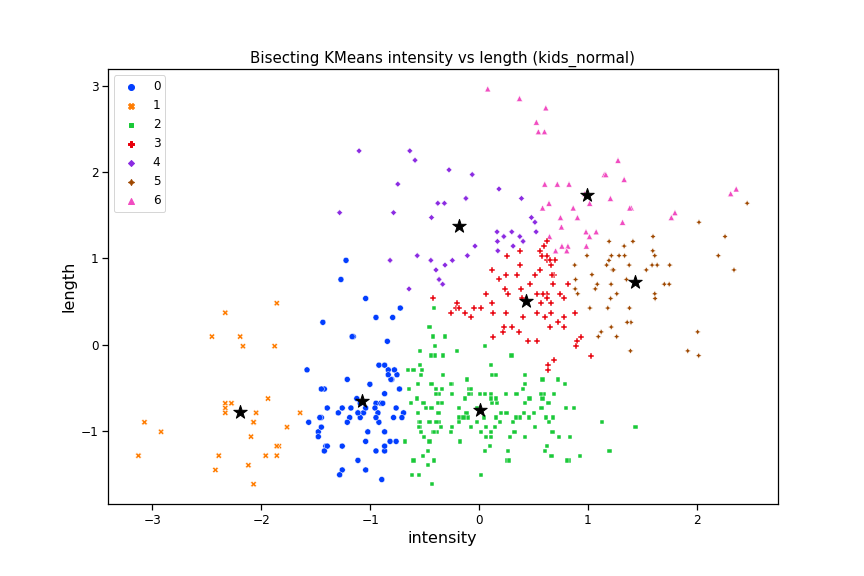
After that, we partition the dataset twice: first time with *statement* values, *0* and *1* which are the numerical transformed values respectively of *Dogs are sitting by the door* and *Kids are talking by the door*, and second time with *emotional\_intensity* values, which are *normal* and *strong*. In this way we get four partitions. In the partitioning function we also normalize the data and eliminate the missing values. Also, before the calculation of KMeans we eliminate the outliers, to which KMeans is sensitive, by calling a dedicated function, named *clean\_outliers()*.

After an exploration of dataset we decide to take in analysis the couple of variables composed by *intensity* and *length*.

## 2.1.2 KMeans and Bisecting KMeans

The first partition is with *emotional\_intensity* = “normal” and the statement *Kids are talking by the door*.

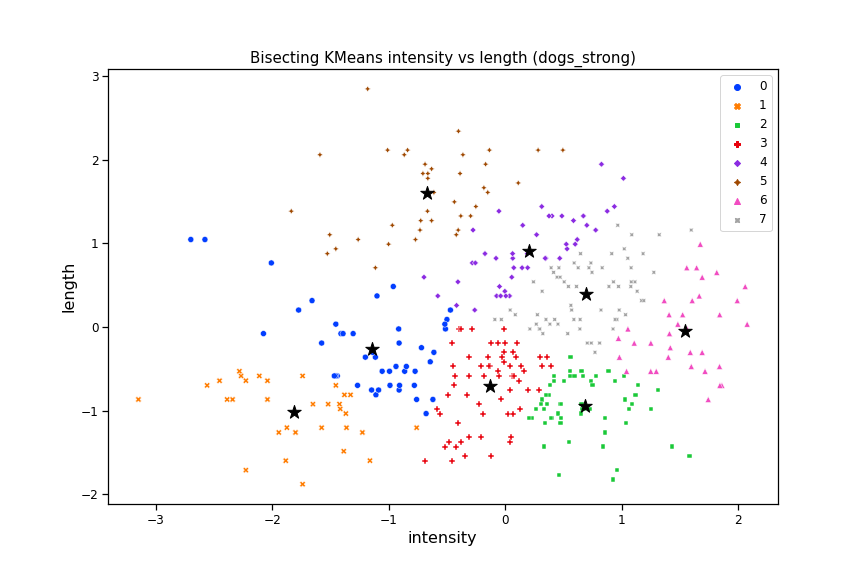
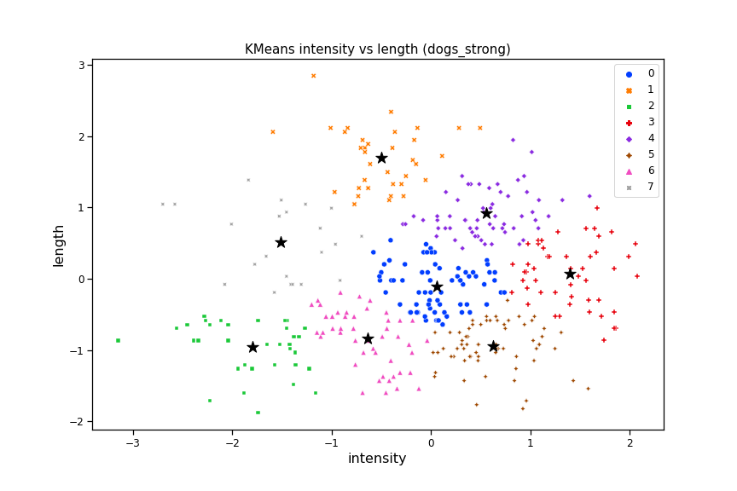
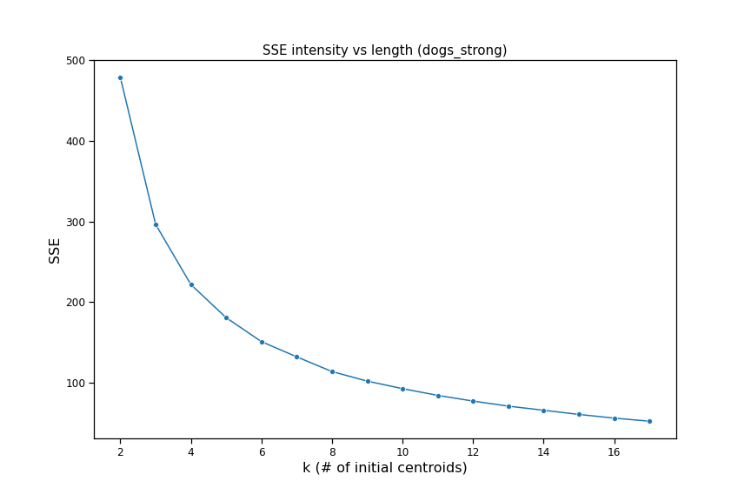
## 



With this plot we decide to use seven clusters because with k >= 7 the curve becomes more linear, as we can see in the [*Figure 1*](#Figure_one).

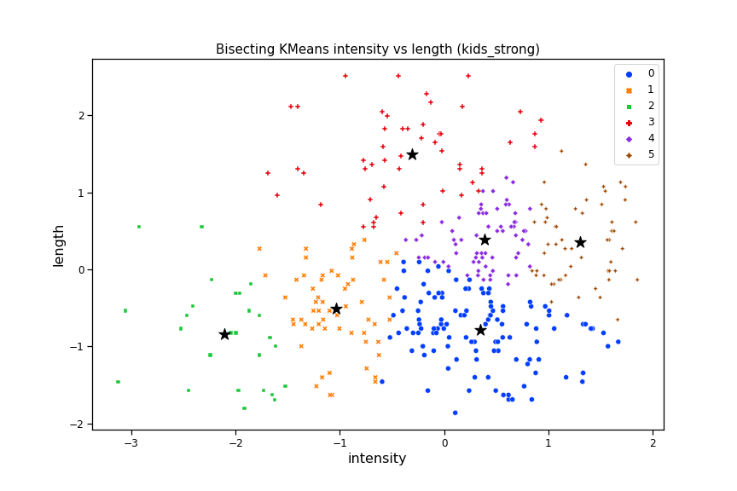
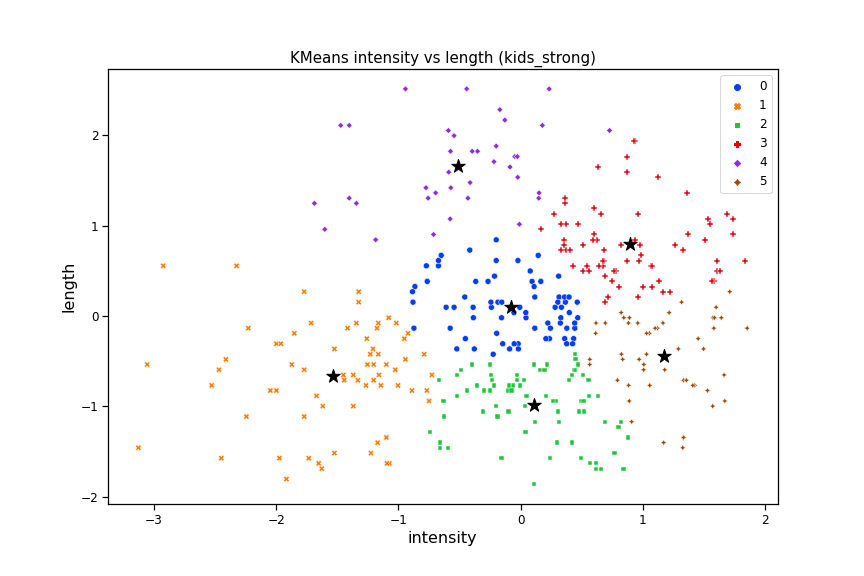
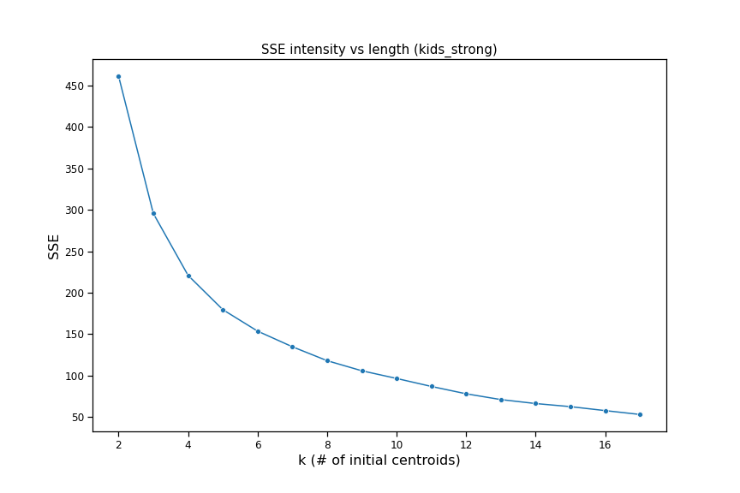
In the [*Figure 2*](#Figure_two) we can see that the recordings are distributed more or less equally between negative and positive intensity, with a slight preference for negative values. Furthermore, there seems to be a weak correlation between the intensity and the length of the recording, so that as one increases, the other increases as well; in fact the highest length values are recorded on average with positive intensity. Finally we can see that the scatterplot is compact, with most values that can be visually split in two macro areas: the first one with low intensity, between -3 and 1, and low length, between -1.5 and 0, and the second one with positive intensity and positive length. We can also see some clustering differences between standard KMeans and Bisecting KMeans, the latter in [*Figure 3*](#Figure_three).

The second partition is with *emotional\_intensity* = “strong” and the statement *Dogs are sitting by the door*.



As before, the number of clusters was chosen through the SSE graph ([*Figure 4*](#Figure_four)) which shows a more linear trend with k >= 8. Here too we note a normal distribution of the values on the intensity axis, with a slightly greater coverage of negative values, with a range of values approximately between -3 and 2. In this case, however, there doesn't seem to be any correlation between intensity and length, since the peaks of the latter are equally distributed between negative and positive intensities. These graphs also are compact, with most values distributed in intensity range between -2 and 2 and length range between -2 and 2. The shape of these graphs tends to be more or less globular. In Bisecting KMeans the pink cluster of standard KMeans is absorbed by two other clusters, the blue one and the red one ([*Figure 5*](#Figure_five) and [*Figure 6*](#Figure_six)).

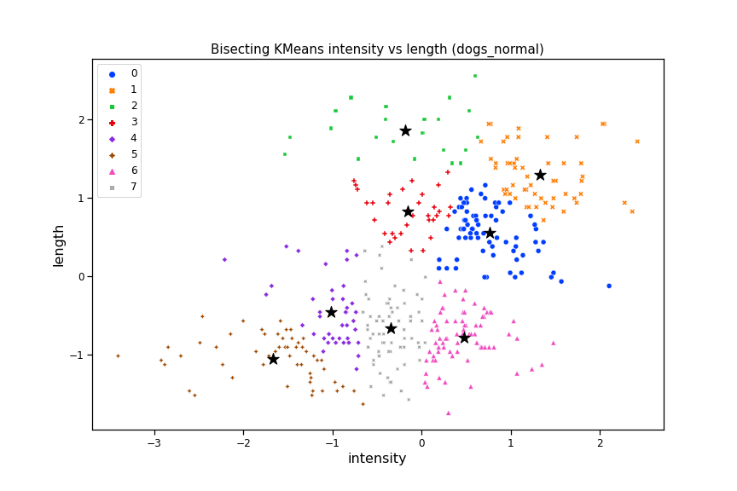
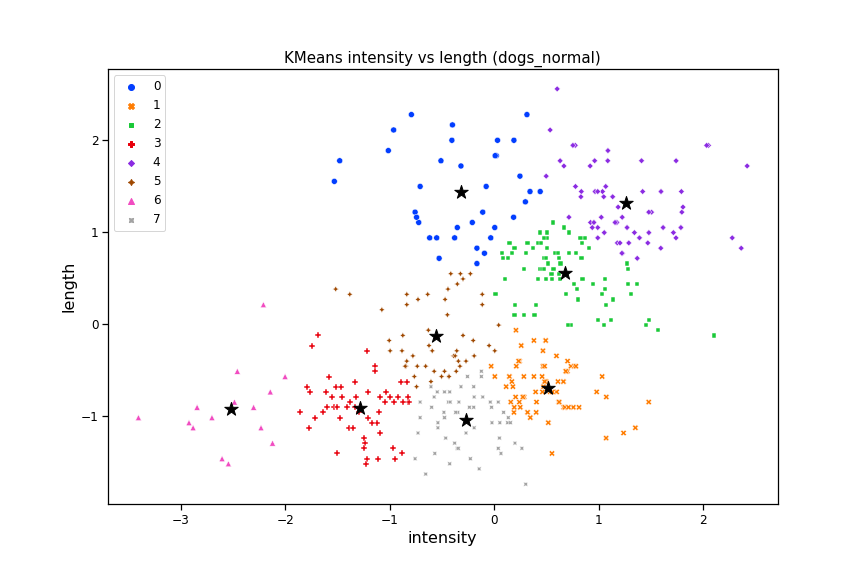
The third partition is with *emotional\_intensity* = “strong” and the statement *Kids are talking by the door*.



From the SSE graph this time we see that it is from k >= 6 that the trend becomes more linear ([*Figure 7*](#Figure_seven)). Partitioning for Kids are talking by the door and emotional\_intensity = "strong" we see that both the distribution of values on the X axis and the absence of correlation between the two attributes with respect to the previous partition have remained unchanged. The values of these graphs are slightly more scattered than the previous ones, with most of the values distributed on the X axis between -2.5 1.5 and on the Y axis between -1.5 and 2. There are also notable differences in clustering between standard KMeans and Bisecting KMeans, as almost all clusters are different The blue clusters of standard KMeans is absorbed by the clusters around in Bisecting KMeans.

The last partition is with *emotional\_intensity* = “normal” and the statement *Dogs are sitting by the door*.

Immagine che contiene testo, sport

Descrizione generata automaticamente

For the number of clusters 8 was chosen because from the SSE graph ([*Figure 10*](#Figure_ten)) we see that the curve begins to flatten with 6 <= k <= 10 and k = 8 is the median of this range of values. We can see that the weak correlation between the two attributes present in the first partition returns and makes us assume that this dependence is given by the partition emotional\_intensity = "normal" and therefore it's independent of the type of statement. Note that Bisecting KMeans distinguish very well low clusters from high clusters, you can see four clusters for low intensity and length and other four clusters for positive intensity and length ([*Figure 12*](#Figure_twelve)), while in standard KMeans there is a brown cluster in the middle ([*Figure 11*](#Figure_eleven)).

# 2.4 Discussion