Group assignment 3

1 Goals

- By conducting an **open-ended unsupervised learning analysis**, build confidence working outside the context of homework prompts.
- Work on collaboration! You get to pick your own group. Groups must be 3-4 students, except in extenuating circumstances (which need to be cleared with me).
- Create an example of your work that you might highlight when applying to internships, jobs, grad school, etc.

2 Directions

- Pick one of the datasets below. You can use other data but need to clear it with me first.
- Track and report your analysis in the **provided template**.
- Conduct a clustering analysis:
  - Write 1–2 paragraphs with some key takeaways. Be sure to address:
    - the various clusters you identify
    - the *features* of these clusters
    - which algorithm you used
  - Provide at least 3 visualizations that support these takeaways. These visualizations must each present unique information, i.e. not simply be different approaches to displaying similar information.
- Conduct a dimension reduction analysis (do PCA, not PCR):
  - Write 1–2 paragraphs with some key takeaways. Be sure to address:
    - the features of the first 2 principal components
    - how many PCs you might retain and why
    - the amount of the original information for which your retained PCs account
  - Provide at least 2 visualizations that support these takeaways. These visualizations must each present unique information, i.e. not simply be different approaches to displaying similar information.

3 Grading

This assignment is due on **Thursday, December 7**. If you'd like an extension, you must communicate with me at least 24 hours in advance of the due date. Passing this assignment requires the following:

- All words and code are your own (no tools like ChatGPT or help from others).
- You use the provided template and submit one knit **HTML** per group (not an Rmd or pdf).
- Your report includes:
  - All code necessary to reproducing your analysis.
  - Code that is structured and commented.
  - Clear discussions and visualizations that appropriately and completely address the prompts in the above Directions section.
- You are a supportive, active collaborator throughout the process. (Review group assignments 1 & 2 for specific expectations if you've forgotten them.)
- You complete the required **collaboration reflection form (click)**.

4 Data options

4.1 Billboard songs

This is a big dataset. My advice is to study ONE artist that has at least 25 songs in the dataset:

```
music <- read.csv("https://ajohns24.github.io/data/billboard.csv")

# Check out artists with at least 25 songs
music %>%
  count(performer) %>%
  filter(n >= 25) %>%
  select(performer)

# Pick just one of these artists to study
my_artist <- music %>%
  filter(performer == "___") %>%
  select(-performer) %>%
  group_by(song) %>%       # The last rows deal w songs that appear more than once
  slice_sample(n = 1) %>%
  ungroup()
```

Original source: TidyTuesday (click)

4.2 The 2019 Kenya census

```
# NA values on farm counts were replaced with 0
kenya_census <- read.csv("https://ajohns24.github.io/data/kenya_census.csv")
```

Original source: TidyTuesday (click)

4.3 NFL teams

```
library(fivethirtyeight)
data(nfl_fav_team)

# Access a codebook and context from the console
?nfl_fav_team
```

4.4 Something else!

You can choose a different dataset. Just run it by me first so that I can confirm whether it will work for this particular assignment.