plots

jabir

2023-09-22

```r
library(spotifyr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lubridate)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(usethis)
library(ggthemes)
```

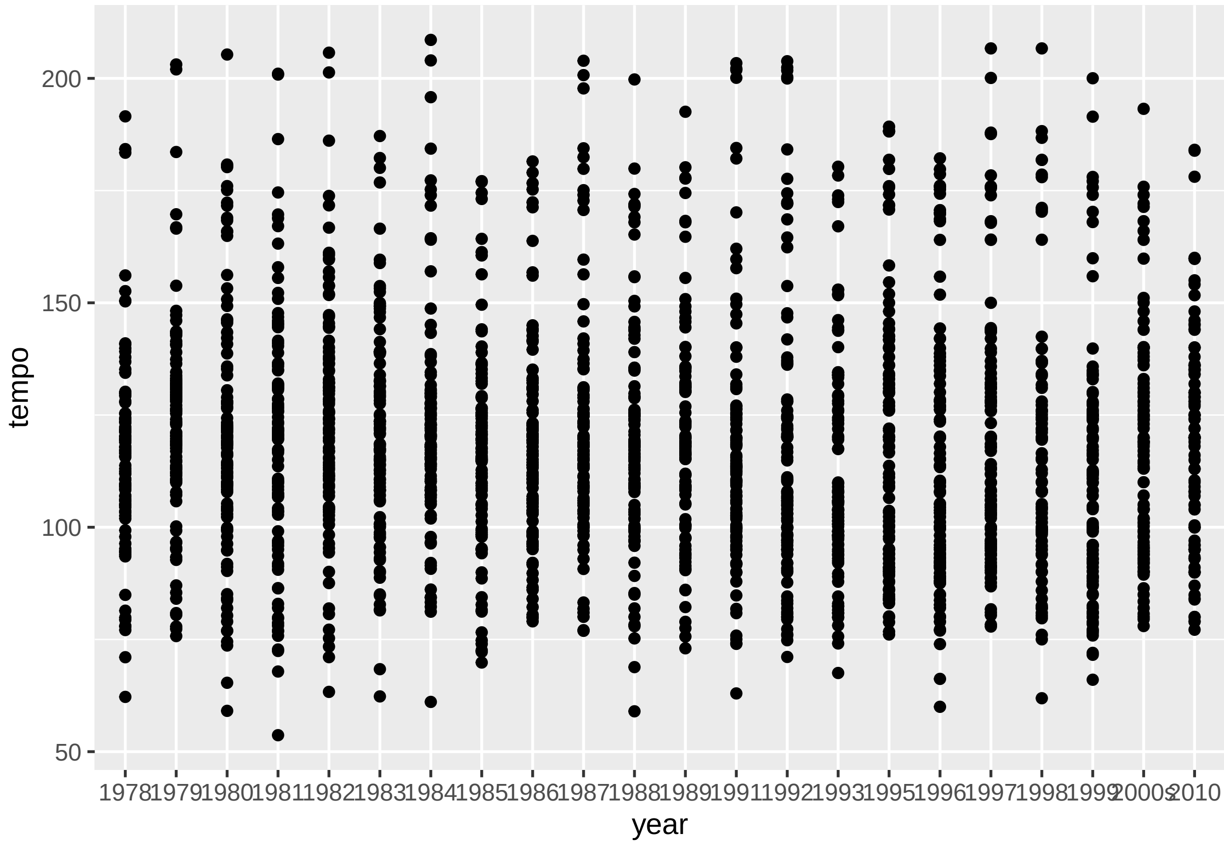
Import data that was set up in the notebook `set_up_data.Rmd`.

```r
trash_songs_df <- read_rds(file.path("data","trash_songs_df.rds"))

good_songs_df <- read_rds(file.path("data","good_songs_df.rds"))

all_songs_df <- readRDS(file.path("data", "all_songs.RDS"))

all_songs_df <- bind_rows(list(unpopular = trash_songs_df, popular = good_songs_df),
                          .id= "song_group")

trash_songs <- read_rds(file.path("data","trash_songs.RDS"))
```

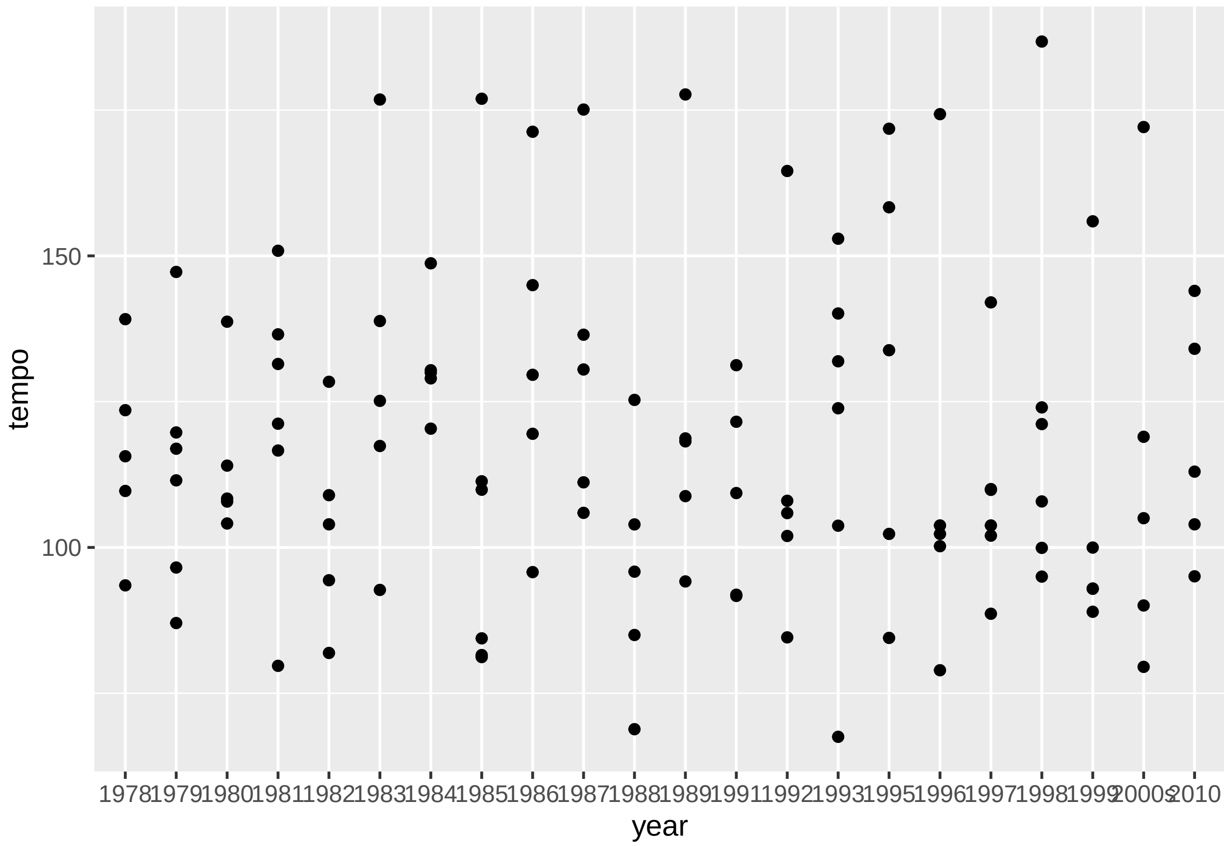
I grabbed the top five songs each year and compared the tempo of all the songs vs the top five and this showed that the tempo was more or less the same throughout each decade.

```r
top_five_all <- good_songs_df |>
  arrange(desc(track.popularity)) |>
  arrange(year) |>
  mutate(year = as.factor(year)) |>
  group_by(year) |>
  top_n(5,track.popularity)
```

```r
good_songs_df |>
  ggplot(aes(x = year, y= tempo)) +
  geom_point()
```




```r
top_five_all |>
  ggplot(aes(x = year, y = tempo)) +
  geom_point()
```

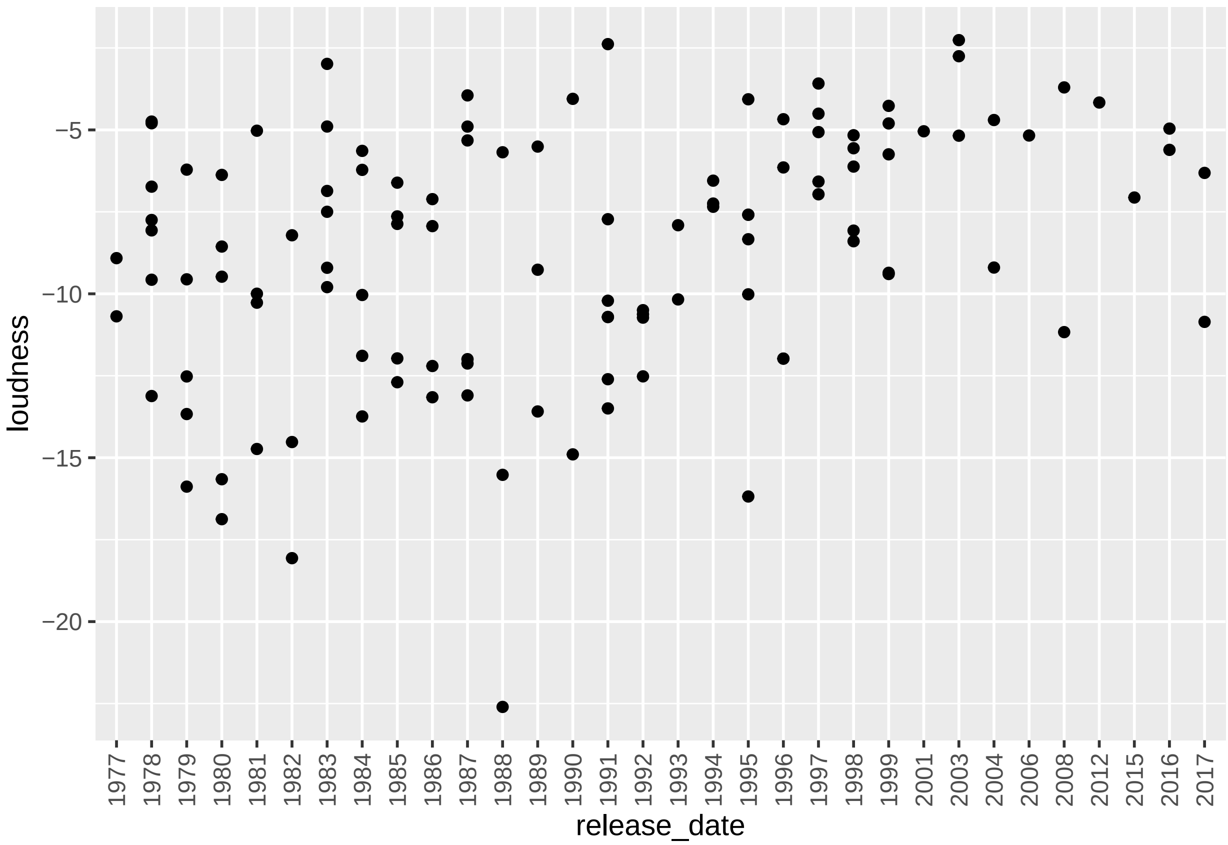


The energy of the songs start to rise half way through the 90s and 2000s so it does look like energy matters.

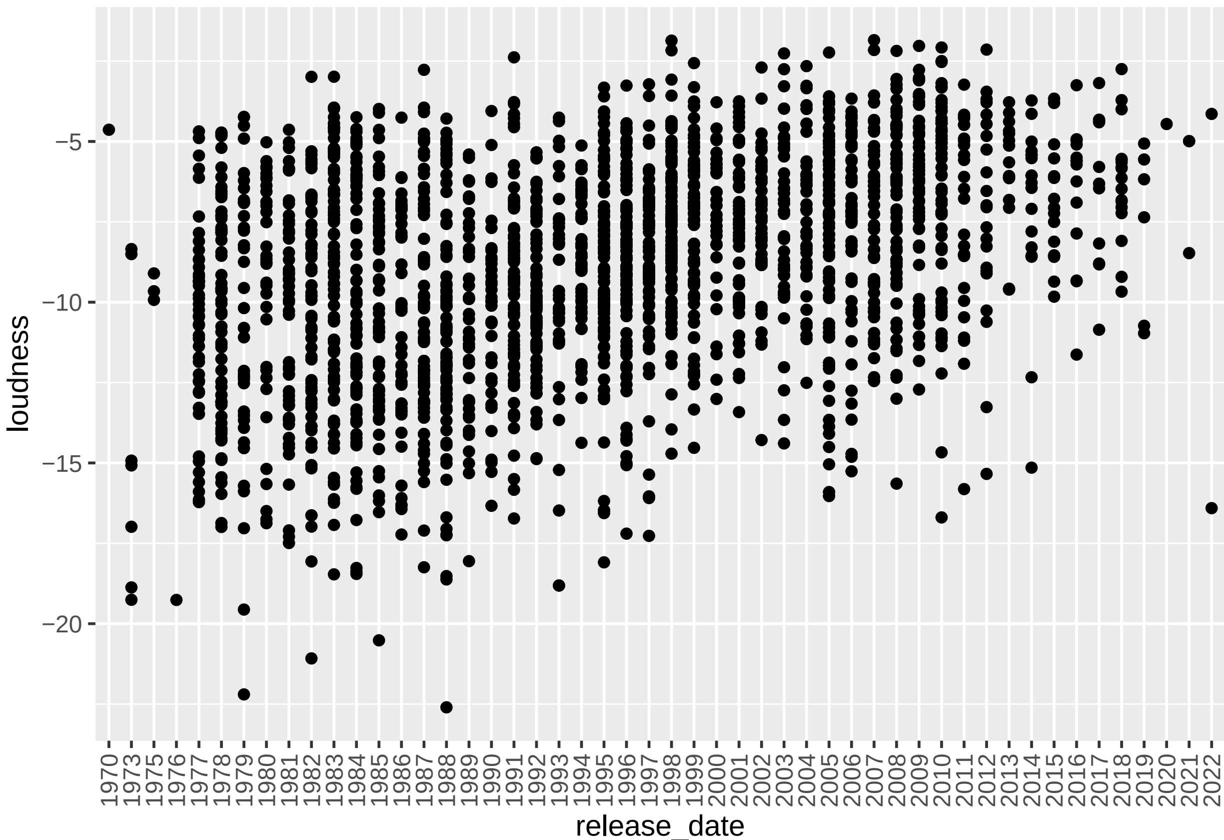
```
top_five_all |>
  ggplot(aes(x = release_date, y = energy)) +
  geom_point() +
   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

The loudness seems to start to increase through out the 90s on into the 2000s.

```
top_five_all |>
  ggplot(aes(x = release_date, y = loudness)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

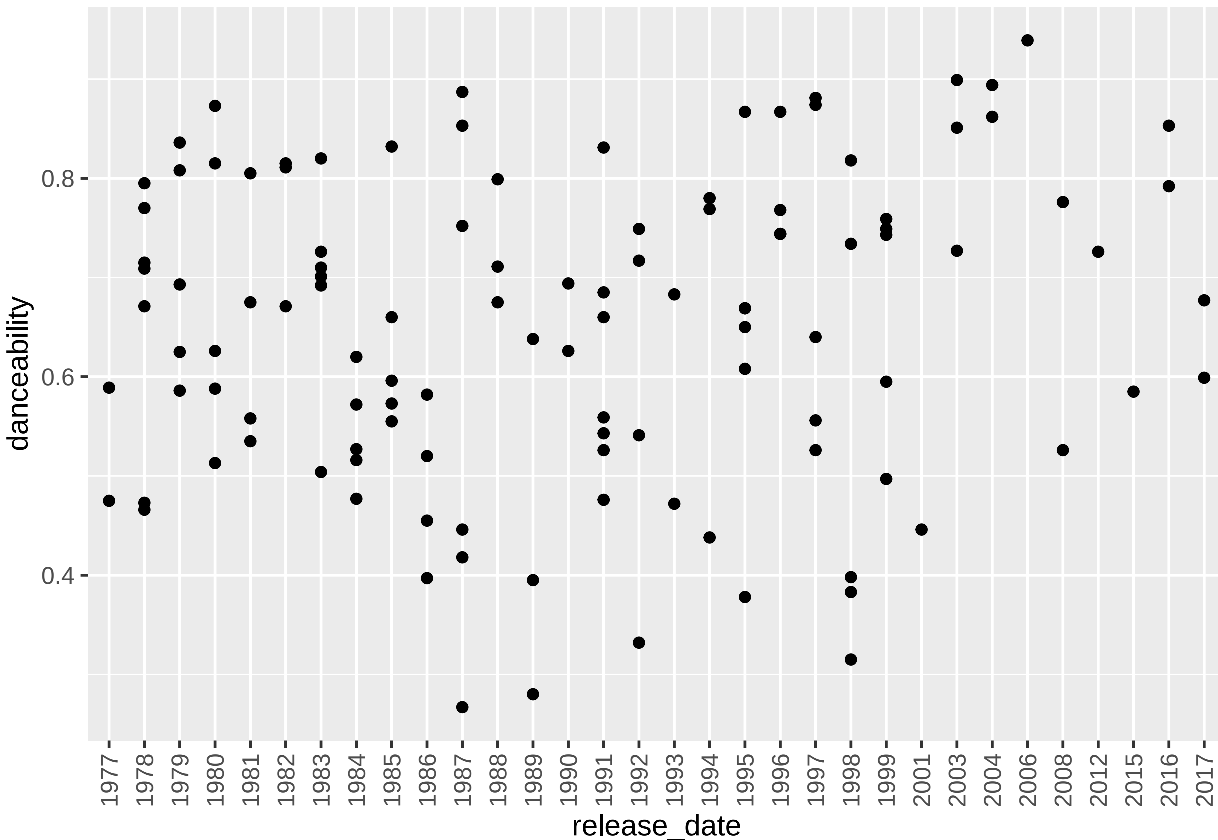



```
good_songs_df |>
  ggplot(aes(x = release_date, y = loudness)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

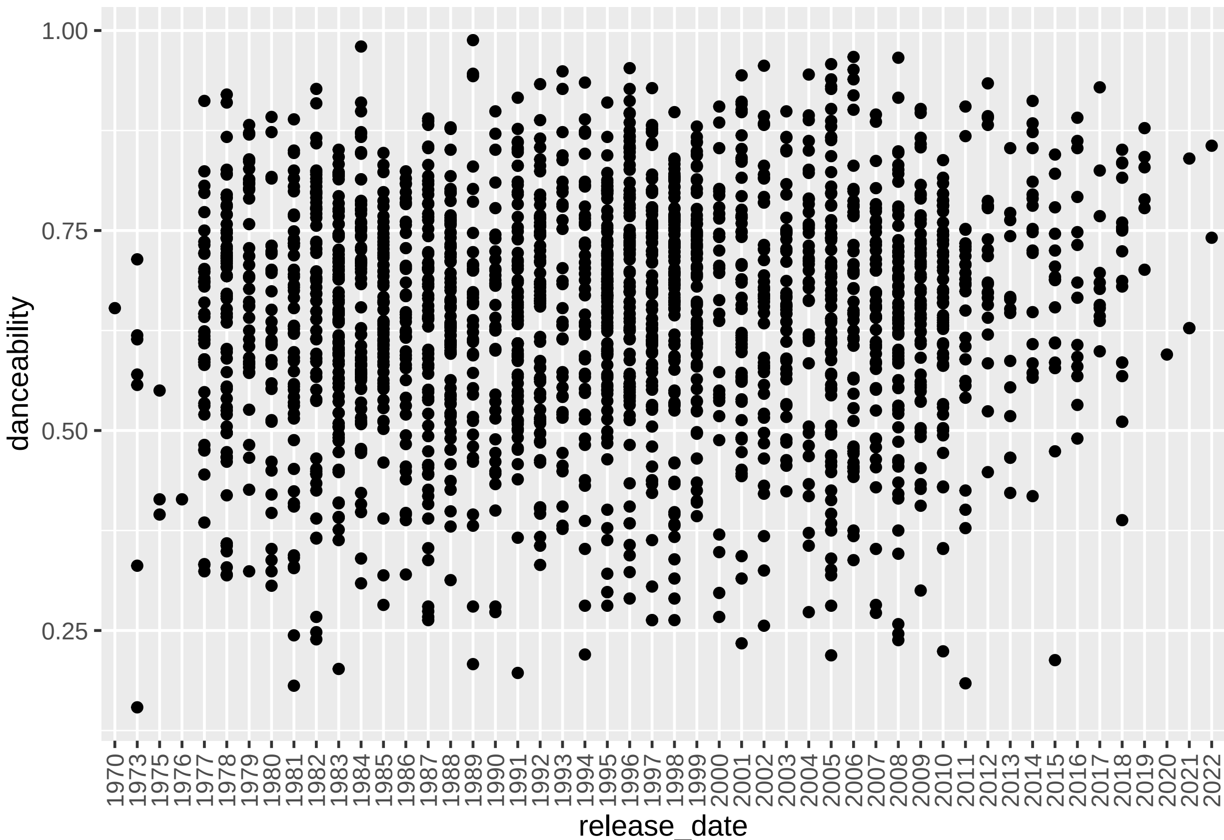


This also seems to increase even though any song can be danced to, also take note that there are less songs as the years increase.

```
top_five_all |>
  ggplot(aes(x = release_date, y = danceability)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



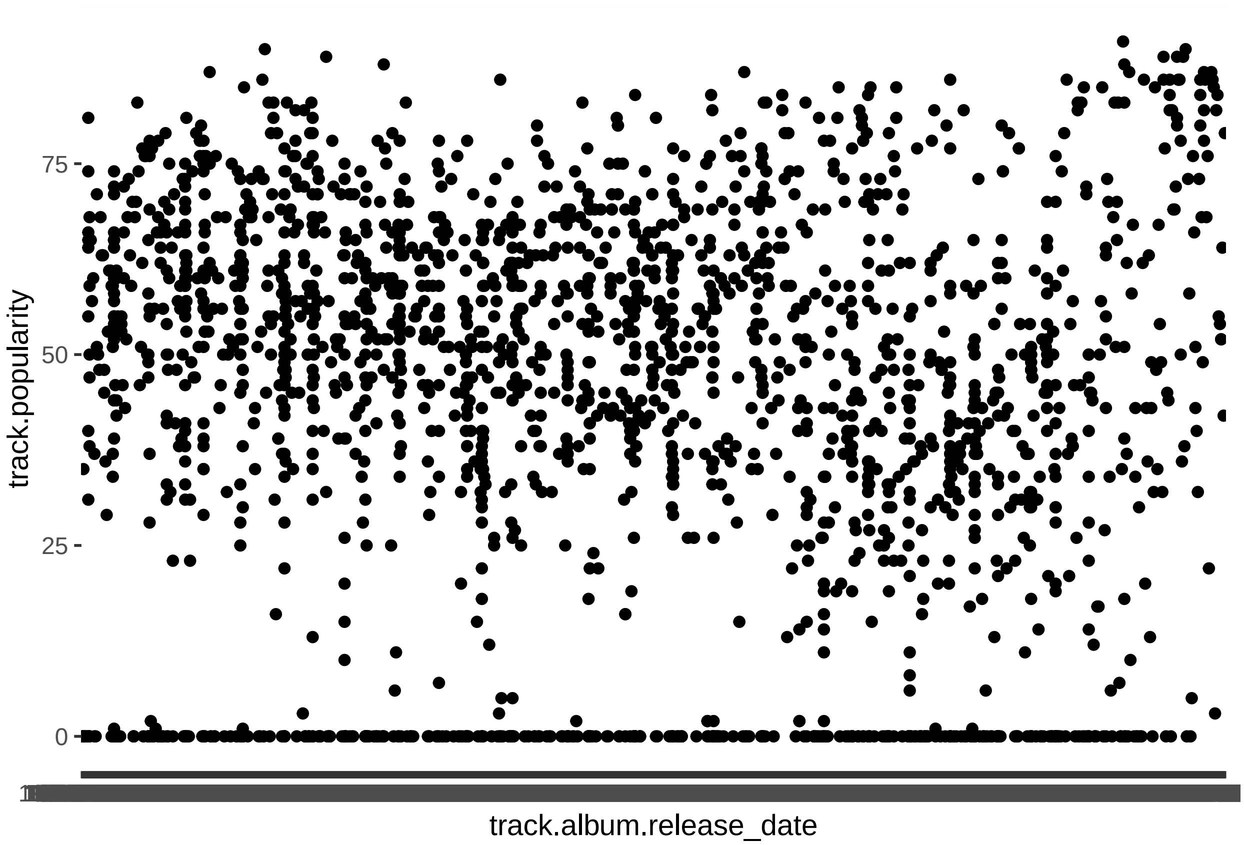

```r
good_songs_df |>
  ggplot(aes(x = release_date, y = danceability)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



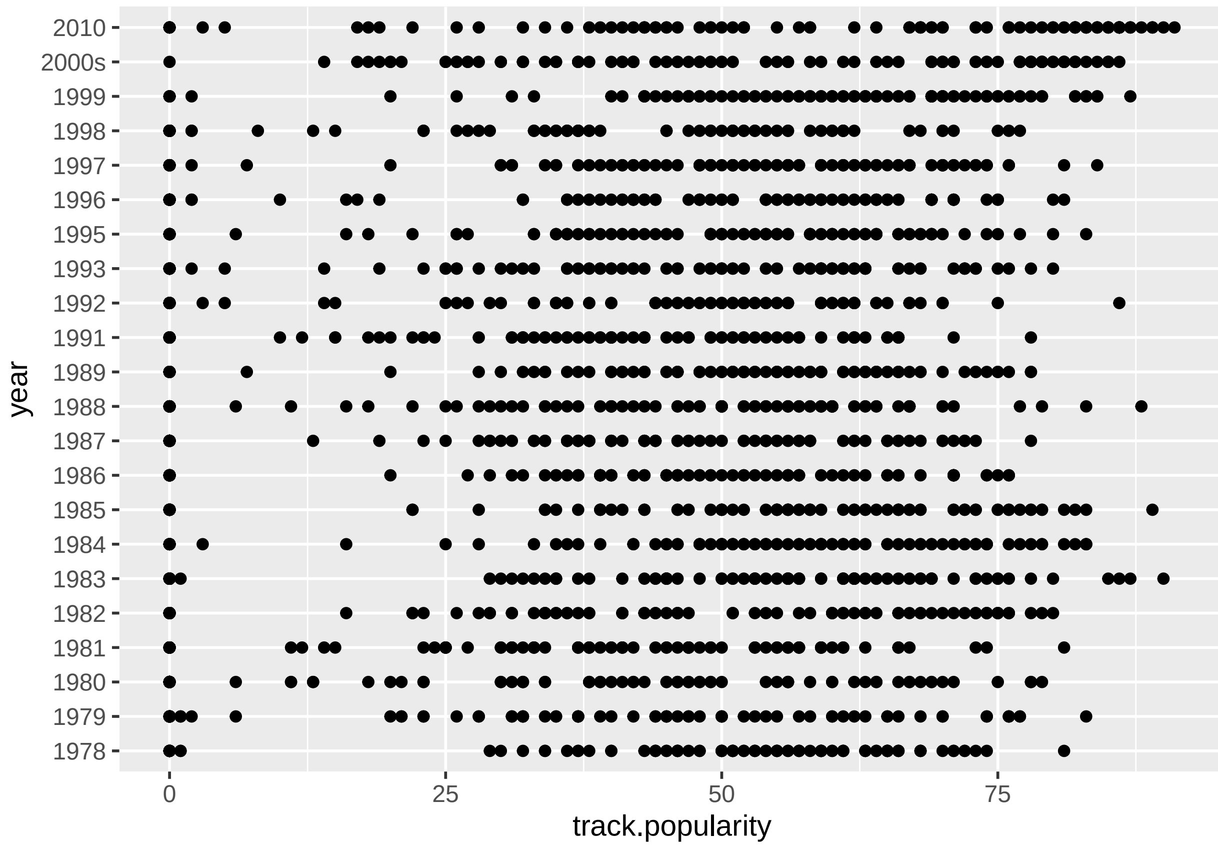
Release year and audio info

The following plots examine the question "Have audio qualities of popular songs changed over time?". To examine this we've plotted the release date against various audio qualities of the most popular songs.

```
# Start making some plots with ggplot2 and explore possible relationships between the year of the song
# Also maybe look at popularity and audio information

ggplot2::ggplot(good_songs_df, ggplot2::aes(track.album.release_date, track.popularity)) +
  ggplot2::geom_point()
```



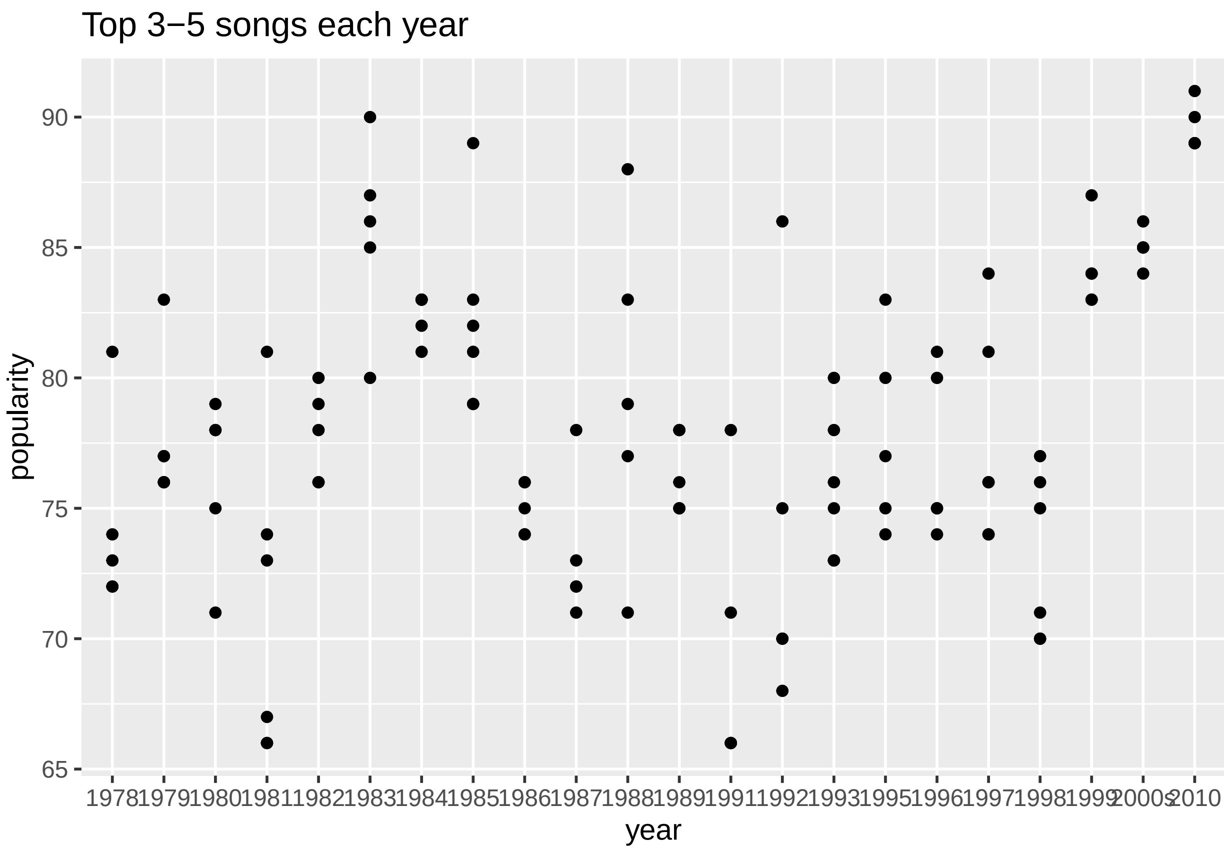

```r
good_songs_df |>
  ggplot(aes(x = track.popularity, y = year)) +
  geom_point()
```




```
top_songs_df <- good_songs_df |>
  select(track.name,track.popularity,year) |>
  arrange(desc(track.popularity)) |>
  arrange(year) |>
  mutate(year = as.factor(year)) |>
  group_by(year) |>
  top_n(5,track.popularity)

 most_pop_song_each_year <- good_songs_df |>
  select(track.name,track.popularity,year) |>
  arrange(desc(track.popularity)) |>
  arrange(year) |>
  mutate(year = as.factor(year)) |>
  group_by(year) |>
  top_n(1,track.popularity)
```

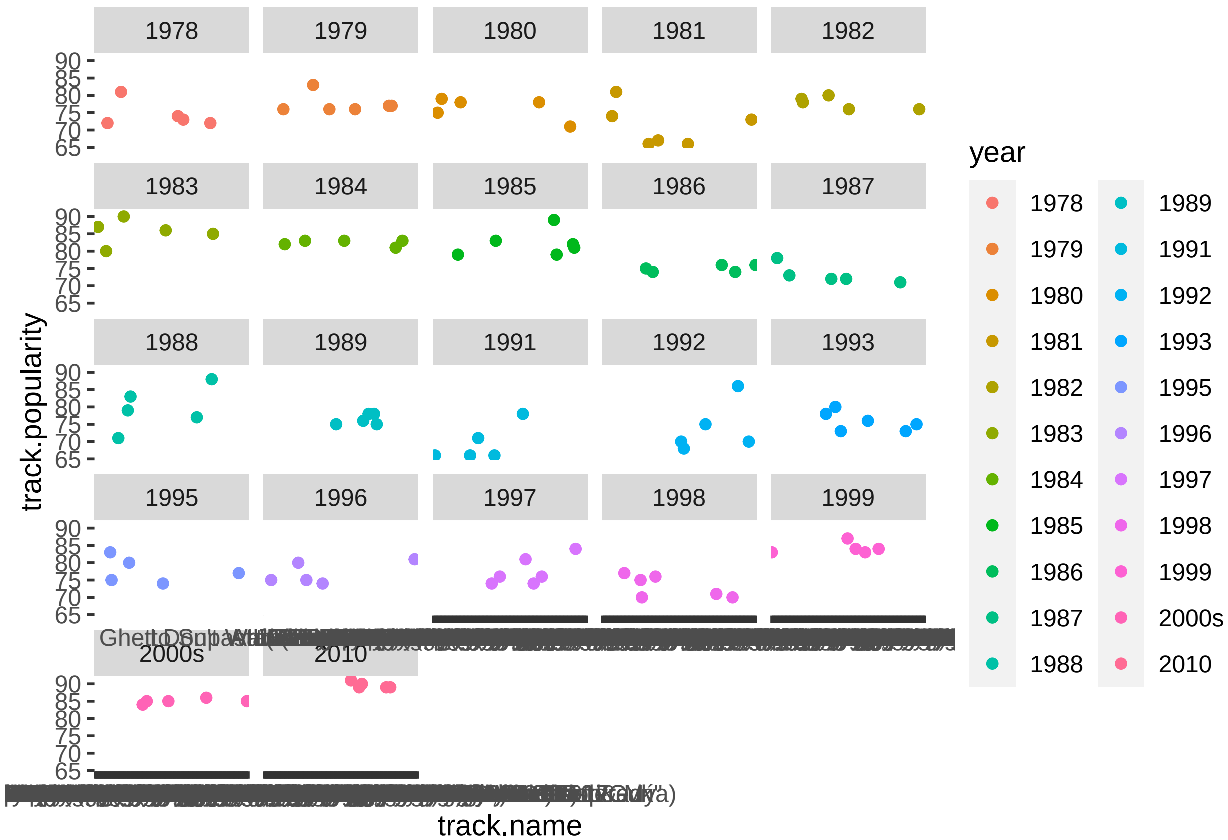
This shows a side by side comparison of the top 3-5 popularity each year, this also shows that songs in the late 90s to the 2010s were very popular.

```
top_songs_df |>
  ggplot(aes(x = year, y = track.popularity)) +
  geom_point() +
  labs(
    title = "Top 3-5 songs each year",
    y = "popularity"
  )
```



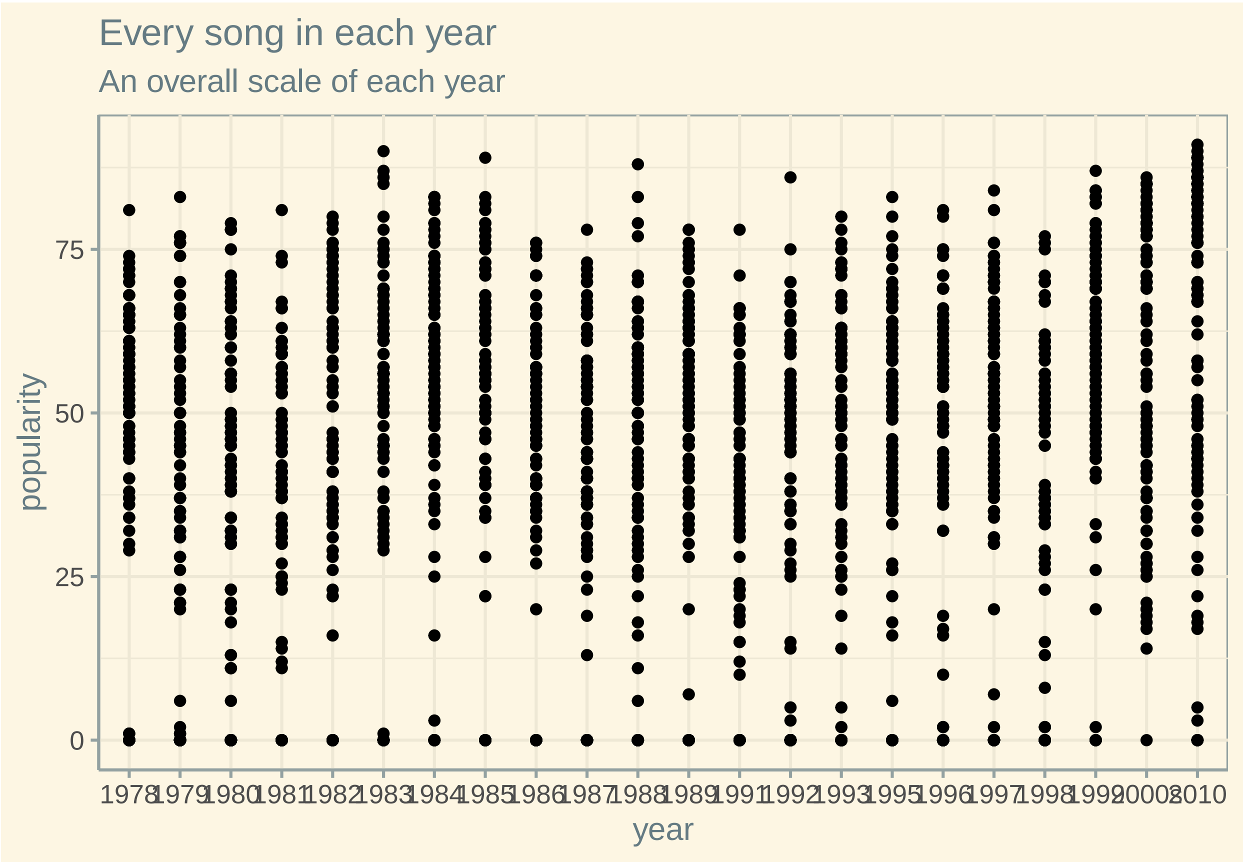
I faceted each year to get a better look at each song group by comparing their popularities side by side, for a better view.

```
top_songs_df |>
  ggplot(aes(x = track.name, y = track.popularity,color = year)) +
  geom_point() +
  facet_wrap(vars(year))
```



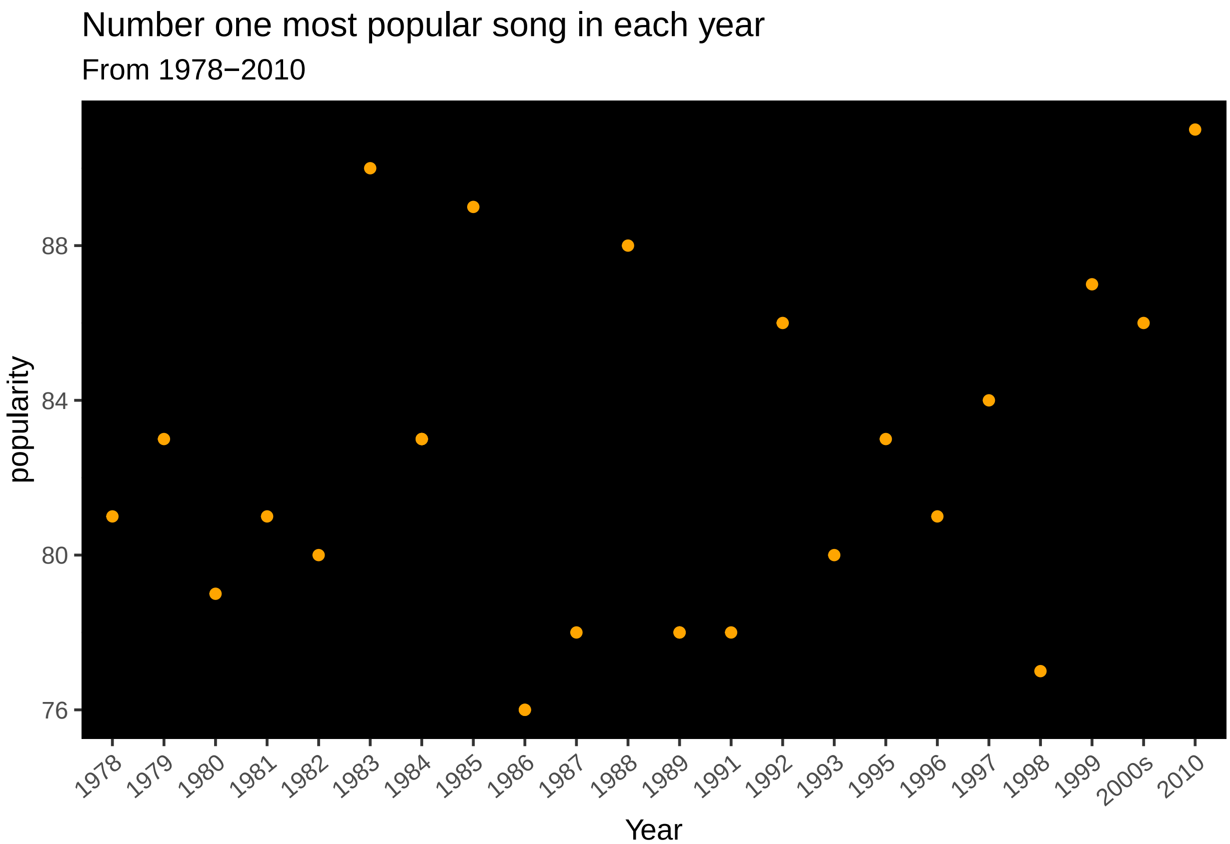
I wanted to see if there were any noticeable differences between every song in each year turns out there wasn't they are pretty consistent.

```r
good_songs_df |>
  ggplot(aes(x = year, y = track.popularity)) +
  geom_point() +
  labs(title = "Every song in each year",
       subtitle = "An overall scale of each year",
       x = "year",
       y = "popularity") +
   theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  ggthemes::theme_solarized()
```



This plot shows the number one song in each year, this plot also shows that as the years go on song popularity rises through the 2000s.

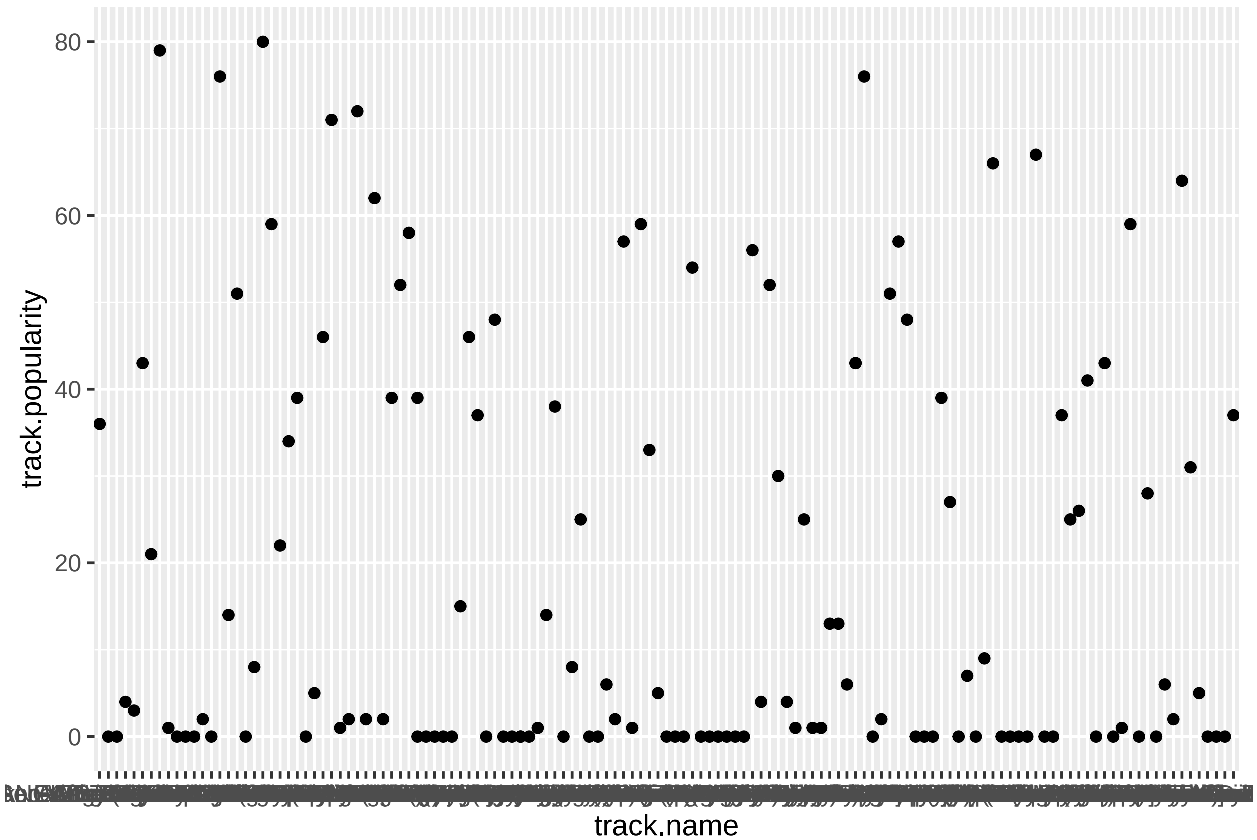
```
most_pop_song_each_year |>
  ggplot(aes(x = year, y = track.popularity)) +
  geom_point(color = "orange") +
  theme(axis.text.x = element_text(angle = 40,hjust = 1)) +
  labs(title = "Number one most popular song in each year",
       subtitle = "From 1978-2010",
       x = "Year",
       y = "popularity") +
 theme(panel.background = element_rect(fill = "black"),
        panel.grid.major = element_line(color = "black"),
        panel.grid.minor = element_line(color = "black"))
```



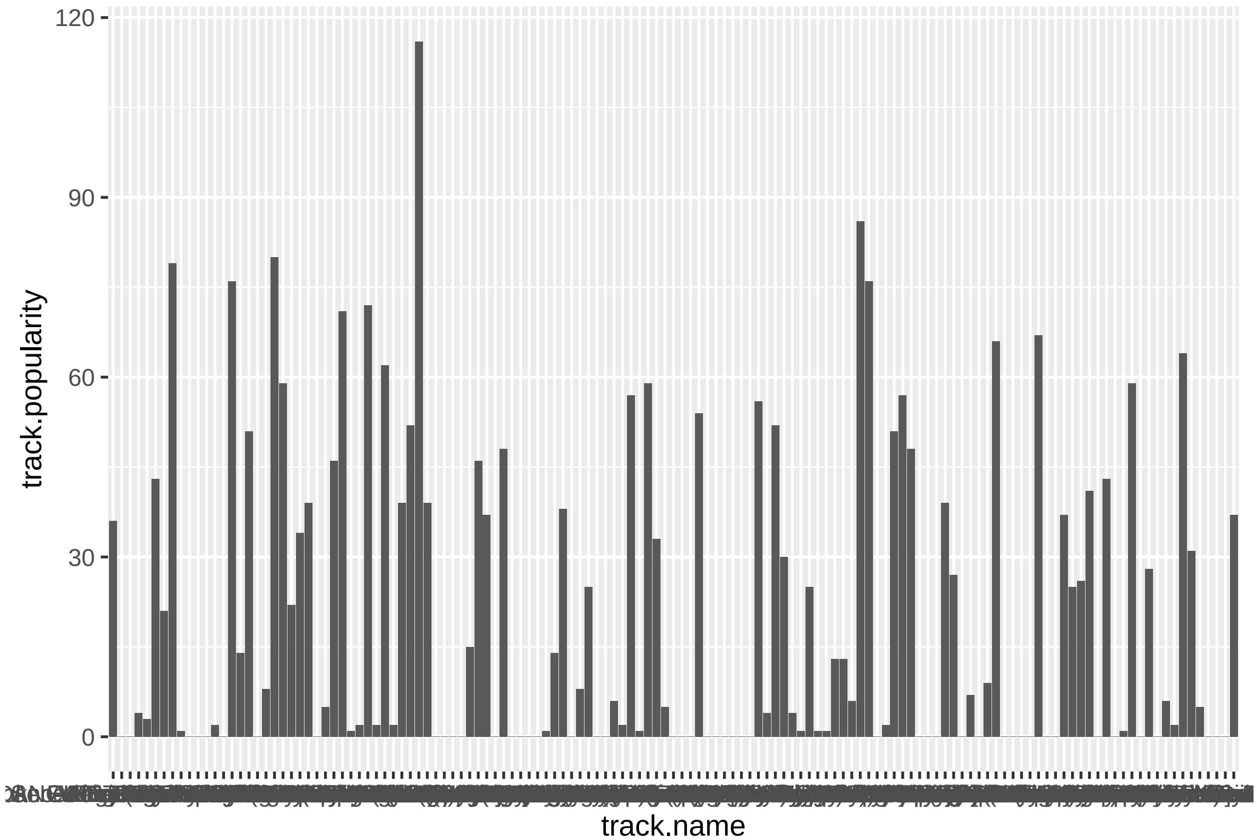
very unpopular songs having low ratings in popularity this playlist is filled with ear stredding ear bleeding horrible covers of songs

First looks at the trash_songs dataset

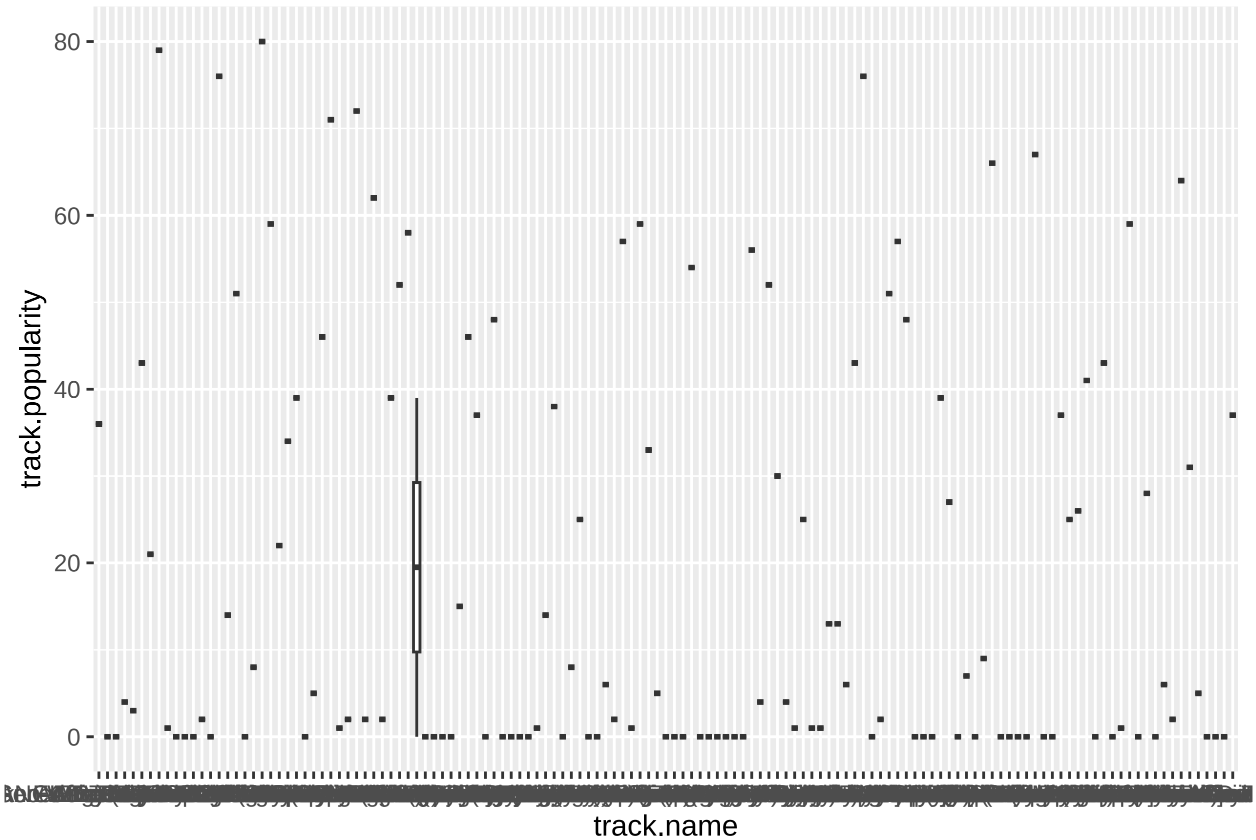
```
trash_songs |>
  ggplot(aes(x = track.name, y = track.popularity)) +
  geom_point()
```




```r
trash_songs |>
  ggplot(aes(x = track.name, y = track.popularity)) +
  geom_bar(stat = "identity")
```




```r
trash_songs |>
  ggplot(aes(x = track.name, y = track.popularity)) +
  geom_boxplot()
```



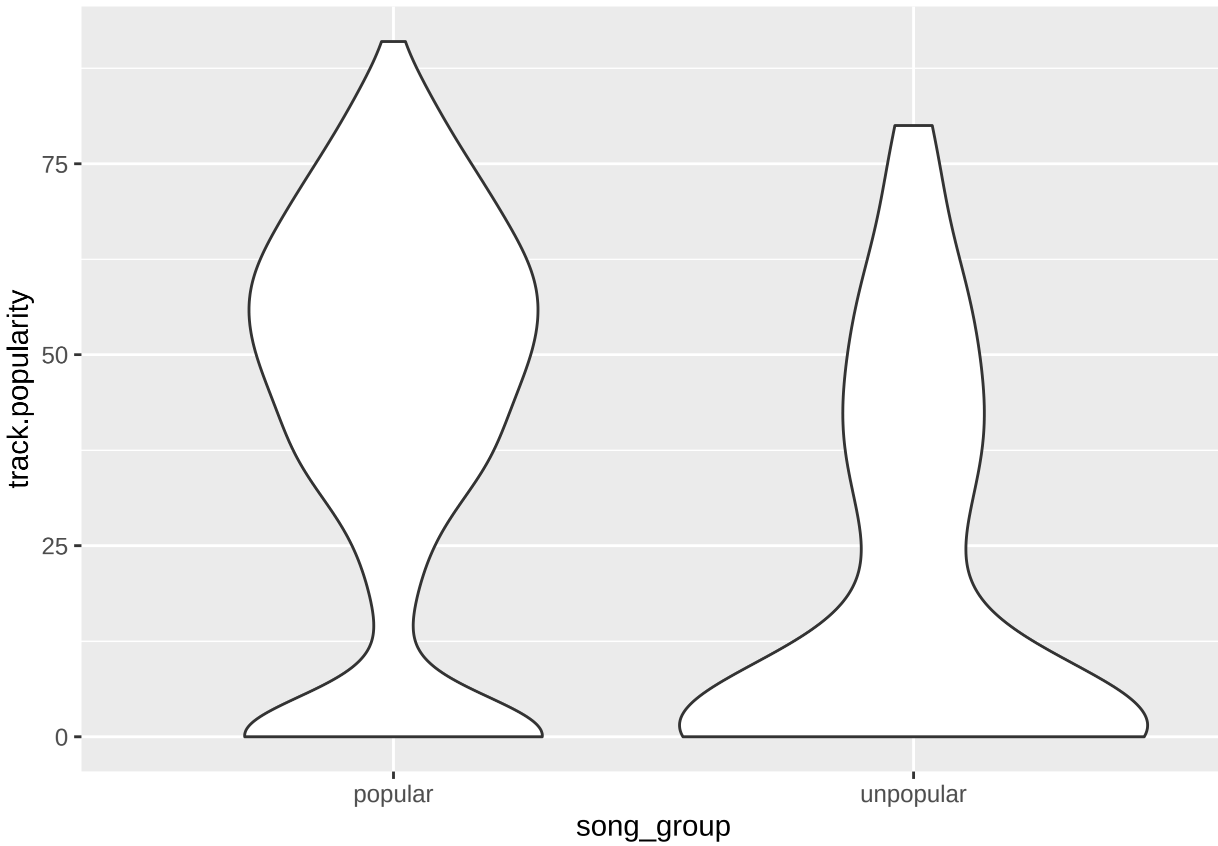
running t test between the popular songs and the unpopular songs

```r
t.test(trash_songs$track.popularity,top_songs_df$track.popularity)
```

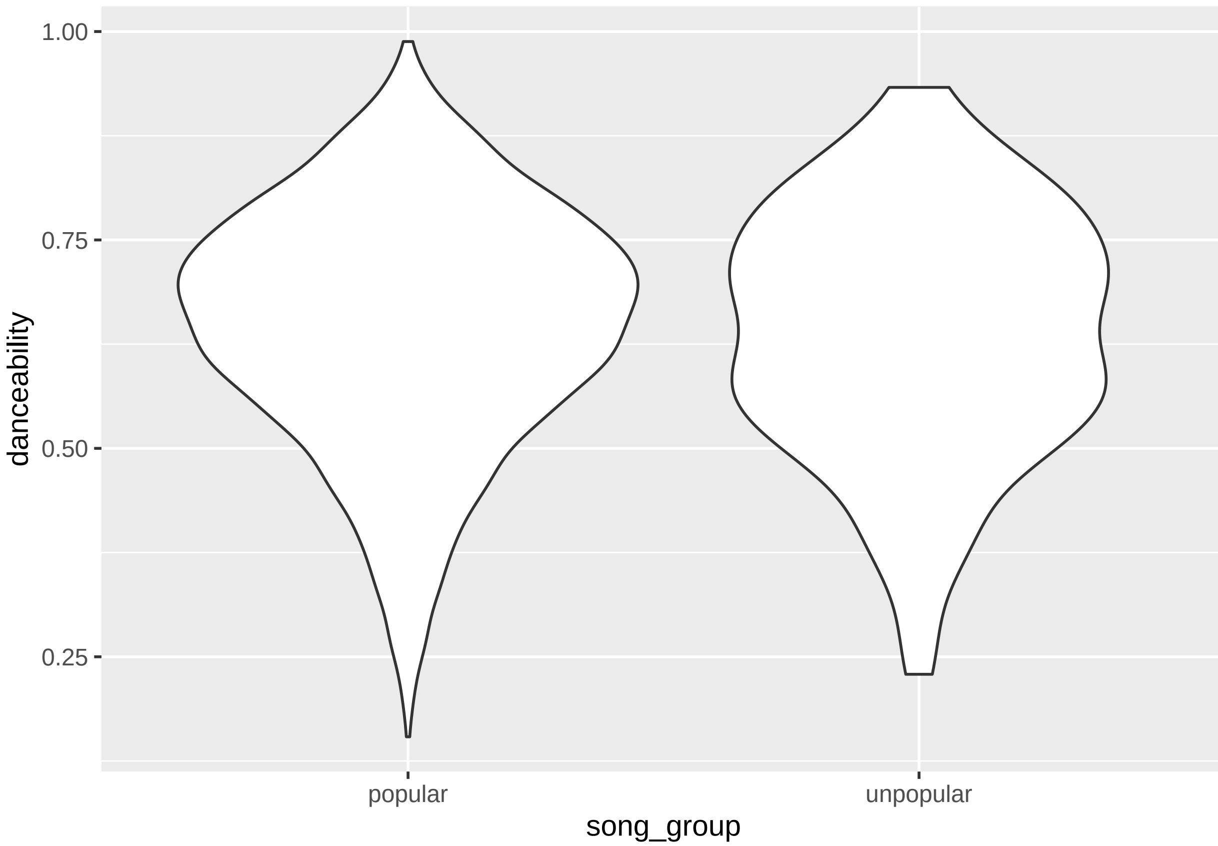
```
##
##  Welch Two Sample t-test
##
## data:  trash_songs$track.popularity and top_songs_df$track.popularity
## t = -27.09, df = 154.84, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -62.57331 -54.06789
## sample estimates:
## mean of x mean of y
##  19.53285  77.85345
```

This shows that there is indeed a difference in popularity in our two song sets. This helps validate our approach of using these two song sets to examine what variables might determine song popularity.

```r
all_songs_df |>
  ggplot(aes(x = song_group, y = track.popularity)) +
  geom_violin()
```

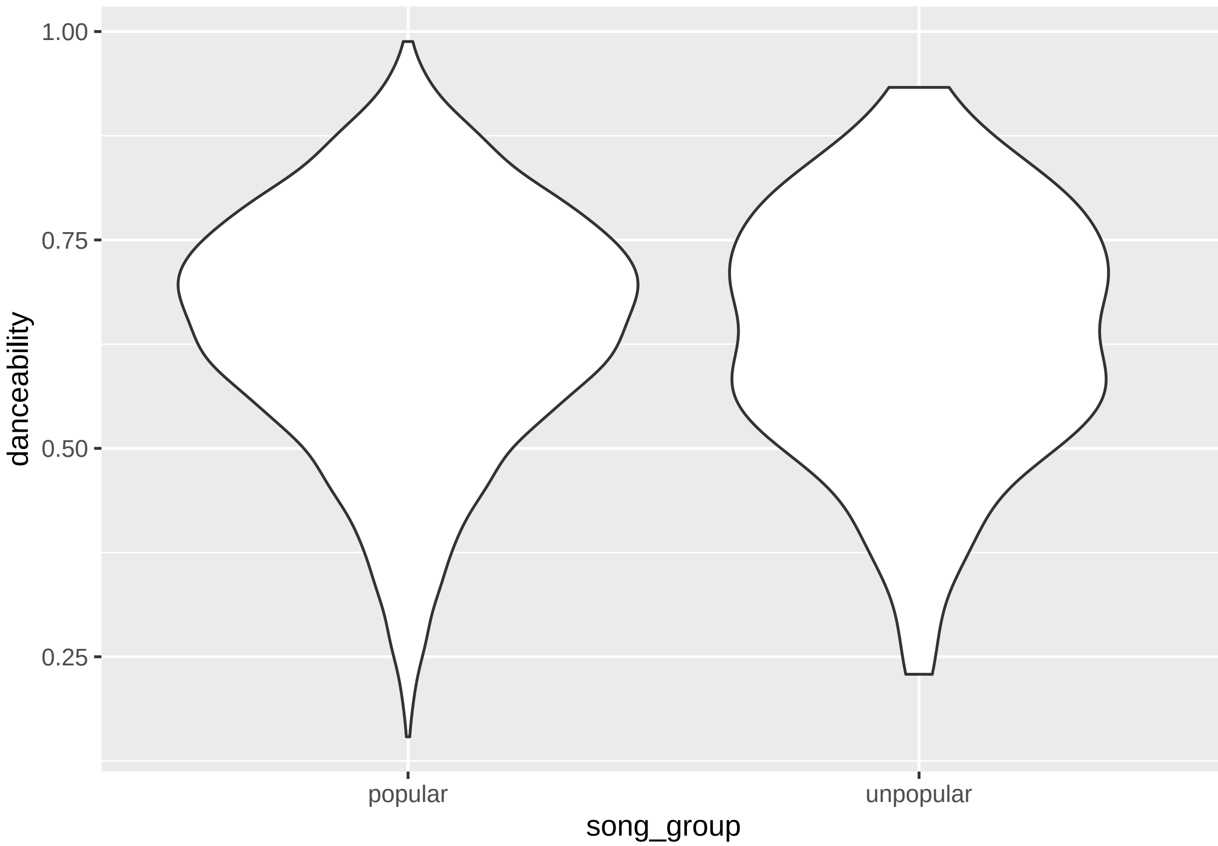



```r
ggplot(data = all_songs_df, aes(x = song_group, y = danceability)) +
  geom_violin()
```



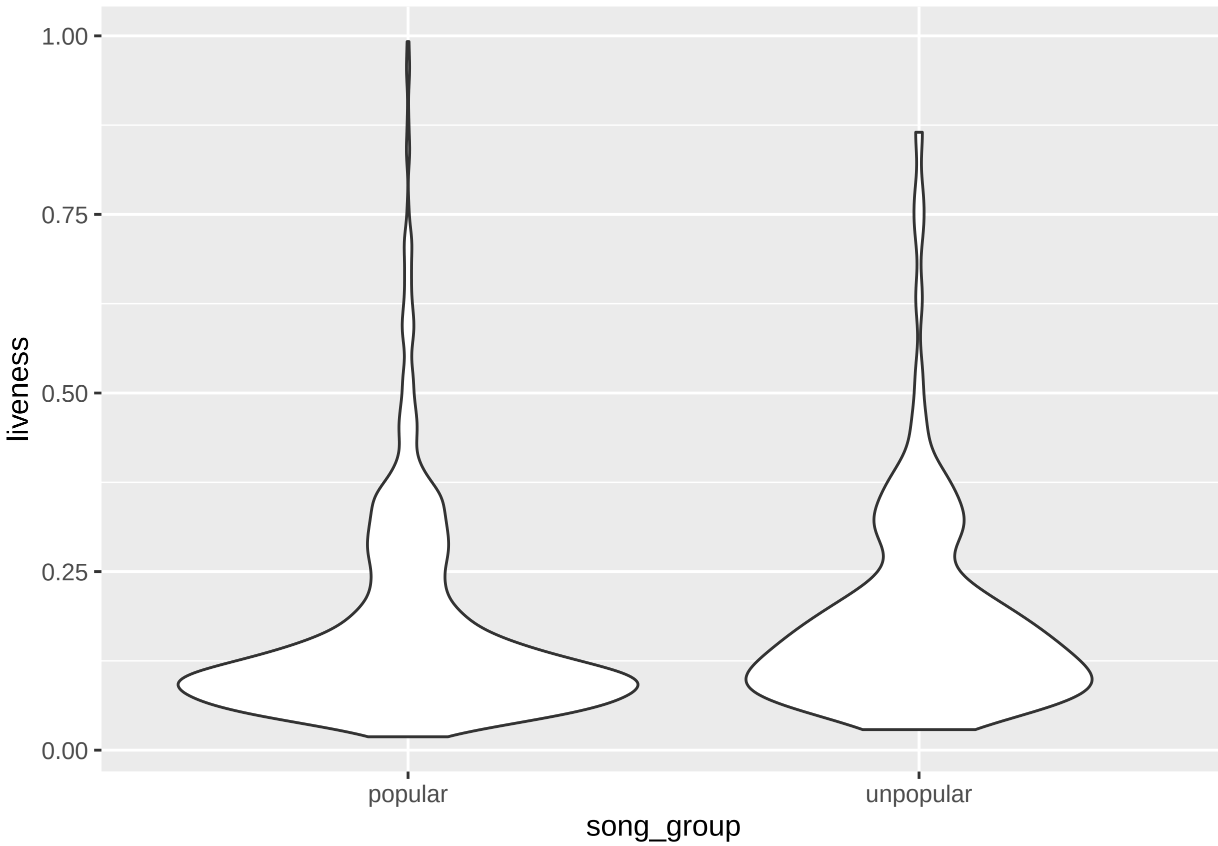
The set of plots below are comparing the popular and the unpopular songs to see if any of them have any noticeable differences but there actually isn't much of a difference between unpopular and popular songs for any of these variables which could possibly mean that unpopular songs try to copy popular songs cadence and flow.

```
all_songs_df |>
  ggplot(aes(x = song_group, y = danceability)) +
  geom_violin()
```

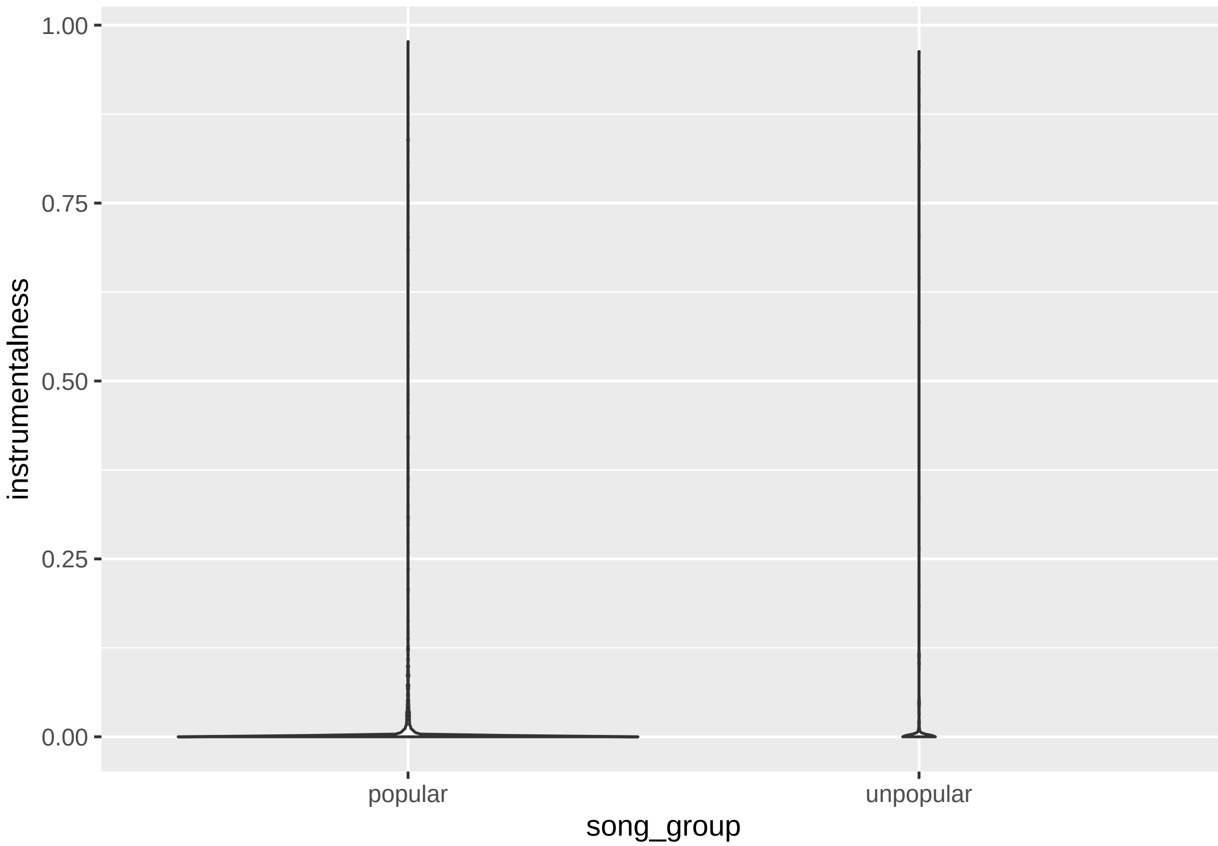



```
all_songs_df |>
  ggplot(aes(x = song_group, y = tempo)) +
  geom_violin()
```

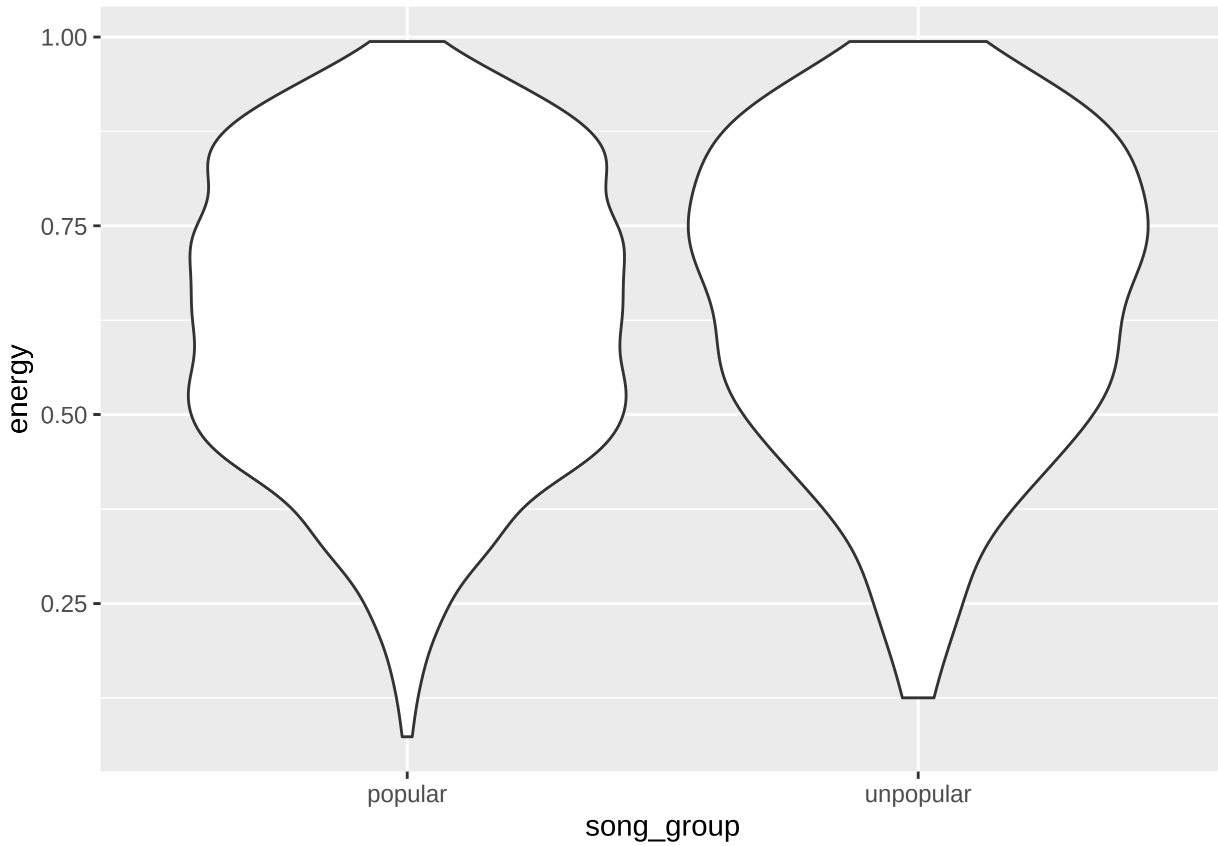

```r
all_songs_df |>
  ggplot(aes(x = song_group, y = liveness)) +
  geom_violin()
```



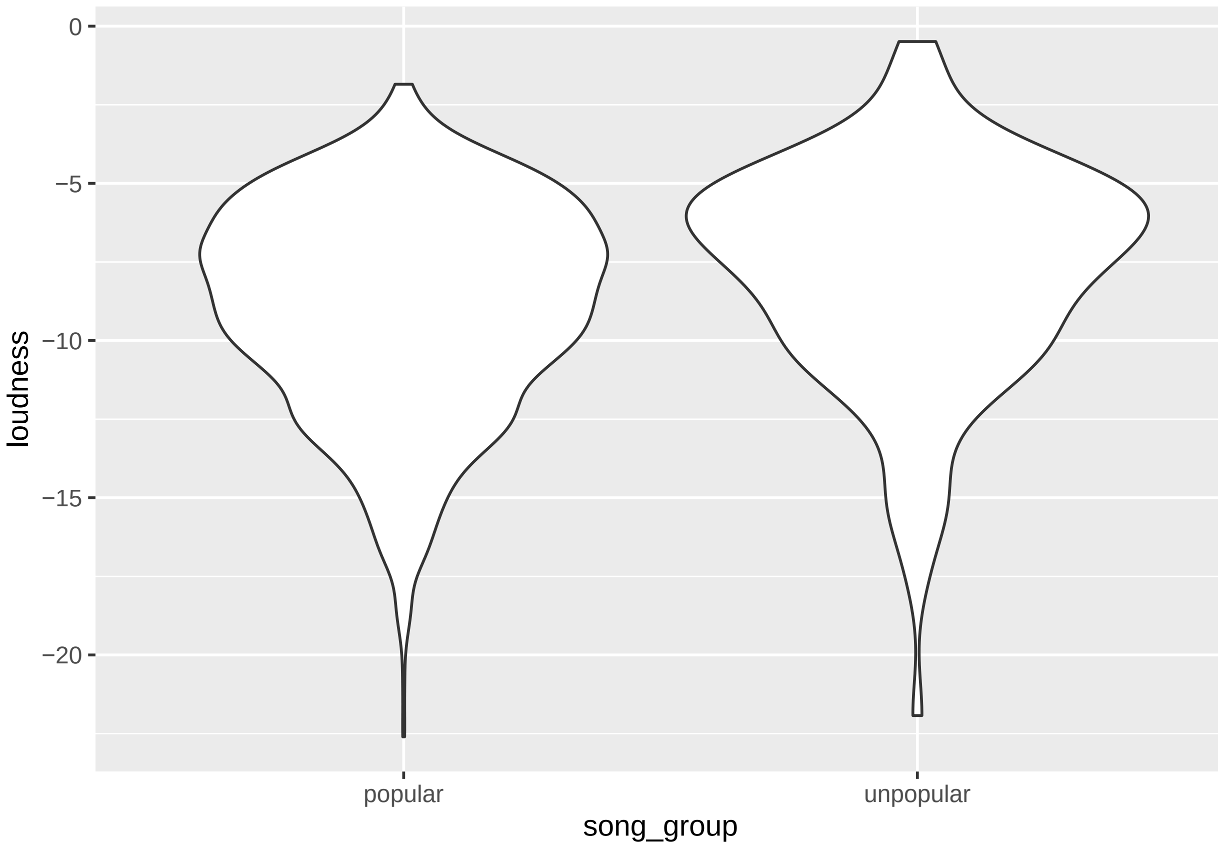

```
all_songs_df |>
  ggplot(aes(x = song_group, y = instrumentalness)) +
  geom_violin()
```



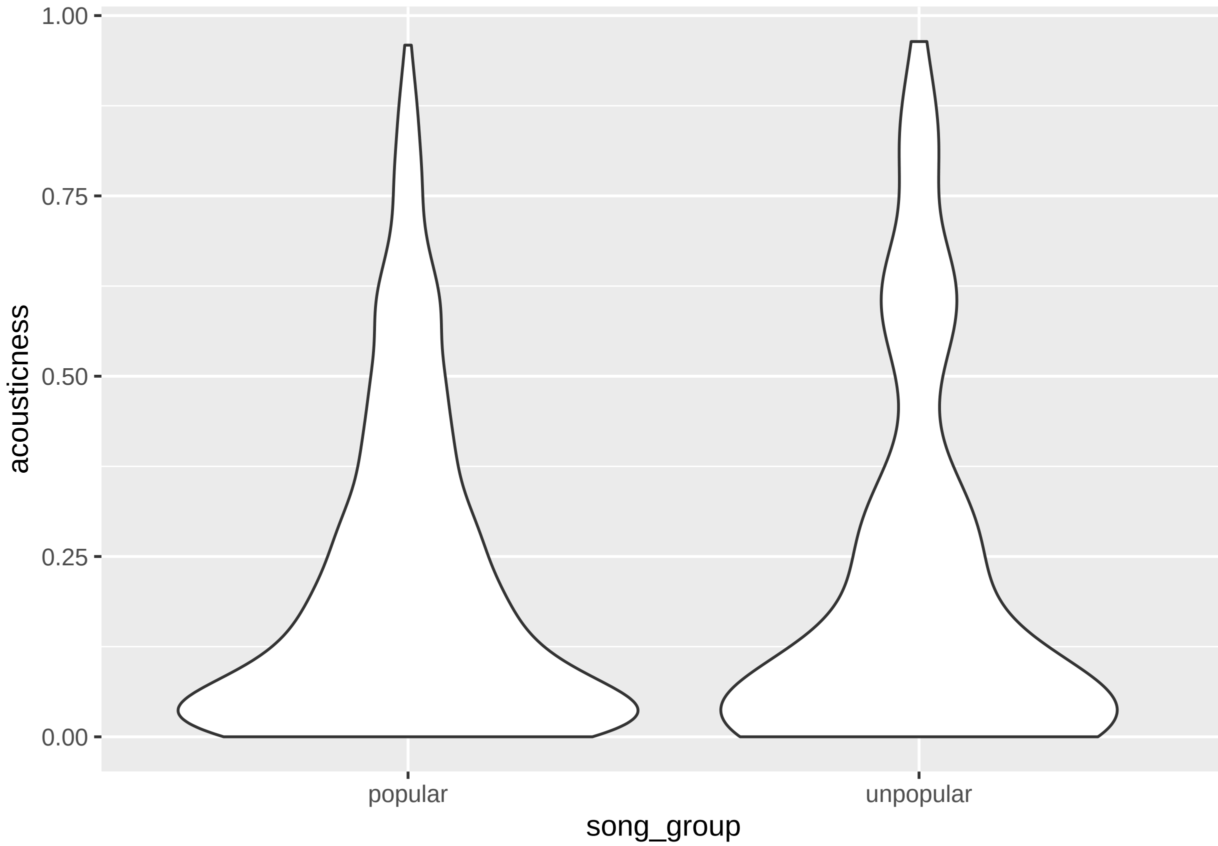

```r
all_songs_df |>
  ggplot(aes(x = song_group, y = energy)) +
  geom_violin()
```




```
all_songs_df |>
  ggplot(aes(x = song_group, y = loudness)) +
  geom_violin()
```




```
all_songs_df |>
  ggplot(aes(x = song_group, y = acousticness)) +
  geom_violin()
```




```r
all_songs_df |>
  ggplot(aes(x = song_group, y = key)) +
  geom_violin()
```


key
9
6
3
0
popular
unpopular
song_group