HW10

YOUR NAME HERE

You only need to submit your knitted pdf file to Moodle, but be sure your RMarkdown file is saved and accessible in your Submit folder on the RStudio server.

> Change author to your name and save your file as HW10_YOURNAMEHERE.Rmd to your Submit folder.

```
library(tidyverse)
library(nycflights13)
library(rvest)
library(lubridate)

flights <- flights %>%
  mutate(month_name = month(time_hour, label = TRUE, abbr = TRUE))

bigspotify <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data
  mutate(year = parse_number(substr(track_album_release_date, 3, 4)),
         decade = (year %/% 10)*10,
         decade = str_c(as.character(decade), "s", sep = "")) %>%
  mutate(year4 = parse_number(substr(track_album_release_date, 1, 4))) %>%
              # year4 is the 4-digit year as a numeric variable
   mutate(decade = fct_recode(decade, "00s" = "0s"),
         decade = fct_relevel(decade, "50s", "60s", "70s", "80s", "90s"))
              # decade is releveled as in HW9 key.


url <- glue::glue("http://www.basketball-reference.com/leagues/NBA_2022_totals.html")

bball <- read_html(url) %>%
  html_nodes("#totals_stats") %>%
  html_table() %>%
  data.frame(check.names = FALSE) %>%
  as_tibble() %>%
  mutate(across(G:PTS, parse_number),
         points_per_minute = PTS/MP,
         Pos = str_sub(Pos, 1,2),
         Pos = str_replace(Pos, "-", "")) %>%
  rename(FGpct = `FG%`,
         FTpct = `FT%`) %>%
  filter(MP > 10)

gss_sm <- read_csv("~/Mscs 264 S23/Class/Data/gss_sm.csv") %>%
  mutate(age_bin = cut_width(age, 10, center = 25))
```

Create two graphs using the bigspotify, bball, gss_sm, or flights datasets. You can use any variables you like or any subset (filter) that you want!

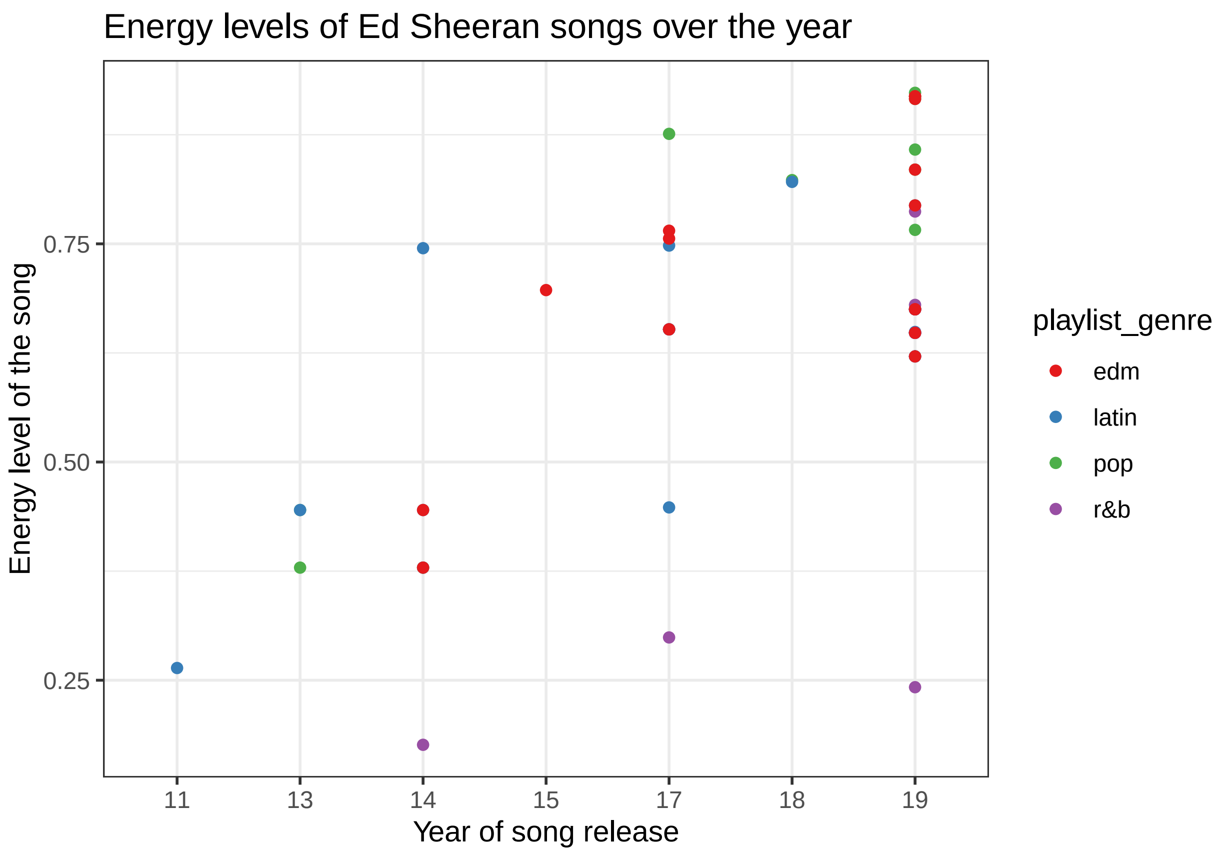
- The two graphs should be of different types (e.g. use different geoms: bar, histogram, density, violin, box, mosaic, line, smooth, point, bin2d, tile, etc)
- You should use group_by at least once (could use group_by/slice to create data subset for graph or labeling, could use group_by/summarize to create new dataset of summary stats for plot. . . .)
- At least one graph should use geom_text or geom_label
- Use a theme_XXX()
- If your graph includes a color or fill aesthetic, use a new scale. (scale_color_viridis_X, scale_fill_viridis_x, scale_color_brewer(), etc)
- For each graph, write a short paragraph describing the trends and relationships observed in the graph. You might also describe why the relationship might exist (or not exist!), if it surprises you, and what implications it has for our understanding of music/spotify.
- Be sure that your graphs have informative axis labels and titles, and that any factor variables are in a useful order.

You MAY work with other students on this homework, and you MAY use the same graphs. You should INDIVIDUALLY write the short paragraph for each. If you use the same graph as another student, please note their name in your homework (e.g. "Marge Simpson and I made these graphs together.")

If you want some inspiration, you could consider one of the following questions:

- Investigate a particular artist, genre, or subgenre. See how characteristics (valence, tempo, etc) relate to each other, or may change over time.

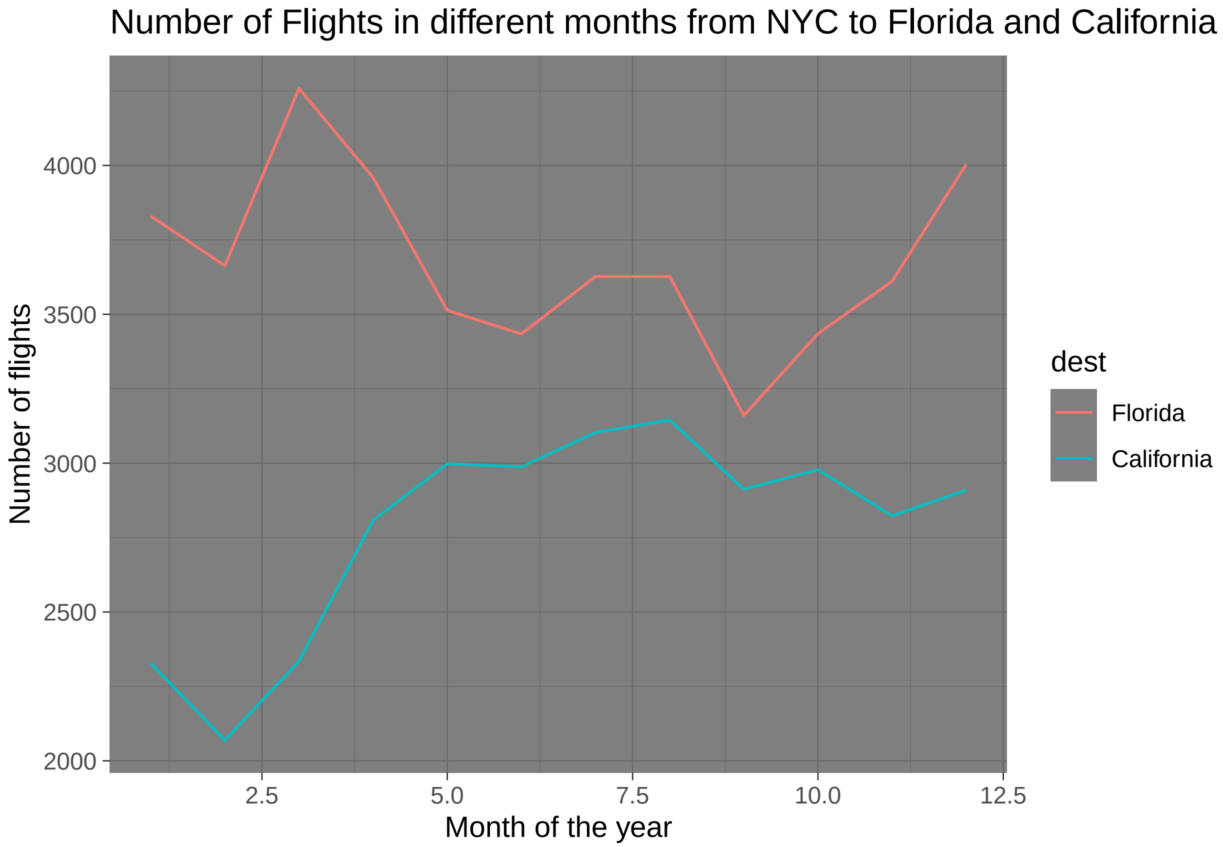
```r
bigspotify|>
  select(track_artist, energy, year, playlist_genre)|>
  filter(track_artist == "Ed Sheeran")|>
  mutate(year = as.factor(year))|>
  ggplot(aes(year, energy, color = playlist_genre))+
  geom_point()+
  scale_color_brewer(palette = "Set1")+
  scale_fill_viridis_d(option = "inferno")+
  labs(title = "Energy levels of Ed Sheeran songs over the year",
       x = "Year of song release", y = "Energy level of the song")+
  theme_bw()
```



Ans: The plot above displays relationships between the energy level and year of release of Ed Sheeran songs. As we can see the more recent song tghe song is the more the energy level the song has. This is because Ed Sheeran has been singing a lot of edm songs lately.

```
flights|>
  filter(dest %in% c("LAX", "SNA", "SFO", "SJC", "SAN", "FLL", "RSW", "JAX", "MIA", "MCO"))|>
  mutate(dest = fct_collapse(dest, "California" = c("LAX", "SNA", "SFO", "SJC", "SAN"), "Florida" = c("
  group_by(month, dest)|>
  summarise(flight_count = n())|>
  ungroup()|>
  ggplot()+
  geom_line(aes(x = month, y = flight_count, color = dest))+
  labs(title = "Number of Flights in different months from NYC to Florida and California", y = "Number 
  theme_dark()
```

```
## ‘summarise()‘ has grouped output by ’month’. You can override using the
## ‘.groups‘ argument.
```



Ans: The graph above is showing how flights to California from NYC reduces during the the summer months decreases while on the other hand the flights to florida during summer months increases. It is most likely due to a lot of people going to florida for summer vacation. And people try to avoid the summer heat in california.