

Song Data Project

Kristian Abad, Steven Truong, Nicole Magallanes

2/13/2022

```
library(readr)
library(tidyr)
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.5    v dplyr   1.0.7
## v tibble  3.1.2    v stringr 1.4.0
## v purrr   0.3.4    v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

Preprocessing

```
data = read_csv("song_data.csv")

## Rows: 18835 Columns: 15

## -- Column specification -----
## Delimiter: ","
## chr  (1): song_name
## dbl (14): song_popularity, song_duration_ms, acousticness, danceability, ene...

##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.

head(data)

## # A tibble: 6 x 15
##   song_name      song_popularity song_duration_ms acousticness danceability energy
##   <chr>          <dbl>          <dbl>          <dbl>          <dbl> <dbl>
## 1 Boulevard o~      73          262333      0.00552      0.496  0.682
## 2 In The End        66          216933      0.0103      0.542  0.853
```

```
## 3 Seven Natio~          76          231733      0.00817      0.737  0.463
## 4 By The Way           74          216933      0.0264      0.451  0.97
## 5 How You Rem~         56          223826      0.000954     0.447  0.766
## 6 Bring Me To~        80          235893      0.00895      0.316  0.945
## # ... with 9 more variables: instrumentalness <dbl>, key <dbl>, liveness <dbl>,
## #   loudness <dbl>, audio_mode <dbl>, speechiness <dbl>, tempo <dbl>,
## #   time_signature <dbl>, audio_valence <dbl>
```

One challenge we need to figure out is addressing the following cases in our data, if there are any:

- Remixes
- Remasters
- Single Versions
- Same name but different artists?
- Radio edits
- Extended versions
- Misc mixes/versions

I think maybe we can leave remixes possibly treating them as reimaginings of songs or somewhat to the same vein that songs have samples from other tracks are in themselves a separate track. Maybe the more difficult is dealing with the other cases. An example that comes to mind is “Smooth Operator” by Sade (seems like only one of the 3 versions is in the data). There’s a single version, a remastered version, and I believe an album version where there’s an immediate difference between the remastered and album version.

I think the duplicated() function finds exact duplicates of rows.

```
duplicates <- data[duplicated(data),]
duplicates
```

```
## # A tibble: 3,909 x 15
##   song_name    song_popularity song_duration_ms acousticness danceability energy
##   <chr>          <dbl>          <dbl>          <dbl>          <dbl> <dbl>
## 1 Sex on Fire      81          203346      0.00172      0.542  0.905
## 2 Use Somebo~     79          230760      0.00552      0.276  0.715
## 3 Hips Don't~     84          218093      0.284        0.778  0.824
## 4 Hotel Cali~     83          391376      0.00574      0.579  0.508
## 5 Me and Bob~     69          271333      0.302        0.453  0.464
## 6 Imagine - ~     77          187866      0.907        0.547  0.257
## 7 Let It Be ~     78          243026      0.631        0.443  0.403
## 8 Rocket Man~     80          281613      0.386        0.602  0.522
## 9 My Sweet L~     78          281226      0.0794       0.538  0.704
## 10 Tangled up~    63          341626      0.414        0.421  0.661
## # ... with 3,899 more rows, and 9 more variables: instrumentalness <dbl>,
## #   key <dbl>, liveness <dbl>, loudness <dbl>, audio_mode <dbl>,
## #   speechiness <dbl>, tempo <dbl>, time_signature <dbl>, audio_valence <dbl>
```

Just testing some cases here...while scrolling through on kaggle, I just picked a random duplicate song to test

```
duplicates %>%
  filter(song_name == 'Zombie')
```

```
## # A tibble: 1 x 15
##   song_name song_popularity song_duration_ms acousticness danceability energy
##   <chr>          <dbl>          <dbl>          <dbl>          <dbl> <dbl>
## 1 Zombie              82          306410          0.0163          0.299 0.613
## # ... with 9 more variables: instrumentalness <dbl>, key <dbl>, liveness <dbl>,
## #   loudness <dbl>, audio_mode <dbl>, speechiness <dbl>, tempo <dbl>,
## #   time_signature <dbl>, audio_valence <dbl>
```

Here's an interesting case where we have 2 of the same rows and 1 with a remix with a track called "8 Letters"

```
duplicates %>%
  filter(song_name == '8 Letters')
```

```
## # A tibble: 1 x 15
##   song_name song_popularity song_duration_ms acousticness danceability energy
##   <chr>          <dbl>          <dbl>          <dbl>          <dbl> <dbl>
## 1 8 Letters              72          190026          0.649          0.607 0.478
## # ... with 9 more variables: instrumentalness <dbl>, key <dbl>, liveness <dbl>,
## #   loudness <dbl>, audio_mode <dbl>, speechiness <dbl>, tempo <dbl>,
## #   time_signature <dbl>, audio_valence <dbl>
```

```
duplicates %>%
  filter(song_name == '8 Letters - R3HAB Remix')
```

```
## # A tibble: 0 x 15
## # ... with 15 variables: song_name <chr>, song_popularity <dbl>,
## #   song_duration_ms <dbl>, acousticness <dbl>, danceability <dbl>,
## #   energy <dbl>, instrumentalness <dbl>, key <dbl>, liveness <dbl>,
## #   loudness <dbl>, audio_mode <dbl>, speechiness <dbl>, tempo <dbl>,
## #   time_signature <dbl>, audio_valence <dbl>
```

So it looks like it just picks up exact duplicates and we'll need to figure out what we're going to do with other cases.

```
data_2 <- data[!duplicated(data),]
nrow(data)
```

```
## [1] 18835
```

```
nrow(data_2)
```

```
## [1] 14926
```

```
nrow(data) - nrow(data_2)
```

```
## [1] 3909
```

Using the grepl function to find any instance of single, remastered, and radio edit/mix versions of tracks.

```
#Function was found via stackoverflow
#https://stackoverflow.com/questions/10128617/test-if-characters-are-in-a-string
data_3 <- data_2[!(grepl('Single',data_2$song_name,fixed=TRUE) |
  grepl('Remaster',data_2$song_name,fixed=TRUE) |
  grepl('Radio',data_2$song_name,fixed=TRUE)),]
data_3
```

```
## # A tibble: 14,276 x 15
##   song_name    song_popularity song_duration_ms acousticness danceability energy
##   <chr>          <dbl>          <dbl>          <dbl>          <dbl> <dbl>
## 1 Boulevard ~      73            262333        0.00552        0.496 0.682
## 2 In The End        66            216933        0.0103         0.542 0.853
## 3 Seven Nati~       76            231733        0.00817        0.737 0.463
## 4 By The Way        74            216933        0.0264         0.451 0.97
## 5 How You Re~       56            223826        0.000954       0.447 0.766
## 6 Bring Me T~       80            235893        0.00895        0.316 0.945
## 7 Last Resort       81            199893        0.000504       0.581 0.887
## 8 Are You Go~       76            213800        0.00148        0.613 0.953
## 9 Mr. Bright~       80            222586        0.00108        0.33 0.936
## 10 Sex on Fire       81            203346        0.00172        0.542 0.905
## # ... with 14,266 more rows, and 9 more variables: instrumentalness <dbl>,
## #   key <dbl>, liveness <dbl>, loudness <dbl>, audio_mode <dbl>,
## #   speechiness <dbl>, tempo <dbl>, time_signature <dbl>, audio_valence <dbl>
```

Checking for missing values

```
data_3[is.na(data_3)]
```

```
## <unspecified> [0]
```

```
data_3[grepl('Mix',data_3$song_name,fixed=TRUE) |
  grepl('Version',data_3$song_name,fixed=TRUE),]
```

```
## # A tibble: 226 x 15
##   song_name    song_popularity song_duration_ms acousticness danceability energy
##   <chr>          <dbl>          <dbl>          <dbl>          <dbl> <dbl>
## 1 Change (In~      69            299533        0.000249       0.269 0.853
## 2 Banana Cli~      57            202373        0.107         0.398 0.66
## 3 Dinero - E~      37            161093        0.00436        0.701 0.779
## 4 Si Una Vez~      54            237860        0.0762         0.62 0.823
## 5 The Way Yo~      68            234000        0.126         0.871 0.597
## 6 Lambada - ~      67            207466        0.211         0.756 0.8
## 7 Bailando --      71            243413        0.0426         0.723 0.777
## 8 The Girl F~      50            325506        0.888         0.66 0.124
## 9 My Generat~      69            198706        0.328         0.397 0.785
## 10 You Really~      71            134253        0.493         0.573 0.939
## # ... with 216 more rows, and 9 more variables: instrumentalness <dbl>,
## #   key <dbl>, liveness <dbl>, loudness <dbl>, audio_mode <dbl>,
## #   speechiness <dbl>, tempo <dbl>, time_signature <dbl>, audio_valence <dbl>
```

```
data_3[grepl('Swimming Pools (Drank)',data_3$song_name,fixed=TRUE),]
```

```
## # A tibble: 2 x 15
##   song_name      song_popularity song_duration_ms acousticness danceability energy
##   <chr>          <dbl>          <dbl>          <dbl>          <dbl> <dbl>
## 1 Swimming Po~      68          313786          0.123          0.716 0.485
## 2 Swimming Po~      45          313786          0.123          0.716 0.485
## # ... with 9 more variables: instrumentalness <dbl>, key <dbl>, liveness <dbl>,
## #   loudness <dbl>, audio_mode <dbl>, speechiness <dbl>, tempo <dbl>,
## #   time_signature <dbl>, audio_valence <dbl>
```