Statistical Tools and Data Exploration for Digitalization

Lab 1 · Introduction to analytic software and data analysis

Szymon Dziewic · AGH University of Krakow

Why statistics in the raw-materials value chain



Material quality varies

Every sample from a deposit is different. Accept or reject a batch?

Labs 6–9



Plants must be controlled

Flotation, grinding, energy use - which setting gives better recovery, and is the difference real?

Labs 8–9



Machines leave event logs

Work orders, sensors, events. Where is time lost? Who waits for whom?

Labs 3–5

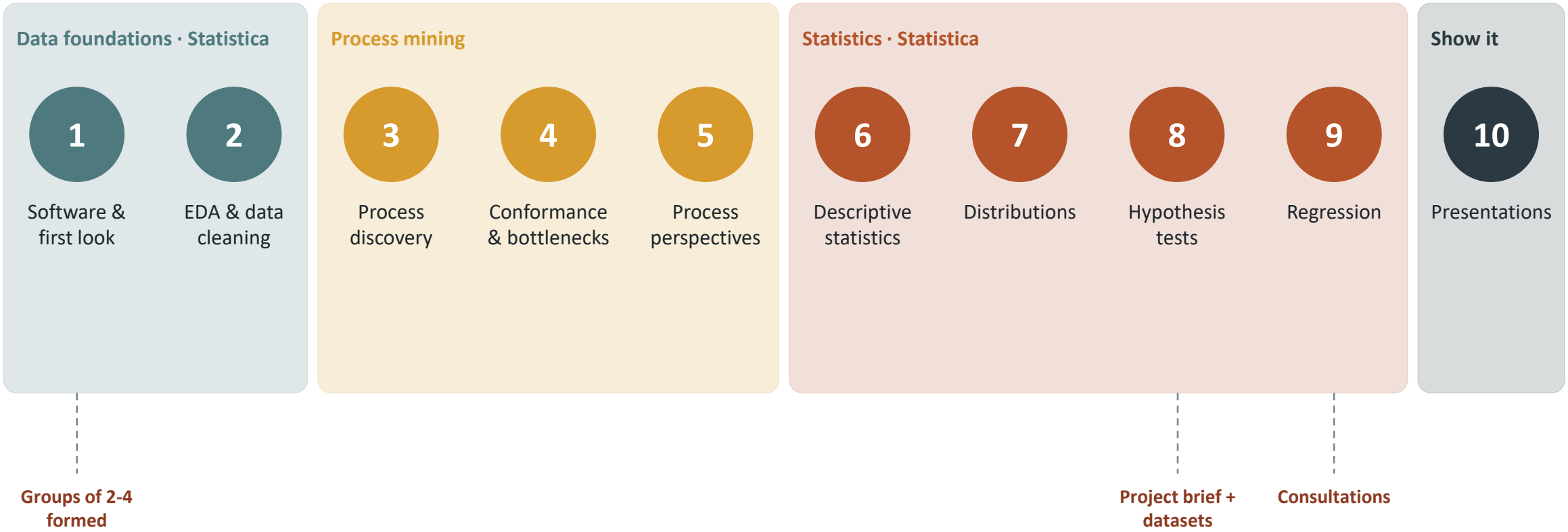
*Digitalization = sensors + logs → **more data than anyone can look at by eye.** Statistics turns measurements into decisions.*

01

How this course works

10 labs · group project · what counts for your grade

10 labs, three blocks



The group project in one slide

A

Optional

Process mining mini-report

- Data: maintenance work orders from a mine
- Discover the process, find rule-breaking cases and the bottleneck

B

Statistical analysis report

- Real public dataset – a different subset for every group
- Clean → describe → distributions → tests → regression
- Report + your Statistica workbook

★

Presentation

- 8 minutes + 4 minutes Q&A
- Every member presents some part

02

Your toolkit

What to install, and what to do if you can't

The tools we use



Statistica

Main tool · Labs 1–2, 6–9 · all reports

- Menus, no coding
- AGH licence for students
- Windows only



Python in Google Colab

Optional · runs in the browser

- Any OS, nothing to install
- Same statistics, written as code
- Good for extra graphs



Process-mining tool

Labs 3–5

- Fluxicon Disco
- Installed before Lab 3

Installing Statistica with the AGH licence

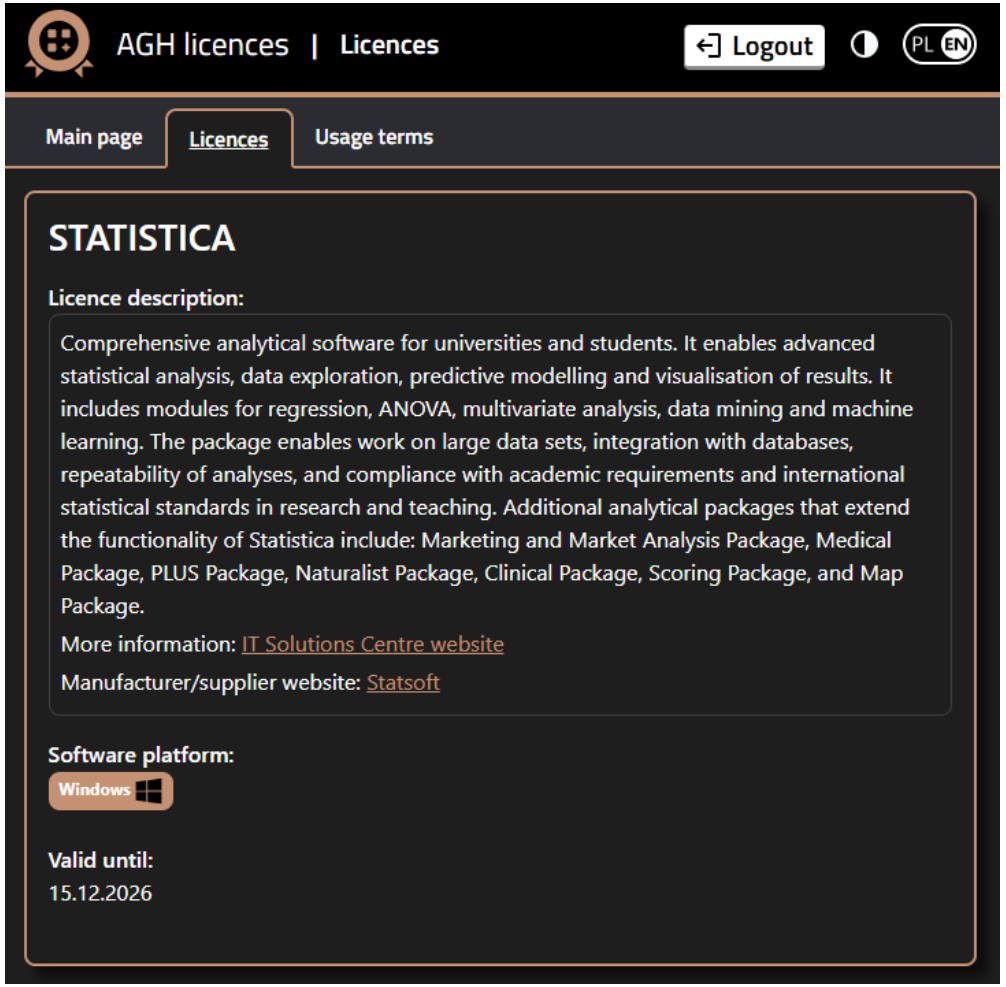
1. **Get your codes**
   AGH Panel → “Software” tab → serial number + product key

2. **Download the installer**
   AGH Cloud (log in with your AGH e-mail). Link: cri.agh.edu.pl → Software → Statistica

3. **Install as administrator**
   Unzip → right-click setup.exe → “Run as administrator”

4. **Enter codes, pick English**
   Paste serial + key. Choose the English interface

5. **Register**
   Usually happens during setup. If not, Statistica stops working after 14 days



No Windows? You still have options

RECOMMENDED

Pair up

Each group of 3 has at least one Windows laptop with Statistica. You run the analysis together.

TEST FIRST

Windows virtual machine

Parallels, VMware Fusion or UTM on a Mac. Test it this week.

FOR IN-CLASS WORK

Follow along in JASP or Colab

JASP is free for Mac/Linux, Colab runs in a browser. Same statistics, different buttons.

TO CONFIRM

University computer

03

What is data analysis for?

process, dataset structure, variable types

Why data mining?

“We are drowning in data, but starving for knowledge.”

Machines and sensors

Shearers, trucks, conveyors, mills log temperature, load, vibration - every second.

Laboratory tests

Every drill core and every plant sample gets Fe %, SiO_2 %, moisture…

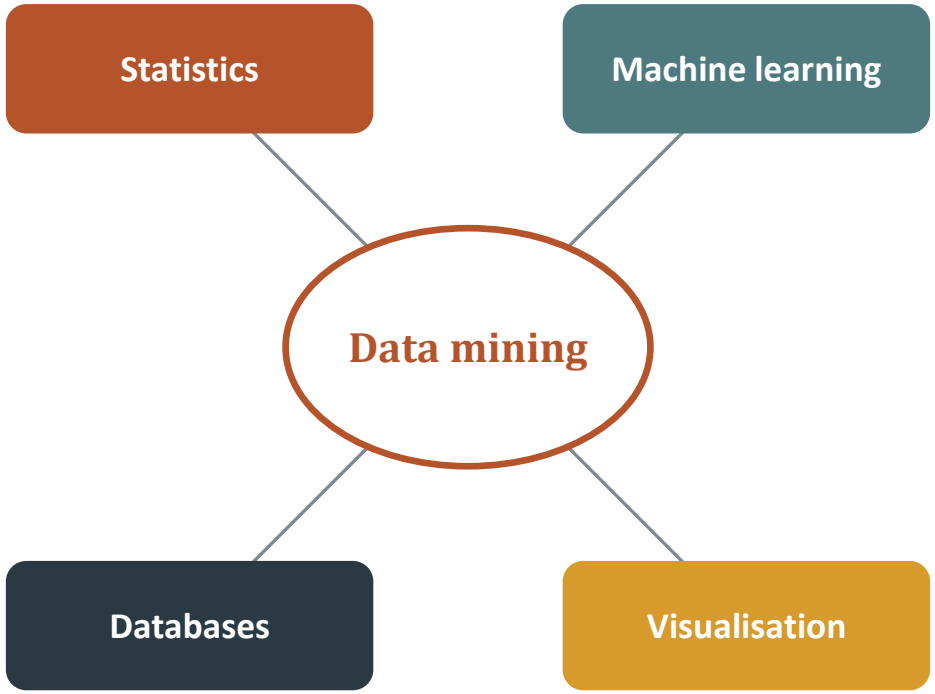
IT systems

Maintenance orders, production reports, shift logs, ERP - all timestamped.

The problem: nobody can look at all of it by eye. **Data mining** = automated ways of turning that pile of numbers into decisions: which material to blend, when to service a machine, why recovery dropped.

What is data mining?

Data mining (knowledge discovery from data) is the extraction of patterns that are **non-trivial, previously unknown and potentially useful** from large amounts of data.



WHERE STATISTICS FITS

1. **Describe** What do the data look like? — Labs 1, 6
2. **Test** Is a difference real or just chance? — Labs 7–8
3. **Model** Can we predict one variable from others? — Lab 9

Two families of data-mining tasks

Descriptive *What is in the data?*

Exploratory data analysis (EDA)
Summaries, graphs, outliers

Clustering
Group similar samples, e.g. material types from tests

Associations / patterns
What tends to happen together?

Predictive *What will happen?*

Regression
Predict a number: recovery from Fe and moisture

Classification
Predict a category: will this machine fail? yes / no

Time-series analysis
Forecast production or prices over time

Plus **process mining**: the same idea applied to event logs - discovering how a process really runs.

Recipe for a good data analysis

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

“Act if you need!” - at every step, fix what you find before moving on.



Why a recipe?

Skipping a step is how analyses go wrong: a test on dirty data, or a test on very non-normal data.

Your project report will follows these 7 steps.

The knowledge discovery process



Example from a mine: “Why is recovery lower some weeks?” → collect sample + plant data → fix typos and missing values → look at histograms and box plots → test whether material from Mine C really differs → model processing recovery → check the model → change the blending plan.

Anatomy of a dataset

sample_id	mine	rock_type	shift	depth_m	fe_pct	sio2_pct	moisture_pct	fracture_count	recovery_pct
S001	**Mine_A**	**Hematite**	**Day**	**263.2**	**60.08**	**7.53**	**4.97**	**2**	**67.56**
S002	Mine_A	Mixed	Day	369.5	54.85	8.86	5.82	4	61.12
S003	Mine_A	Hematite	Night	272.0	57.75	9.93	4.29	2	64.31
S004	Mine_A	Mixed	Day	323.0	59.47	8.02	8.19	6	60.88
…	…	…	…	…	…	…	…	…	…

One row = one case (record, observation): one drill-core sample

One column = one variable (feature): something we measured

Target variable (response): what we want to predict



Which column would a plant manager most want to predict - and why?

Our dataset: ore_samples_clean.csv

240
drill-core samples

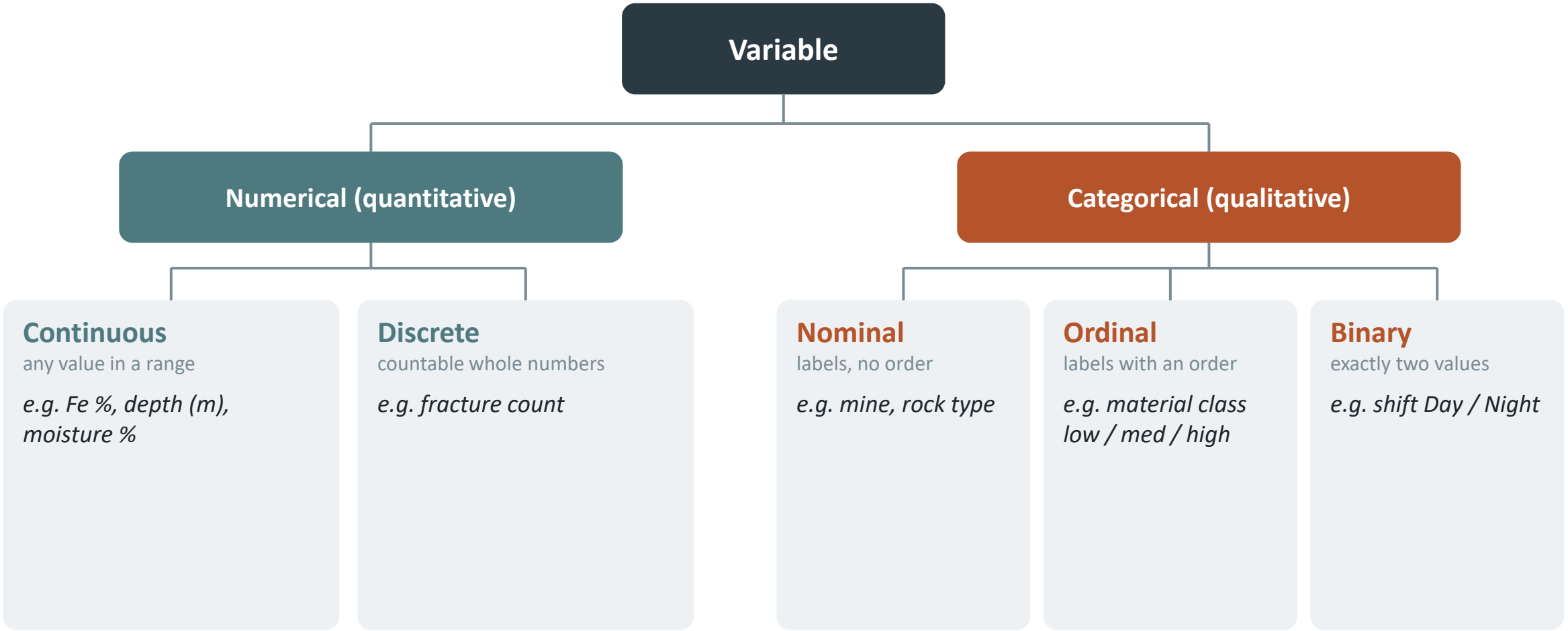
3
iron mines

10
variables

A teaching dataset modelled on a real iron-ore operation - one row per drill-core sample sent to the lab.

Column	What it means	Values
`sample_id`	Label of the sample	S001 … S240
`mine`	Which mine the sample comes from	Mine_A, Mine_B, Mine_C
`rock_type`	Main iron mineral: hematite (Fe_2O_3), magnetite (Fe_3O_4, magnetic) or a mix	Hematite / Magnetite / Mixed
`shift`	Shift during which the sample was taken	Day / Night
`depth_m`	Depth of the sample below surface	21.6 – 399.5 m
`fe_pct`	Iron grade - % of iron in the rock. Higher = more valuable	44.9 – 66.1 %
`sio2_pct`	Silica (quartz) - the main impurity.	4.2 – 14.1 %
`moisture_pct`	Water content - you pay to transport and dry it	1.9 – 15.3 %
`fracture_count`	Number of cracks in the core - how broken the rock is	0 – 9
`recovery_pct`	Share of the iron the plant recovers into concentrate - money	43.4 – 77.4 %

Variable types



The type decides the tools

Type	Summarise with	Graph with	Don't
Continuous	mean, median, SD	histogram, box plot	—
Discrete	mean / median, counts	bar chart, histogram	—
Nominal	counts, %, mode	bar chart	**mean, SD**
Ordinal	counts, median, percentiles	ordered bar chart	**mean (careful)**
Binary	proportion (%)	bar chart	**mean of labels**

"The mean of `mine` is 2.0"

…is nonsense. Mine B is not an average of A and C.

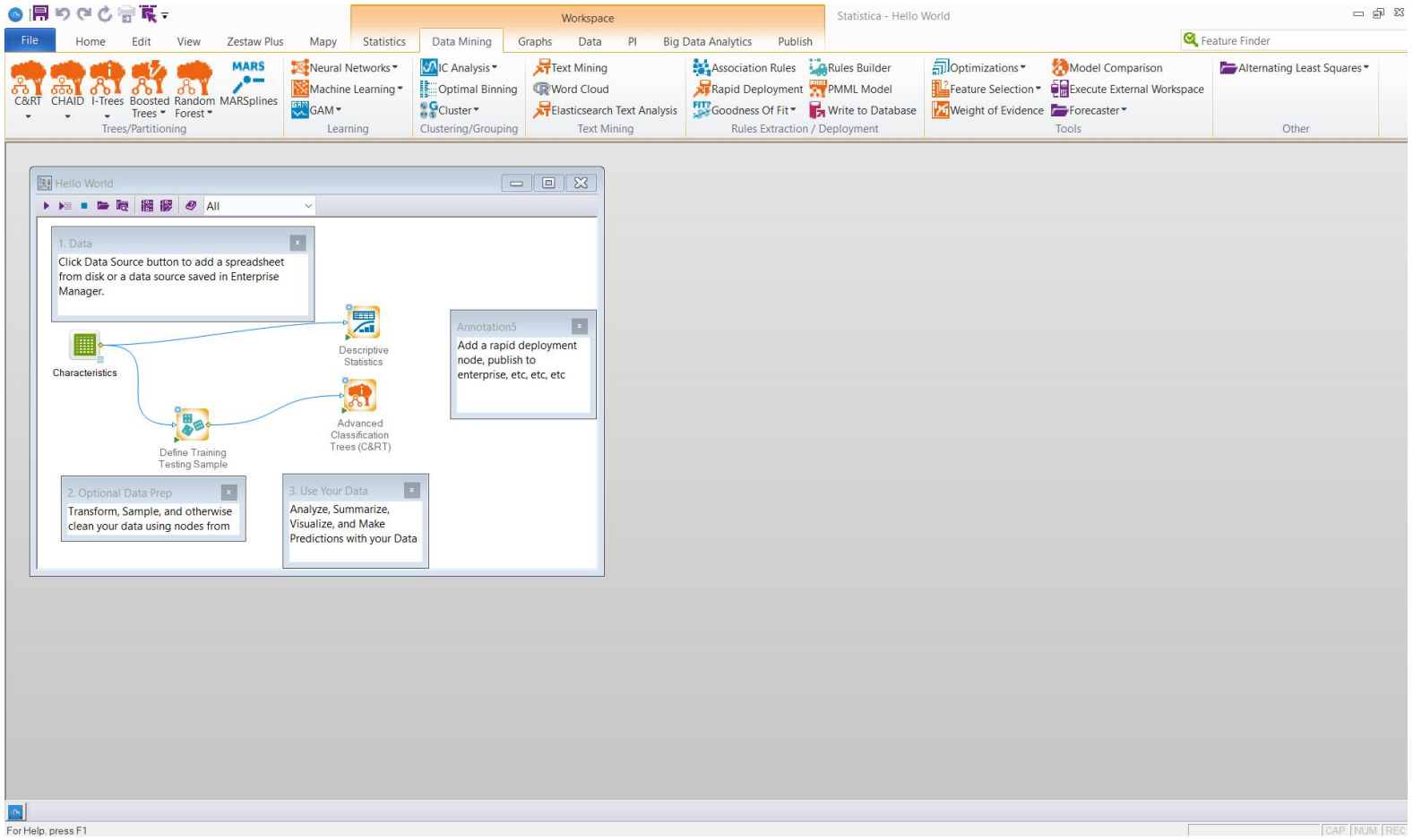
Statistica will happily compute it if mine is coded 1/2/3 – the software does not know what your numbers mean. You do.

04

Statistica: a guided tour

Live demo

The Statistica window



1 **Ribbon**
Tabs: Home, Edit, View, Insert, Format, Statistics, Data Mining, Graphs, Tools, Data

2 **Spreadsheet**
Cases in rows, variables in columns — like Excel, but each column has a type

3 **Workbook**
Tree on the left; every table and graph you make lands here

4 **Analysis dialog**
Choose variables → set options → press Summary

Step 1 · Open the CSV

File › **Open** › **ore_samples_clean.csv** › **Import: Delimited**

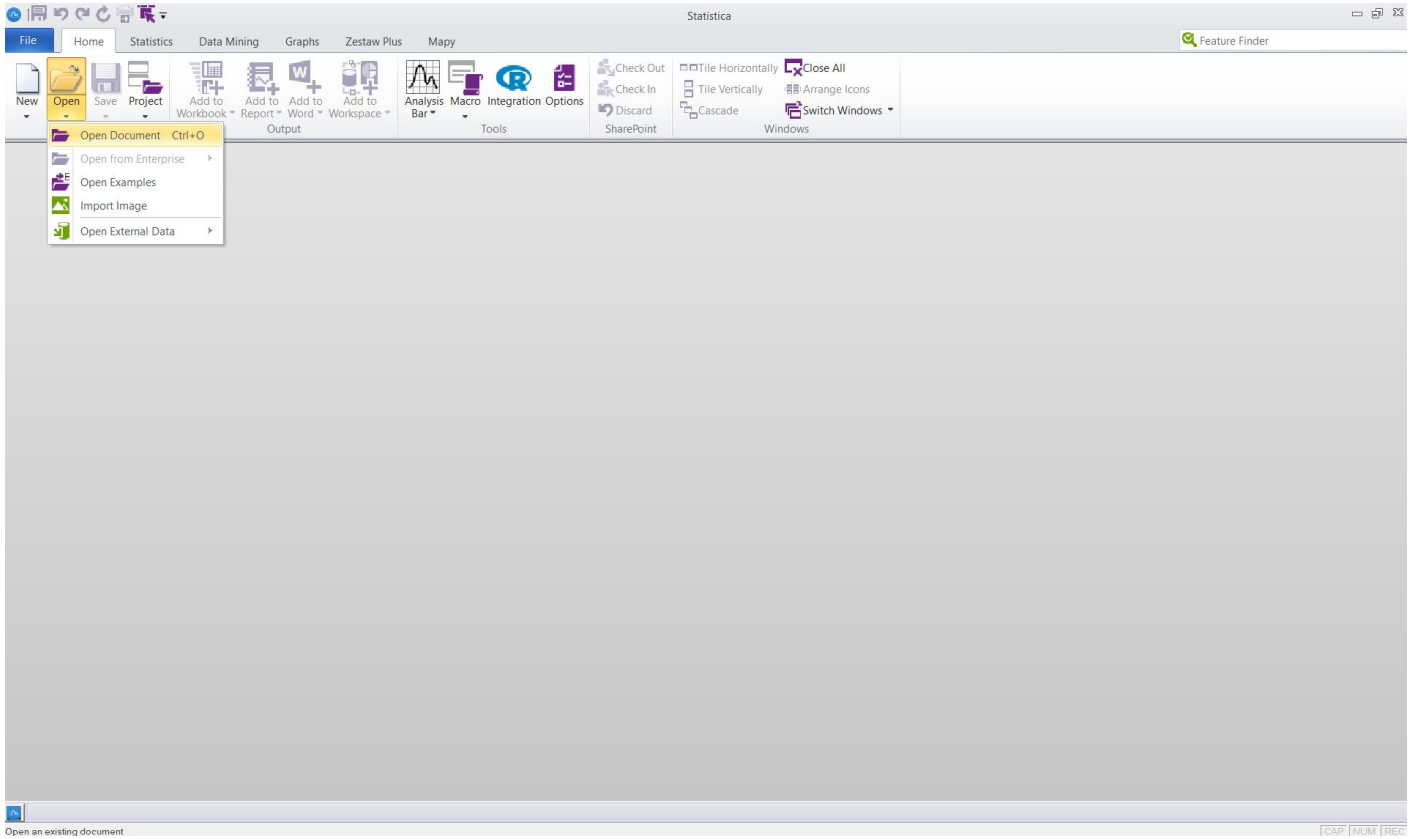
Field separator: comma

☑ Get variable names from first row

☐ Case names from first column (optional)

Decimal separator: dot (.)

Polish/European Windows uses a comma as decimal separator. If yours does, Fe % 53.85 is read as text – and no statistics will work.



Step 1 · Open the CSV

File › **Open** › **ore_samples_clean.csv** › **Import: Delimited**

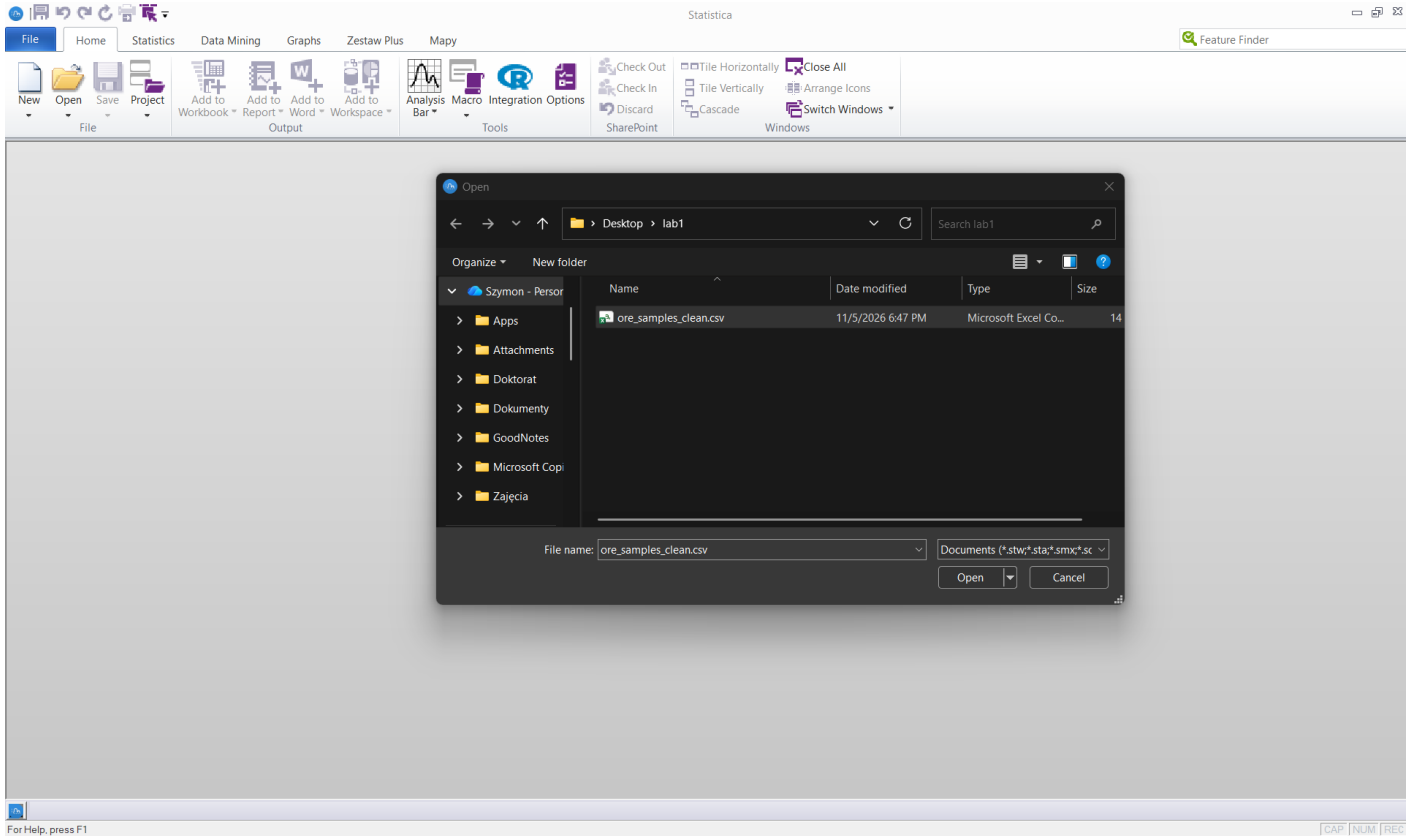
Field separator: comma

☑ Get variable names from first row

☐ Case names from first column (optional)

Decimal separator: dot (.)

Polish/European Windows uses a comma as decimal separator. If yours does, Fe % 53.85 is read as text – and no statistics will work.



Step 1 · Open the CSV

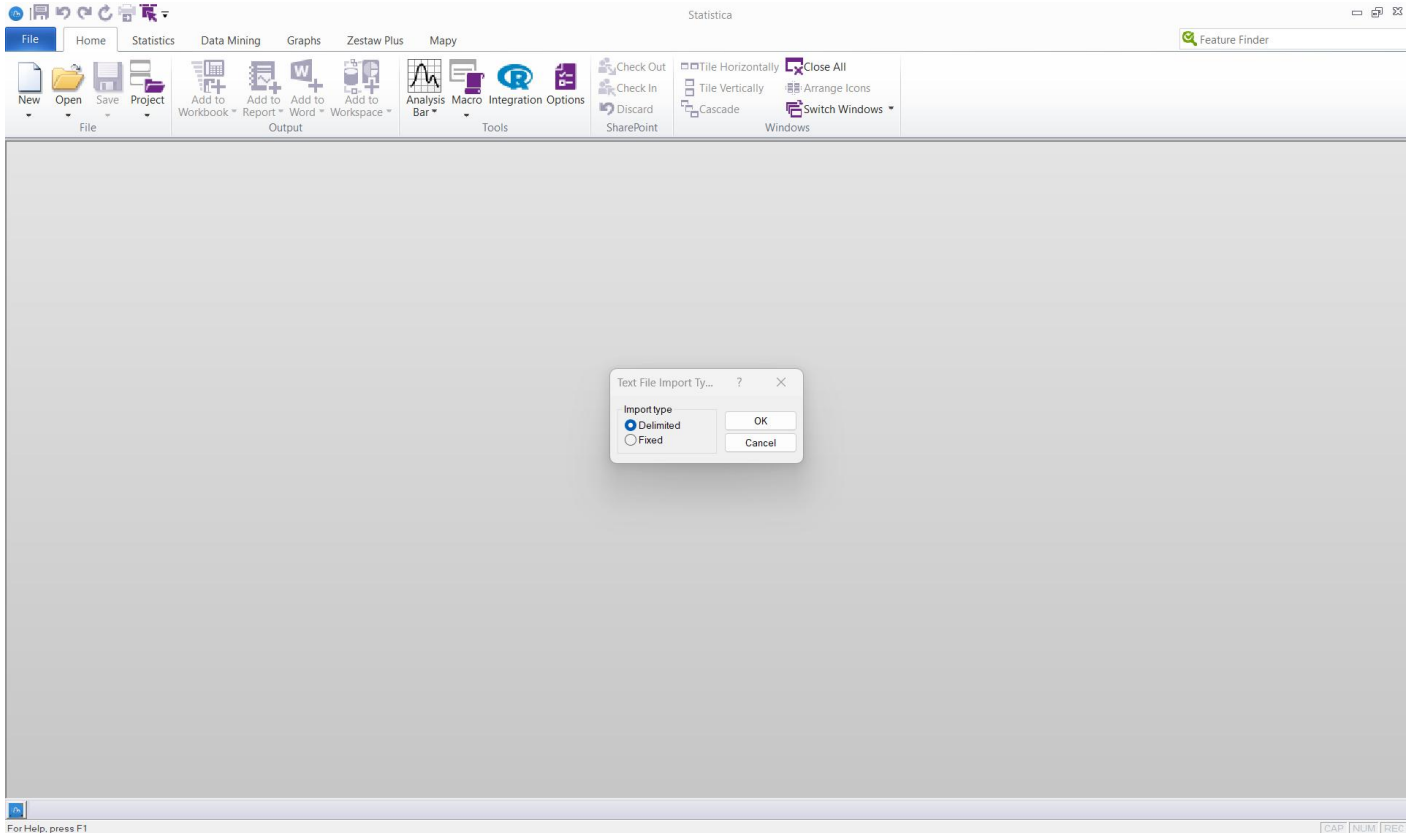
File › **Open** › **ore_samples_clean.csv** › **Import: Delimited**

Field separator: comma

☑ Get variable names from first row

☐ Case names from first column (optional)

Decimal separator: dot (.)



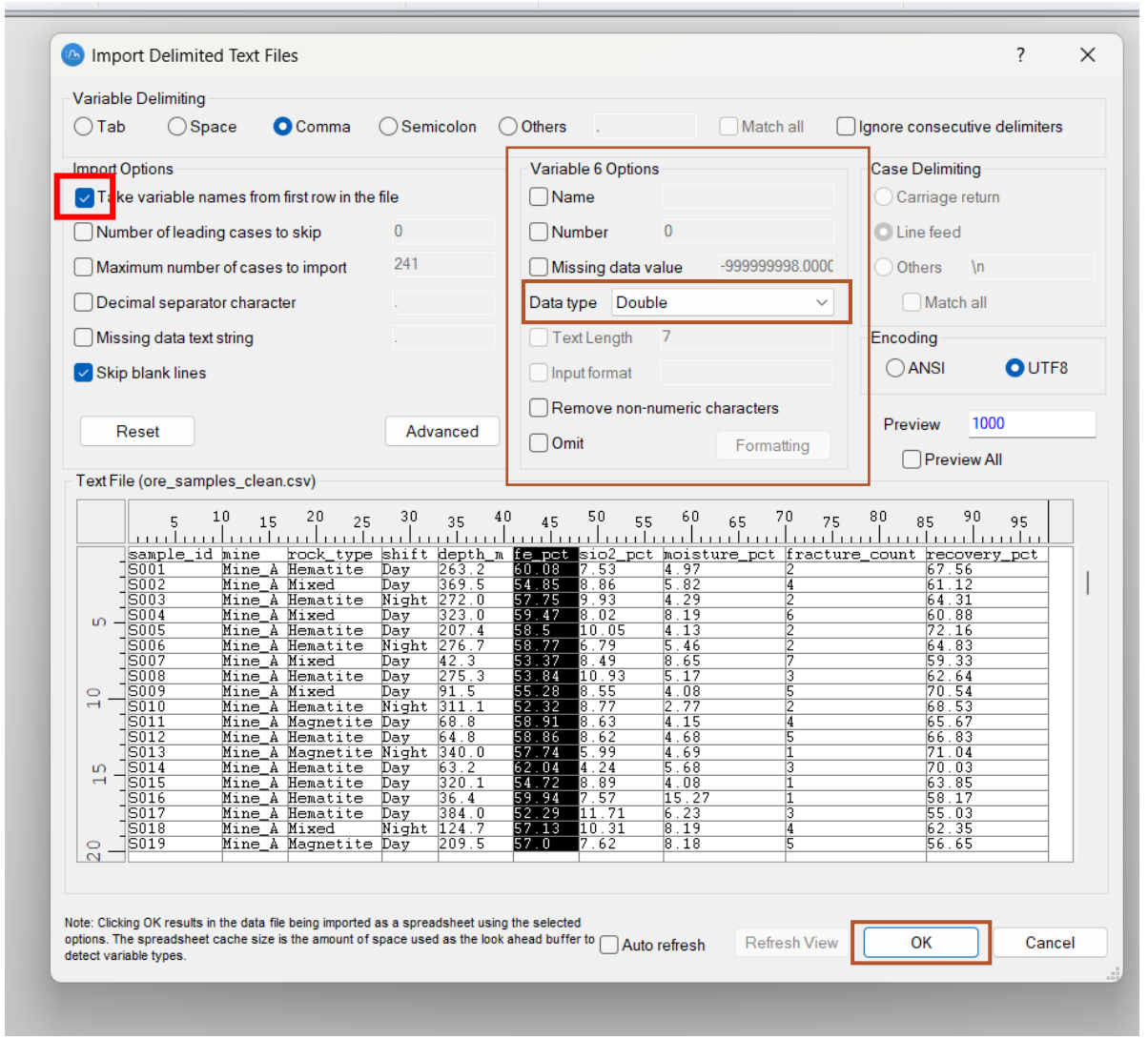
Polish/European Windows uses a comma as decimal separator. If yours does, Fe % 53.85 is read as text – and no statistics will work.

Step 2 · Check every variable's type

Click on column header › **Check Variable Options** › **OK**

Type: Double	numbers you can compute with - fe_pct, depth_m…
Type: Text	labels - mine, rock_type, shift
MD code: –9999	Statistica's default code for a missing value

Quick test: if `fe_pct` shows as Text, go back to Step 1 and fix the decimal separator.



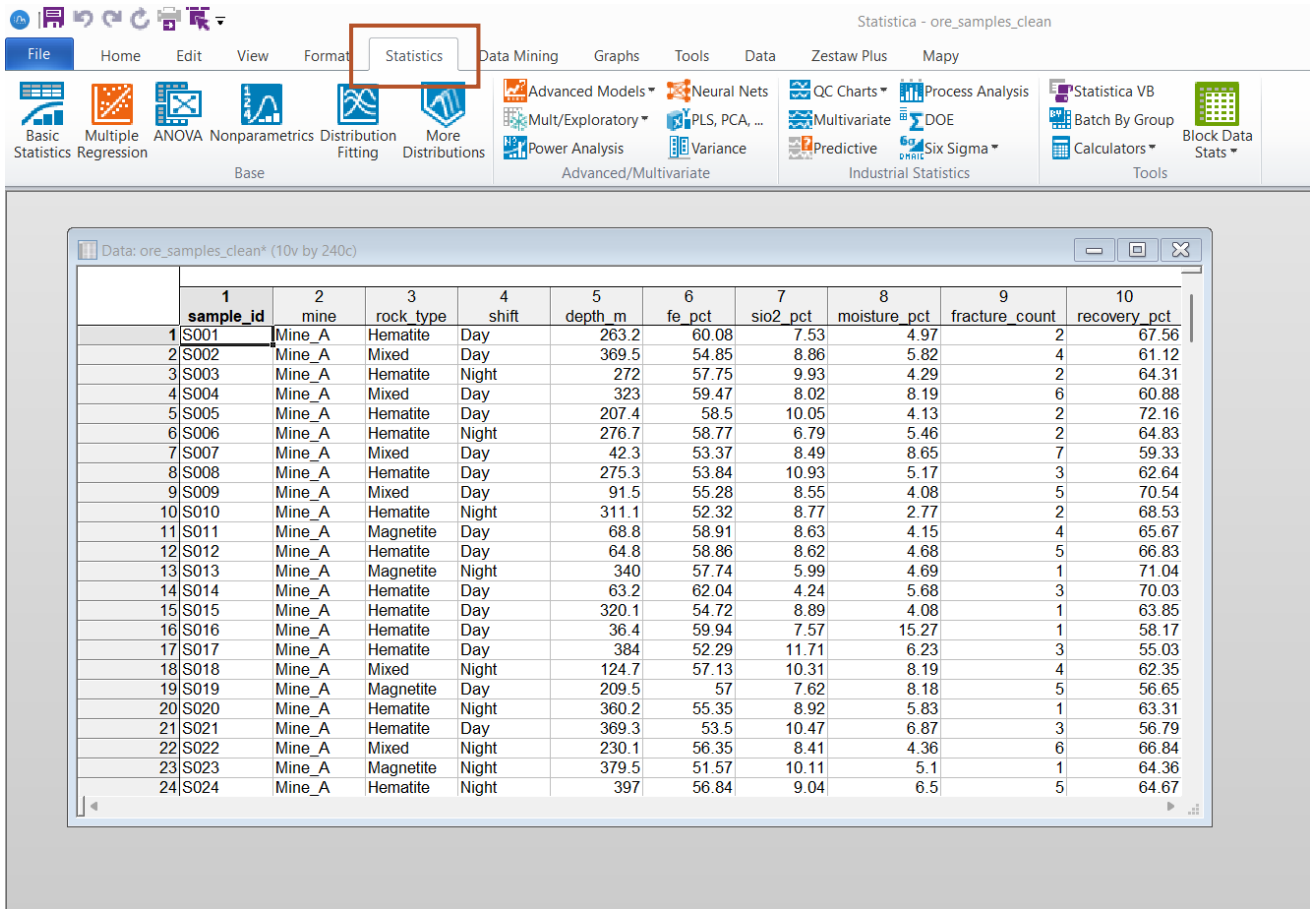
Step 3 · Descriptive statistics

Statistics › **Basic Statistics** › **Descriptive statistics** › **Variables** › **Advanced tab** › **Summary**

On the Advanced tab, tick:

- ☑ Mean
- ☑ Median
- ☑ Std. Dev.
- ☑ Minimum & maximum
- ☑ Lower & upper quartiles
- ☑ Skewness

Look for: is the mean close to the median? If not, the variable is skewed - a long tail on one side. Which of Fe, SiO_2, moisture looks skewed?



	1 sample_id	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	S001	Mine_A	Hematite	Day	263.2	60.08	7.53	4.97	2	67.56
2	S002	Mine_A	Mixed	Day	369.5	54.85	8.86	5.82	4	61.12
3	S003	Mine_A	Hematite	Night	272	57.75	9.93	4.29	2	64.31
4	S004	Mine_A	Mixed	Day	323	59.47	8.02	8.19	6	60.88
5	S005	Mine_A	Hematite	Day	207.4	58.5	10.05	4.13	2	72.16
6	S006	Mine_A	Hematite	Night	276.7	58.77	6.79	5.46	2	64.83
7	S007	Mine_A	Mixed	Day	42.3	53.37	8.49	8.65	7	59.33
8	S008	Mine_A	Hematite	Day	275.3	53.84	10.93	5.17	3	62.64
9	S009	Mine_A	Mixed	Day	91.5	55.28	8.55	4.08	5	70.54
10	S010	Mine_A	Hematite	Night	311.1	52.32	8.77	2.77	2	68.53
11	S011	Mine_A	Magnetite	Day	68.8	58.91	8.63	4.15	4	65.67
12	S012	Mine_A	Hematite	Day	64.8	58.86	8.62	4.68	5	66.83
13	S013	Mine_A	Magnetite	Night	340	57.74	5.99	4.69	1	71.04
14	S014	Mine_A	Hematite	Day	63.2	62.04	4.24	5.68	3	70.03
15	S015	Mine_A	Hematite	Day	320.1	54.72	8.89	4.08	1	63.85
16	S016	Mine_A	Hematite	Day	36.4	59.94	7.57	15.27	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	51.57	10.11	5.1	1	64.36
24	S024	Mine_A	Hematite	Night	397	56.84	9.04	6.5	5	64.67

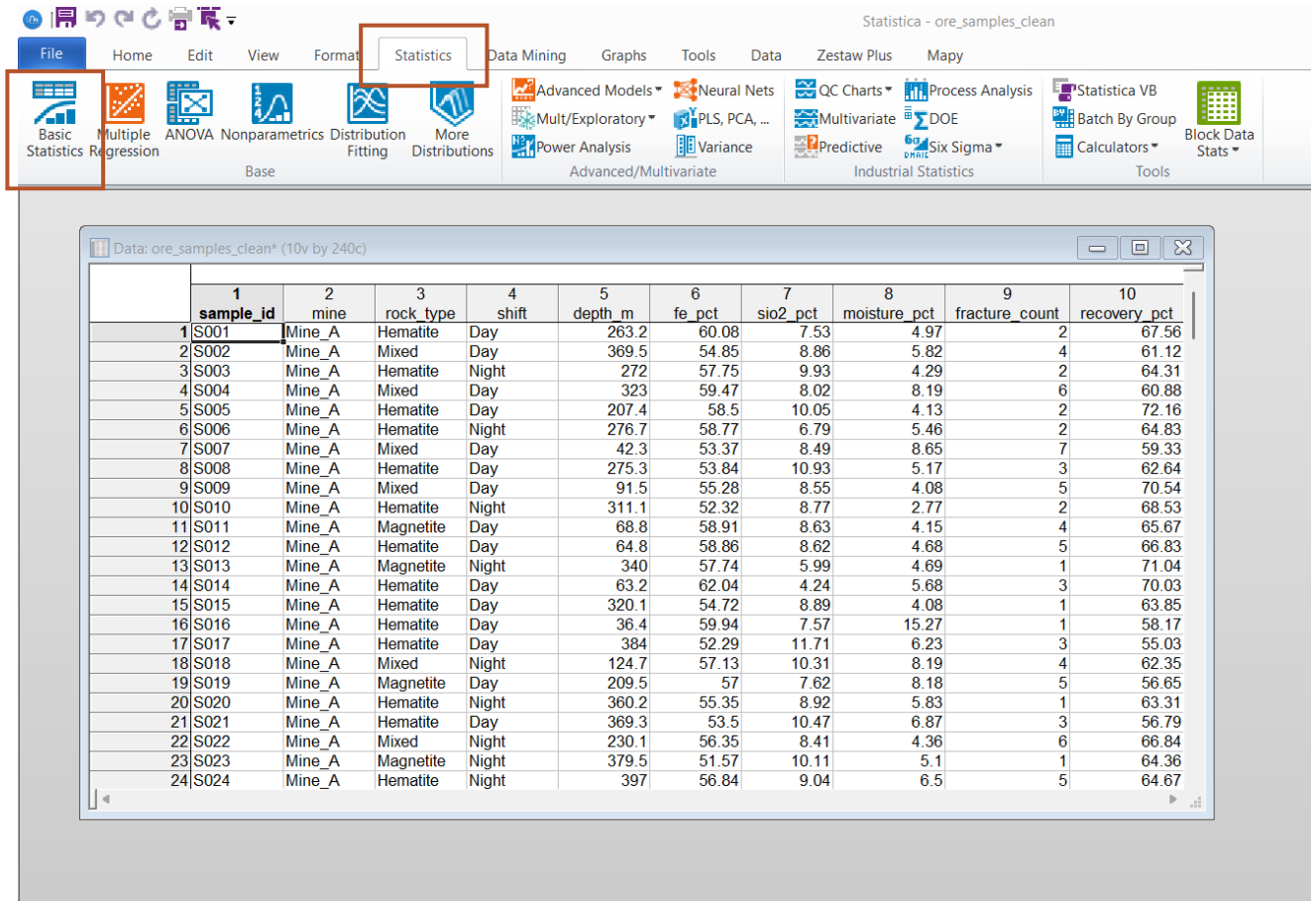
Step 3 · Descriptive statistics

Statistics › **Basic Statistics** › **Descriptive statistics** › **Variables** › **Advanced tab** › **Summary**

On the Advanced tab, tick:

- ☑ Mean
- ☑ Median
- ☑ Std. Dev.
- ☑ Minimum & maximum
- ☑ Lower & upper quartiles
- ☑ Skewness

Look for: is the mean close to the median? If not, the variable is skewed - a long tail on one side. Which of Fe, SiO_2, moisture looks skewed?



	1 sample_id	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	S001	Mine_A	Hematite	Day	263.2	60.08	7.53	4.97	2	67.56
2	S002	Mine_A	Mixed	Day	369.5	54.85	8.86	5.82	4	61.12
3	S003	Mine_A	Hematite	Night	272	57.75	9.93	4.29	2	64.31
4	S004	Mine_A	Mixed	Day	323	59.47	8.02	8.19	6	60.88
5	S005	Mine_A	Hematite	Day	207.4	58.5	10.05	4.13	2	72.16
6	S006	Mine_A	Hematite	Night	276.7	58.77	6.79	5.46	2	64.83
7	S007	Mine_A	Mixed	Day	42.3	53.37	8.49	8.65	7	59.33
8	S008	Mine_A	Hematite	Day	275.3	53.84	10.93	5.17	3	62.64
9	S009	Mine_A	Mixed	Day	91.5	55.28	8.55	4.08	5	70.54
10	S010	Mine_A	Hematite	Night	311.1	52.32	8.77	2.77	2	68.53
11	S011	Mine_A	Magnetite	Day	68.8	58.91	8.63	4.15	4	65.67
12	S012	Mine_A	Hematite	Day	64.8	58.86	8.62	4.68	5	66.83
13	S013	Mine_A	Magnetite	Night	340	57.74	5.99	4.69	1	71.04
14	S014	Mine_A	Hematite	Day	63.2	62.04	4.24	5.68	3	70.03
15	S015	Mine_A	Hematite	Day	320.1	54.72	8.89	4.08	1	63.85
16	S016	Mine_A	Hematite	Day	36.4	59.94	7.57	15.27	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	51.57	10.11	5.1	1	64.36
24	S024	Mine_A	Hematite	Night	397	56.84	9.04	6.5	5	64.67

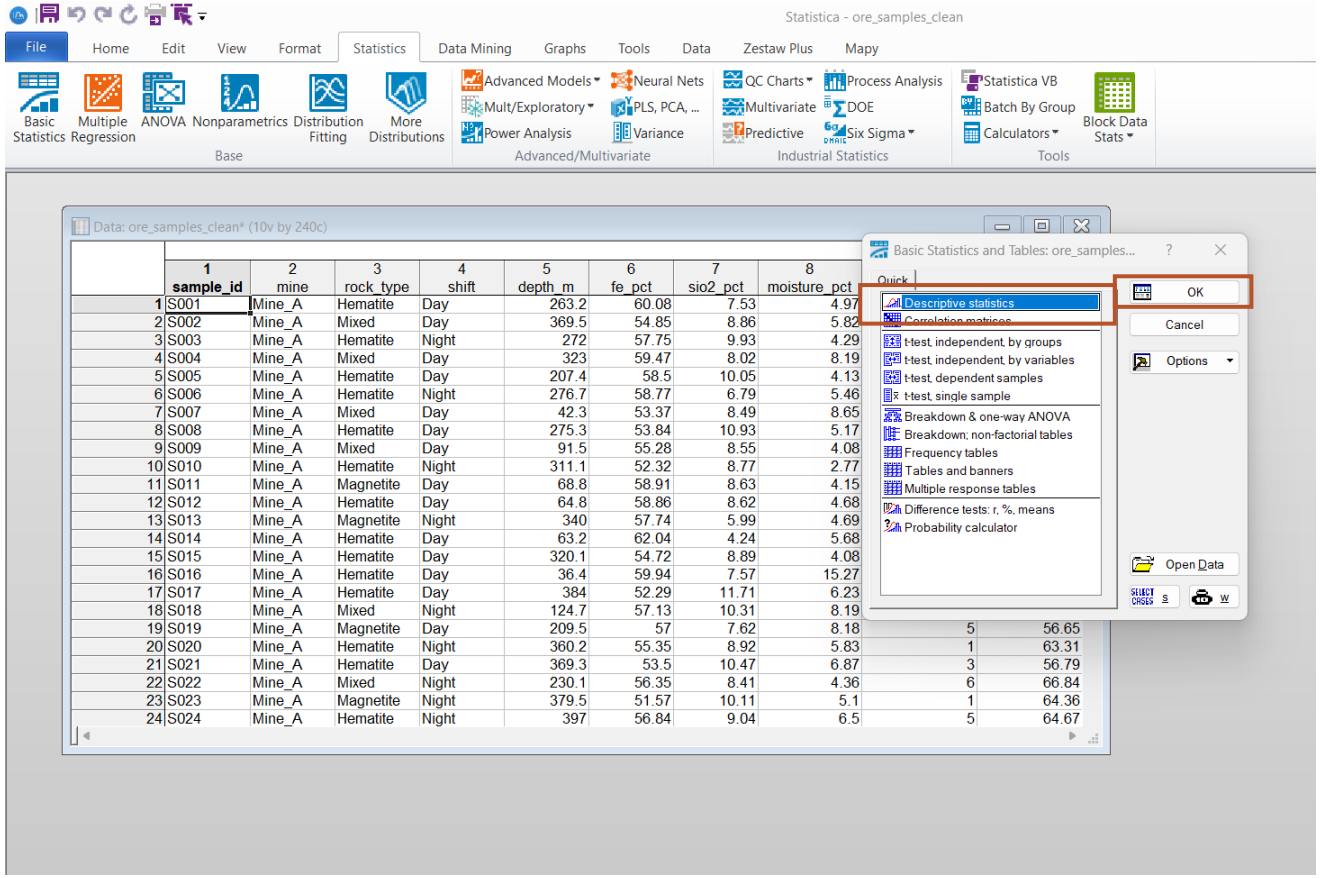
Step 3 · Descriptive statistics

Statistics › Basic Statistics › Descriptive statistics › Variables › Advanced tab › Summary

On the Advanced tab, tick:

- ☑ Mean
- ☑ Median
- ☑ Std. Dev.
- ☑ Minimum & maximum
- ☑ Lower & upper quartiles
- ☑ Skewness

Look for: is the mean close to the median? If not, the variable is skewed - a long tail on one side. Which of Fe, SiO_2, moisture looks skewed?



	1 sample_id	2 mine	3 rock_type	4 shift	5 depth_m	6 fe_pct	7 sio2_pct	8 moisture_pct		
1	S001	Mine_A	Hematite	Day	263.2	60.08	7.53	4.97		
2	S002	Mine_A	Mixed	Day	369.5	54.85	8.86	5.82		
3	S003	Mine_A	Hematite	Night	272	57.75	9.93	4.29		
4	S004	Mine_A	Mixed	Day	323	59.47	8.02	8.19		
5	S005	Mine_A	Hematite	Day	207.4	58.5	10.05	4.13		
6	S006	Mine_A	Hematite	Night	276.7	58.77	6.79	5.46		
7	S007	Mine_A	Mixed	Day	42.3	53.37	8.49	8.65		
8	S008	Mine_A	Hematite	Day	275.3	53.84	10.93	5.17		
9	S009	Mine_A	Mixed	Day	91.5	55.28	8.55	4.08		
10	S010	Mine_A	Hematite	Night	311.1	52.32	8.77	2.77		
11	S011	Mine_A	Magnetite	Day	68.8	58.91	8.63	4.15		
12	S012	Mine_A	Hematite	Day	64.8	58.86	8.62	4.68		
13	S013	Mine_A	Magnetite	Night	340	57.74	5.99	4.69		
14	S014	Mine_A	Hematite	Day	63.2	62.04	4.24	5.68		
15	S015	Mine_A	Hematite	Day	320.1	54.72	8.89	4.08		
16	S016	Mine_A	Hematite	Day	36.4	59.94	7.57	15.27		
17	S017	Mine_A	Hematite	Day	384	52.29	11.71	6.23		
18	S018	Mine_A	Mixed	Night	124.7	57.13	10.31	8.19		
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	51.57	10.11	5.1	1	64.36
24	S024	Mine_A	Hematite	Night	397	56.84	9.04	6.5	5	64.67

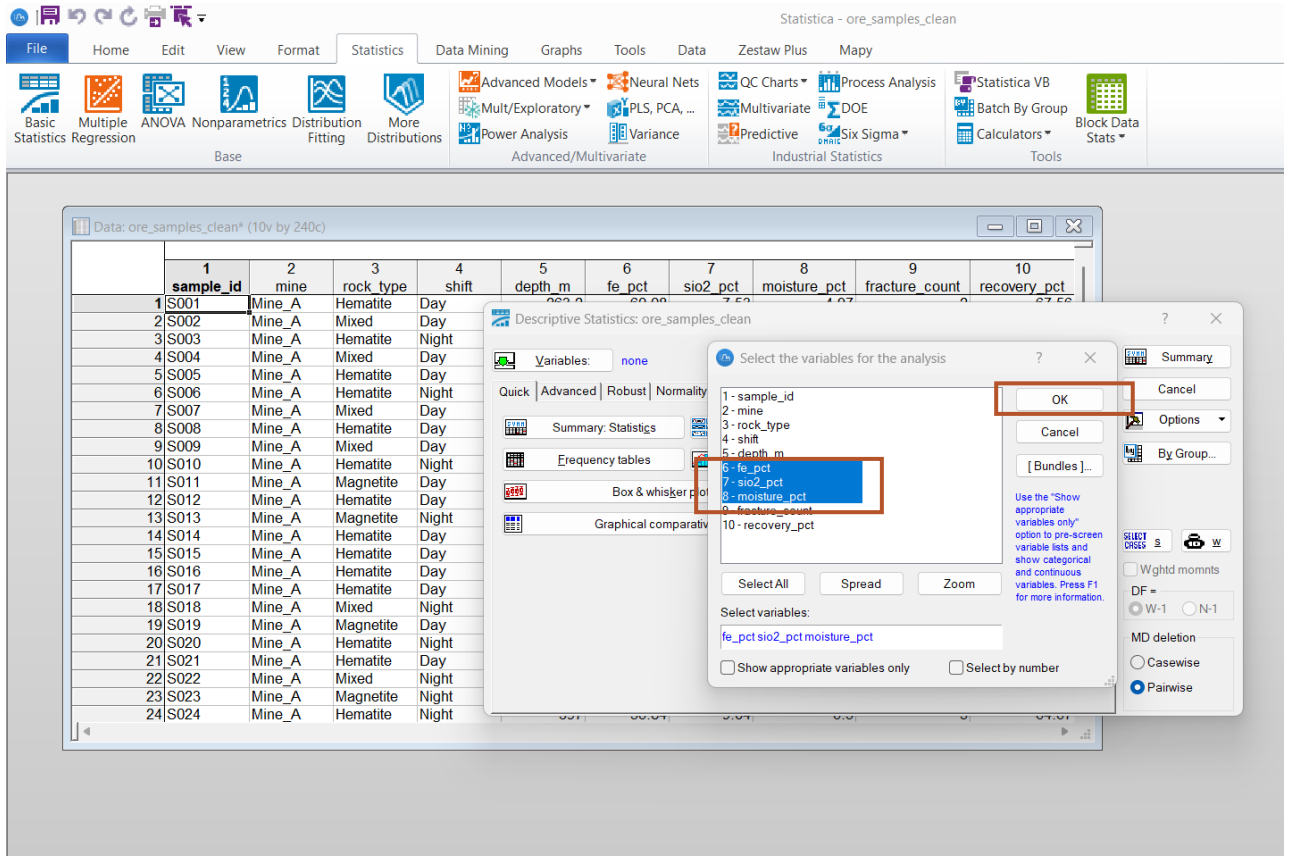
Step 3 · Descriptive statistics

Statistics › **Basic Statistics** › **Descriptive statistics** › **Variables** › **Advanced tab** › **Summary**

On the Advanced tab, tick:

- ☑ Mean
- ☑ Median
- ☑ Std. Dev.
- ☑ Minimum & maximum
- ☑ Lower & upper quartiles
- ☑ Skewness

Look for: is the mean close to the median? If not, the variable is skewed - a long tail on one side. Which of Fe, SiO_2, moisture looks skewed?



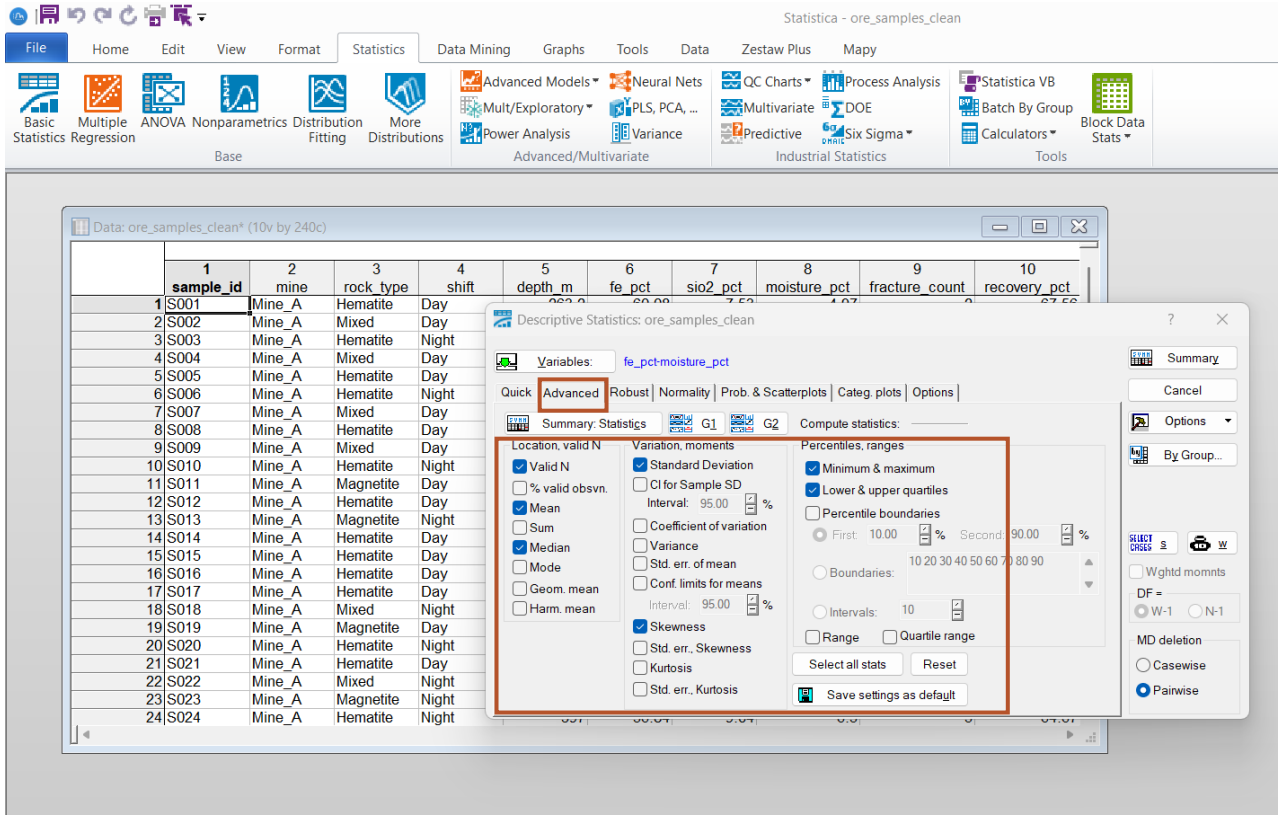
Step 3 · Descriptive statistics

Statistics › **Basic Statistics** › **Descriptive statistics** › **Variables** › **Advanced tab** › **Summary**

On the Advanced tab, tick:

- ☑ Mean
- ☑ Median
- ☑ Std. Dev.
- ☑ Minimum & maximum
- ☑ Lower & upper quartiles
- ☑ Skewness

Look for: is the mean close to the median? If not, the variable is skewed - a long tail on one side. Which of Fe, SiO_2, moisture looks skewed?



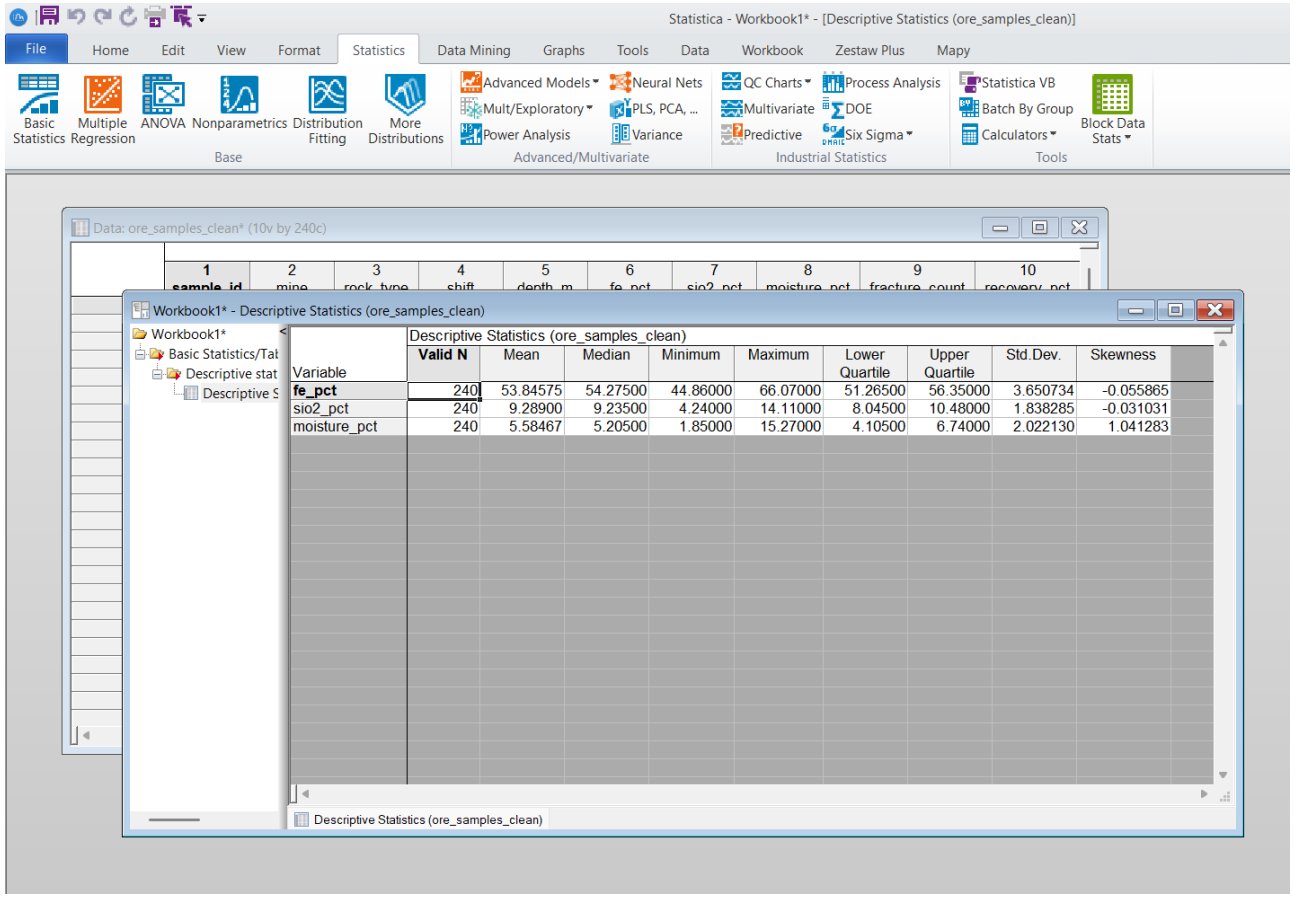
Step 3 · Descriptive statistics

Statistics › Basic Statistics › Descriptive statistics › Variables › Advanced tab › Summary

On the Advanced tab, tick:

- ☑ Mean
- ☑ Median
- ☑ Std. Dev.
- ☑ Minimum & maximum
- ☑ Lower & upper quartiles
- ☑ Skewness

Look for: is the mean close to the median? If not, the variable is skewed - a long tail on one side. Which of Fe, SiO_2, moisture looks skewed?

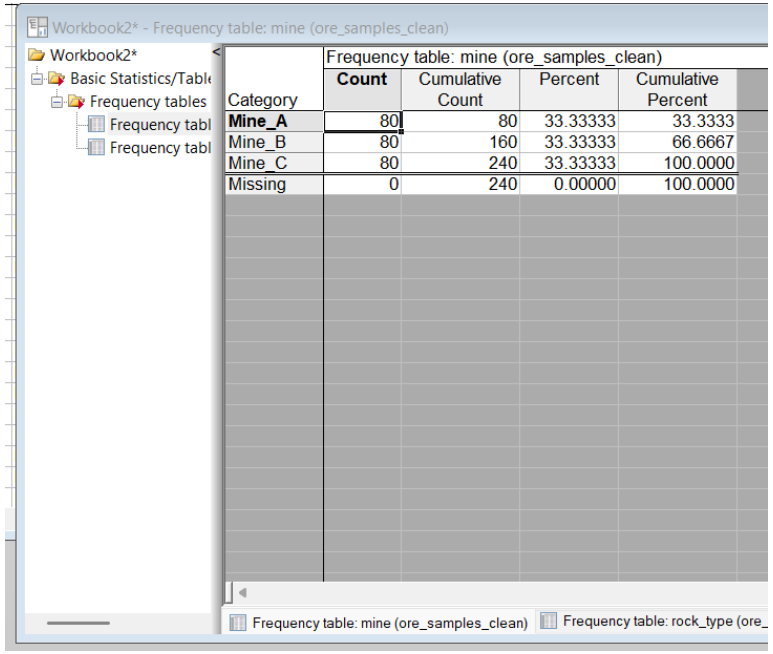


Step 4 · Frequencies and graphs

Frequency table

Statistics › Basic Statistics › Frequency tables

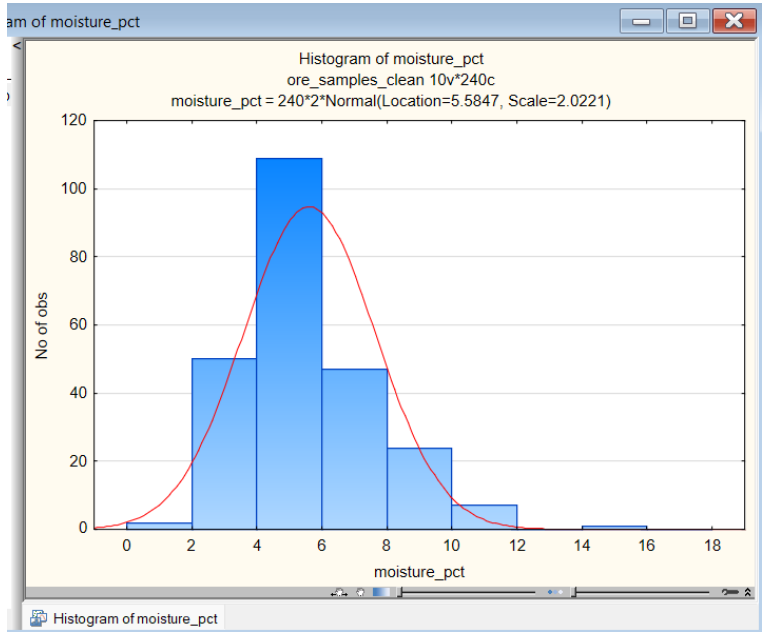
`mine, rock_type → Summary`



Histogram

Graphs › Histogram

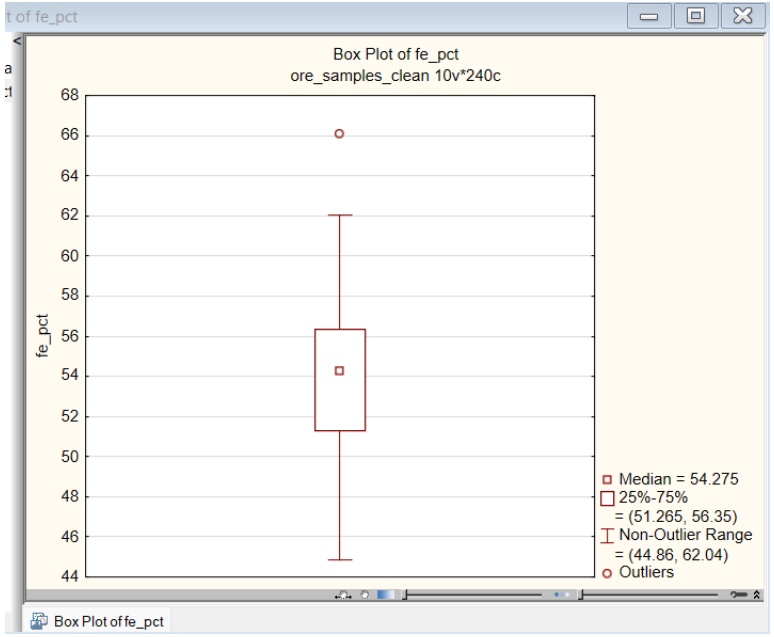
`moisture_pct`



Box plot by group

Graphs › Box Plot

`dependent fe_pct · grouping mine`



Step 5 · Save your work



.sta

Data file

The spreadsheet itself: cases, variables, types. Save after cleaning.



.stw

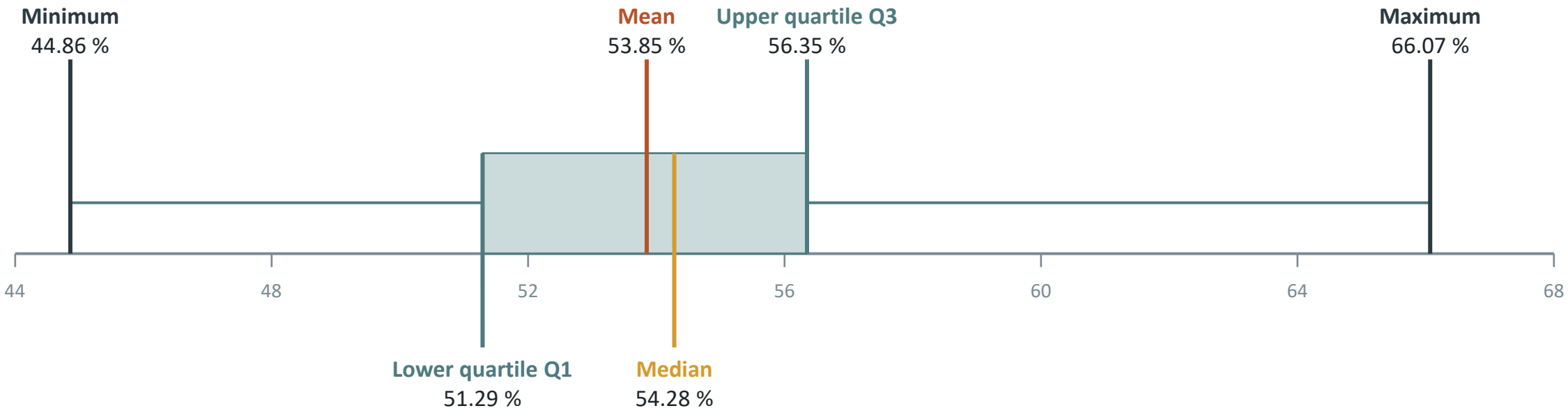
Workbook

Every result table and graph you made. This is what you submit with your report.

File › Save As › Lab1_YourName.stw

Six numbers that describe a variable

Example: iron grade `fe_pct` of all 240 samples



Standard deviation 3.65 % = typical distance from the mean. **Skewness −0.06** = almost perfectly symmetric.

Mean - the average

WHAT IT IS

Add all values and divide by how many there are. The **balance point** of the data.

FORMULA

$$\bar{x} = (x_1 + x_2 + \ldots + x_n) / n$$

WORKED EXAMPLE · 5 SAMPLES OF FE %

52	54	55	53	56

52 + 54 + 55 + 53 + 56 = **270**

270 / 5 = **54.0 %**

53.85 %

Mean iron grade of our 240 samples

Use it for numerical variables that are roughly symmetric (Fe, SiO_2, recovery).

In mining: the average grade of a stockpile is what the buyer pays for.

Watch out: one extreme value pulls the mean towards it - and a mean of a category (mine, rock type) is meaningless.

Median - the middle value

WHAT IT IS

Sort the values; the median is the one in the middle. **Half the samples are below it, half above.**

HOW

Odd n: the middle value. Even n: the average of the two middle values.

WORKED EXAMPLE · 5 SAMPLES OF MOISTURE %, SORTED

3.5	4.1	5.2	5.8	15.3

Median = **5.2 %** (the 3rd of 5)

Mean = 33.9 / 5 = **6.8 %**

5.20 %
Median moisture

5.58 %
Mean moisture

Rule of thumb: if mean and median differ noticeably, the data are skewed - report the **median**. That's why house prices and salaries are reported as medians.

Watch out: the median ignores how extreme the extremes are - one very wet sample changes the mean, not the median.

Standard deviation - the typical spread

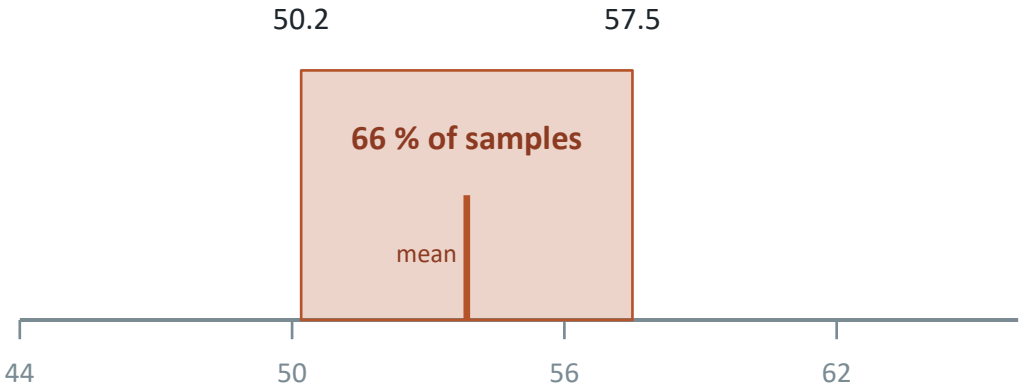
WHAT IT IS

How far, **on average, values lie from the mean**. Small SD = samples are similar. Large SD = they vary a lot.

FORMULA

$$s = \sqrt{[\ \Sigma\ (x_i - \bar{x})^2\ /\ (n - 1)\ ]}$$

OUR FE DATA: MEAN ± 1 SD



± 3.65 %

SD of iron grade, in the same unit as the data (% Fe)

Rule of thumb for bell-shaped data: about ⅔ of values lie within ±1 SD of the mean, about 95 % within ±2 SD.

Watch out: Statistica also prints “Std. err.” (SE = SD/√n) - that is the precision of the mean, not the spread of the data.

Minimum and maximum - the extremes

WHAT IT IS

The smallest and the largest value. **Range = maximum − minimum.**



Your first data-quality check

Before any analysis, ask: is every minimum and maximum physically possible?

Negative moisture, a depth of 9999 m, Fe = 540 % - they all show up here first.

OUR DATA

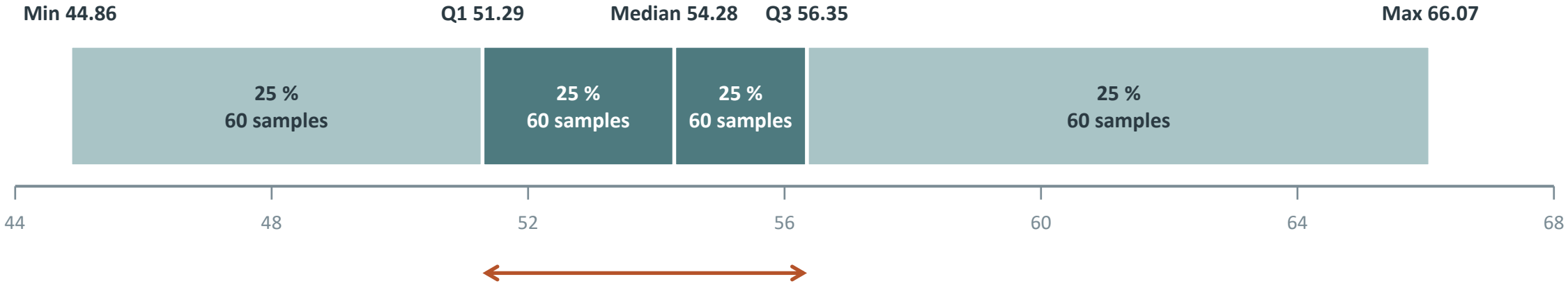
`fe_pct`	**44.86** → **66.07** %	range 21.21
`sio2_pct`	**4.24** → **14.11** %	range 9.87
`moisture_pct`	**1.85** → **15.27** %	range 13.42
`recovery_pct`	**43.36** → **77.39** %	range 34.03
`depth_m`	**21.6** → **399.5** m	range 377.9

Watch out: the range depends on just two samples. One unusual sample (moisture 15.27 %, S016) makes it look huge - check if it's real or a typo.

Quartiles - cutting the data into four

WHAT IT IS

Sort the data and cut it into four equal parts. **Q1** (lower quartile): 25 % of values below. **Q2** = the median. **Q3** (upper quartile): 75 % below.



Interquartile range IQR = Q3 − Q1 = **5.06 %** → the middle half of the samples

Why it's useful: like the median, quartiles don't care about extreme values - and they are exactly what a box plot draws (the box = Q1 to Q3).

Watch out: Statistica, Excel and Python use slightly different quartile formulas —-small differences in the 2nd decimal are normal.

Skewness - is the shape lopsided?

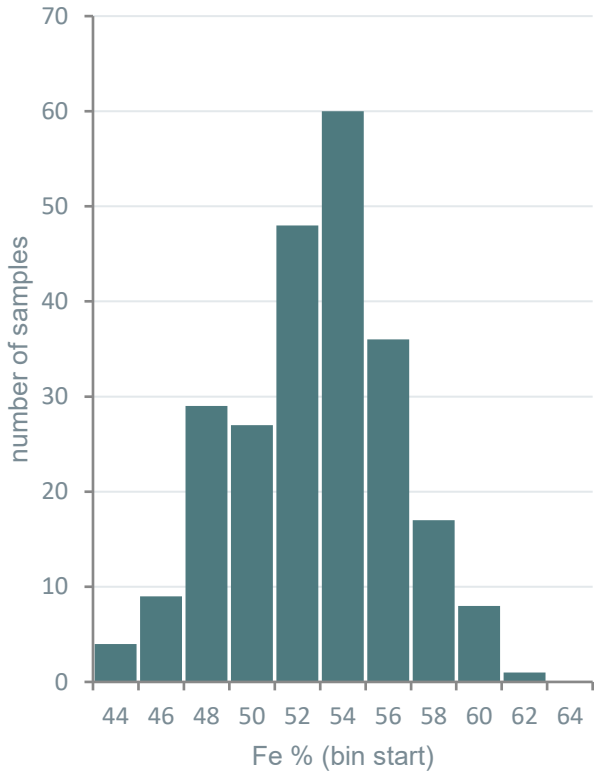
WHAT IT IS

A number for asymmetry.

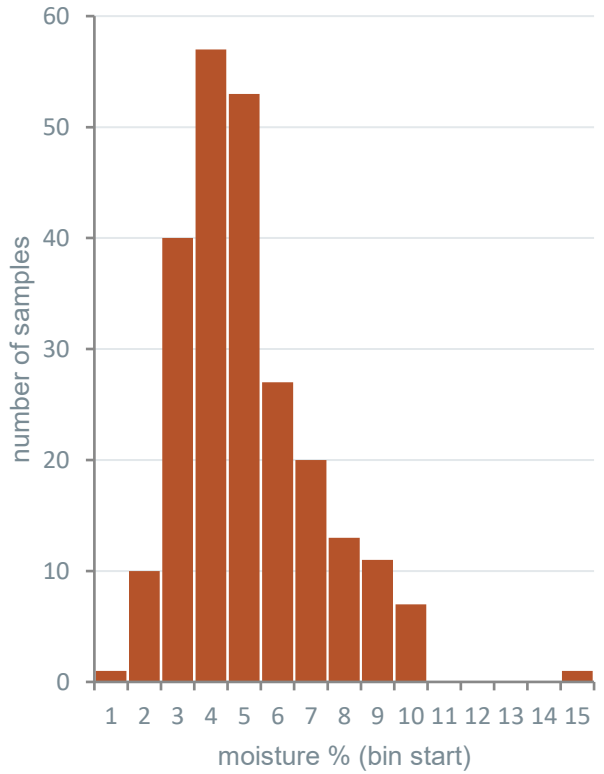
≈ 0	**symmetric** mean ≈ median
> 0	**long tail to the RIGHT** mean > median
< 0	**long tail to the LEFT** mean < median

Rough guide: |skew| < 0.5 fairly symmetric · 0.5–1 moderate · > 1 strong

Fe % · skew −0.06



Moisture % · skew 1.04



Watch out: strong skew means the mean is misleading - report the median and quartiles and expect trouble with tests that assume normality.

Frequency table - counting categories

WHAT IT IS

For each category: **how many cases** and **what share** of the total. The basic summary for categorical variables.

USE IT FOR

nominal, ordinal and binary variables - and small discrete counts (fracture_count).

Read it like this: hematite is the most common rock type (38 %), but all three are well represented. Mines are perfectly balanced: 80 / 80 / 80.

ROCK_TYPE - AS STATISTICA SHOWS IT

Category	Count	Cumul. count	Percent	Cumul. %
Hematite	91	91	37.9	37.9
Magnetite	77	168	32.1	70.0
Mixed	72	240	30.0	100.0
Missing	0	240	0.0	100.0

Check the 'Missing' row every time - it tells you how many cases have no value.

Watch out: a typo creates a new category - “Mine_a” or “Mine A ” would appear as separate rows. That's how you'll catch label errors in Lab 2.

Bar chart - categories at a glance

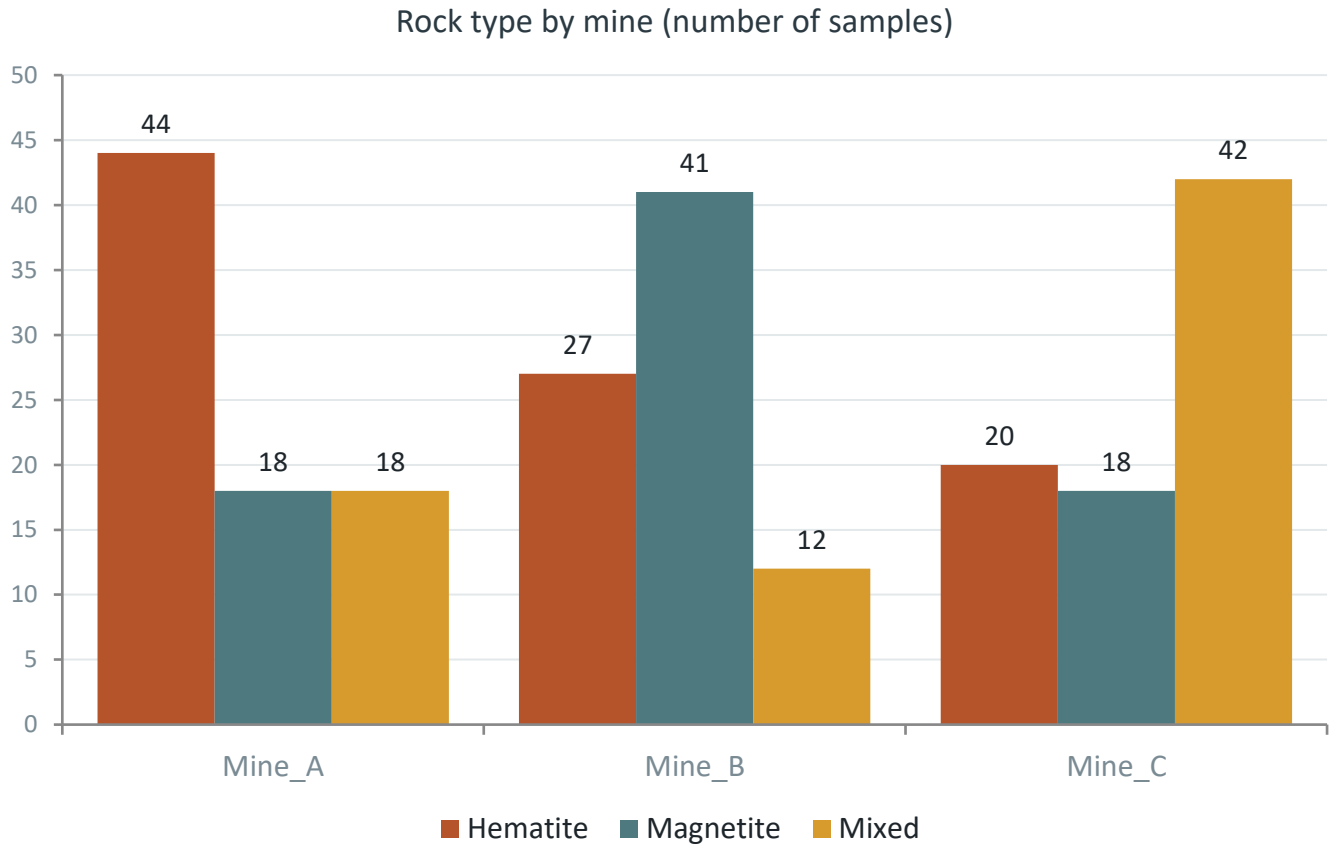
WHAT IT IS

One bar per category, height = count (or %). **Gaps between bars** - the categories are separate things.

USE IT FOR

categorical variables. Comparing two categorical variables? Put one in the colors.

What this one shows: each mine is dominated by a different rock type - Mine_A hematite, Mine_B magnetite, Mine_C mixed. (Is that real?)



Watch out: a bar chart is not a histogram - bars for categories have gaps and their order doesn't matter; histogram bars touch.

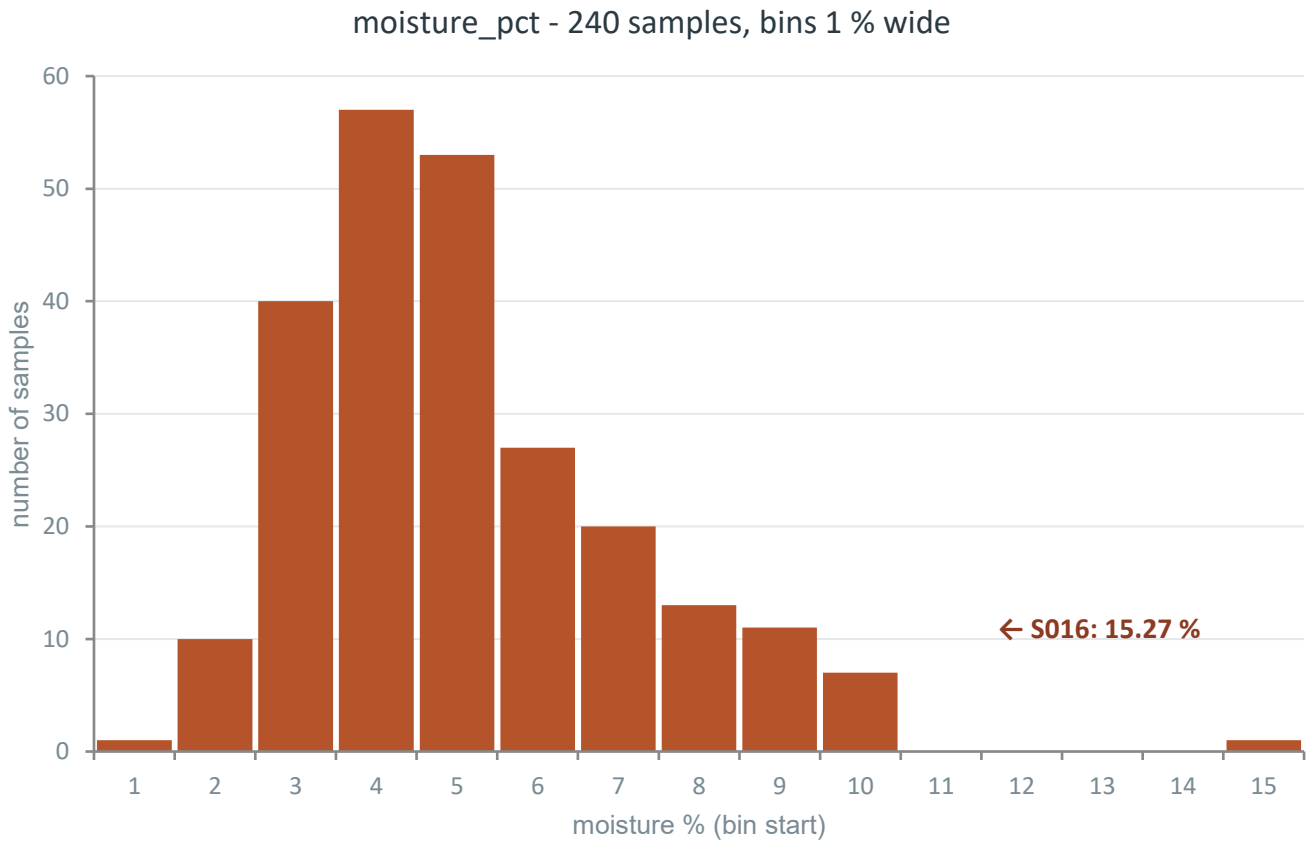
Histogram - the shape of a numerical variable

WHAT IT IS

Cut the scale into equal **bins** (e.g. 1 % wide) and count samples in each. Bars **touch** - the scale is continuous.

READ FOUR THINGS

1. **Centre** where is the peak? (4–5 %)
2. **Spread** how wide is it? (mostly 2–10 %)
3. **Shape** symmetric or a tail? (long right tail)
4. **Outliers** anything far away? (one at 15.3 %)



Watch out: the bin width changes the picture - too few bins hide the shape; too many make it noisy. Statistica picks a default; try changing it.

Box plot - comparing groups

HOW TO READ IT

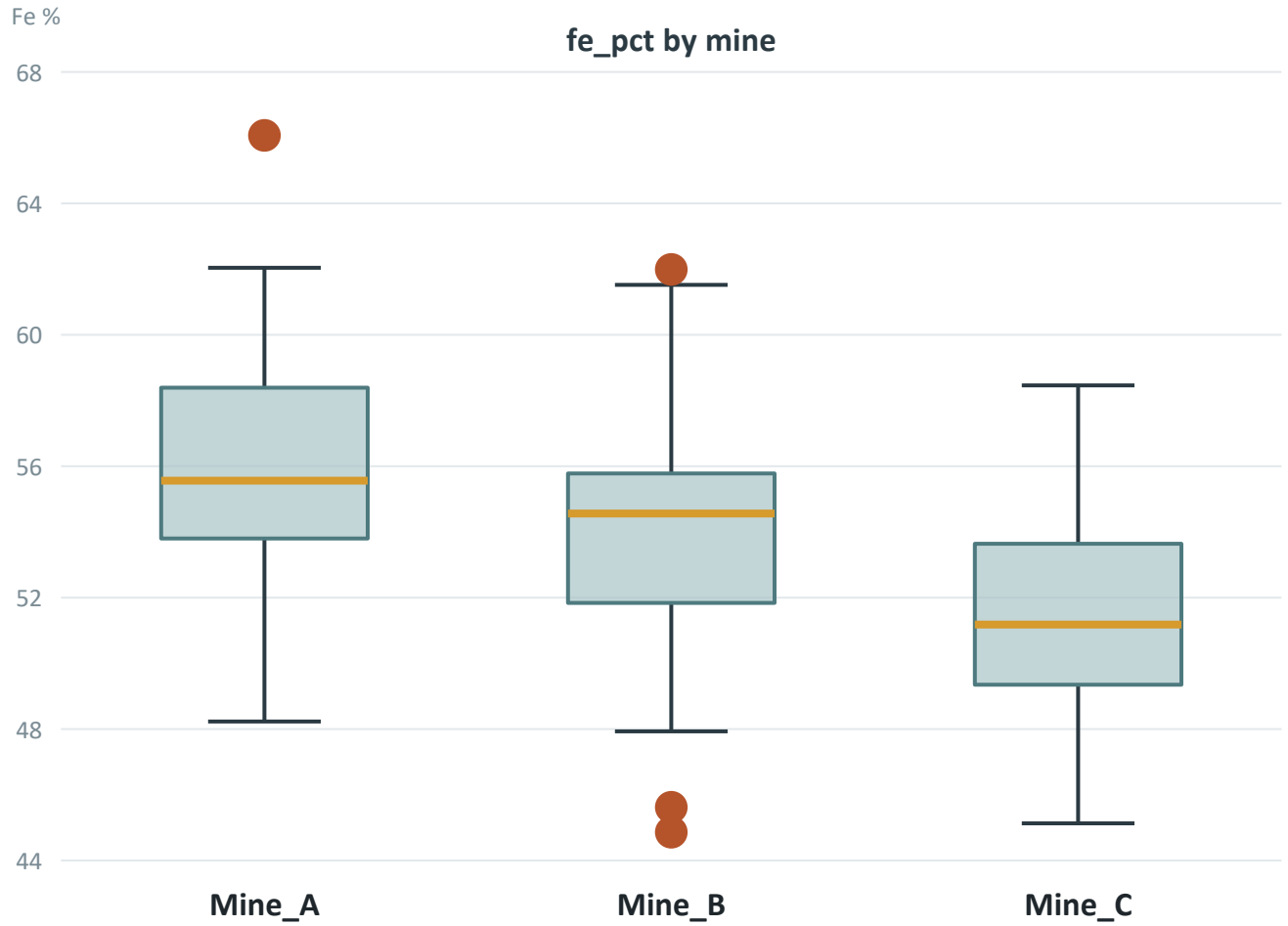
Line in the box
median

Box
Q1 to Q3 — the middle 50 %

Whiskers
the rest of the normal range

Dots
outliers - beyond 1.5 × IQR from the box

What this one shows: Mine_A has the richest samples, Mine_C the poorest.



Watch out: in Statistica's Box Plot dialog pick “Median / Quartiles / Range” - another option draws mean ± SD, which looks similar but means something else.

Same data, three tools, same answer

Excel
Formulas in cells

```
=AVERAGE(F2:F241)
```

→ 53.85

Python (Colab)
Code in a notebook

```
df["fe_pct"].mean()
```

→ 53.85

Statistica
Menus and dialogs

```
Statistics › Basic
Statistics › Descriptive
```

→ 53.85

The statistics don't change with the tool. What changes: how reproducible your work is, how fast you repeat it on new data, and how easy it is to make a mistake you can't see.

Python in 8 lines

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("ore_samples_clean.csv")

fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(14, 5))

# Left plot: Histogram
df["moisture_pct"].hist(bins=20, ax=ax1, edgecolor="black", grid=False)
ax1.set_title("Distribution of Moisture (%)")
ax1.set_xlabel("Moisture (%)")
ax1.set_ylabel("Frequency")

# Right plot: Boxplot
df.boxplot(column="fe_pct", by="mine", ax=ax2, grid=False)
ax2.set_title("Iron Percentage (Fe %) by Mine")
ax2.set_xlabel("Mine")
ax2.set_ylabel("Fe (%)")

plt.suptitle("")  # Clear automatic pandas suptitle
plt.tight_layout()
plt.show()
```

1. Open colab.research.google.com
2. New notebook (Google account)
3. Files panel → upload the CSV
4. Paste one line per cell, Shift + Enter

Before next lab

Statistica working
Installed and registered — or a plan with your Windows teammate

Read
Larose & Larose, Discovering Knowledge in Data, ch. 1 - or the “Data preparation” slides on UPeL

Download
ore_samples_messy.csv

NEXT · LAB 2

Exploratory data analysis

Real data is messy. You'll get the same ore samples - with typos, missing values, impossible depths and duplicates - and clean them.

Questions?

See you in Lab 2 - bring your laptop

sdziewic@agh.edu.pl

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

email/MS Teams

