In-Class Problem Set: Totals vs Proportions in Political and Health Data (R + GitHub)

Goal. Practice the difference between visualizing *totals* and *proportions*, and understand how these choices change interpretation. You will use state-level election and public health data to construct total counts, proportions, and grouped comparisons, then visualize them using appropriate color encodings.

Dataset. The provided dataset contains one row per U.S. state with election results and cumulative death counts through November 7, 2020.

Key variables (excerpt).

- `totalvotes`, `biden_votes`, `trump_votes`
- `biden_share`, `trump_share`, `biden_margin`
- `covid_deaths_to_2020_11_07`
- `pneumonia_deaths_to_2020_11_07`

What to submit (in your GitHub repo).

- A script file: `scripts/lab.R`
- A short write-up: `outputs/writeup.md`
- Saved figures in `figures/` (see requirements below)

Rules.

- Work inside an **R Project**.
- Use a **sequential, hard-coded workflow** (no user-defined functions).
- Save figures using `ggsave()` (no screenshots).
- Git commands must be run in the **Terminal tab**, not the R Console.
- Color choices must be intuitive and accessibility-conscious.

Questions

1. **Pull the data and set up your workflow (proof required).**
   (a) In the **Terminal tab**, run:

   ```
   git status
   git pull
   ```

   (b) Confirm the dataset exists in your repo (path specified in the course GitHub).
   (c) Create the standard folder structure if missing: `scripts/`, `outputs/`, `figures/`.
   (d) **Proof (write-up):** In `outputs/writeup.md`, paste:
   - the output of `getwd()`,
   - the output of `list.files("data")`.
2. **Load and inspect the dataset.**
   (a) Load the dataset into an object named `df`.

(b) Summarize the data structure.
Suggested edit: Use `dim(df)`, `names(df)`, and a focused summary of vote and death variables.

(c) **Proof (write-up):** Report:
- number of rows and columns,
- the range of `totalvotes`,
- the range of `covid_deaths_to_2020_11_07`.

3. **Create a total illness death variable.**
(a) Create a new variable:

```
total_illness_deaths = covid_deaths + pneumonia_deaths
```

(b) Compute total illness deaths aggregated across all states.
(c) **Proof (write-up):** Report the national total and the minimum/maximum state totals.

4. **Visualize totals: bar plot of total illness deaths by state.**
Create a bar plot showing **total illness deaths by state**.

Required:
- Order states by total illness deaths.
- Clearly label axes.
- State explicitly that this is a *total*, not a proportion.

Save as:

`figures/total_illness_deaths_by_state.png`

5. **Visualize proportions: Trump vs Biden.**
Create a categorical variable indicating the election winner:
- `winner = "Biden"` if `biden_share > trump_share`
- `winner = "Trump"` otherwise

Then:
(a) Create a bar plot showing the **proportion of total illness deaths** accounted for by Trump- vs Biden-won states.
(b) Use an appropriate color scheme:
- red–blue for categorical winner, or
- red–purple–blue if you choose to encode `biden_margin`.

Required:
- Colors must be intuitive and colorblind-friendly.
- The plot must clearly communicate that values are *proportions*, not raw counts.

Save as:

`figures/illness_deaths_by_winner_proportion.png`

6. **Interpretation (write-up required).**
In `outputs/writeup.md`, write 12–16 sentences addressing:
- How interpretation changes when moving from totals to proportions.
- What the bar plot of totals emphasizes that the proportional plot hides.
- What the proportional plot emphasizes that the totals plot hides.
- Why your color choices are appropriate for the task.
- One way these plots could be misinterpreted if shown without context.

7. **Push your work to GitHub (proof required).**
(a) In the **Terminal tab**, run:

```
git status
git add .
git commit -m "Totals vs proportions lab"
git push
```

(b) **Proof (write-up):** Paste:
- the output of `git status` after committing,
- the output of `git log -1`.

Optional challenge (if you finish early)

Create an alternative visualization where:
- the same data are shown using **faceting** instead of color, or
- you encode `biden_margin` as a continuous color scale (red–purple–blue).

In 5–7 sentences, explain which version is clearer and for what audience.

Checklist (before you leave)

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