Lecture 2:

Data Visualization

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com

Lecture Contents

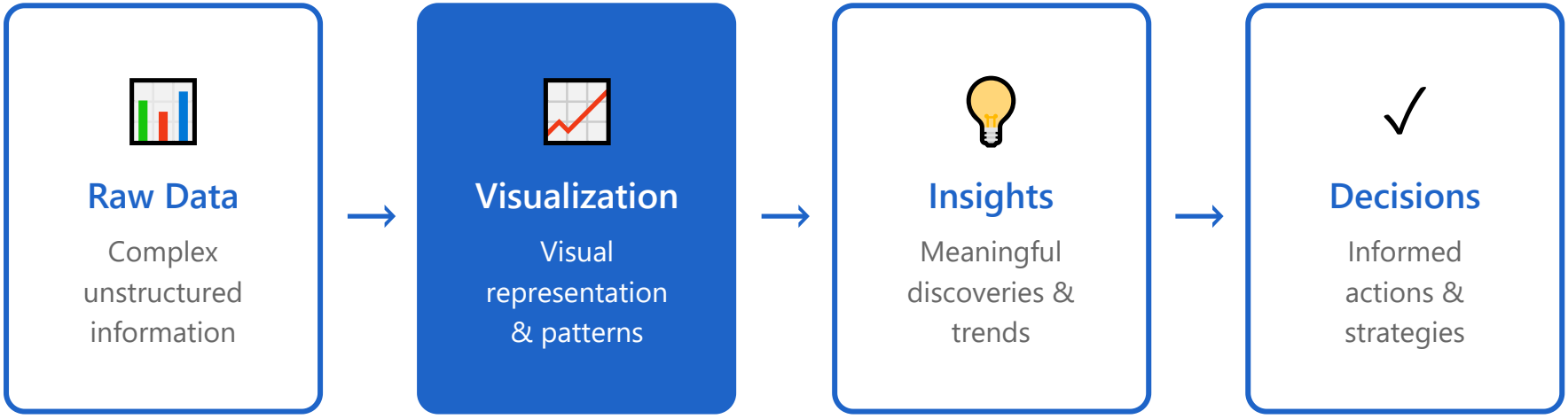
Part 1/3:

Visualization Fundamentals

1. Importance and Goals of Data Visualization
2. Visual Encoding Principles
3. Gestalt Principles and Perception
4. Color Theory and Color Blindness Accessibility
5. Typography and Layout
6. Information Density and Data-Ink Ratio
7. Grammar of Graphics
8. Good vs Bad Visualization Examples

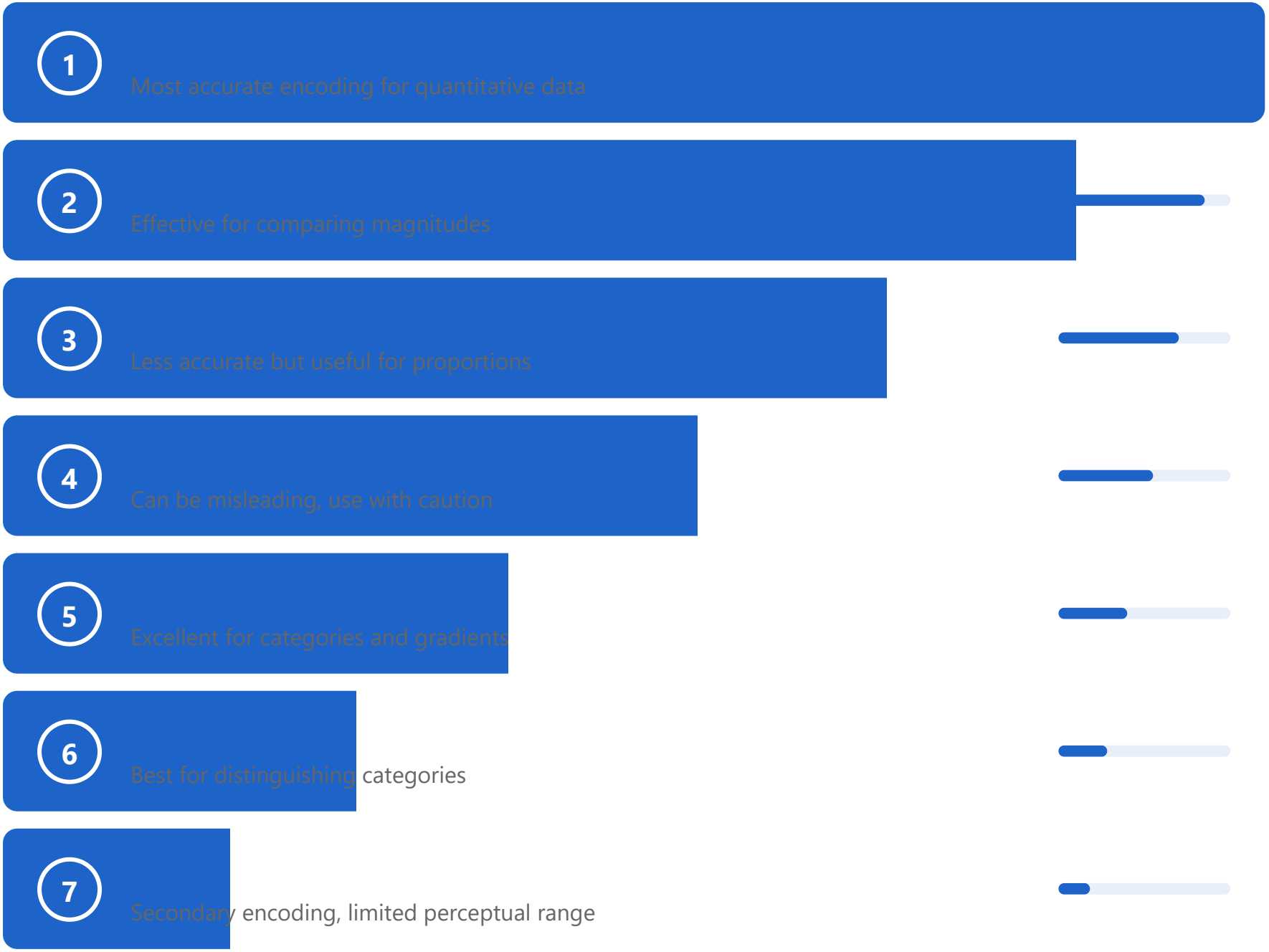
Data Visualization Process

Transform Data into Actionable Insights



Visual Encoding Hierarchy

Effectiveness Ranking from Most to Least Accurate



Gestalt Principles and Perception

Visual Organization Principles for Effective Data Visualization

Proximity

Objects close together are grouped

Similarity

Similar elements grouped together

Continuity

Eyes follow smooth paths

Closure

Mind completes shapes

Figure-Ground

Objects vs background

Common Fate

Moving together = related

Symmetry

Balanced compositions

Application

Guide viewer attention

Color Theory and Accessibility

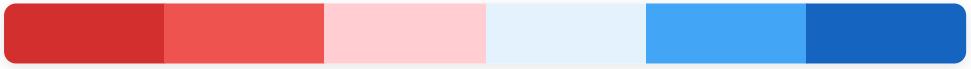
Choosing the Right Color Palette for Inclusive Visualization

S Sequential

Single hue, varying lightness (temperatures)

D Diverging

Two hues meeting at neutral (gains/losses)



C Categorical

Distinct hues for different categories

~8%

of males have color vision deficiency

⚠ Accessibility Guidelines

- **Avoid** red-green combinations
- Use **ColorBrewer** or **Viridis** palettes
- Add **texture/patterns** as redundant encoding
- Test with **color blindness simulators**

✓ Color Contrast Examples

Typography and Layout Principles

Creating Clear and Professional Visualizations

Visual Hierarchy

TITLE (24PT)

Main Visualization Title

SUBTITLE (20PT)

Supporting Context

BODY (16PT)

Detailed information and annotations with proper spacing and clarity

Grid System & White Space

Chart | Legend | Metrics

Key Principles

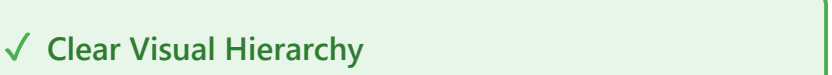
1. **Sans-serif** for digital, **serif** for print
2. **12-14pt minimum** for readability
3. **2-3 font families** maximum
4. **Consistent alignment** is professional

✓ Best Practices

- ✓ White space improves clarity
- ✓ Grid systems ensure balance
- ✓ Annotations enhance content
- ✗ Avoid cluttering with text

Concrete Examples: Good Design vs Bad Design

✓ Clear Visual Hierarchy



✗ Unclear Visual Hierarchy

Main Title

Subtitle

Body text is smallest and organized for easy reading.

Size and weight express information priority

Main Title

Subtitle

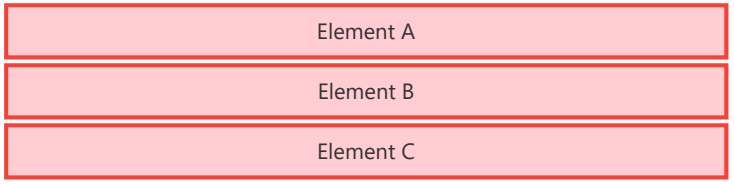
Body text is smallest and organized for easy reading.

Mixed sizes make it hard to identify priority

✓ Proper Spacing

Adequate spacing makes each element clear

✗ Insufficient Spacing



Cramped spacing feels cluttered and hard to read

Data-Ink Ratio & Information Density

Maximize Data-Ink, Minimize Chartjunk (Tufte's Principle)

Key Principles to Remember

- Remove chartjunk & decorations
- Eliminate redundant labels

High density ≠ cluttered

Iterate until nothing can be removed

Grammar of Graphics

A Systematic Framework for Building Visualizations

1. **Data**
   Foundation: Raw dataset to visualize
2. **Aesthetics**
   Map data to visual properties (x, y, color)
3. **Geometries**
   Visual marks (points, lines, bars)
4. **Statistics**
   Transform data (mean, count, regression)
5. **Scales**
   Control mapping from data to aesthetics
6. **Coordinates**
   Map to 2D plane (Cartesian, polar)
7. **Facets**

Framework

ggplot2

by Leland Wilkinson

Key Concepts

- Build graphics incrementally
- Separates data from visuals
- Systematic & composable
- Layer-based architecture

Split data into multiple subplots

Step-by-Step Implementation

Building Visualizations Layer by Layer

📊 Practical Example: Car Fuel Efficiency Analysis

Using the mtcars dataset, we'll visualize the relationship between car weight (wt) and fuel efficiency (mpg). Watch how each layer progressively builds the visualization from an empty canvas to a complete, insightful graph.

⬇️

✨ Complete Visualization Achieved

By layering all 7 components, we've built a sophisticated multi-dimensional visualization that reveals:

📌 **Key Insights:**

- Heavier cars have worse fuel efficiency (negative correlation)
- More cylinders → heavier cars → worse MPG (shown by red colors)
- Manual transmission cars are generally lighter and more efficient

This is the power of Grammar of Graphics: systematic, composable, and infinitely flexible!

Good vs Bad Visualization Examples

Common Pitfalls and Best Practices

1 Chart Type Selection

✗ Bad		✓ Good
3D pie charts distort proportions and values	vs	Simple 2D bar charts for clear comparisons

2 Axis Management

✗ Bad		✓ Good
Dual y-axes mislead by arbitrary scaling	vs	Normalized or separate charts for clarity

3 Y-Axis Baseline

✗ Bad		✓ Good
Truncated y-axis exaggerates differences	vs	Start at zero or clearly indicate breaks

4 Color Application

✗ Bad		✓ Good
Too many colors causing confusion	vs	Intentional color use highlighting key insights

Part 2/3:

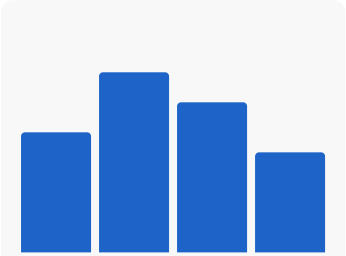
Mastering Basic Charts

9. 1D Data - Histogram, KDE
10. 2D Relationships - Scatter Plot, Bubble Chart
11. Categorical Data - Bar Chart, Pie Chart
12. Distribution Comparison - Boxplot, Violin Plot
13. Time Series - Line Graph, Area Chart
14. Correlation - Heatmap, Correlation Matrix
15. Multidimensional - Parallel Coordinates, Radar Chart
16. Geographic Data - Choropleth, Bubble Map

1D Data: Histogram & KDE

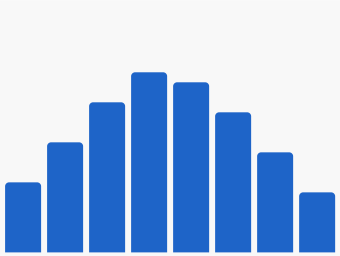
Understanding Distribution Through Binning and Smoothing

Few Bins



⚠ Loses detail

Optimal Bins



✓ Clear pattern

KDE Curve

✓ Smooth & continuous

📊 Key Concepts

- Histogram: Binned frequency distribution
- KDE: Smooth density curve
- Bandwidth affects smoothness
- Identify: normal, skewed, bimodal
- Overlay multiple distributions

📐 Bin Size Rules

Sturges' Rule: $\text{bins} \approx \log_2(n) + 1$

Freedman-Diaconis: Based on IQR & sample size

When to Use

- ✓ Exploring distribution
- ✓ Detecting outliers
- ✓ Understanding data spread
- ✓ Comparing distributions

🔍 How KDE (Kernel Density Estimation) Works

1. **Place a kernel at each data point**
2. **Sum all kernels**
3. **Normalize the curve**

$$KDE(x) = (1/nh) \times \Sigma K((x - x_i)/h)$$

n = number of data points

- A small bell curve (Gaussian) is centered at every observation

- Add up the contributions from all kernels at each position

- Scale so the total area under the curve equals 1

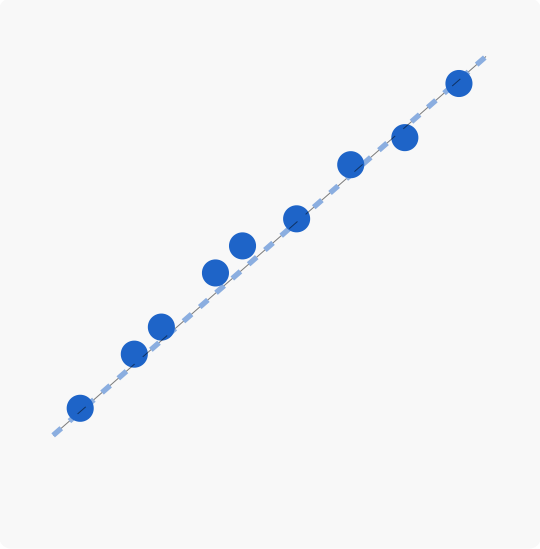
h = bandwidth (smoothing parameter)

K = kernel function (usually Gaussian)

2D Relationships: Scatter Plot & Bubble Chart

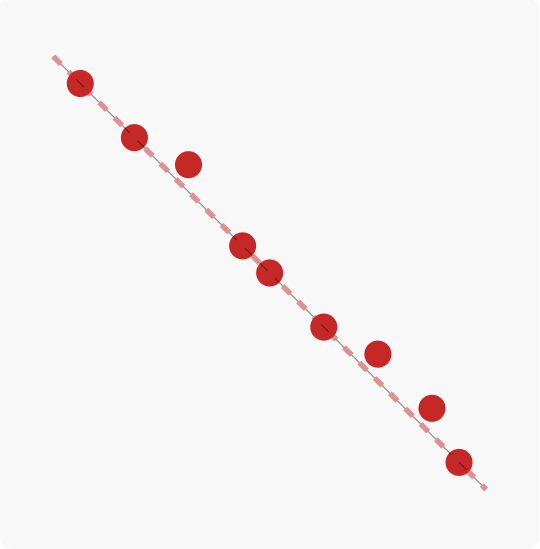
Exploring Relationships Between Continuous Variables

Positive Correlation



Strong Positive (r ≈ +0.9)

Negative Correlation



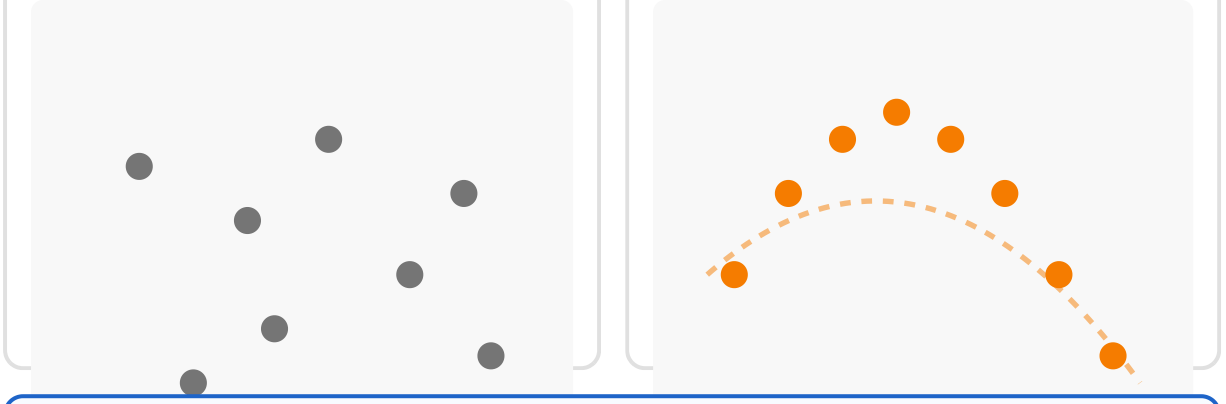
Strong Negative (r ≈ -0.9)

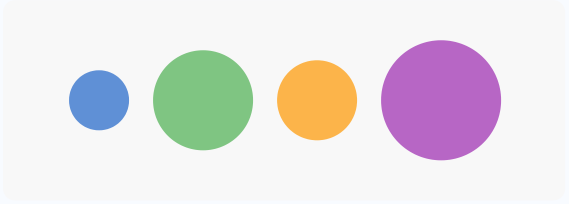
No Correlation

Non-linear Pattern

Key Features

- Reveals correlation & clusters
- Identifies outliers
- Shows non-linear patterns
- Add trend lines for clarity





Bubble Chart

Add 3rd dimension (size) + optional 4th (color)

Best Practices

- Use alpha transparency
- Try hexbin for large data
- Annotate key points

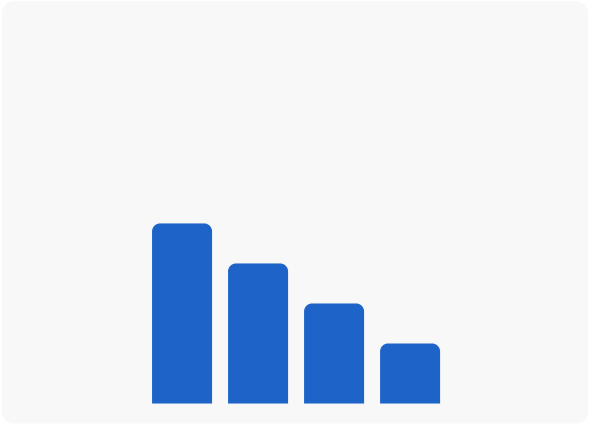
When to Use

- ✓ Correlation analysis
- ✓ Trend detection
- ✓ Multivariate exploration

Categorical Data: Bar Chart & Pie Chart

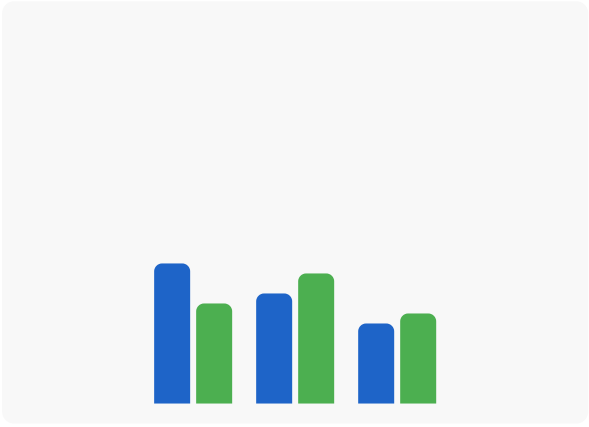
Comparing Categories and Showing Proportions

Simple Bar Chart



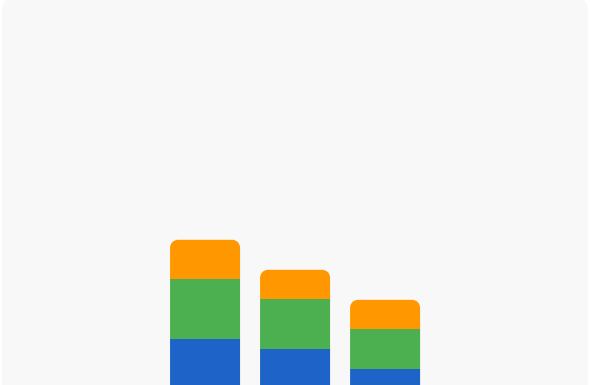
Best for comparing categories

Grouped Bar Chart

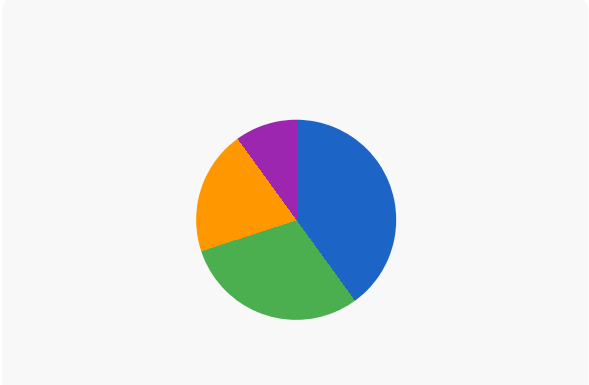


Compare subcategories

Stacked Bar Chart



Pie Chart



Guidelines

- Horizontal bars for long names
- Sort by value (descending)
- Stacked bars harder to compare
- Pie only for simple proportions

Perception Accuracy

Length (Bar)	High ✓
Angle (Pie)	Low ✗

When to Use

Show composition

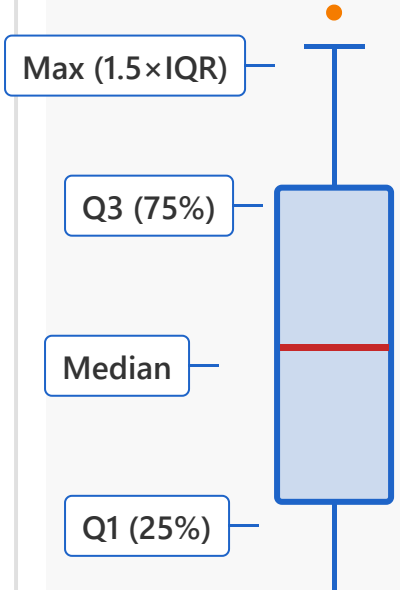
Use sparingly (<5 categories)

- ✓ Category comparison
- ✓ Ranking & ordering
- ✓ Simple proportions

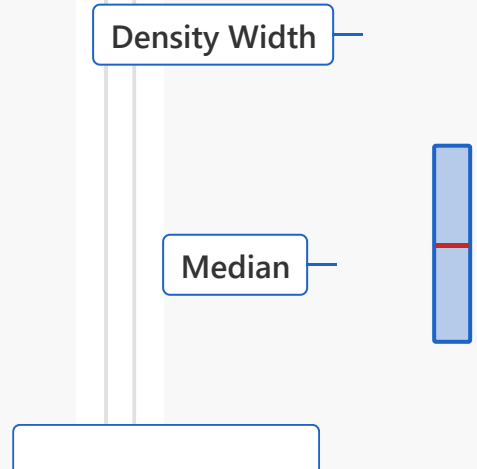
Distribution Comparison: Boxplot & Violin Plot

Anatomy and Components of Distribution Visualizations

Boxplot Anatomy



Violin Plot Anatomy



Key Features

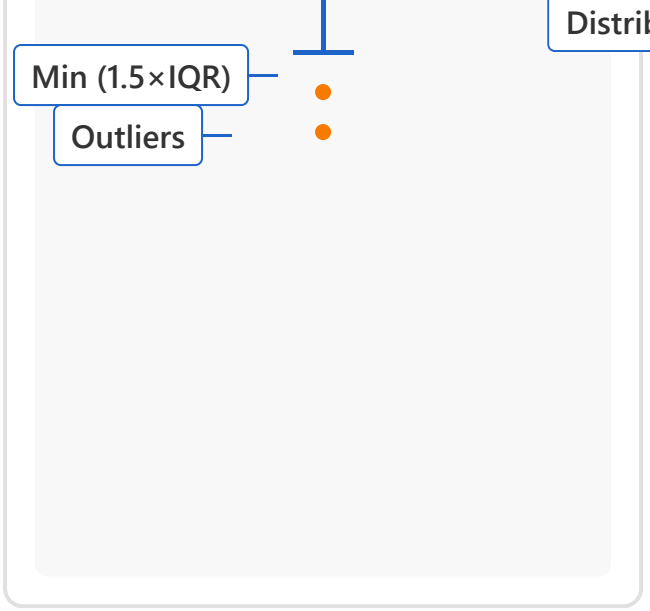
- Boxplot shows 5-number summary
- Violin reveals full distribution
- IQR = Q3 - Q1
- Width = density at each value

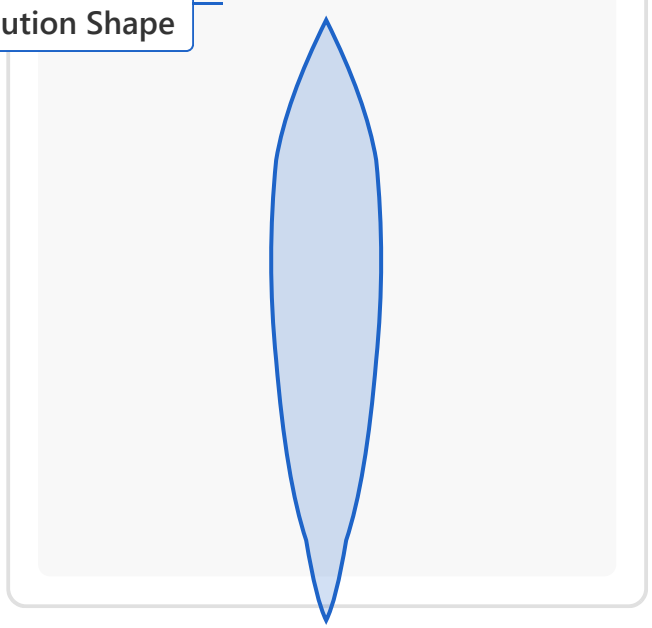
Whisker Formula

Upper: `Q3 + 1.5×IQR`

Lower: `Q1 - 1.5×IQR`

Comparison





Boxplot: Summary stats

Violin: Full shape + KDE

When to Use

- ✓ Comparing groups
- ✓ Identifying outliers
- ✓ Seeing data spread
- ✓ Detecting multimodality

Time Series: Line Graph & Area Chart

Visualizing Trends and Changes Over Time

Single Line

Multiple Lines

Stacked Area

Features

Line Graph

- Sequential continuity
- Trend comparison

Area Chart

- Emphasizes magnitude
- Shows composition

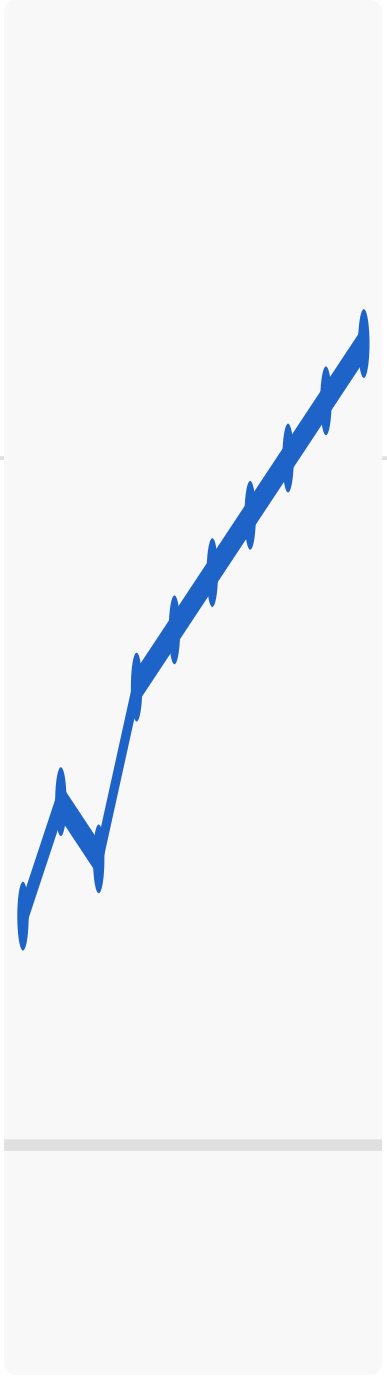
Best Practices

- Distinct colors for lines
- Add reference lines
- Appropriate granularity

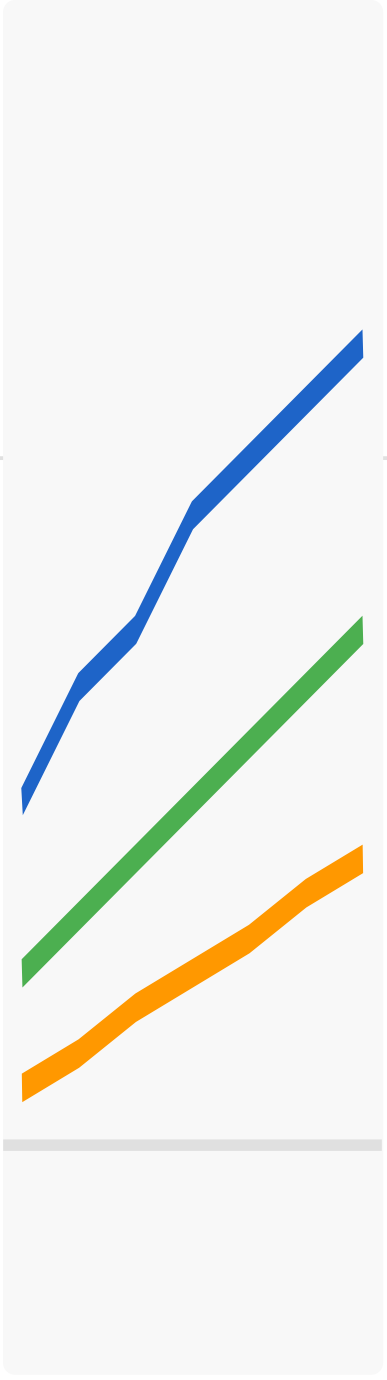
- Avoid over/undersampling

When to Use

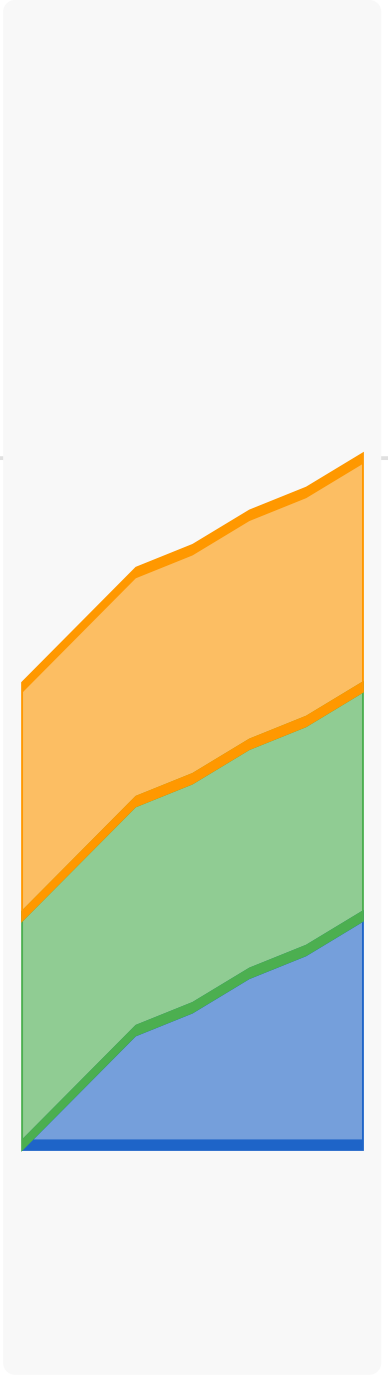
- ✓ Trend analysis
- ✓ Forecasting
- ✓ Seasonal patterns
- ✓ Change over time



Shows continuous trend over time



Compares multiple trends with distinct colors

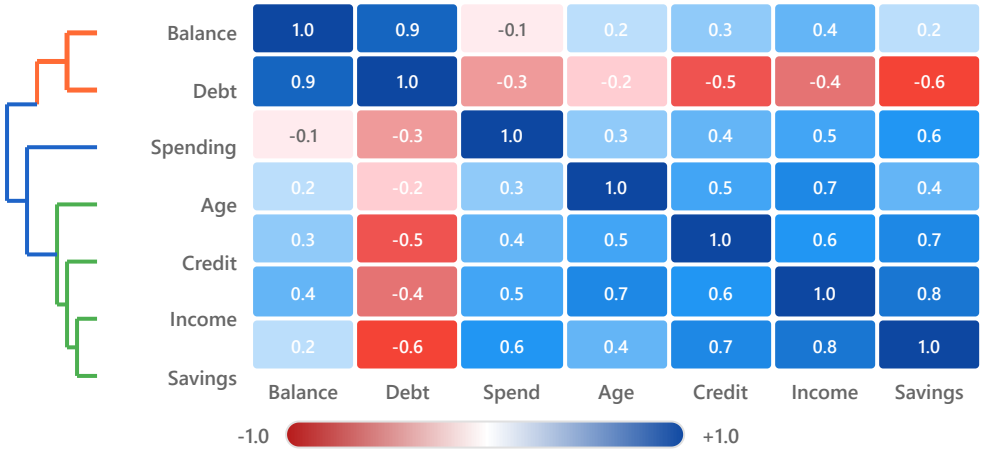


Shows composition changes over time

Correlation Analysis

Heatmap & Correlation Matrix

Correlation Heatmap with Dendrogram



Interpretation

- Positive: Variables move together
- Negative: Inverse relationship
- Zero: No linear relationship
- Dendrogram groups similar features

Correlation Matrix (Original Order)

	Age	Income	Credit	Balance	Spend	Savings	Debt
Age	1.00	0.70	0.50	0.20	0.30	0.40	-0.20
Income	0.70	1.00	0.60	0.40	0.50	0.80	-0.40
Credit	0.50	0.60	1.00	0.30	0.40	0.70	-0.50
Balance	0.20	0.40	0.30	1.00	-0.10	0.20	0.90
Spend	0.30	0.50	0.40	-0.10	1.00	0.60	-0.30
Savings	0.40	0.80	0.70	0.20	0.60	1.00	-0.60

Key Insights

- **Balance-Debt**: 0.9 (very strong)
- **Income-Savings**: 0.8 (strong)
- **Age-Income**: 0.7 (strong)
- Financial health cluster visible

✓ Best Practices

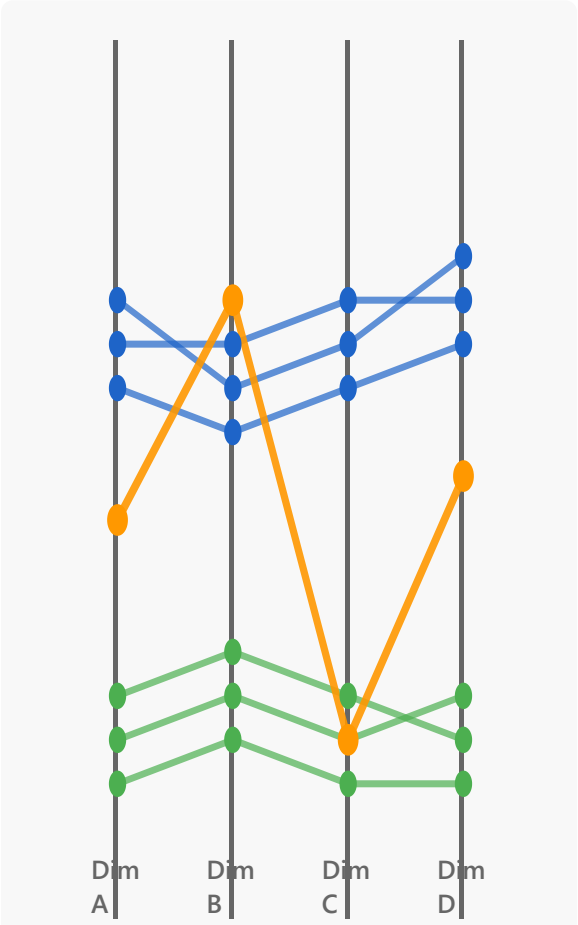
- ✓ Use diverging colormap
- ✓ Cluster similar variables
- ✓ Identify redundant features

✓ Look for multicollinearity

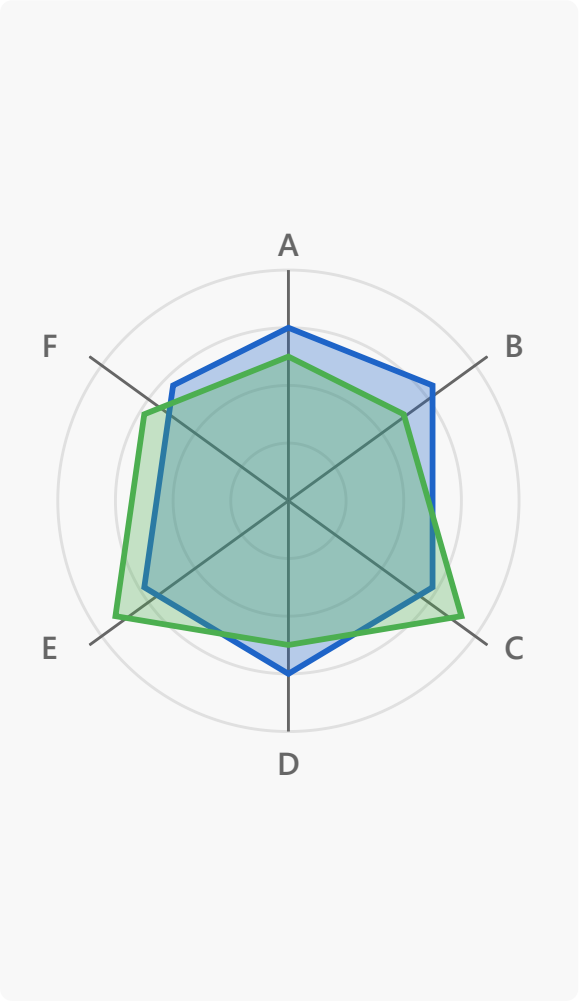
Multidimensional Visualization

Parallel Coordinates & Radar Chart

Parallel Coordinates



Radar Chart



Comparison

Parallel Coords

- Many dimensions
- Reveals clusters
- Color by class

Radar Chart

- 6-8 dimensions max
- Profile comparison
- Circular symmetry

Limitations

- Parallel: Axis order matters
- Radar: Limited dimensions
- Both: Can be cluttered

When to Use

- ✓ High-dim exploration
- ✓ Pattern detection
- ✓ Profile comparison

Lines connect observations across dimensions

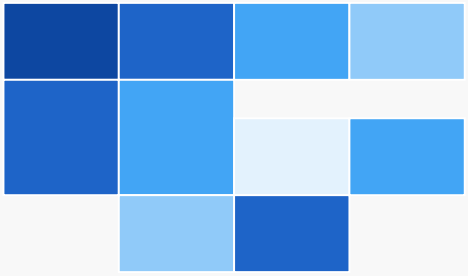
Circular layout comparing entity profiles

✓ Player/product stats

Geographic Data Visualization

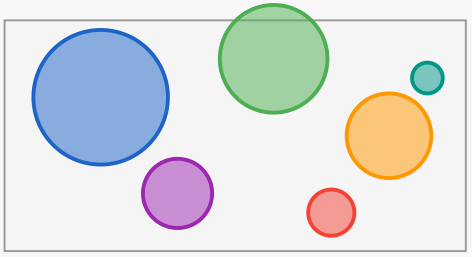
Choropleth Map & Bubble Map

Choropleth Map



Very High Medium Low

Bubble Map



Large Medium Small

Features

Choropleth

- Color-encode regions
- Sequential/diverging scales

Bubble Map

- Size at coordinates
- Combine size + color

Best Practices

- Normalize by population/area
- Consider projection distortion
- Add borders & labels
- Use appropriate color scale

When to Use

- ✓ Spatial patterns
- ✓ Regional comparison
- ✓ Location-based insights

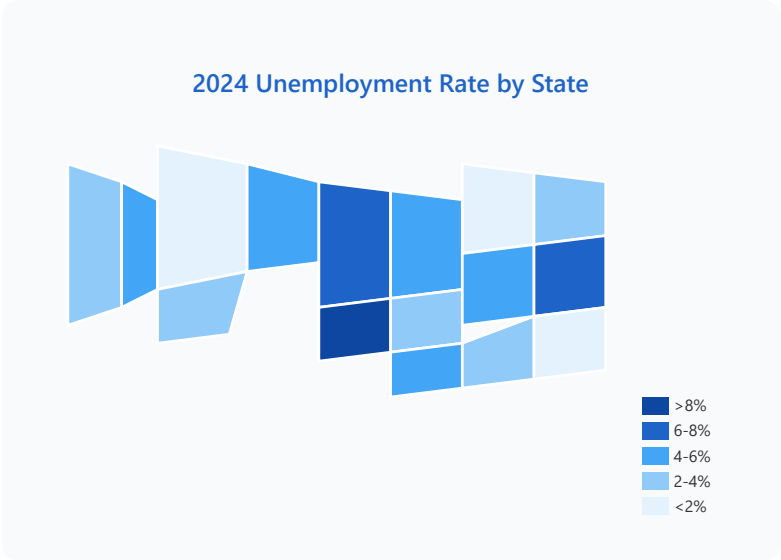
Color encodes regional statistics

Size encodes values at coordinates

✓ Demographics analysis

Real-World Examples

📊 Choropleth: US Unemployment Rate



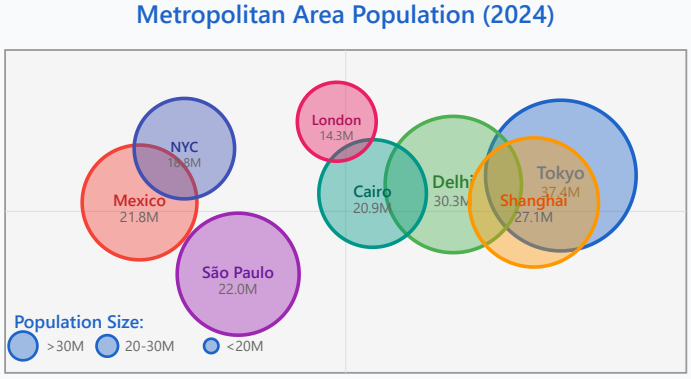
Use Case

This choropleth map visualizes unemployment rates across US states using color intensity. Darker blues indicate higher unemployment rates.

Key Insight: Regional patterns show economic disparities

Color Scale: Sequential (light to dark blue)

🌍 Bubble: Global City Populations



Use Case

This bubble map displays population sizes of major cities worldwide. Bubble size represents population, and colors distinguish different cities.

Key Insight: Asia dominates with largest urban centers

Encoding: Size = population, Position = geography

Advantage: Shows both location and magnitude

Data Type: Normalized percentages

Part 3/3:

Advanced Visualization for ML

17. Visualization Strategy for EDA
18. Feature Distribution and Outlier Detection
19. Visualizing Feature Interactions
20. Dimensionality Reduction Visualization (PCA, t-SNE, UMAP)
21. Model Performance Visualization - Learning Curves
22. Classification Model Evaluation - Confusion Matrix, ROC, PR
23. Regression Model Evaluation - Residuals, QQ Plot
24. Model Interpretation - SHAP, LIME, Attention
25. Dashboard Design and Storytelling

Visualization Strategy for EDA

Systematic Workflow for Exploratory Data Analysis

1

Data Loading

Import and understand data structure

↓

2

Univariate Analysis

Distributions, missing values, outliers

↓

3

Bivariate Relationships

Correlations, dependencies, patterns

↓

Multivariate Patterns

Key Principles

- Systematic pipeline
- Automate repetitive plots
- Use small multiples (facets)
- Document insights iteratively

Best Practices

- Start simple, add complexity
- Maintain visualization journal

4

Interactions, clusters, dimensionality

- Compare across categories
- Iterate hypothesis → viz → insight

↓

5

Insights & Hypotheses

Generate actionable insights

Balance

Breadth ⟷ Depth

Many features ⟷ Detail analysis

Feature Distribution & Outlier Detection

Assessing Normality and Identifying Anomalies

Histogram

Q-Q Plot

Box Plot

Key Insights

- Reveals skewness & modality
- Identifies outliers & quartiles
- Q-Q: diagonal = normal

Statistical Tests

Shapiro-Wilk

Anderson-Darling

Decision

Errors	Remove

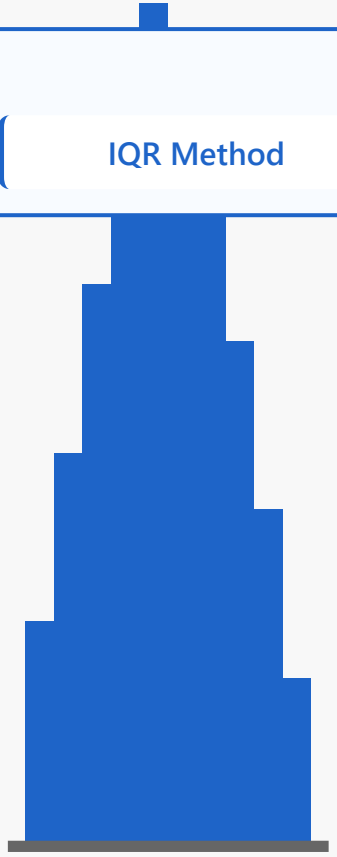
Extreme**Keep/Investigate**

Transformations

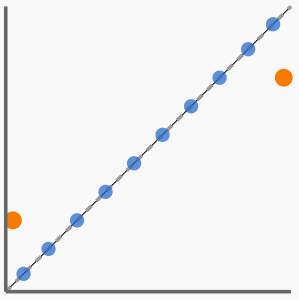
- ✓ Log transform
- ✓ Box-Cox
- ✓ Feature scaling

Outlier Detection Methods

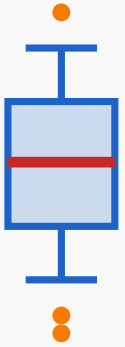
- IQR Method
- Z-Score
- Isolation Forest



Shows distribution shape



Tests normality assumption



Highlights outliers & IQR

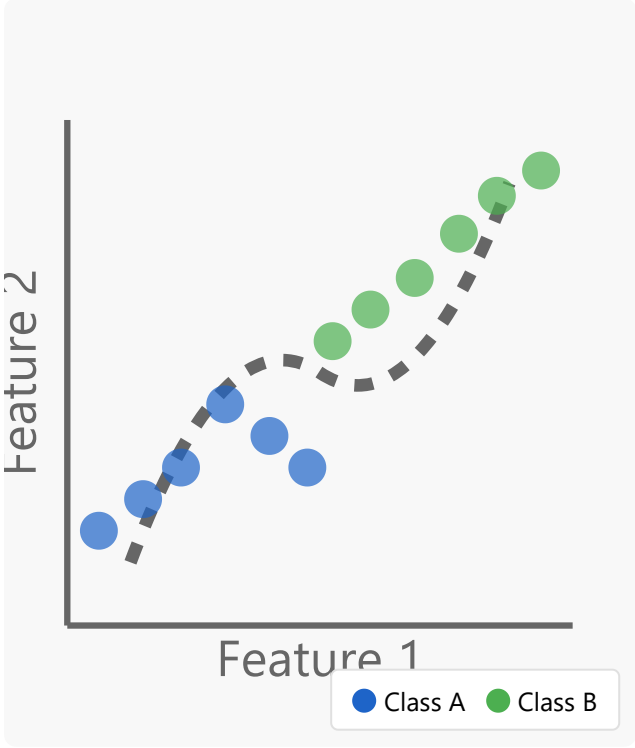
Visualizing Feature Interactions

Understanding How Features Jointly Affect Target

Scatter Plot Matrix

All pairwise relationships

Interaction Plot



Non-linear decision boundary

Contour Plot

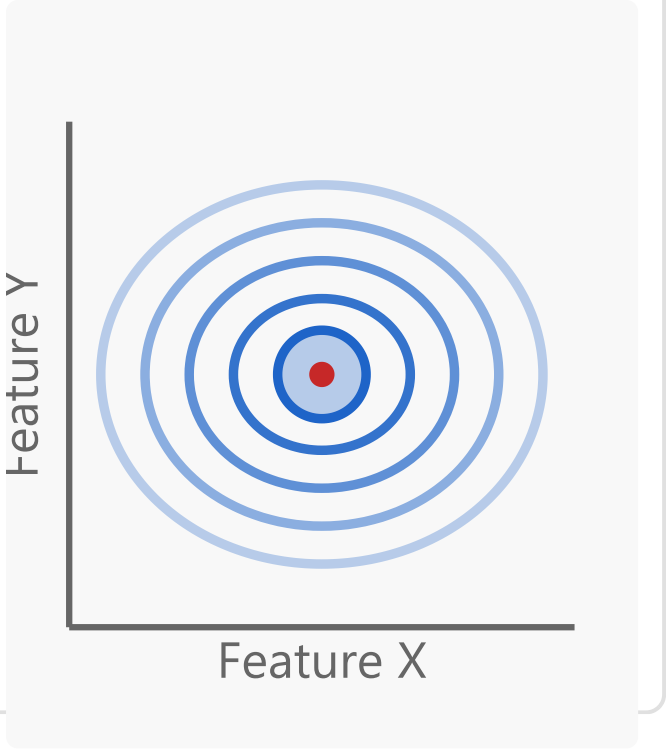
Conditional Plot

Methods

- Pairplot: All relationships
- Interaction: Joint effects
- Contour: Response surface
- Conditional: Fixed features

Key Insights

- Color-encode target variable
- Reveal discriminative patterns



Response surface visualization

Fixed feature conditions

- Identify non-linear interactions
- Feature crosses for new space

✓ Best Practices

- ✓ Use stratification
- ✓ Separate by categories
- ✓ Explore feature crosses
- ✓ Catch what linear models miss

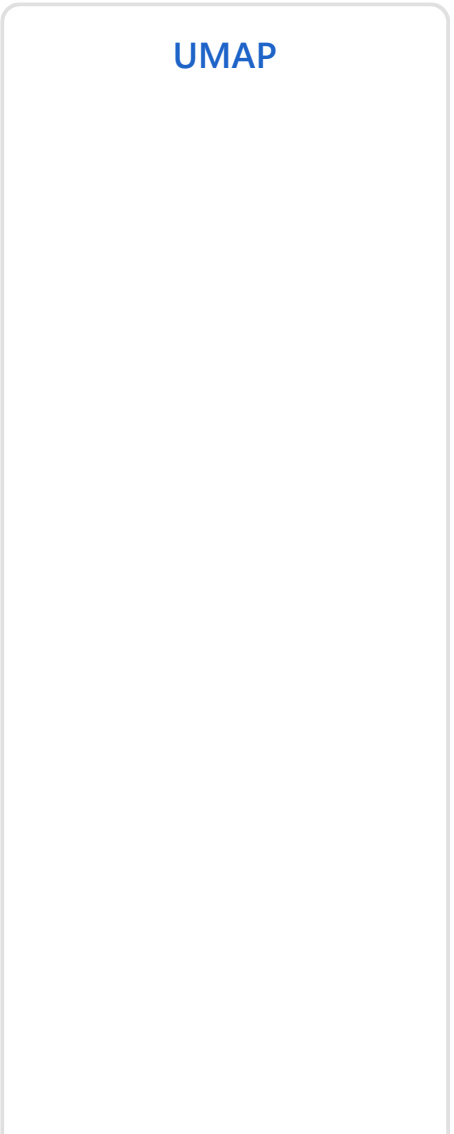
Dimensionality Reduction: PCA, t-SNE, UMAP

Projecting High-Dimensional Data to 2D

PCA

t-SNE

UMAP



Comparison

PCA

- Linear, interpretable
- Fast, deterministic

t-SNE

- Non-linear, local focus
- Perplexity: 5-50

UMAP

- Faster than t-SNE
- Local + global

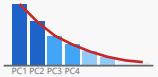
Cautions

- 2D projections lose info
- Can create false patterns
- Validate with multiple methods

✓ Best Practices

- ✓ Color by class labels
- ✓ Try continuous variables
- ✓ Combine with clustering
- ✓ Use scree plot for PCA

PCA Scree Plot: Explained Variance Ratio



Linear Projection

Non-linear

Balanced

Maximizes variance

Preserves local structure

Local + global structure

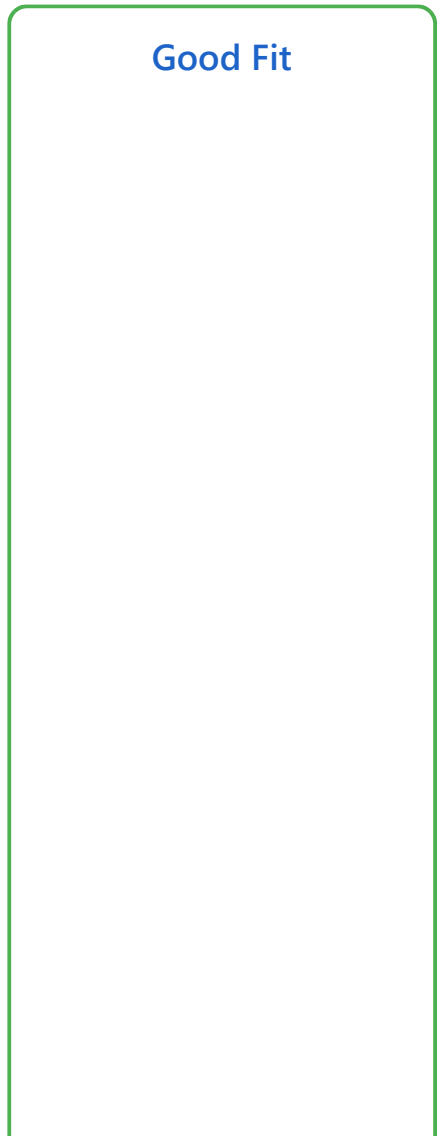
Model Performance: Learning Curves

Diagnosing Bias-Variance Tradeoff

Underfitting

Overfitting

Good Fit



Diagnosis

Underfitting

- Model too simple
- Add complexity/features

Overfitting

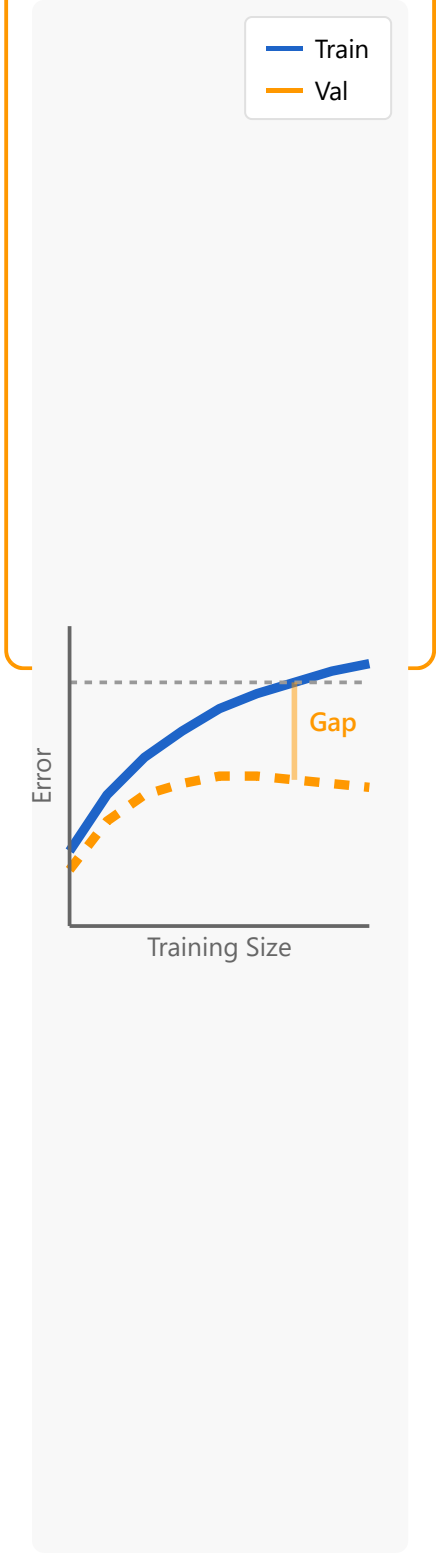
- More data helps
- Add regularization

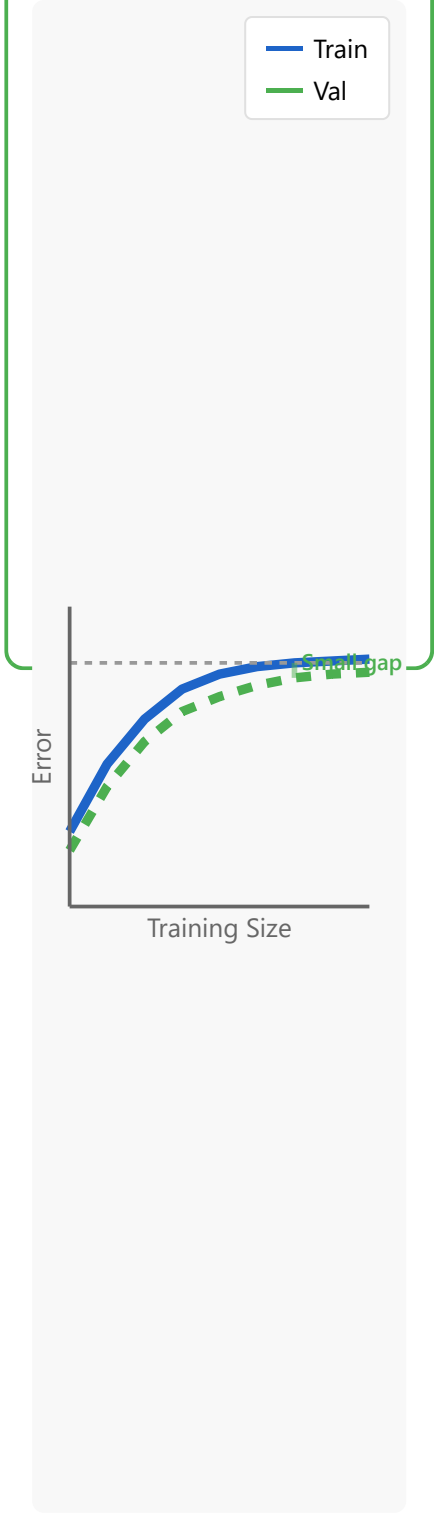
Good Fit

- Balanced complexity
- Near optimal performance

What to Plot

- Loss curves





- Accuracy / F1
- Confidence intervals
- Convergence monitoring

✓ Best Practices

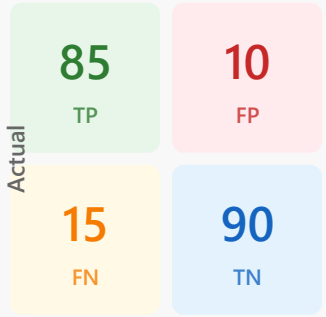
- ✓ Use cross-validation
- ✓ Plot multiple metrics
- ✓ Monitor early stopping
- ✓ Compare train/val gap

High Bias	High Variance	Well-Balanced
Both plateau at poor performance	Large gap between curves	Curves converge near optimal

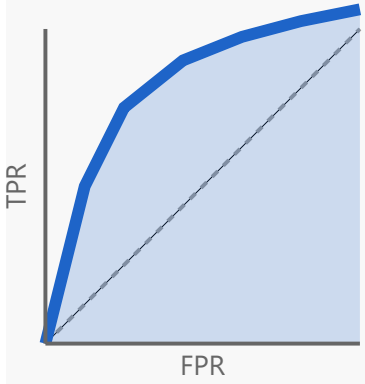
Classification Model Evaluation

Confusion Matrix, ROC, and Precision-Recall Curves

Confusion Matrix



ROC Curve



PR Curve

Metrics

Confusion Matrix

- TP, FP, FN, TN
- Normalize by row/col

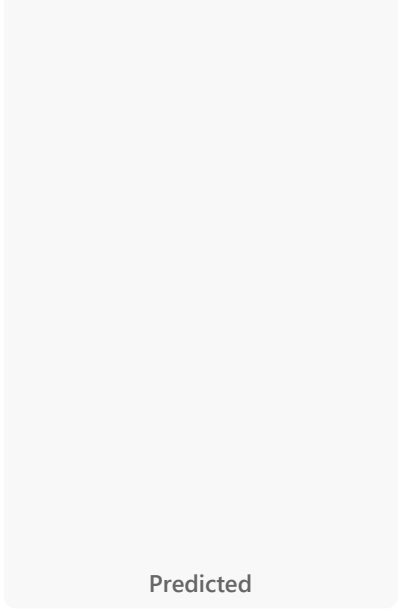
ROC Curve

- Balanced classes
- AUC summarizes

PR Curve

- Imbalanced data
- Focus on positive

When to Use



Prediction outcomes grid

Accuracy: 87.5%

TPR vs FPR

AUC: 0.92

Precision vs Recall

AP: 0.88

- ROC: balanced classes
- PR: imbalanced datasets
- CM: detailed breakdown

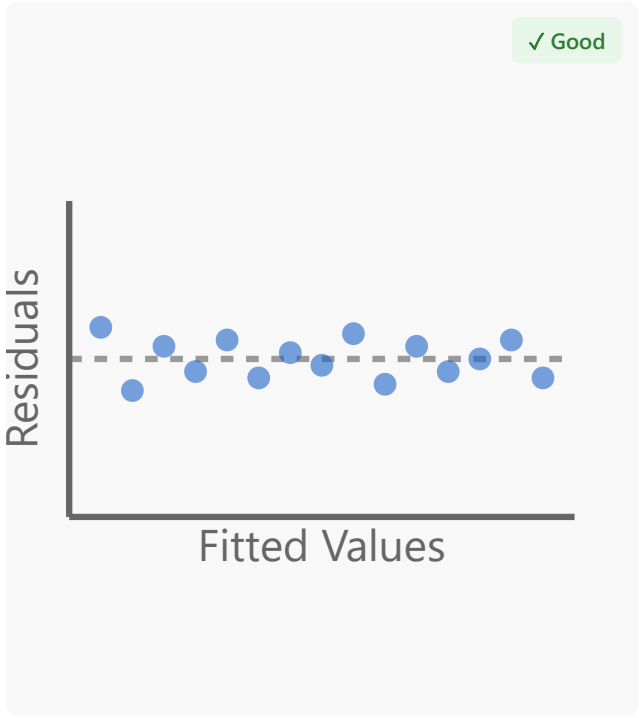
✓ Best Practices

- ✓ Compare models
- ✓ Select threshold
- ✓ Balance P & R by cost
- ✓ Multi-class: one-vs-rest

Regression Evaluation: Residuals & Diagnostics

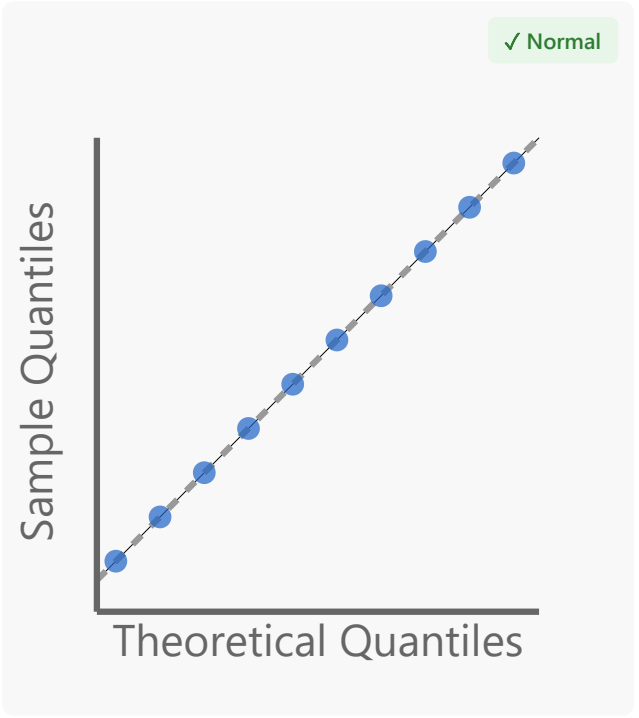
Checking Model Assumptions and Fit Quality

Residuals vs Fitted



Random scatter = good fit

Q-Q Plot (Residuals)



Points on line = normality

Scale-Location

Predicted vs Actual

✓ Good Signs

Residuals

- Random scatter
- Zero mean

Q-Q Plot

- Points on diagonal

Scale-Location

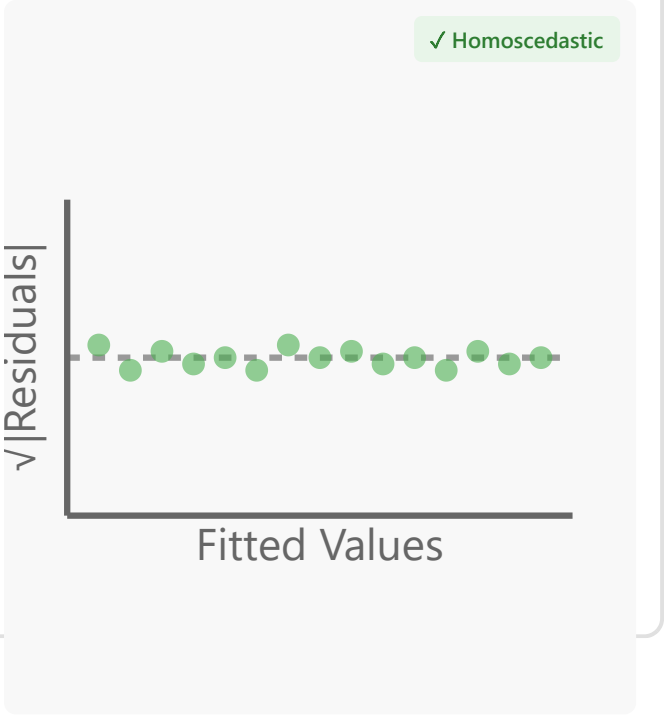
- Horizontal band

Pred vs Actual

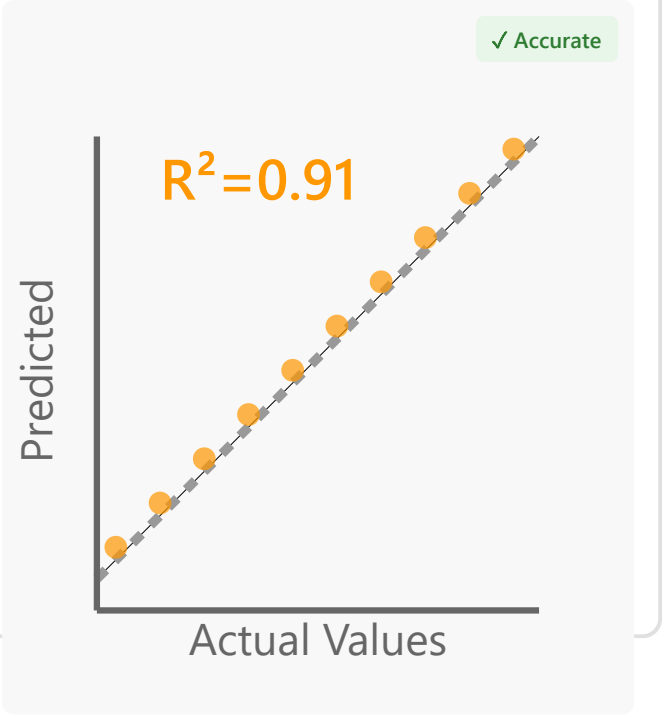
- Points near diagonal

⚠ Warning Signs

- Patterns → model issues
- Funnel → heteroscedasticity



Constant variance



Close to diagonal = good

- Curved → non-linearity
- Q-Q deviates → non-normal

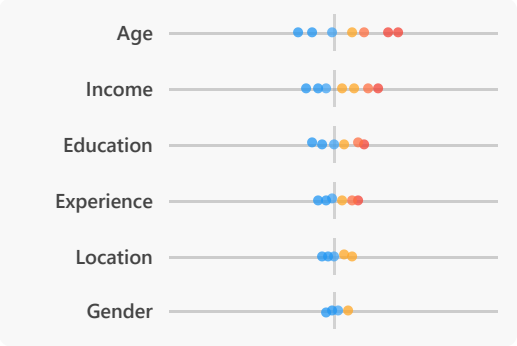
✓ Best Practices

- ✓ Check all 4 plots
- ✓ Look for patterns
- ✓ Identify influential points
- ✓ Transform if needed

Model Interpretation Methods

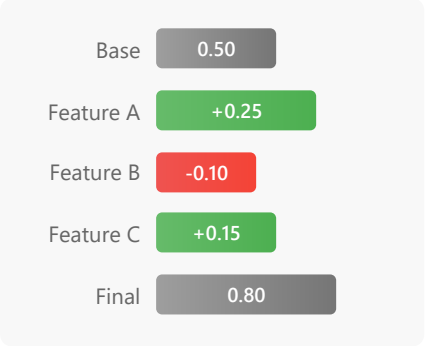
SHAP, LIME & Attention Mechanisms

SHAP Summary Plot



Feature importance ranking

SHAP Waterfall

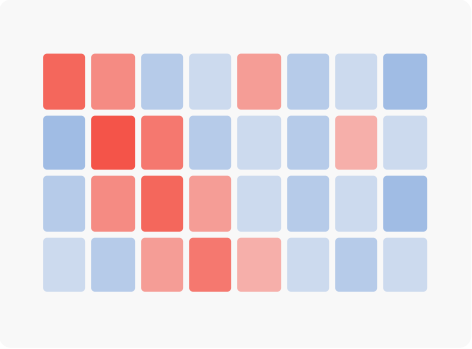


Individual contribution

LIME Explanation

Local linear approximation

Attention Heatmap



Model focus visualization

Methods

SHAP

- Game-theory based
- Global + local

LIME

- Model-agnostic
- Local perturbation

Attention

- Built-in mechanism
- Text/image focus

Scope

SHAP	Global/Local
LIME	Local only
Attention	Instance-level

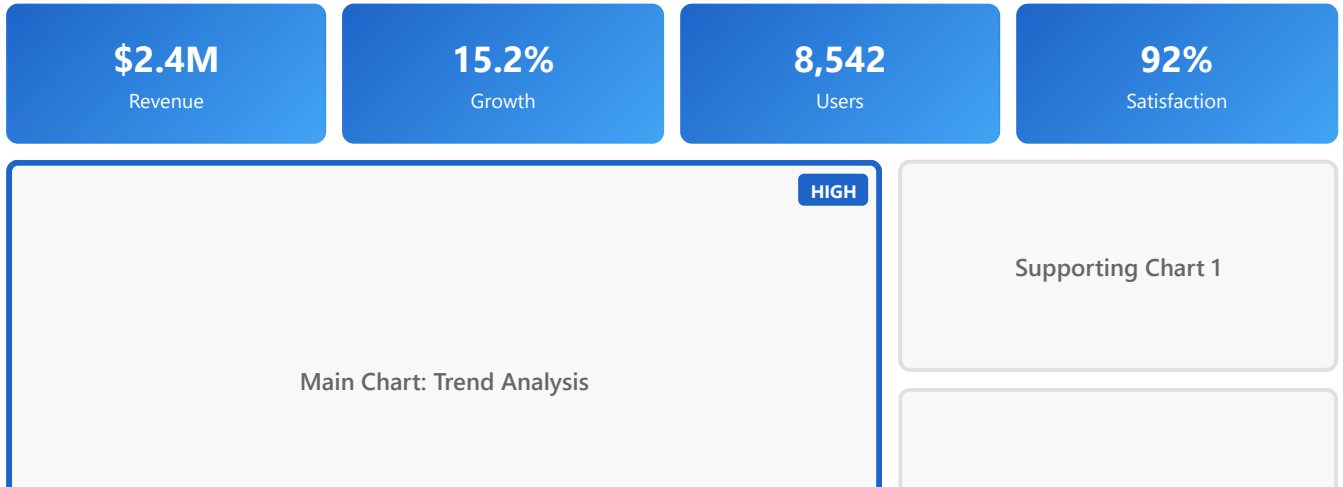
✓ Best Practices

- ✓ Use SHAP for importance
- ✓ LIME for quick insights
- ✓ Attention for NLP/vision
- ✓ Compare with PDP

Dashboard Design & Storytelling

Information Hierarchy and Narrative Structure

Dashboard Layout: Information Hierarchy



Design Principles

1. Define audience & message
2. KPIs at top
3. Consistent layout & colors

Supporting Chart 2

Detail View 1

Detail View 2

Detail View 3

Data Storytelling Structure

1 **Setup**
Context & background

→

2 **Conflict**
Problem or insight

→

3 **Resolution**
Data-driven solution

4 Progressive disclosure

Interactive Elements

- Filters
- Drill-downs
- Hover details
- Dynamic updates

✓ Best Practices

- ✓ Test with users
- ✓ Iterate based on feedback
- ✓ Measure engagement
- ✓ Keep it simple

Thank you

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com