

Data, Data Storage, Data Collection

Lecture 8: From Data Analysis to Communication

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

Introduction

Recap: Data Storage

- ① **A format is a choice on three layers:** data model, physical organization, encoding.
- ② **Analyzing (OLAP):** columns and Parquet; the format matters at the boundary.
- ③ **Operating (OLTP):** hashing finds a record by its key; a DBMS answers the four failures with constraints and transactions (ACID).
- ④ **Keeping:** open, documented formats, checksums, and planned retention and erasure.
- ⑤ **No storage is best for every workload.**

Session Objectives

By the end of this session, you should be able to:

Analysis

- Explain what data analysis is and its place in the lifecycle.
- Uncover insights with summary statistics and visualizations.
- Choose visual encodings suited to each feature type.

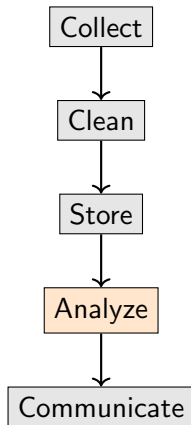
Interpretation

- Distinguish an observation from an interpretation.
- Check what survives noise, confounding and cleaning choices.

Communication

- Explain the role of communication within the data lifecycle.
- Structure a data-driven narrative around a clear insight.
- Identify and avoid pitfalls in misleading or unethical visuals.

Within The Lifecycle



What Is Data Analysis?

Data Analysis

Transformation of **cleaned data** into **useful information** to:

- Draw conclusions.
- Support decision-making.

It is driven by **questions**, not tools.



Descriptive

What happened?



Diagnostic

Why did it happen?



Predictive

What will happen?



Prescriptive

What should we do?

What Is Data Analysis?

Data Analysis

Transformation of **cleaned data** into **useful information** to:

- Draw conclusions.
- Support decision-making.

It is driven by **questions**, not tools.



Descriptive

What happened?



Diagnostic

Why did it happen?



Predictive

What will happen?



Prescriptive

What should we do?

Exploratory Data Analysis (EDA)

“Exploratory data analysis is **actively incisive**, rather than passively descriptive, with real emphasis on the **discovery of the unexpected**.” – John Tukey.

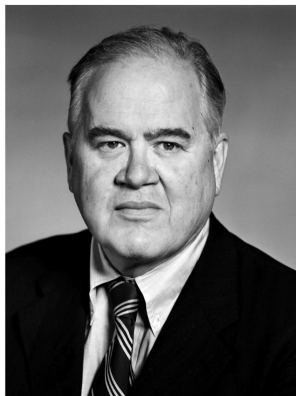


Photo Credit: Princeton University, Robert Matthews

John Tukey (1915–2000)

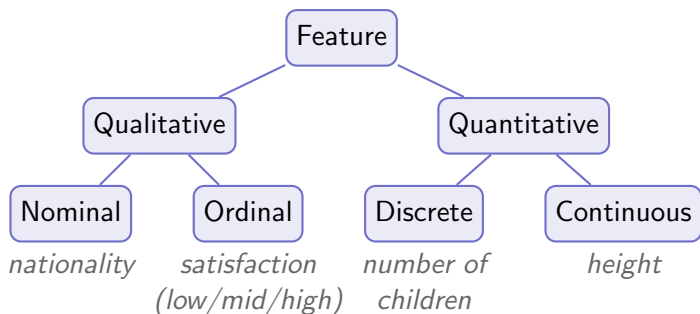
Questions Driving EDA:

- 🔍 How is Feature X distributed?
- 🔗 How do X and Y relate?
- 📊 Does X behave differently across Z?
- ⚠️ Any unusual values in X?
- ✂️ Are transformations helpful?

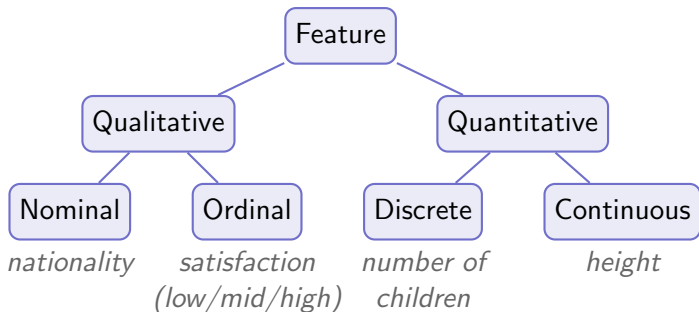
Summary statistics and visualizations

Describing Data

Feature Types



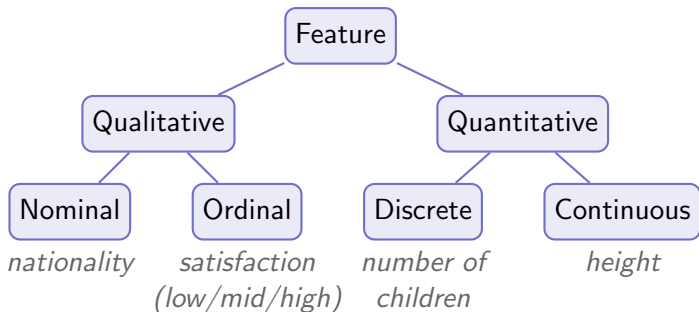
Feature Types



Why feature types matter in EDA?

- Determine **summaries**: mean for heights; not for nationalities.

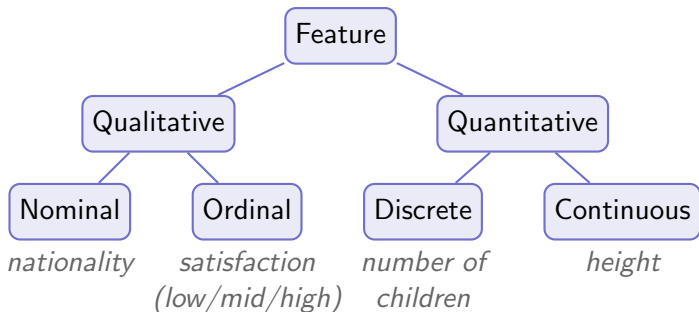
Feature Types



Why feature types matter in EDA?

- Determine **summaries**: mean for heights; not for nationalities.
- Guide **visualizations**: histogram vs bar plot.

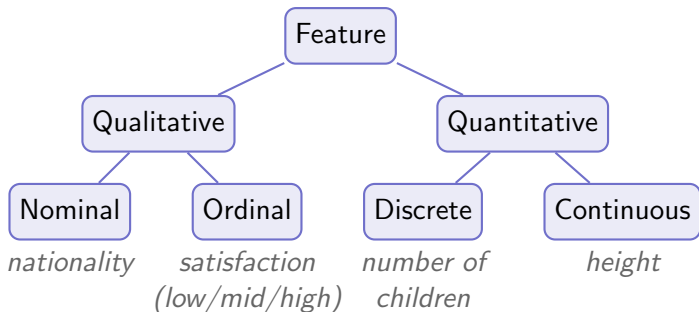
Feature Types



Why feature types matter in EDA?

- Determine **summaries**: mean for heights; not for nationalities.
- Guide **visualizations**: histogram vs bar plot.
- Constrain **transformations**: log-transform only for positive quantitative values.

Feature Types



Why feature types matter in EDA?

- Determine **summaries**: mean for heights; not for nationalities.
- Guide **visualizations**: histogram vs bar plot.
- Constrain **transformations**: log-transform only for positive quantitative values.
- Shape **interpretation**: meaning of the feature

Feature Type vs Data Type vs Encoding

Feature type

(what the values **mean**)

- Nominal
- Ordinal
- Quantitative

Data type

(language manipulations)

- int
- float
- str

Encoding

(encoding in bytes)

- CSV: "3" (text)
- SQL: VARCHAR, INTEGER

 "1", "2", "3" may represent an **ordinal** scale (low/mid/high).

The **meaning** of a feature – not how it is stored or encoded – determines how we analyze it.

Summary Statistics: Center and Spread

🎯 Center

Mean	✗
Median (also ordinal)	✓
Mode (any type)	✓

↔ Spread

Variance and SD	✗
IQR	✓
Range	✗

✓ robust to extreme values

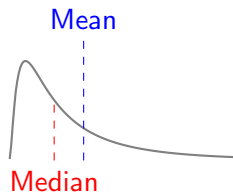
✗ moved by one extreme value

In EDA, statistics are **summaries**: they describe the structure of our data, they do not infer.

Shape of a Distribution

A distribution shows **how values are spread**: look at its shape before summarizing.

Skewed

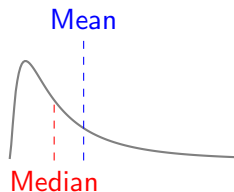


The mean is pulled
toward the tail:
prefer the **median**.

Shape of a Distribution

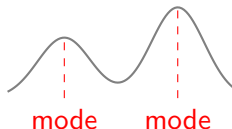
A distribution shows **how values are spread**: look at its shape before summarizing.

Skewed



The mean is pulled toward the tail:
prefer the **median**.

Several peaks

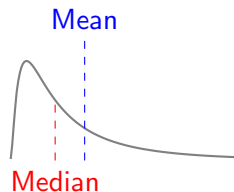


Possibly several subpopulations:
split before summarizing.

Shape of a Distribution

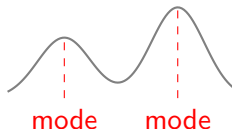
A distribution shows **how values are spread**: look at its shape before summarizing.

Skewed



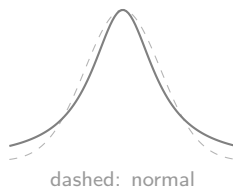
The mean is pulled toward the tail: prefer the **median**.

Several peaks



Possibly several subpopulations: **split** before summarizing.

Heavy tail



Extreme values are common: prefer **median and IQR** over mean and SD.

Box Plot

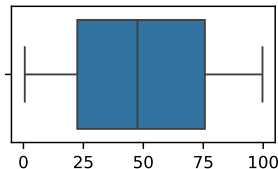
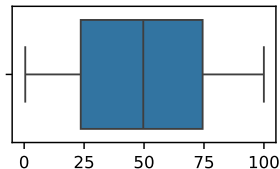
Five-number summary:

- 1 Minimum (Q_0)
- 2 First quartile (25%, Q_1)
- 3 Median (Q_2)
- 4 Third quartile (75%, Q_3)
- 5 Maximum (Q_4)

IQR directly: ($Q_3 - Q_1$).

Sometimes, whiskers extend to the most extreme observations **within** $1.5 \times \text{IQR}$ of the quartiles.

Points beyond them are flagged as **potential** outliers.



Box Plot

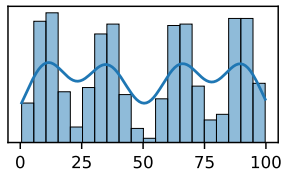
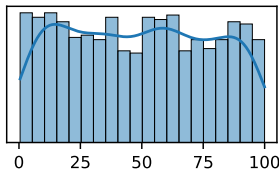
Five-number summary:

- 1 Minimum (Q_0)
- 2 First quartile (25%, Q_1)
- 3 Median (Q_2)
- 4 Third quartile (75%, Q_3)
- 5 Maximum (Q_4)

IQR directly: ($Q_3 - Q_1$).

Sometimes, whiskers extend to the most extreme observations **within** $1.5 \times \text{IQR}$ of the quartiles.

Points beyond them are flagged as **potential** outliers.

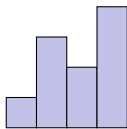


Same box plot, different distributions.

Histogram vs Bar Plot

Histogram

Y-axis = count or density

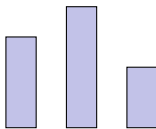


Width = bin size

- Continuous data (or discrete, many values).
- Width matters (bin size).
- **Density** histogram: total area is 1.

Bar Plot

Y-axis = count/proportion



Width = arbitrary

- Qualitative data (or discrete, few values).
- Width carries no meaning.
- Bar order: free for nominal, fixed for ordinal.



Do not interpret bar plots like histograms.

Transforming Features

Why?

Transforming Features

Why? To reveal patterns or make summaries more meaningful.

Transforming Features

Why? To reveal patterns or make summaries more meaningful.



Collapse categories

Merge rare levels
into "Other".

e.g., countries
below 1% of the sample

Transforming Features

Why? To reveal patterns or make summaries more meaningful.



Collapse categories

Merge rare levels
into “Other”.

e.g., countries
below 1% of the sample



Log transformation

Spreads out
right-skewed values.

e.g., income, file sizes,
transaction amounts

Transforming Features

Why? To reveal patterns or make summaries more meaningful.



Collapse categories

Merge rare levels
into "Other".

e.g., countries
below 1% of the sample



Log transformation

Spreads out
right-skewed values.

e.g., income, file sizes,
transaction amounts

⚠ Keep the **original feature**: add the transformed one as a new column (never edit the raw data).

From Variables to Relationships

Two Quantitative Features

How should we analyze the relationship between two quantitative features?

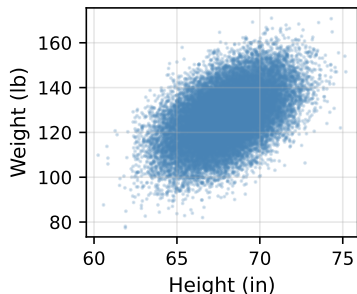
Two Quantitative Features

How should we analyze the relationship between two quantitative features?

Univariate plots alone **cannot** reveal joint structure.

Scatterplots reveal:

- **Strength** of association: strong, weak, none
- **Shape**: linear, curved, clustered
- **Outliers** that may distort trends.



*From Statistics Online
Computational Resource, UCLA*

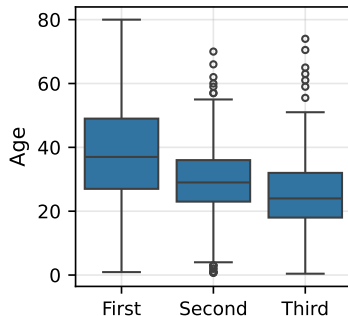
Quantitative vs Qualitative

How should we analyze the relationship between a quantitative and a qualitative feature?

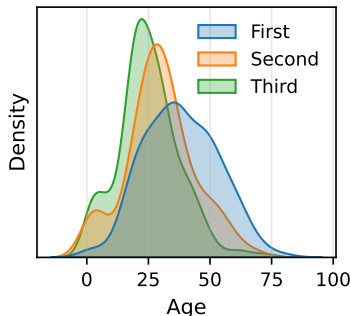
Quantitative vs Qualitative

How should we analyze the relationship between a quantitative and a qualitative feature?

Compare **distributions across groups**.



Box Plots



Overlaid Density curves

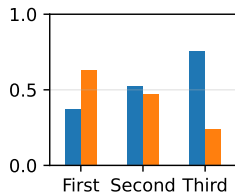
Two Qualitative Features

How should we analyze the relationship between two qualitative features?

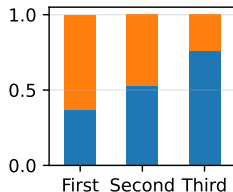
Two Qualitative Features

How should we analyze the relationship between two qualitative features?

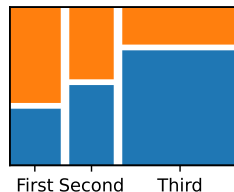
Compare **proportions across groups**.



Bar Charts



Stacked Bar Charts

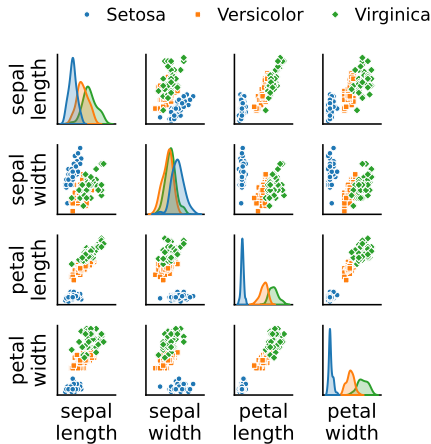


Mosaic Plot

■ Did not survive ■ Survived

Look at composition, frequency, and potential **confounding**.

Pair Plot: Look at Every Pair



Pair plot: scatter plots off the diagonal, distributions on it.

Beyond two features:

- **Color or symbol:** a third feature (here, the species).
- **Faceting:** one subplot per category.
- **Conditioning:** look within each value of a feature.

⚠ Each subgroup is smaller: patterns get noisier.

Correlation Matrix: Measure Every Pair

sepal length	1.00	-0.12	0.87	0.82
sepal width	-0.12	1.00	-0.43	-0.36
petal length	0.87	-0.43	1.00	0.96
petal width	0.82	-0.36	0.96	1.00
	sepal length	sepal width	petal length	petal width

Correlation matrix: the PCC of each pair.

- Pearson correlation coefficient (PCC)
 $r = \text{Cov}(X, Y) / (\sigma_X \sigma_Y)$:
strength of a **linear** association, from -1 to $+1$.
- $r \approx 0$ does not mean no relationship: a U-shape gives $r \approx 0$.
- Sensitive to outliers and to mixtures of subgroups.
- Correlation is not causation (confounding).

Which Chart for Which Question?



One quantitative
histogram, box plot



One qualitative
bar plot



Over time
line chart



Two quantitative
scatter plot



One of each
box plots per group



Two qualitative
stacked bars, mosaic



A third feature
color or facets



Many features
pair plot



Many quantitative
correlation matrix

Start from the **question** and the **types**, not the chart.

Interpretation

Observation or Interpretation?

Observation: what the data shows.

Interpretation: a claim about the world beyond the data.



Poll

Are the following claims observations or interpretations?

- (1) Vaccinated groups showed fewer hospitalizations.
- (2) Vaccination reduced hospitalizations.
- (3) First-class passengers survived at a higher rate.
- (4) Wealth caused survival.

Observation or Interpretation?

Observation: what the data shows.

Interpretation: a claim about the world beyond the data.



Are the following claims observations or interpretations?

- (1) Vaccinated groups showed fewer hospitalizations. **Observation**
- (2) Vaccination reduced hospitalizations.
- (3) First-class passengers survived at a higher rate.
- (4) Wealth caused survival.

Observation or Interpretation?

Observation: what the data shows.

Interpretation: a claim about the world beyond the data.



Are the following claims observations or interpretations?

- (1) Vaccinated groups showed fewer hospitalizations. **Observation**
- (2) Vaccination reduced hospitalizations. **Interpretation**
- (3) First-class passengers survived at a higher rate.
- (4) Wealth caused survival.

Observation or Interpretation?

Observation: what the data shows.

Interpretation: a claim about the world beyond the data.



Are the following claims observations or interpretations?

- (1) Vaccinated groups showed fewer hospitalizations. **Observation**
- (2) Vaccination reduced hospitalizations. **Interpretation**
- (3) First-class passengers survived at a higher rate. **Observation**
- (4) Wealth caused survival.

Observation or Interpretation?

Observation: what the data shows.

Interpretation: a claim about the world beyond the data.



Are the following claims observations or interpretations?

- (1) Vaccinated groups showed fewer hospitalizations. **Observation**
- (2) Vaccination reduced hospitalizations. **Interpretation**
- (3) First-class passengers survived at a higher rate. **Observation**
- (4) Wealth caused survival. **Interpretation**

Observation or Interpretation?

Observation: what the data shows.

Interpretation: a claim about the world beyond the data.



Are the following claims observations or interpretations?

(1) Vaccinated groups showed fewer hospitalizations. **Observation**

(2) Vaccination reduced hospitalizations. **Interpretation**

(3) First-class passengers survived at a higher rate. **Observation**

(4) Wealth caused survival. **Interpretation**

 Causal claims from observational data are interpretations.

Bigger Than Noise?

Two groups: **48%** vs **52%**. Is the 4-point gap real?

Bigger Than Noise?

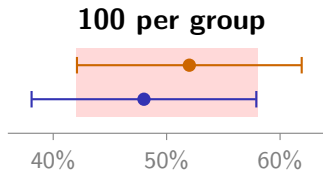
Two groups: **48%** vs **52%**. Is the 4-point gap real?

Recall sampling: margin $\approx \pm 2 \text{ SE}$, with $\text{SE}(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}$.

Bigger Than Noise?

Two groups: **48%** vs **52%**. Is the 4-point gap real?

Recall sampling: margin $\approx \pm 2 \text{ SE}$, with $\text{SE}(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}$.



$\text{SE} \approx 0.05$: ± 10 points

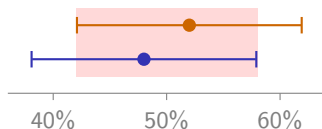
Intervals overlap: **within noise**.

Bigger Than Noise?

Two groups: **48%** vs **52%**. Is the 4-point gap real?

Recall sampling: margin $\approx \pm 2 \text{ SE}$, with $\text{SE}(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}$.

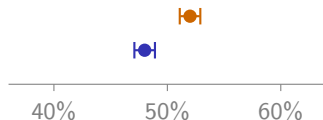
100 per group



$\text{SE} \approx 0.05$: ± 10 points

Intervals overlap: **within noise**.

10 000 per group



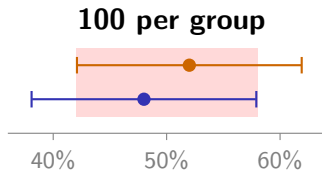
$\text{SE} \approx 0.005$: ± 1 point

Intervals apart: **beyond noise**.

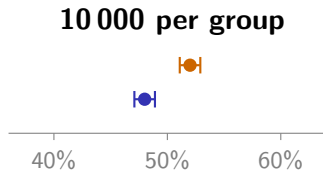
Bigger Than Noise?

Two groups: **48%** vs **52%**. Is the 4-point gap real?

Recall sampling: margin $\approx \pm 2 \text{ SE}$, with $\text{SE}(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}$.



SE ≈ 0.05 : ± 10 points
Intervals overlap: **within noise**.



SE ≈ 0.005 : ± 1 point
Intervals apart: **beyond noise**.

⚠ Overlapping intervals are a rough check, not a test.

Beyond noise $\neq$ important: the gap is real, but does it matter?

Simpson's Paradox¹: Confounding Variables

Definition (Simpson's Paradox)

A trend in groups of data **reverses when groups are combined**.

¹E. H. Simpson. "The Interpretation of Interaction in Contingency Tables".
In: **Journal of the Royal Statistical Society: Series B (Methodological)**
13.2 (July 1951).

Simpson's Paradox¹: Confounding Variables

Definition (Simpson's Paradox)

A trend in groups of data **reverses when groups are combined**.

Example: Berkeley Admissions Fall 1973 (Simplified)

Department	Men			Women		
	Total	Admitted	Perc.	Total	Admitted	Perc.
Overall	1500	750	50%	750	180	24%
Engineering	1000	720	72%	150	120	80%
English	500	30	6%	600	60	10%

Overall: Males admitted at higher rates.

¹E. H. Simpson. "The Interpretation of Interaction in Contingency Tables".
In: **Journal of the Royal Statistical Society: Series B (Methodological)**
13.2 (July 1951).

Simpson's Paradox¹: Confounding Variables

Definition (Simpson's Paradox)

A trend in groups of data **reverses when groups are combined**.

Example: Berkeley Admissions Fall 1973 (Simplified)

Department	Men			Women		
	Total	Admitted	Perc.	Total	Admitted	Perc.
Overall	1500	750	50%	750	180	24%
Engineering	1000	720	72%	150	120	80%
English	500	30	6%	600	60	10%

Overall: Males admitted at higher rates.

Within departments: Females admitted at higher rates in both!

¹E. H. Simpson. "The Interpretation of Interaction in Contingency Tables".
In: **Journal of the Royal Statistical Society: Series B (Methodological)**
13.2 (July 1951).

Simpson's Paradox¹: Confounding Variables

Definition (Simpson's Paradox)

A trend in groups of data **reverses when groups are combined**.

Example: Berkeley Admissions Fall 1973 (Simplified)

Department	Men			Women		
	Total	Admitted	Perc.	Total	Admitted	Perc.
Overall	1500	750	50%	750	180	24%
Engineering	1000	720	72%	150	120	80%
English	500	30	6%	600	60	10%

Overall: Males admitted at higher rates.

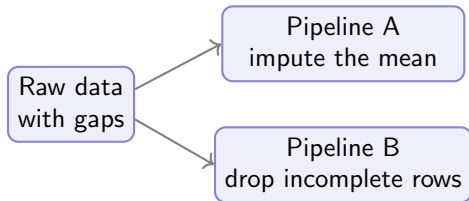
Within departments: Females admitted at higher rates in both!

Always check relationships **within subgroups**: use **stratification**!

¹E. H. Simpson. "The Interpretation of Interaction in Contingency Tables".
In: **Journal of the Royal Statistical Society: Series B (Methodological)**
13.2 (July 1951).

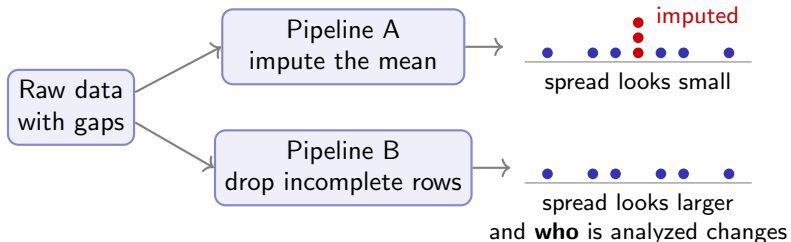
Does It Survive Your Cleaning?

Rerun the pipeline with the **other** cleaning decision: a reproducible pipeline makes it cheap.



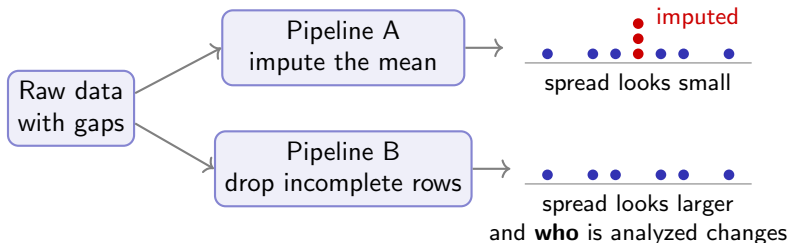
Does It Survive Your Cleaning?

Rerun the pipeline with the **other** cleaning decision: a reproducible pipeline makes it cheap.



Does It Survive Your Cleaning?

Rerun the pipeline with the **other** cleaning decision: a reproducible pipeline makes it cheap.



Neither choice is automatically right.
If the conclusion **flips**, it is **fragile**: report it as a limitation.

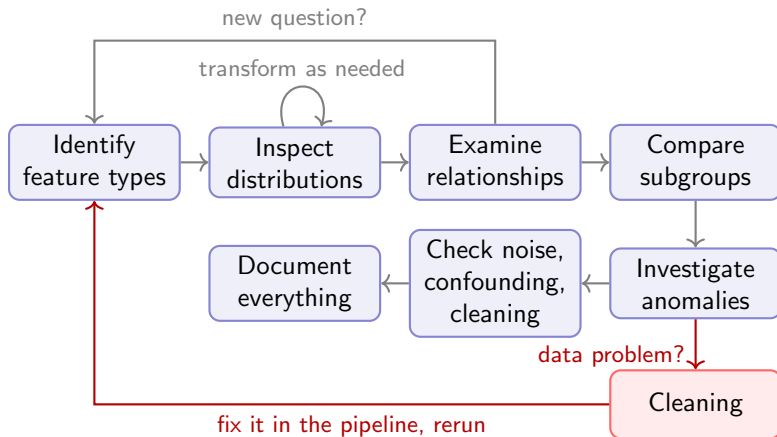
Limitations of EDA

Be cautious about:

-  **Overinterpreting noise:** Not every pattern is meaningful.
-  **Data dredging:** Looking at enough plots shows a pattern.
-  **Hindsight bias:** “I knew it all along” effect.
-  **Unreported choices:** Lack of transparency.

Keep a trail: the cleaning audit table and the wrangling pipeline.

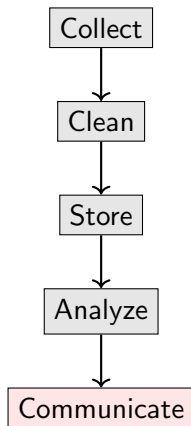
EDA Workflow



At each step: What do I see? Why does it matter? What next?

Communication

Within The Lifecycle



Principles of Communication



What
do we say?

Know your audience
technical or not,
deciding or
implementing



How
do we show it?

Tell a story
key insight first,
then evidence,
then action



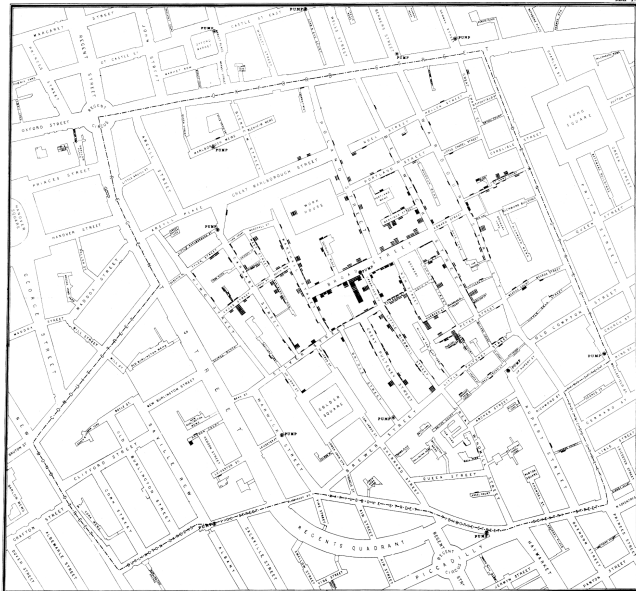
How do we say it
responsibly?

Show the limits
uncertainty,
limitations,
honesty

Turn data into **actionable insights** that drive decisions.

and the Ugly

"The correlation coefficient is 0.76."

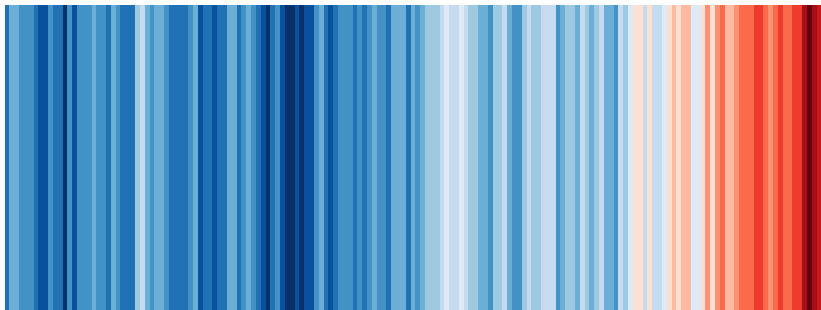


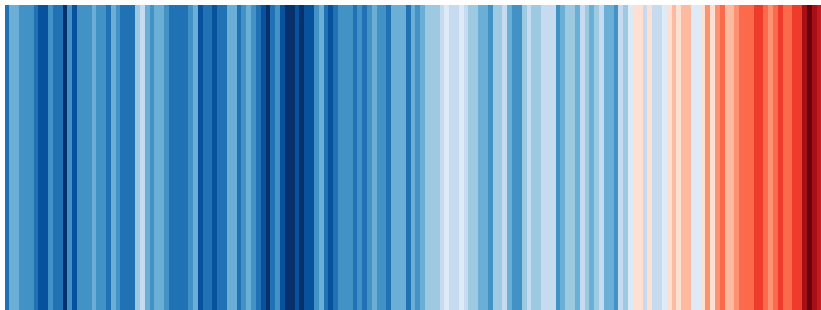
C. J. Griffin Ltd. Southampton S.W. 1, London.

SCALE 50 INCHES TO A MILE.

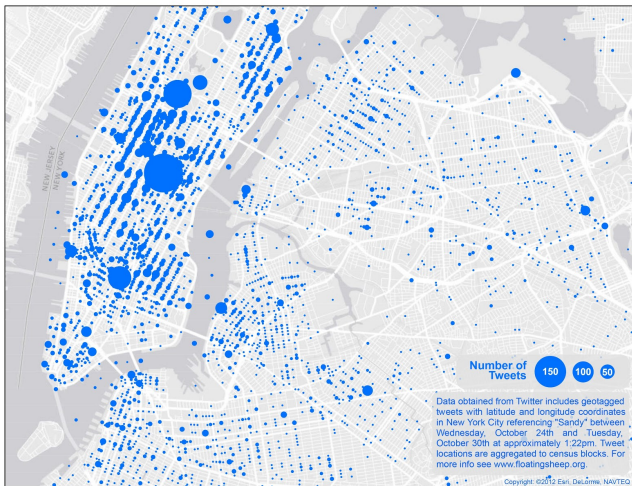


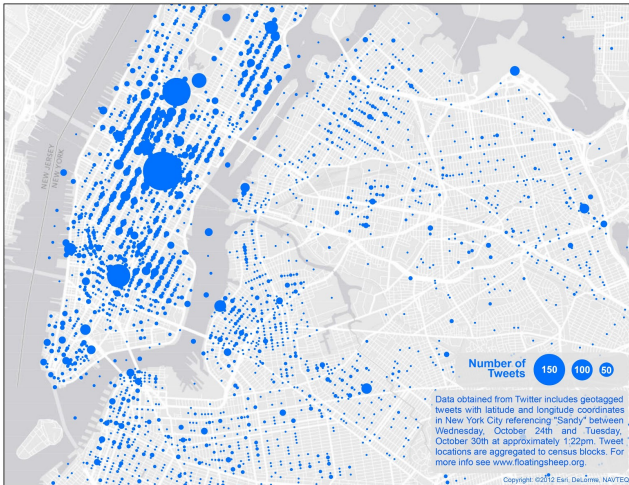
1854 Broad Street cholera outbreak by John Snow





Warming Stripes (1850 – 2018) by Ed Hawkins  

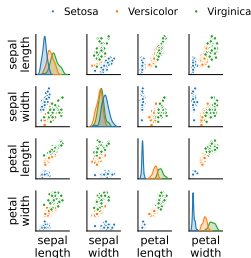




The Red Cross used tweets to dispatch help during a hurricane. “A social media blackhole meant the area needed help” – Andy Kirk

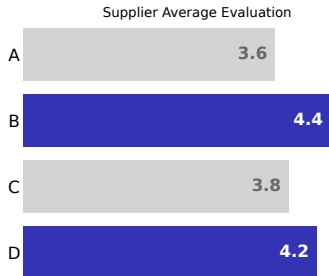
Different Stages, Different Purposes

Analysis



Many quick plots, for you.
“Good enough” is enough.

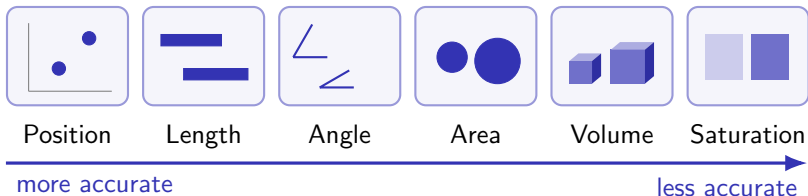
Communication



One chosen chart, for others.
Understandable without you.

The Cleveland-McGill² Ranking of Graphical Perceptions

How accurately do we read each visual encoding?



✓ Prefer

- Bars to compare values
- Lines for trends

✗ Avoid

- Pies to compare values
- 3D effects

²William Cleveland and Robert McGill. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". In: **Journal of the American Statistical Association** 79.387 (Sept. 1984).

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio¹
- Avoid chart junk¹

Scores of our suppliers:

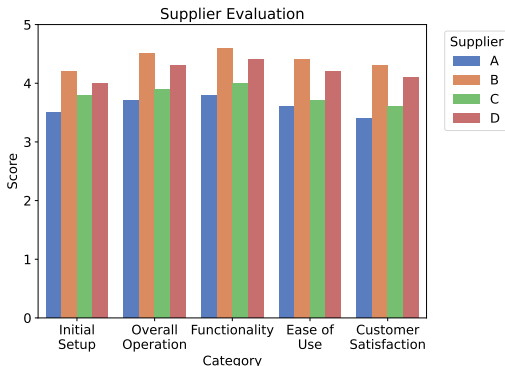
Criterion	A	B	C	D
Initial Setup	3.5	4.2	3.8	4.0
Overall Operation	3.7	4.5	3.9	4.3
Functionality	3.8	4.6	4.0	4.4
Ease of Use	3.6	4.4	3.7	4.2
Satisfaction	3.4	4.3	3.6	4.1

¹Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio¹
- Avoid chart junk¹

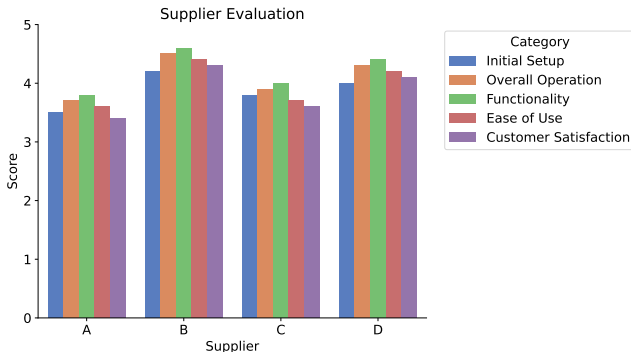


¹Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio¹
- Avoid chart junk¹

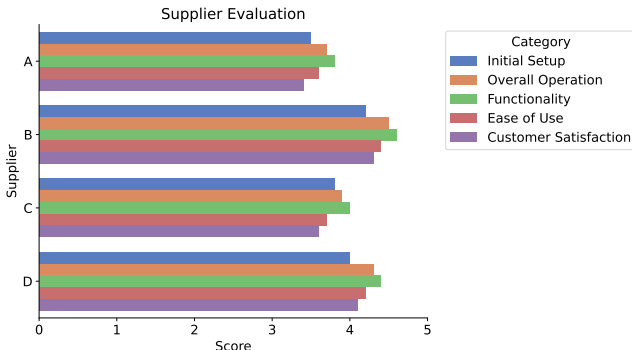


¹Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio¹
- Avoid chart junk¹

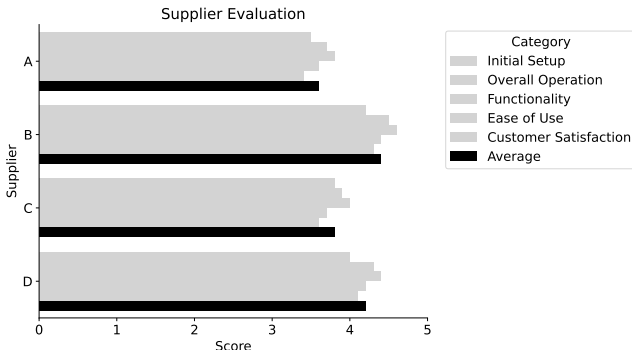


¹Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio¹
- Avoid chart junk¹

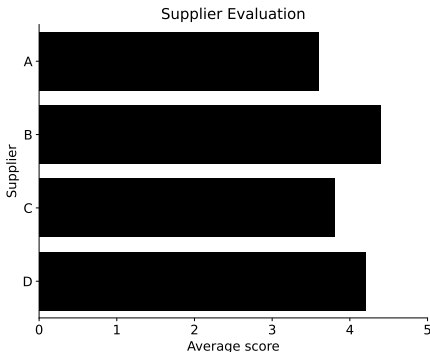


¹Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Maximize data-ink ratio¹
- Use clean, simple designs
- Avoid chart junk¹

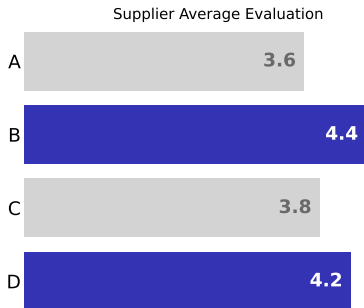


¹Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

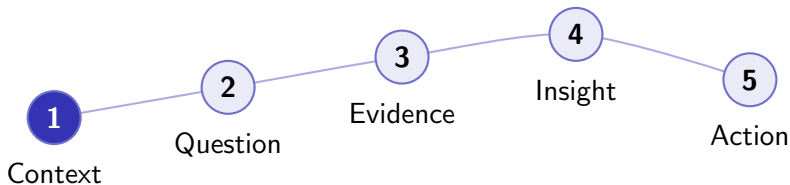
- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio¹
- Avoid chart junk¹



¹Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

Data Storytelling Framework

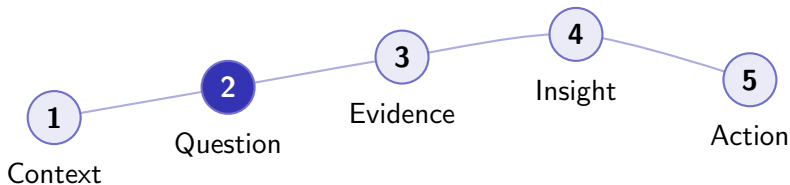
The supplier scores, told as a story:



“We renew one supplier contract next quarter. Four suppliers were scored on five criteria.”

Data Storytelling Framework

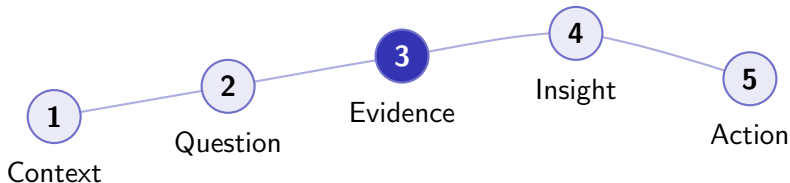
The supplier scores, told as a story:



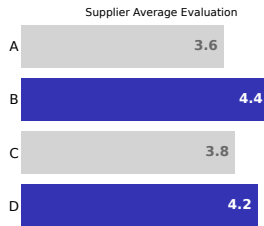
“Which supplier should we keep?”

Data Storytelling Framework

The supplier scores, told as a story:

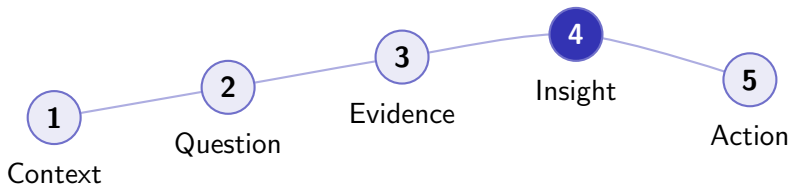


“B has the highest average score, 4.4 out of 5, and it leads on every criterion.”



Data Storytelling Framework

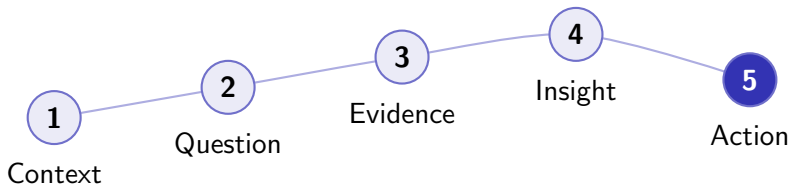
The supplier scores, told as a story:



“Keeping B means no trade-off: it is the best on everything we measured. D is a close second (4.2).”

Data Storytelling Framework

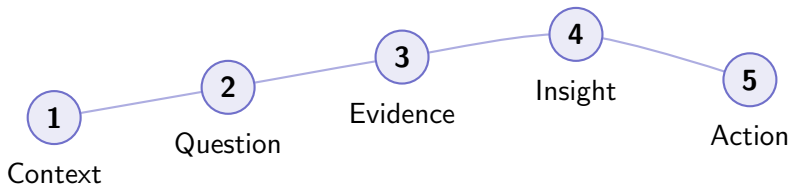
The supplier scores, told as a story:



“We recommend renewing B, with D as the fallback.”

Data Storytelling Framework

The supplier scores, told as a story:



“We recommend renewing B, with D as the fallback.”

Data storytelling is about **impact**, not just information.

Communicating Uncertainty

Data is never exact: say how confident we can be.

Communicating Uncertainty

Data is never exact: say how confident we can be.

How to Acknowledge Uncertainty Clearly

- State the **scope** of the data
(sample size, time window, missing data).
- Indicate the **stability** of findings
(variance, ranges).
- Describe **limitations**
(possible biases, small subgroups, unmeasured variables)

Transparency earns trust.

Ethical Communication of Data

Data communication influences decisions. Six ways to mislead:



Misleading axes
truncated or
inconsistent scales



Cherry-picking
only the
supportive subsets



Over-smoothing
variability
smoothed away



**Hiding
uncertainty**
or assumptions



Opaque methods
unclear steps,
missing context



Blurring claims
an observation
shown as a cause

Ethical Communication of Data

Data communication influences decisions. Six ways to mislead:



Misleading axes
truncated or
inconsistent scales



Cherry-picking
only the
supportive subsets



Over-smoothing
variability
smoothed away



**Hiding
uncertainty**
or assumptions



Opaque methods
unclear steps,
missing context



Blurring claims
an observation
shown as a cause

Honest communication means showing the data **as it is**,
not as we wish it were.

Conclusion

Takeaways: From Data Analysis to Communication

- 1 **EDA is driven by questions:** it is iterative, and each answer can raise a new question.
- 2 **Meaning, not encoding:** the feature type decides the summary and the chart.
- 3 **One feature, then several:** examine distributions before relationships.
- 4 **An observation is not an interpretation:** check it against noise, confounding and your cleaning choices.
- 5 **Communicate for an audience:** clear, minimal charts, tailored to the decision at hand.
- 6 **Show the data as it is:** state uncertainty and limitations, never hide them.



Why should we care about Data?

What is data and what we can do with it

DATA



SORTED



ARRANGED



PRESENTED VISUALLY



EXPLAINED WITH A STORY



ACTIONABLE (USEFUL)

