Data, Data Storage, Data Collection

Lecture 11: From Data Analysis to Communication

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

Recap

Recap: Normalization and Denormalization

Normalization

1. Eliminates redundancy and ensures **data integrity**.
2. **1NF → 2NF → 3NF → BCNF → 4NF → 5NF**.
3. Solves update, deletion, and insertion anomalies.
4. Best for **OLTP systems** (frequent writes).

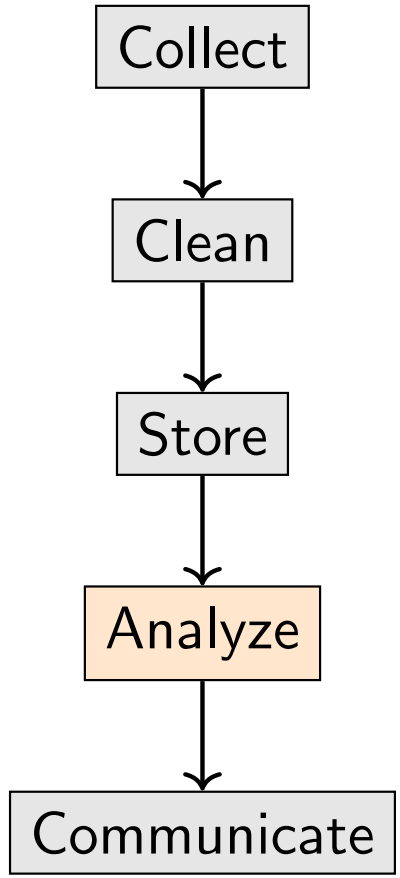
Denormalization

1. **Intentional violation** of normal forms for practical or performance reasons.
2. Best for **OLAP systems** (frequent reads).

Normalize for **consistency**, denormalize for **performance**.

Introduction

Within the lifecycle



Session Objectives

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

Analysis

- Explain what data analysis is and how it fits in the data lifecycle.
- Use descriptive statistics and visualizations to uncover insights.
- Identify appropriate visual encodings for different types of data.

Communication

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

What is Data Analysis?

Data Analysis

Transformation of **raw, 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?

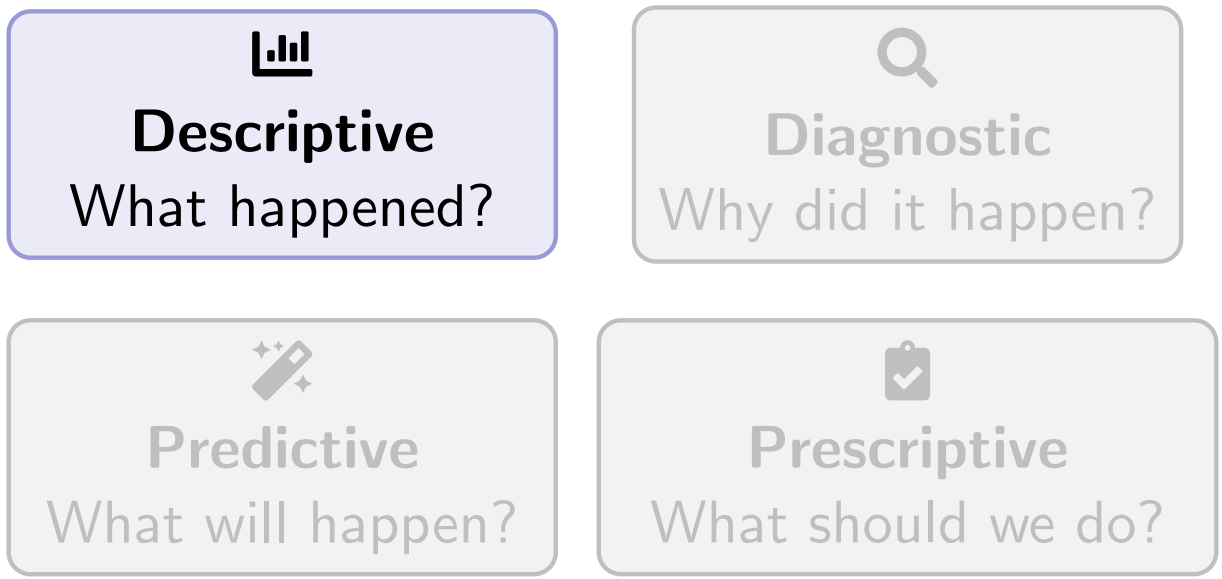
What is Data Analysis?

Data Analysis

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

- Draw conclusions.
- Support decision-making.

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



Example: COVID-19 Data Analysis

Raw Data: Daily case counts, hospitalizations, vaccinations.

Example: COVID-19 Data Analysis

Raw Data: Daily case counts, hospitalizations, vaccinations.

Questions:

- How is the virus spreading over time?
- Which age groups are most affected?
- Are vaccinations reducing hospitalizations?

Example: COVID-19 Data Analysis

Raw Data: Daily case counts, hospitalizations, vaccinations.

Questions:

- How is the virus spreading over time?
- Which age groups are most affected?
- Are vaccinations reducing hospitalizations?

Insights:

- Trends: Cases peaked in winter 2021.
- Patterns: Elderly had higher hospitalization rates.
- Impact: Vaccinated groups showed significantly fewer hospitalizations.

Example: COVID-19 Data Analysis

Raw Data: Daily case counts, hospitalizations, vaccinations.

Questions:

- How is the virus spreading over time?
- Which age groups are most affected?
- Are vaccinations reducing hospitalizations?

Insights:

- Trends: Cases peaked in winter 2021.
- Patterns: Elderly had higher hospitalization rates.
- Impact: Vaccinated groups showed significantly fewer hospitalizations.

Action: Prioritize vaccine rollout for elderly populations.

Example: COVID-19 Data Analysis

Raw Data: Daily case counts, hospitalizations, vaccinations.

Questions:

- How is the virus spreading over time?
- Which age groups are most affected?
- Are vaccinations reducing hospitalizations?

Insights:

- Trends: Cases peaked in winter 2021.
- Patterns: Elderly had higher hospitalization rates.
- Impact: Vaccinated groups showed significantly fewer hospitalizations.

Action: Prioritize vaccine rollout for elderly populations.

⚠ Data analysis informs but does not determine policy.

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.

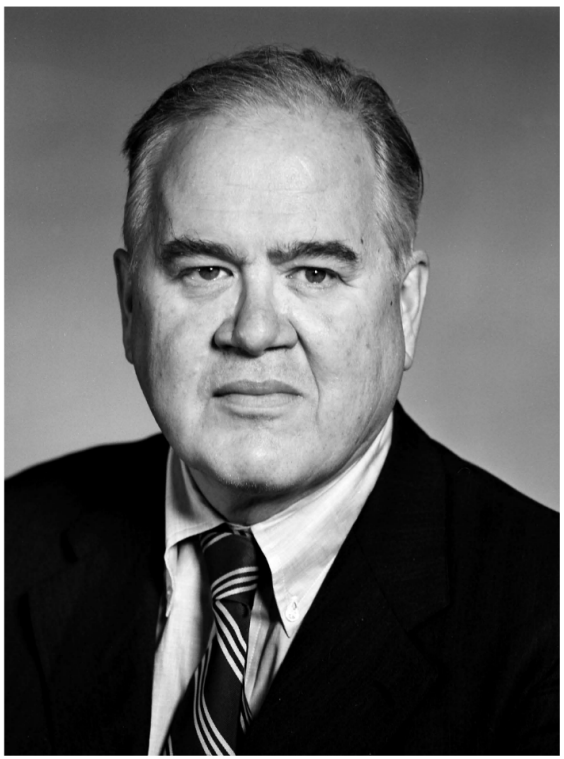


Photo Credit: Princeton University, Robert Matthews

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?

EDA is not a set of fixed recipes.

Exploratory Data Analysis relies on . . .

Summary statistics and **visualizations** . . .

To discover **patterns**, **structure**, and **anomalies**.

Exploratory Data Analysis relies on ...

Summary statistics and **visualizations** ...

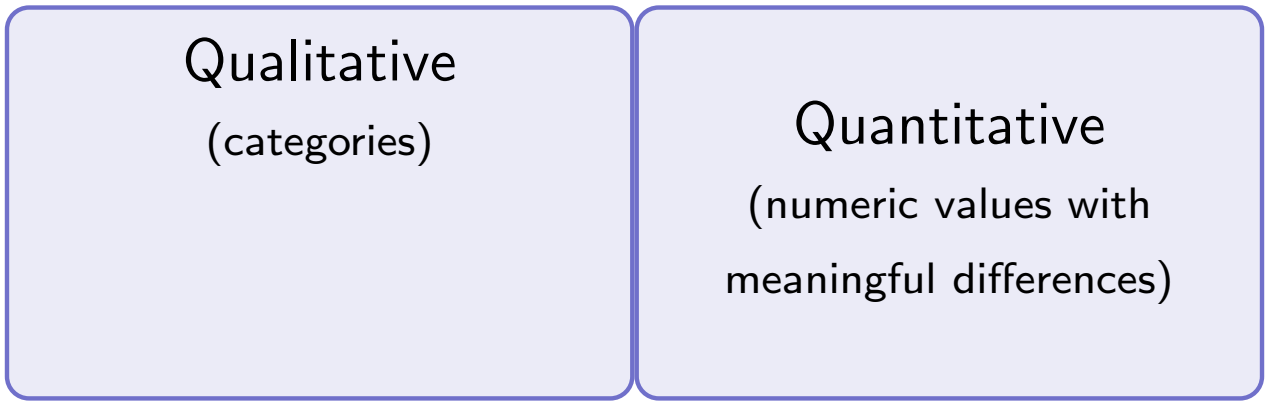
To discover **patterns**, **structure**, and **anomalies**.

Neither is sufficient alone: they offer **different insights**.

Features & Meaning

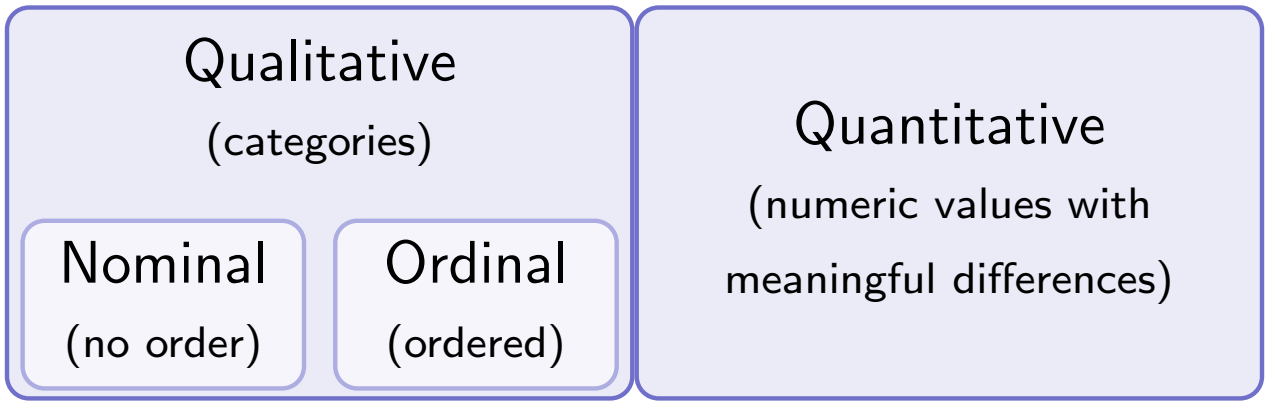
Feature Types

Taxonomy (recall Data Collection):



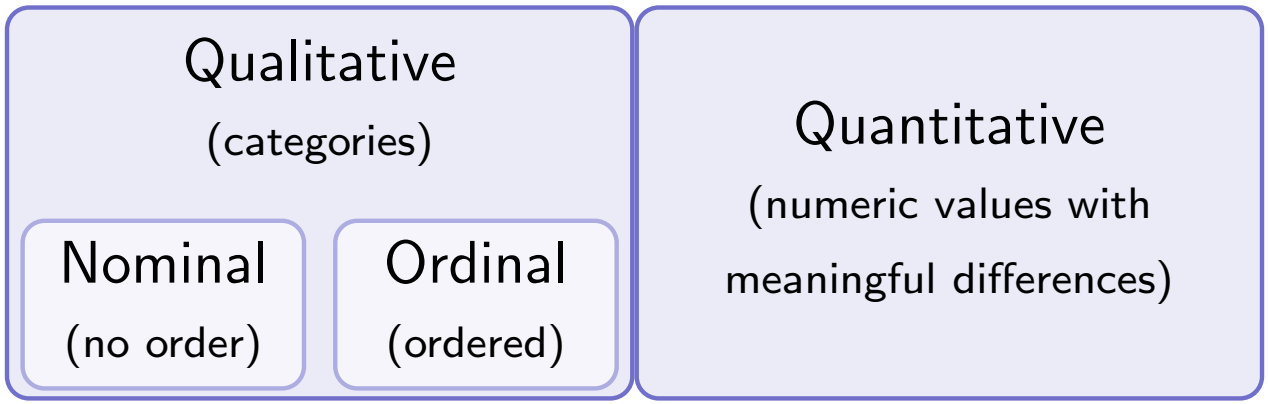
Feature Types

Taxonomy (recall Data Collection):



Feature Types

Taxonomy (recall Data Collection):

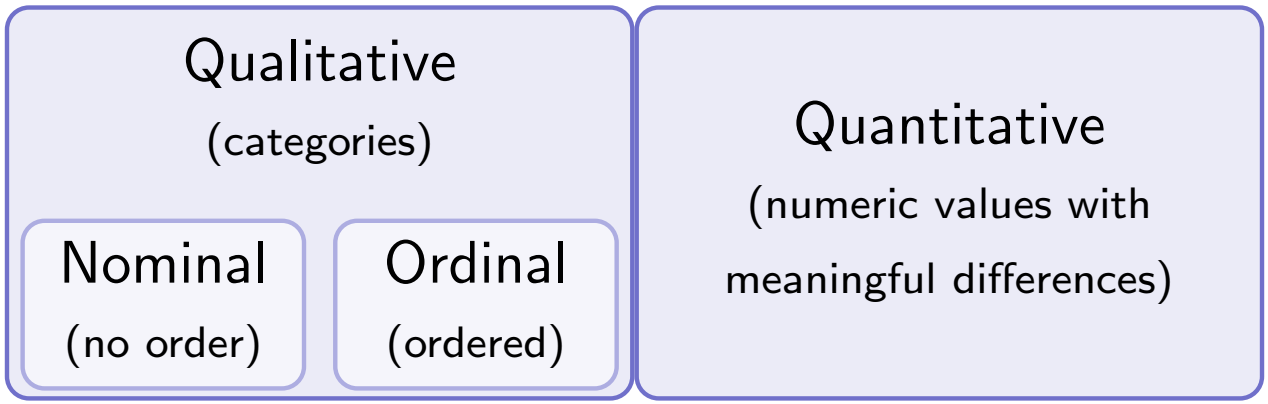


Why feature types matter in EDA?

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

Feature Types

Taxonomy (recall Data Collection):

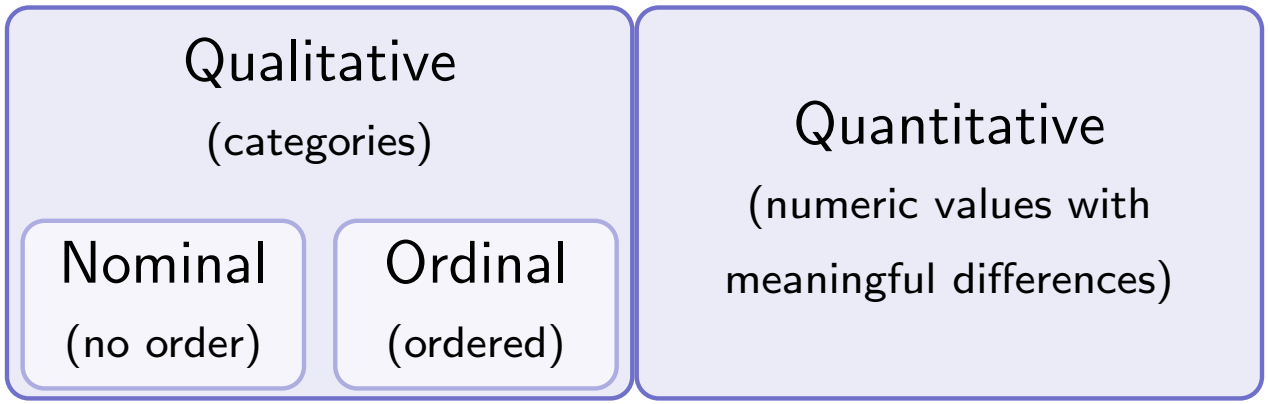


Why feature types matter in EDA?

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

Feature Types

Taxonomy (recall Data Collection):

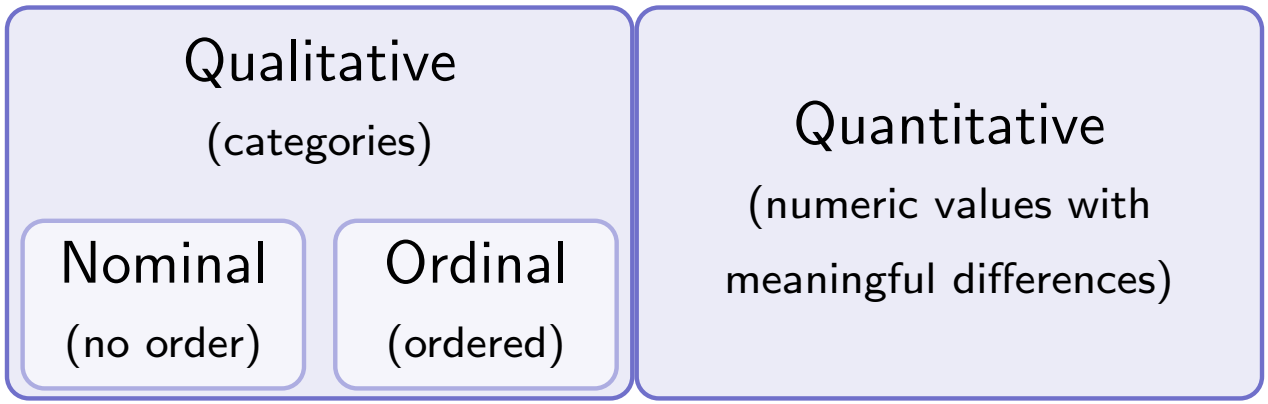


Why feature types matter in EDA?

- 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

Taxonomy (recall Data Collection):



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 Storage Type

Feature type 💡
(what the values **mean**)

- Nominal
- Ordinal
- Quantitative

Data type </>
(language manipulations)

- `int`,
- `float`
- `string`

Storage type 🗄
(encoding in bytes)

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

⚠ Same encoding, different meaning:
`"1"`, `"2"`, `"3"` may represent an **ordinal** scale (low/medium/high).

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

Summary Statistics

Statistics in Data Analysis

Why Statistics?

In Exploratory Data Analysis (EDA), statistics are **summaries** that help us understand the structure of our data:

- How large are typical values?
- How much do values vary?
- How do different groups compare?
- Do unusual or extreme values appear?

Statistics in Data Analysis

Why Statistics?

In Exploratory Data Analysis (EDA), statistics are **summaries** that help us understand the structure of our data:

- How large are typical values?
- How much do values vary?
- How do different groups compare?
- Do unusual or extreme values appear?

⚠ We use statistics here for **description**, not for inference.
No hypothesis testing, no confidence intervals.

Samples and Populations

Most datasets we analyze are **samples** from a broader population.

Key Terms (Practical View)

- **Statistic**: a numerical summary of the sample you have.
- **Parameter**: a numerical summary of the population (usually unknown).

Samples and Populations

Most datasets we analyze are **samples** from a broader population.

Key Terms (Practical View)

- **Statistic**: a numerical summary of the sample you have.
- **Parameter**: a numerical summary of the population (usually unknown).

Example: Car Ownership at ESSEC

- "45% of all ESSEC students own a car" $\rightarrow$
- "34% of AIDAMS students own a car" $\rightarrow$

Why does this matter? Even with careful sampling, our summaries depend on what we observe: no dataset represents the full population perfectly.

Samples and Populations

Most datasets we analyze are **samples** from a broader population.

Key Terms (Practical View)

- **Statistic**: a numerical summary of the sample you have.
- **Parameter**: a numerical summary of the population (usually unknown).

Example: Car Ownership at ESSEC

- "45% of all ESSEC students own a car" $\rightarrow$ **parameter**.
- "34% of AIDAMS students own a car" $\rightarrow$

Why does this matter? Even with careful sampling, our summaries depend on what we observe: no dataset represents the full population perfectly.

Samples and Populations

Most datasets we analyze are **samples** from a broader population.

Key Terms (Practical View)

- **Statistic**: a numerical summary of the sample you have.
- **Parameter**: a numerical summary of the population (usually unknown).

Example: Car Ownership at ESSEC

- "45% of all ESSEC students own a car" $\rightarrow$ **parameter**.
- "34% of AIDAMS students own a car" $\rightarrow$ **statistic**.

Why does this matter? Even with careful sampling, our summaries depend on what we observe: no dataset represents the full population perfectly.

Measures of Center

Measures of center summarize data with a **single value**.

Common Measures

- **Mean**: arithmetic average.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **Median**: middle value (robust to outliers).
- **Mode**: most frequent category or value.

Measures of Center

Measures of center summarize data with a **single value**.

Common Measures

- **Mean**: arithmetic average.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **Median**: middle value (robust to outliers).
- **Mode**: most frequent category or value.

How Are These Used in EDA?

- Income data: median is more informative than mean (few extreme values dominate the mean).
- Heights of athletes: mean works well (roughly symmetric).
- Survey answers: mode and proportions summarize preferences.

Measures of Spread

Measures of spread summarize how **variable** the data is.

Common Measures

- **Range**: $\max - \min$ (sensitive to outliers).
- **Interquartile Range**: $\text{IQR} = Q_3 - Q_1$ (robust to outliers).
- **Variance**:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Standard Deviation**: $\sigma = \sqrt{\sigma^2}$.

Measures of Spread

Measures of spread summarize how **variable** the data is.

Common Measures

- **Range**: $\max - \min$ (sensitive to outliers).
- **Interquartile Range**: $\text{IQR} = Q_3 - Q_1$ (robust to outliers).
- **Variance**:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Standard Deviation**: $\sigma = \sqrt{\sigma^2}$.

Interpreting Spread in EDA

- Two neighborhoods can have the same median income but very different IQRs.
- A dataset with large SD may contain subgroups or outliers.

Understanding Data Through Distributions

What is a Distribution?

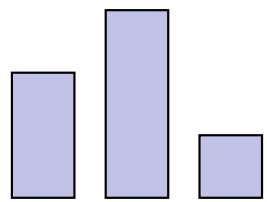
A distribution shows **how values are spread** in the data.

Quantitative data:

- Shape (e.g., symmetric, skewed)
- Modes (peaks)
- Tails (extreme values)

Qualitative data:

- Frequencies (counts)
- Proportions (%)



Why Does It Matter? Distributions Help You:

- Understand the **central tendency**.
- Identify **patterns**.
- Detect **anomalies** or **unusual** values.

Symmetric vs Skewed Distributions

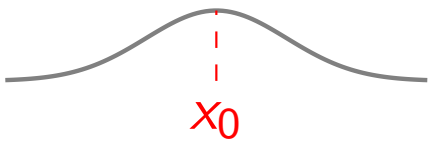
Definition (Symmetry)

A distribution is **symmetric** if there exists x_0 such that:

$$f(x_0 + x) = f(x_0 - x) \quad \forall x.$$

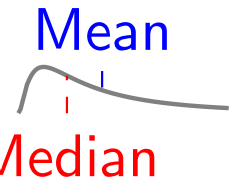
then x_0 is both the mean and the median. Otherwise it is **skewed**.

Symmetric



Shape balanced around center.

Right-Skewed



Mean pulled toward tail.

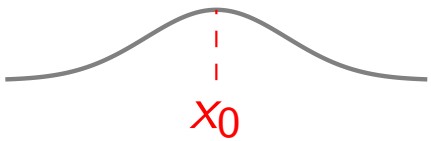
Skewness ($\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$) describes the **asymmetry** of a distribution. It guides the choice between mean and median.

Unimodal vs Multimodal Distributions

Definition (Mode)

A **mode** of a distribution is a value x for which the distribution has a local maximum.

Unimodal



One peak.
Single center.

Multimodal

Multiple peaks.
Several clusters.

Interpretation

The number of modes in a distribution can reveal **subpopulations** or **mixtures** of data.

Light-Tailed vs Heavy-Tailed

Definition (Light-Tailed)

A distribution is **light-tailed** if its probability density function (PDF) decays **exponentially** or faster:

$$\lim_{x \to \pm\infty} e^{|x|} f(x) = 0.$$

Otherwise, it is **heavy-tailed**.

Light-Tailed

$f(x) \sim e^{-x^2}$

Extreme values are rare.
SD is meaningful.

Heavy-Tailed

$f(x) \sim \frac{1}{1+x^2}$

Outliers common or severe.
Median + IQR more reliable.

Interpretation

Tailedness describes the **frequency and severity** of extreme values. It affects robustness of summary statistics.

Choosing the Right Summary

Salaries

Choosing the Right Summary

Salaries

Salaries are typically right-skewed with extreme high values.

- **Median** and **IQR** resist distortion from outliers.

Test scores

Choosing the Right Summary

Salaries

Salaries are typically right-skewed with extreme high values.

- **Median** and **IQR** resist distortion from outliers.

Test scores

Scores are often roughly symmetric.

- **Mean** and **SD** capture both level and variability well.

Survey answers

Choosing the Right Summary

Salaries

Salaries are typically right-skewed with extreme high values.

- **Median** and **IQR** resist distortion from outliers.

Test scores

Scores are often roughly symmetric.

- **Mean** and **SD** capture both level and variability well.

Survey answers

Responses are often qualitative: counts and proportions are the natural summaries.

- **Frequencies** or **mode**.

From Variables To Relationships

Univariate EDA

First, understand each each variable on its own.

- 🔍 How is Feature X distributed?
- ⚠️ Any unusual values in X?
- 🪄 Are transformations helpful?

Box Plot

Five-number summary:

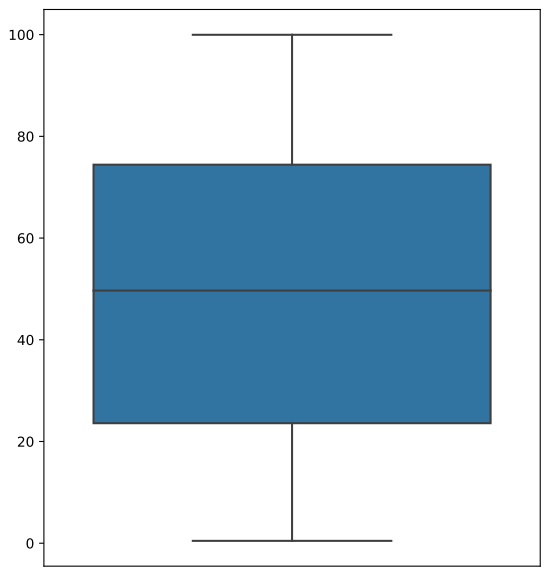
1. Minimum (Q_0)
2. First quartile (25%, Q_1)
3. Median (Q_2)
4. Third quartile (75%, Q_3)
5. Maximum (Q_4)

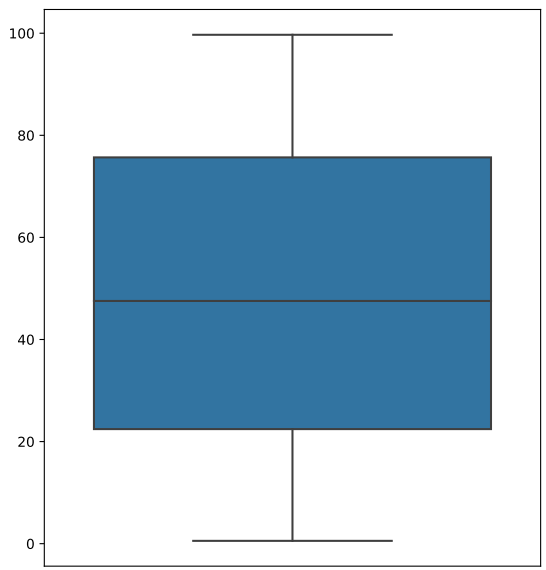
IQR directly: $(Q_3 - Q_1)$.

Sometimes whiskers extend to

- $Q_1 - 1.5 \times \text{IQR}$ and
- $Q_3 + 1.5 \times \text{IQR}$.

Points beyond them are shown as outliers.





Box Plot

Five-number summary:

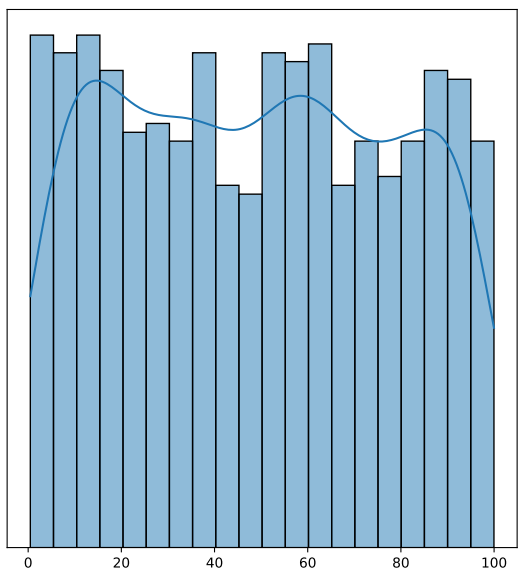
1. Minimum (Q_0)
2. First quartile (25%, Q_1)
3. Median (Q_2)
4. Third quartile (75%, Q_3)
5. Maximum (Q_4)

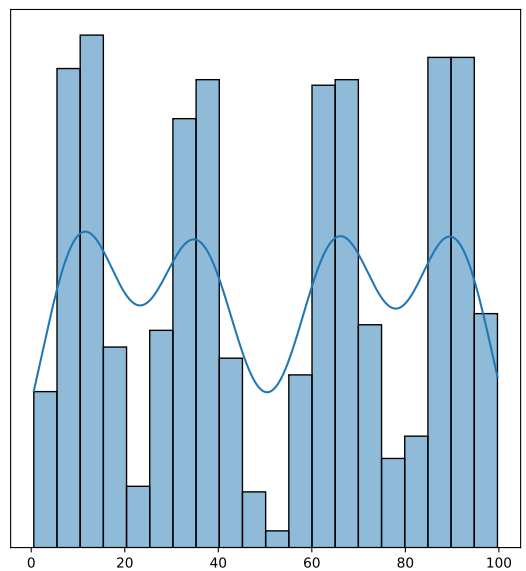
IQR directly: $(Q_3 - Q_1)$.

Sometimes whiskers extend to

- $Q_1 - 1.5 \times \text{IQR}$ and
- $Q_3 + 1.5 \times \text{IQR}$.

Points beyond them are shown as outliers.

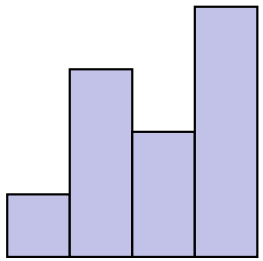




Histogram vs Bar Plot

Histogram

Y-axis = density

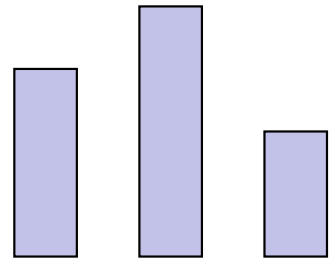


Width = bin size

- Quantitative data.
- Width matters (bin size).

Bar Plot

Y-axis = count/proportion



Width = arbitrary (area = 1)

- Qualitative data.
- Width carries no meaning.

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 to stabilizes proportions and simplifies visuals.

- e.g., countries with $< 1\%$ of the sample $\rightarrow$ "Other".

Transforming Features

Why? To reveal patterns or make summaries more meaningful.

Collapse categories

Merge rare levels to stabilizes proportions and simplifies visuals.

- e.g., countries with $< 1\%$ of the sample $\rightarrow$ "Other".

Log transformation

Apply log to right-skewed variables to reveals structure hidden by extreme values.

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

Transforming Features

Why? To reveal patterns or make summaries more meaningful.

Collapse categories

Merge rare levels to stabilizes proportions and simplifies visuals.

- e.g., countries with $< 1\%$ of the sample $\rightarrow$ "Other".

Log transformation

Apply log to right-skewed variables to reveals structure hidden by extreme values.

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

⚠ Always keep the **original feature**: transformations reveal structure, but also distort.

Bivariate EDA

Second, understand how pairs of variables behave together.

- 🔗 How do X and Y relate?
- 📚 Does X behave differently across Z?

Two Quantitative Features

How should we analyze the relationship between two quantitative features?

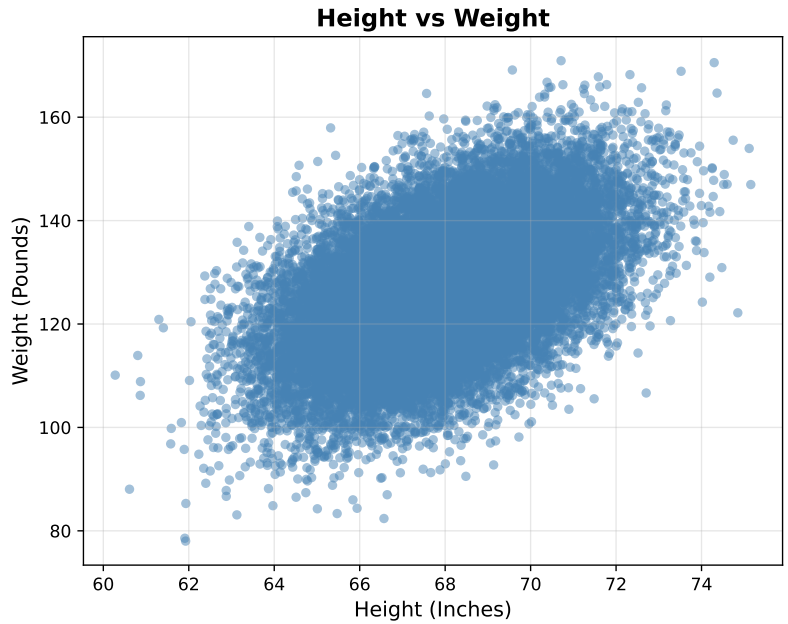
Two Quantitative Features

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

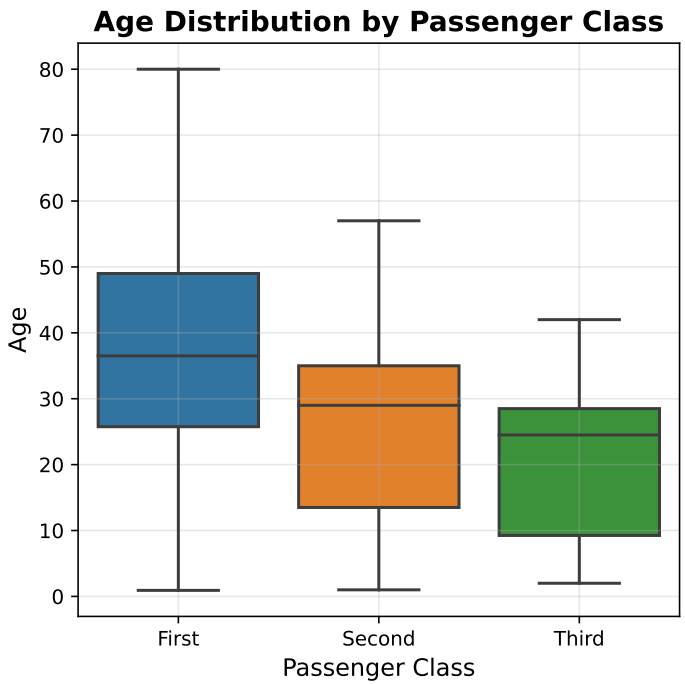
How should we analyze the relationship between a quantitative and a qualitative feature?

Compare **distributions across groups**.

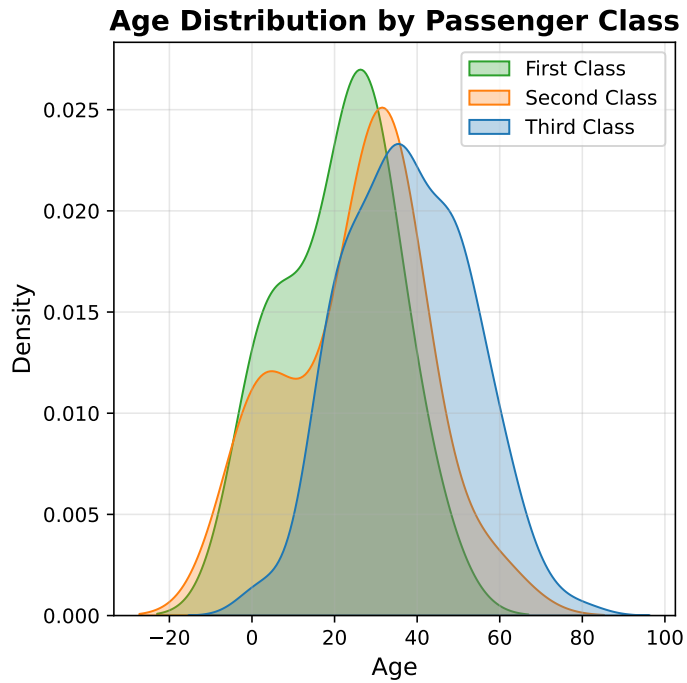
Quantitative vs Qualitative

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

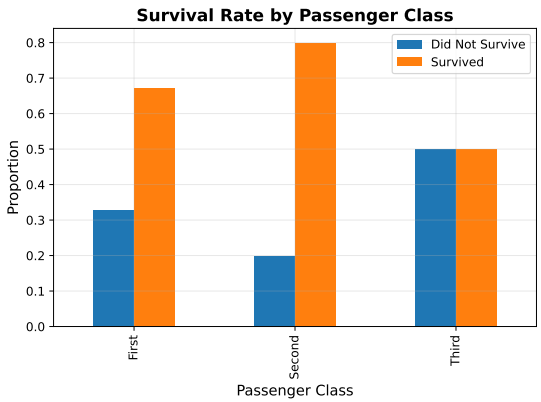
How should we analyze the relationship between two qualitative features?

Compare **proportions across groups**.

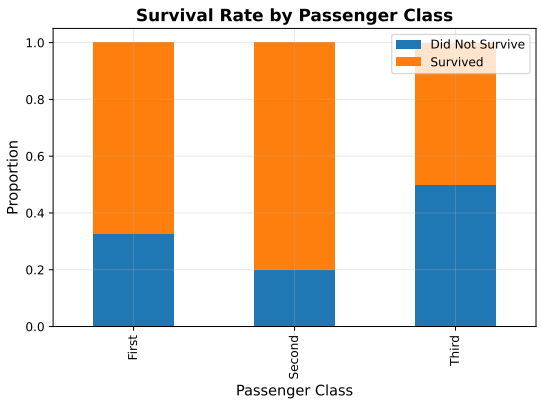
Two Qualitative Features

How should we analyze the relationship between two qualitative features?

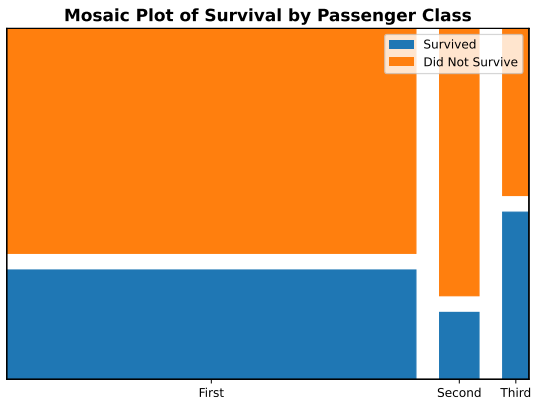
Compare **proportions across groups**.



Bar Charts



Stacked Bar Charts



Mosaic Plot

Look at:

- Changes in composition
- Comparative frequency
- Potential **confounding** (remember data collection).

Multivariate EDA

To examine **three or more variables**, use:

- **Faceting**: multiple subplots by category
- **Color or symbol encodings**
- **Conditioning** on a variable

Multivariate EDA

To examine **three or more variables**, use:

- **Faceting**: multiple subplots by category
- **Color or symbol encodings**
- **Conditioning** on a variable

What challenges arise when we add more variables?

Multivariate EDA

To examine **three or more variables**, use:

- **Faceting**: multiple subplots by category
- **Color or symbol encodings**
- **Conditioning** on a variable

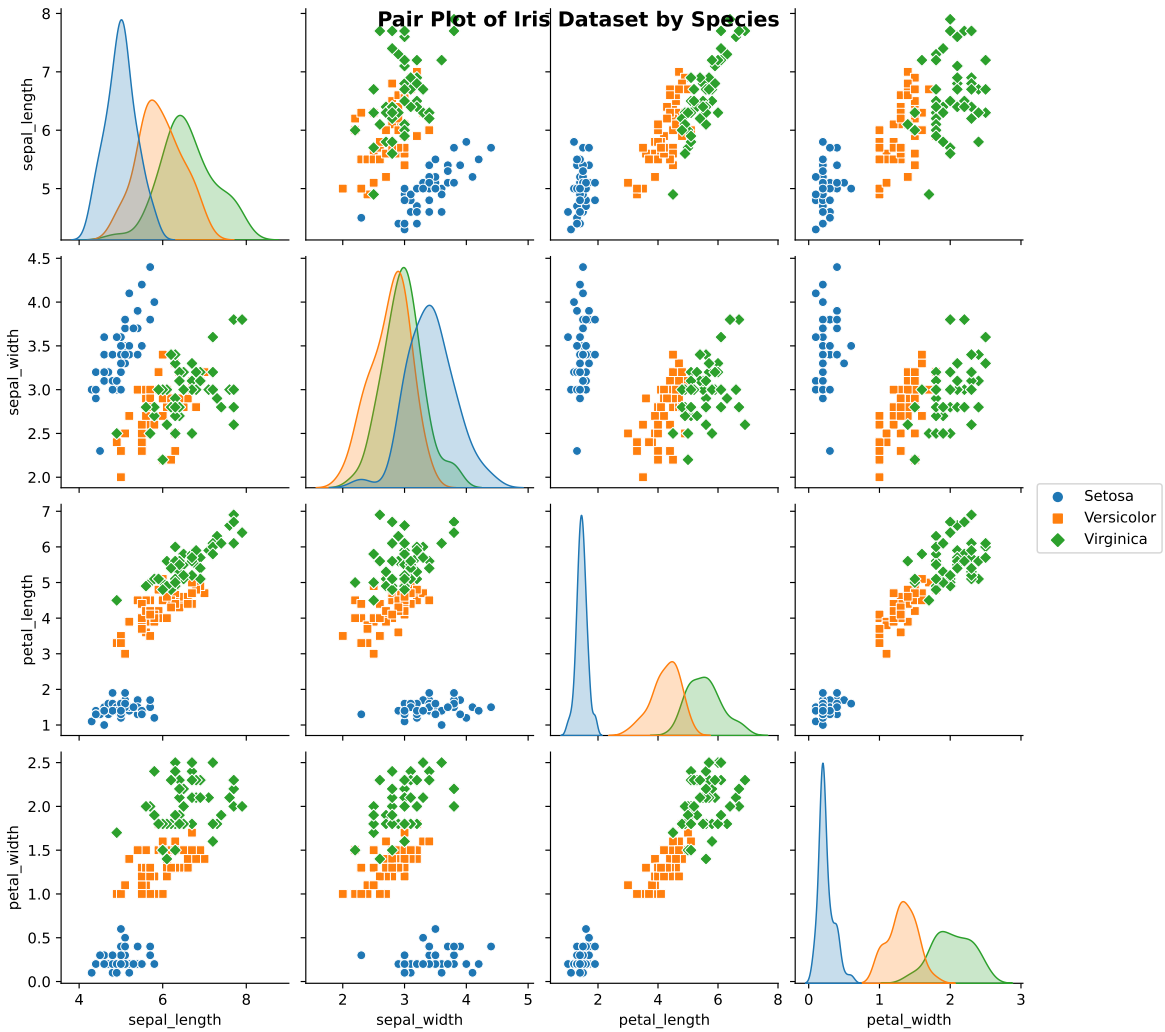
What challenges arise when we add more variables?

⚠ **Curse of dimensionality**: More features → sparser data

⚠ **Small subgroup sizes**: Harder to interpret differences

Pair Plots

- Visualize pairwise relationships between quantitative features
- Scatterplots (off-diagonal) and histograms plots (diagonal)



Correlation Matrices

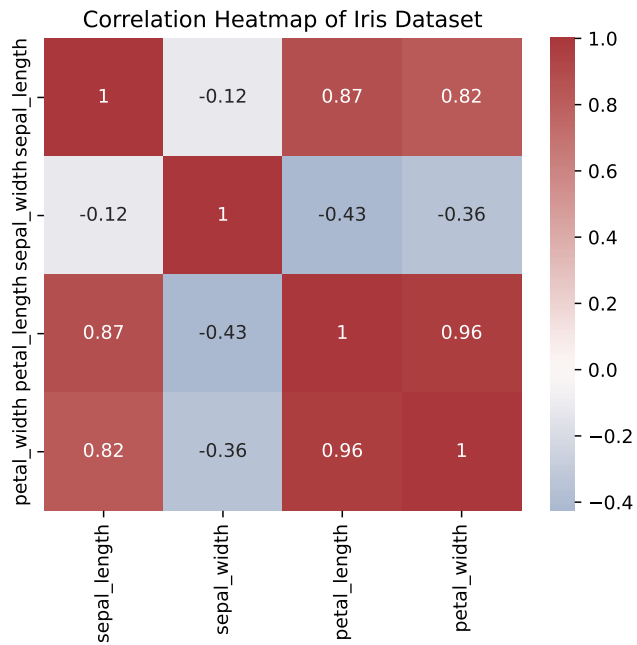
Pearson Correlation Coefficient

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where:

- $\text{Cov}(X, Y)$ is the covariance between X and Y
- σ_X and σ_Y are the standard deviations of X and Y

- Quantitative measure of **linear relationships** between features
- Values range from -1 (perfect negative correlation) to $+1$ (perfect positive correlation)



Anomalies and Pitfalls

Detecting Data Problems

EDA helps detect and address:

Data Problems

- **Impossible values** (e.g., negative age).
- **Miscoded categories** (e.g., "Male" coded as 3).
- **Gaps** indicating missing data.
- **Duplicates** or incorrect granularity.

Structural Issues

- **Suspicious patterns** (e.g., unexpected trends).
- **Mixed types** (e.g., numbers stored as text).
- **Missingness mechanisms** (e.g., not at random).
- **Strange subpopulations** (e.g., outliers).

Use EDA **early and often** to catch and address these issues!

Simpson's Paradox[1]

Definition (Simpson's Paradox)

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

[1] E. 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 Paradox[1]

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.

[1] E. 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 Paradox[1]

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!

[1] E. 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 Paradox[1]

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**!

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

Limitations of EDA

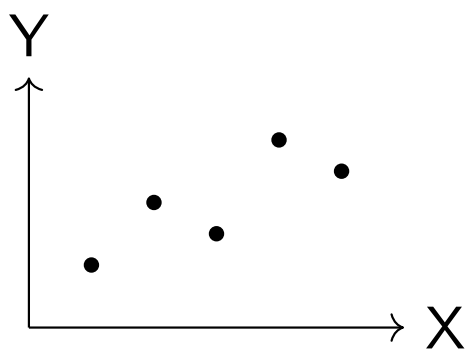
Be cautious about:

- **Overinterpreting noise**: Not every pattern is meaningful.
- **Data dredging**: Testing too many hypotheses.
- **Hindsight bias**: "I knew it all along" effect.
- **Unreported choices**: Lack of transparency.

Always keep a **notebook trail or log** for reproducibility.

Always Look at Joint Behavior

- Not just individual distributions.
- Visualizations (e.g., scatterplots) **guide interpretation**.
- Relationships can be **misleading** if you ignore a third variable.



Joint behavior reveals patterns!

EDA Workflow

Iterative Workflow

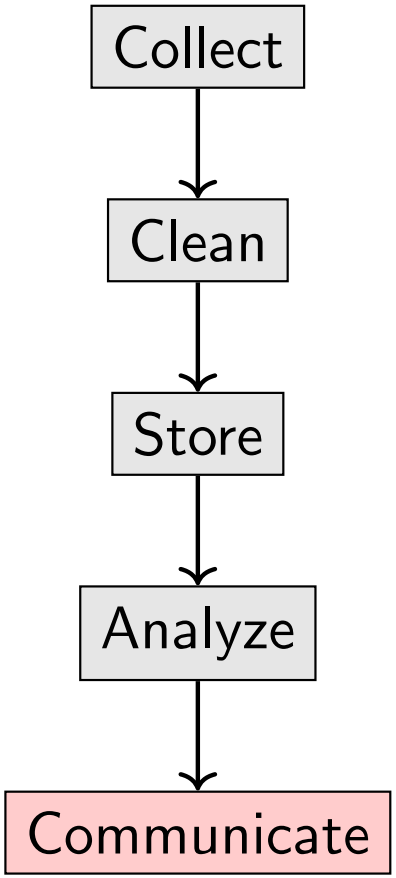
1. Identify feature types.
2. Inspect distributions.
3. Examine relationships.
4. Transform as needed.
5. Compare across subgroups.
6. Investigate anomalies.
7. Document everything.

Guiding Questions

- What do I see?
- Why does it matter?
- What should I look at next?

Communication

Within the lifecycle



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

Principles of Communication

1. Know Your Audience: Adjust Language and Depth

- Technical vs. non-technical stakeholders
- Decision-makers vs. implementers

2. Tell a Story

- Start with the key insight
- Provide context and relevance
- Use a logical flow: problem $\rightarrow$ analysis $\rightarrow$ insight $\rightarrow$ action

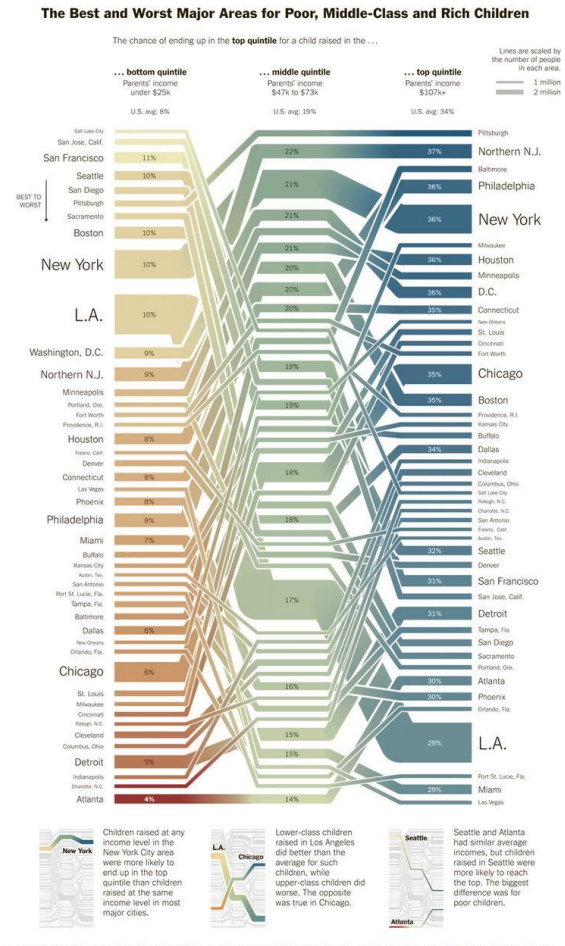
The Good,

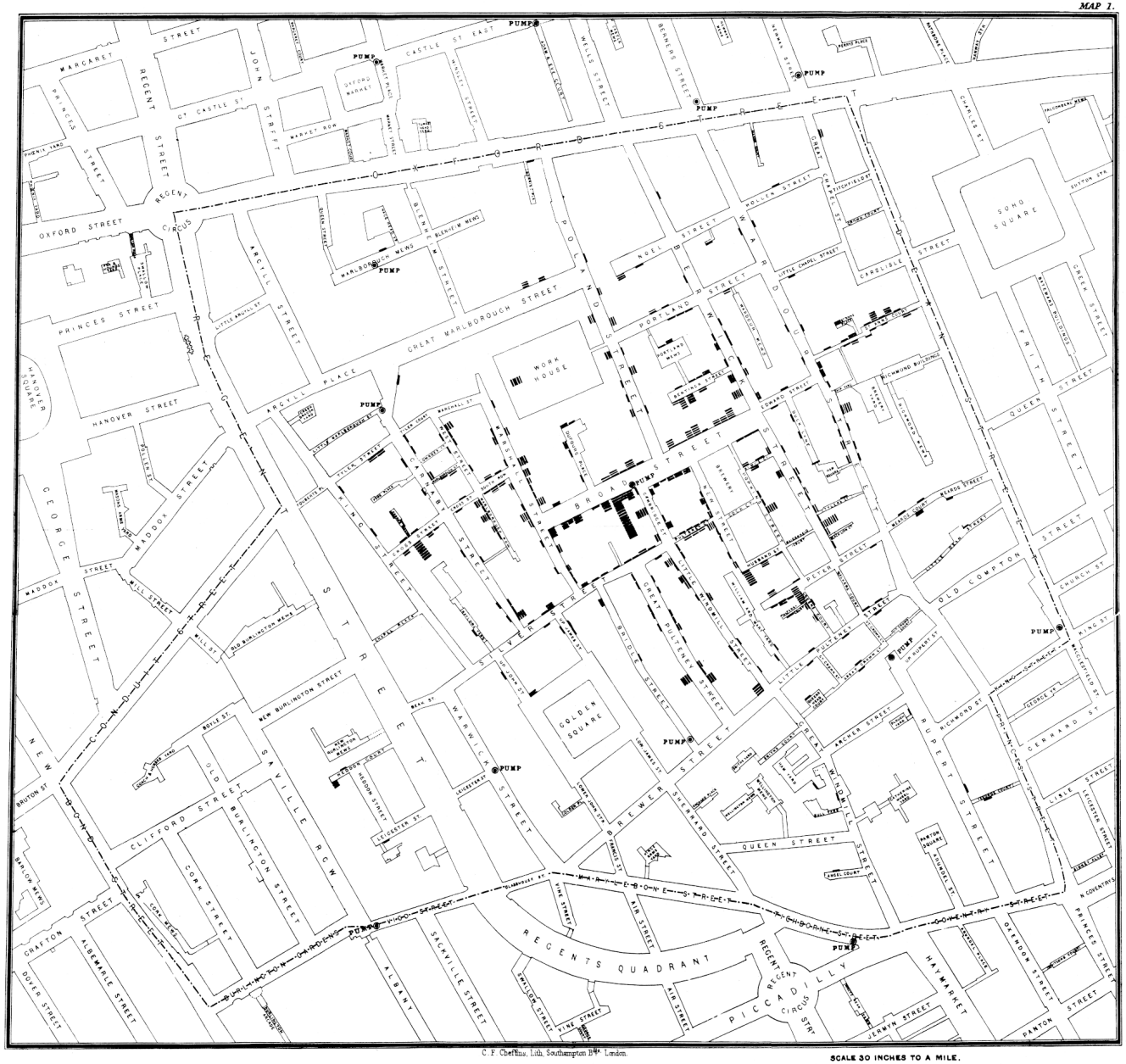
"Our analysis shows a strong relationship between X and Y, suggesting that [action] could improve [outcome]."

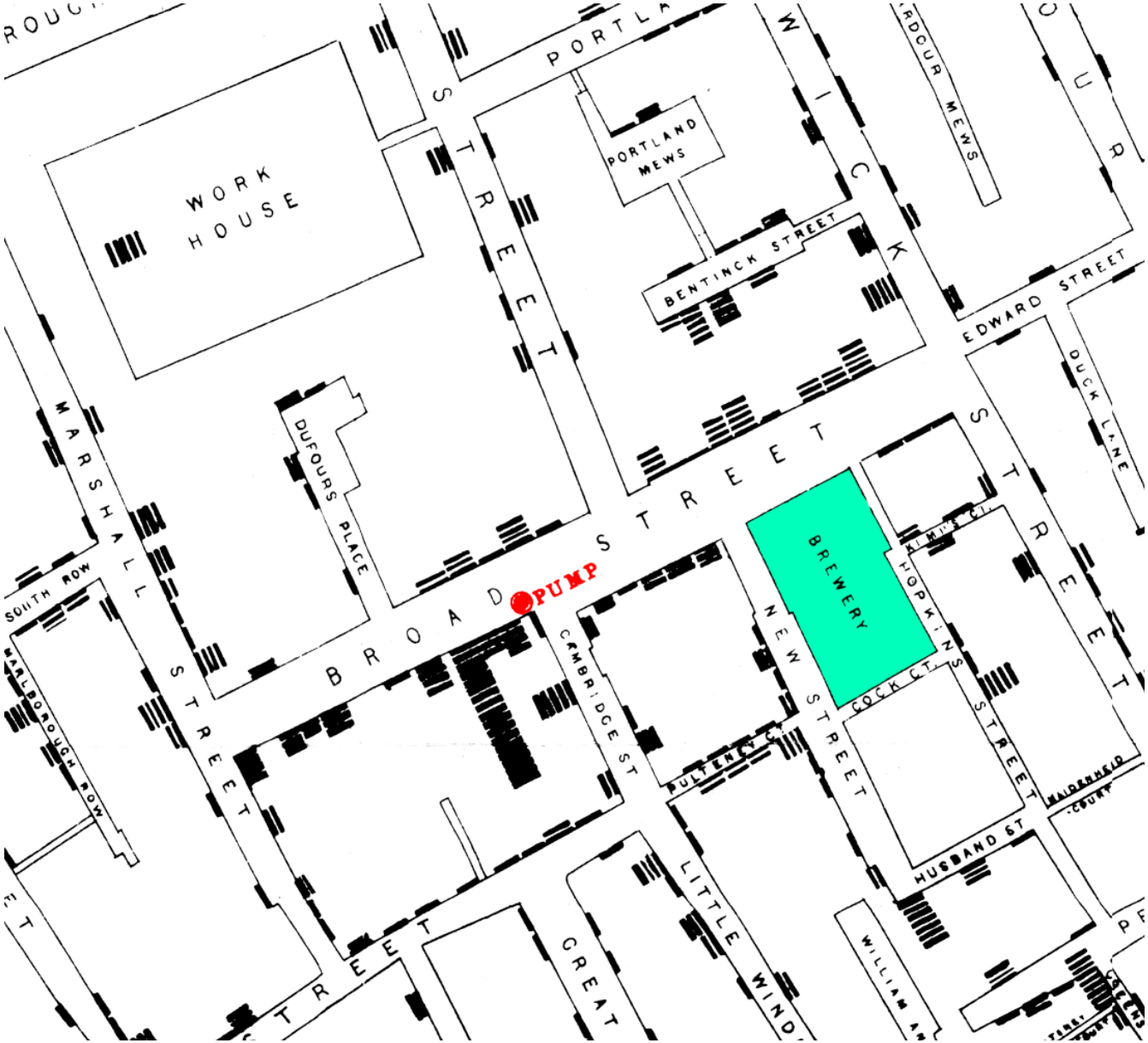
the Bad

"The correlation coefficient is 0.76."

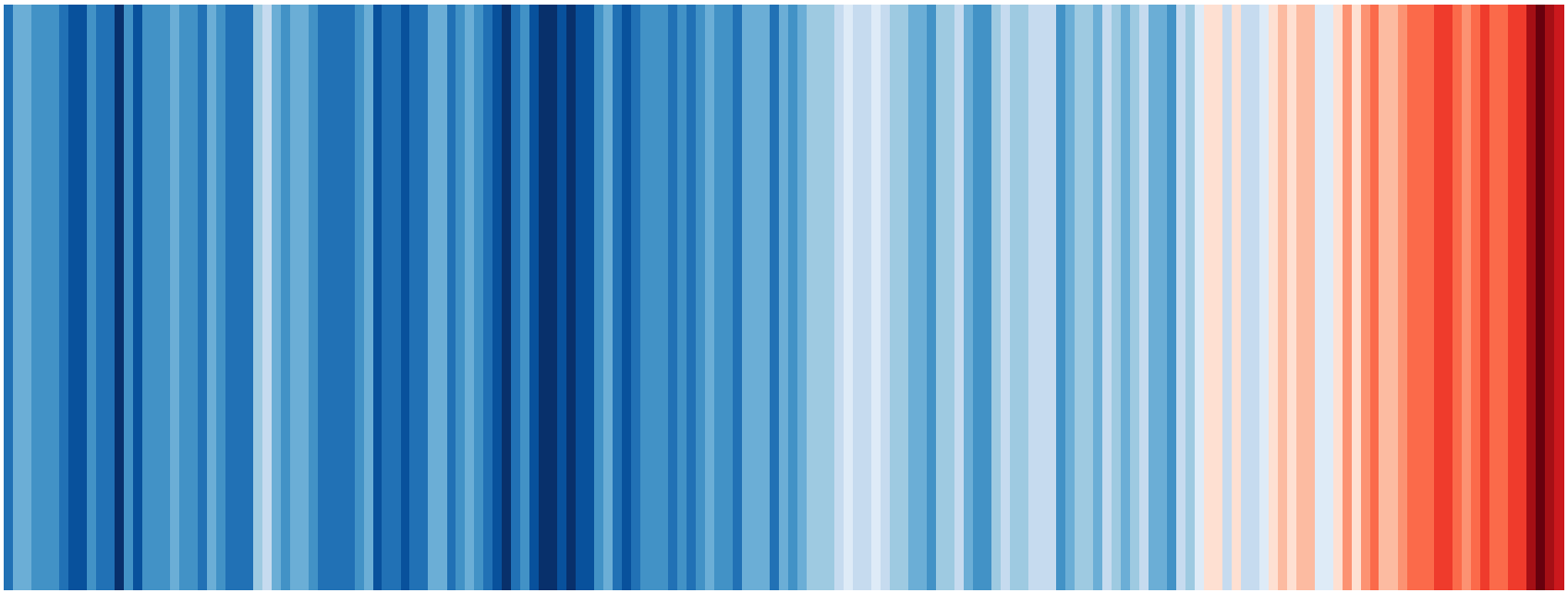
and the Ugly

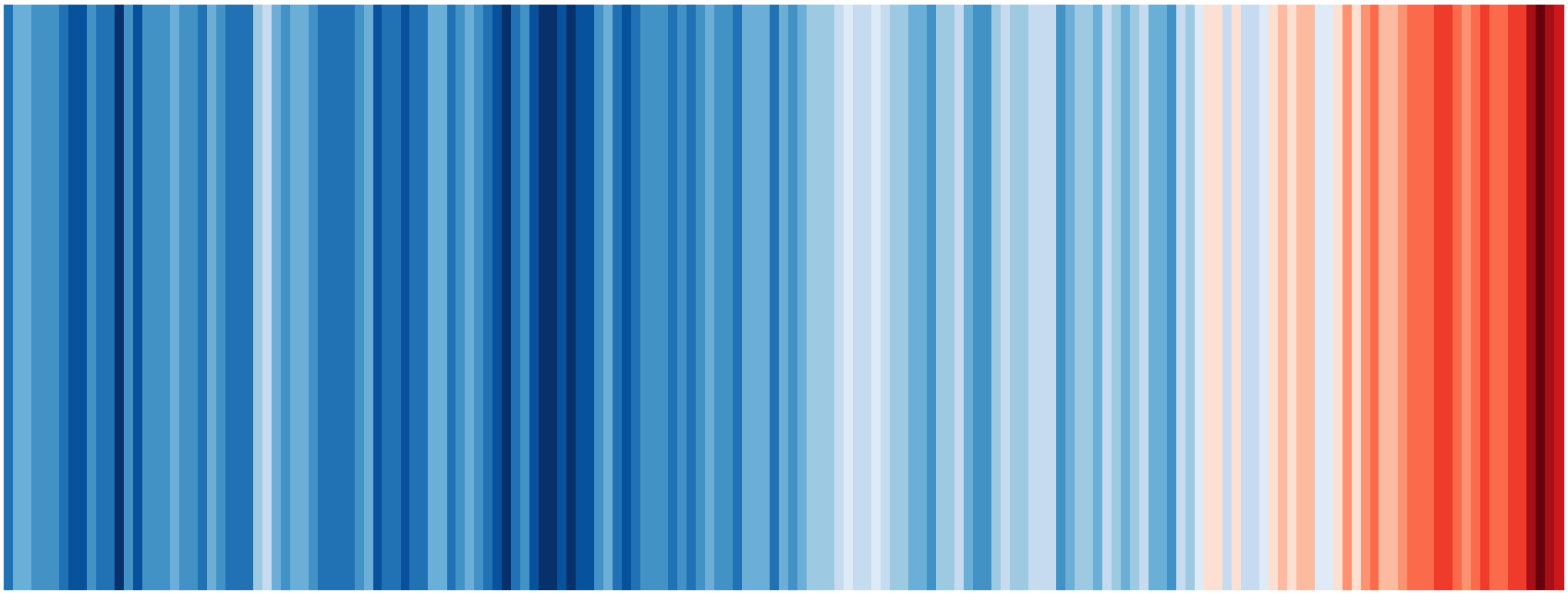



MAP 1.
BROAD STREET
GOLDEN SQUARE
REGENTS QUADRANT
PICCADILLY
OXFORD STREET
REGENT STREET
GREAT MARLBOROUGH STREET
PRINCES STREET
HANOVER STREET
CARLISLE STREET
QUEEN STREET
OLD COMPTON
SOHO SQUARE
WORK HOUSE
PUMP
C. F. Cheffins, Lith. Southampton Bldgs. London.
SCALE 30 INCHES TO A MILE.

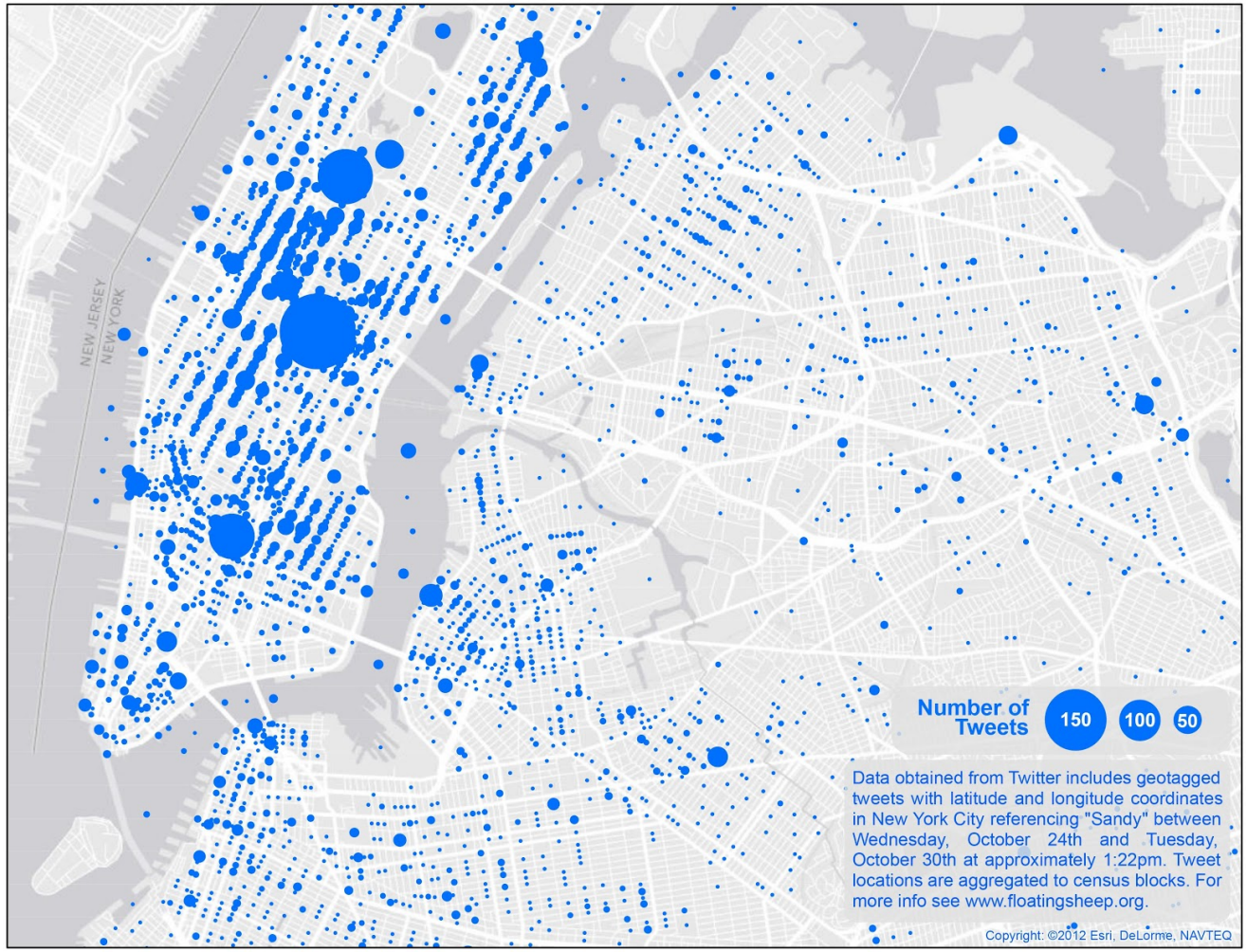


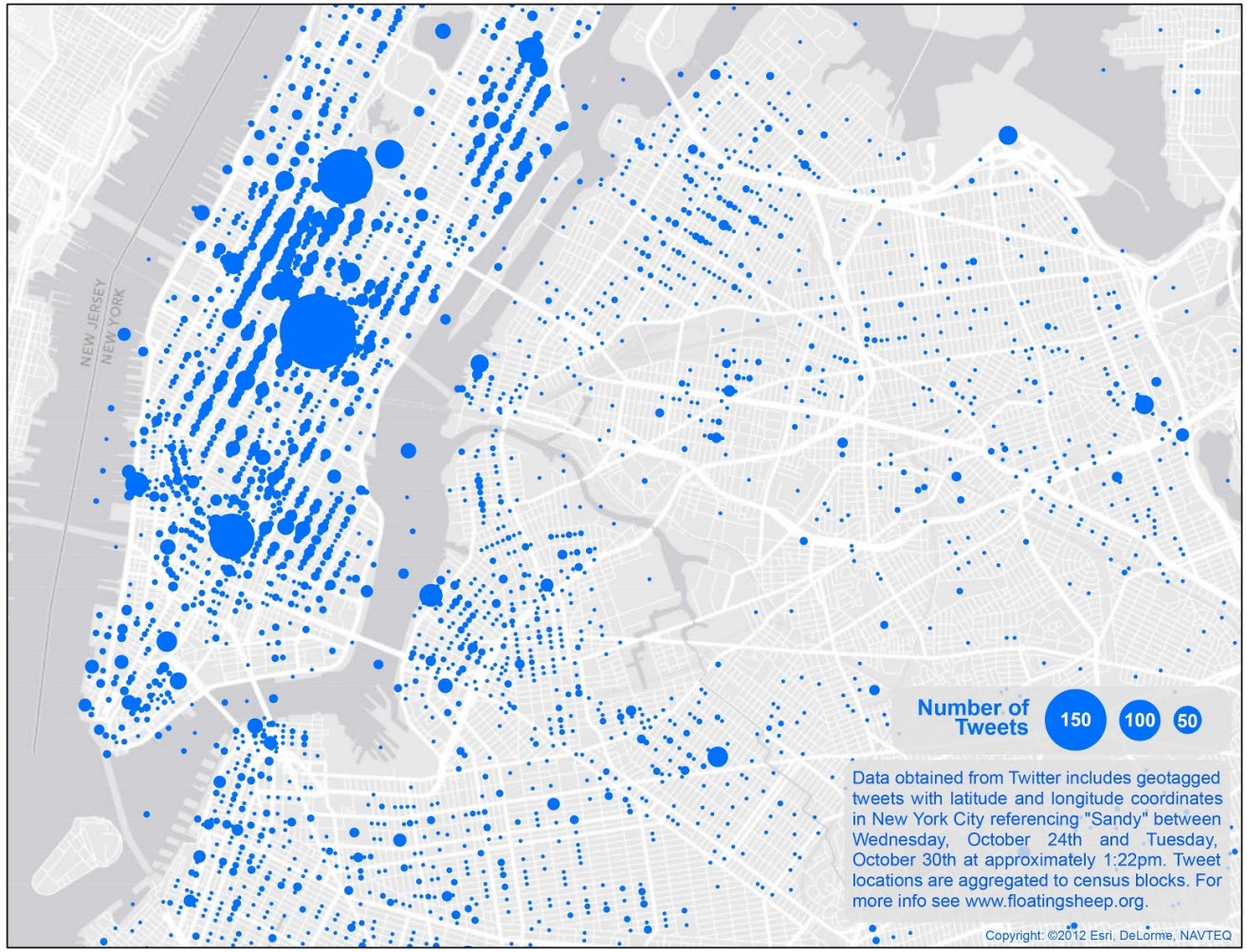
1854 Broad Street cholera outbreak by John Snow





Warming Stripes (1850 – 2018) by Ed Hawkins


NEW JERSEY
NEW YORK
Number of Tweets
150
100
50
Data obtained from Twitter includes geotagged tweets with latitude and longitude coordinates in New York City referencing "Sandy" between Wednesday, October 24th and Tuesday, October 30th at approximately 1:22pm. Tweet locations are aggregated to census blocks. For more info see www.floatingsheep.org.
Copyright: ©2012 Esri, DeLorme, NAVTEQ



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

Visualization serves different purposes at different stages

Analysis

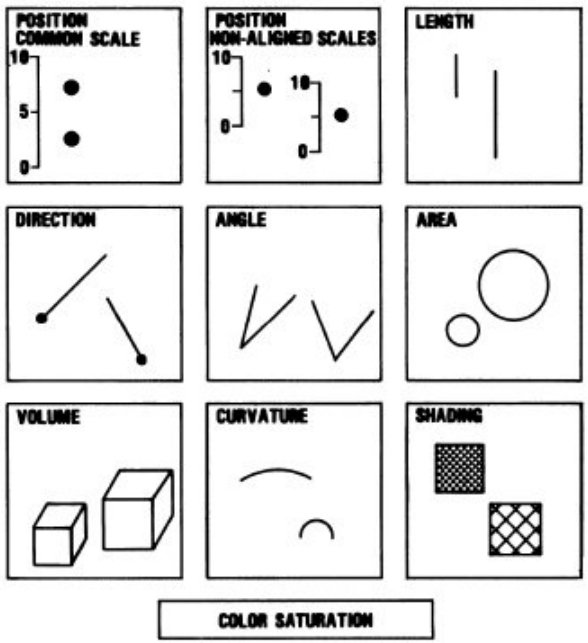
- Discover relationships, distributions, comparisons
- Use exploratory tools (e.g., histograms, scatterplots)
- Focus on insight generation
- "Good enough" visuals

Communication

- Present final results
- Clear, understandable communication
- Chart type, layout, and design matter
- Visuals must "speak for themselves"

The Cleveland-McGill[2] Ranking of Graphical Perceptions

How effectively do we perceive different visual encodings?



Practical Implications

- Prefer bar charts over pie charts for comparisons
- Use line charts for trends over time
- Avoid 3D charts that distort perception

[2] William 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 ratio[1]
- Avoid chart junk[1]

Scores of our suppliers:

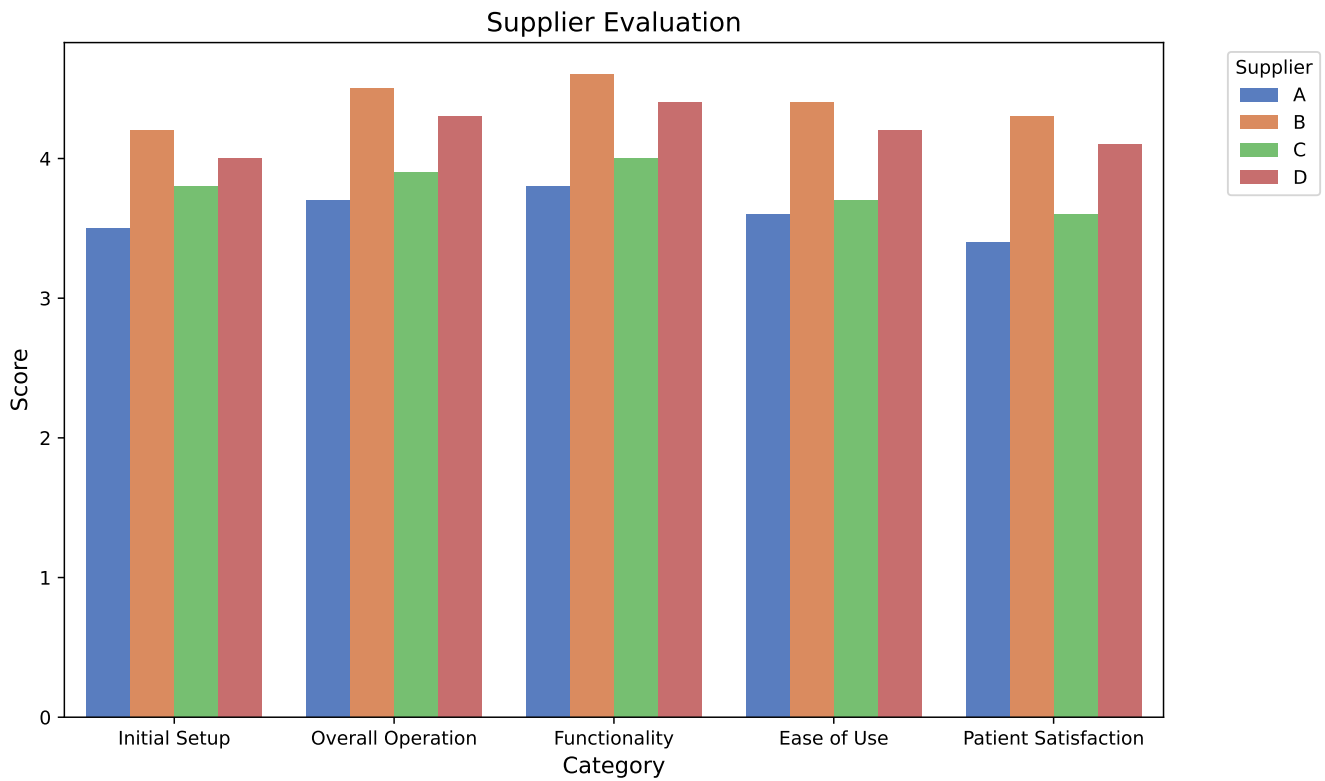
Criterion	A	B	C	D
Initial Setup	3.5	4.2	3.8	4.0
Overall Operation	3.7	4.5	3.9	4.3
Functionality	3.8	4.6	4.0	4.4
Ease of Use	3.6	4.4	3.7	4.2
Patient Satisfaction	3.4	4.3	3.6	4.1

[1] Edward 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 ratio[1]
- Avoid chart junk[1]

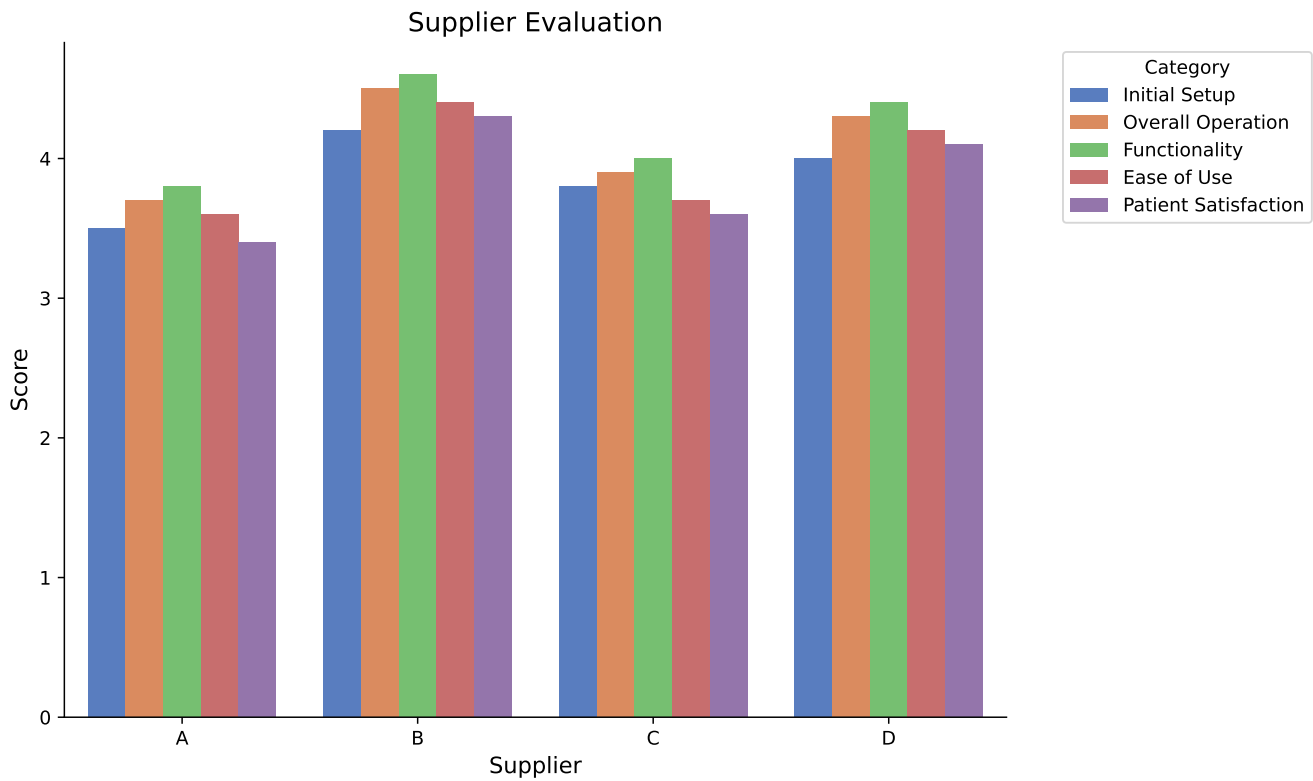


[1] Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio[1]
- Avoid chart junk[1]

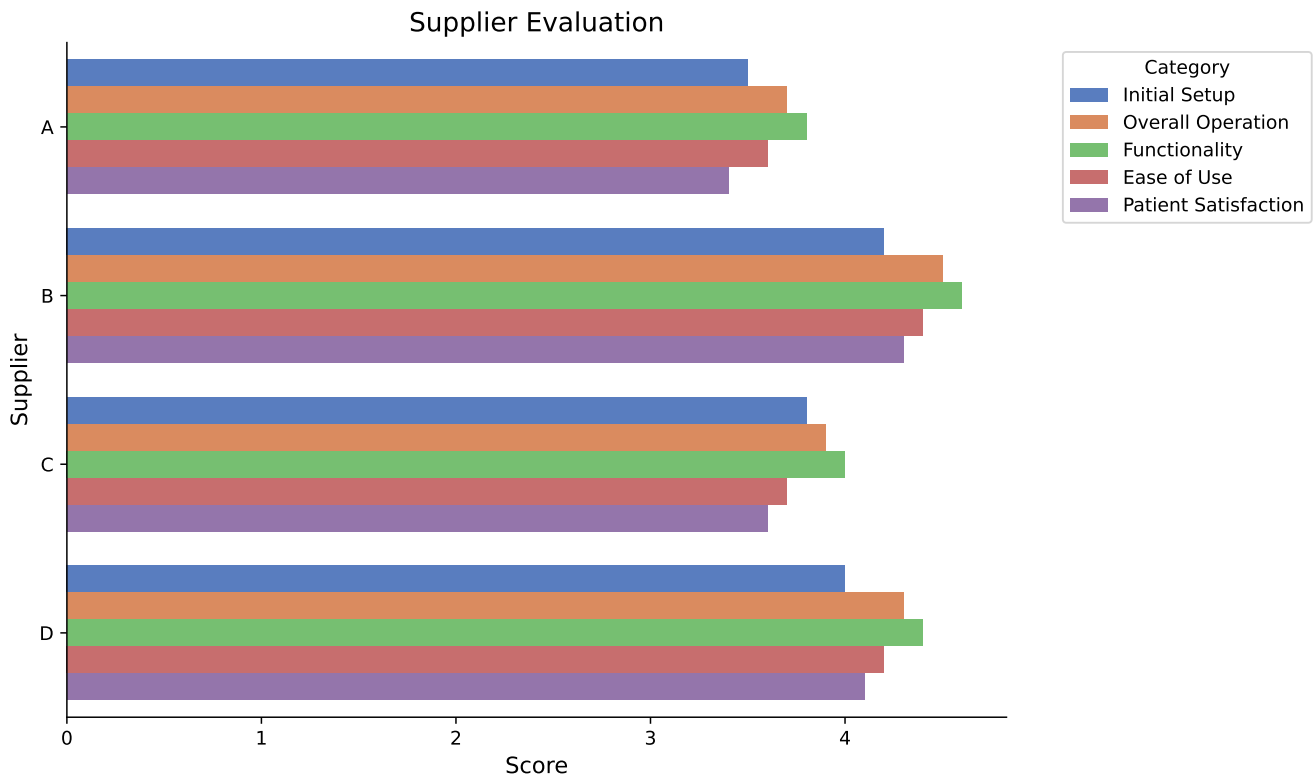


[1] Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio[1]
- Avoid chart junk[1]

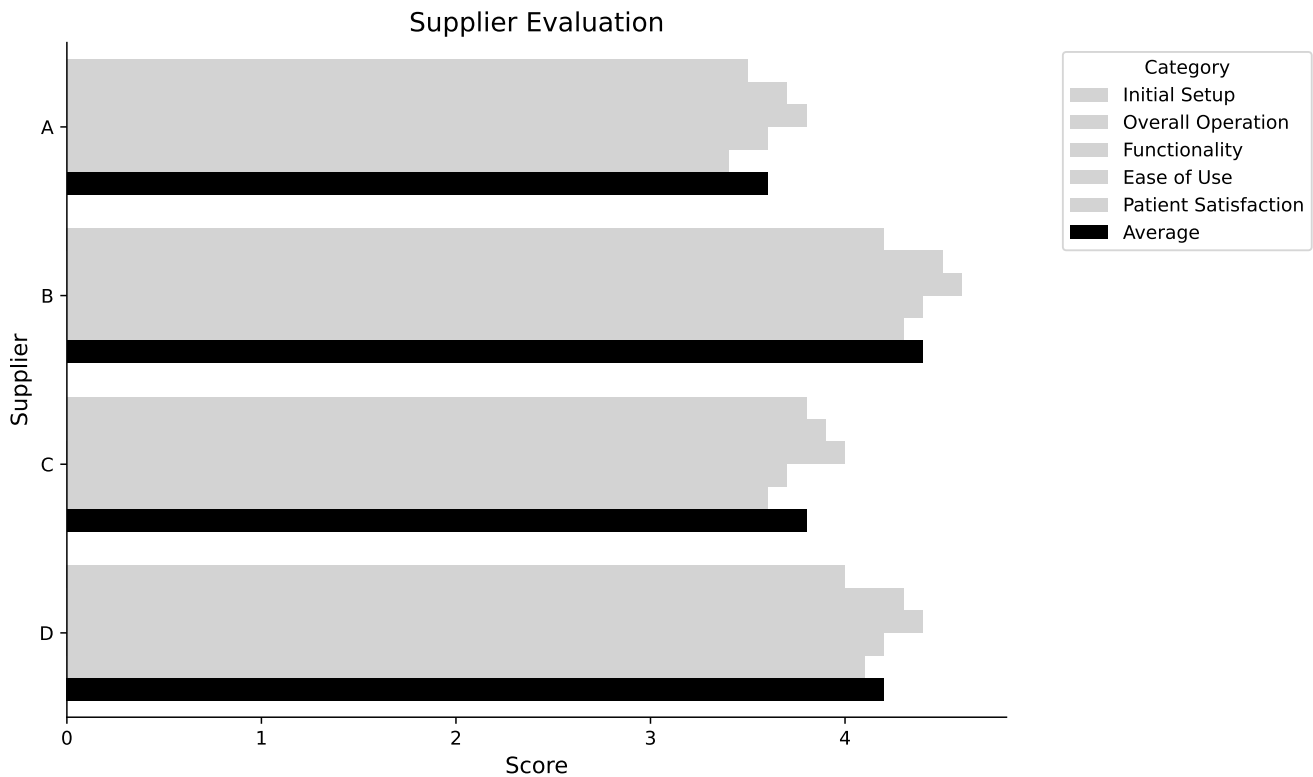


[1] Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio[1]
- Avoid chart junk[1]

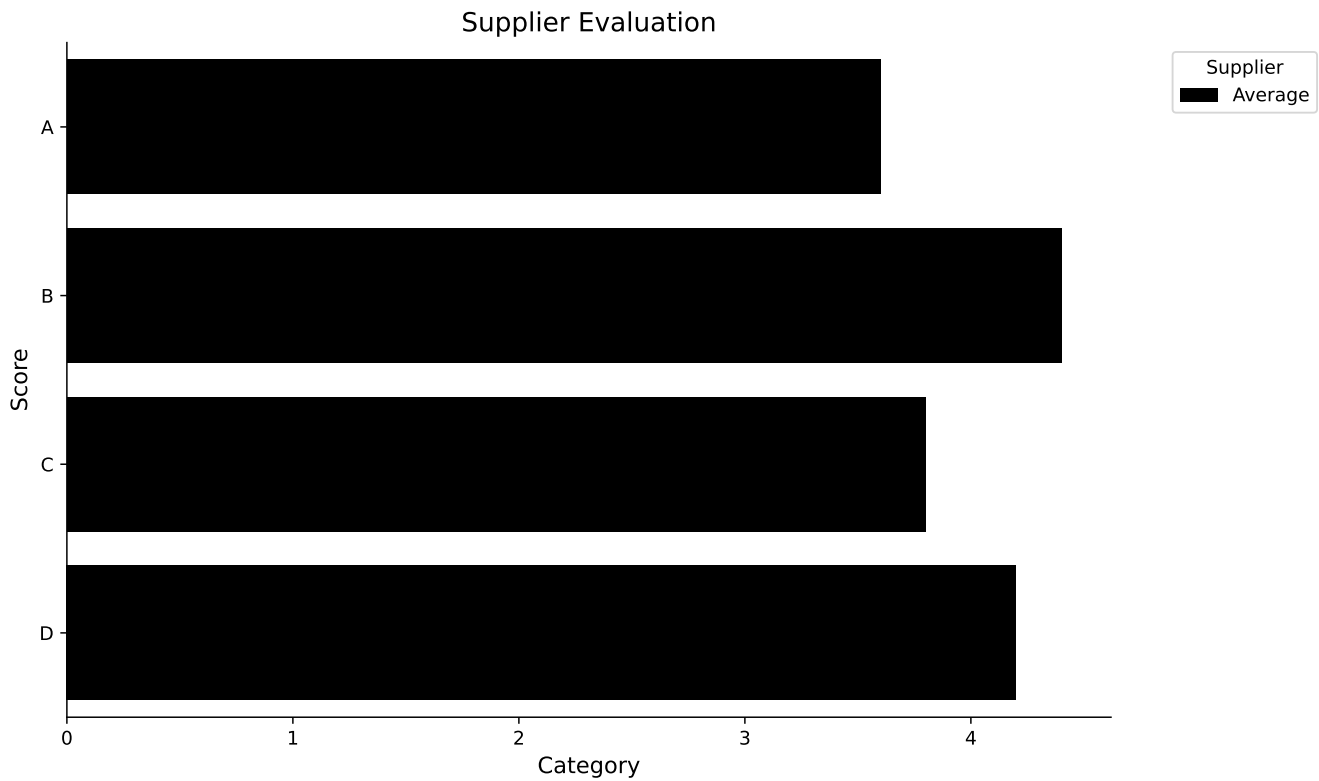


[1] Edward R. Tufte. **The Visual Display of Quantitative Information**. 2nd. Cheshire, Conn: Graphics Pr, 2001

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio[1]
- Avoid chart junk[1]



[1] Edward 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 ratio[1]
- Avoid chart junk[1]

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

Data Storytelling Framework

1. Set the context
   "Our customer churn has increased by 15% this quarter."

Data Storytelling Framework

1. Set the context
   "Our customer churn has increased by 15% this quarter."
2. Ask the question
   "What factors are contributing to this increase?"

Data Storytelling Framework

1. Set the context
   "Our customer churn has increased by 15% this quarter."
2. Ask the question
   "What factors are contributing to this increase?"
3. Reveal evidence incrementally and logically
   "60% of churned customers had contacted support 3+ times."

Data Storytelling Framework

1. Set the context
   "Our customer churn has increased by 15% this quarter."
2. Ask the question
   "What factors are contributing to this increase?"
3. Reveal evidence incrementally and logically
   "60% of churned customers had contacted support 3+ times."
4. Reveal the insight
   "This suggests customer service issues are a major driver of churn."

Data Storytelling Framework

1. Set the context
   "Our customer churn has increased by 15% this quarter."
2. Ask the question
   "What factors are contributing to this increase?"
3. Reveal evidence incrementally and logically
   "60% of churned customers had contacted support 3+ times."
4. Reveal the insight
   "This suggests customer service issues are a major driver of churn."
5. Recommend action
   "We recommend a customer success program for at-risk accounts."

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

Communicating Uncertainty

- Data is never exact.
- Decisions depend on understanding confidence.
- Transparency earns trust.

Communicating Uncertainty

- Data is never exact.
- Decisions depend on understanding confidence.
- Transparency earns trust.

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)

Ethical Communication of Data

Data communication influences decisions.

Avoid:

- **Misleading axes** (truncated or inconsistent scales)
- **Cherry-picking** supportive subsets
- **Over-smoothing** to hide variability
- **Hiding uncertainty** or assumptions
- **Opaque methods** (unclear steps, missing context)

Ethical Communication of Data

Data communication influences decisions.

Avoid:

- **Misleading axes** (truncated or inconsistent scales)
- **Cherry-picking** supportive subsets
- **Over-smoothing** to hide variability
- **Hiding uncertainty** or assumptions
- **Opaque methods** (unclear steps, missing context)

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

Conclusion

Takeaways: Exploratory Data Analysis

1. EDA is an iterative, question-driven process centered on discovery.
2. Understanding feature types is essential to choosing valid summaries and visualizations.
3. Histograms, scatterplots, box plots, densities form the core toolkit.
4. Always examine distributions before relationships.
5. Documenting each step ensures transparency and reproducibility.

Takeaways: Communication

1. Tailor your message to the audience and the decision at hand.
2. Visualizations for communication must be clear, purposeful, yet minimal.
3. Highlight uncertainty and limitations: never hide them.
4. Build a narrative to convince your audience.
5. Ethical communication means showing the data *as it is*, not as we wish it to be.

Why should we care about Data?

What is data and what we can do with it