# SongPopularityAnalysis

Yasko

2022-10-07

```
library(readr)
library(plyr)
library(dplyr)
library(ggplot2)
library(formattable)
library(wordcloud)
library(RWeka)
library(qdap)
library(tm)
```

```
spotify_data <- read_csv('featuresdf.csv')
daily_spotify <- read_csv("data.csv")
```

```
glimpse(spotify_data)
```

```
## Rows: 100
## Columns: 16
## $ id               <chr> "7qiZfU4dY1lWllzX7mPBI", "5CtI0qwDJkDQGwXD1H1cL", "4a~
## $ name             <chr> "Shape of You", "Despacito - Remix", "Despacito (Feat~
## $ artists          <chr> "Ed Sheeran", "Luis Fonsi", "Luis Fonsi", "The Chains~
## $ danceability     <dbl> 0.825, 0.694, 0.660, 0.617, 0.609, 0.904, 0.640, 0.72~
## $ energy           <dbl> 0.652, 0.815, 0.786, 0.635, 0.668, 0.611, 0.533, 0.76~
## $ key              <dbl> 1, 2, 2, 11, 7, 1, 0, 6, 1, 0, 11, 2, 5, 3, 2, 6, 1, ~
## $ loudness         <dbl> -3.183, -4.328, -4.757, -6.769, -4.284, -6.842, -6.59~
## $ mode             <dbl> 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1,~
## $ speechiness      <dbl> 0.0802, 0.1200, 0.1700, 0.0317, 0.0367, 0.0888, 0.070~
## $ acousticness     <dbl> 0.581000, 0.229000, 0.209000, 0.049800, 0.055200, 0.0~
## $ instrumentalness <dbl> 0.00e+00, 0.00e+00, 0.00e+00, 1.44e-05, 0.00e+00, 2.0~
## $ liveness         <dbl> 0.0931, 0.0924, 0.1120, 0.1640, 0.1670, 0.0976, 0.086~
## $ valence          <dbl> 0.9310, 0.8130, 0.8460, 0.4460, 0.8110, 0.4000, 0.515~
## $ tempo            <dbl> 95.977, 88.931, 177.833, 103.019, 80.924, 150.020, 99~
## $ duration_ms      <dbl> 233713, 228827, 228200, 247160, 288600, 177000, 22078~
## $ time_signature   <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,~
```

```
glimpse(daily_spotify)
```

```
## Rows: 3,441,197
## Columns: 7
## $ Position     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ `Track Name` <chr> "Reggaeton Lento (Bailemos)", "Chantaje", "Otra Vez (feat~
## $ Artist       <chr> "CNCO", "Shakira", "Zion & Lennox", "Ricky Martin", "J Ba~
## $ Streams      <dbl> 19272, 19270, 15761, 14954, 14269, 12843, 10986, 10653, 9~
## $ URL          <chr> "https://open.spotify.com/track/3AEZUABDXNtecAOSC1qTfo", ~
## $ Date         <date> 2017-01-01, 2017-01-01, 2017-01-01, 2017-01-01, 2017-01-~
```

```
## $ Region       <chr> "ec", "ec", "ec", "ec", "ec", "ec", "ec", "ec", "ec", "ec~
```

```
summary(spotify_data)
```

```
##       id                name              artists           danceability
##  Length:100         Length:100         Length:100         Min.   :0.2580
##  Class :character   Class :character   Class :character   1st Qu.:0.6350
##  Mode  :character   Mode  :character   Mode  :character   Median :0.7140
##                                                           Mean   :0.6968
##                                                           3rd Qu.:0.7702
##                                                           Max.   :0.9270
##      energy            key           loudness            mode
##  Min.   :0.3460   Min.   : 0.00   Min.   :-11.462   Min.   :0.00
##  1st Qu.:0.5565   1st Qu.: 2.00   1st Qu.: -6.595   1st Qu.:0.00
##  Median :0.6675   Median : 6.00   Median : -5.437   Median :1.00
##  Mean   :0.6607   Mean   : 5.57   Mean   : -5.653   Mean   :0.58
##  3rd Qu.:0.7875   3rd Qu.: 9.00   3rd Qu.: -4.327   3rd Qu.:1.00
##  Max.   :0.9320   Max.   :11.00   Max.   : -2.396   Max.   :1.00
##   speechiness       acousticness      instrumentalness       liveness
##  Min.   :0.02320   Min.   :0.000259   Min.   :0.000e+00   Min.   :0.04240
##  1st Qu.:0.04312   1st Qu.:0.039100   1st Qu.:0.000e+00   1st Qu.:0.09828
##  Median :0.06265   Median :0.106500   Median :0.000e+00   Median :0.12500
##  Mean   :0.10397   Mean   :0.166306   Mean   :4.796e-03   Mean   :0.15061
##  3rd Qu.:0.12300   3rd Qu.:0.231250   3rd Qu.:1.335e-05   3rd Qu.:0.17925
##  Max.   :0.43100   Max.   :0.695000   Max.   :2.100e-01   Max.   :0.44000
##     valence           tempo         duration_ms     time_signature
##  Min.   :0.0862   Min.   : 75.02   Min.   :165387   Min.   :3.00
##  1st Qu.:0.3755   1st Qu.: 99.91   1st Qu.:198491   1st Qu.:4.00
##  Median :0.5025   Median :112.47   Median :214106   Median :4.00
##  Mean   :0.5170   Mean   :119.20   Mean   :218387   Mean   :3.99
##  3rd Qu.:0.6790   3rd Qu.:137.17   3rd Qu.:230543   3rd Qu.:4.00
##  Max.   :0.9660   Max.   :199.86   Max.   :343150   Max.   :4.00
```

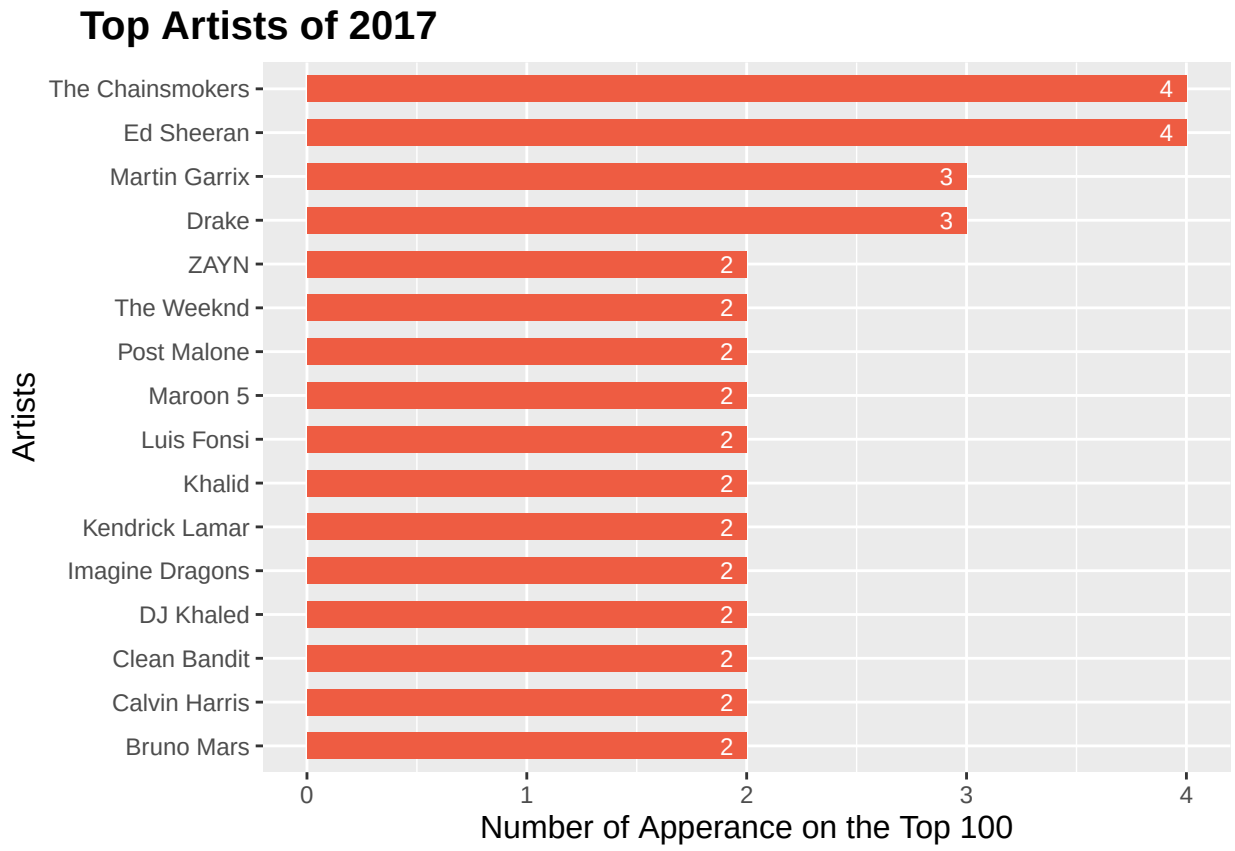
```
spotify_data$duration_ms <- round(spotify_data$duration_ms / 1000)
colnames(spotify_data)[15] <- "duration"
summary(spotify_data)
```

```
##       id                name              artists           danceability
##  Length:100         Length:100         Length:100         Min.   :0.2580
##  Class :character   Class :character   Class :character   1st Qu.:0.6350
##  Mode  :character   Mode  :character   Mode  :character   Median :0.7140
##                                                           Mean   :0.6968
##                                                           3rd Qu.:0.7702
##                                                           Max.   :0.9270
##      energy            key           loudness            mode
##  Min.   :0.3460   Min.   : 0.00   Min.   :-11.462   Min.   :0.00
##  1st Qu.:0.5565   1st Qu.: 2.00   1st Qu.: -6.595   1st Qu.:0.00
##  Median :0.6675   Median : 6.00   Median : -5.437   Median :1.00
##  Mean   :0.6607   Mean   : 5.57   Mean   : -5.653   Mean   :0.58
##  3rd Qu.:0.7875   3rd Qu.: 9.00   3rd Qu.: -4.327   3rd Qu.:1.00
##  Max.   :0.9320   Max.   :11.00   Max.   : -2.396   Max.   :1.00
##   speechiness       acousticness      instrumentalness       liveness
##  Min.   :0.02320   Min.   :0.000259   Min.   :0.000e+00   Min.   :0.04240
##  1st Qu.:0.04312   1st Qu.:0.039100   1st Qu.:0.000e+00   1st Qu.:0.09828
##  Median :0.06265   Median :0.106500   Median :0.000e+00   Median :0.12500
```

```
##  Mean   :0.10397   Mean   :0.166306   Mean   :4.796e-03   Mean   :0.15061
##  3rd Qu.:0.12300   3rd Qu.:0.231250   3rd Qu.:1.335e-05   3rd Qu.:0.17925
##  Max.   :0.43100   Max.   :0.695000   Max.   :2.100e-01   Max.   :0.44000
##     valence           tempo           duration     time_signature
##  Min.   :0.0862   Min.   : 75.02   Min.   :165.0   Min.   :3.00
##  1st Qu.:0.3755   1st Qu.: 99.91   1st Qu.:198.8   1st Qu.:4.00
##  Median :0.5025   Median :112.47   Median :214.0   Median :4.00
##  Mean   :0.5170   Mean   :119.20   Mean   :218.4   Mean   :3.99
##  3rd Qu.:0.6790   3rd Qu.:137.17   3rd Qu.:230.2   3rd Qu.:4.00
##  Max.   :0.9660   Max.   :199.86   Max.   :343.0   Max.   :4.00
```

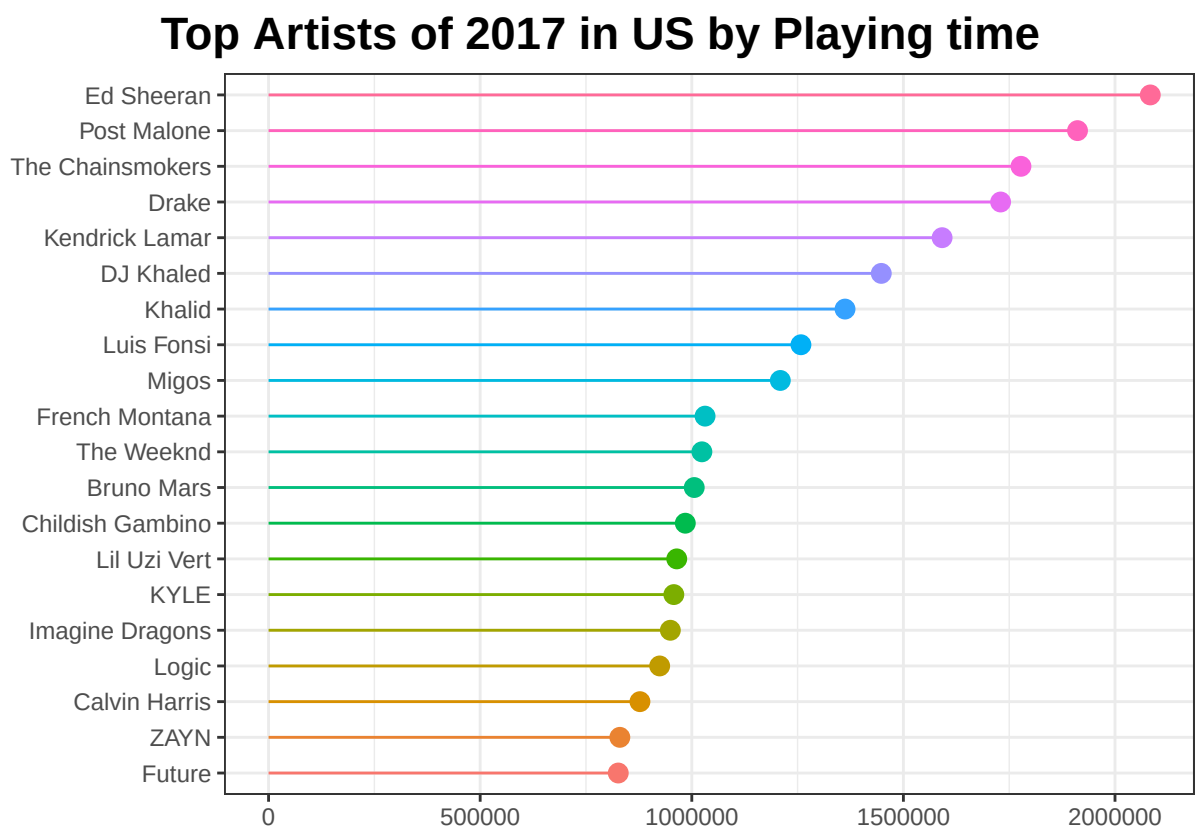
```
top_artist <- spotify_data %>% group_by(artists) %>% summarise(n_apperance = n()) %>% filter(n_apperanc
top_artist$artists <- factor(top_artist$artists, levels = top_artist$artists[order(top_artist$n_apperan
```

```
ggplot(data = top_artist, aes(x = artists, y = n_apperance)) +
  geom_bar(stat = "identity", fill = "tomato2", width = 0.6) +
  labs(title = "Top Artists of 2017", x = "Artists", y = "Number of Apperance on the Top 100") +
    theme(plot.title = element_text(size=15,hjust=-.3,face = "bold"), axis.title = element_text(size=12
    geom_text(aes(label=n_apperance), hjust = 2, size = 3, color = 'white') +
    coord_flip()
```




```
us_daily_spotify <- daily_spotify %>% filter(Region == 'us') %>% group_by(`Track Name`) %>% summarise(t
names(us_daily_spotify)[1] <- paste('name')
top_by_playtime <- spotify_data %>% left_join(us_daily_spotify, by = "name") %>% select(name, artists, 
  mutate(total_time = duration * total_streams / 60000)
top20_by_playtime <- top_by_playtime %>% group_by(artists) %>% summarise(n_time = sum(total_time)) %>% 
top20_by_playtime$artists <- factor(top20_by_playtime$artists, levels = top20_by_playtime$artists [orde
```

```
ggplot(top20_by_playtime, aes(x=artists, y=n_time, color=artists)) +
    geom_point(size=3) +
    geom_segment(aes(x=artists,xend=artists, y=0, yend=n_time)) +
    labs(title = "Top Artists of 2017 in US by Playing time", x='',y='') +
    theme_bw() +
    theme(legend.position = 'none', plot.title = element_text(size=17,hjust = -0.7, face = "bold"), axis
    coord_flip()
```

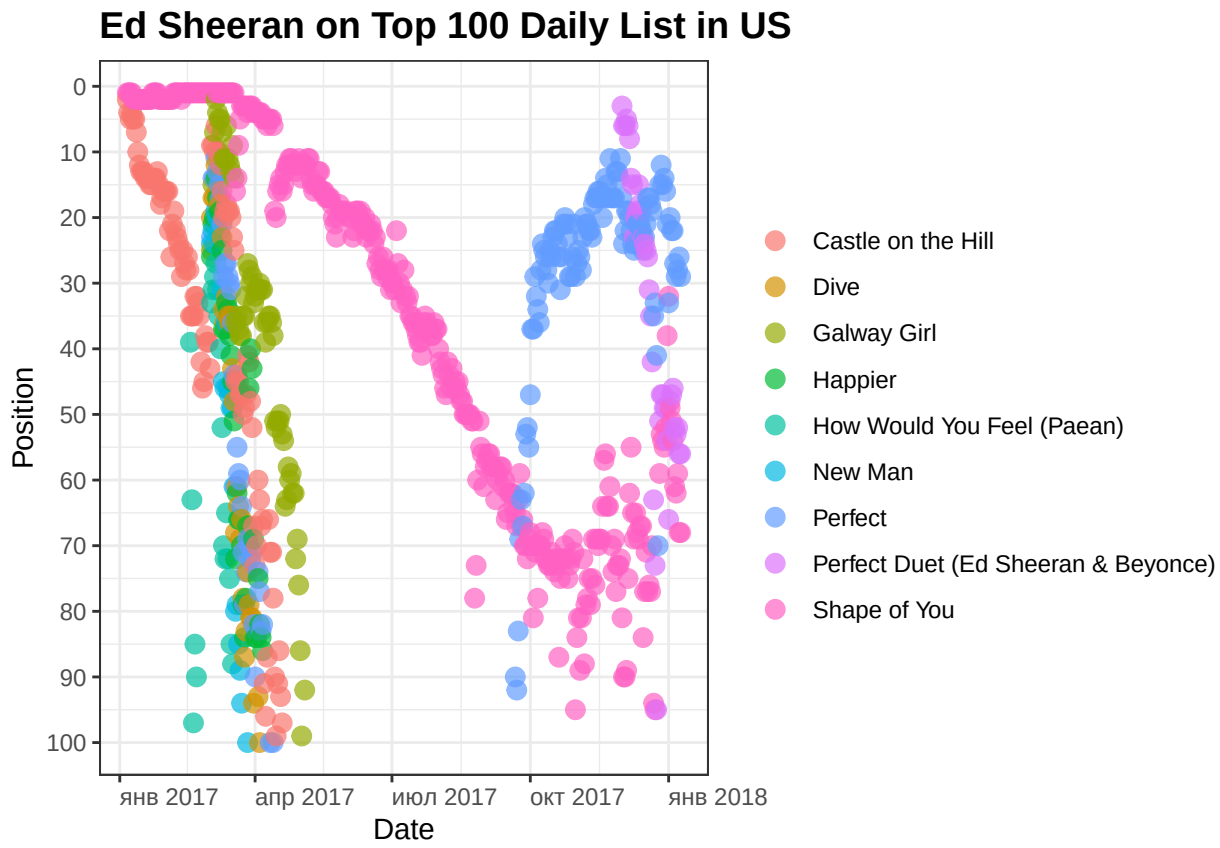



```
ed_sheeran_daily <- daily_spotify %>% filter(Region == 'us', Artist == 'Ed Sheeran', Position <= 100)
formatted_ed <- ed_sheeran_daily %>% group_by(`Track Name`) %>% summarise(n_daily = n()) %>% arrange(des
formatted_ed
```

```
## # A tibble: 19 x 2
##    `Track Name`                          n_daily
##    <chr>                                   <int>
##  1 Shape of You                              364
##  2 Perfect                                   148
##  3 Castle on the Hill                        104
##  4 Galway Girl                                62
##  5 Perfect Duet (Ed Sheeran & Beyonce)        39
##  6 Happier                                    35
##  7 Dive                                       31
##  8 New Man                                    22
##  9 How Would You Feel (Paean)                 20
## 10 What Do I Know?                            19
## 11 Barcelona                                  17
```

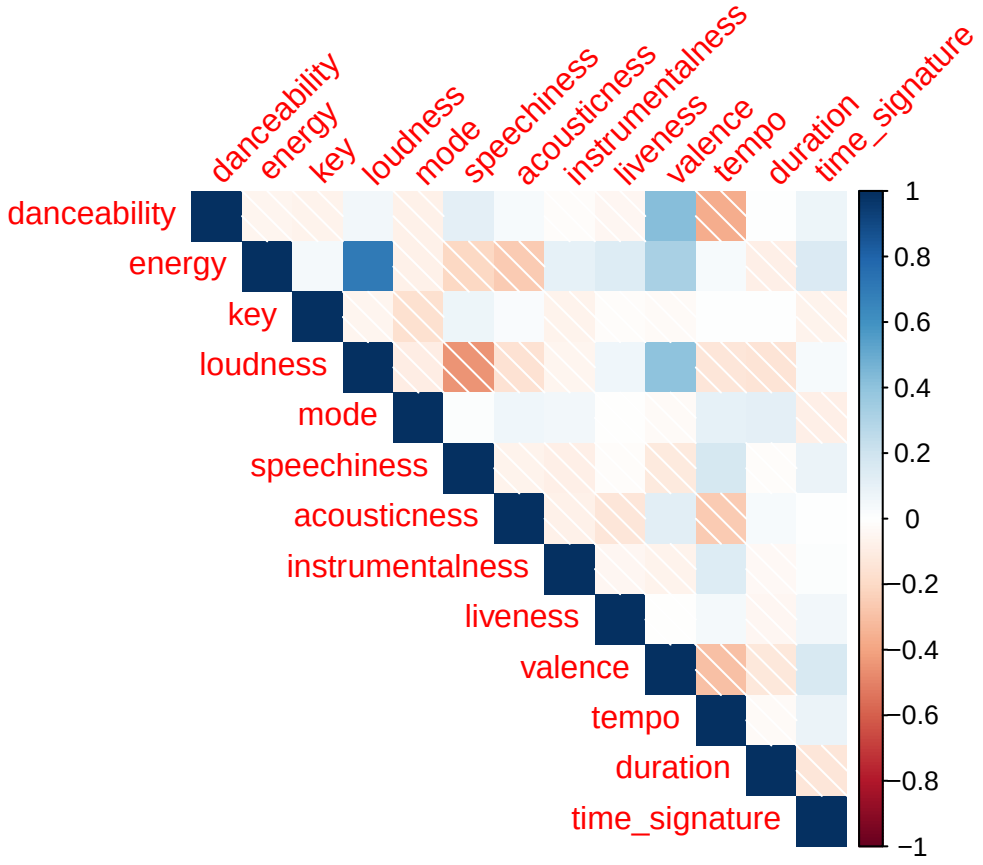
```
## 12 Supermarket Flowers                      17
## 13 Nancy Mulligan                           16
## 14 Hearts Don't Break Around Here           15
## 15 Bibia Be Ye Ye                           14
## 16 Eraser                                   14
## 17 Save Myself                              14
## 18 Thinking Out Loud                         6
## 19 Photograph                                5
```

```
ed_20 <- ed_sheeran_daily %>% group_by(`Track Name`) %>% summarise(n_daily = n()) %>% filter(n_daily >=
ed_20 <- ed_20 %>% collect %>% .[["Track Name"]]
ed_daily_plot <- ed_sheeran_daily %>%
    filter(`Track Name` %in% ed_20) %>%
    ggplot(aes(x = Date, y = Position, col = `Track Name`)) +
    geom_point(alpha = 0.7, size = 3) +
    scale_y_reverse(breaks = seq(0,100,10)) +
    scale_x_date() +
    ggtitle("Ed Sheeran on Top 100 Daily List in US") +
    theme_bw() +
    theme(plot.title = element_text(size = 14, face = "bold")) +
    theme(legend.title=element_blank())

ed_daily_plot
```

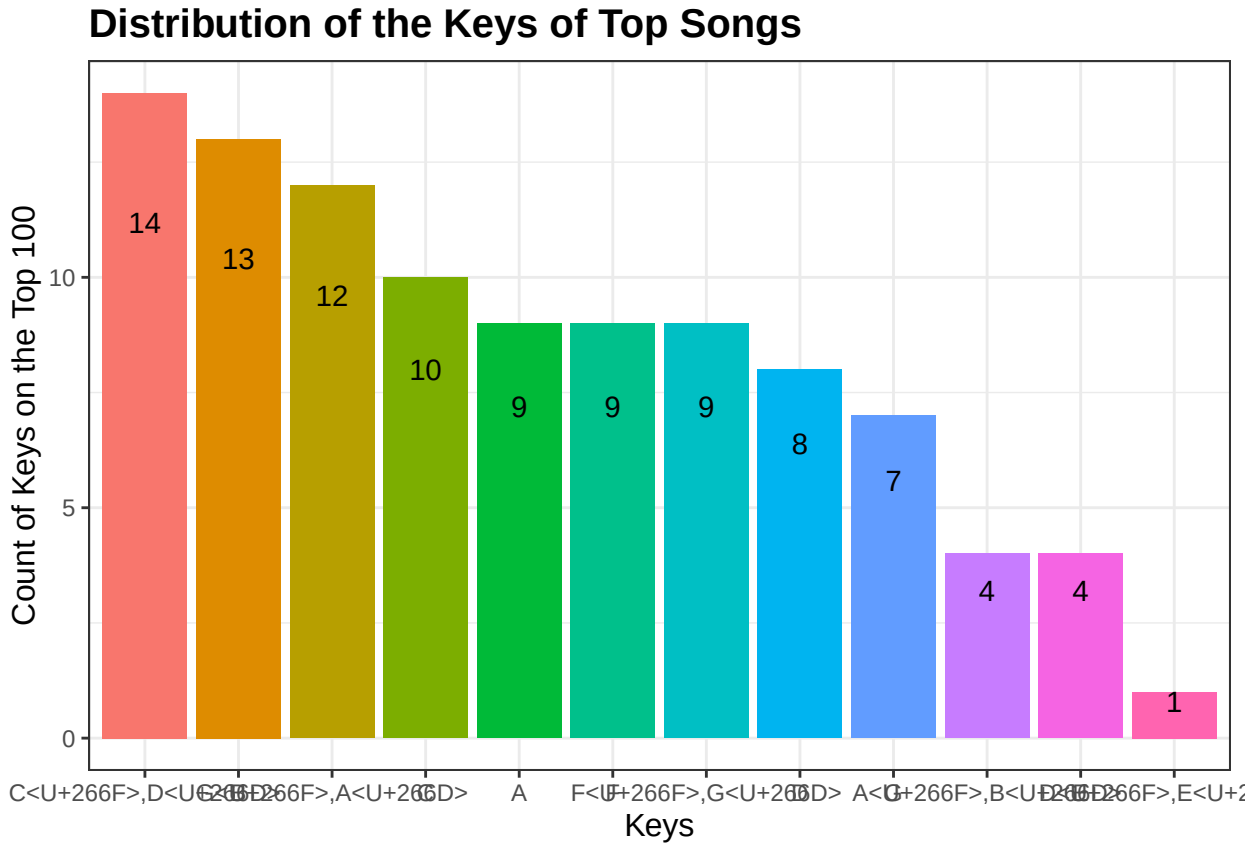



```
library(corrplot)
corrData <- spotify_data[, -(1:3)]
mtC <- cor(corrData)
```

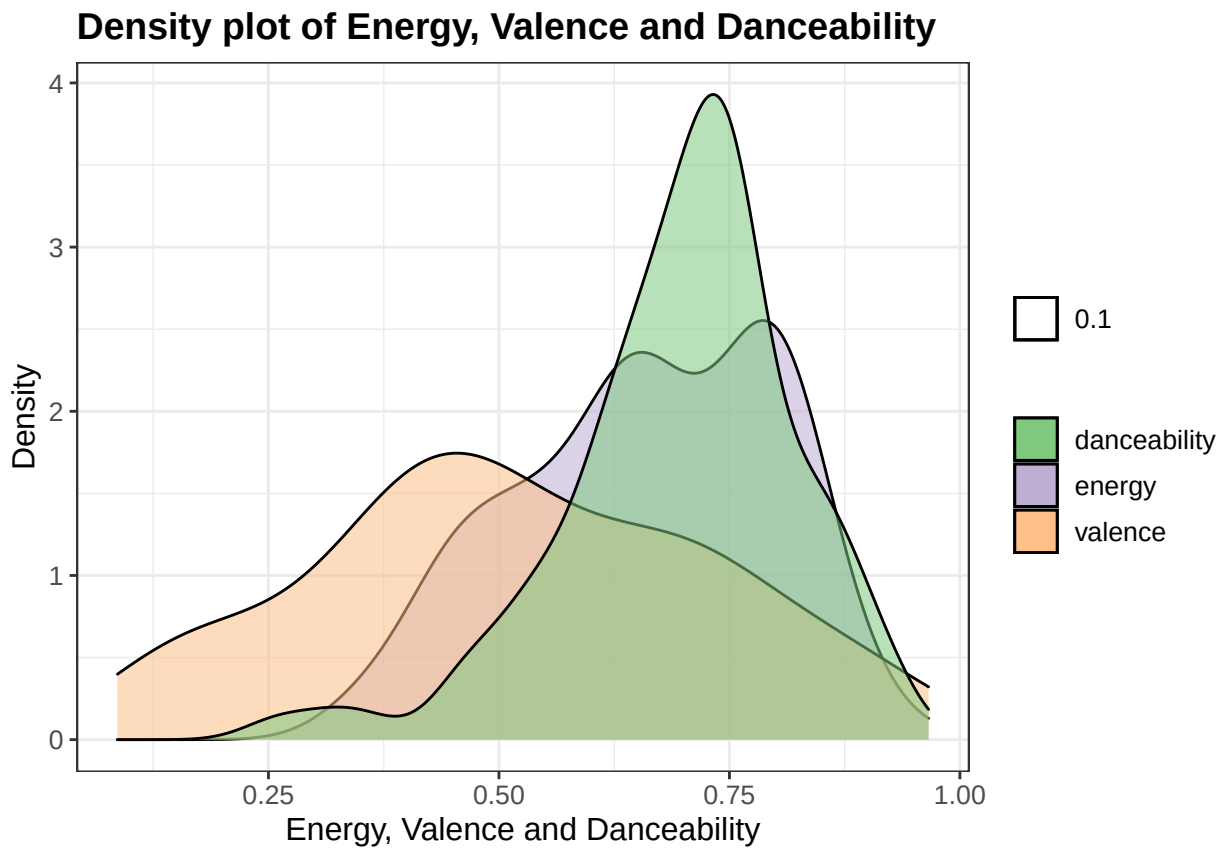
```
corrplot(mtC, method = "shade", type = "upper", tl.srt = 45)
```




```
spotify_data$key <- as.character(spotify_data$key)
spotify_data$key <- revalue(spotify_data$key, c("0" = "C", "1" = "C ,D ", "2" = "D", "3" = "D ,E ", "4" =

song_keys <- spotify_data %>%
    group_by(key) %>%
    summarise(n_key = n()) %>%
    arrange(desc(n_key))

song_keys$key <- factor(song_keys$key, levels = song_keys$key[order(song_keys$n_key)]) # in order to vi

ggplot(song_keys, aes(x = reorder(key,-n_key), y = n_key, fill = reorder(key,-n_key))) +
    geom_bar(stat = "identity") +
    labs(title = "Distribution of the Keys of Top Songs", x = "Keys", y = "Count of Keys on the Top 100
    geom_text(aes(label=n_key), position = position_stack(vjust = 0.8)) +
    theme_bw() +
    theme(plot.title = element_text(size=15,face = "bold"), axis.title = element_text(size=12)) +
    theme(legend.position="none")
```

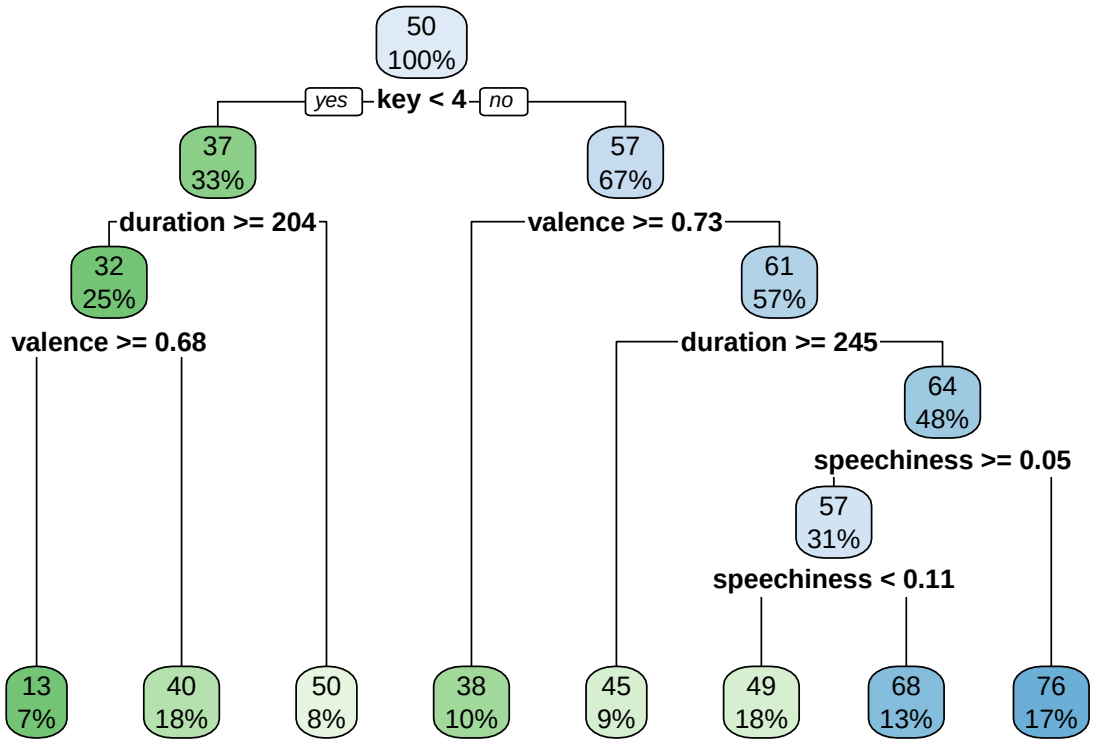



```
correlated_density <- ggplot(spotify_data) +
    geom_density(aes(energy, fill ="energy", alpha = 0.1)) +
    geom_density(aes(valence, fill ="valence", alpha = 0.1)) +
    geom_density(aes(danceability, fill ="danceability", alpha = 0.1)) +
    scale_x_continuous(name = "Energy, Valence and Danceability") +
    scale_y_continuous(name = "Density") +
    ggtitle("Density plot of Energy, Valence and Danceability") +
    theme_bw() +
    theme(plot.title = element_text(size = 14, face = "bold"),
          text = element_text(size = 12)) +
    theme(legend.title=element_blank()) +
    scale_fill_brewer(palette="Accent")

correlated_density
```




```
loudness_density <- ggplot(spotify_data) +
    geom_density(aes(loudness, fill ="loudness")) +
    scale_x_continuous(name = "Loudness") +
    scale_y_continuous(name = "Density") +
    ggtitle("Density plot of Loudness") +
    theme_bw() +
    theme(plot.title = element_text(size = 14, face = "bold"),
            text = element_text(size = 12)) +
    theme(legend.title=element_blank()) +
    scale_fill_brewer(palette="Paired")

print(loudness_density)
```


```
library(rpart)
library(rpart.plot)
corrData$standing <- c(1:100)
tree_model <- rpart(standing ~ ., data = corrData)
rpart.plot(tree_model, box.palette = "GnBu")
```




```
qdap_clean <- function(x) {
    x <- replace_abbreviation(x)
    x <- replace_contraction(x)
    x <- replace_number(x)
    x <- replace_ordinal(x)
    x <- tolower(x)

    return(x)

    }



tm_clean <- function(corpus) {
    corpus <- tm_map(corpus, content_transformer(strip), char.keep="$")
    corpus <- tm_map(corpus, stripWhitespace)
    corpus <- tm_map(corpus, removeWords,
                     c(stopwords("en"), "with", "feat", "ty"))
    return(corpus)

}
tokenizer <- function(x)
    NGramTokenizer(x, Weka_control(min = 2, max = 3))
```

```
us_top100_titles <- daily_spotify %>%
    filter(Region == "us", Position <= 100) %>%
```

```
    select(`Track Name`) %>%
    filter(grepl('feat|with', `Track Name`))

us_top100_titles <- us_top100_titles[!duplicated(us_top100_titles$`Track Name`),]
us_top100_titles <- qdap_clean(us_top100_titles)
us_top100_corp <- VCorpus(VectorSource(us_top100_titles))
us_top100_corp_tm <- tm_clean(us_top100_corp)
us_top100_tdm <- TermDocumentMatrix(us_top100_corp_tm, control = list(tokenize = tokenizer))
us_top100_tdm_m <- as.matrix(us_top100_tdm)
us_top100_freq <- rowSums(us_top100_tdm_m)
wordcloud(names(us_top100_freq),us_top100_freq, min.freq = 2, max.words = 100, scale = c(3,.3), colors =
```