CDC-data-analysis

Justin Rivera

2023-09-30

```
library(tidyverse)
```

```
## Warning: package ’tidyverse’ was built under R version 4.1.3

## Warning: package ’tibble’ was built under R version 4.1.3

## Warning: package ’tidyr’ was built under R version 4.1.3

## Warning: package ’readr’ was built under R version 4.1.3

## Warning: package ’purrr’ was built under R version 4.1.3

## Warning: package ’dplyr’ was built under R version 4.1.3

## Warning: package ’stringr’ was built under R version 4.1.3

## Warning: package ’forcats’ was built under R version 4.1.3

## Warning: package ’lubridate’ was built under R version 4.1.3
```

```
library(knitr)
```

```
cdc <- read.csv('CDC-spotify.csv')
```

```
cdc1 = cdc |>
  rename('Available Markets' = Available.Markets, 'Duration (sec)' = Duration..sec., 'Track Name' = Tra
```

```
spotify_data = cdc1 |>
  select(-X)
```

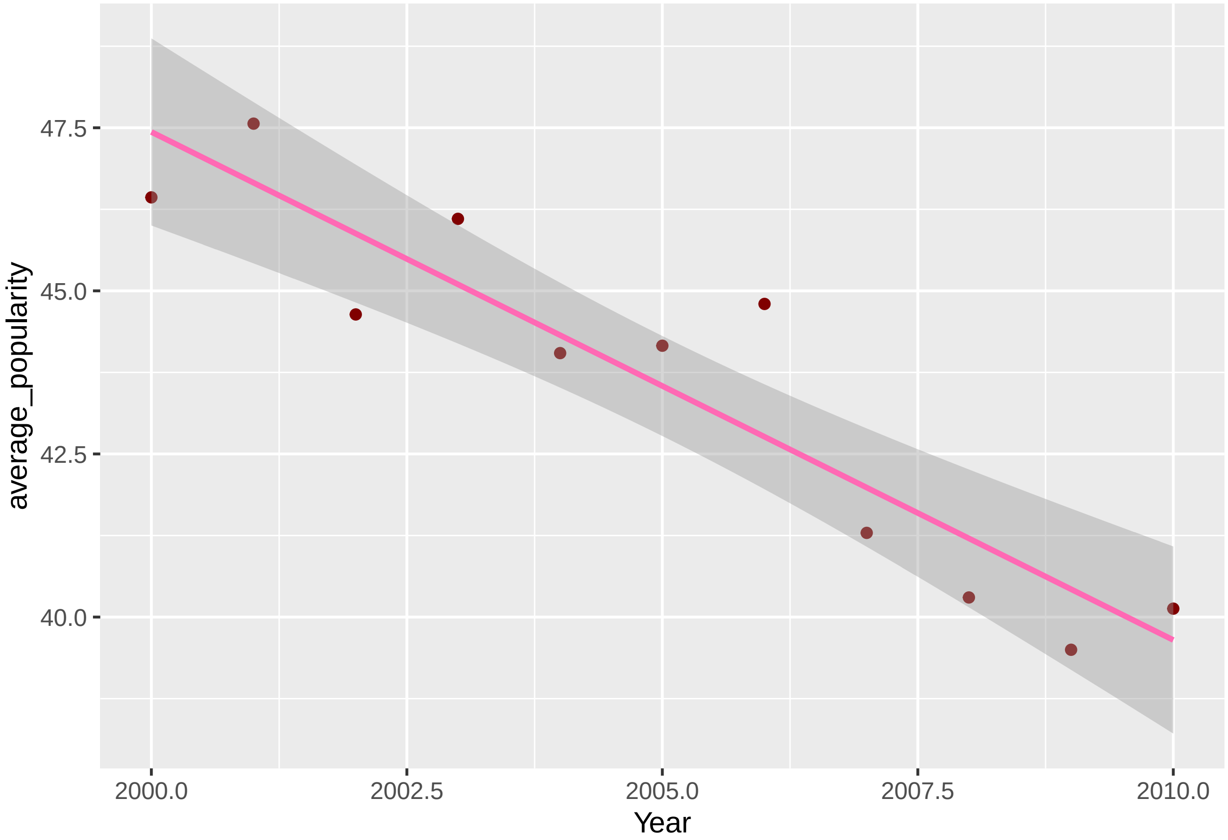
```
average_years = spotify_data |>
  group_by(Year) |>
  summarize(average_popularity = mean(Popularity))

average_years
```

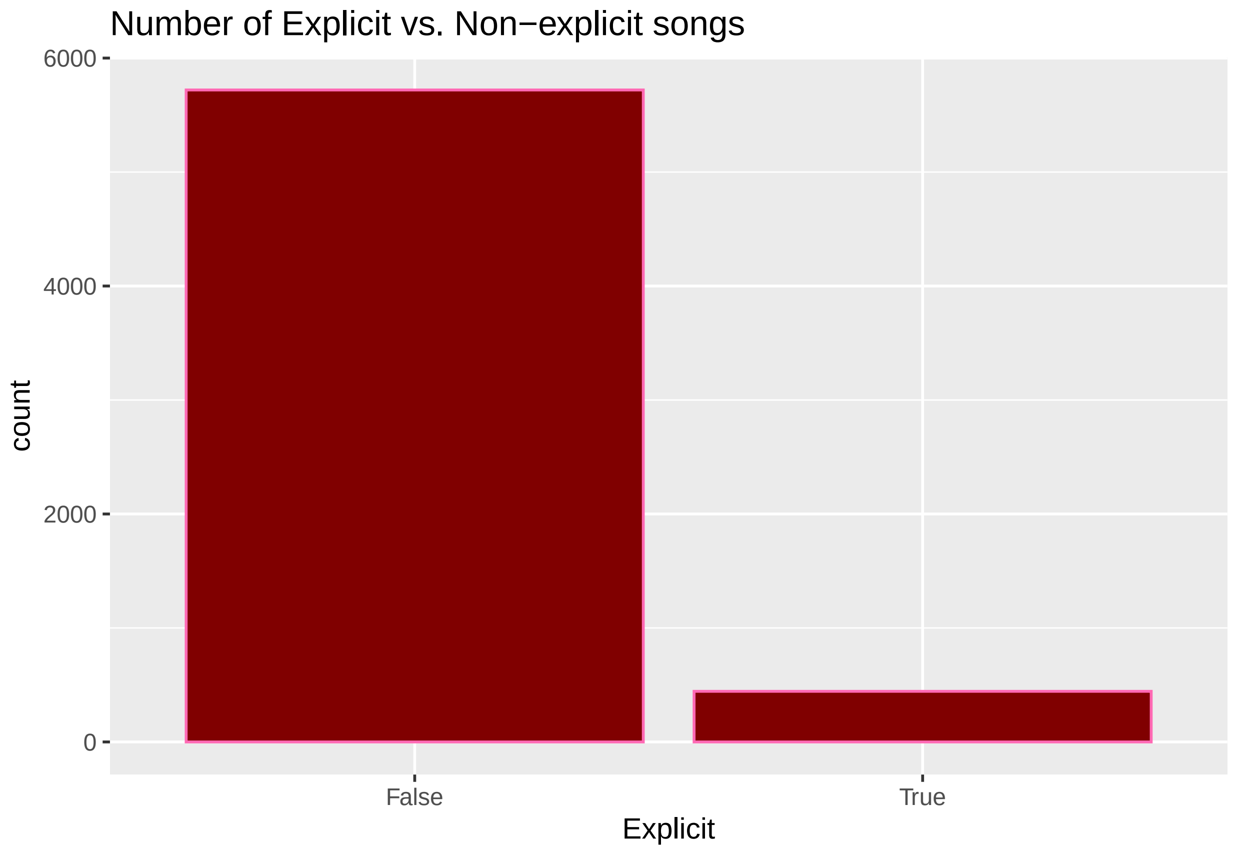
```
## # A tibble: 11 x 2
##     Year average_popularity
##    <int>              <dbl>
##  1  2000               46.4
##  2  2001               47.6
##  3  2002               44.6
##  4  2003               46.1
##  5  2004               44.0
##  6  2005               44.2
##  7  2006               44.8
##  8  2007               41.3
##  9  2008               40.3
## 10  2009               39.5
## 11  2010               40.1
```

```
average_years |>
  ggplot(mapping = aes(x= Year, y = average_popularity)) +
  geom_point(color = "#800000") +
  geom_smooth(method = "lm", color = '#FF69B4', na.rm = TRUE)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
spotify_data |>
  ggplot(mapping = aes(x = Explicit)) +
  geom_bar(color = "#FF69B4", fill = "#800000") +
  labs(title = "Number of Explicit vs. Non-explicit songs")
```

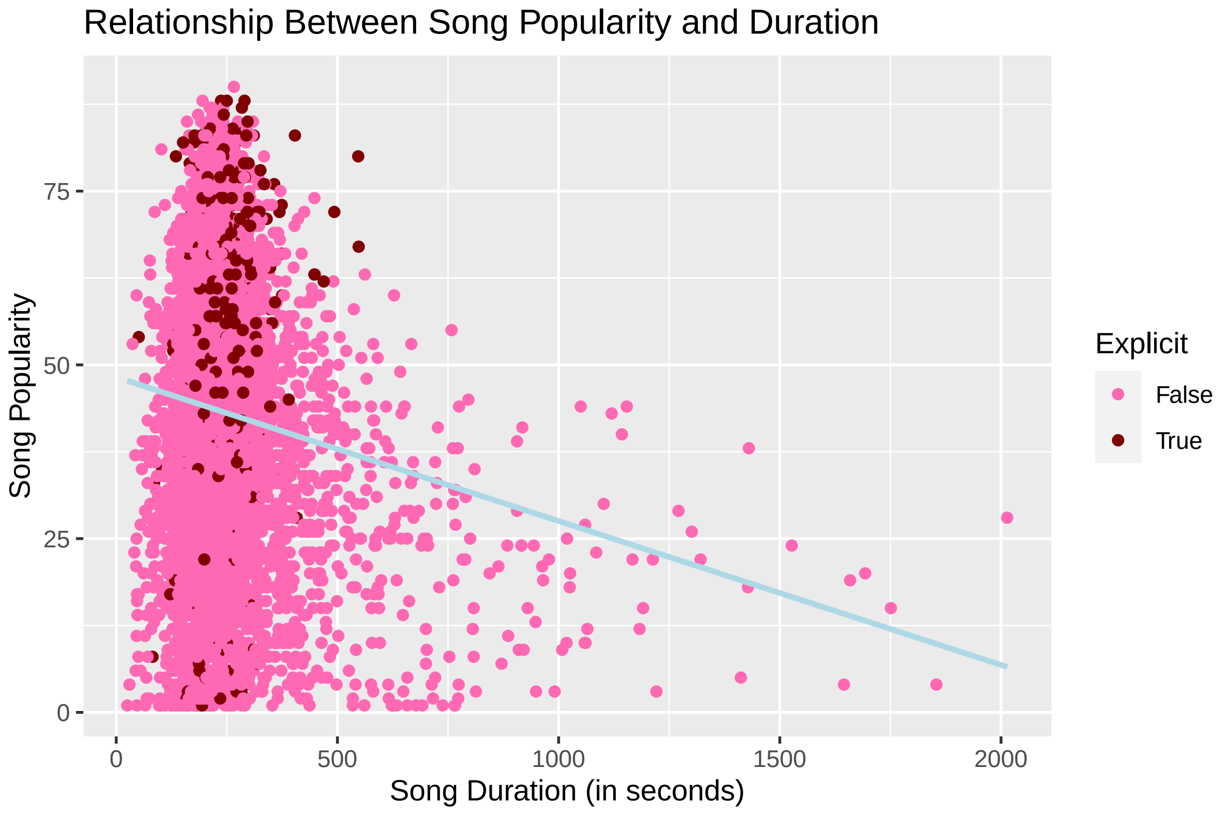


```
average_explicit = spotify_data |>
  group_by(Explicit) |>
  summarize(Explicit_Popularity = mean(Popularity))
kable(average_explicit)
```

Explicit	Explicit_Popularity
False	41.96031
True	56.49550

```
spotify_data |>
  filter(`Duration (sec)` < 3000) |>
  ggplot(mapping = aes(x = `Duration (sec)`, y = Popularity)) +
  geom_point(aes(color = Explicit)) +
   scale_color_manual(values = c("#FF69B4", "#800000")) +
  geom_smooth(method = 'lm', color = "#ADD8E6", se = FALSE) +
    labs(title = "Relationship Between Song Popularity and Duration",
       x="Song Duration (in seconds)",
      y="Song Popularity",
       caption = "Source: Spotify API ")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Source: Spotify API

```
total_mean_duration = mean(spotify_data$`Duration (sec)`)
average_artists = spotify_data |>
  group_by(Artist) |>
  summarize(`Total Popularity` = sum(Popularity),
            Count = n(),
            `Average Duration` = mean(`Duration (sec)`)) |>
  arrange(desc(`Total Popularity`))
```

```
average_artists |>
  ggplot()
```