Session 4: Data Transformation II — Pen-and-Paper Pair Exercise

PSY 410 | Data Science for Psychology

Name: _____ **Date:** _____

No laptop today? No problem. This handout lets you practice the same skills on paper. Work with a partner who has a laptop and compare your work at the end.

The data: `flights`

Here are 12 rows from the `flights` dataset, showing just `carrier` and `dep_delay`:

carrier	dep_delay
AA	8
AA	-2
AA	14
DL	-3
DL	5
DL	-1
F9	25
F9	38
F9	-4
UA	11
UA	-6
UA	22

Key: Negative delays mean the flight departed early. Delays are in minutes.

The task (same as the slide exercise)

1. Calculate the **average departure delay** for each carrier
2. Which airline has the **worst** average delay?
3. **Bonus:** Also calculate the number of flights per carrier. Does the worst airline just have fewer flights?

Your pen-and-paper version

Step 1: Group the data. Draw a line between each group of rows by carrier. (They're already sorted for you in the table above.)

Step 2: Summarize each group. Calculate the mean `dep_delay` for each carrier. Show your work:

carrier	dep_delay values	sum	n	mean (sum / n)
AA				
DL				
F9				
UA				

Which carrier has the worst (highest) average delay? _____

Does that carrier just have fewer flights? _____

Step 3: Write the code. Fill in the blanks to produce this summary:

```
flights |>
  ____________(carrier) |>
  ____________(
    avg_delay = _______(dep_delay, na.rm = TRUE),
    n_flights = ___()
  ) |>
  arrange(_______(avg_delay))
```

Step 4: Think about it. Why do we need `na.rm = TRUE` inside `mean()`?

Your answer: _____

Check your work

Compare your summary table and code with your partner's screen. Do your answers match?