What makes a song hit top 100 on Spotify?

Installing necessary packages:

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(CARS)
```

Summarize the data to see what variables we are working with:

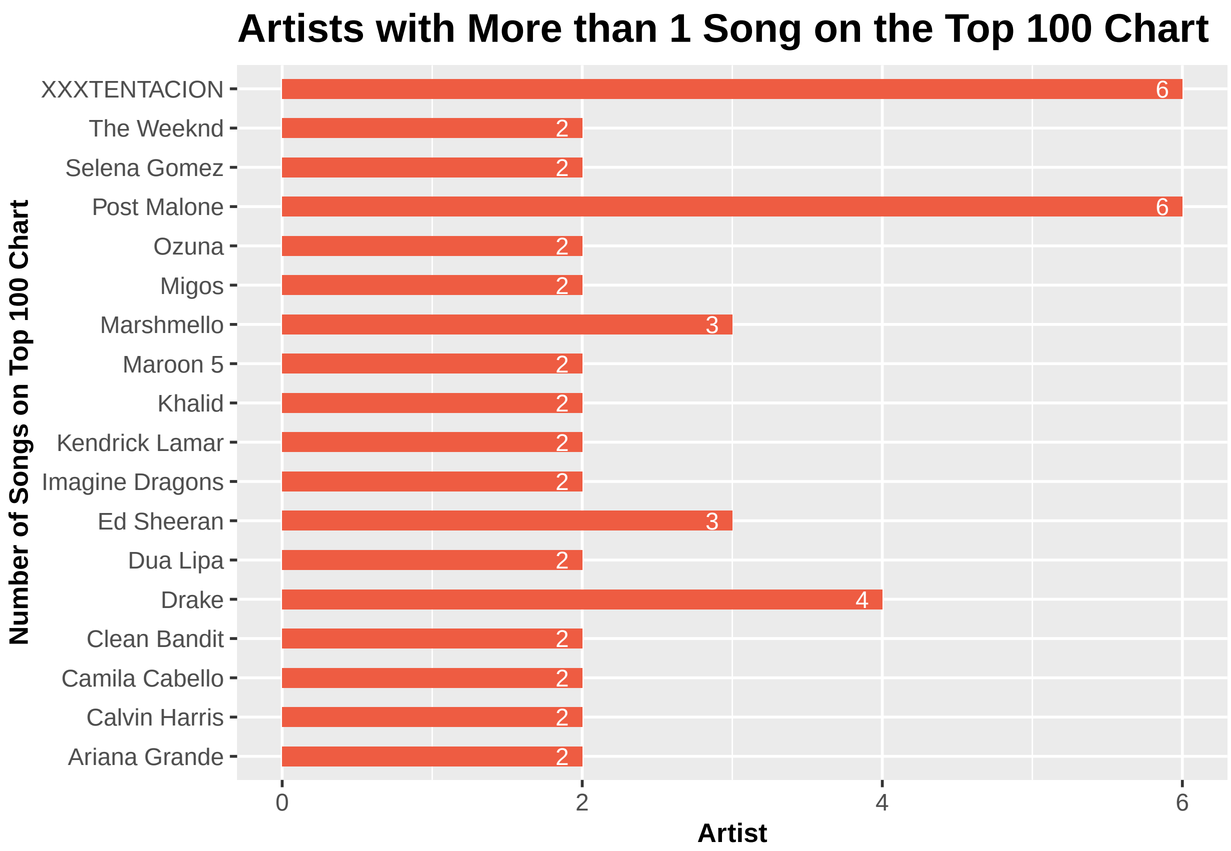
```r
spotify_data <- read.csv("~/Desktop/projects/stats_projects/spotify_top100/top2018.csv")
summary(spotify_data)
```

```
##                       id
##  08bNPGLD8AhKpnnERrAc6: 1
##  09IStsImFySgyp0pIQdqA: 1
##  0d2iYfpKoM0QCKvcLCkBa: 1
##  0e7ipj03S05BNilyu5bRz: 1
##  0E9ZjEAyAw0XZ7wJC0PD3: 1
##  0JP9xo3adEtGSdUEISisz: 1
##  (Other)              :94
##                                                        name
##  ?chame La Culpa                                        : 1
##  1, 2, 3 (feat. Jason Derulo & De La Ghetto)            : 1
##  2002                                                   : 1
##  All The Stars (with SZA)                               : 1
##  Back To You - From 13 Reasons Why ? Season 2 Soundtrack: 1
##  Be Alright                                             : 1
##  (Other)                                                :94
##           artists    danceability        energy             key
##  Post Malone  : 6   Min.   :0.2580   Min.   :0.2960   Min.   : 0.00
##  XXXTENTACION : 6   1st Qu.:0.6355   1st Qu.:0.5620   1st Qu.: 1.75
##  Drake        : 4   Median :0.7330   Median :0.6780   Median : 5.00
##  Ed Sheeran   : 3   Mean   :0.7165   Mean   :0.6591   Mean   : 5.33
##  Marshmello   : 3   3rd Qu.:0.7983   3rd Qu.:0.7722   3rd Qu.: 8.25
##  Ariana Grande: 2   Max.   :0.9640   Max.   :0.9090   Max.   :11.00
##  (Other)      :76
##     loudness            mode       speechiness       acousticness
##  Min.   :-10.109   Min.   :0.00   Min.   :0.02320   Min.   :0.000282
##  1st Qu.: -6.651   1st Qu.:0.00   1st Qu.:0.04535   1st Qu.:0.040225
##  Median : -5.566   Median :1.00   Median :0.07495   Median :0.109000
##  Mean   : -5.678   Mean   :0.59   Mean   :0.11557   Mean   :0.195701
##  3rd Qu.: -4.364   3rd Qu.:1.00   3rd Qu.:0.13700   3rd Qu.:0.247750
##  Max.   : -2.384   Max.   :1.00   Max.   :0.53000   Max.   :0.934000
##
##  instrumentalness       liveness          valence           tempo
##  Min.   :0.000e+00   Min.   :0.02150   Min.   :0.0796   Min.   : 64.93
##  1st Qu.:0.000e+00   1st Qu.:0.09467   1st Qu.:0.3410   1st Qu.: 95.73
##  Median :0.000e+00   Median :0.11850   Median :0.4705   Median :120.12
##  Mean   :1.584e-03   Mean   :0.15830   Mean   :0.4844   Mean   :119.90
##  3rd Qu.:3.088e-05   3rd Qu.:0.17075   3rd Qu.:0.6415   3rd Qu.:140.02
##  Max.   :1.340e-01   Max.   :0.63600   Max.   :0.9310   Max.   :198.07
##
##   duration_ms     time_signature
##  Min.   : 95467   Min.   :3.00
##  1st Qu.:184680   1st Qu.:4.00
##  Median :205048   Median :4.00
##  Mean   :205207   Mean   :3.98
##  3rd Qu.:221493   3rd Qu.:4.00
##  Max.   :417920   Max.   :5.00
##
```

Determine which artists have more than 1 song on the top 100 songs list:

```
artists <- spotify_data$artists
n_occur <- data.frame(table(artists))
top_artists <- n_occur[n_occur$Freq > 1,]

ggplot(top_artists,aes(x=top_artists$artists,y=top_artists$Freq,label=top_artists$artists)) + geom_bar(
coord_flip() +
labs(x='Number of Songs on Top 100 Chart',y='Artist',title='Artists with More than 1 Song on the Top 10
theme(plot.title=element_text(face='bold',size=15),axis.title=element_text(face='bold',size=10))
```



Exploring the correlation between "danceability" with different variables of the top 100 songs: (insert spotify's definition of song factors)

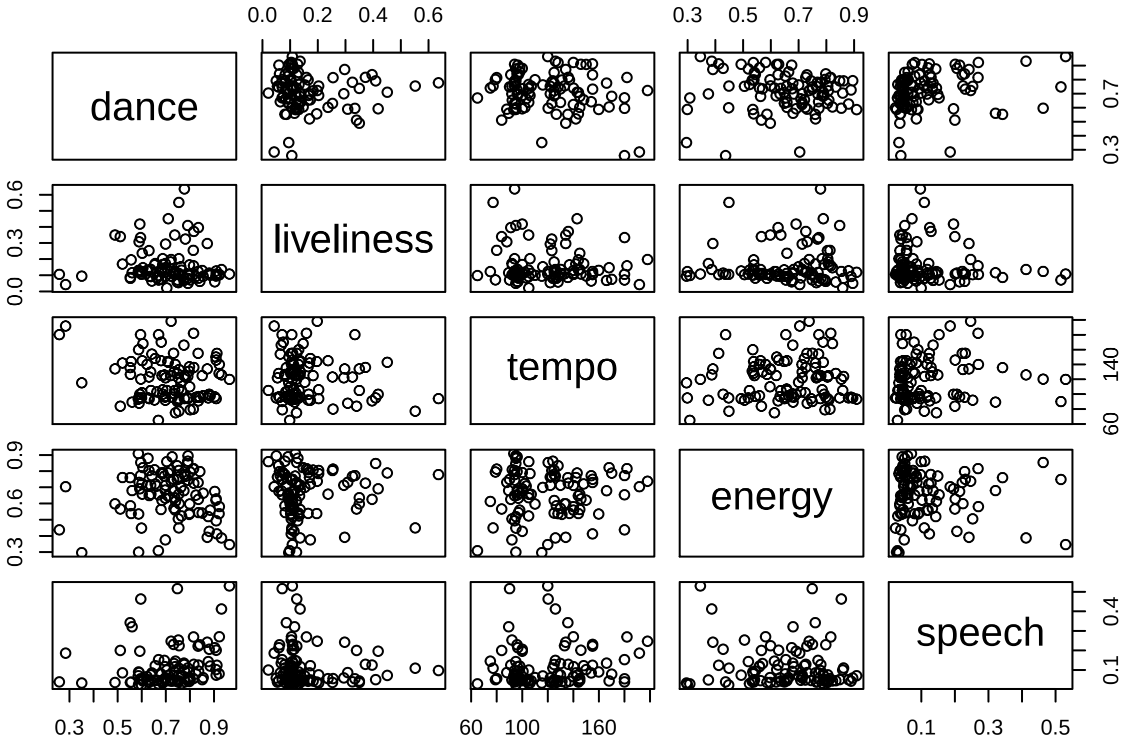
```
dance <- spotify_data$danceability
energy <- spotify_data$energy
loudness <- spotify_data$loudness
speech <- spotify_data$speechiness
acoustics <- spotify_data$acousticness
liveliness <- spotify_data$liveness
valence <- spotify_data$valence
tempo <- spotify_data$tempo
duration <- spotify_data$duration_ms
```

Build an initial model based on our intuition about "danceability":

```
# Guess: danceability is correlated to liveliness, tempo, energy, and speechiness
model01 <- lm(dance ~ liveliness + tempo + energy + speech)
summary(model01)
```

```
##
## Call:
## lm(formula = dance ~ liveliness + tempo + energy + speech)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38492 -0.06004  0.01572  0.08802  0.22203
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8321300  0.0813538  10.229   <2e-16 ***
## liveliness  -0.0427891  0.1153536  -0.371   0.7115
## tempo       -0.0010085  0.0004482  -2.250   0.0267 *
## energy      -0.0352632  0.0883805  -0.399   0.6908
## speech       0.3052108  0.1233625   2.474   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1267 on 95 degrees of freedom
## Multiple R-squared:  0.1027, Adjusted R-squared:  0.06495
## F-statistic: 2.719 on 4 and 95 DF,  p-value: 0.03414
```

```r
pairs(dance ~ liveliness + tempo + energy + speech)
```



(analysis of AVplots + pair plots)

It appears that liveliness, energy, and speech are weakly correlated to the "danceability" of a song.

```r
model02 <- lm(dance~tempo+speech)
summary(model02)
```

```
##
```

```
## Call:
## lm(formula = dance ~ tempo + speech)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38238 -0.06152  0.01630  0.08734  0.22814
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8007320  0.0547357  14.629   <2e-16 ***
## tempo       -0.0010048  0.0004408  -2.279   0.0248 *
## speech       0.3132473  0.1214455   2.579   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1256 on 97 degrees of freedom
## Multiple R-squared:  0.09977,    Adjusted R-squared:  0.08121
## F-statistic: 5.375 on 2 and 97 DF,  p-value: 0.00611
```