midterm project

Danya Zhang

2022-11-30

R Markdown

Reading in data

```
top100 <- read.csv("/Users/dz/Downloads/archive (1)/Spotify 2010 - 2019 Top 100.csv")
top100 <- top100[-c(1001:1003),] #last 3 rows NA
```

Cleaning data

```
subgenre_df <- as.data.frame(table(top100$subgenre))
#rename top.genre column to subgenre
names(top100)[names(top100) == "top.genre"] <- "subgenre"

#divide into 10 general categories
pop_rows <- grep(paste(c("pop", "neo mellow", "talent show", "indietronica", "adult standards", "boy ba
hiphop_rows <- grep(paste(c("hip hop", "rap", "trap", "g funk","uk drill"), collapse="|"), top100$subge
rock_rows <- grep(paste(c("rock", "permanent wave", "icelandic indie", "emo"), collapse="|"), top100$su
country_rows <- grep("country", top100$subgenre, ignore.case=TRUE)
latin_rows <- grep(paste(c("latin", "reggae"), collapse="|"), top100$subgenre, ignore.case=TRUE)
randb_rows <- grep(paste(c("soul", "r&b"), collapse="|"), top100$subgenre, ignore.case=TRUE)
edm_rows <- grep(paste(c("house", "grime", "edm", "australian dance", "tronica",
                         "dancefloor dnb", "french shoegaze", "big room", "techno",
                         "electro", "brostep", "complextro","alternative dance"), collapse="|"), top100
metal_rows <- grep("metal", top100$subgenre, ignore.case=TRUE)

#make new column for parent genre
#10 genres
top100$genre <- ""
top100 <- top100[, c(1,2,18,3:17)]
top100[pop_rows, 3] <- "pop"
top100[hiphop_rows, 3] <- "hip hop"
top100[rock_rows, 3] <- "rock"
top100[country_rows, 3] <- "country"
top100[latin_rows, 3] <- "latin"
top100[c(21,177, 111), 3] <- "folk"
top100[randb_rows, 3] <- "r&b"
top100[edm_rows, 3] <- "edm"
top100[metal_rows, 3] <- "metal"
top100$genre <- sub("^$", "other", top100$genre)
```