Assignment 1

2023-09-09

```
knitr::opts_chunk$set(echo = TRUE)
```

(1)Data-set name:“Most Streamed Spotify Songs 2023”

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
#The dataset used is taken from "Kaggle" site
#"https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023"

# (2) Dataset loading

spotify <- read_csv("/Users/ritikakalyani/Downloads/ritika/spotify-2023.csv")
```

```
## Rows: 953 Columns: 24
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): track_name, artist(s)_name, streams, key, mode
## dbl (17): artist_count, released_year, released_month, released_day, in_spot...
## num  (2): in_deezer_playlists, in_shazam_charts
##
## 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(spotify)
```

```
## # A tibble: 6 × 24
##   track_name          `artist(s)_name` artist_count released_year released_month
##   <chr>               <chr>                   <dbl>         <dbl>          <dbl>
## 1 Seven (feat. Latto… Latto, Jung Kook            2          2023              7
## 2 LALA                Myke Towers                 1          2023              3
## 3 vampire             Olivia Rodrigo              1          2023              6
## 4 Cruel Summer        Taylor Swift                1          2019              8
## 5 WHERE SHE GOES      Bad Bunny                   1          2023              5
## 6 Sprinter            Dave, Central C…            2          2023              6
## # ℹ 19 more variables: released_day <dbl>, in_spotify_playlists <dbl>,
## #   in_spotify_charts <dbl>, streams <chr>, in_apple_playlists <dbl>,
## #   in_apple_charts <dbl>, in_deezer_playlists <dbl>, in_deezer_charts <dbl>,
## #   in_shazam_charts <dbl>, bpm <dbl>, key <chr>, mode <chr>,
## #   `danceability_%` <dbl>, `valence_%` <dbl>, `energy_%` <dbl>,
## #   `acousticness_%` <dbl>, `instrumentalness_%` <dbl>, `liveness_%` <dbl>,
## #   `speechiness_%` <dbl>
```

```
spec(spotify)
```

```
## cols(
##   track_name = col_character(),
##   `artist(s)_name` = col_character(),
##   artist_count = col_double(),
##   released_year = col_double(),
##   released_month = col_double(),
##   released_day = col_double(),
##   in_spotify_playlists = col_double(),
##   in_spotify_charts = col_double(),
##   streams = col_character(),
##   in_apple_playlists = col_double(),
##   in_apple_charts = col_double(),
##   in_deezer_playlists = col_number(),
##   in_deezer_charts = col_double(),
##   in_shazam_charts = col_number(),
##   bpm = col_double(),
##   key = col_character(),
##   mode = col_character(),
##   `danceability_%` = col_double(),
##   `valence_%` = col_double(),
##   `energy_%` = col_double(),
##   `acousticness_%` = col_double(),
##   `instrumentalness_%` = col_double(),
##   `liveness_%` = col_double(),
##   `speechiness_%` = col_double()
## )
```

```
#(3)(a) Descriptive statistics for selection of quantitative variables

summary(spotify[,3:7])
```

```
##   artist_count   released_year  released_month    released_day
##  Min.   :1.000   Min.   :1930   Min.   : 1.000   Min.   : 1.00
##  1st Qu.:1.000   1st Qu.:2020   1st Qu.: 3.000   1st Qu.: 6.00
##  Median :1.000   Median :2022   Median : 6.000   Median :13.00
##  Mean   :1.556   Mean   :2018   Mean   : 6.034   Mean   :13.93
##  3rd Qu.:2.000   3rd Qu.:2022   3rd Qu.: 9.000   3rd Qu.:22.00
##  Max.   :8.000   Max.   :2023   Max.   :12.000   Max.   :31.00
##  in_spotify_playlists
##  Min.   :   31
##  1st Qu.:  875
##  Median : 2224
##  Mean   : 5200
##  3rd Qu.: 5542
##  Max.   :52898
```

```
##(3)(b) Descriptive statistics for selection of qualitative variables

summary(spotify[,1:2])
```

```
##   track_name         artist(s)_name
##  Length:953         Length:953
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
summary(spotify[,16:17])
```

```
##      key                mode
##  Length:953         Length:953
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
#(4) Variable transformation

spotify$Log_in_spotify_playlists <- log(spotify$in_spotify_playlists)
print(head(spotify$Log_in_spotify_playlists))
```

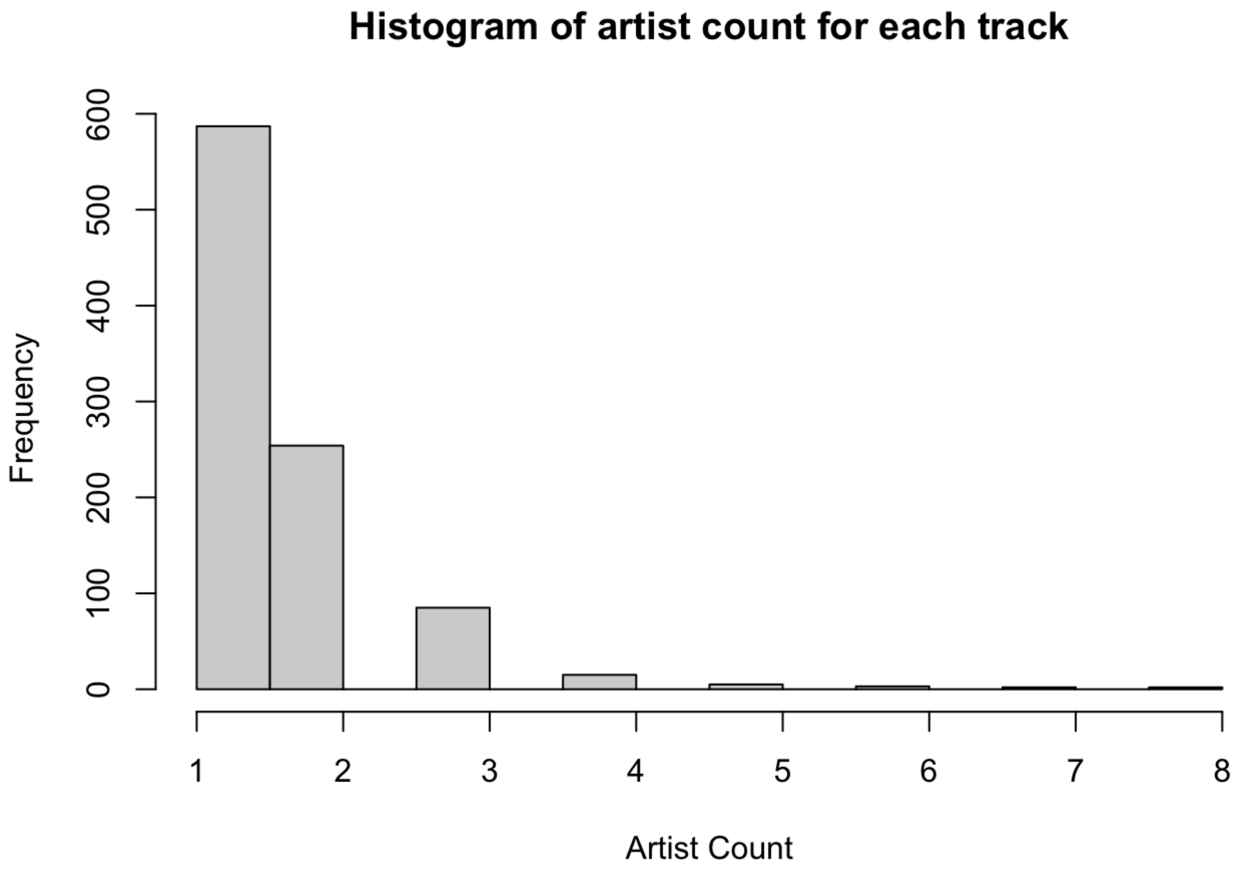
```
## [1] 6.315358 7.295735 7.242082 8.969287 8.049746 7.689829
```

```
spotify$beats_per_minute <- sqrt(spotify$bpm)
print(head(spotify$beats_per_minute))
```

```
## [1] 11.180340  9.591663 11.747340 13.038405 12.000000 11.874342
```

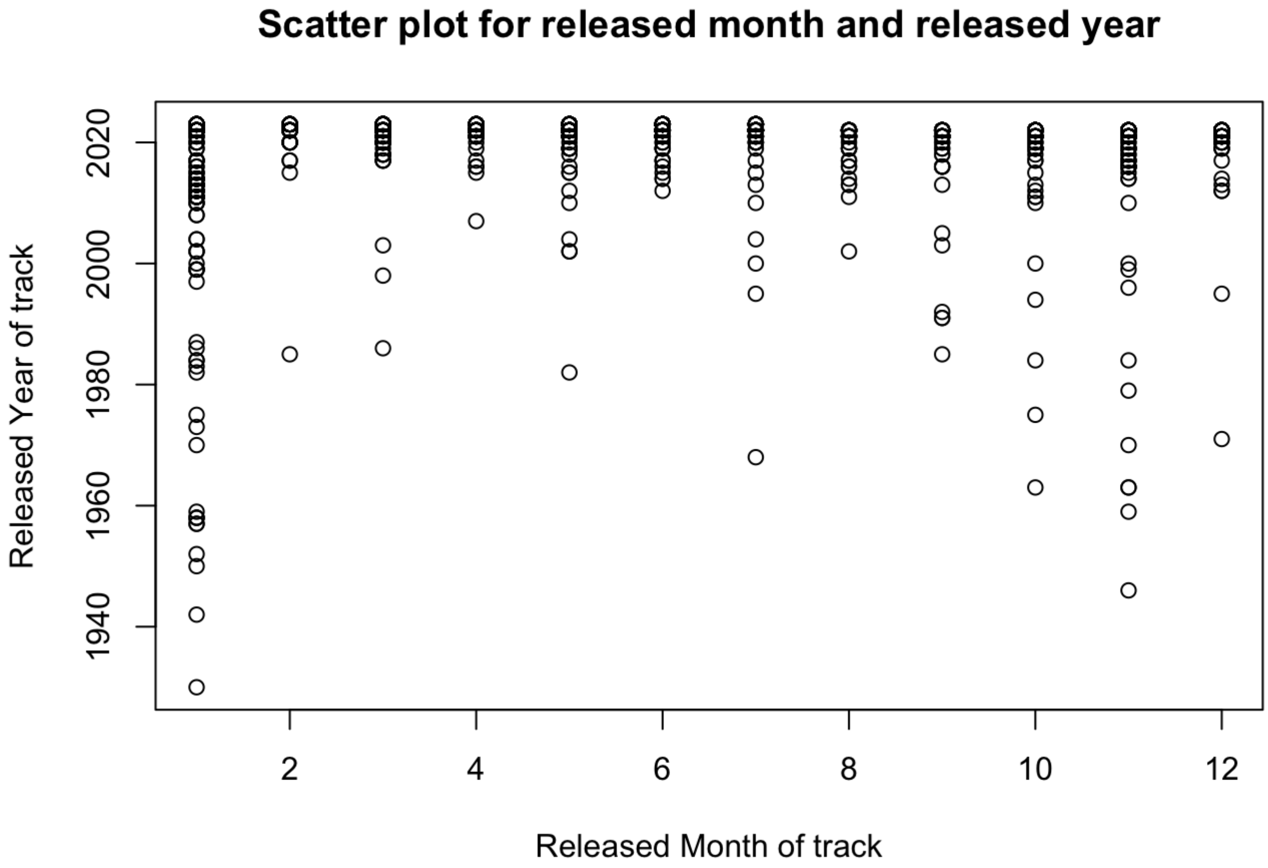
```
#(5)(a) Plotting of one quantitative variable

hist(spotify$artist_count, main = "Histogram of artist count for each track", xlab = "Ar
tist Count")
```



```
#(5)(b)Scatter plot

plot(spotify$released_month, spotify$released_year,main= "Scatter plot for released mont
h and released year", xlab="Released Month of track",ylab = "Released Year of track")
```


Scatter plot for released month and released year
Released Year of track
1940
1960
1980
2000
2020
Released Month of track
2
4
6
8
10
12