Practical 5 Report

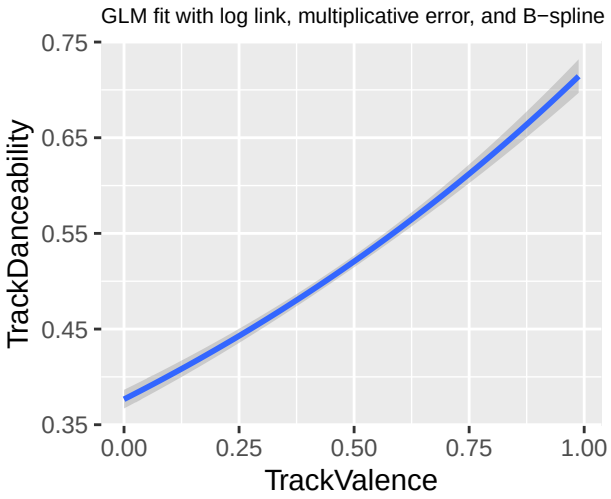
Jakub Domasik, 1632905

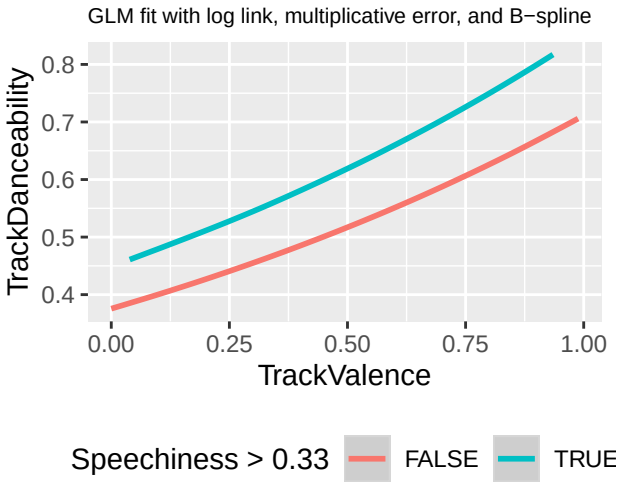
Introduction

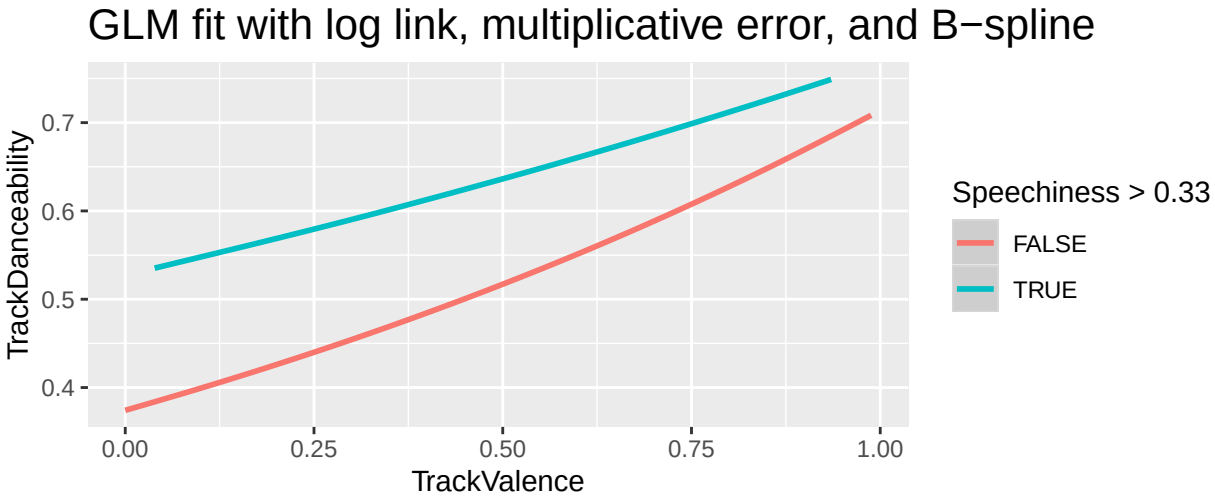
In this report, we will examine how relationships between valence and danceability differ for songs containing speech. We will distinguish two groups of songs: speechy(rap) and non-speechy(non-rap). In my work, libraries *splines* and *ggplot2*, *readxl* and *gridExtra* are required. We will investigate this relationship on the *edited_spotify.xlsx* dataset for variables: *TrackValence*, *TrackDanceability* and *TrackSpeechiness*.

Analysis

The intuition suggests that there is a positive relationship between valence and danceability as more valent tracks tend to be easier to dance to.
To further investigate this difference, three models will be constructed. The first one is a simple linear model of the dependance between *TrackDanceability* and *TrackValence*. The second one allows the intercept to be different for speechy and non-speechy tracks. In the last model, all parameters may be different for rap and non-rap songs.





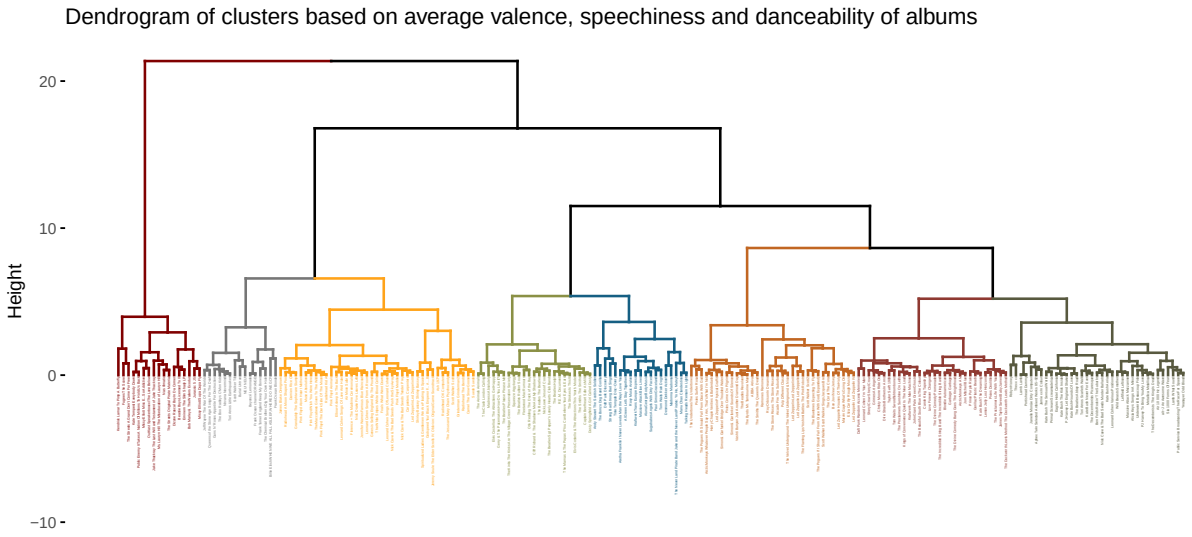


In the first graph, we see a very high correlation between valance of the track and its danceability. The fit looks more or less as a straight line with a slight bend in the middle facing downwards.
From the second model, we learn that rap albums have a larger intercept which means that on average they have larger danceability score by around 0.08.
Last plot shows us that the line representing non-rap songs have much smaller intercept(by around 0.15), however it has a much steeper slope so non-rap songs which are very valent have similar danceability as rap songs with high valence. We may say that the difference in danceability between rap and non-rap songs is diminishing as the valence of the tracks is growing.

Table 1: Analysis of Deviance Table

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
2565	271.9	NA	NA	NA	NA
2564	268.9	1	3.052	32.84	1.119e-08
2563	268.6	1	0.3285	3.535	0.06019

Using ANOVA table on our 3 models, we get P value very close to 0 for the second model and 0.06 for the third model. We claim that all of the models significantly improve our knowledge about the relationship between valence and danceability in two groups: speechy and non-speechy songs. Therefore, we may use the third model as it is the most complex and explains the reality in the most detailed way.
Next, I will try to cluster the albums based on three variables: *TrackValence*, *AlbumSpeechiness* and *AlbumDanceability*. We have 211 rows and it is reasonable to define 8 clusters, because people usually define 8 main genres of music. Then I will create a dendrogram of those clusters using the function *fviz_dend* from the library *factoextra*.



We see that the first, highest division of our dendrogram creates two, very unequal groups. The first one(on the left) consists only of less than 10% and only one cluster, while the other group contains all 7 other clusters. We may conclude that this small group of songs is very different than others in terms of the chosen variables. Furthermore, we see that the clusters are roughly similar in the size and that around 50% of the data is contained in half of the clusters(4).