AIM learning

centria University of Applied Sciences ammattikorkeakoulu

kpedu

Kokkola Karleby

DATA ANALYSIS — THE BASICS

Learning Package– AIMLearning project

Teemu Voutilainen
RDI-Expert, Centria University of Applied Sciences

Euroopan unionin osarahoittama

Elinvoimakeskus

DATA ANALYSIS — THE BASICS

A self-study guide for complete beginners

This guide was produced with the assistance of a large language model.

Before you start

This guide teaches the core ideas of data analysis from zero. No math background is needed beyond school arithmetic, and no software is required to follow along — though a spreadsheet program (Excel, Google Sheets or LibreOffice) will let you try things yourself, which is strongly recommended.

Work through the sections in order; each builds on the previous one. Small **Exercise** boxes appear throughout. Do them — reading about analysis is like reading about swimming. Answers are at the end of the guide.

Throughout the guide we use one running example: a small kiosk that sells drinks and snacks in a park. Here are ten days of its sales:

Day	Weekday	Drinks sold	Temperature (°C)	Rain
1	Monday	40	15	no
2	Tuesday	48	16	no
3	Wednesday	56	18	no
4	Thursday	44	16	yes
5	Friday	60	19	no
6	Saturday	82	20	no
7	Sunday	74	19	no
8	Monday	30	13	yes
9	Tuesday	46	16	no
10	Wednesday	180	17	no

Keep a finger on this table — we will return to it many times. (Yes, day 10 looks strange. That is on purpose.)

1. WHAT DATA ANALYSIS ACTUALLY IS

Data analysis is **finding patterns in recorded observations and using them to make better decisions.** That is the whole field in one sentence. Everything else — the tools, the statistics, the charts, the machine learning — is machinery in service of that sentence.

Notice what the sentence does *not* say. It does not mention computers, programming, or mathematics beyond what you already know. Humans were analyzing data thousands of years before any of those existed. Ancient Egyptian priests noticed that the star Sirius appeared just before dawn at the same time the Nile began to flood — an observation, recorded and repeated over generations, that let them predict the flood that fed their civilization (Encyclopaedia Britannica). Babylonian astronomers turned centuries of sky-watching into calendars and astronomical tables of remarkable accuracy (Neugebauer 1969, 97–144), and the Maya built a calendar system that could count days far into the past and future and predict astronomical events such as the movements of Venus (Fraknoi, Morrison & Wolff 2022, section 4.4). A hunter who knows which valley the deer favor after the first frost is doing data analysis. So are you, when you notice that your bus is always late on Fridays and leave home earlier.

All of these follow the same loop, and so will everything in this guide:

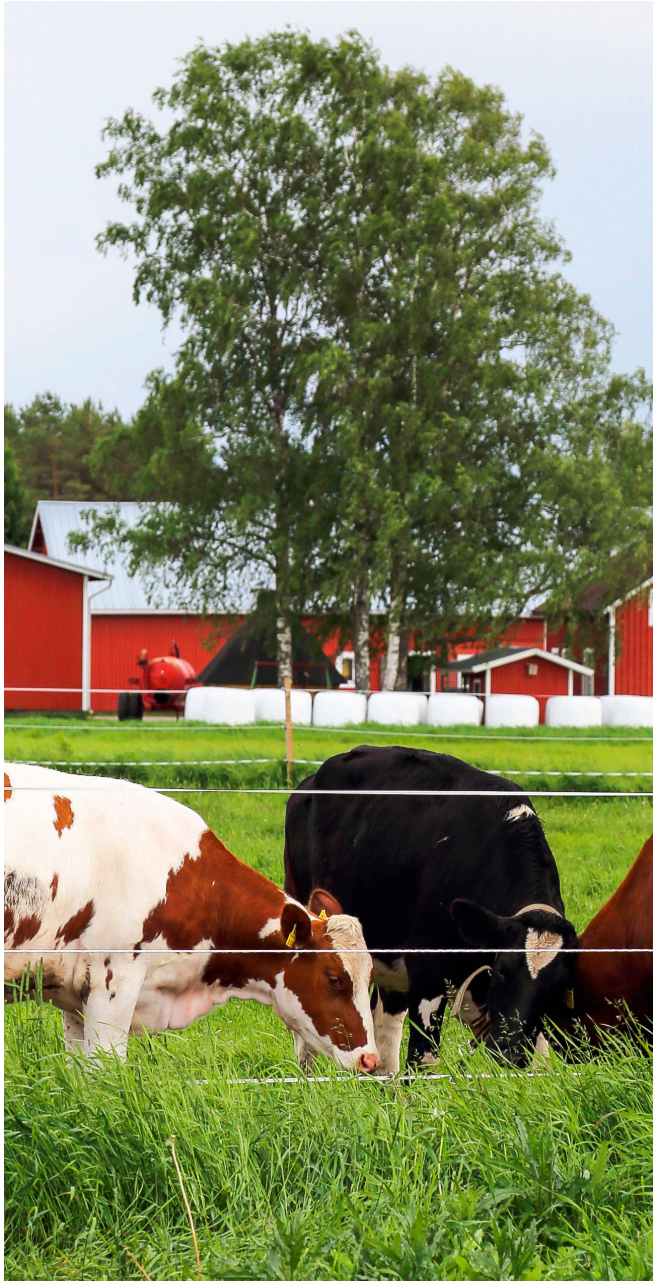
> ***observe → record → find the pattern → predict → decide***

Adobe Stock

Why it became a profession

If humans always did this, why is "data analyst" suddenly a job title? Two things changed. First, the amount of recorded data exploded: every purchase, click, sensor reading, and bus ride now leaves a trace. Second, computers made it cheap to process all of it. A dataset of seventeen thousand rows — trivial by modern standards — is already far beyond what anyone can eyeball.

This shifted where the expertise lives. The Egyptian priest was an expert on the Nile who happened to use data. The modern analyst is an expert on *met-*



Adobe Stock

hods who can move between bike rentals, hospital queues, and video-game economies, because the methods are the same everywhere. Knowledge of the topic (**called domain knowledge**) still matters enormously — we will see exactly why in Section 6 — but it is now shared between the analyst and the people they work with.

The five steps of any analysis

In practice, analysis projects move through five stages. You will do all five in miniature in this guide.

1. **Collect** — get the data: measure it, download it, export it from some system.
2. **Clean** — fix the mess: typos, missing values, duplicates, impossible numbers. Analysts joke that this is 80% of the work (Wickham 2014, 1 [Dasu & Johnson 2003]). It is only half a joke.
3. **Explore** — compute summaries, draw charts, hunt for patterns.
4. **Conclude** — test whether the patterns are real and decide what they mean.
5. **Communicate** — explain the finding so that someone can act on it. An analysis nobody understands changes nothing.

2. UNDERSTANDING DATA ITSELF

Before analyzing data you need to be able to *read* it. Nearly all data you will ever meet is arranged as a table, and tables have a grammar.

Each **row** is one observation: one day, one customer, one game match, one measurement. Each **column** is one variable: something measured about every observation. In the kiosk table, each row is a day, and the four variables are the weekday, drinks sold, temperature, and rain. This “rows are things, columns are properties of things” arrangement is called **tidy data**, and getting messy data into this shape is a large part of real analytical work (Wickham 2014, 4–5).

Adobe Stock

Types of variables

Not all columns are the same kind of thing, and the type determines what you can meaningfully do with it.

Numeric variables are quantities: temperature, drinks sold, price, age. You can add and average them. Some are *continuous* (temperature can be 16.4) and some are *discrete* counts (you cannot sell 16.4 drinks).

Categorical variables are labels: weekday, city, favorite genre. Averaging them is meaningless — there is no "average weekday" — but you can *count* them and *group by* them, which turns out to be one of the most powerful moves in analysis.

Ordinal variables are categories with an order: survey answers like bad / okay / good, or clothing sizes S / M / L. The order carries information, but the gaps between steps are not necessarily equal.

Boolean (yes/no) variables, like our Rain column, are categories with exactly two values. They are often written as 1 and 0, which enables a neat trick: the *average* of a 1/0 column is the *share* of yeses. If ten days have three rainy ones coded as 1, the mean is 0.3 — that is, 30% of days were rainy.

Dates and times deserve respect: they look simple but hide structure (weekday, month, season, holiday) that often explains more than any other variable.

One more distinction you will hear constantly: the **population** is everything you care about (all days the kiosk will ever be open); a **sample** is the part you actually have (these ten days). Everything we compute describes the sample, and we quietly hope it resembles the population. The bigger and more representative the sample, the safer that hope. Ten days is a very small sample — real conclusions would want a full season or more.

EXERCISE 1.

In the kiosk table, classify each of the four variables (weekday, drinks sold, temperature, rain) by type. Then: can you compute an "average weekday"? What can you do with the weekday column instead?

3. DESCRIBING DATA: THE HONEST SUMMARY

You cannot show anyone 17,000 rows, and even our 10-row table takes effort to absorb. The first job of analysis is to **summarize**: reduce many numbers to a few that tell the story. The classic summaries answer two questions — where is the *middle*, and *how spread out* are the values?

The middle: mean, median, mode

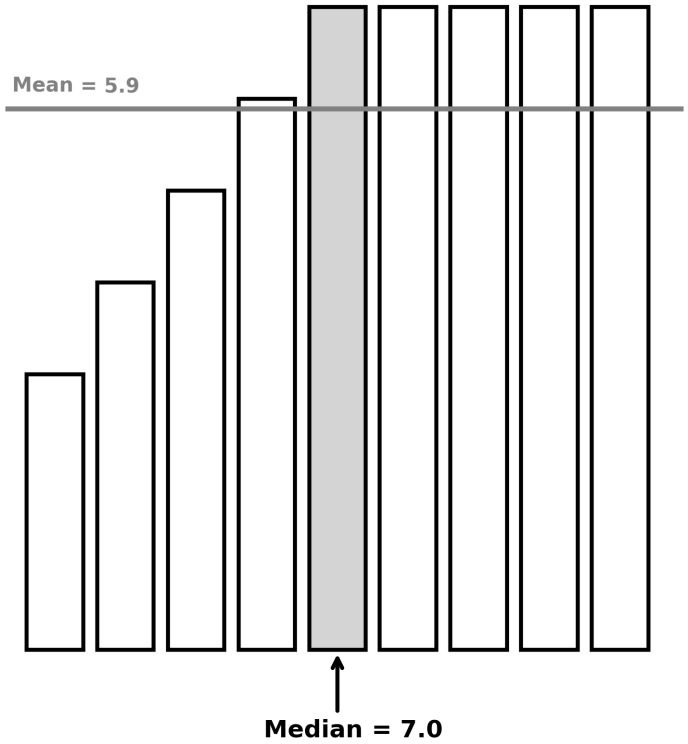
The **mean** (everyday "average") is the sum divided by the count. For drinks sold: 40+48+56+44+60+82+74+30+46+180 = 660, and 660 ÷ 10 = **66 drinks per day**.

Stop and look at that number against the table. Sixty-six is more than the kiosk sold on eight of the ten days. As a "typical day" it feels wrong — and it is, because day 10's freakish 180 dragged the mean upward. The mean is democratic in the worst way: every value votes, and extreme values shout.

The **median** is the middle value when you sort the data. Sorted, our drinks are 30, 40, 44, 46, 48, 56, 60, 74, 82, 180; with ten values the median is the average of the 5th and 6th, (48+56)/2 = 52. Notice how calm the median is: replace 180 with 1,800 and it does not move at all. The median is your defense against extreme values.

The **mode** is the most common value — most useful for categories ("the most common weather was: no rain").

The gap between mean and median is itself information. When you read that "average salary" in some group is surprisingly high, check the median: a handful of enormous salaries pull the mean up while the median stays put. Any honest description of skewed data (salaries, house prices, social-media followers) reports the median.



Mean: How high are the bars on average?
Median: Which bar is in the middle, and how high is it?

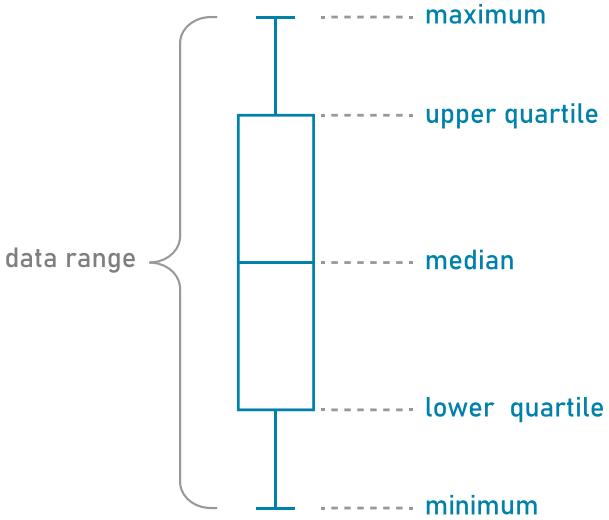
The spread: range and standard deviation

Two kiosks can both average 66 drinks — one selling 60–70 every day, the other swinging between 10 and 150. Same middle, utterly different businesses. Spread matters.

The **range** (max − min) is the crudest measure: 180 − 30 = 150 for our kiosk. Simple, but it depends entirely on the two most extreme values.

The **standard deviation** measures, roughly, how far a typical day sits from the mean. You will rarely compute it by hand (every tool has it built in: =STDEV.S(...) in Excel), but you should be able to read it: a mean of 66 with a standard deviation of 43 (our kiosk) describes a wildly variable business; a mean of 66 with a standard deviation of 5 describes a boringly steady one. As a rule of thumb, in many datasets about two-thirds of values fall within one standard deviation of the mean.

Percentiles generalize the median: the 25th percentile is the value below which a quarter of the data falls, the 75th percentile three-quarters. The median is the 50th. Together the 25th–75th range (**the interquartile range**) tells you where the middle half of the data lives — for our kiosk, roughly 45 to 70 drinks.



Adobe Stock

Box-and-whisker plot showing the five-number summary of a dataset: minimum, lower quartile (Q1), median, upper quartile (Q3), and maximum.

EXERCISE 2.

Compute the mean and median temperature for the ten days. Are they close together or far apart? What does that tell you, given what you just read about day 10's sales?

4. PICTURES BEAT TABLES: VISUALIZATION

Human eyes are pattern-finding machines, but only when information is shaped for them. A column of numbers hides its story; the right chart makes the story impossible to miss. Each basic chart type answers a different question, and choosing the right one is half the skill.

The bar chart answers "*how do categories compare?*" Categories on one axis, a numeric summary on the other: average sales per weekday, players per game genre. Bars invite comparison, which is why they should normally start at zero — more on that below.

The line chart answers "*how did it change over time?*" Time flows left to right; the line's shape reveals trends, seasons, and sudden breaks. Two years of daily bike-rental data drawn as a line instantly shows a summer-winter wave that no table would reveal (Fanaee-T 2013).

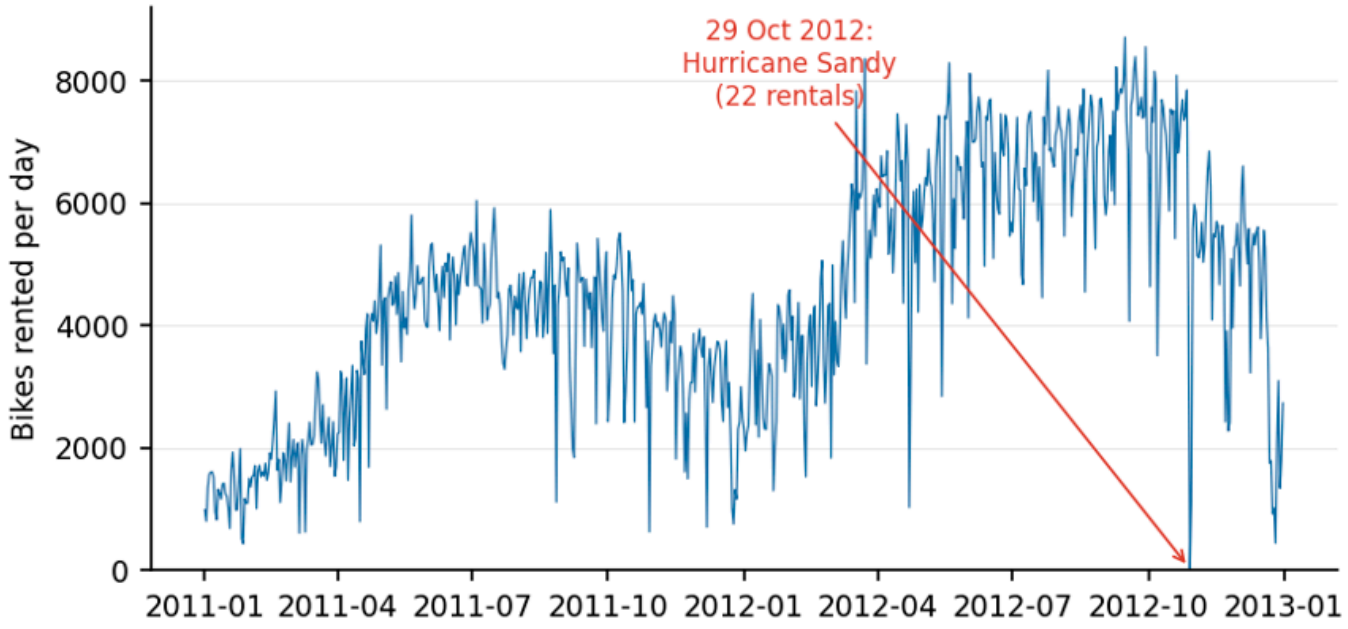


FIGURE 1. Daily bike rentals in Washington, D.C., 2011–2012, drawn as a line chart: the summer–winter wave is immediate, and one day collapses to 22 rentals — Hurricane Sandy, discussed in Section 7 (data: Fanaee-T 2013).

The scatter plot answers "*are these two numbers related?*" Every observation becomes one dot, positioned by its two values — for the kiosk, temperature on the horizontal axis, drinks sold on the vertical. If the dots drift upward from left to right, warm days sell more. The scatter plot is the analyst's favorite tool, because it shows *every single observation* with no summarizing at all: outliers, clusters, curves and gaps are all visible at once.

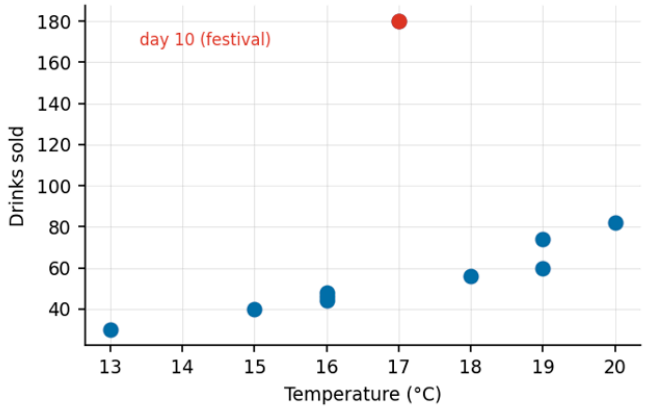


FIGURE 2. Scatter plot of the kiosk data: temperature against drinks sold. The upward drift is visible, and day 10 stands apart from the pattern.

The histogram answers "*how are the values of one variable distributed?*" It chops the range into bins and counts observations per bin, revealing whether values cluster in one hump, two humps, or trail off in a long tail. That long tail is exactly the skewness that makes means misleading — a histogram of incomes shows immediately why economists prefer the median.

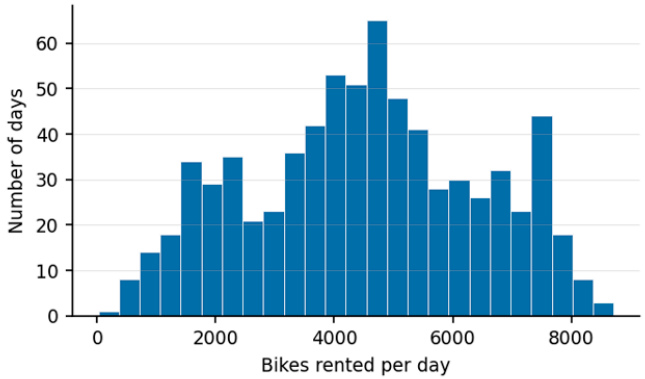


FIGURE 3. Histogram of the same 731 days of bike rentals as in FIGURE 1: most days cluster around the middle, with tails toward quiet and busy days (data: Fanaee-T 2013).

Draw first, summarize second. A famous teaching dataset, Anscombe's quartet, contains four datasets with identical means and correlations that look utterly different when plotted — one is a clean line, one a curve, one wrecked by a single outlier (Anscombe 1973, 19–20). Summary numbers can lie by omission; a good plot rarely does.

How charts lie

Because charts are persuasive, they are also the favorite weapon of people who want to mislead you (Huff 1954). Three classics to watch for, both in others' charts and your own:

The truncated axis. A bar chart whose vertical axis starts at 95 instead of 0 makes a difference between 96 and 99 look enormous. Bars encode quantity by length, so cutting the axis is visual lying — sadly common in news graphics and sales decks.

The cherry-picked window. "Prices have fallen since January!" — shown on a chart that conveniently starts in January, at an unusual peak. Always ask what happened before the left edge.

The double y-axis. Two lines, two different scales, stretched until they seem to move together. With freedom to scale each axis independently, you can make almost any two variables appear related.

The test for your own charts is simple: would the impression survive if the reader saw the full, honest picture? If not, fix the chart, not the caption.

Adobe Stock

EXERCISE 3.

For each question, name the best chart type: (a) Did the kiosk sell more on weekends than weekdays? (b) Is temperature related to sales? (c) What does a typical day's sales look like, and are there weird days? (d) How did sales develop over the ten days?

5. COMPARING GROUPS

Here is where analysis starts producing real answers. Most practical questions are comparisons in disguise: *Do rainy days hurt sales? Do players spend more after the update? Which store layout sells more?* The move is always the same — **split the rows into groups by a categorical variable, then compute the same summary for each group.**

For the kiosk: on the two rainy days, sales were 44 and 30, mean **37**. On the eight dry days, the mean is (40+48+56+60+82+74+46+180)/8 = **73.25**. Dry days appear to sell twice as much. In spreadsheet terms this is one function: =AVERAGEIF(rain_column, ”yes”, sales_column) — filtering and averaging in a single move. Analysts do this so often that every tool has it built in (in Python it is called groupby).

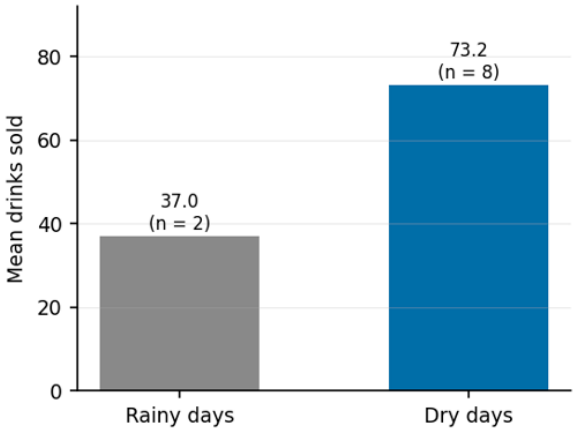


FIGURE 4. Bar chart of the kiosk group comparison: mean drinks sold on rainy versus dry days. The group sizes are part of the picture — two rainy days is a very thin basis for a conclusion.

Three cautions make this simple move trustworthy.

Check group sizes. Our rain group has *two* days. Two! Any average of two values deserves deep suspicion — one odd day swings it completely. Always report how many observations each group contains; “n = 2” is a health warning on any conclusion.

Check what else differs between groups. Our rainy days were also cool days (13°C and 16°C, at the cold end of the range). So is it rain that hurts sales, or cold? With this data we cannot tell — the two explanations are tangled together. This tangling has a name, **confounding,** and it is the central villain of the next section.

Beware of averages of averages. If Saturday’s average comes from 50 observations and Monday’s from 3, averaging those two numbers as if equal is wrong. When groups have very different sizes, combined results can even reverse direction — a phenomenon called Simpson’s paradox (Simpson 1951). You don’t need to master it now; you need to remember that *group sizes matter*.

EXERCISE 4.

Compute the mean sales for weekend days (Saturday + Sunday) versus weekdays in the kiosk table. Before trusting the result: how many observations are in each group, and does anything else differ between the groups besides being the weekend?

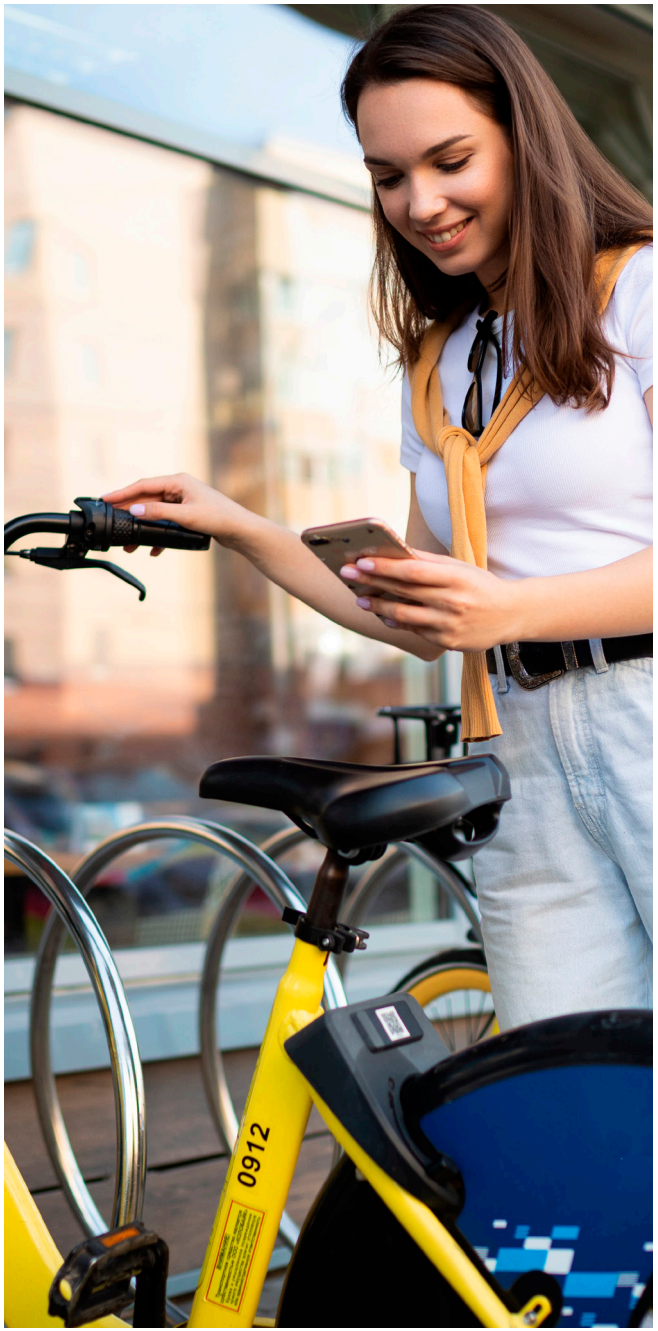
6. RELATIONSHIPS: CORRELATION AND ITS FAMOUS TRAP

Group comparison works when one variable is categorical. When *both* variables are numeric — temperature and sales, hours played and money spent — we ask instead: *do they move together?* This is **correlation**.

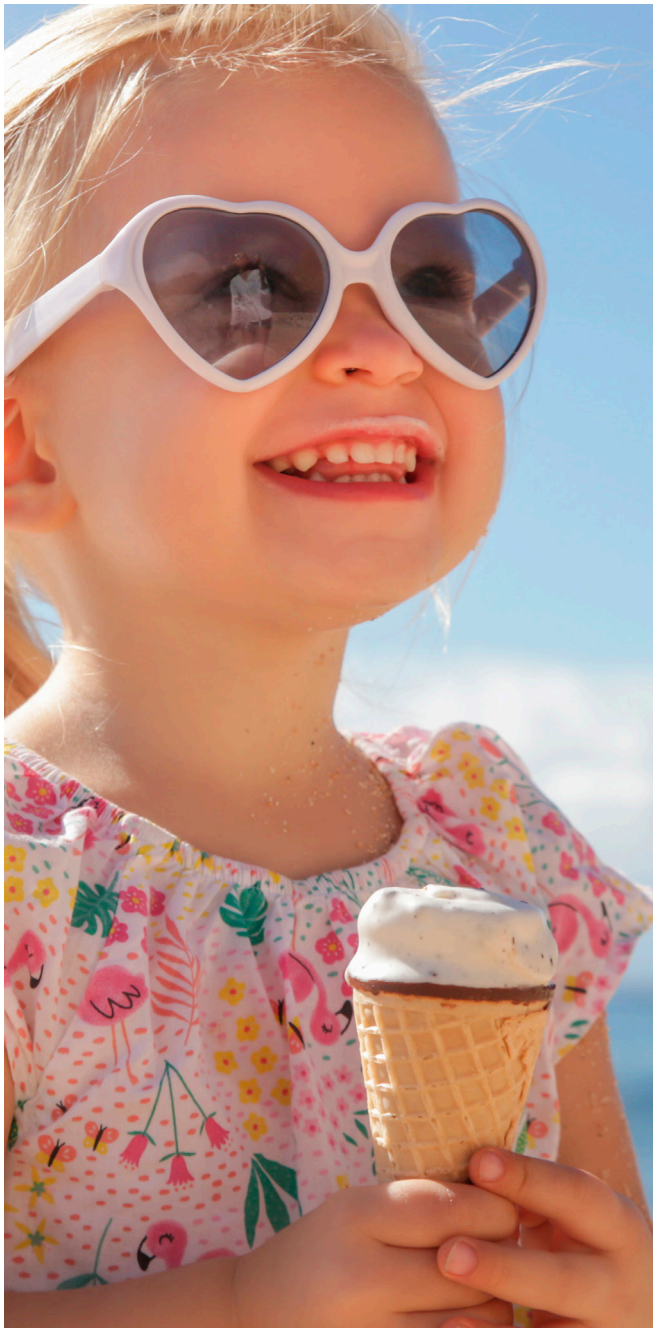
Start with the picture. Plot the dots (a scatter plot); if they drift upward together, the correlation is positive; downward, negative; a shapeless cloud, near zero. Then, if you want a number, the **correlation coefficient r** compresses that drift into a single value between −1 and +1. In Excel it is =CORREL(column1, column2); every analysis tool has it.

Reading r takes calibration, because textbook examples miscalibrate people. Real-world data is noisy: an r of ±0.9 or higher almost never happens outside physics labs; ±0.5 to ±0.7 is a strong, valuable signal in messy human data (temperature and bike rentals across two years: 0.63); ±0.2 to ±0.4 is weak but sometimes meaningful; and near 0 means no *linear* relationship. For the kiosk's temperature and sales, r ≈ 0.36 — but excluding the strange day 10, it jumps to about 0.95 (a reminder that single outliers can wreck summary numbers; more in Section 7).

That word *linear* hides a trap: r only measures straight-line relationships. Bike rentals rise steadily with temperature — until about 30°C, where they fall again because it is too hot to pedal. The true relationship is an arch, but r, forced to summarize the arch as a line, reports a mediocre value and hides the most interesting fact. The cure is the same as ever: **look at the scatter plot first, compute second.**



Adobe Stock



Adobe Stock

Correlation is not causation

Now the most famous sentence in all of data analysis. Suppose you discover that ice-cream sales correlate strongly with drowning deaths, month by month. Does ice cream cause drowning? Should we ban it? Obviously not: *summer* causes both — more ice cream eaten, more people swimming. Summer is a **confounder**: a third variable driving both of the ones you measured, manufacturing a correlation between them.

When you find that A correlates with B, there are always at least four possibilities: A causes B; B causes A (the direction is reversed — maybe stores that already sell a lot can afford better locations, rather than locations causing sales); some C causes both; or it is coincidence. With enough variables, coincidences are guaranteed — entire websites collect absurd "spurious correlations" like per-capita cheese consumption tracking deaths by bedsheet entanglement (r = 0.95 over a decade, and meaning nothing) (Hrabač & Trkulja 2020, 293 [Vigen 2015]).

Correlation is still precious: it tells you *where to look*. It generates suspects, not verdicts. Establishing causation needs stronger tools — the gold standard is the **controlled experiment**, where you *change* one thing for a random half of your subjects and compare (websites run these constantly as "A/B tests": half of users see the new button, half the old, and randomness guarantees no confounder separates the groups).

This is also where domain knowledge earns its keep. The data alone cannot tell you that summer explains the ice-cream–drowning link — *you* have to know the world well enough to think of it. An analyst who knows only methods and no context will confidently report nonsense.

EXERCISE 5.

A study finds that children's shoe size strongly correlates with reading ability. What is the confounder? And a harder one: your kiosk's sales correlate with the number of parked bicycles outside the park. List the four possible explanations.tä.

7. OUTLIERS, MISSING DATA, AND THE ART OF CLEANING

Time to deal with day 10 and its 180 drinks — well over two standard deviations above the mean. A value far outside the pattern of the rest is an **outlier**, and what you do with it is one of the most consequential judgment calls in analysis.

The first rule: **an outlier is a question, not an error**. There are three possibilities, and each demands a different response. It might be *a mistake* — a typo (18 became 180), a broken sensor, a double-counted register. Investigate; if confirmed wrong, fix or remove it, and write down that you did. *It might be real but irrelevant* — a one-off festival day, meaningful in itself but poisonous if your goal is to describe normal operations; you might analyze it separately and exclude it from the "typical day" summaries, again documenting the choice. Or it might be **the discovery itself**. A famous real example: in two years of Washington D.C. bike-share data, one October day shows 22 rentals against a norm of about 5,000. That outlier is Hurricane Sandy, recorded by accident in a bicycle dataset. (Fanaee-T & Gama 2014, section 3.2.) Delete it as a "data error" and you delete the most interesting fact the data contains.

What you must never do is silently delete outliers because they spoil a tidy conclusion. That is not cleaning; that is fraud with extra steps.

Missing data is the other everyday nuisance: empty cells, "N/A", or — far worse — missing values disguised as real ones (0 where nothing was recorded, or placeholder nonsense like a temperature of −999). Disguised missing values silently poison every average computed over them, which is why step one with any new dataset is the same: compute the minimum and maximum of every column, and stare at anything impossible. Whether to drop incomplete rows or fill gaps with estimates depends on *why* the data is missing — a sensor that fails randomly is an annoyance; a sensor that fails precisely during storms deletes exactly the most interesting rows, biasing everything you compute afterward.

The general lesson: real data is dirty, cleaning it is most of the work, and every cleaning decision is an analytical decision that can change the conclusion. Keep the raw data untouched, work on a copy, and keep a note of every change — future-you, redoing the analysis in three months, will be grateful.

EXERCISE 6.

Suppose you learn day 10 was the day of a big park festival. Give the mean and median of drinks sold without day 10, and compare them to the values from Section 3 (mean 66, median 52). Which summary was more robust to the outlier? Should day 10 be deleted from the dataset entirely?

8. FROM PATTERNS TO PREDICTIONS

Everything so far describes the past. The payoff comes from pointing it at the future: *tomorrow is a dry Wednesday, 18°C — how many drinks should the kiosk stock?* The moment you answer, you have built a **model**: a simplified, mathematical description of a pattern, used to make predictions.

Build one now, in three steps of increasing refinement, using only ideas you already have. The crudest prediction is the overall typical day: median 52 (we exclude festival-day 10 from all of this — a cleaning decision, documented). Refine it with a group comparison: tomorrow is dry, and dry days (without day 10) average about 58. Refine again with correlation: 18°C sits toward the warm end of our range, and warmer days sell more, so nudge upward — perhaps 60–65. That reasoning chain — start from a baseline, adjust for the conditions you know — is what formal methods like *linear regression* do with more rigor: they find the exact line through the temperature-sales scatter and read the prediction off it at 18°C.

Two ideas separate people who use models well from people who get burned by them.

Predictions are always uncertain, and honest ones say how uncertain. "About 60" and "somewhere between 45 and 80, most likely near 60" are different claims; the second is more useful *because* it admits what it does not know. Even a rough range beats a falsely confident point. Weather forecasts learned this long ago — "70% chance of rain" is a model being honest.

A model can fit the past too well. With enough flexibility you can build a model that explains your ten days *perfectly* — day 3 sold 56 because it was a dry Wednesday at 18°C — and predicts day 11 terribly, *because* it memorized noise instead of learning the pattern. This failure is called **overfitting**, and it is why serious analysts always test a model on data it has never seen before judging it. If someone shows you a model with a perfect track record, your first question should be: perfect on data it was trained on, or on genuinely new data?

Machine learning, for all its glamour, is this same loop — baseline, pattern, prediction, error-checking — automated and scaled to millions of rows and thousands of variables. If you understand the kiosk model, you understand the skeleton of the thing.

EXERCISE 7.

Using the same three-step logic, predict sales for a rainy Monday at 13°C. Which historical days most resemble it? How confident are you, and why should you not be very?

9. COMMUNICATING RESULTS

An analysis only matters if someone acts on it, and people act on what they understand. This final skill is underrated because it looks easy. It is not.

State the finding first, in one plain sentence a non-analyst could repeat: *"Dry days sell roughly twice as much as rainy days, so we should staff and stock by the weather forecast."* Note the shape: pattern, then consequence. Lead with it — don't force your audience to sit through your whole journey of exploration before hearing the point. Then show one honest chart that makes the finding visible (usually the one that convinced *you*), state the numbers behind it, and be candid about limitations: ten days of data, two rainy days, rain tangled with cold. Admitting limits does not weaken an analysis — it is what makes the rest believable, and it is what separates analysis from advertising.

A practical checklist for any result you are about to share: Can I say it in one sentence? Does the chart survive the honesty tests from Section 4? Did I report group sizes? Did I check for confounders and say which ones remain? Would I bet lunch money on this holding up in new data?

Adobe Stock

10. YOUR TOOLBOX, AND WHERE TO GO NEXT

Every idea in this guide maps to a one-liner in the two tools you are most likely to meet. Use this table as a bridge whenever you sit down to practice.

Nobody memorizes these — analysts look syntax up constantly. The skill you have been building in this guide is knowing *what to ask*; the syntax is just spelling.

Idea (section)	Excel / Sheets	Python (pandas)
Mean (3)	=AVERAGE(B2:B11)	df['sales'].mean()
Median (3)	=MEDIAN(B2:B11)	df['sales'].median()
Spread (3)	=STDEV.S(B2:B11)	df['sales'].std()
Min / Max (7)	=MIN(...), =MAX(...)	df['sales'].min(), .max()
Count rows (2)	=COUNT(B2:B11)	len(df)
Group average (5)	=AVERAGEIF(D2:D11,"yes",B2:B11)	df.groupby('rain')['sales'].mean()
Correlation (6)	=CORREL(B2:B11,C2:C11)	df['sales'].corr(df['temp'])
Scatter plot (4)	Insert → Scatter chart	df.plot.scatter(x='temp',y='sales')
Sort / find outliers (7)	Data → Sort	df.sort_values('sales')

To keep going: practice on real datasets from **kaggle.com** (thousands of free ones, from football matches to Spotify songs — pick a topic you already care about, because domain knowledge makes analysis dramatically easier and more fun). Run Python free in the browser at **colab.google** with nothing to install. And when you meet statistics proper — hypothesis tests, confidence intervals, regression — you will find they are careful, formal versions of questions you can now already ask: *is this difference real or luck? how sure are we? what does the line through the scatter predict?*

MINI-GLOSSARY

Correlation — the degree to which two numeric variables move together, from −1 to +1.

Confounder — a hidden third variable driving two others, faking a relationship between them.

Domain knowledge — understanding of the real-world topic behind the data.

Distribution — the overall shape of a variable's values (see histogram).

Mean / median / mode — three versions of "the middle": arithmetic average / middle value when sorted / most frequent value.

Model — a simplified description of a pattern, used for prediction.

Outlier — an observation far outside the pattern of the rest.

Overfitting — a model memorizing its training data instead of learning the pattern, failing on new data.

Sample vs. population — the data you have vs. everything you would want to know about.

Standard deviation — typical distance of values from their mean.

Tidy data — one observation per row, one variable per column.

ANSWERS TO THE EXERCISES

1. Weekday: categorical. Drinks sold: numeric (discrete count). Temperature: numeric (continuous). Rain: boolean/categorical. There is no average weekday — but you can group by weekday and average the *sales* within each group, which is Section 5's whole trick.

2. Temperatures: 15+16+18+16+19+20+19+13+16+17 = 169, mean 16.9. Sorted middle pair is 16 and 17 → median 16.5. Mean and median nearly agree, so the *temperature* data has no wild outlier — day 10's sales were extreme, but its temperature (17°C) was ordinary. Outliers live in specific columns, not whole rows.

3. (a) Bar chart (categories compared). (b) Scatter plot (two numeric variables). (c) Histogram (distribution and weird days). (d) Line chart (change over time).

4. Weekend: (82+74)/2 = 78. Weekdays: (40+48+56+44+60+30+46+180)/8 = 63. Weekend looks higher — but n = 2 for weekends, and the weekday group contains the festival outlier, which pulls its mean up; without day 10 the weekday mean drops to about 46.3, making the weekend effect look much larger. Small groups and unhandled outliers can both flip a conclusion.

5. Shoe size and reading: the confounder is *age* — older children have bigger feet and read better. Kiosk and bicycles: (1) bikes bring customers (A→B); (2) people who came to buy drinks parked their bikes (B→A); (3) sunny weather causes both cycling and thirst (confounder C); (4) coincidence across only ten days. Note that (1), (2) and (3) could all be partly true at once.

6. Without day 10: sum 480, nine days, mean ≈ 53.3 (was 66 — moved by almost 13). Median of the nine sorted values is 48 (was 52 — barely moved). The median was far more robust. Day 10 should *not* be deleted from the dataset — it is real and interesting (analyze festivals separately!) — but it should be excluded from summaries that claim to describe a *normal* day, with the exclusion stated.

7. The most similar day is day 8 (rainy Monday, 13°C): 30 drinks. The other rainy day (44) and the general "rain roughly halves sales" pattern support a prediction around 30–40. Confidence should be low: exactly two rainy days in the data, rain confounded with cold, and Mondays are underrepresented. A wide range, honestly stated — "probably 25–45" — is the professional answer.

REFERENCES

Anscombe, F. J. 1973. Graphs in statistical analysis. *The American Statistician*, 27(1), 17–21. Saatavissa: https://doi.org/10.1080/00031305.1973.10478966. Viitattu 9.7.2026.

Dasu, T. & Johnson, T. 2003. *Exploratory Data Mining and Data Cleaning.* Hoboken: John Wiley & Sons.

Encyclopaedia Britannica. *Sirius.* Saatavissa: https://www.britannica.com/place/Sirius-star. Viitattu 9.7.2026.

Fanaee-T, H. 2013. *Bike Sharing.* Aineisto. UCI Machine Learning Repository. Saatavissa: https://doi.org/10.24432/C5W894. Viitattu 9.7.2026.

Fanaee-T, H. & Gama, J. 2014. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence,* 2(2–3), 113–127. Saatavissa: https://doi.org/10.1007/s13748-013-0040-3. Viitattu 9.7.2026.

Fraknoi, A., Morrison, D. & Wolff, S. 2022. *Astronomy 2e.* Houston: OpenStax, Rice University. Saatavissa: https://openstax.org/books/astronomy-2e/pages/4-4-the-calendar. Viitattu 10.7.2026.

Hrabač, P. & Trkulja, V. 2020. Of cheese and bedsheets – some notes on correlation. *Croatian Medical Journal,* 61(3), 293–295. Saatavissa: https://doi.org/10.3325/cmj.2020.61.293. Viitattu 10.7.2026.

Huff, D. 1954. *How to Lie with Statistics.* New York: W. W. Norton.

Neugebauer, O. 1969. *The Exact Sciences in Antiquity.* 2. painos. New York: Dover Publications.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society,* Series B, 13(2), 238–241.

Vigen, T. 2015. *Spurious Correlations.* New York: Hachette Books.

Wickham, H. 2014. Tidy data. *Journal of Statistical Software,* 59(10), 1–23. Saatavissa: https://doi.org/10.18637/jss.v059.i10. Viitattu 9.7.2026.