In-Class Problem Set: Exploring Movie Data with Distribution and Color (R + GitHub *or* Canvas)

Goal. Use real movie data to practice visualizing distributions, comparing groups, and encoding multiple variables in a single plot. You will obtain the dataset from the course materials (GitHub *or* Canvas), build a reproducible workflow, generate several plots, interpret what they show, and submit your work via **GitHub *or* Canvas**.

What to submit (GitHub *or* Canvas).
- A script file: `scripts/lab.R`
- A short write-up: `outputs/writeup.md`
- Saved figures in `figures/` (see requirements below)

If you submit via Canvas, upload the same files listed above as individual files (or upload a single zip that preserves the directory structure).

Rules.
- Work inside an **R Project**.
- Use a **sequential, hard-coded workflow** (no user-defined functions).
- Save figures using code (`ggsave`); do not use screenshots.
- If you choose the **GitHub submission option**, Git commands must be run in the **Terminal tab**, not the R Console.

Submission options

You may submit this assignment using **either** of the following methods:
- **GitHub submission (recommended):** Commit and push your work to your GitHub repository. You will include Git proof (`git status` and `git log`) in your write-up.
- **Canvas submission:** Upload the required files directly to Canvas. You do *not* need to use GitHub if you choose this option.

Both submission methods are graded using the same rubric.

Mini codebook (use this; do not guess)

Each row represents one movie. Relevant variables include:
- `budget`: Production budget in USD (0 if unavailable).
- `revenue`: Worldwide box office revenue in USD (0 if unavailable).
- `director`: Director of the movie.
- `runtime`: Movie length in minutes.
- `vote_average`: Average user rating (0–10).
- `vote_count`: Number of user votes.
- `popularity`: Popularity score based on user engagement.

Questions

1. **Get the movie dataset (proof required).**
   (a) Choose **one** method:
   - **GitHub option:** In the **Terminal tab**, run:
     ```
     git status
     git pull
     ```
   - **Canvas option:** Download the movie dataset from Canvas and place it in your project `data/` folder.

   (b) **Pseudo-code (follow structure; fill in details):**
   ```
   # confirm working directory
   __________()

   # list files in data directory
   __________("data")

   # visually confirm expected movie file exists
   ```
   (c) Confirm the movie dataset exists in your project.
   (d) **Proof (write-up):** In `outputs/writeup.md`, paste:
   - the output of `getwd()`,
   - the output of `list.files("data")`.
2. **Load and summarize the dataset.**
   (a) Load the movie dataset into an object named `df`.
   (b) **Pseudo-code (intentionally incomplete):**
   ```
   df <- read.csv("data/__________.csv")

   # quick structure checks (choose at least two)
   __________(df)
   __________(df)
   __________(df)
   ```
   (c) Summarize the dataset to understand its structure.
   (d) **Proof (write-up):** Report:
   - number of rows and columns,
   - the range of `budget`,
   - the range of `revenue`.
3. **Plot distributions: movie budget and revenue.**
   Create two histograms:
   - one for `budget`,
   - one for `revenue`.
   - Repeat the process again and chose alternative bin sizes to assess how that changes the interpretation. Use one of the following formulas:
     - **Freedman–Diaconis (bin width):**
     $$h = 2 \cdot \mathrm{IQR}(x) \cdot n^{-1/3}$$
     - **Scott's Rule (bin width):**
     $$h = 3.5 \cdot s(x) \cdot n^{-1/3}$$

- **Sturges' Rule (number of bins):**

$$k = \lceil \log_2(n) + 1 \rceil$$

4. **Pseudo-code (structure only):**

```
# budget histogram
ggplot(df, aes(x = ________)) +
  geom_histogram(__________)

# revenue histogram
ggplot(df, aes(x = ________)) +
  geom_histogram(__________)
```

   Save the plots as:
   - `figures/budget_hist.png`
   - `figures/revenue_hist.png`
5. **Identify top-grossing directors and compare revenue.**
   (a) Identify the **top three directors** by total box office revenue.
   (b) Subset the data to movies directed by these three directors.
   (c) Create a **boxplot** showing the distribution of `revenue` for each director.
6. **Pseudo-code (leave blanks):**

```
top_directors <- df %>%
  group_by(__________) %>%
  summarize(total_revenue = ________) %>%
  arrange(__________) %>%
  slice(____)

df_top <- df %>%
  filter(__________ %in% top_directors$__________)
```

   Save the plot as:

   `figures/revenue_by_director.png`
7. **Scatter plot with size and color encodings.**
   Create a scatter plot with:
   - x-axis: `budget`
   - y-axis: `revenue`

   Then:
   - map one quantitative variable to **point size**,
   - map one categorical variable to **color**.
8. **Pseudo-code (structure only):**

```
ggplot(df, aes(x = ________, y = ________,
               size = ________, color = ________)) +
  geom_point(__________)
```

   Save the plot as:

   `figures/budget_revenue_scatter.png`

9. **Interpretation (write-up required).**
   In `outputs/writeup.md`, write 10–14 sentences addressing:
   - what the budget and revenue distributions reveal,
   - how revenues differ across top directors,
   - what relationships are most salient in the scatter plot,
   - how size and color encodings affect interpretation.
10. **Submit your work (GitHub *or* Canvas) + proof required.**
    (a) **Choose ONE submission method:**
    - **GitHub option:** Commit and push your work to GitHub.
    - **Canvas option:** Upload `scripts/lab.R`, `outputs/writeup.md`, and all required figures to Canvas.

    (b) **Pseudo-code (submission skeleton):**

    ```
    # GitHub option
    git status
    git add ______
    git commit -m "__________"
    git push
    ```

    (c) **Proof (write-up):**
    - If using **GitHub**: paste `git status` and `git log -1`.
    - If using **Canvas**: paste `list.files("scripts")`, `list.files("outputs")`, and `list.files("figures")`, and state that you submitted via Canvas.

Optional challenge (if you finish early)

Choose one plot and create an alternative version optimized for a **general public** audience. In 5–7 sentences, explain what design choices you changed and why.

Checklist (before you leave)

- `scripts/lab.R` runs top-to-bottom
- `outputs/writeup.md` exists and includes interpretation + proofs
- Required figures exist in `figures/`
- Work is either committed and pushed to GitHub *or* uploaded to Canvas