CDC-data-analysis

Justin Rivera

2023-09-30

```r
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
```

```r
library(knitr)
```

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

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

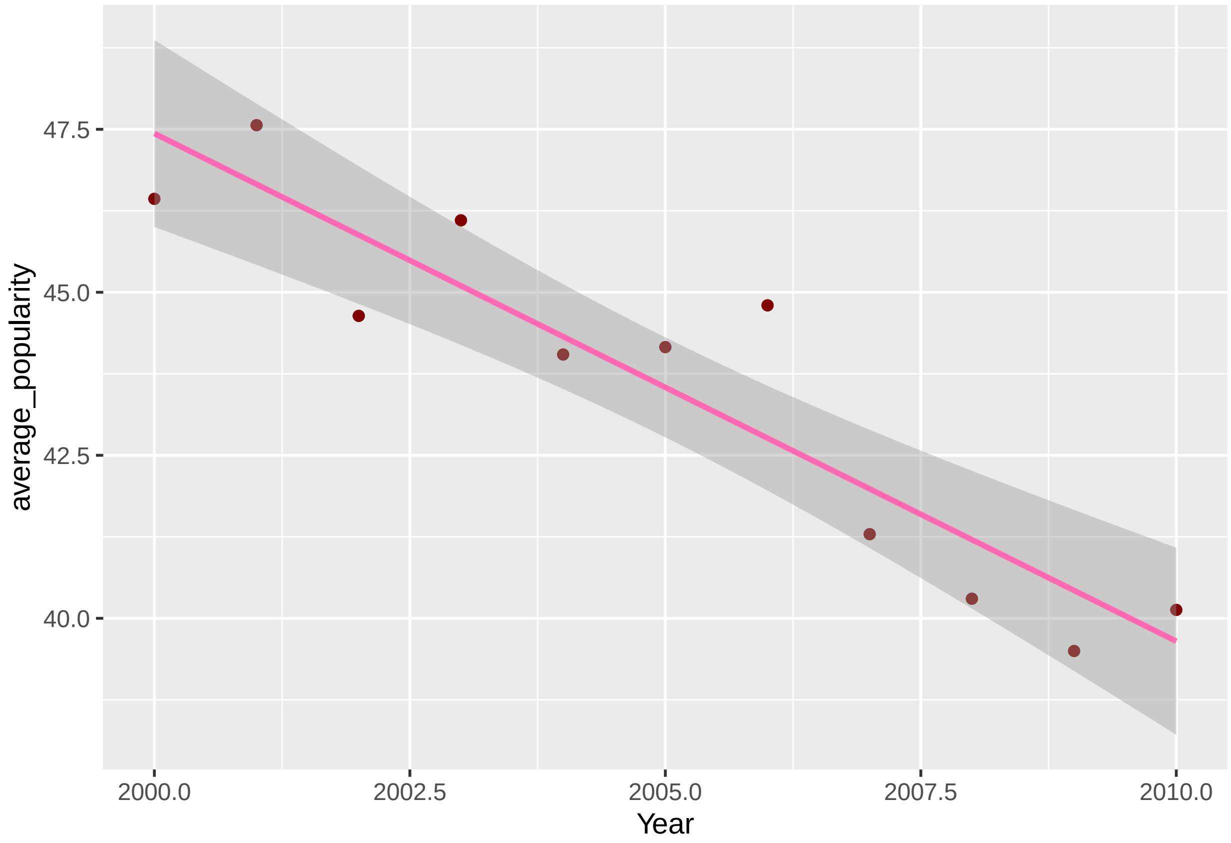
```r
spotify_data = cdc1 |>
  select(-X)
```

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

average_years
```

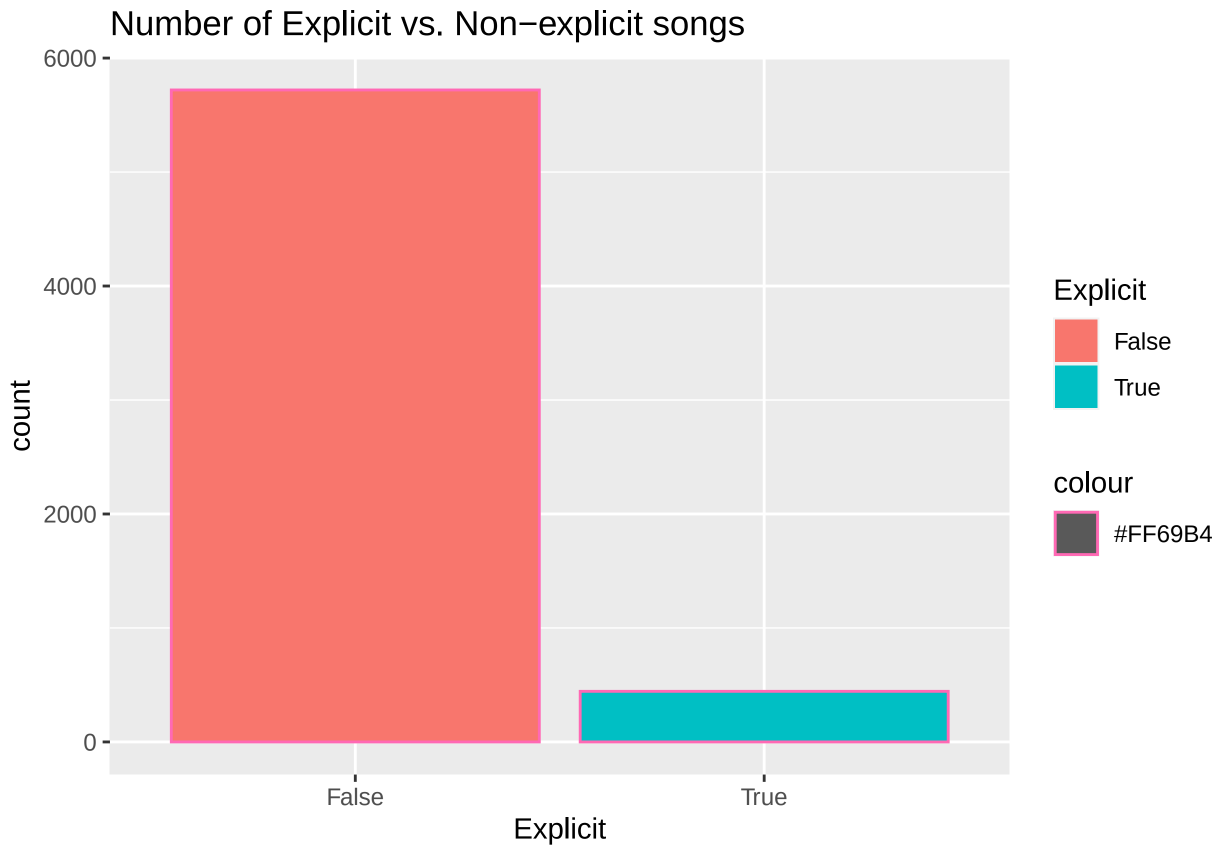
```
## # A tibble: 11 x 2
##     Year average_popularity
##    <int>              <dbl>
##  1  2000               46.4
##  2  2001               47.6
##  3  2002               44.6
##  4  2003               46.1
##  5  2004               44.0
##  6  2005               44.2
##  7  2006               44.8
##  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(aes(color = "#FF69B4", fill = Explicit)) +
  scale_color_manual(values = c("#FF69B4", "#800000")) +
  labs(title = "Number of Explicit vs. Non-explicit songs")
```

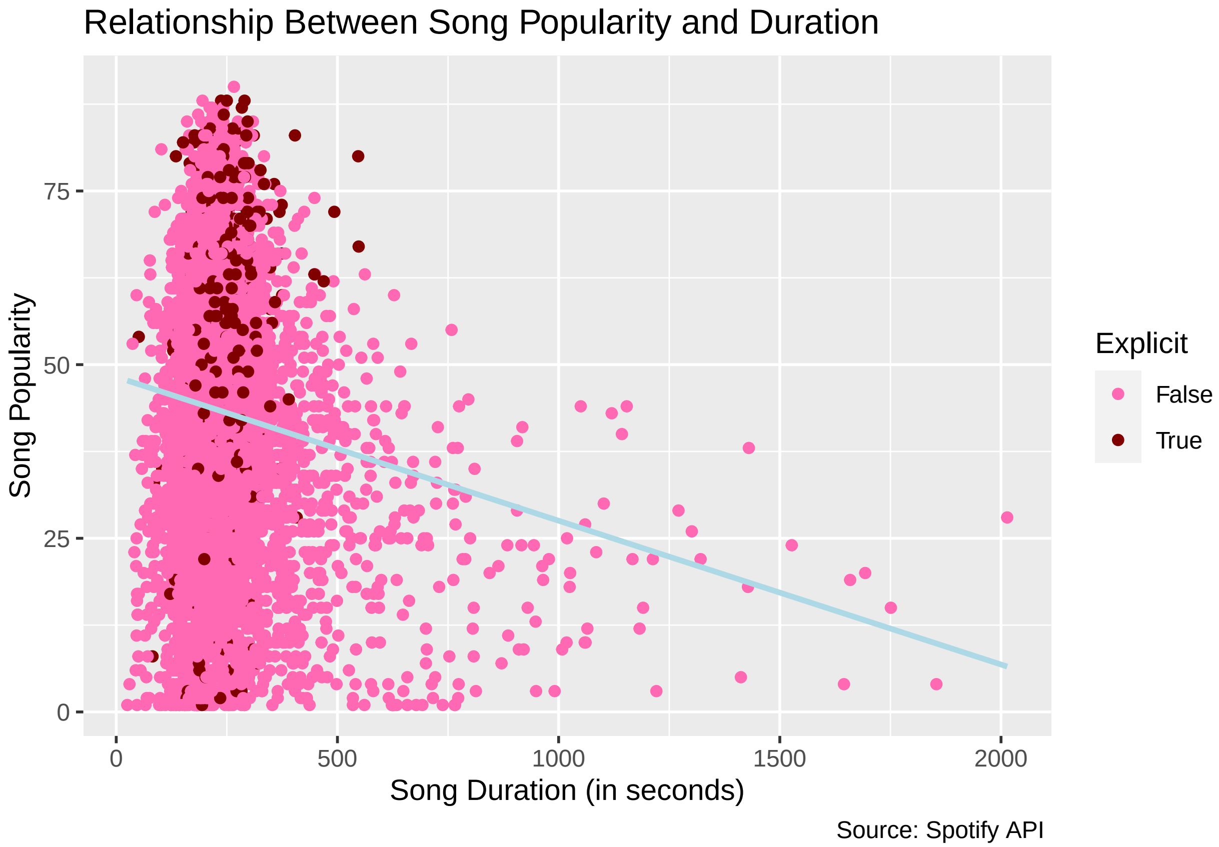


```
average_explicit = spotify_data |>
  group_by(Explicit) |>
  summarize(Explicit_Popularity = mean(Popularity))
kable(average_explicit)
```

Explicit	Explicit_Popularity
False	41.96031
True	56.49550

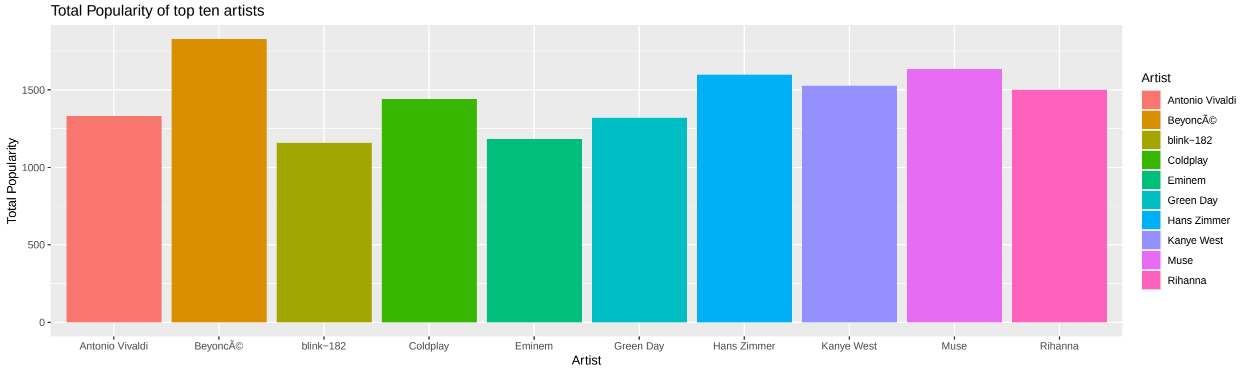
```
spotify_data |>
  filter(`Duration (sec)` < 3000) |>
  ggplot(mapping = aes(x = `Duration (sec)`, y = Popularity)) +
  geom_point(aes(color = Explicit)) +
   scale_color_manual(values = c("#FF69B4", "#800000")) +
  geom_smooth(method = 'lm', color = "#ADD8E6", se = FALSE) +
    labs(title = "Relationship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

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

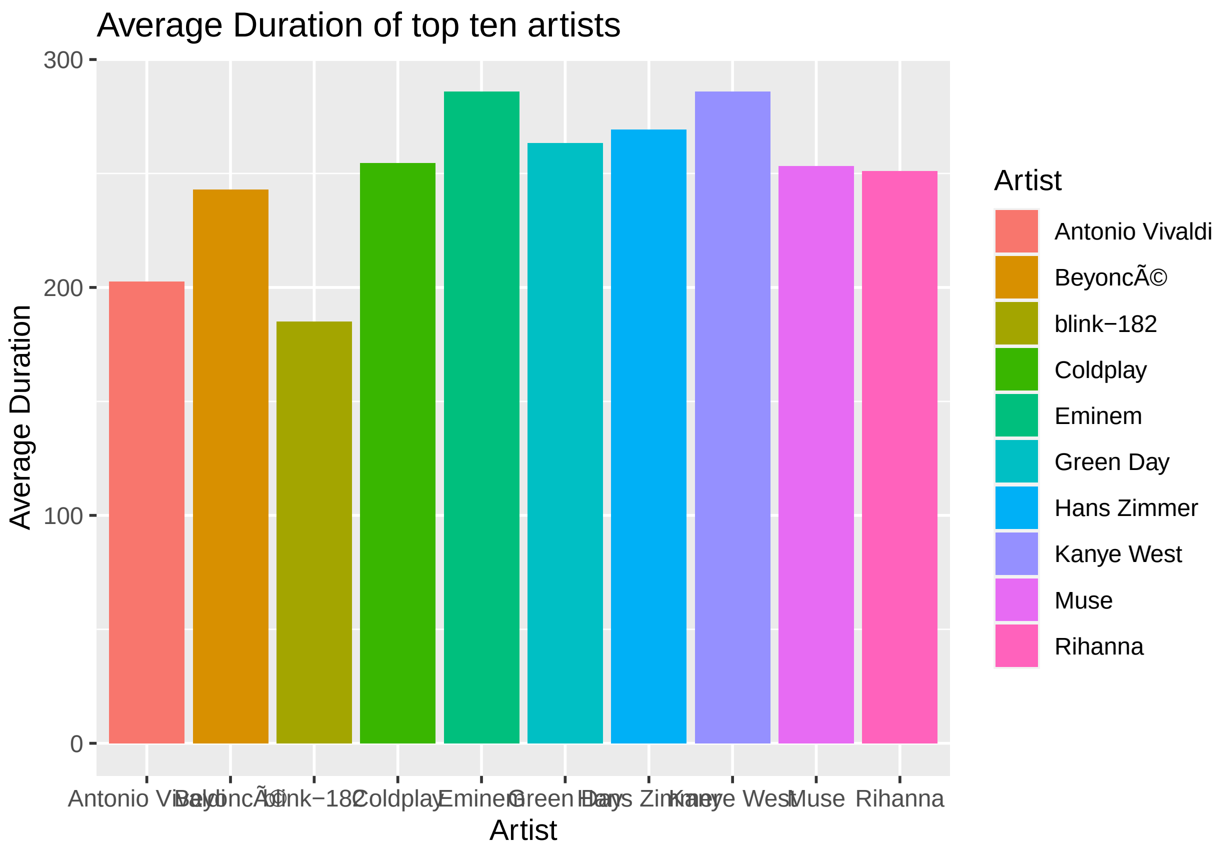


```
total_mean_duration = mean(spotify_data$`Duration (sec)`)
average_artists = spotify_data |>
  group_by(Artist) |>
  summarize(`Total Popularity` = sum(Popularity),
            Count = n(),
            `Average Duration` = mean(`Duration (sec)`) ) |>
  arrange(desc(`Total Popularity`))
```

```
average_artists |>
  filter(`Total Popularity` > 1146) |>
  ggplot(mappping = aes(x = Artist, y = `Total Popularity`)) +
  geom_col(aes(x = Artist, y = `Total Popularity`, fill = Artist)) +
  labs(title = "Total Popularity of top ten artists")
```



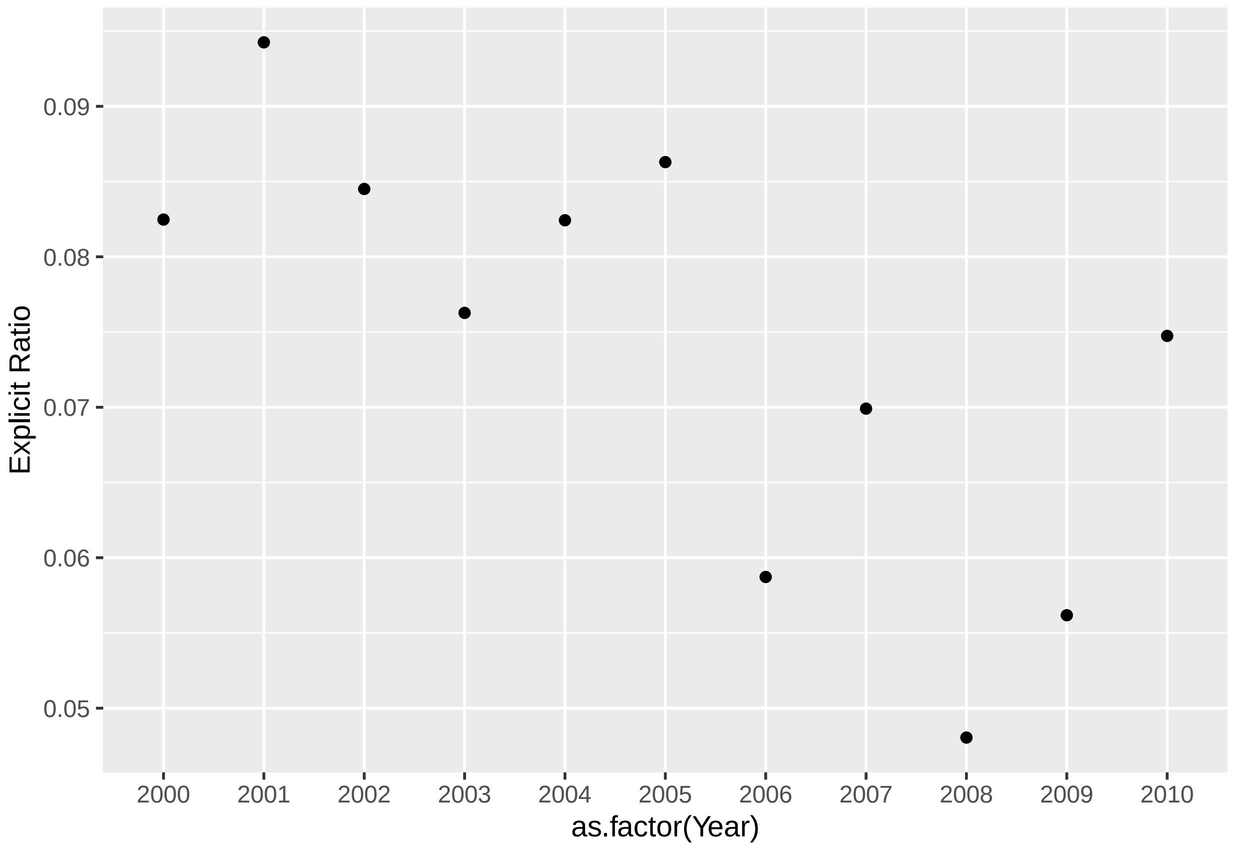
```
average_artists |>
  filter(`Total Popularity` > 1146) |>
  ggplot(mappping = aes(x = Artist, y = `Average Duration`)) +
  geom_col(aes(x = Artist, y = `Average Duration`, fill = Artist)) +
  labs(title = "Average Duration of top ten artists")
```



```
Explicit_ratio = spotify_data |>
  group_by(Year) |>
  summarize(`Explicit Ratio` = sum(Explicit == "True")/n())

kable(Explicit_ratio)
```

Year	Explicit Ratio
2000	0.0824742
2001	0.0942529
2002	0.0845070
2003	0.0762712
2004	0.0824295
2005	0.0862944
2006	0.0587219
2007	0.0699088
2008	0.0480480
2009	0.0561798
2010	0.0747423

```
Explicit_ratio |>
  ggplot(mapping = aes(x = as.factor(Year), y = `Explicit Ratio`)) +
  geom_point()
```

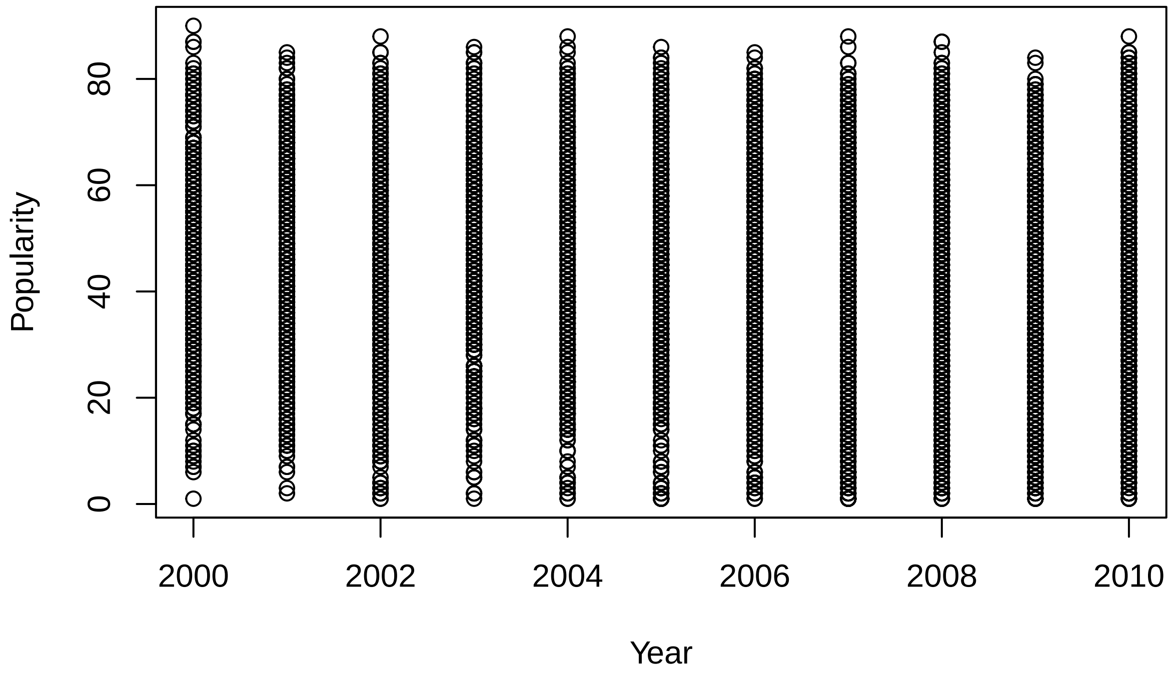


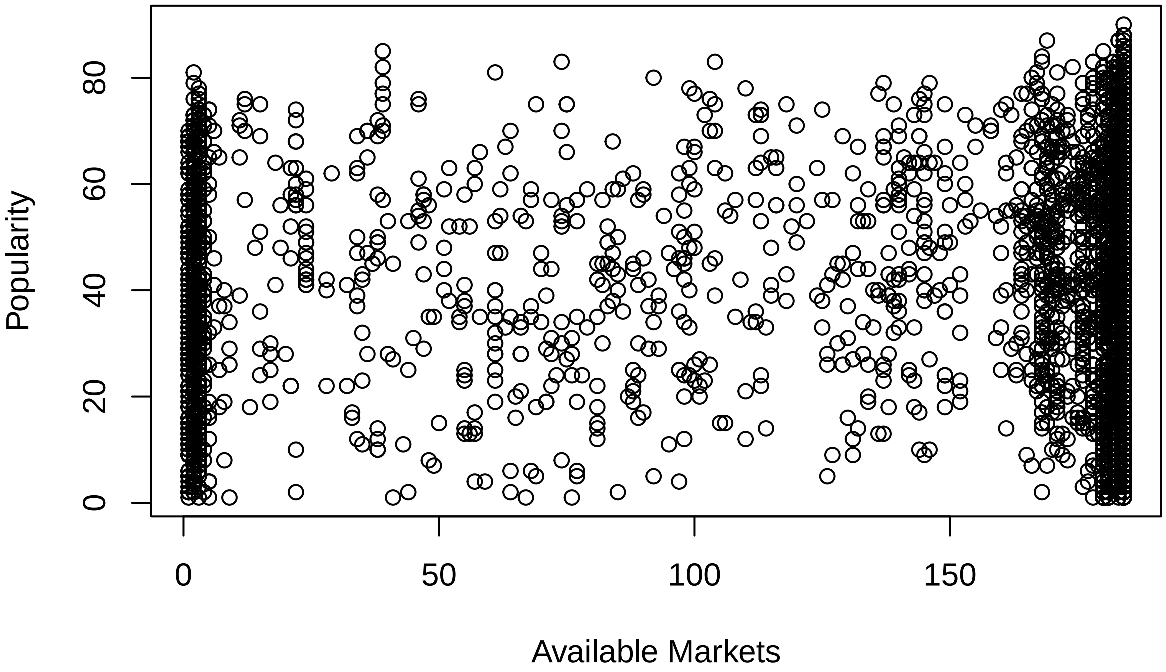
```
popularity_model = lm(Popularity~Year + `Available Markets` + `Duration (sec)` + as.factor(Explicit), da

summary(popularity_model)
```

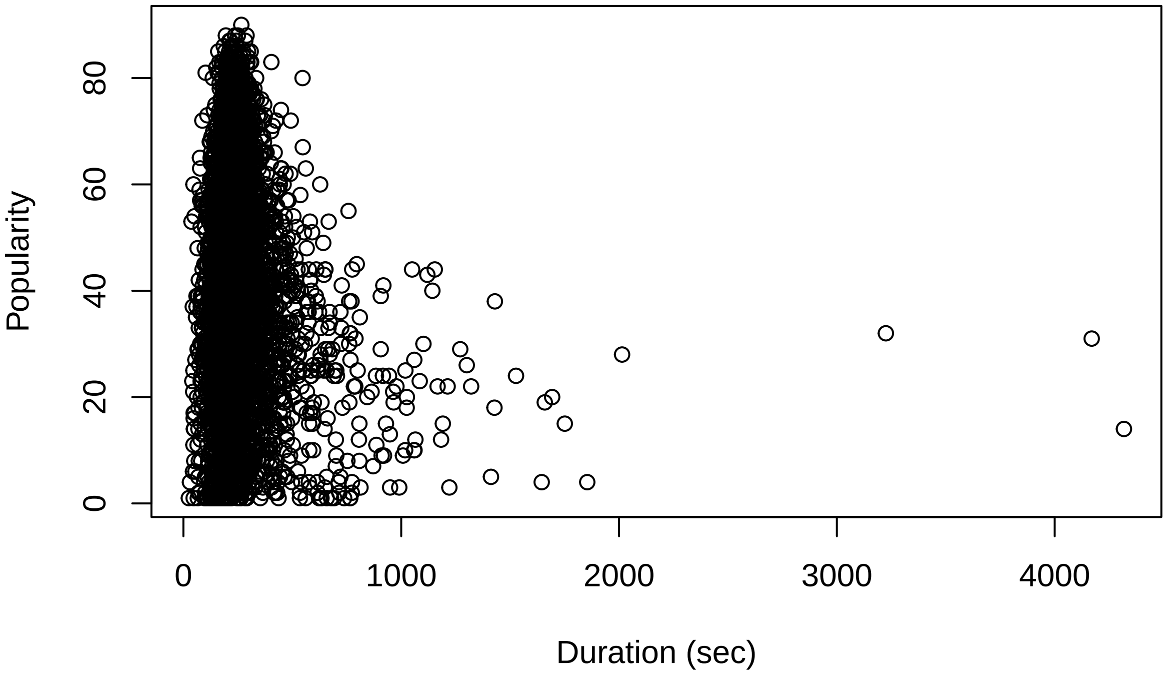
```
##
## Call:
```

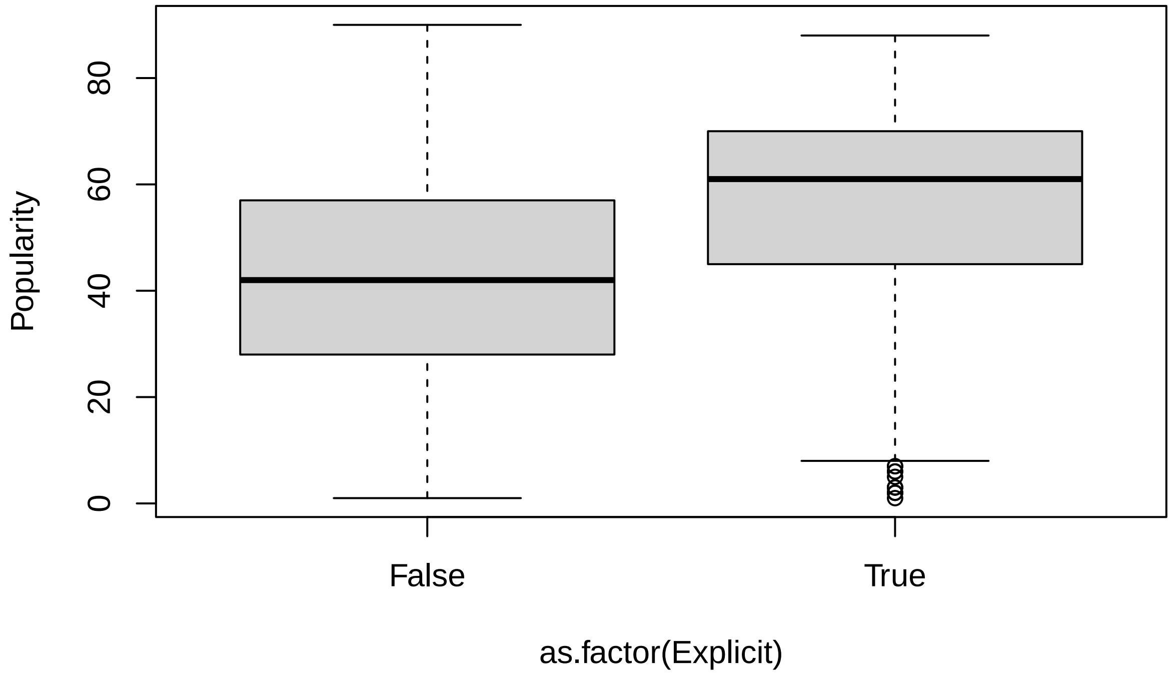
```
## lm(formula = Popularity ~ Year + ‘Available Markets‘ + ‘Duration (sec)‘ +
##     as.factor(Explicit), data = spotify_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.374 -13.670  -0.098  14.528  53.510
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.526e+03  1.574e+02   9.697  < 2e-16 ***
## Year                     -7.389e-01  7.844e-02  -9.421  < 2e-16 ***
## ‘Available Markets‘       1.412e-02  4.030e-03   3.504 0.000461 ***
## ‘Duration (sec)‘         -1.595e-02  1.613e-03  -9.889  < 2e-16 ***
## as.factor(Explicit)True   1.420e+01  9.441e-01  15.041  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 19.12 on 6159 degrees of freedom
## Multiple R-squared:  0.06659,    Adjusted R-squared:  0.06598
## F-statistic: 109.8 on 4 and 6159 DF,  p-value: < 2.2e-16
```

```
plot(Popularity~Year + `Available Markets` + `Duration (sec)` + as.factor(Explicit), data = spotify_dat
```




Popularity
0
20
40
60
80
0
50
100
150
Available Markets


Popularity
0
20
40
60
80
0
1000
2000
3000
4000
Duration (sec)



```
summary(popularity_model)
```

```
##
## Call:
## lm(formula = Popularity ~ Year + `Available Markets` + `Duration (sec)` +
##     as.factor(Explicit), data = spotify_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.374 -13.670  -0.098  14.528  53.510
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.526e+03  1.574e+02   9.697  < 2e-16 ***
## Year                   -7.389e-01  7.844e-02  -9.421  < 2e-16 ***
## `Available Markets`     1.412e-02  4.030e-03   3.504 0.000461 ***
## `Duration (sec)`       -1.595e-02  1.613e-03  -9.889  < 2e-16 ***
## as.factor(Explicit)True  1.420e+01  9.441e-01  15.041  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 6159 degrees of freedom
## Multiple R-squared:  0.06659,    Adjusted R-squared:  0.06598
## F-statistic: 109.8 on 4 and 6159 DF,  p-value: < 2.2e-16
```