Homework for Advanced R programin - Class 7 - Script and report automation (RMarkdown)

Rafał Rysiejko

09/04/2020

Introduction

The recorded music industry was worth $19.1 billion in 2018, which was almost a double-digit gain (9.7%) from the year prior [1]. In this entertainment market, an increasingly large number of song products are introduced each year. However, only a small group achieves mainstream success, and among these, the distribution of market success is becoming increasingly skewed. By analyzing the top songs from the popular music streaming service provider, we can investigate if some underlying patterns make a song truly hit.

This assignment will try to determine whether there are some distinctive features of song that make it successful and if so, are there any patterns among those features. Dataset [2] used for this assignment contains information about the top 50 most listened songs in the world on music streaming platform Spotify in 2019. Each song has a set of 13 variables:

Table 1: Variable description

Variable	Description
Track.Name	Song Title
Artist.Name	Artist performing the song
Genre	The genre of the track
Beats.Per.Minute	Variable describing the tempo of the song.
Energy	The energy of a song - the higher the value, the more energtic song.
Danceability	The higher the value, the easier it is to dance to this song.
Loudness..dB.	The higher the value, the louder the song.
Liveness	The higher the value, the more likely the song is a live recording.
Valence	The higher the value, the more positive mood for the song.
Length	The duration of the song.
Acousticness	The higher the value the more acoustic the song is.
Speechiness	The higher the value the more spoken word the song contains.
Popularity	The higher the value the more popular the song is.

Exploratory data analysis

Installing and running the libraries:

```r
requiredPackages = c("tidyverse","factoextra","stats","clustertend","flexclust","ggforce"
,"fpc","cluster","ClusterR","knitr","kableExtra","DataExplorer","caret",
"reshape2","corrplot","labdsv","smacof","clusterSim","pastecs","psych","pca3d","pls")
for(i in requiredPackages){if(!require(i,character.only = TRUE)) install.packages(i)}
for(i in requiredPackages){library(i,character.only = TRUE) }
```

Loading the data:

[1] https://www.forbes.com/sites/hughmcintyre/2019/04/02/the-global-music-industry-hits-19-billion-in-sales-in-2018-jumping-by-almost-10/#52c7da3e18a9

[2] Dataset available at: https://www.kaggle.com/leonardopena/top50spotify2019/data

```
data_full <- read.csv("top50.csv",stringsAsFactors = F)
data_full <- data_full[,2:14]
data_init <- data_full
```

Table 2: Summary statistics of the dataset

	min	max	range	median	mean	SE.mean	CI.mean.0.95	var	std.dev
Beats.Per.Minute	85	190	105	104.5	120.06	4.37	8.78	954.71	30.90
Energy	32	88	56	66.5	64.06	2.01	4.04	202.55	14.23
Danceability	29	90	61	73.5	71.38	1.69	3.39	142.32	11.93
Loudness..dB..	-11	-2	9	-6.0	-5.66	0.29	0.58	4.23	2.06
Liveness	5	58	53	11.0	14.66	1.57	3.16	123.62	11.12
Valence.	10	95	85	55.5	54.60	3.16	6.35	498.90	22.34
Length.	115	309	194	198.0	200.96	5.54	11.12	1532.24	39.14
Acousticness..	1	75	74	15.0	22.16	2.69	5.40	360.83	19.00
Speechiness.	3	46	43	7.0	12.48	1.58	3.17	124.58	11.16
Popularity	70	95	25	88.0	87.50	0.64	1.28	20.17	4.49

Songs in the data set are grouped into 21 genres.

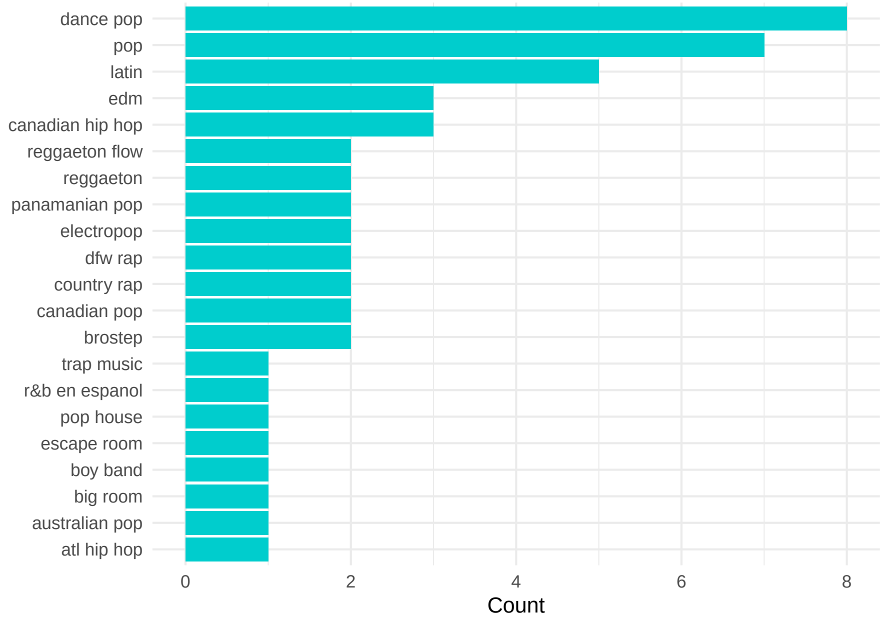

Figure 1: Genre distribution.

As there a lot of subgenres with a low number of instances, a new variable `aggregate.genres` is created.

```
data_init$aggregate.genre <- NA
data_init$aggregate.genre <- ifelse(data_init$Genre  %in% c("canadian pop","pop","dance pop",
                        "electropop","panamanian pop","pop house","australian pop"),"pop",
ifelse(data_init$Genre  %in% c("dfw rap","country rap","canadian hip hop","atl hip hop"),"hip hop",
ifelse(data_init$Genre  %in% c("r&b en espanol","latin"),"latin",
ifelse(data_init$Genre  %in% c("reggaeton flow","reggeaeton"),"reggaeton",
ifelse(data_init$Genre  %in% c("edm","trap music","brostep"),"electronic",
ifelse(data_init$Genre  %in% c("escape room","boy band","big room"),"other",data_init$Genre))))))
```

```r
data_init$aggregate.genre <- as.factor(data_init$aggregate.genre)
```

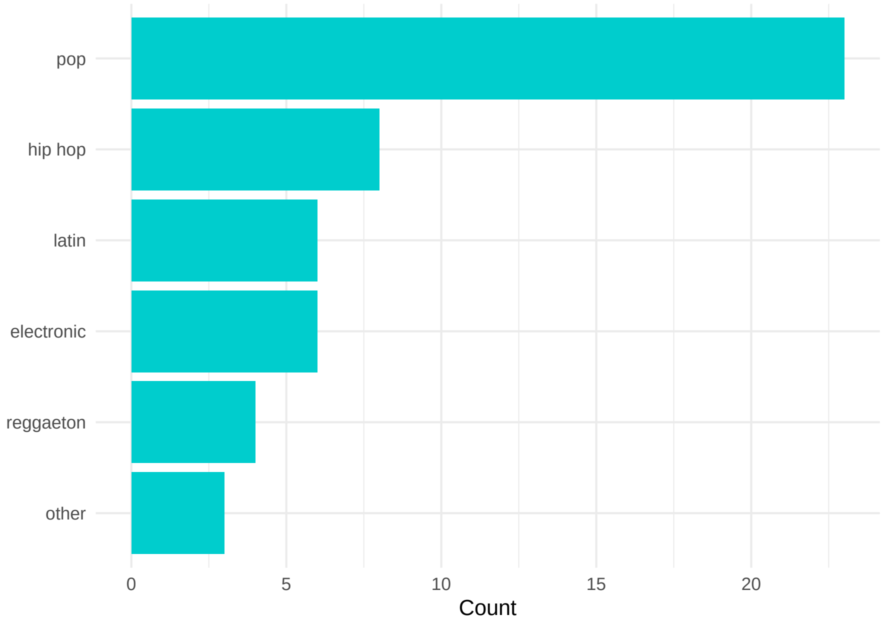

Figure 2: Aggregated genre distribution.

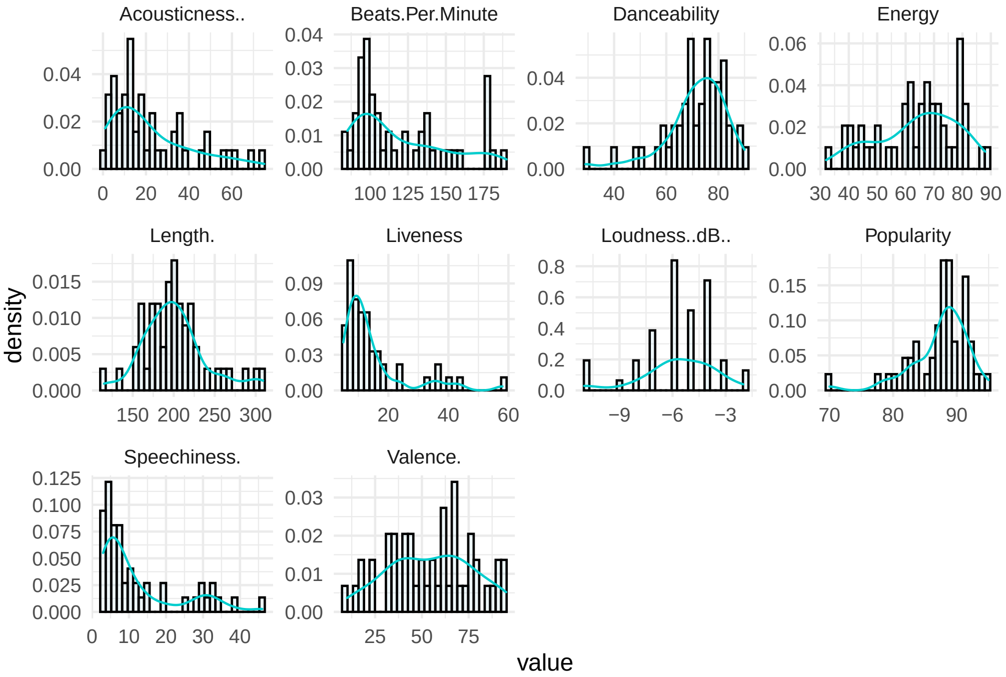

Figure 3: Histograms of numerical variables.

Variables `Liveness` and `Speechiness.` are highly positively skewed, suggesting the existence of outliers.

```
transformed_var <- c("Liveness", "Speechiness.")
data_init <- data_init %>% mutate_at(vars(transformed_var), log)
```

Distribution of variables `Liveness` and `Speechiness.` after log transformation

To further investigate the relationship between variables, a correlation analysis was carried using a Pearson correlation coefficient.

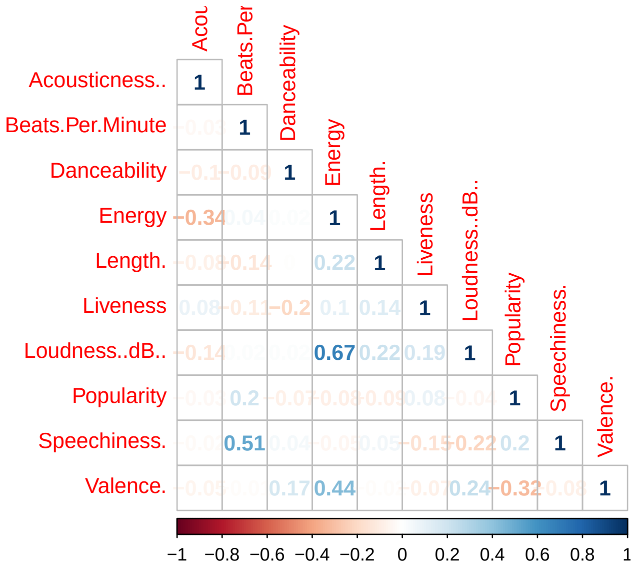

Figure 4: Correlation matrix.

From Figure 4, we can see a high positive correlation between variables `Speechiness` and `Beats.Per.Minute`, typical for rap songs. Another relationship is visible between variables `Energy` and `Loudness.dB.` suggesting that highly energetic songs tend to be louder. Also, a more energetic song, on average less acoustic.