

Session 3: Data Transformation I — Pen-and-Paper Pair Exercise

PSY 410 | Data Science for Psychology

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

This dataset has information about 336,776 flights departing NYC in 2013. Here are 10 rows with the columns you'll need:

carrier	origin	dest	dep_delay	arr_delay
UA	LGA	IAH	-7	-38
UA	JFK	LAX	161	154
UA	JFK	LAX	94	149
UA	JFK	LAX	197	169
UA	EWR	LAX	-2	-26
UA	JFK	LAX	-4	-21
UA	EWR	SNA	14	62
AA	EWR	DFW	111	NA
AA	JFK	LAX	-5	-19
VX	JFK	LAX	-9	-21

Key: Negative delays mean the flight was early. `arr_delay` is in minutes. NA means the data is missing.

The task (same as the slide exercise)

Using the `flights` dataset:

1. Find all **United Airlines** ("UA") flights
2. that were **more than 2 hours late** arriving
3. and were flying **to Los Angeles** ("LAX")

How many flights match? Which origin airport had the most?

Your pen-and-paper version

Step 1: Identify your filter conditions. Write down each condition as an R expression:

- United Airlines: `carrier` _____ "UA"
- More than 2 hours late arriving: `arr_delay` _____ _____ (hint: delays are in minutes)
- Flying to LAX: `dest` _____ "_____"

Step 2: Apply the filters by hand. Go through the 10 rows above. For each row, check all three conditions. Circle the rows that match ALL three.

Row	<code>carrier == "UA"?</code>	<code>arr_delay > 120?</code>	<code>dest == "LAX"?</code>	Keeps?
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

How many rows (out of these 10) match? _____

Step 3: Write the code. Fill in the blanks to write the `filter()` call:

```
flights |>  
  filter(____ == "UA", ____ > ___, ____ == "LAX")
```

Bonus: Which origin airport had the most matching flights? Tally the origins from your circled rows:

- EWR: _____
 - JFK: _____
 - LGA: _____
-

Check your work

Compare your circled rows and your code with your partner's screen.

Expected filter conditions: `carrier == "UA", arr_delay > 120, dest == "LAX"`

Matching rows from the sample above: Rows 2, 3, and 4 (UA, JFK, LAX, with arr_delay of 154, 149, and 169 — all > 120 minutes).

From the full dataset: 129 flights match. EWR had 82, JFK had 47.

Expected code:

```
flights |>  
  filter(carrier == "UA", arr_delay > 120, dest == "LAX")
```