In-Class Problem Set: Distributions and Overplotting with Flight Data (R + GitHub *or* Canvas)

Goal. Practice visualizing and interpreting distributions using real transportation data. You will use histograms and density plots to explore skewed variables, compare groups, diagnose overplotting, and assess how bin width and smoothing adjustments change interpretation. You may submit via **GitHub or Canvas**.

Dataset. This problem set uses the `nycflights13` dataset.

Key variables.

- `dep_delay`: Departure delay (minutes)
- `arr_delay`: Arrival delay (minutes)
- `air_time`: Time in the air (minutes)
- `distance`: Flight distance (miles)
- `carrier`: Airline carrier code
- `origin`: Airport of origin (JFK, LGA, EWR)

What to submit (GitHub or Canvas).

- `scripts/lab.R`
- `outputs/writeup.md`
- Required figures in `figures/`

Rules.

- Work inside an **R Project**.
- Use a **sequential, hard-coded workflow** (no user-defined functions).
- Save all plots using `ggsave()`.
- If submitting via GitHub, Git commands must be run in the **Terminal tab**.

Questions

1. **Load data and confirm setup (proof required).**

```
# load required packages
library(__________)
library(__________)

# access dataset
df <- nycflights13::flights

# quick inspection
dim(df)
names(df)
head(df)
```

Include in your write-up:

- number of rows and columns,
- confirmation that dep_delay and arr_delay exist.

2. **Histograms and bin width comparison.**

   Create histograms for:

   - dep_delay
   - arr_delay

   You must:

   - Use at least **two different bin widths**.
   - Compare how bin width changes apparent skew, modality, and tail behavior.

   Pseudo-code structure:

   ```
   # Step 1: choose variable
   # Step 2: remove NA values
   # Step 3: create histogram with binwidth = ___
   # Step 4: create histogram with different binwidth = ___
   # Step 5: save figure

   ggplot(df, aes(x = __________)) +
     geom_histogram(binwidth = ________) +
     labs(title = "Histogram with binwidth = ___")

   ggplot(df, aes(x = __________)) +
     geom_histogram(binwidth = ________) +
     labs(title = "Histogram with binwidth = ___")
   ```

   Save as:

   - figures/dep_delay_hist.png
   - figures/arr_delay_hist.png

3. **Histogram vs Density (including smoothing adjustments).**

   For dep_delay, create:

   - One histogram
   - One density plot
   - A second density plot with a different smoothing adjustment (adjust)

   Pseudo-code structure:

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

   # density (default smoothing)
   ggplot(df, aes(x = ________)) +
     geom_density()

   # density with different smoothing
   ggplot(df, aes(x = ________)) +
     geom_density(adjust = ________)
   ```

   Compare:

   - How smoothing affects perceived multimodality.

- Whether density exaggerates or hides tail behavior relative to histogram.

Save as: `figures/dep_delay_hist_vs_density.png`

4. **Grouped distributions (color and faceting).**

   Compare `dep_delay` across:

   - `origin` or
   - at least three major `carriers`.

   You must:

   - Create one version using color mapping.
   - Create one version using `facet_wrap()`.

   Pseudo-code structure:

   ```
   # grouped with color
   ggplot(df, aes(x = ________, fill = ________)) +
     geom_histogram(position = "identity", alpha = 0.4)

   # grouped with faceting
   ggplot(df, aes(x = ________)) +
     geom_histogram(binwidth = ________) +
     facet_wrap(˜ ________)
   ```

   Save as: `figures/grouped_dep_delay.png`

5. **Overplotting and transparency.**

   Create scatter plots:

   - x-axis: `distance`
   - y-axis: `air_time`

   Produce:

   - One raw version
   - One version using `alpha`

   ```
   # raw scatter
   ggplot(df, aes(x = ________, y = ________)) +
     geom_point()

   # transparent scatter
   ggplot(df, aes(x = ________, y = ________)) +
     geom_point(alpha = ________)
   ```

   Save as:

   - `figures/airtime_distance_raw.png`
   - `figures/airtime_distance_alpha.png`

6. **Interpretation (write-up required).**

   Write 12–16 sentences addressing:

   - How skew appears in delay variables.
   - How bin width changes visible patterns.
   - How density smoothing alters interpretation.
   - When histograms are preferable to density plots.
   - How faceting changes clarity relative to color overlays.
   - Why transparency matters in scatter plots.

7. **Submit your work (GitHub or Canvas; proof required).**

- **GitHub option:**

  ```
  git status
  git add .
  git commit -m "Distributions and overplotting lab"
  git push
  ```

- **Canvas option:** Upload required files to Canvas.

Include proof in write-up:

- GitHub: output of `git status` and `git log -1`
- Canvas: confirmation note + list of uploaded files

Optional Extension: Faceting and Scaling

Choose `dep_delay` and create:

- A faceted histogram by `origin`
- A faceted density plot by `origin`

Then compare:

- Whether scale differences distort interpretation.
- Whether fixed vs free scales change conclusions.

```
ggplot(df, aes(x = dep_delay)) +
  geom_histogram(binwidth = ________) +
  facet_wrap(~ origin, scales = "fixed")

ggplot(df, aes(x = dep_delay)) +
  geom_histogram(binwidth = ________) +
  facet_wrap(~ origin, scales = "free")
```

Checklist

- Script runs top-to-bottom
- Required figures saved
- Write-up complete
- Submitted via GitHub or Canvas