Lab 1 worksheet: exploring ore samples

Statistical tools and data exploration for digitalization · AGH University of Krakow ·

How to work: in pairs, with a pen and the calculator on your phone. Write your answers on this sheet. We will compare answers at the end.

Names: _____

1. The dataset

You are analysts for a company with **three iron-ore mines**. The laboratory measured **240 drill-core samples** (80 per mine). Each row of the file `ore_samples_clean.csv` is one sample; each column is one variable.

This is a teaching dataset, modelled on a real iron-ore operation.

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 the surface	21.6 – 399.5 m
`fe_pct`	Iron grade: % of iron in the rock. Higher = more valuable ore	44.9 – 66.1 %
`sio2_pct`	Silica (quartz): the main impurity. Steel mills pay less for it	4.2 – 14.1 %
`moisture_pct`	Water content. You pay to transport and dry the water	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 processing plant actually recovers. This is money	43.4 – 77.4 %

A few rows of the data (12 of 240)

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
S081	Mine_B	Hematite	Night	52.9	55.83	5.61	3.74	0	65.20
S082	Mine_B	Magnetite	Night	25.4	55.76	10.07	8.89	4	60.69
S083	Mine_B	Magnetite	Night	158.9	52.67	8.56	4.32	8	65.20
S084	Mine_B	Magnetite	Night	160.3	48.89	10.38	3.43	3	64.12
S161	Mine_C	Mixed	Day	88.4	53.26	10.22	3.49	4	64.64
S162	Mine_C	Mixed	Day	312.1	46.29	12.39	7.87	2	52.57
S163	Mine_C	Hematite	Day	122.8	55.20	8.97	5.20	2	66.44
S164	Mine_C	Mixed	Day	305.7	52.38	9.11	6.95	5	56.99

2. Exercises

Exercise 1 · Variable types

For each variable, circle its type. Remember: **continuous** = any value in a range, **discrete** = whole-number counts, **nominal** = labels without order, **ordinal** = labels with an order, **binary** = exactly two values.

Variable	Type (circle one)
`sample_id`	continuous discrete nominal ordinal binary
`mine`	continuous discrete nominal ordinal binary
`rock_type`	continuous discrete nominal ordinal binary
`shift`	continuous discrete nominal ordinal binary
`depth_m`	continuous discrete nominal ordinal binary
`fe_pct`	continuous discrete nominal ordinal binary
`sio2_pct`	continuous discrete nominal ordinal binary
`moisture_pct`	continuous discrete nominal ordinal binary
`fracture_count`	continuous discrete nominal ordinal binary
`recovery_pct`	continuous discrete nominal ordinal binary

Which variable would a plant manager most want to predict, and why?

_____

Exercise 2 · Calculate by hand

Iron grade `fe_pct` of nine samples from Mine_C:

S161	S162	S163	S164	S165	S166	S167	S168	S169
53.26	46.29	55.20	52.38	49.14	56.95	55.28	51.91	50.73

a) Sort the values from smallest to largest:

_____

b) **Mean** = sum ÷ 9 = _____

c) **Median** (the middle value) = _____

d) **Minimum** = _____ **Maximum** = _____ **Range** = max − min = _____

e) **Quartiles.** Q1 = median of the 4 values below the median = _____ Q3 = median of the 4 values above it = _____

f) **IQR** = Q3 − Q1 = _____. What does it tell you, in one sentence?

_____

Exercise 3 · Mean or median?

Moisture `moisture_pct` of seven samples:

S170	S171	S172	S173	S174	S175	S176
5.38	3.98	4.51	10.86	5.71	4.60	3.70

a) Mean = _____ Median = _____

b) Which one is pulled away by a single value? Which sample is it?

c) A manager asks: "What is the typical moisture?" Which number do you give, and why?

Exercise 4 · Read the computer's output

This is what Statistica gives for **all 240 samples** (Statistics → Basic Statistics → Descriptive statistics):

Variable	N	Mean	Median	Min	Max	Q1	Q3	SD	Skewness
depth_m	240	210.49	203.55	21.60	399.50	114.05	306.05	113.65	0.02
fe_pct	240	53.85	54.28	44.86	66.07	51.29	56.35	3.65	−0.06
sio2_pct	240	9.29	9.24	4.24	14.11	8.05	10.48	1.84	−0.03
moisture_pct	240	5.58	5.20	1.85	15.27	4.11	6.72	2.02	1.04
fracture_count	240	3.19	3.00	0	9	2	5	2.05	0.55
recovery_pct	240	61.83	62.35	43.36	77.39	58.44	65.35	5.16	−0.23

a) Which variable is clearly skewed? Give two pieces of evidence from the table.

---

b) The SD of fe_pct is 3.65. Between which two values do about two-thirds of the samples lie? _____ to _____ %

c) Look at every minimum and maximum. Is any value physically impossible? Is any value suspicious?

---

d) Why is there no row for mine or rock_type in this table?

Exercise 5 · Frequency table

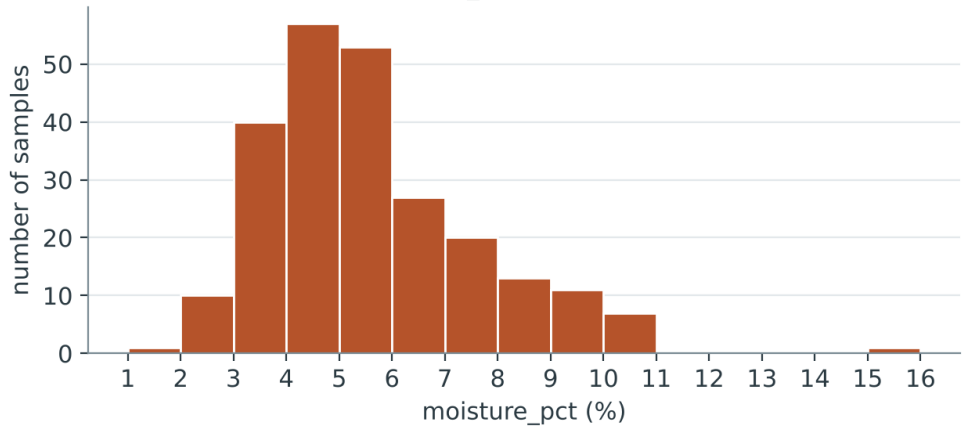
Fill in the percent column (count ÷ 240 × 100):

rock_type	Count	Percent	Cumulative %
Hematite	91		
Magnetite	77		
Mixed	72		
Total	240	100.0	

Why can't you calculate a mean of rock_type? _____

Exercise 6 · Read the graphs

Figure 1. Histogram of moisture_pct (240 samples)

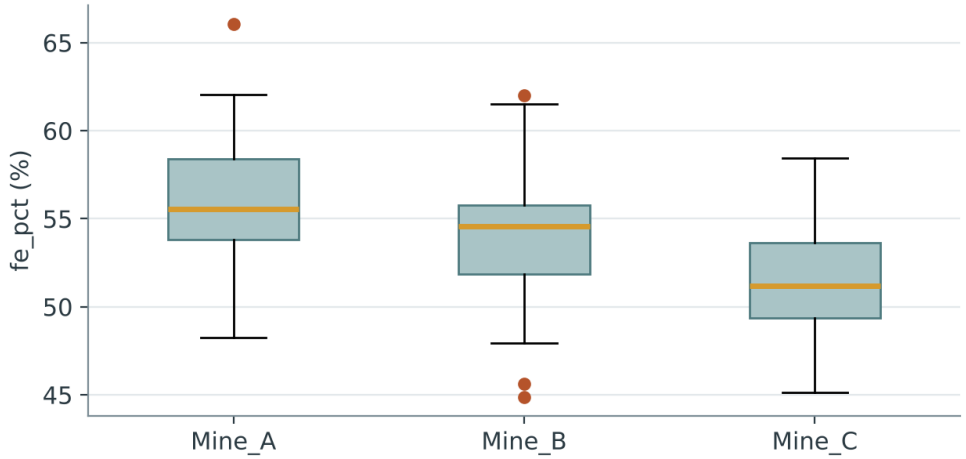


a) Where is the peak of the histogram (most common moisture)? _____ %

b) Is the shape symmetric? If not, on which side is the long tail? _____

c) One sample is far away from all the others. About what value is it? _____ %

Figure 2. Box plot of fe_pct by mine



d) Which mine has the richest ore (highest median)? _____ The poorest? _____

e) The box shows the middle 50 % of samples. Do the boxes of Mine_A and Mine_C overlap? _____

f) What do the dots mean? _____

g) If you could buy ore from only one mine, which would you choose, and what else would you want to know first?

Bonus · Think like an analyst

Write **one question** about these mines that the data could answer, e.g. "Does deeper ore contain less iron?"

Have a phone? Download the CSV and open it in the Google Sheets app. Try `=AVERAGE(F2:F241)` and `=MEDIAN(F2:F241)` on the fe_pct column.

Before next lab: install Statistica (AGH Panel → "Software" tab for the codes; installer on the AGH Cloud) or team up with a Windows user.