Statistical Tools and Data Exploration for Digitalization

Lab 2 · Exploratory data analysis: finding and fixing bad data



Szymon Dziewic · AGH University of Krakow

Same samples, different file

ORE_SAMPLES_CLEAN.CSV · LAST WEEK

53.85 %	3.65	54.28 %
Mean Fe	Std. dev.	Median

ORE_SAMPLES_MESSY.CSV · TODAY

59.77 %	52.84	54.30 %
Mean Fe	Std. dev.	Median

Three typos out of 2 440 cells (a moved decimal point: 58.45 → 584.5) shifted the mean by 6 points of iron and made the SD 14× bigger. **The median did not care.**

Why it matters in a mine: the average grade of a stockpile is what the buyer pays for. Every statistic in Labs 6-9 is only as good as the file it is computed on.

01

Why clean data?

where errors come from and why they matter

Preparation is most of the work

Business question	Data	**Preparation**	Exploration	Model	Evaluation	Deployment

TODAY

≈ 60 %

of the effort in a data-mining project goes into cleaning and preprocessing (Han & Kamber)



Garbage in, garbage out

Statistica never refuses dirty data. A mean or a regression on a broken file runs perfectly - and gives you a **precise, confident, wrong** answer. The software checks the syntax - only you can check the meaning.

The recipe: today is steps 3 and 5

1	Identify numerical and categorical variables	Lab 1
2	**Calculate basic statistics**	Lab 1 · today as a detector
3	**Check missing data and errors**	**TODAY**
4	**Create basic visualisations**	Lab 1 · today as a detector
5	**Check the outliers**	**TODAY**
6	Check normality of numerical variables	Lab 7
7	Check dependencies between variables	Labs 8–9



“Act if you need!”

Each step is a check. When a check fails - a typo, an impossible value, a duplicate - you fix it before the next step.

A normality test (Lab 7) on a file with Fe = 584.5 % is meaningless.

Where do errors come from in a mine lab?



Typing by hand

Lab sheet → Excel. A decimal point moves: 58.45 becomes 584.5.



Many people, many habits

“Mine A”, “MA”, “mine_a” - three spellings, one mine.



Codes for ‘no value’

Instruments and systems write −1, 9999 or −9999 instead of leaving a gap.



Export and locale

“n/a” typed in a number column; 53,85 vs 53.85 on Polish Windows.



Merging files

The same core sent twice by two shifts → duplicated rows.

All five are hidden in today's file. Your job: find them before they find your report.

02

Four kinds of dirty data

missing · incorrect · duplicated · outliers - and what to do with each

A map of data-quality problems

Missing

The value is unknown

empty cell · “n/a” · a code like −1

Incorrect

The value is impossible or inconsistent

Fe = 584.5 % · depth −1 m · “Mine A” vs “Mine_A”

Duplicated

The same record appears twice

S030 twice · same core, two IDs

Outlier

Far from the rest - maybe real

moisture 15.27 % after heavy rain?

ERRORS → fix, or set to missing (MD)

INVESTIGATE → decide

Basic statistics are your first detector

Variable	Valid N	Mean	Median	Min	Max
fe_pct	234	59.77	54.30	44.86	584.50
depth_m	244	249.13	203.55	−1.0	9999.0
moisture_pct	238	5.47	5.19	−10.57	15.27
sio2_pct	244	9.32	9.26	4.24	14.11
recovery_pct	244	61.75	62.21	43.36	77.39

ore_samples_messy.csv, 244 rows, before cleaning

READ IT LIKE A DETECTIVE

N = 244	we expect 240 cores → duplicates?
Fe N = 234	10 values missing
Fe max 584.5 %	no rock is 584 % iron
depth −1 / 9999	codes, not depths
moisture −10.57	water can't be negative

Watch out: check min and max of every numerical variable first - impossible values show up there before anywhere else. Then compare N with what you expect.

Domain knowledge sets the limits

Variable	Possible values	Why
fe_pct	0 – 72.4 %	Pure magnetite Fe_3O_4 = 72.4 % Fe, pure hematite Fe_2O_3 = 69.9 % Fe. No ore is richer.
sio2_pct	0 – 100 %	A share of the rock; Fe % + SiO_2 % must stay below 100 %.
moisture_pct	0 – ~20 %	Water content can't be negative; > 20 % is mud.
depth_m	> 0 and ≤ mine depth	Our deepest sample is ~400 m. −1 or 9999 is a code.
fracture_count	0, 1, 2, …	A count: whole numbers only. 2.5 cracks don't exist.
recovery_pct	0 – 100 %	A share of the iron recovered. 0 % usually means ‘not measured’.

Write the limits down before you look at the data. Then every value outside them is an error by definition - no arguing with yourself later.

Missing data: why is it missing? - three main mechanisms

MCAR

Completely at random

A sample bag was dropped on the way to the lab.

Harmless - you just have fewer data.

MAR

At random, given other data

The night-shift lab skips the moisture test. Missingness depends on shift - which we know.

Fixable: compare or model by shift.

MNAR

Not at random

Poor-looking cores are not sent for Fe assay. The missing values ARE the low grades.

Dangerous: deleting makes the deposit look richer.

Watch out: you cannot see from the numbers why they are missing. Ask how the data were collected - the lab, the shift, the system.

Missing data: what can you do?

Option	How	Good	Bad
Delete the case (casewise)	drop the whole row	simple, honest	loses every other value in that row
Pairwise deletion	each statistic uses what it can	keeps more data	different N for each result
Constant	e.g. put 0	easy	almost always wrong for measurements
Mean / mode	fill with the average / most common	keeps N	shrinks SD - fake certainty
Random from distribution	draw a plausible value (hot deck)	keeps the spread	different result every run
Predicted value	regression / MICE from other variables	best guess	complex - still a guess

Rule for this course: leave missing values as MD unless you have a reason to fill them - and document whatever you do in your cleaning log.

Worked example: our 10 missing Fe values

53.83 · 3.66

Leave as MD · mean · SD, N = 230

53.83 · 3.58

Fill with the mean · N = 240

53.85 · 3.65

The truth (last week's clean file)

Mean fill looks perfect… the mean can't change - but ten identical values at the centre make the SD smaller. Later tests will look more certain than the data allow.

Casewise deletion is greedy. 21 missing cells sit in 21 different rows: drop every incomplete row and only **219 of 240** samples survive. 21 cells cost you 9 % of the data.

Watch out: with 4 % missing, any sensible choice gives almost the same mean. With 30 % missing, the choice is the result.

Missing data in Statistica

Empty cell

→ MD automatically

Shown as an empty cell. MD code in Variable specs: −9999.

Text in a number column

→ the whole column becomes Text

“n/a”, “lost”, “64.5%”: no statistics for that variable until you remove the text and set Type: Double.

MD deletion option

→ in analysis dialogs

Casewise · Pairwise · Mean substitution. Choose deliberately and write it down.

Watch out: a cell containing −9999 is ‘missing’ in Statistica but a real number in Excel or Python - one such cell drags the mean Fe down by about 40 points there.

Incorrect data 1: one mine, many names

mine	Count
Mine_A	77
Mine_B	77
Mine_C	77
MC	3
Mine A	2
mine_b	2
MA	1
mine_a	1
Mine B	1
MB	1
Mine C	1
mine_c	1

12 ≠ 3

categories for three mines. Every comparison ‘by mine’ would now compare 12 groups.

Software compares text letter by letter. “Mine A”, “Mine_A” and “mine_a” are three different mines for Statistica, Excel and Python.

FIX IT

1 **Pick one official spelling:** Mine_A / Mine_B / Mine_C.

2 **Recode every variant:** Data › Recode, or Edit › Replace on the column.

3 **Run the frequency table again:** only 3 rows may remain.

Watch out: rock_type can have the same disease: Hematite / hematite, Mixed / MIXED. Check every text variable, not just the first one.

Incorrect data 2: numbers that mean ‘no value’

−1	0	999	9999	−9999
depth, counts	recovery, grades	counts	depth, IDs	Statistica's own MD code

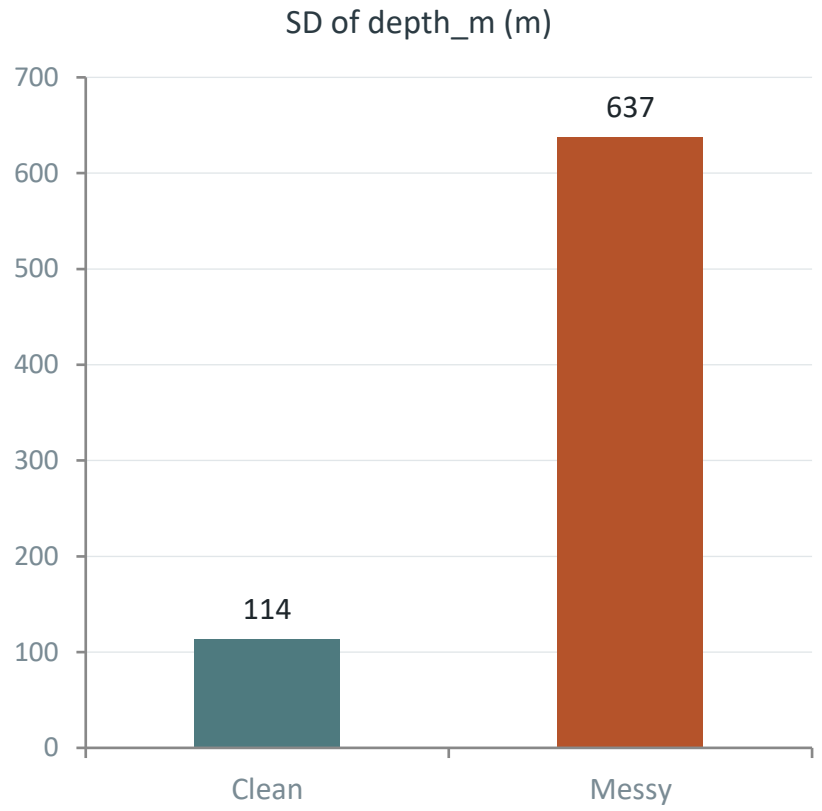
Typical sentinel codes - they look like numbers, they mean “nothing”.

210.5 → 249.1 m

mean depth: clean file → messy file

113.7 → 637.2 m

standard deviation - 5.6× larger because of one 9999



Watch out: replace a code with MD - never with a guessed value. You don't know the real depth of S222.

Fix it or set it missing? The ×10 typo

THREE CELLS IN FE_PCT

S077 **584.5**	S105 **484.9**	S107 **487.9**

“Be careful with corrections!”
Both choices are fine if you write them down. A silent change is never fine.

A · Divide by 10

584.5 → 58.45. Assumes the decimal point moved one place. Plausible values, but a guess.

Fe: N = 230 · mean 53.83 · SD 3.66

B · Set to MD

We know the value is wrong; we don't know the right one. Lose 3 values, invent nothing.

Fe: N = 227 · mean 53.85 · SD 3.64

Watch out: negative moisture (S025 −5.30, S194 −10.57) is the same dilemma: absolute value or MD? Decide, then document.

Duplicates: three kinds

Exact copy

Same ID, same values in every column.

ACTION

Delete one copy.

HOW TO FIND

Sort by sample_id → twins sit next to each other.

Same ID, different values

The core was assayed twice (re-assay).

ACTION

Keep one or average - and document which.

HOW TO FIND

Sort by sample_id; rows differ in a few columns.

Hidden duplicate

New ID, but every measurement identical.

ACTION

Delete the copy.

HOW TO FIND

Sort by several numeric columns; IDs beyond the expected range.

The quickest check of all: does N match the number of samples you expect? 244 rows for 240 cores = trouble.

Outliers: unusual is not the same as wrong

An error

Fe = 584.5 %

Impossible → fix or MD (slides before).

Real but rare

Moisture 15.27 % (S016)

Possible: a very wet core after heavy rain → keep it, report it.

A different population

A sample from another ore body

Real, but not ‘our’ material → analyse separately.

Outlier ≠ error. Never delete a point only because it is far from the others - the wettest or richest sample may be the most important one in the file.

Finding outliers 1: the z-score rule

THE RULE

$$z = (x - \bar{x}) / s$$

|z| > 3 → suspicious. For bell-shaped data that happens to about 3 in 1 000 values.

WORKED EXAMPLE · CLEANED FILE

S016 moisture (15.27 − 5.56) / 2.02 = **4.8**

S043 Fe (66.07 − 53.83) / 3.66 = **3.3**

Masking: outliers hide other outliers

On the raw file the three typos blow the SD up to 52.84. Then S043 gets z = (66.07 − 59.77) / 52.84 = **0.12** - invisible. The rule uses the mean and SD, and both are pulled by the very outliers you are looking for.

Watch out: the z-rule assumes roughly normal data. Moisture is right-skewed (skewness 1.04) - use the IQR rule there.

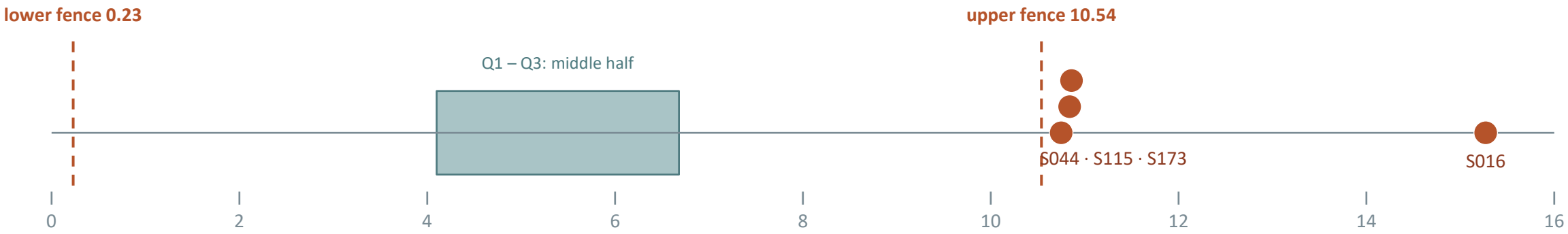
Finding outliers 2: the IQR rule

THE RULE

$$IQR = Q3 - Q1$$
$$\text{outlier if } x < Q1 - 1.5 \cdot IQR$$
$$\text{or } x > Q3 + 1.5 \cdot IQR$$

MOISTURE, CLEANED FILE

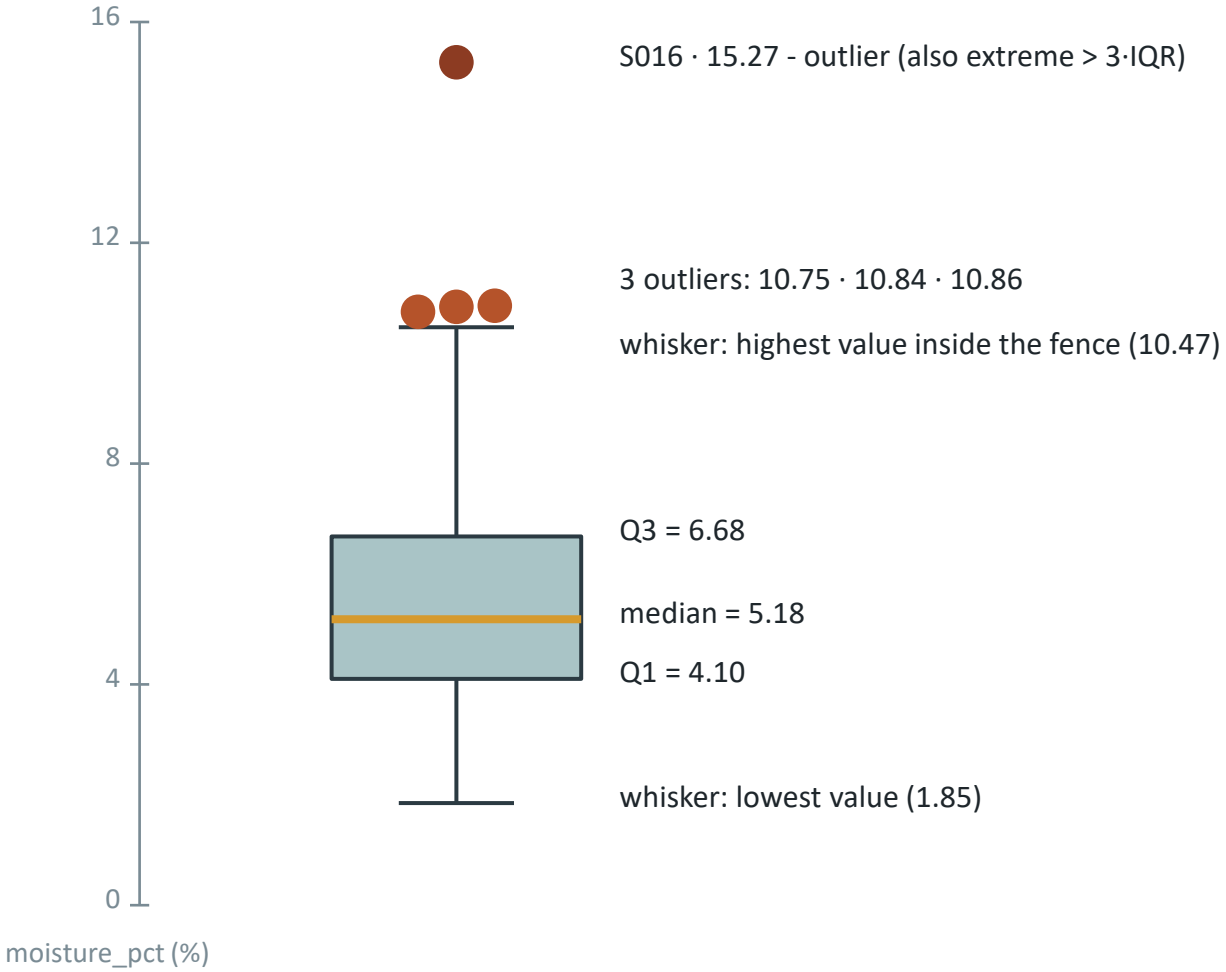
Q1	Q3	IQR	Lower fence	Upper fence
4.10	6.68	2.58	0.23	**10.54**



4 outliers by IQR, 1 by z-score. Quartiles ignore how extreme the extremes are, so the IQR rule can't be masked - and it doesn't assume a bell curve.

Statistica, Excel and Python compute quartiles slightly differently - fences may differ in the 2nd decimal.

The box plot draws the IQR rule for you



IN STATISTICA

Graphs › Box Plot

Middle point: Median

Box: Quartiles (25 %–75 %)

Whisker: Non-outlier range

Outlier coefficient: 1.5

Box plot by group (moisture by mine) finds outliers within each mine - a value can be normal in Mine C and unusual in Mine A.

Found an outlier. Now what?

1	**Is the value physically possible?**	No → it's an error: fix it (if you know how) or set MD.
2	**Can you check the source?**	Lab sheet, ask the shift. Confirmed → keep. Wrong → fix.
3	**Does it belong to the same population?**	Different sample body, different instrument → analyse separately.
4	**Still unsure?**	Keep it, report it, and run the analysis with and without it.

Example: S016, moisture 15.27 % → possible · can't re-check today · same ore → keep it and report it.

The cleaning log: your proof of honesty

#	sample_id	variable	problem	what I did	why
1	S017	mine	“Mine A”	recoded to Mine_A	same mine, other spelling
2	S077	fe_pct	584.5 %	divided by 10 → 58.45	decimal point moved; < 72.4 % limit
3	S222	depth_m	9999	set to MD	sentinel code; true depth unknown
…					

Never overwrite the raw file

Save the cleaned data as a new file.

One line per change

Which sample, what, why.

End with the final N

“244 rows → 240 cases.”

In your project report this table becomes the section “Data preparation”.

03

Live demo

cleaning ore_samples_messy in Statistica - five steps

Step 0 · Open the right file

ore_samples_messy.csv

English Windows

Separator: comma · decimal: dot **53.85**

ore_samples_messy.txt

Polish Windows

Separator: TAB · decimal: comma **53,85**

File › Open › ore_samples_messy › Delimited › ☑ names from first row

Watch out: numbers left-aligned after import = read as text. Close, and open the other version of the file. (moisture_pct is Text on purpose - next step.)

Step 1 · Fix the variable types

WHAT YOU SEE

sample_id	moisture_pct
S062	4.21
S063	n/a
S064	2.0
S065	5.21

One word in 244 cells turns the whole column into Text: no mean, no histogram, nothing.

FIX

1. **Select the column**
   click the moisture_pct header
2. **Replace “n/a” with nothing**
   Edit › Replace › Replace All
3. **Change the type**
   double-click header › Variable specs › Type: Double
4. **Check**
   empty cells = MD; statistics work again

Watch out: check the type of EVERY numerical column. In the advanced file other columns have the same disease.

Step 2 · Find odd labels, then recode

Statistics › Basic Statistics › Frequency tables › mine, rock_type, shift › Summary

OPTION A · RECODE

Data › Recode
New value: Mine_A
Condition: v2 = "MA" or v2 = "Mine A" or v2 = "mine_a"

Good for many variants at once; repeatable.

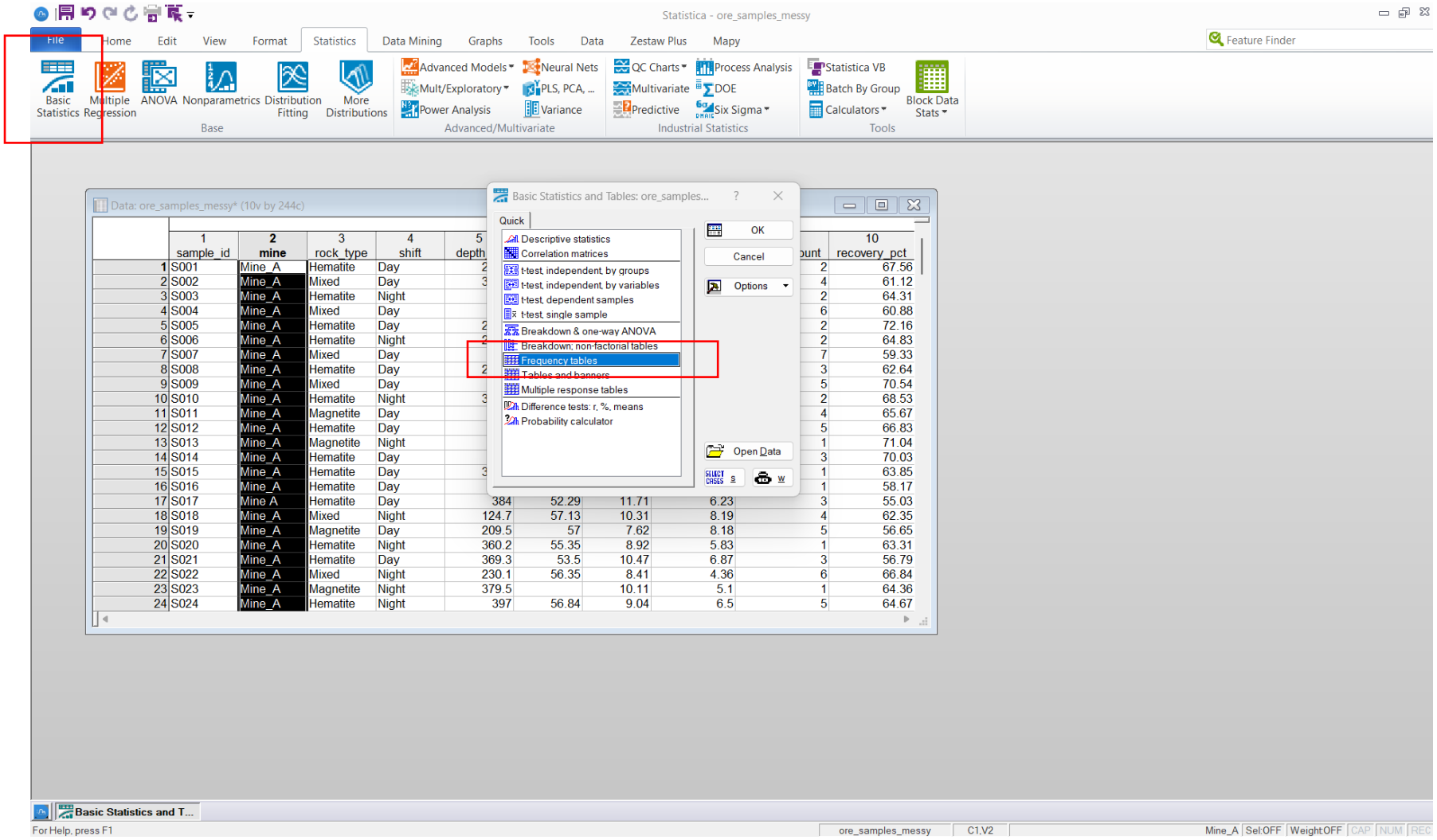
OPTION B · REPLACE

Select the mine column
Edit › Replace
Find: MA Replace with: Mine_A
☑ Match entire cell!

Fast for one or two variants.

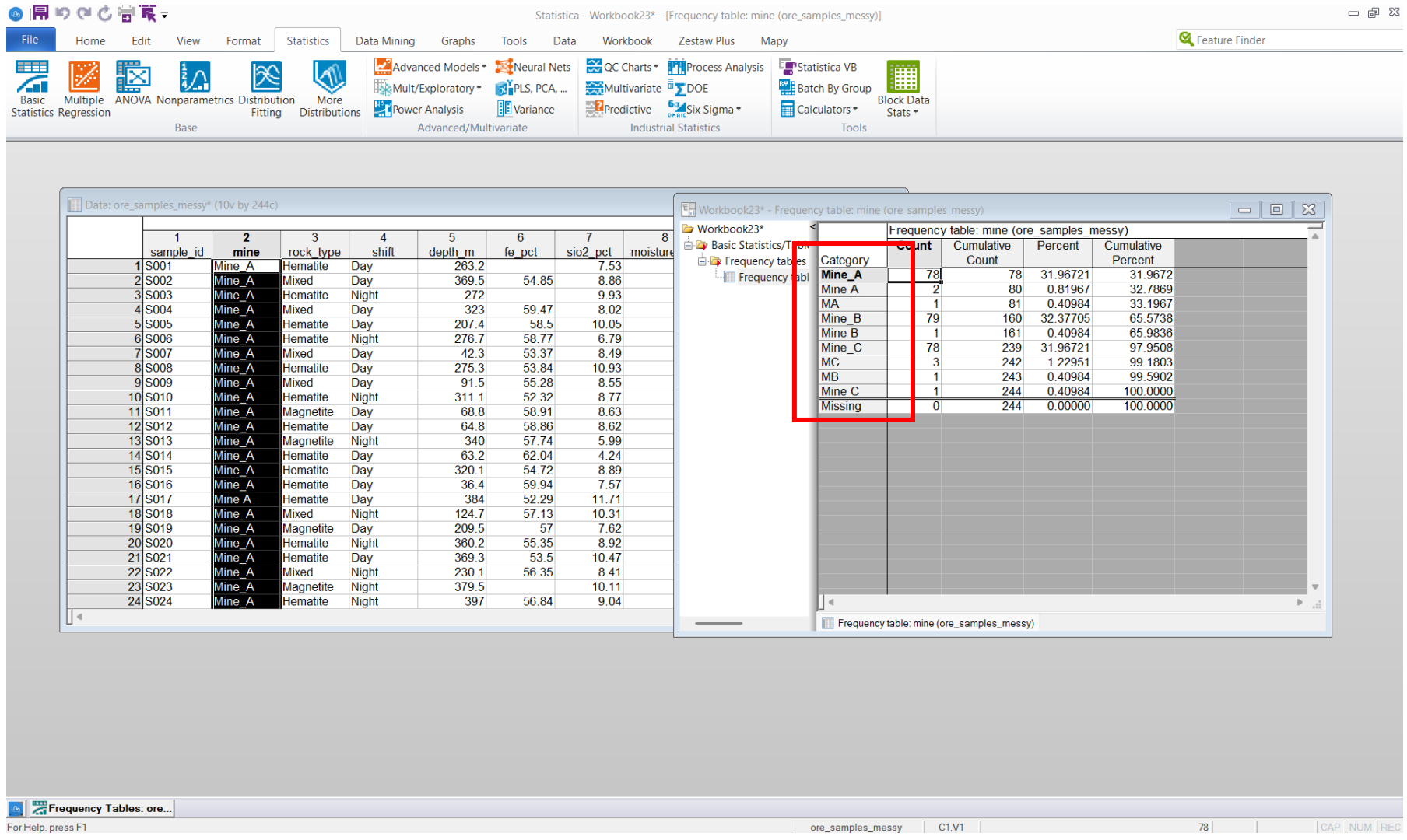
Watch out: without “match entire cell”, replacing MA would also change “MAgnetite”. Always run the frequency table again after recoding.

Step 2 · Find odd labels, then recode

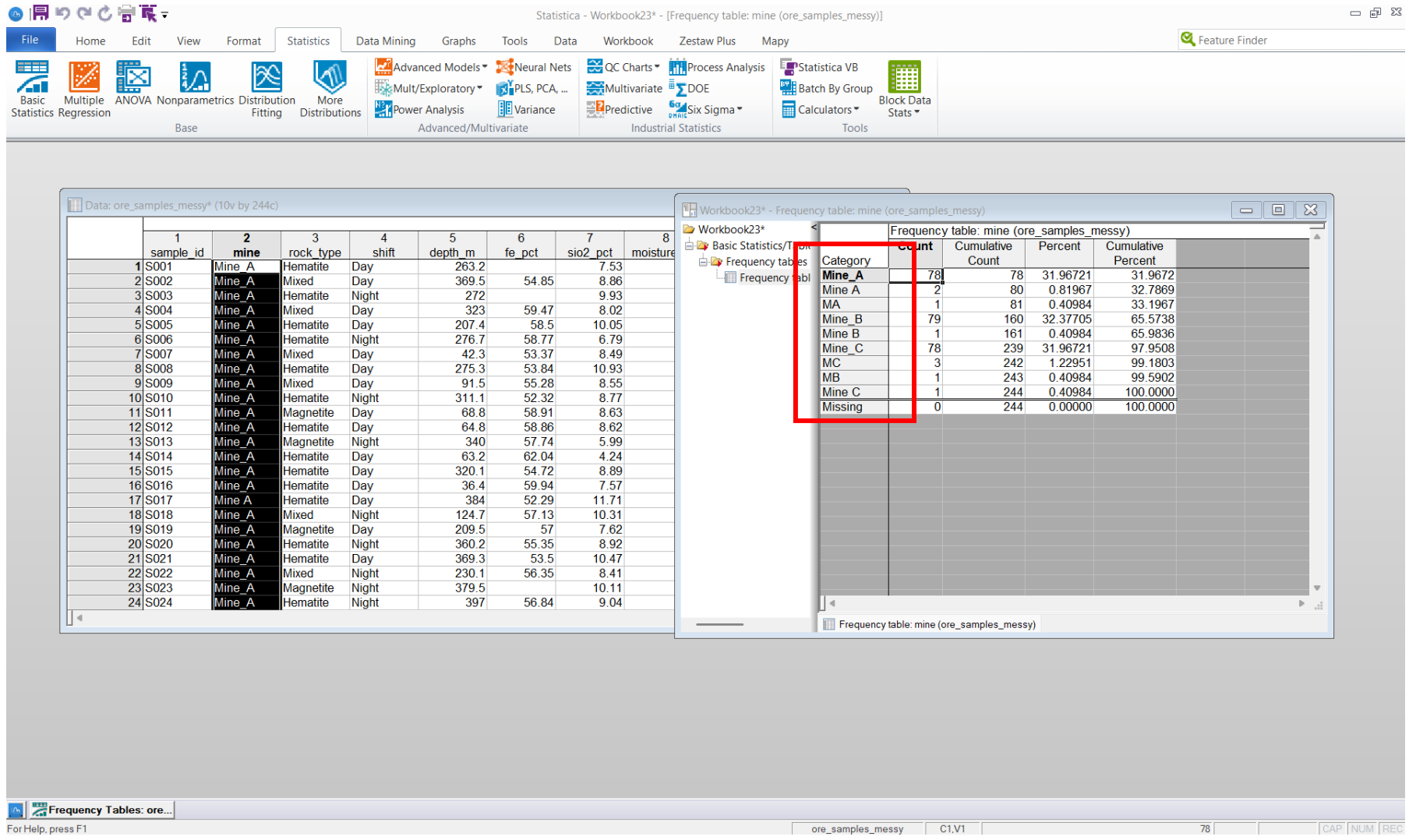


	1 sample_id	2 mine	3 rock_type	4 shift	5 depth					10 recovery_pct
1	S001	Mine_A	Hematite	Day	2				2	67.56
2	S002	Mine_A	Mixed	Day	3				4	61.12
3	S003	Mine_A	Hematite	Night					2	64.31
4	S004	Mine_A	Mixed	Day					6	60.88
5	S005	Mine_A	Hematite	Day	2				2	72.16
6	S006	Mine_A	Hematite	Night	2				2	64.83
7	S007	Mine_A	Mixed	Day					7	59.33
8	S008	Mine_A	Hematite	Day	2				3	62.64
9	S009	Mine_A	Mixed	Day					5	70.54
10	S010	Mine_A	Hematite	Night	3				2	68.53
11	S011	Mine_A	Magnetite	Day					4	65.67
12	S012	Mine_A	Hematite	Day					5	66.83
13	S013	Mine_A	Magnetite	Night					1	71.04
14	S014	Mine_A	Hematite	Day					3	70.03
15	S015	Mine_A	Hematite	Day	3				1	63.85
16	S016	Mine_A	Hematite	Day					1	58.17
17	S017	Mine A	Hematite	Day	384	52.29	11.71	6.23	3	55.03
18	S018	Mine_A	Mixed	Night	124.7	57.13	10.31	8.19	4	62.35
19	S019	Mine_A	Magnetite	Day	209.5	57	7.62	8.18	5	56.65
20	S020	Mine_A	Hematite	Night	360.2	55.35	8.92	5.83	1	63.31
21	S021	Mine_A	Hematite	Day	369.3	53.5	10.47	6.87	3	56.79
22	S022	Mine_A	Mixed	Night	230.1	56.35	8.41	4.36	6	66.84
23	S023	Mine_A	Magnetite	Night	379.5		10.11	5.1	1	64.36
24	S024	Mine_A	Hematite	Night	397	56.84	9.04	6.5	5	64.67

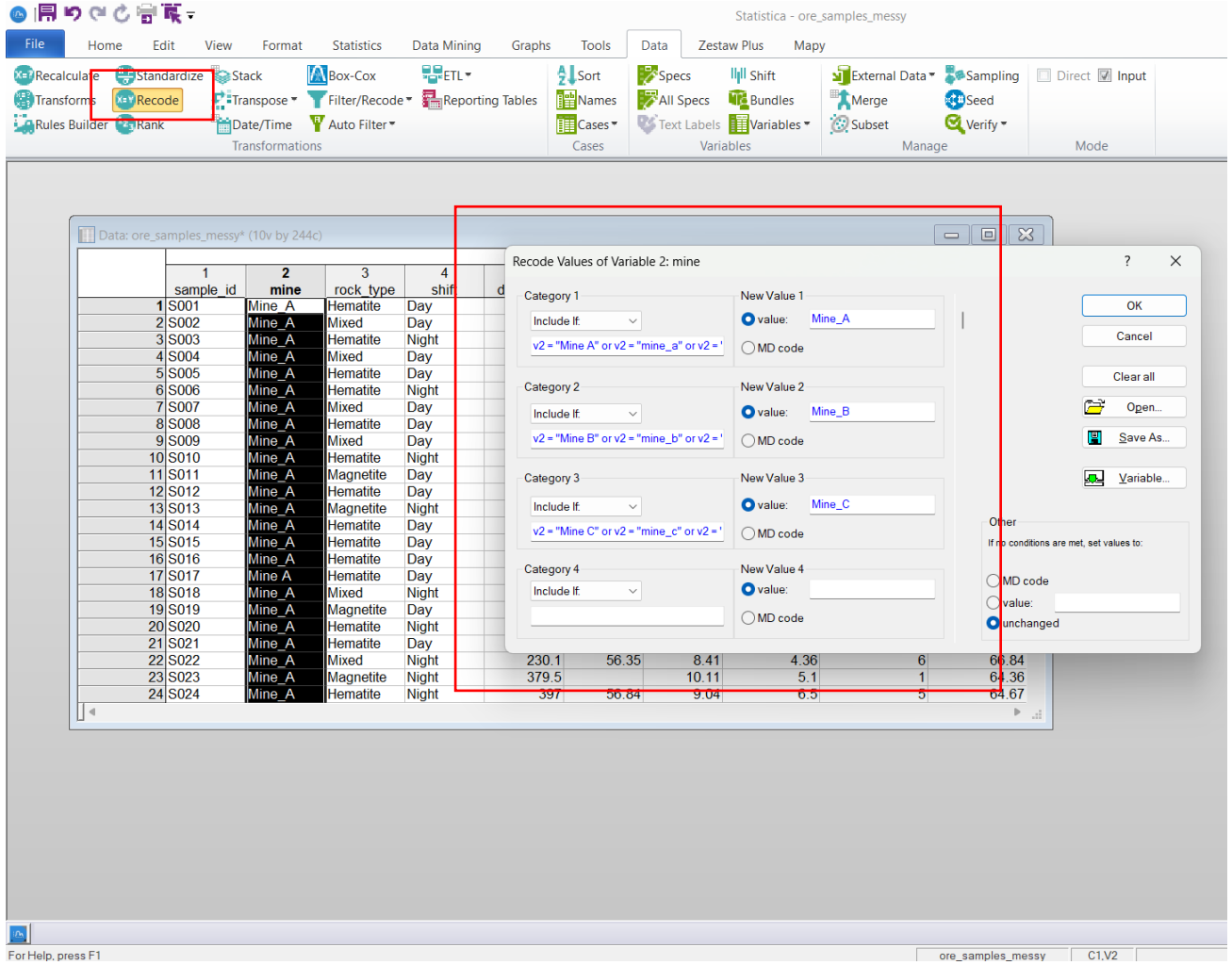
Step 2 · Find odd labels, then recode



Step 2 · Find odd labels, then recode



Step 2 · Find odd labels, then recode



New value: Mine_A
Include If: v2 = "MA" or v2 = "Mine A" or v2 = "mine_a"

New value: Mine_B
Include If: v2 = "MB" or v2 = "Mine B" or v2 = "mine_b"

New value: Mine_C
Include If: v2 = "MC" or v2 = "Mine C" or v2 = "mine_c"

Step 3 · Find impossible values, set MD

Statistics › Basic Statistics › Descriptive statistics › Advanced: Min & max › Summary

Find the row

Data › Sort by the suspicious variable. Descending → the 584.5 is on top; ascending → the −1.

Set to MD

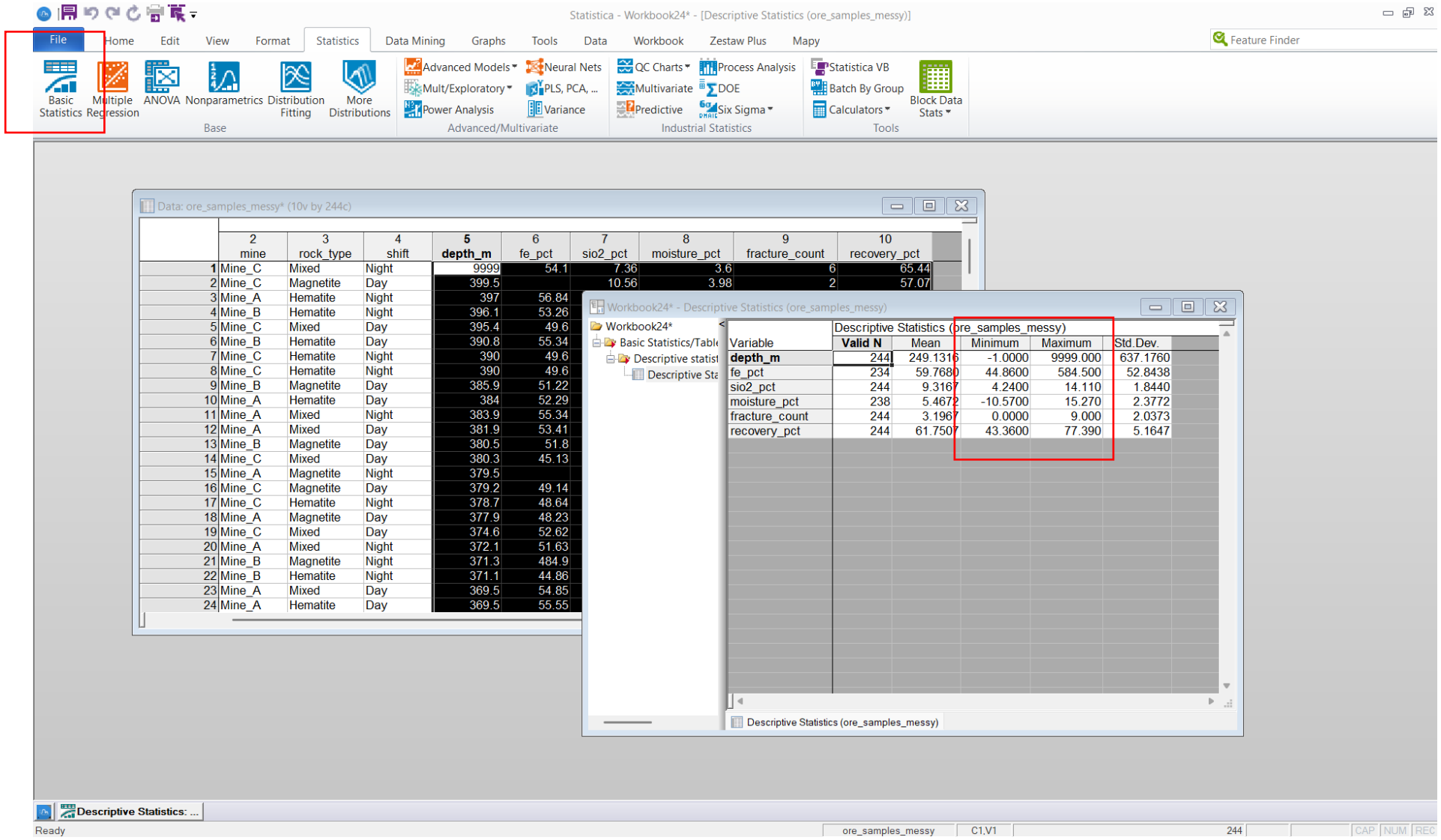
Select the cell → Delete. The empty cell becomes MD.

…or correct it

Only if you are sure: 584.5 → 58.45. Write it in the log.

Watch out: sorting changes the row order - that's fine, sample_id keeps each sample's identity. Sort by sample_id to get the original order back.

Step 3 · Find impossible values, set MD

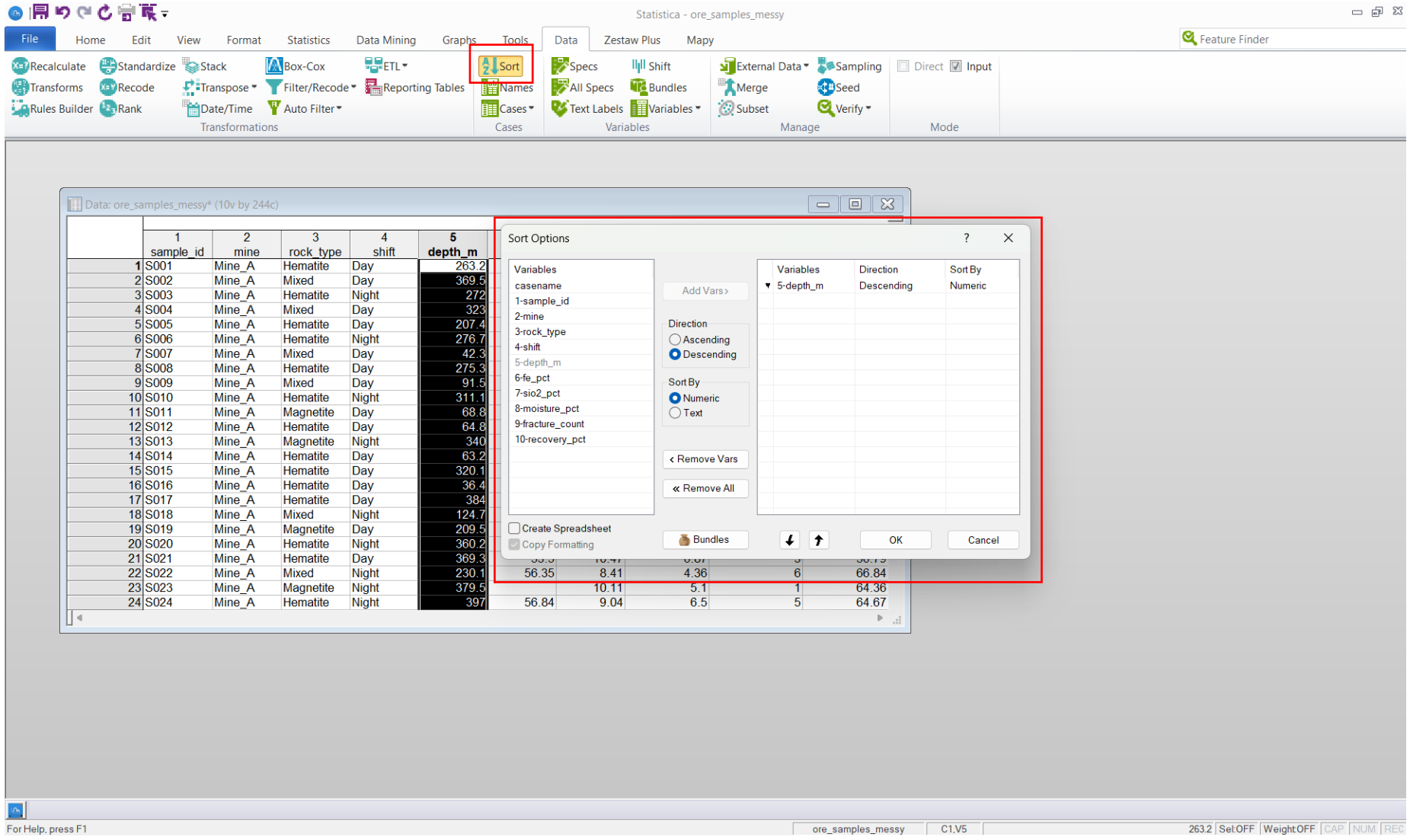


	2 mine	3 rock_type	4 shift	5 depth_m	6 fe_pct	7 sio2_pct	8 moisture_pct	9 fracture_count	10 recovery_pct
1	Mine_C	Mixed	Night	9999	54.1	7.36	3.6	6	65.44
2	Mine_C	Magnetite	Day	399.5		10.56	3.98	2	57.07
3	Mine_A	Hematite	Night	397	56.84				
4	Mine_B	Hematite	Night	396.1	53.26				
5	Mine_C	Mixed	Day	395.4	49.6				
6	Mine_B	Hematite	Day	390.8	55.34				
7	Mine_C	Hematite	Night	390	49.6				
8	Mine_C	Hematite	Night	390	49.6				
9	Mine_B	Magnetite	Day	385.9	51.22				
10	Mine_A	Hematite	Day	384	52.29				
11	Mine_A	Mixed	Night	383.9	55.34				
12	Mine_A	Mixed	Day	381.9	53.41				
13	Mine_B	Magnetite	Day	380.5	51.8				
14	Mine_C	Mixed	Day	380.3	45.13				
15	Mine_A	Magnetite	Night	379.5					
16	Mine_C	Magnetite	Day	379.2	49.14				
17	Mine_C	Hematite	Night	378.7	48.64				
18	Mine_A	Magnetite	Day	377.9	48.23				
19	Mine_C	Mixed	Day	374.6	52.62				
20	Mine_A	Mixed	Night	372.1	51.63				
21	Mine_B	Magnetite	Night	371.3	484.9				
22	Mine_B	Hematite	Night	371.1	44.86				
23	Mine_A	Mixed	Day	369.5	54.85				
24	Mine_A	Hematite	Day	369.5	55.55				

Descriptive Statistics (ore_samples_messy)

Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
depth_m	244	249.1316	-1.0000	9999.000	637.1760
fe_pct	234	59.7680	44.8600	584.500	52.8438
sio2_pct	244	9.3167	4.2400	14.110	1.8440
moisture_pct	238	5.4672	-10.5700	15.270	2.3772
fracture_count	244	3.1967	0.0000	9.000	2.0373
recovery_pct	244	61.7507	43.3600	77.390	5.1647

Step 3 · Find impossible values, set MD



Step 4 · Duplicates: sort and compare

Data › Sort › sample_id ↑ › OK

AFTER SORTING

sample_id	mine	depth_m	fe_pct	sio2_pct	…
S029	Mine_A	235.3	53.35	10.19	…
S030	Mine_A	57.1	54.43	10.44	…
S030	Mine_A	57.1	54.43	10.44	…
S031	Mine_A	359.3	52.34	9.92	…



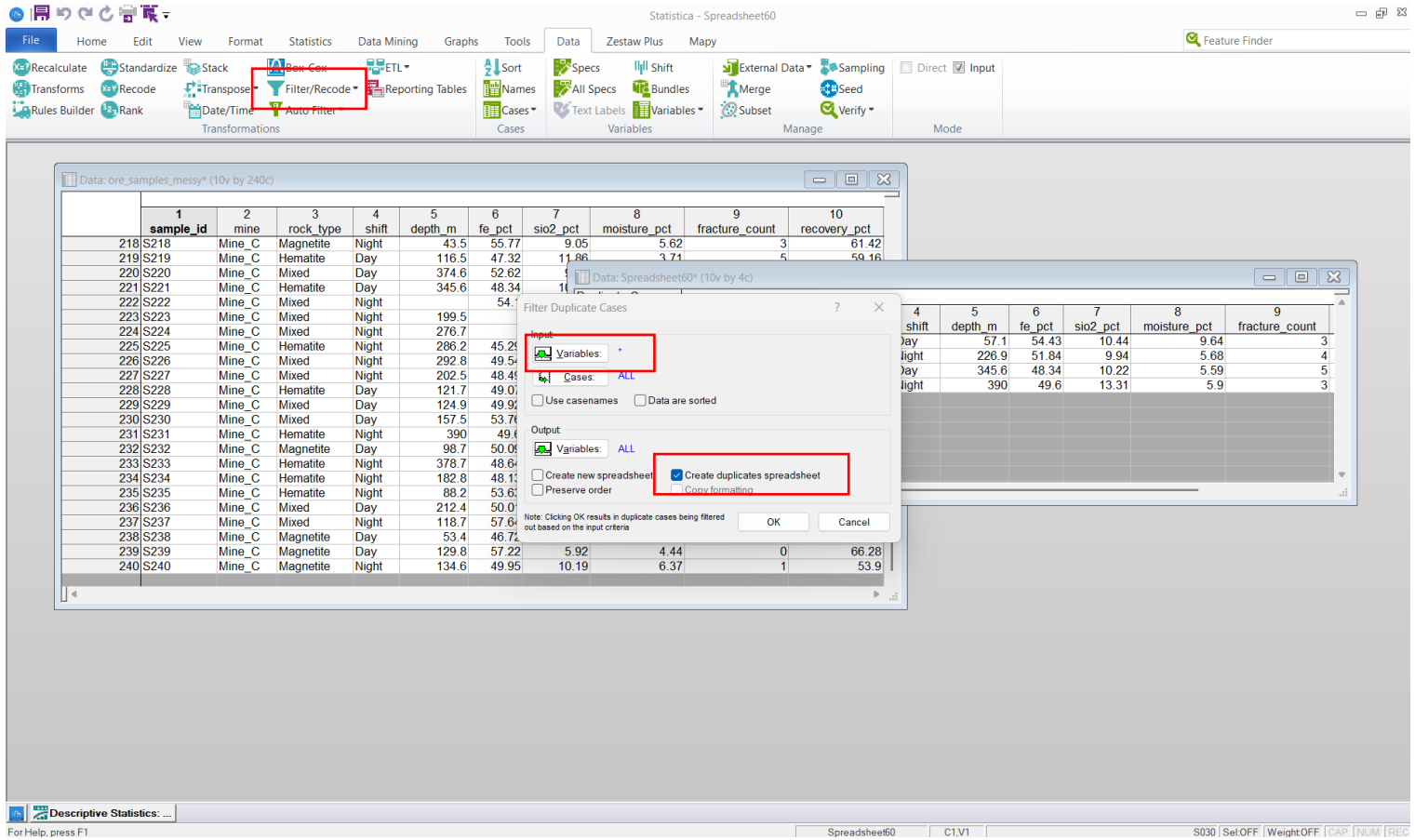
Delete the copy

Click the row header of one copy → Delete
(or Edit › Delete › Cases).
Then check the number of cases: 240.

Watch out: only delete a row if EVERY column is identical. Same ID with different values is a decision, not a deletion.

Step 4 · Duplicates: sort and compare

Data › **Filter/Recode** › **ALL Variables** › **Duplicates spreadsheet** › **OK**



Step 5 · Box plot for outliers, then save

Graphs › Box Plot › moisture_pct › Non-outlier range · 1.5 › OK

4 dots

moisture outliers above 10.54 % - all physically possible → keep and report

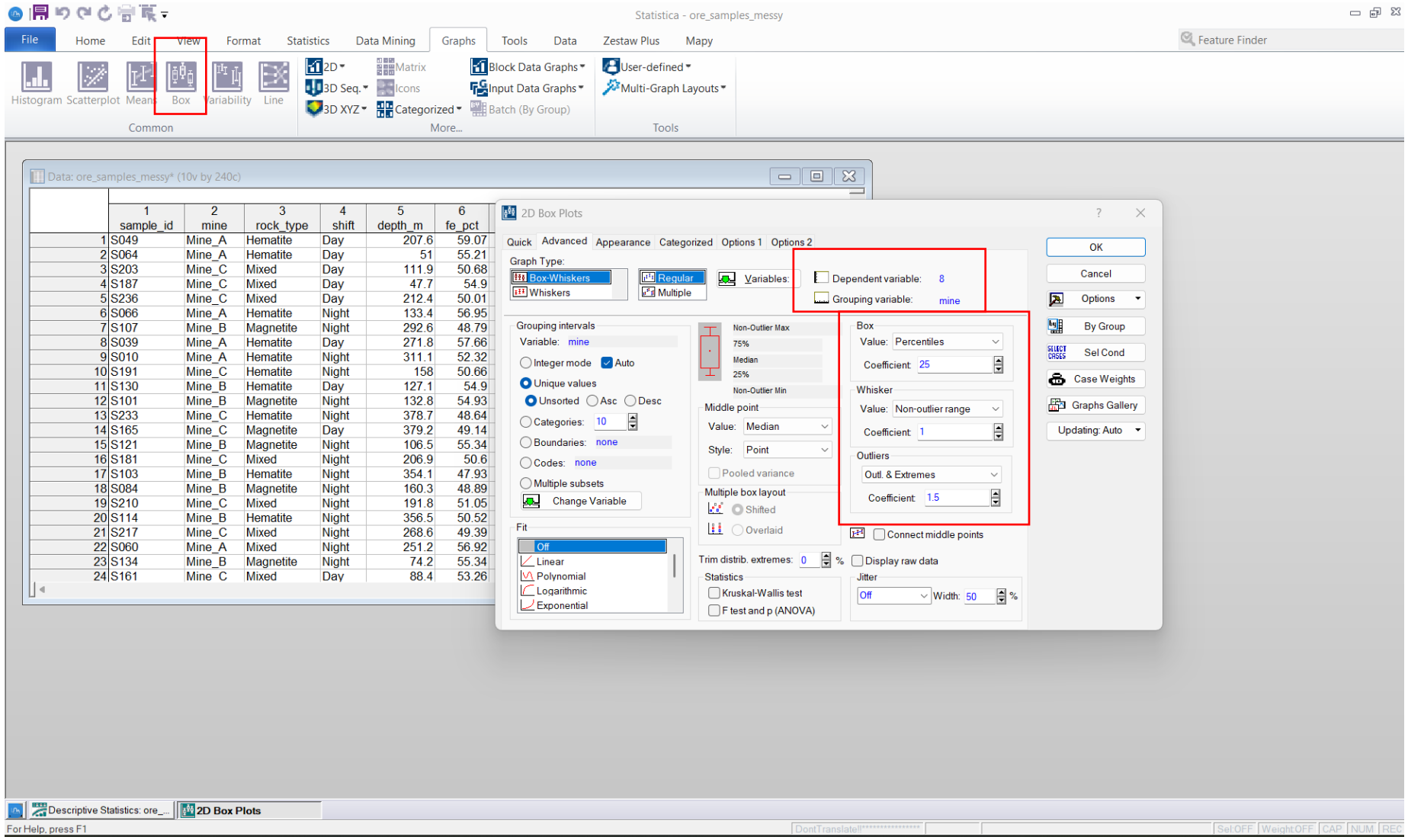
Save two files

`ore_clean_<names>.sta` - the cleaned data (new file - the raw file stays untouched)

`Lab2_<names>.stw` - the workbook: every table and graph

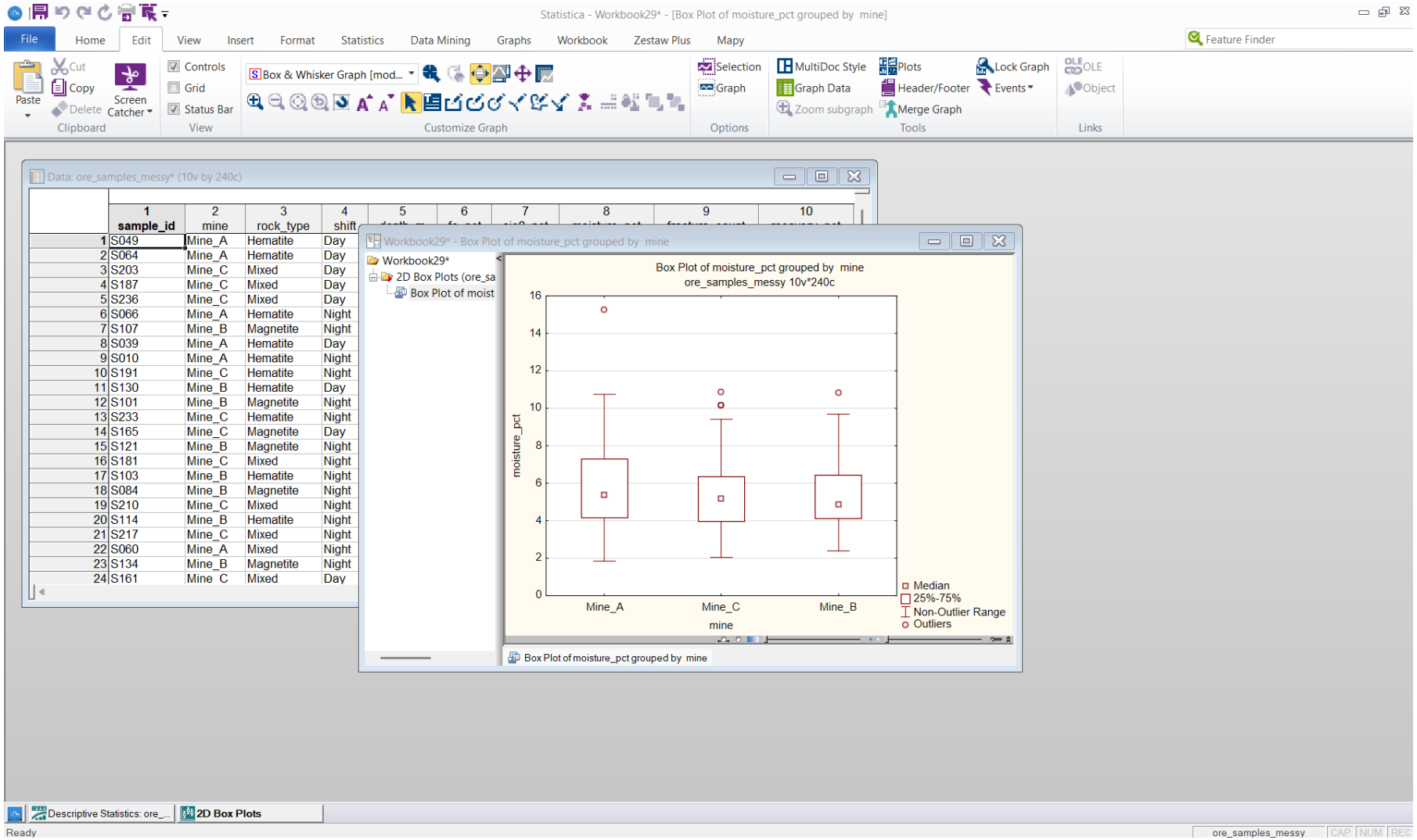
Demo done. I showed you two problems of each kind. There are more than 35 problem cells in the file - now you find all of them.

Step 5 · Box plot for outliers, then save



	1 sample_id	2 mine	3 rock_type	4 shift	5 depth_m	6 fe_pct
1	S049	Mine_A	Hematite	Day	207.6	59.07
2	S064	Mine_A	Hematite	Day	51	55.21
3	S203	Mine_C	Mixed	Day	111.9	50.68
4	S187	Mine_C	Mixed	Day	47.7	54.9
5	S236	Mine_C	Mixed	Day	212.4	50.01
6	S066	Mine_A	Hematite	Night	133.4	56.95
7	S107	Mine_B	Magnetite	Night	292.6	48.79
8	S039	Mine_A	Hematite	Day	271.8	57.66
9	S010	Mine_A	Hematite	Night	311.1	52.32
10	S191	Mine_C	Hematite	Night	158	50.66
11	S130	Mine_B	Hematite	Day	127.1	54.9
12	S101	Mine_B	Magnetite	Night	132.8	54.93
13	S233	Mine_C	Hematite	Night	378.7	48.64
14	S165	Mine_C	Magnetite	Day	379.2	49.14
15	S121	Mine_B	Magnetite	Night	106.5	55.34
16	S181	Mine_C	Mixed	Night	206.9	50.6
17	S103	Mine_B	Hematite	Night	354.1	47.93
18	S084	Mine_B	Magnetite	Night	160.3	48.89
19	S210	Mine_C	Mixed	Night	191.8	51.05
20	S114	Mine_B	Hematite	Night	356.5	50.52
21	S217	Mine_C	Mixed	Night	268.6	49.39
22	S060	Mine_A	Mixed	Night	251.2	56.92
23	S134	Mine_B	Magnetite	Night	74.2	55.34
24	S161	Mine_C	Mixed	Day	88.4	53.26

Step 5 · Box plot for outliers, then save



04

Your turn

the detective exercise · then Level 2

Detective exercise: clean ore_samples_messy

IN PAIRS · ONE LOG PER PAIR

1	**Open**	the .csv or the .txt (Polish Windows) version
2	**Detect**	frequency table for every text variable, min/max for every number, N of rows
3	**Fix**	recode labels, set impossible values and codes to MD, delete exact duplicates
4	**Log**	one line per change: sample_id · variable · problem · what I did · why
5	**Report**	final number of cases and mean fe_pct; save ore_clean_<names>.sta

Hints

More than 35 problem cells
+ some duplicated rows

At least 6 different kinds of problem

Not every unusual value is an error.

Stuck? Check this first

Symptom	Likely cause	Fix
Numbers left-aligned, statistics greyed out	Decimal comma vs dot	Open the other file version (.txt for Polish Windows)
moisture_pct has no statistics	“n/a” text in the column	Replace n/a with empty → Variable specs → Double
Type change turns numbers into MD	Decimal separator again	Wrong file version; re-import
Recode also changed “MAgnetite”	Partial-cell replace	Undo; tick “match entire cell”
Still 244 cases	Duplicates not removed	Data › Sort by sample_id, delete identical twins
Frequency table still shows 4+ mines	A spelling left over	Look for case and spaces: “mine_b”, “Mine B”
Can't find the row with 584.5	Big file	Data › Sort descending by fe_pct

Reveal: what was hidden in the file

Problem	Cells	Where	Action
Mine label variants (Mine A, mine_b, MC …)	12	mine	recode → Mine_A / B / C
Rock type case typos (hematite, MIXED …)	3	rock_type	recode
Blank Fe	10	fe_pct	leave MD
Text “n/a”	6	moisture_pct	→ MD, type Double
Decimal moved (×10)	3	fe_pct	÷10 or MD - documented
Sentinel codes −1 / 9999	3	depth_m	MD
Negative moisture	2	moisture_pct	MD (or |x|, documented)
Total problem cells	**39**		
Exact duplicate rows (S030, S106, S221, S231)	4 rows	whole row	delete → 240 cases

Real outliers - NOT errors: S016 moisture 15.27 %, S043 Fe 66.07 %. If you deleted them, put them back.

Check your numbers

240
cases after removing 4 duplicates

80 / 80 / 80
Mine_A / Mine_B / Mine_C

91 / 77 / 72
Hematite / Magnetite / Mixed

53.83 %
mean Fe, N = 230 (÷10 policy)

Variable	Valid N	Mean	Median	Min	Max
fe_pct (typos ÷10)	230	53.83	54.30	44.86	66.07
fe_pct (typos → MD)	227	53.85	54.30	44.86	66.07
moisture_pct	232	5.56	5.18	1.85	15.27
depth_m	237	210.01	203.00	21.6	399.5

Last week's clean file: mean Fe 53.85 % with N = 240. Close - but honest missing data has a price.

Level 2: ore_samples_advanced - three quick tasks

IN PAIRS · AT HOME

Task 1 · Labels

- Frequency tables for mine, rock_type and shift.
- How many different labels does each have?
- Recode them. One mine label is “?” - can you still tell which mine it is?

Report: counts per category after cleaning

Task 2 · Numbers

- Which numeric columns import as Text, and why?
- Min / max of every number: list each kind of problem.
- Which can you repair, which become MD?

Report: N and mean of fe_pct after cleaning

Task 3 · Rows & outliers

- Find all duplicates - there are three different kinds.
- How many cases remain?
- IQR rule on cleaned moisture: which samples are outliers - errors or real?

Report: final N + outlier list

Six rules for clean data

Look before you analyse

Frequency tables for text, min/max for numbers, N for rows.

Never overwrite the raw file

Clean into a new file; the original is your evidence.

Missing ≠ 0 ≠ text

An empty cell is MD; 0, −1, 9999 and “n/a” are not.

Outlier ≠ error

Impossible → fix or MD. Possible → keep and report.

Don't guess silently

Every correction goes into the cleaning log, with a reason.

Check N at the end

You know how many samples there should be.

The same cleaning in Python (optional)

```
import pandas as pd
df = pd.read_csv("ore_samples_messy.csv", na_values=["n/a"])
df = df.drop_duplicates()                          # 244 -> 240 rows
fix = {"Mine A":"Mine_A","mine_a":"Mine_A","MA":"Mine_A",
       "Mine B":"Mine_B","mine_b":"Mine_B","MB":"Mine_B",
       "Mine C":"Mine_C","mine_c":"Mine_C","MC":"Mine_C"}
df["mine"] = df["mine"].replace(fix)
df["rock_type"] = df["rock_type"].str.capitalize()
df.loc[df.fe_pct > 100, "fe_pct"] /= 10          # x10 typos
df.loc[(df.depth_m < 0) | (df.depth_m > 2000), "depth_m"] = None
df.loc[df.moisture_pct < 0, "moisture_pct"] = None
q1, q3 = df.moisture_pct.quantile([.25, .75])
print(df[df.moisture_pct > q3 + 1.5*(q3-q1)])    # outliers
print(len(df), df.fe_pct.mean())                 # 240  53.83
```



The script is the log

Every fix is a line you can read, share and rerun on next month's file. In Statistica the same role is played by your cleaning log + workbook.

na_values=["n/a"] does in one argument what took three clicks in Statistica.

Before Lab 3: install Fluxicon Disco

Download

fluxicon.com/disco → Windows, macOS Intel or macOS Apple Silicon

Install

Windows: unzip → Disco-Setup.exe. Mac: open the .dmg → drag to Applications

Register with your AGH e-mail

Use your university address → you'll be offered an academic licence → accept the terms

Test it

Open maintenance_log.csv (sent by e-mail) - if you see a process map, you're ready

WHICH LAPTOP?

Windows
✓

macOS (Intel & M-chips)
✓

Linux
✗ pair up / VM

Unlike Statistica, Disco runs on Macs.

Watch out: register with your UNIVERSITY e-mail - only an AGH address is offered the free academic licence.

Before next lab



Disco installed

and registered with your AGH e-mail (previous slide)



Check your e-mail

Level 2 answer key + maintenance_log.csv for Lab 3



Read (optional)

Larose & Larose, Discovering Knowledge in Data, ch. 2 - Data preprocessing

NEXT · LAB 3

Process discovery

From a table of samples to a log of events: 120 maintenance work orders from a mine. Who did what, when, to which machine - and how does the process really run?

Questions?

See you in Lab 3 - bring your laptop with Disco installed

sdziewic@agh.edu.pl

Online, MS Teams, Thursday 8:30-10:00

e-mail / MS Teams