# 2.1 Centroid-based methods

## 2.1.1 Pre-processing and partitioning

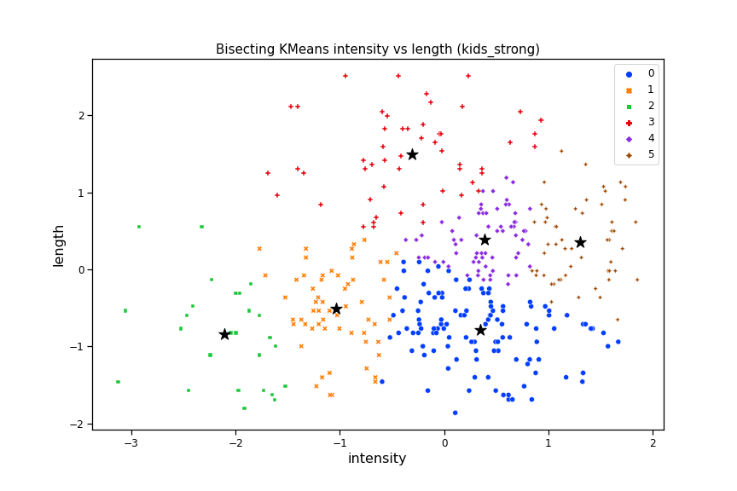
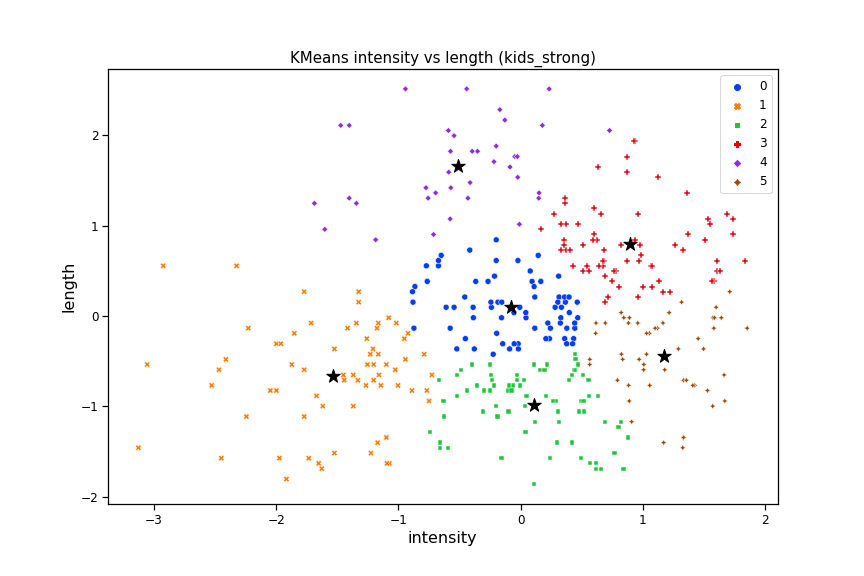
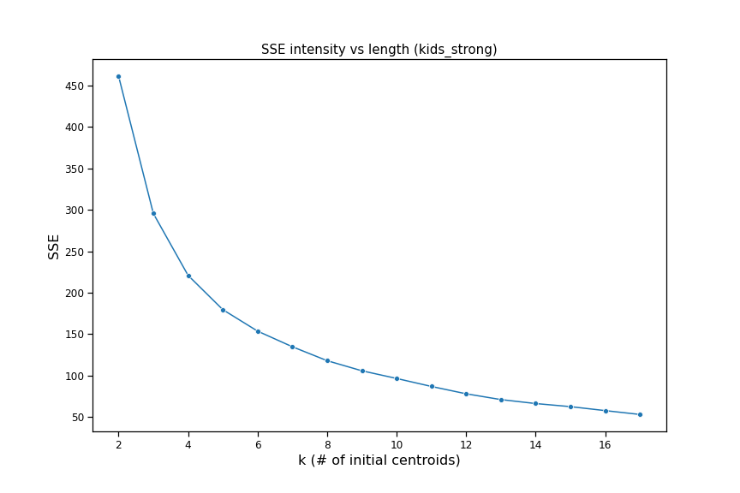
For centroid-based methods we define a function that draw three plots: the SSE/K plot, which can help us to estimate the number of clusters and to better understanding the choices made by us, the KMeans plot and the Bisecting KMeans plot. Then, the function can be invoked to draw these plots between 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

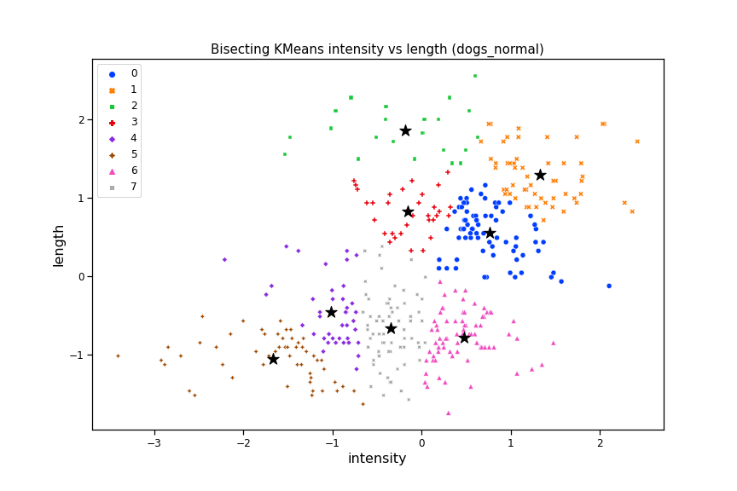
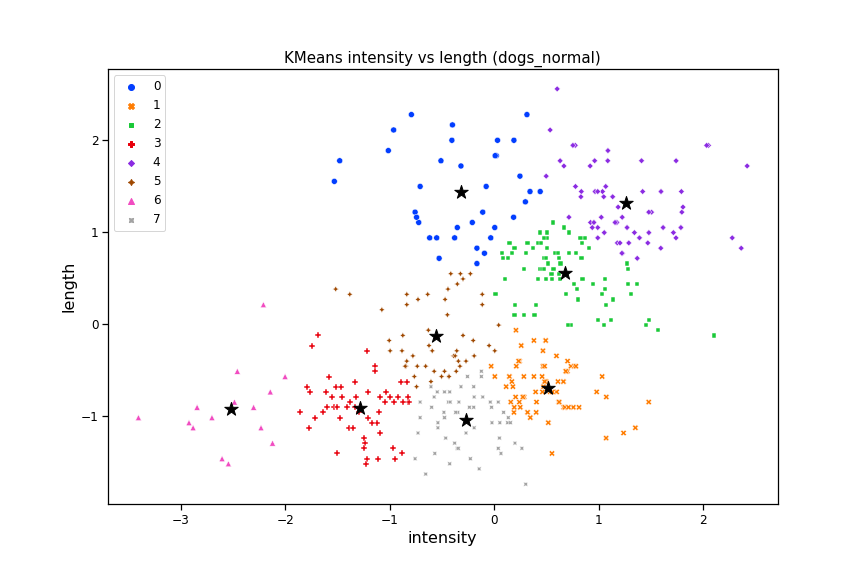
We skip the first and second partition and let's focus on the third and fourth partition. The third partition is with *emotional\_intensity* = “strong” and the statement *Kids are talking by the door*.



From the SSE graph this time we see that it is from k >= 6 that the trend becomes more linear ([*Figure 1*](#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 ([*Figure 2*](#Figure_eight) and [*Figure 3*](#Figure_nine)).

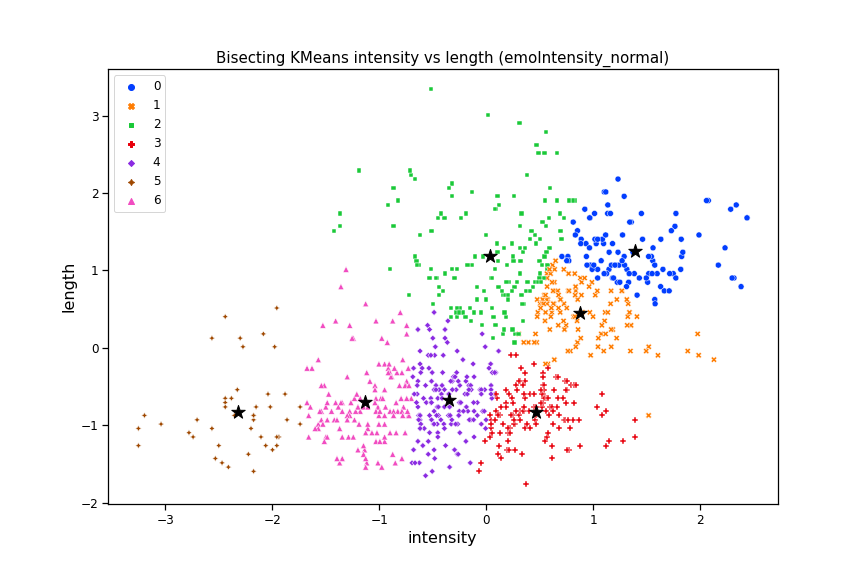
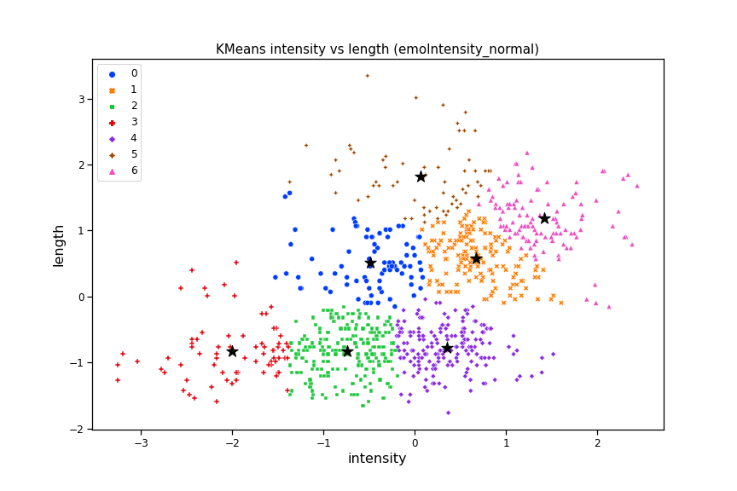
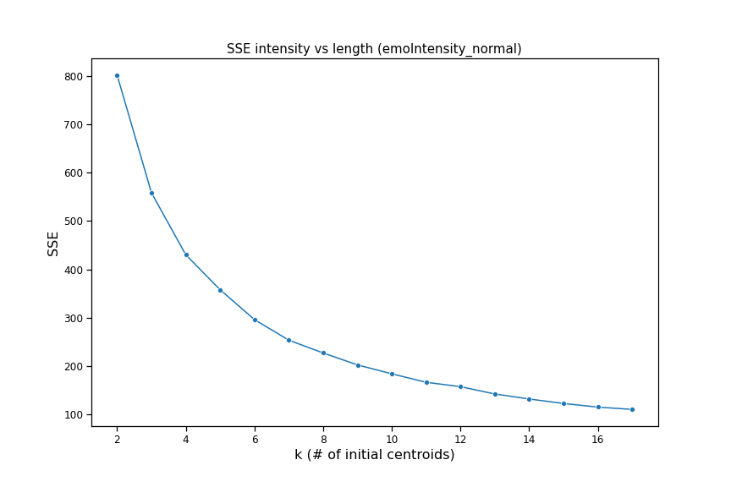
The fourth partition is with *emotional\_intensity* = “normal” and the statement *Dogs are sitting by the door*.

Immagine che contiene testo, sport

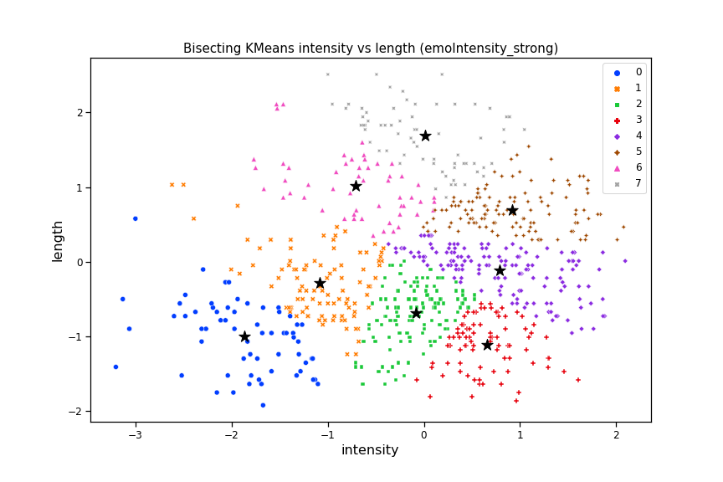
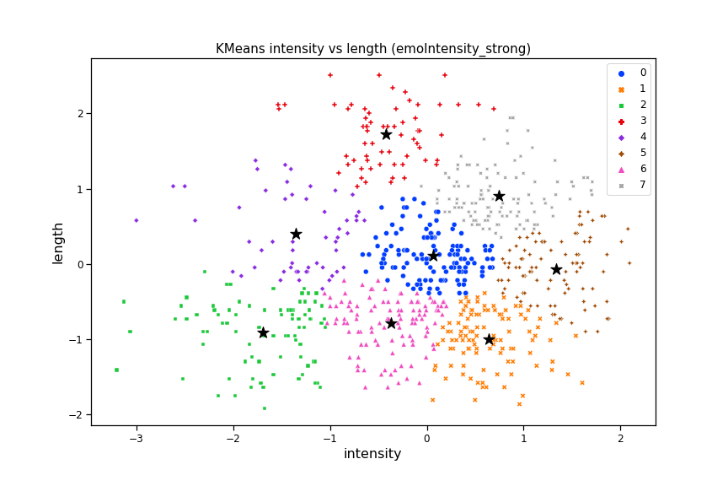
Descrizione generata automaticamente

For the number of clusters 8 was chosen because from the SSE graph ([*Figure 4*](#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 6*](#Figure_twelve)), while in standard KMeans there is a brown cluster in the middle ([*Figure 5*](#Figure_eleven)).

# 2.4 Discussion (my part of partition in common)



7 clusters were chosen because according to the SSE graph with k >= 7 the trend of the curve becomes more linear ([*Figure 7*](#Figure_thirteen)). It can be seen that Bisecting KMeans clearly distinguishes the lower clusters in the graph, that is those with negative intensity and length values, from the higher ones, that is those with positive values of both attributes ([*Figure 9*](#Figure_fifteen)). While the KMeans standard performs a more "gradual" clustering, with two “intermediate” clusters in the center of graph. The shape of the scatter plot is similar to those seen in partitions with emotional\_intensity = "normal " ([*Figure 8*](#Figure_fourteen)).



The same goes for the previous partition, in emotional\_intensity = "strong" we find more or less the shape we observed in the corresponding partitions with the statements. From here we have the proof that the shape and, consequently, the presence of correlations between the attributes doesn't depend on the statement but on the value of emotional\_intensity. Note that also here Bisecting KMeans tends to distinguish more clearly clusters between negative and positive values for both attributes ([*Figure 11*](#Figure_eighteen)), while the KMeans standard doesn't make this distinction very well and it makes a more “gradual” clustering for both attributes ([*Figure 10*](#Figure_seventeen)).