Data Exploration

Data Science Project: An Inductive Learning Approach

Prof. Dr. Filipe A. N. Verri

About these slides

These slides are companion material for the book

Data Science Project: An Inductive Learning Approach

Prof. Dr. Filipe A. N. Verri

`https://leanpub.com/dsp`

All intellectual content comes from the book and is not AI-generated. Slides were produced with the assistance of Claude Code.

The greatest value of a picture is when it forces us to notice what we never expected to see.

— John W. Tukey, Exploratory Data Analysis

Overview

Contents

- Introduction and goals
- Descriptive statistics
- Data visualization
- Graphical integrity and data-ink ratio
- Turning charts into stories

Objectives

- Understand what to explore and what not to
- Use descriptive statistics effectively
- Create honest, clear visualizations
- Communicate findings as stories

Introduction

When does exploration happen?

Two main moments:

1. **Before data organization** — understand the original phenomena and data sources
2. **After data organization** — validate that transformations achieved expected results

In practice, also during intermediate steps as a form of **debugging**.

Exploratory Data Analysis (EDA)

- Term coined by Tukey (1977)[1]
- Informal study of data: graphics and elaborate statistical summaries
- Counterpoint to dominance of confirmatory statistical methods
- Original context: statistical modeling, not machine learning
- Two classical goals: data description and model formulation
- In our approach: model formulation is delegated to the ML algorithm

[1] J. W. Tukey (1977). ***Exploratory data analysis.*** Pearson, p. 712. ISBN: 978-0201076165.

The fine line

Key tension

Deepen understanding enough to know the data and the problem, but not so much as to bias the final solution.

- ML principle: minimize human interference in the learning process
- Excessive domain-knowledge influence can limit discovery
- A variable that experts deem unimportant may contain valuable information

Goal: understand units and variables

Fixed (index) variables:

- Identify each one's meaning, individually and combined
- Possible observational units (e.g. enrollment, class)
- Domain: categorical, ideally finite and known *a priori*
- Inspect unique values — but beware: sample $\neq$ full domain

Measured (non-index) variables:

- Used directly in the ML problem
- Analyze type, domain, and distribution
- Descriptive statistics and visualizations

Goal: investigate inconsistencies

Fixed variables:

- No duplicate rows (unique identification)
- No missing keys
- Missing observations vs. missing key parts

Measured variables:

- Extreme values — measurement errors or irrelevant phenomena?
- Missing values — which preprocessing techniques are needed?
- Compare observed characteristics with prior knowledge

Goal: validate data manipulations

- Data handling operations can produce counter-intuitive results
- Explore both original and transformed data
- Example: after pivoting, check unique values and distributions
- Interactive environments (e.g. Jupyter) as debugging tools
- Step-by-step inspection of intermediate transformations

Non-goals of data exploration

Do not model or infer relationships

- Correlation analysis to decide variable importance
- Fitting regression models to understand relationships
- These tasks belong to the solution search algorithm

Do not test statistical hypotheses

- Testing if a variable follows a specific distribution adds little
- ML algorithms are robust to deviations
- Prefer experimental validation of the solution

Do not manipulate data

- Exploration and manipulation are distinct tasks
- Exploratory transformations must not persist

Communication in the exploration report

Three-step approach:

1. Formulate a **question** about the data
2. Choose **tools and techniques** to answer it
3. **Communicate** the answer via statistics and visualizations

Key principles:

- **Simplicity** — avoid excess information and complex charts
- **Relevance** — the right answer to the wrong question is useless
- **Clarity** — no ambiguity, no unnecessary jargon
- **Objectivity** — each piece answers one question only

Descriptive statistics

Descriptive statistics

Numeric summaries of specific properties of the data distribution:

- **Location**: mean, median, mode, min, max
- **Frequency**: counts per category
- **Dispersion**: standard deviation, entropy, kurtosis

Main uses in exploration:

- Validate the domain of variables against expectations
- Detect class imbalance or underrepresented categories
- Assess information content of each variable

Dispersion and information content

Low dispersion (std. dev. ≈ 0, low entropy):

- Variable carries little discriminative information
- Most observations share similar values
- May need to collect more diverse data

High dispersion:

- Too many unique or spread-out values
- Model may struggle to generalize
- Excess information can be as harmful as lack of it
- Address during preprocessing

Always present statistics within a **storytelling** context — never as isolated tables without interpretation.

Data visualization

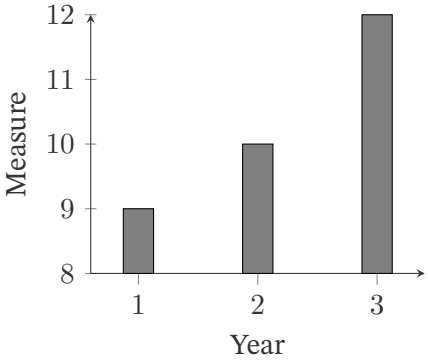
Why visualize?

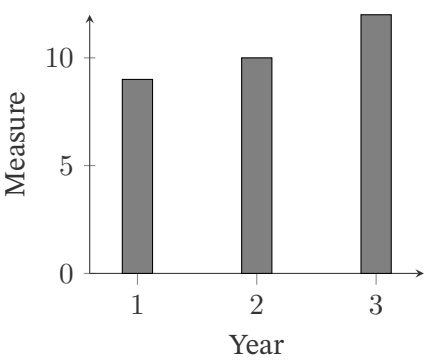
- The visual cortex absorbs large amounts of information quickly
- Charts and diagrams make data more intuitive than raw numbers
- Powerful for identifying patterns, anomalies, and relationships

But beware

Visualizations can be ambiguous, poorly chosen, or intentionally misleading (“lying with statistics”).

Graphical integrity — misleading scales





Left: axis not starting at zero exaggerates differences.
Right: axis starting at zero gives an honest representation.

Tufte's six principles of graphical integrity

1. Physical representation must be proportional to values
2. Labels, explanations, and events must be clear and visible
3. Show variation in data, never in design or scale
4. For monetary values, adjust for inflation/deflation
5. Dimensions in chart $\leq$ dimensions in data
6. Charts must not omit context relevant for interpretation

Tufte (2001)[2]

[2] E. R. Tufte (2001). ***The Visual Display of Quantitative Information***. 2nd ed. Cheshire, CT, USA: Graphics Press, p. 197. ISBN: 978-0961392147.

Data-ink ratio

Core principle: **above all, show the data.**

Definition (Tufte, 2001)

$$\text{DIR} = \frac{\text{DI}}{\text{DI} + \text{RI} + \text{NDI}}$$

- DI: non-erasable data-ink
- RI: redundant data-ink
- NDI: non-data ink

$1 - \text{DIR}$ = proportion of the chart that can be removed without losing information.

Maximizing data-ink ratio

Erase non-data ink (within reason):

- Remove unnecessary axis lines, grids, legends, labels
- DIR increases as NDI $\to 0$
- But: grids can help reading without a ruler

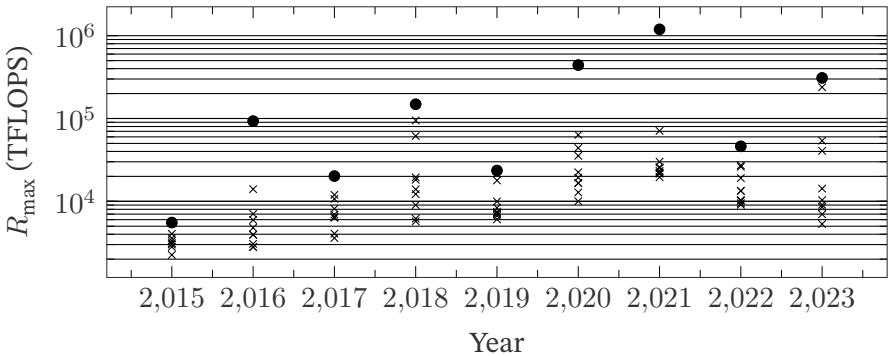
Erase redundant data-ink (within reason):

- Labeled axes $\to$ no need to label each bar
- Exception: periodic data may benefit from repetition
- Highlighting specific data points can be useful

Chart editing — iterative process

1. Idealize the visualization and map variables to dimensions
2. Produce an initial chart
3. Iteratively remove ink and improve integrity
4. Adjust until the chart clearly communicates its intention

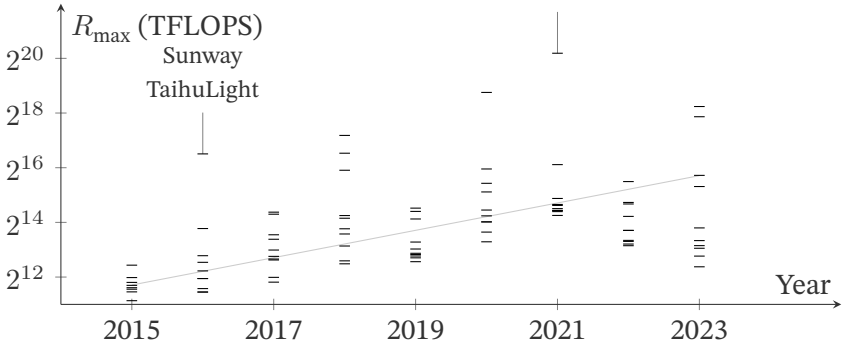
Example: supercomputer performance (before editing)



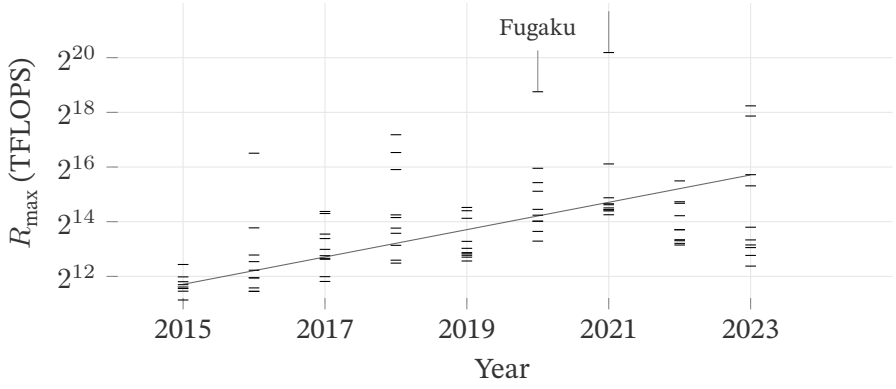
Editing steps

- Display performance as powers of two (2^{12}, 2^{14}, ...) to emphasize doubling
- Remove vertical grid — log base-2 scale is self-explanatory
- Use uniform markers (short horizontal lines) — focus on trend, not individual bests
- Fix axis labels (years without thousands separator)
- Replace bounding box with axis arrows for cleaner design
- Add Moore's Law trend line and annotations

Example: supercomputer performance (after editing)



Alternative version with discrete grid



Intentionality of graphical elements

- Keep redundant/non-data elements when they help communicate **intent**
- Grids $\rightarrow$ “compare values”
- Annotations $\rightarrow$ “highlight this point”
- Trend lines $\rightarrow$ “observe this relationship”
- A chart becomes cluttered when it tries to communicate **too many things at once**
- Prefer simpler charts highlighting a single phenomenon

Colors and contrast

- Color variation must be **perceptually uniform**[3]
- Neighboring colors should have equal perceptual differences
- Must remain legible for colorblind readers and in grayscale
- Use palettes with monotonic lightness gradient[4]
- Scientific color maps available for most programming languages

[3] F. Crameri, G. E. Shephard, and P. J. Heron (2020). **"The misuse of colour in science communication".** In: *Nature Communications* 11.1, p. 5444. DOI: 10.1038/s41467-020-19160-7.

[4] F. Crameri (2023). ***Scientific colour maps.*** Version 8.0.1. DOI: 10.5281/zenodo.8409685.

Practical aspects (Wilke, 2019)

Wilke's guide for digital data visualization[5]:

- Compares versions of the same chart
- Labels problematic figures as:
  - **Ugly** — aesthetic problems, but clear and informative
  - **Bad** — perception problems (confusing, misleading)
  - **Wrong** — mathematical errors or misinterpretation
- Some classifications are subjective
- Focus: honest, clear, professional design

[5] C. O. Wilke (2019). ***Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.*** 1st ed. Shelter Island, NY, USA: O'Reilly Media, p. 387. ISBN: 978-1492031086. URL: https://clauswilke.com/dataviz/.

Turning charts into stories

A story creates an emotional reaction through ordered observations.

Story arc: tension of a question $\rightarrow$ resolution.

OCAR format (Opening–Challenge–Action–Resolution):

1. **Opening** — introduce the topic
2. **Challenge** — raise a question
3. **Action** — present the chart
4. **Resolution** — state the conclusion

Other formats: LDR (Lead–Development–Resolution),
ABDCE (Action–Background–Development–Climax–Ending).

OCAR example

O Moore's Law states that the number of transistors in a microprocessor doubles every two years, typically reflecting exponential performance growth.

C Does supercomputer performance follow this trend?

A The chart shows R_{max} of the top 10 supercomputers per year (2015–2023).

R Despite outliers (Sunway TaihuLight, Frontier), the trend line confirms performance doubles roughly every two years.

Final comments

Advanced methods

- PCA, t-SNE, clustering, association rules
- Valuable for EDA, but use with caution
- When the goal is predictive, exploration can become an end in itself
- May find *interesting* but not *important* features
- Prefer simple, widely known methods

Takeaways

- Explore to **understand** and **validate**, not to model
- Descriptive statistics and visualizations within a storytelling context
- Graphical integrity: proportional, labeled, contextual
- Maximize data-ink ratio — remove what does not inform
- Choose colors that are perceptually uniform and accessible
- Structure reports as stories (OCAR)
- The exploration report must be concise and objective — it communicates progress to stakeholders

Questions?