

Week 2 · DataFrames and profiling

STAT163 — Data Manipulation Essentials

Oleh Omelchenko

2026-09-07

Pre-lecture quiz

The number to check

A website that lists flats and rooms for rent has published a result from its own research.

The average rent for a one-room flat in Kyiv is 14,000 UAH a month.

DISCUSS

You want to check that number for yourself.

- What data would you need before you could check this number at all?
- One row of that data — what is it, and what has to be written in it?
- Which listings might belong in the calculation, and which do not?

The listings you were given

listing	deal	rooms	area	price	currency	agent
L-101	rent	1.0	38.0	16400	UAH	Homeline
L-102	rent	1.0	38.0	16400	UAH	Kyiv Estate
L-103	daily rent	2.0	54.0	1350	UAH	Kyiv Estate
L-104	rent	1.0	31.0	13200	UAH	Homeline
L-105	rent	3.0	96.0	1750	USD	Kyiv Estate
L-106	sale	1.0	36.0	372000	UAH	Kyiv Estate
L-107	rent	1.0	42.0	12800	UAH	Homeline
L-108	rent	1.0		11500	UAH	owner
L-109	daily	2.0	57.0	980	UAH	Kyiv Estate
L-110	rent		35.0	14300	UAH	owner
L-111	rent	2.0	62.0	22500	UAH	Homeline
L-112	rent	1.0	40.0	420	USD	Kyiv Estate
L-113	rent	1.0	44.0	13900	UAH	homeline
L-114	rent	2.0	68.0	19700	UAH	Homeline
14 rows	4 values	1 missing	1 missing	average 36,943	2	4 as typed

Fourteen listings from one week. The **deal** column says what kind of offer each row is, and the **agent** column says who posted it.

Agenda

- **01 The average** · Cleaning data moves the number. It does not settle it.
- **02 Outside the data** · Many important things are not in the table
- **03 Who fixes it** · The person who can fix it decides what you do next
- **04 Duplicates** · The tool picks which copy survives unless you control it

1 · Cleaning data moves the number. It does not settle it.

Which rows do not belong in that average?

listing	deal	rooms	area	price	currency	agent
L-101	rent	1.0	38.0	16400	UAH	Homeline
L-102	rent	1.0	38.0	16400	UAH	Kyiv Estate
L-103	daily rent	2.0	54.0	1350	UAH	Kyiv Estate
L-104	rent	1.0	31.0	13200	UAH	Homeline
L-105	rent	3.0	96.0	1750	USD	Kyiv Estate
L-106	sale	1.0	36.0	372000	UAH	Kyiv Estate
L-107	rent	1.0	42.0	12800	UAH	Homeline
L-108	rent	1.0		11500	UAH	owner
L-109	daily	2.0	57.0	980	UAH	Kyiv Estate
L-110	rent		35.0	14300	UAH	owner
L-111	rent	2.0	62.0	22500	UAH	Homeline
L-112	rent	1.0	40.0	420	USD	Kyiv Estate
L-113	rent	1.0	44.0	13900	UAH	homeline
L-114	rent	2.0	68.0	19700	UAH	Homeline

DECIDE

Five things are true of this table. Which of them change the average?

- One flat appears twice.
- One row is not a rent at all.
- Two rows are a price per night.
- Two prices are not in hryvnia.
- One row is missing its area, another its room count.

What each decision does to the number

listing	deal	rooms	area	price	currency	agent
L-101	rent	1.0	38.0	16400	UAH	Homeline
L-102	rent	1.0	38.0	16400	UAH	Kyiv Estate
L-103	daily rent	2.0	54.0	1350	UAH	Kyiv Estate
L-104	rent	1.0	31.0	13200	UAH	Homeline
L-105	rent	3.0	96.0	1750	USD	Kyiv Estate
L-106	sale	1.0	36.0	372000	UAH	Kyiv Estate
L-107	rent	1.0	42.0	12800	UAH	Homeline
L-108	rent	1.0		11500	UAH	owner
L-109	daily	2.0	57.0	980	UAH	Kyiv Estate
L-110	rent		35.0	14300	UAH	owner
L-111	rent	2.0	62.0	22500	UAH	Homeline
L-112	rent	1.0	40.0	420	USD	Kyiv Estate
L-113	rent	1.0	44.0	13900	UAH	homeline
L-114	rent	2.0	68.0	19700	UAH	Homeline

36,943
AVERAGE

13,550
MEDIAN

Every row counted, exactly as the price column arrived. **14 rows.**

What each decision does to the number

listing	deal	rooms	area	price	currency	agent
L-101	rent	1.0	38.0	16400	UAH	Homeline
L-102	rent	1.0	38.0	16400	UAH	Kyiv Estate
L-103	daily rent	2.0	54.0	1350	UAH	Kyiv Estate
L-104	rent	1.0	31.0	13200	UAH	Homeline
L-105	rent	3.0	96.0	1750	USD	Kyiv Estate
L-106	sale	1.0	36.0	372000	UAH	Kyiv Estate
L-107	rent	1.0	42.0	12800	UAH	Homeline
L-108	rent	1.0		11500	UAH	owner
L-109	daily	2.0	57.0	980	UAH	Kyiv Estate
L-110	rent		35.0	14300	UAH	owner
L-111	rent	2.0	62.0	22500	UAH	Homeline
L-112	rent	1.0	40.0	420	USD	Kyiv Estate
L-113	rent	1.0	44.0	13900	UAH	homeline
L-114	rent	2.0	68.0	19700	UAH	Homeline

38,523
AVERAGE

13,200
MEDIAN

The same flat, listed by two agencies. Count it once. **13 rows.**

What each decision does to the number

listing	deal	rooms	area	price	currency	agent
L-101	rent	1.0	38.0	16400	UAH	Homeline
L-102	rent	1.0	38.0	16400	UAH	Kyiv Estate
L-103	daily rent	2.0	54.0	1350	UAH	Kyiv Estate
L-104	rent	1.0	31.0	13200	UAH	Homeline
L-105	rent	3.0	96.0	1750	USD	Kyiv Estate
L-106	sale	1.0	36.0	372000	UAH	Kyiv Estate
L-107	rent	1.0	42.0	12800	UAH	Homeline
L-108	rent	1.0		11500	UAH	owner
L-109	daily	2.0	57.0	980	UAH	Kyiv Estate
L-110	rent		35.0	14300	UAH	owner
L-111	rent	2.0	62.0	22500	UAH	Homeline
L-112	rent	1.0	40.0	420	USD	Kyiv Estate
L-113	rent	1.0	44.0	13900	UAH	homeline
L-114	rent	2.0	68.0	19700	UAH	Homeline

10,733
AVERAGE

13,000
MEDIAN

One row is a sale, not a rent. It answers a different question. **12 rows.**

What each decision does to the number

listing	deal	rooms	area	price	currency	agent
L-101	rent	1.0	38.0	16400	UAH	Homeline
L-102	rent	1.0	38.0	16400	UAH	Kyiv Estate
L-103	daily rent	2.0	54.0	1350	UAH	Kyiv Estate
L-104	rent	1.0	31.0	13200	UAH	Homeline
L-105	rent	3.0	96.0	1750	USD	Kyiv Estate
L-106	sale	1.0	36.0	372000	UAH	Kyiv Estate
L-107	rent	1.0	42.0	12800	UAH	Homeline
L-108	rent	1.0		11500	UAH	owner
L-109	daily	2.0	57.0	980	UAH	Kyiv Estate
L-110	rent		35.0	14300	UAH	owner
L-111	rent	2.0	62.0	22500	UAH	Homeline
L-112	rent	1.0	40.0	420	USD	Kyiv Estate
L-113	rent	1.0	44.0	13900	UAH	homeline
L-114	rent	2.0	68.0	19700	UAH	Homeline

12,647 **13,550**
AVERAGE MEDIAN

Two rows are a price per night, and they are not spelled the same. **10 rows.**

What each decision does to the number

listing	deal	rooms	area	price	currency	agent
L-101	rent	1.0	38.0	16400	UAH	Homeline
L-102	rent	1.0	38.0	16400	UAH	Kyiv Estate
L-103	daily rent	2.0	54.0	1350	UAH	Kyiv Estate
L-104	rent	1.0	31.0	13200	UAH	Homeline
L-105	rent	3.0	96.0	1750	USD	Kyiv Estate
L-106	sale	1.0	36.0	372000	UAH	Kyiv Estate
L-107	rent	1.0	42.0	12800	UAH	Homeline
L-108	rent	1.0		11500	UAH	owner
L-109	daily	2.0	57.0	980	UAH	Kyiv Estate
L-110	rent		35.0	14300	UAH	owner
L-111	rent	2.0	62.0	22500	UAH	Homeline
L-112	rent	1.0	40.0	420	USD	Kyiv Estate
L-113	rent	1.0	44.0	13900	UAH	homeline
L-114	rent	2.0	68.0	19700	UAH	Homeline

21,544

AVERAGE

15,350

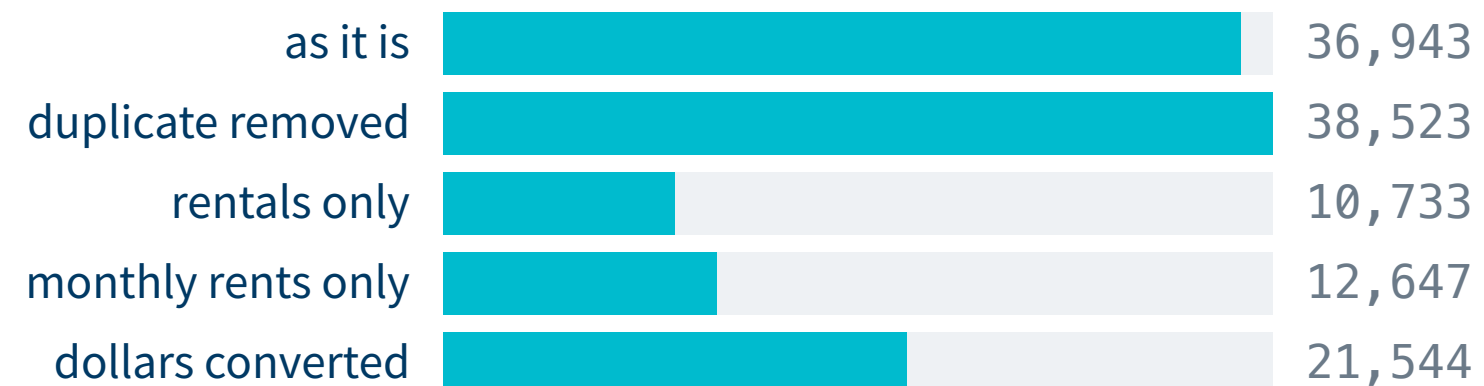
MEDIAN

Two prices are in dollars. At 42 to the hryvnia, the rate on the day they were posted, 1750 becomes 73,500 and 420 becomes 17,640. **10 rows.**

Four decisions, two ways of answering

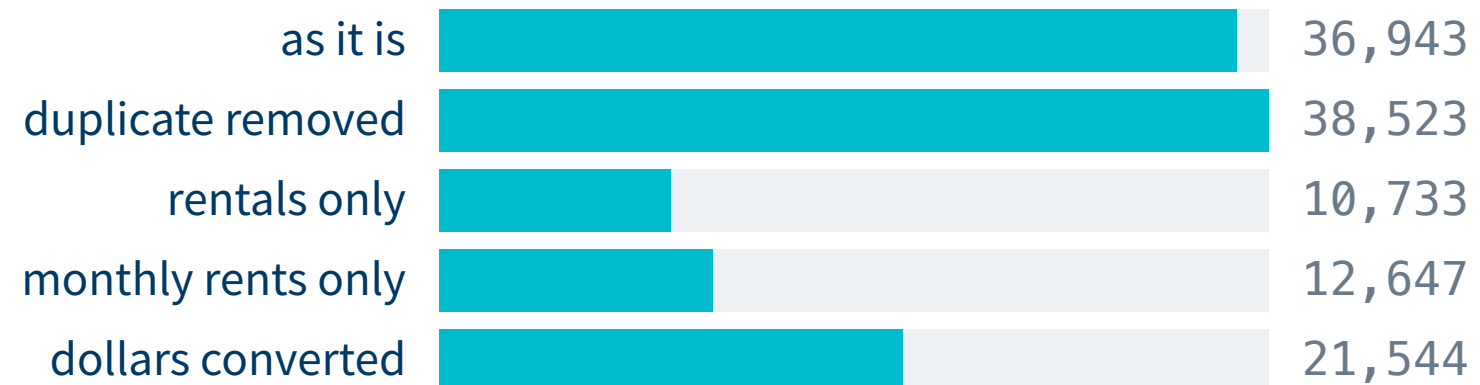
THE AVERAGE

THE MEDIAN



Four decisions, two ways of answering

THE AVERAGE

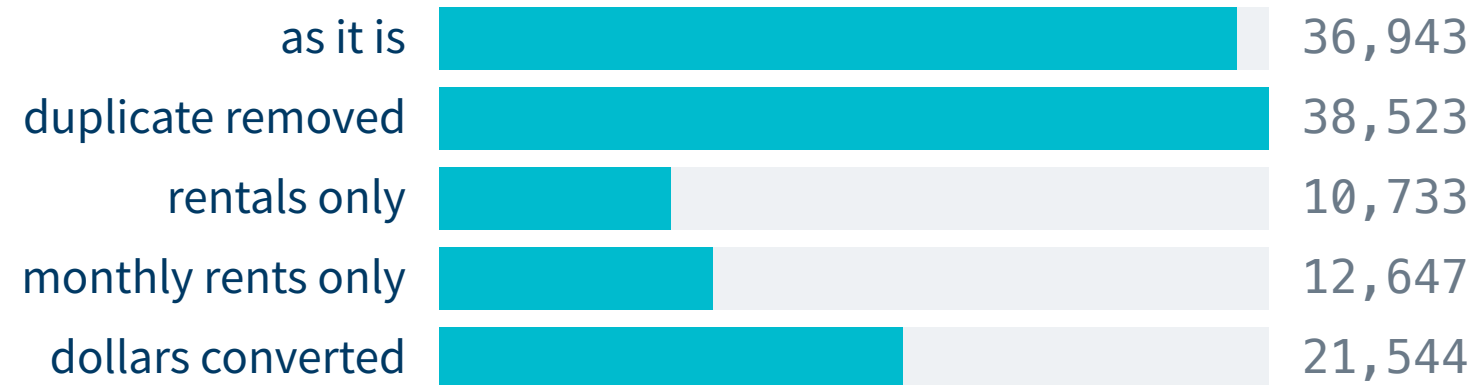


THE MEDIAN



Four decisions, two ways of answering

THE AVERAGE

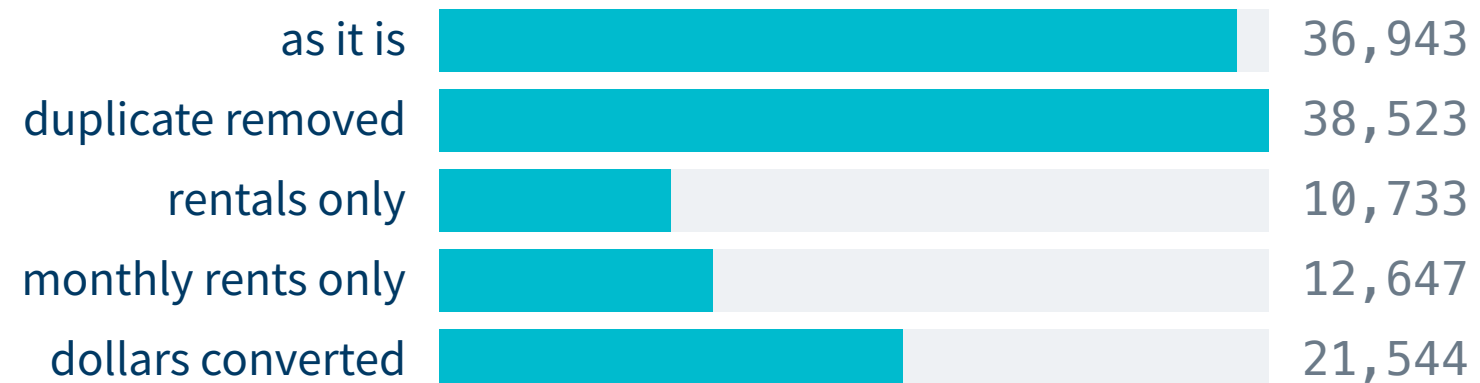


THE MEDIAN

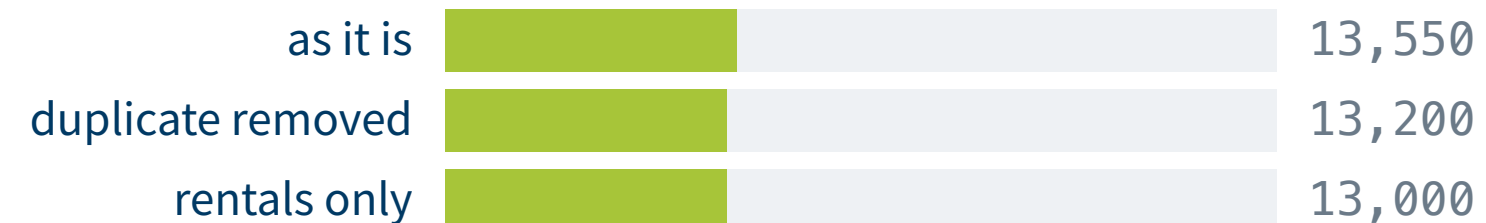


Four decisions, two ways of answering

THE AVERAGE

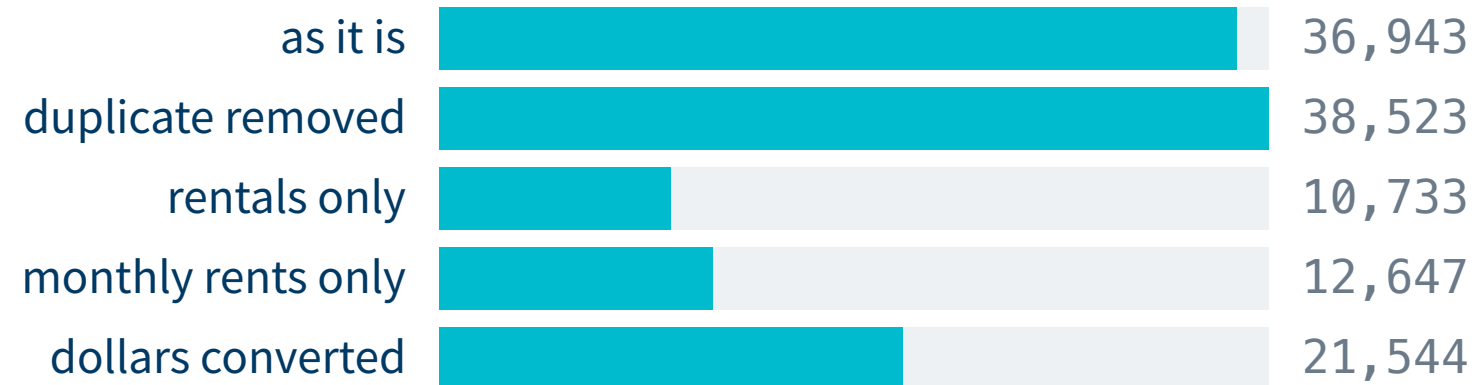


THE MEDIAN

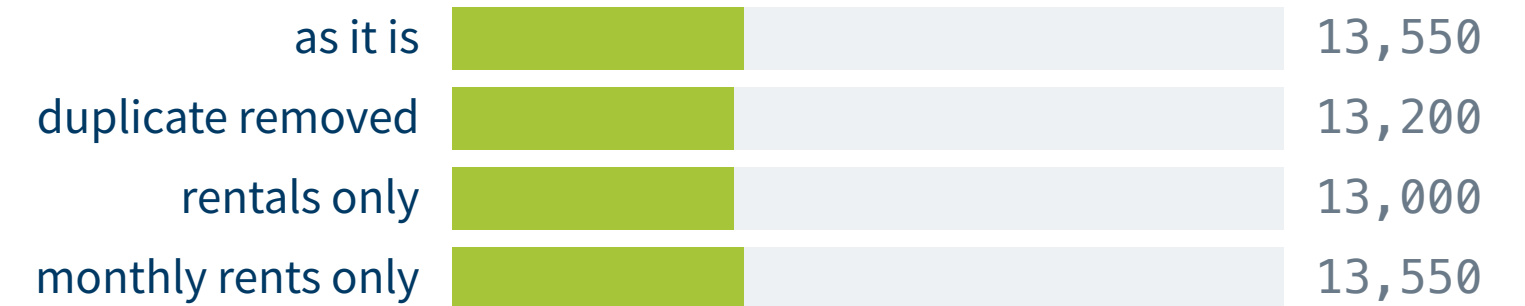


Four decisions, two ways of answering

THE AVERAGE

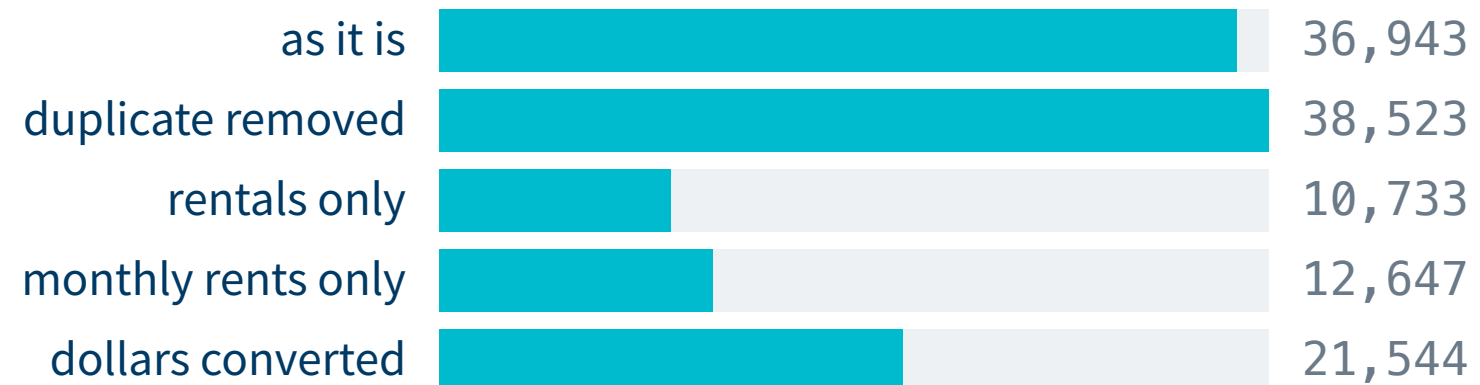


THE MEDIAN

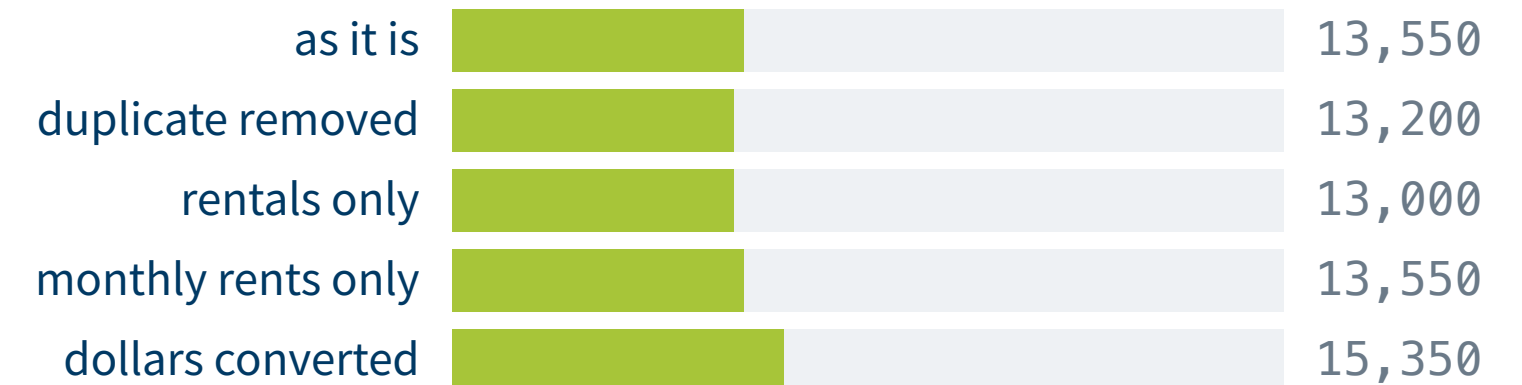


Four decisions, two ways of answering

THE AVERAGE

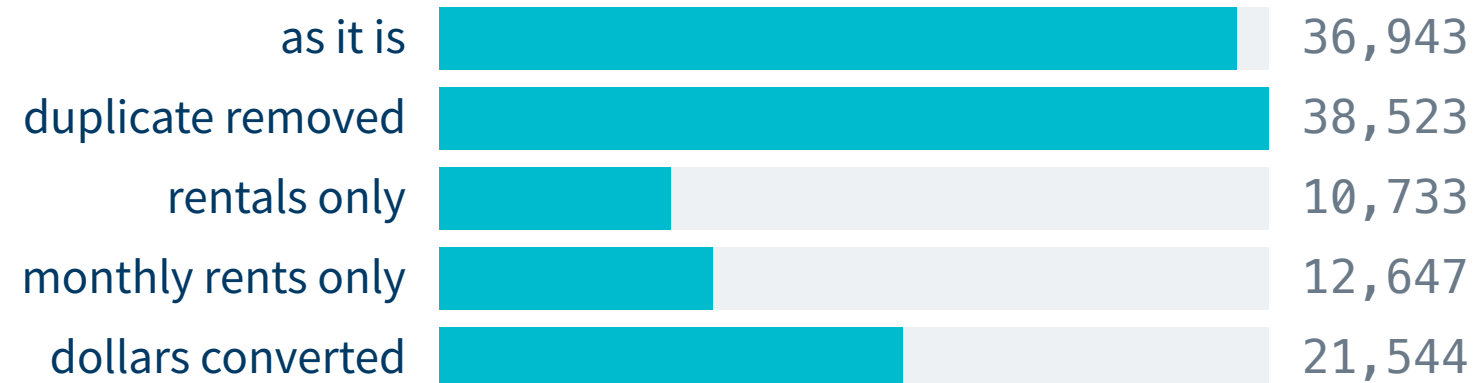


THE MEDIAN

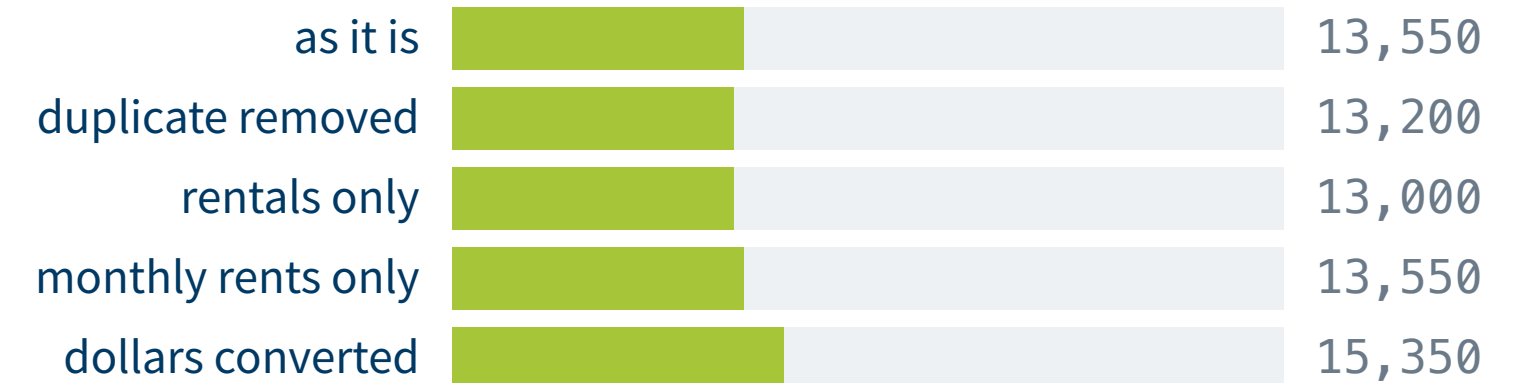


Four decisions, two ways of answering

THE AVERAGE

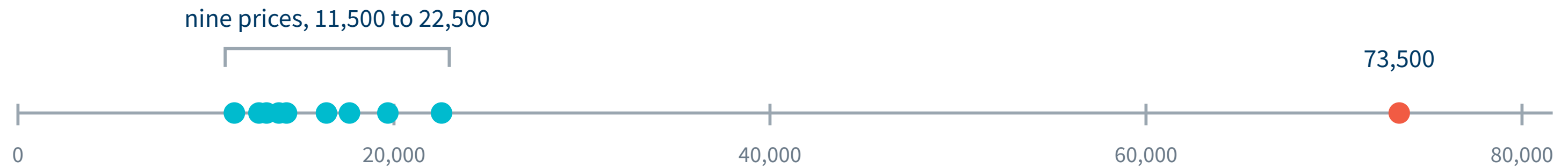


THE MEDIAN



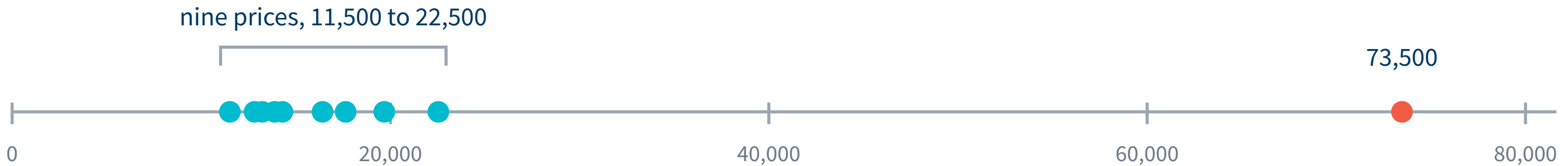
The average moves by a factor of nearly four across four decisions anybody could defend. The median stays between 13,000 and 15,350 the whole way.

Where the median sits



The ten monthly rents, each one a dot on the scale. Nine sit close together and one does not.

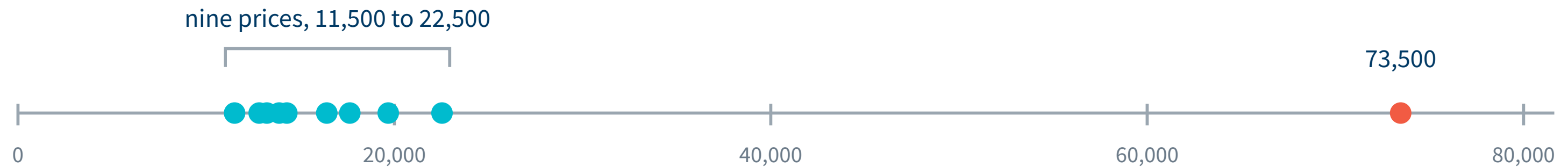
Where the median sits



The ten monthly rents, each one a dot on the scale. Nine sit close together and one does not.

1	2	3	4	5	6	7	8	9	10
11,500	12,800	13,200	13,900	14,300	16,400	17,640	19,700	22,500	73,500

Where the median sits



The ten monthly rents, each one a dot on the scale. Nine sit close together and one does not.

1	2	3	4	5	6	7	8	9	10
11,500	12,800	13,200	13,900	14,300	16,400	17,640	19,700	22,500	73,500

Sorted, there are ten values and no single one in the middle. The median is the fifth and the sixth averaged: 15,350. The average is 21,544, which is larger than eight of the ten.

Which one to report

A single very large value raises the average and leaves the median almost unchanged. When the question is what a typical flat costs, report the median, and state how many listings it was computed from.

Does the 14,000 hold?

LISTINGS, PRICES IN HRYVNIA

listing	rooms	area	price
L-101	1.0	38.0	16400
L-104	1.0	31.0	13200
L-107	1.0	42.0	12800
L-108	1.0		11500
L-112	1.0	40.0	17640
L-113	1.0	44.0	13900
L-110		35.0	14300

14,240

AVERAGE

13,550

MEDIAN

The 73,500 flat has three rooms and 96 square metres. The claim was about one-room flats, so that row never belonged. Both numbers land near the published 14,000.

Does the 14,000 hold?

LISTINGS, PRICES IN HRYVNIA

listing	rooms	area	price
L-101	1.0	38.0	16400
L-104	1.0	31.0	13200
L-107	1.0	42.0	12800
L-108	1.0		11500
L-112	1.0	40.0	17640
L-113	1.0	44.0	13900
L-110		35.0	14300

14,240
AVERAGE

13,550
MEDIAN

The 73,500 flat has three rooms and 96 square metres. The claim was about one-room flats, so that row never belonged. Both numbers land near the published 14,000.

L-110 has no room count, so it is out. If it is a one-room flat, there are seven rows and the median is 13,900. Nothing in the table says which.

2 · Many important things are not in the table

Which of these can the table answer?

DECIDE

Can the fourteen rows answer it, or does the answer sit outside them?

- Do two rows carry the same room count, area and price?
- Are those two rows one flat, or two identical flats?
- Does **price** include the bills?
- Is **area** the whole flat, or the living space only?
- Are these every listing that week, or only the ones an agency posted?

Which of these can the table answer?

DECIDE

Can the fourteen rows answer it, or does the answer sit outside them?

- Do two rows carry the same room count, area and price?
in the data
- Are those two rows one flat, or two identical flats?
- Does **price** include the bills?
- Is **area** the whole flat, or the living space only?
- Are these every listing that week, or only the ones an agency posted?

Which of these can the table answer?

DECIDE

Can the fourteen rows answer it, or does the answer sit outside them?

- Do two rows carry the same room count, area and price?
in the data
- Are those two rows one flat, or two identical flats?
outside the data
- Does **price** include the bills?
- Is **area** the whole flat, or the living space only?
- Are these every listing that week, or only the ones an agency posted?

Which of these can the table answer?

DECIDE

Can the fourteen rows answer it, or does the answer sit outside them?

- Do two rows carry the same room count, area and price?
in the data
- Are those two rows one flat, or two identical flats?
outside the data
- Does **price** include the bills? **outside the data**
- Is **area** the whole flat, or the living space only?
- Are these every listing that week, or only the ones an agency posted?

Which of these can the table answer?

DECIDE

Can the fourteen rows answer it, or does the answer sit outside them?

- Do two rows carry the same room count, area and price?
in the data
- Are those two rows one flat, or two identical flats?
outside the data
- Does **price** include the bills? **outside the data**
- Is **area** the whole flat, or the living space only? **outside the data**
- Are these every listing that week, or only the ones an agency posted?

Which of these can the table answer?

DECIDE

Can the fourteen rows answer it, or does the answer sit outside them?

- Do two rows carry the same room count, area and price?
in the data
- Are those two rows one flat, or two identical flats?
outside the data
- Does **price** include the bills? **outside the data**
- Is **area** the whole flat, or the living space only? **outside the data**
- Are these every listing that week, or only the ones an agency posted? **outside the data**

Which of these can the table answer?

DECIDE

Can the fourteen rows answer it, or does the answer sit outside them?

- Do two rows carry the same room count, area and price?
in the data
- Are those two rows one flat, or two identical flats?
outside the data
- Does `price` include the bills? **outside the data**
- Is `area` the whole flat, or the living space only? **outside the data**
- Are these every listing that week, or only the ones an agency posted? **outside the data**

One of five.

The checks you run inside a table tell you whether it is in good order.

What a column means, and which rows reached the table at all, is written down somewhere else: in the documentation, in a data dictionary, or in the head of whoever supplied the file. No operation on the table reaches it.

When the answer is nowhere

DECIDE

Two agencies posted the same rooms, the same area and the same price on the same day. Neither agency can say whether that is one flat or two, and no documentation covers it.

- What could you assume, and what does each assumption do to the count of flats?
- Which one would you defend, and on what grounds?
- What has to be written down so that somebody else can check your work?

Recording an assumption

Not every question has an answer. Choose an assumption, state why you chose it, and record it beside the number. A number that carries its assumptions can be checked by somebody else. A number without them cannot.

3 · The person who can fix it decides what you do next

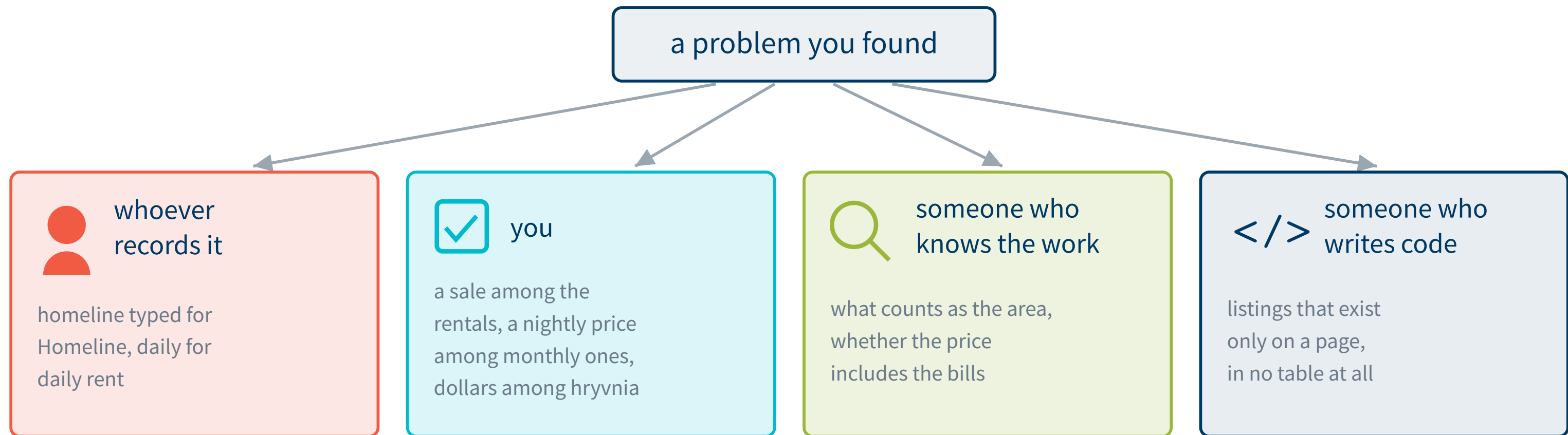
Five things in this table, and one that is not

What you see	What it is
daily rent and daily	two names for the same kind of offer
Homeline and homeline	a capital letter turns one agency into two
16400 UAH beside 420 USD	prices nobody can compare until one is converted
rent beside sale	rows that answer different questions
area and price, with nothing that says what they cover	two columns whose meaning nobody wrote down
a flat whose price is given only on request	a listing that never reached the table

DECIDE

For each one: who can fix it, and what happens if nobody does?

Where each problem has to go



Only the second box is yours to repair alone, and naming the box is already the decision about what you do next. Some problems reach no box at all. Those are the ones you close with a recorded assumption.

4 • The tool picks which copy survives unless you control it

The same flat, or two flats?

LISTINGS — ONE DISTRICT, WITH THE DATE EACH LISTING WAS POSTED

listing	posted	rooms	area	price	agent
L-101	2026-03-02	1.0	38.0	16400	Homeline
L-102	2026-03-02	1.0	38.0	16400	Kyiv Estate
L-104	2026-03-03	1.0	31.0	13200	Homeline
L-121	2026-03-14	1.0	38.0	15900	Homeline
L-122	2026-03-14	1.0	38.0	15900	Homeline
L-111	2026-03-09	2.0	62.0	22500	Homeline

DECIDE

Three pairs could be the same flat twice.

- Two agencies posted the same flat on the same day. One flat, or two?
- One agency posted the same flat twice on one day. One flat, or two?
- The same flat appears again twelve days later, cheaper. One listing, or two?
- If you keep one row of a pair, which one, and what makes it the right one?

Which copy the check flags

```
listings.duplicated()
```

listing	posted	rooms	area	price	agent
L-101	2026-03-02	1.0	38.0	16400	Homeline
L-102	2026-03-02	1.0	38.0	16400	Kyiv Estate
L-104	2026-03-03	1.0	31.0	13200	Homeline
L-121	2026-03-14	1.0	38.0	15900	Homeline
L-122	2026-03-14	1.0	38.0	15900	Homeline
L-111	2026-03-09	2.0	62.0	22500	Homeline

None. Every row carries its own `listing` id, so no two rows are equal on every column. An id column hides every duplicate there is.

Which copy the check flags

```
listings.duplicated(subset=["rooms", "area", "price"])
```

listing	posted	rooms	area	price	agent
L-101	2026-03-02	1.0	38.0	16400	Homeline
L-102	2026-03-02	1.0	38.0	16400	Kyiv Estate
L-104	2026-03-03	1.0	31.0	13200	Homeline
L-121	2026-03-14	1.0	38.0	15900	Homeline
L-122	2026-03-14	1.0	38.0	15900	Homeline
L-111	2026-03-09	2.0	62.0	22500	Homeline

Two. The check reads only the three columns that describe the flat, and marks the second row of each pair, as the file has them.

Which copy the check flags

```
listings.duplicated(  
  subset=["rooms", "area", "price"],  
  keep=False)
```

listing	posted	rooms	area	price	agent
L-101	2026-03-02	1.0	38.0	16400	Homeline
L-102	2026-03-02	1.0	38.0	16400	Kyiv Estate
L-104	2026-03-03	1.0	31.0	13200	Homeline
L-121	2026-03-14	1.0	38.0	15900	Homeline
L-122	2026-03-14	1.0	38.0	15900	Homeline
L-111	2026-03-09	2.0	62.0	22500	Homeline

Four. Every row of every matching pair. This is the one to run when you want to look at them before you decide.

Takeaways

- Cleaning a column moves the number it produces. It does not settle what the number should be.
- The median moves less than the average.
- A value far from all the others may not belong in the question you asked.
- To say what a typical value is, you have to work several things out. Some are inside the data. Some are only outside it, in the documentation or with whoever supplied the file.
- Sorting a problem by who can act on it decides what you do next. Most problems are not yours to repair.
- Some questions have no answer anywhere. Choose an assumption, say why, and record it beside the number.
- A duplicate check compares only the columns you name. Include an id column and every row looks unique.
- If you do not choose which copy to keep, the tool keeps whichever came first in the file.

Before your practice

Two files this time. Work out what each one holds, find the days they share, run the checks, and write the profile.

Sources

McGregor, *Practical Python Data Wrangling and Data Quality* — Ch. 3, fit and integrity; Ch. 6, assessing data quality.

McKinney, *Python for Data Analysis*, 3rd ed. — Ch. 5, pandas basics.

Quartz, *The Quartz Guide to Bad Data* — github.com/Quartz/bad-data-guide.

Questions