In-Class Problem Set: Scatterplots and Association (R + GitHub)

Goal. Use NBA player data to practice visualizing relationships between **two quantitative variables** using scatterplots, smoothers, polynomial fits, and careful interpretation. You will pull the data from GitHub, clean variable types, generate required figures, interpret what they show, and submit via GitHub.

Dataset. `basketball` (569 rows; 23 variables). Key variables you may use:

- IDs and labels: `PLAYER_NAME`, `TEAM_ABBREVIATION`
- Player attributes: `AGE`, `PLAYER_HEIGHT_INCHES`, `PLAYER_WEIGHT`
- Games + performance: `GP`, `PTS`, `REB`, `AST`, `NET_RATING`
- Rates (0–1): `OREB_PCT`, `DREB_PCT`, `USG_PCT`, `TS_PCT`, `AST_PCT`
- Draft info: `DRAFT_YEAR` (contains "Undrafted" for some players)

Important data note. Many columns are stored as **character strings**. You must convert variables you analyze into appropriate numeric types before plotting.

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.
- Use `theme_classic()` unless you explicitly justify an alternative.
- Handle missing values defensibly (state what you did).

Questions

1. **Pull the repo and confirm the dataset (proof required).**
   (a) In the **Terminal tab**, run:
   ```
   git status
   git pull
   ```
   (b) Confirm the dataset file exists in your repo (path posted in the course repository).
   (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()` from the project root, and
      - the output of `list.files("data")` showing the dataset file.
2. **Load and clean `basketball` (proof required).**

(a) Load the dataset into an object named `basketball`.
(b) Create a cleaned object named `basketball_clean` where you convert the following columns to numeric:

`AGE`, `PLAYER_HEIGHT_INCHES`, `PLAYER_WEIGHT`, `GP`, `PTS`, `REB`, `AST`, `NET_RATING`,

`OREB_PCT`, `DREB_PCT`, `USG_PCT`, `TS_PCT`, `AST_PCT`.

(c) Create a simple draft indicator:

`draft_status` = "Undrafted" vs "Drafted"

where "Drafted" means `DRAFT_YEAR` is not "Undrafted".
(d) **Proof (write-up):** Report:
- the number of rows in `basketball` and `basketball_clean`,
- a quick summary of at least three numeric columns (e.g., `AGE`, `PTS`, `USG_PCT`),
- one sentence describing how you handled missing or non-numeric values after conversion.

3. **Relationship 1: usage and scoring (scatterplot baseline).**
Make a scatterplot of `USG_PCT` (x) vs `PTS` (y).
   - Use `geom_point()` with an overplotting fix (e.g., `alpha` and/or smaller `size`).
   - Use clear labels. Since `USG_PCT` is a proportion, format the x-axis as percent if you can.
   - Use `theme_classic()`.

   Save as:

   `figures/usg_pts_scatter.png`

4. **Relationship 1 (extension): add a linear smoother with standard errors.**
Using the same x/y pairing as the previous question, add a linear fit:
   - `geom_smooth(method = "lm", se = TRUE)`
   - Keep the points visible (do not remove them).

   Write-up (3–5 sentences):
   - What does the fitted line claim?
   - What does the shaded band represent (in plain language)?
   - Does the band make you more or less confident about the trend?

   Save as:

   `figures/usg_pts_lm_se.png`

5. **Relationship 1B: usage and scoring efficiency.**
Make a scatterplot of `USG_PCT` (x) vs `TS_PCT` (y).
   - Use `geom_point()` with an overplotting fix (e.g., `alpha` and/or smaller `size`).
   - Format `USG_PCT` and `TS_PCT` as percents if you can.
   - Add a linear smoother with standard errors: `geom_smooth(method = "lm", se = TRUE)`.
   - Use `theme_classic()`.

   Write-up (3–5 sentences):
   - Does the trend look linear, or do you suspect curvature?
   - What does the SE ribbon suggest about uncertainty across usage levels?

   Save as:

   `figures/usg_ts_eff_lm_se.png`

6. **Relationship 1C: assists (raw vs rate).**
Make a scatterplot of `AST` (x) vs `AST_PCT` (y).

- Use `geom_point()` with an overplotting fix (e.g., `alpha` and/or smaller `size`).
- Format `AST_PCT` as a percent if you can.
- Add a linear smoother with standard errors: `geom_smooth(method = "lm", se = TRUE)`.
- Use `theme_classic()`.

Write-up (3–5 sentences):

- Are `AST` and `AST_PCT` close to a one-to-one relationship, or are there notable exceptions?
- Give one plausible basketball reason you might see players with similar `AST` but different `AST_PCT`.

Save as:

`figures/ast_astpct_lm_se.png`

7. **Final interpretation (write-up required).**
   In `outputs/writeup.md`, write 12–16 sentences addressing:
   - For each figure, what is the main pattern (direction + strength + shape)?
   - Name one outlier or "surprising" point pattern and what it could imply (without claiming causality).
   - Name one concrete plotting choice you made (alpha, axis formatting, legend placement, polynomial fit) and why it helped interpretability.
8. **Git workflow and submission (proof required).**
   You must show evidence of both **pull** and **push**, plus at least two commits.
   (a) After you finish cleaning the data (Question 2), commit and push:

```
git status
git add .
git commit -m "NBA relationships: clean basketball  types"
git push
```

   (b) Before your final push, run a fresh pull (to catch updates):

```
git pull
```

   (c) Commit and push your figures + write-up:

```
git status
git add .
git commit -m "NBA relationships: scatterplots + smoothers + writeup"
git push
```

   (d) **Proof (write-up):** Paste:
   - the output of `git status` after the final push (clean working tree), and
   - the output of `git log -2`.

Optional challenge (if you finish early)

Choose one:

- Create a small-multiple version of one relationship by faceting on `TEAM_ABBREVIATION` for **only** the teams with at least 12 players in the dataset. (State your rule and why you chose it.)
- Create a correlation heatmap for the numeric columns you used (and in 4–6 sentences, explain what the heatmap hides that a scatterplot reveals).

Checklist (before you leave)

- `scripts/lab.R` runs top-to-bottom
- Required figures exist in `figures/`:
  - `usg_pts_scatter.png`, `usg_pts_lm_se.png`
  - `usg_ts_eff_lm_se.png`, `ast_astpct_lm_se.png`
  - `size_reb_scatter.png`
  - `age_ts_poly2_se.png`
- `outputs/writeup.md` includes required proofs + interpretation
- At least two commits + pushed to GitHub