Stage 2 report

Introduction

[space for text]

Exploratory Data Analysis

Originally, we had planned to use the peak position on the billboard as a response variable, and model that using features (covariates) of the actual song as described in the data. However, this response variable is not an optimal one to use, and through plots, logical reasoning, and exploratory data analysis we will show why and how we arose to to this conclusion and led to our final decision of response variable.

We first cleaned our dataset by removing obviously non-useful variables like names. (However, while we recognize that different genres may have differing levels of popularity, for simplicity we will leave out this variable as the "genre" of certain songs may be ambiguous and there are too many categories to keep track of.)

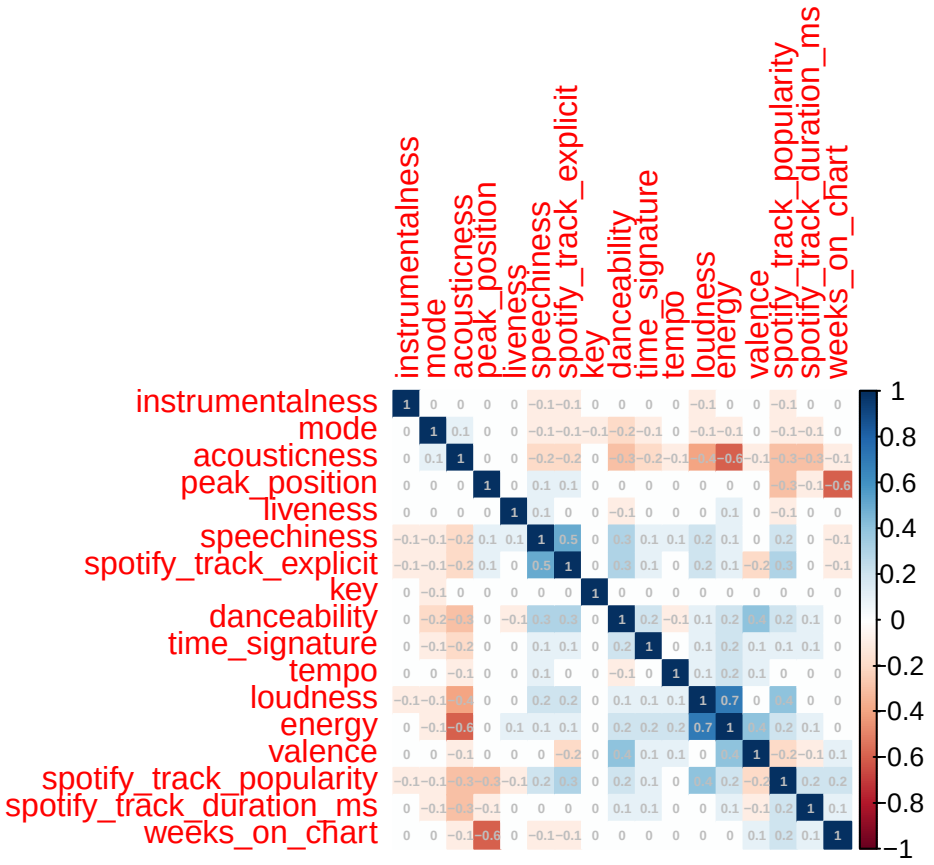
```
bd <- read.csv("joined_billboard_audiofeature.csv")
bd_clean <- select(bd, -X, -performer, -song, -spotify_genre,
    -spotify_track_album)
```

Next we make a correlation plot to have a broad overview of what variables may be connected to our response as well as with each other:

```
library(ggcorrplot)
```

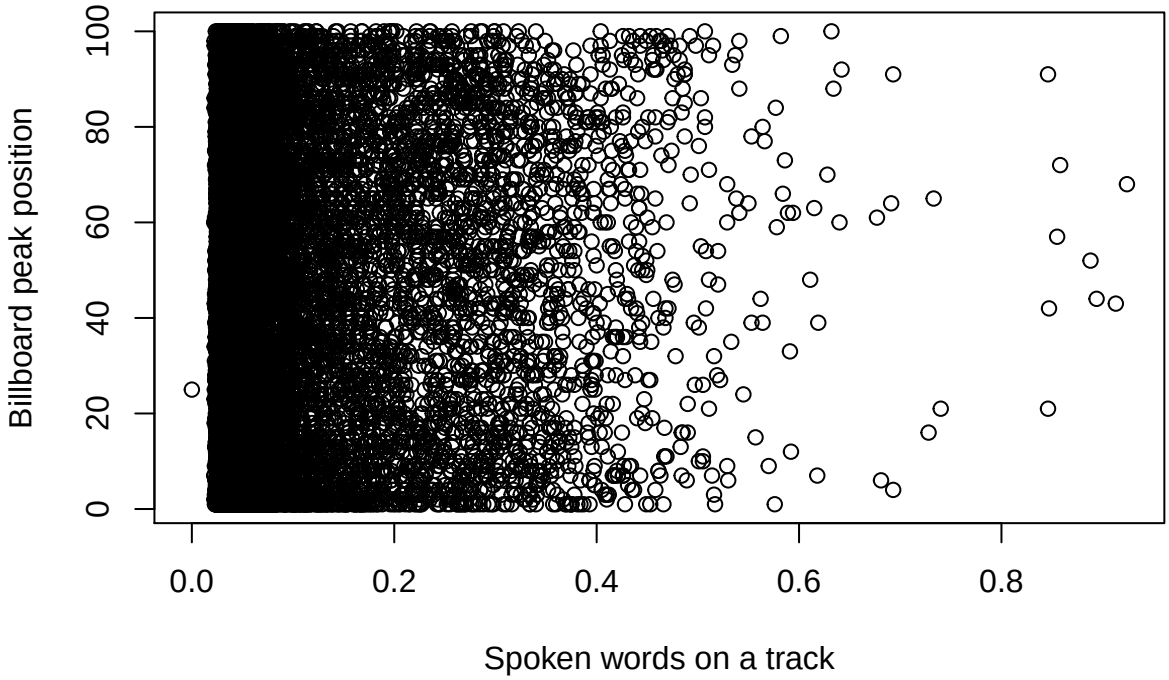
```
## Warning: package 'ggcorrplot' was built under R version 4.0.5
```

```
options(repr.plot.width = 15, repr.plot.height = 15)
cor_matrix <- round(cor(bd_clean), 1)
corrplot(cor_matrix, method = "color", addCoef.col = "grey",
    order = "AOE", number.cex = 0.4)
```



This is where we became very skeptical about our selected response variable: peak_position. The chart shows that a song's peak position on the billboards have little to no correlation to audio features. Even with actual audio features that our response variable had little correlation with, take "speechiness" for example, has a plot that looks like:

```
plot(x = bd_clean$speechiness, y = bd_clean$peak_position, xlab = "Spoken words on a track",
    ylab = "Billboard peak position")
```



Even a quick glance at these plots show that these audio features have little to no inherent effect on where the song ends up on the chart. Sure we had some correlated features like the weeks on the billboard and spotify song popularity, but those variables and their relation to peak_position is quite obvious and doesnt quite cover our goal of wanting to predict a songs position through mostly its AUDIO features.

Upon closer inspection of the chart however, we realized that spotify track popularity did indeed have many correlations with audio features, and is still a proper response variable when it comes to questions like if people wanted to get big on spotify, what kind of songs they should create. Hence this is why we decided to drop peak_position as our response for spotify track popularity.

Zack's Section

```
# Insert code here
```

[space for text]

Anirudh's Section

```
# Insert code here
```

[space for text]