Exploratory Data Analysis (EDA)

CMSC 173 - Machine Learning

Course Lecture

Outline

Introduction to EDA

What is Exploratory Data Analysis?

Definition

EDA is the process of investigating datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

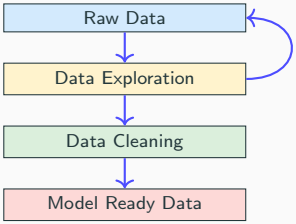
Primary Goals:

- **Understand** data structure and quality
- **Discover** patterns and relationships
- **Identify** anomalies and outliers
- **Guide** feature engineering decisions
- **Inform** modeling strategy

Key Questions EDA Answers:

- What does my data look like?
- Is my data clean and complete?
- What patterns exist?
- Which features are important?

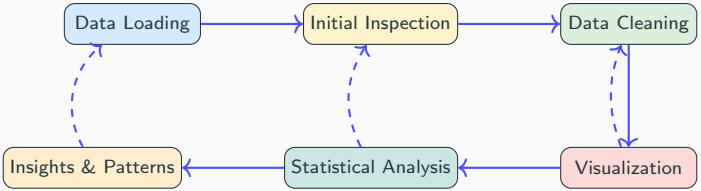
EDA Process Overview:



Key Insight

EDA is iterative! Insights from one analysis often lead to new questions and deeper investigations.

The EDA Workflow



Data Loading

- Import datasets
- Check file formats
- Handle encoding issues

Statistical Analysis

- Descriptive statistics
- Correlation analysis
- Distribution testing

Key Outcome

- Clean, understood data
- Feature insights
- Modeling strategy

Why EDA is Critical for Machine Learning

Without EDA

Common Pitfalls:

- Garbage In, Garbage Out
- Poor model performance
- Biased predictions
- Overfitting to noise
- Missing important patterns
- Wasted computational resources

With Proper EDA

Benefits Achieved:

- High-quality, clean data
- Optimal feature selection
- Appropriate model choice
- Better generalization
- Actionable insights
- Efficient resource usage

Statistical Foundation:

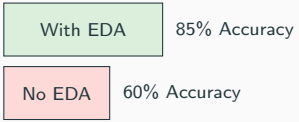
For dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$:

$$\text{Data Quality} = f(\text{Completeness, Accuracy, Consistency}) \quad (1)$$

$$\text{Model Performance} \propto \text{Data Quality} \quad (2)$$

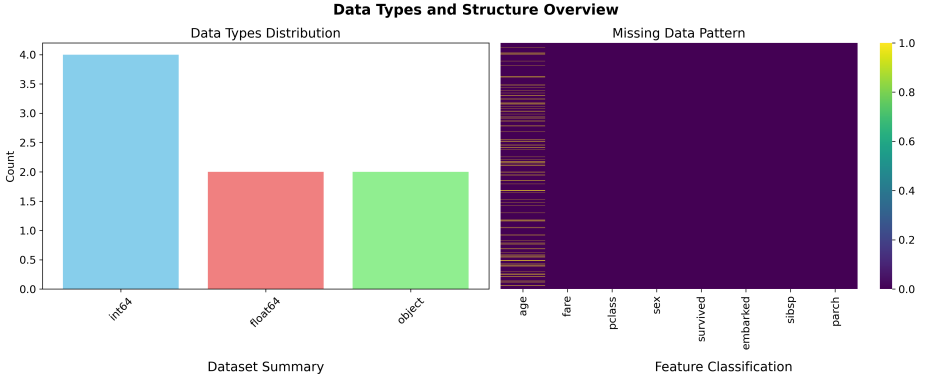
Impact Quantification:

Studies show that proper EDA can improve model performance by 15-30% and reduce development time by 40-60%.



Data Understanding & Types

Understanding Your Dataset



Dataset Shape: 891 rows × 8 columns

Numerical Features: 6
Categorical Features: 2

Memory Usage: 132.1 KB
Missing Values: 89 (1.2%)

Numerical Features:
age, fare, pclass, survived, sibsp, parch

Categorical Features:
sex, embarked

Binary Features:
sex, survived

First Steps

Data Types Classification

Numerical Data

Continuous Variables:

- Can take any value in a range
- Examples: age, salary, temperature
- Mathematical operations meaningful

Discrete Variables:

- Countable, distinct values
- Examples: number of children, cars owned
- Often integers

Categorical Data

Nominal Variables:

- No natural ordering
- Examples: color, gender, city
- Cannot perform arithmetic

Ordinal Variables:

- Natural ordering exists
- Examples: education level, rating
- Ranking meaningful, differences may not be

Mathematical Representation

For numerical variable X:

$$X \in \mathbb{R} \text{ (continuous) or } X \in \mathbb{Z} \text{ (discrete)}$$

Mathematical Representation

For categorical variable C:

$$C \in \{\text{category}_1, \text{category}_2, \ldots, \text{category}_k\}$$

Practical Tip

Encoding Strategy: Numerical $\rightarrow$ Keep as-is; Nominal $\rightarrow$ One-hot encoding; Ordinal $\rightarrow$ Label encoding

Sample Dataset: Titanic Survival Analysis

Dataset Overview

Contains information on 891 passengers aboard the Titanic. Goal: Predict passenger survival based on their attributes.

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
1	0	3	male	22.0	1	0	7.25	S
2	1	1	female	38.0	1	0	71.28	C
3	1	3	female	26.0	0	0	7.92	S
4	1	1	female	35.0	1	0	53.10	S
5	0	3	male	35.0	0	0	8.05	S

Numerical Features

- **Age:** Continuous (0-80)
- **Fare:** Continuous (0-512)
- **SibSp:** Discrete count
- **Parch:** Discrete count

Categorical Features

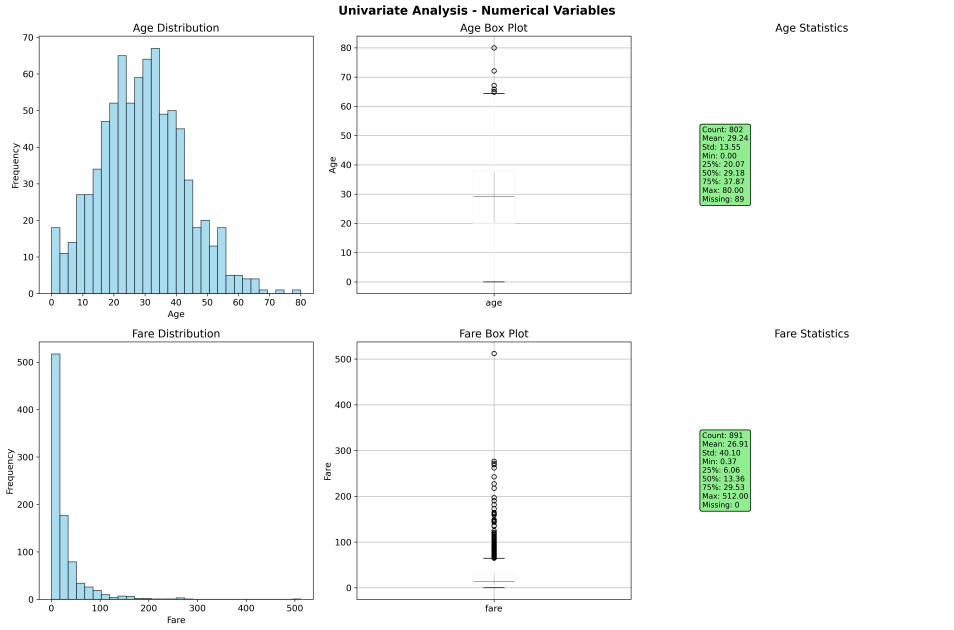
- **Sex:** Nominal (M/F)
- **Embarked:** Nominal (C/Q/S)
- **Pclass:** Ordinal (1st, 2nd, 3rd)

Target Variable

- **Survived:** Binary (0/1)
- **Classification Problem**
- 38.4% survival rate

Univariate Analysis

Univariate Analysis - Numerical Variables



Key Observations

Statistical Measures for Numerical Data

Central Tendency

For variable $X = \{x_1, x_2, \ldots, x_n\}$:

Mean (Arithmetic):

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

Median: Middle value when sorted

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ odd} \\ \frac{x_{n/2} + x_{n/2+1}}{2} & \text{if } n \text{ even} \end{cases}$$

Mode: Most frequent value

Dispersion Measures

Variance:

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

Standard Deviation:

$$\sigma = \sqrt{\sigma^2}$$

Interquartile Range:

$$\text{IQR} = Q_3 - Q_1$$

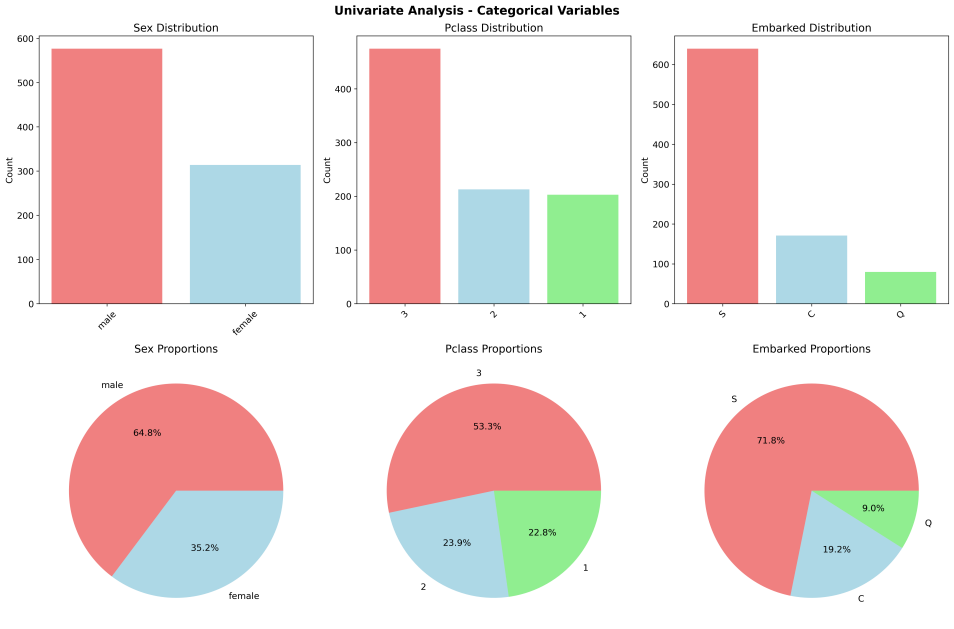
Range:

$$\text{Range} = x_{\text{max}} - x_{\text{min}}$$

Practical Guidelines

Skewed data: Use median & IQR; **Normal data:** Use mean & standard deviation

Univariate Analysis - Categorical Variables



Key Insights

Statistical Measures for Categorical Data

Frequency Analysis

For categorical variable C with categories $\{c_1, c_2, \ldots, c_k\}$:

Frequency:

$$f_i = \text{count}(C = c_i)$$

Relative Frequency:

$$p_i = \frac{f_i}{n} \text{ where } \sum_{i=1}^{k} p_i = 1$$

Mode: Category with highest frequency

$$\text{Mode} = \arg\max_{c_i} f_i$$

Entropy Measure

Information content:

$$H(C) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Higher entropy = more uniform distribution

Visualization Guidelines

Bar Charts:

- Best for comparing categories
- Order by frequency for impact
- Use consistent colors

Pie Charts:

- Good for showing proportions
- Limit to ≤ 5 categories
- Start largest slice at 12 o'clock

Chi-Square Test

Test for uniform distribution:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i = observed, E_i = expected

Distribution Analysis & Normality Testing

Common Distributions

Normal Distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Log-Normal Distribution:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

Skewness:

$$\text{Skew} = \frac{E[(X-\mu)^3]}{\sigma^3}$$

Normality Tests

Shapiro-Wilk Test:

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

Kolmogorov-Smirnov Test:

$$D = \sup_x |F_n(x) - F(x)|$$

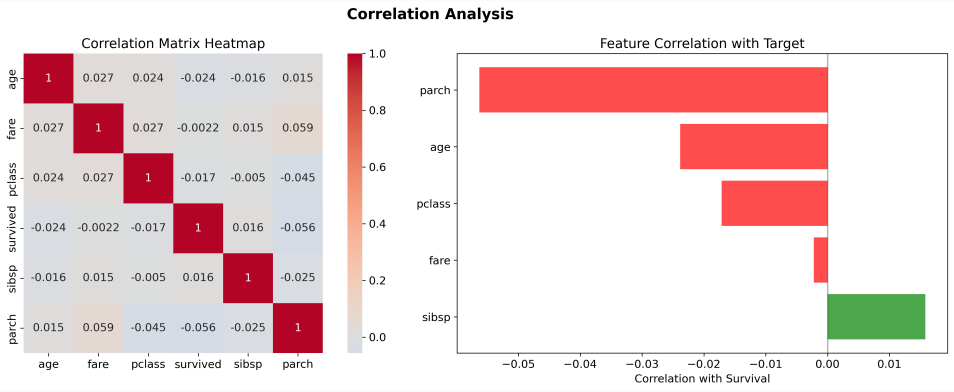
Anderson-Darling Test: More sensitive to tail deviations

Decision Rule

If $p < 0.05$: Reject normality assumption; Consider transformations (log, square root, Box-Cox)

Bivariate Analysis

Correlation Analysis



Correlation Insights

Strong correlations: Fare-Survival (0.26), Age-Survival (-0.07); **Weak correlations:** SibSp-Parch (0.41)

Correlation Coefficients & Interpretation

Pearson Correlation

For linear relationships:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Range: $r \in [-1, 1]$

- $r = 1$: Perfect positive correlation
- $r = 0$: No linear correlation
- $r = -1$: Perfect negative correlation

Significance Test

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

Non-Linear Correlations

Spearman Rank Correlation:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i = rank difference

Kendall's Tau:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

where n_c = concordant pairs, n_d = discordant pairs

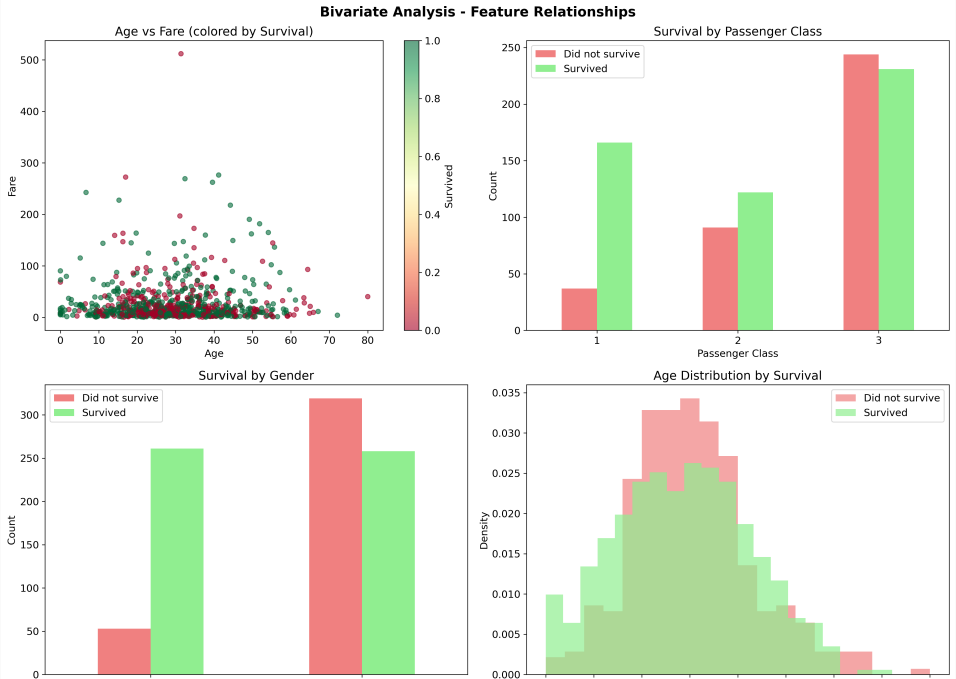
Interpretation Guide

- $|r| < 0.3$: Weak relationship
- $0.3 \leq |r| < 0.7$: Moderate relationship
- $|r| \geq 0.7$: Strong relationship

Remember

Correlation $\neq$ Causation! Always investigate the underlying mechanisms.

Bivariate Analysis - Feature Relationships



Cross-Tabulation & Contingency Tables

Contingency Table

For categorical variables A and B:

	B_1	B_2	Total
A_1	n_{11}	n_{12}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Joint Probability:

$$P(A_i, B_j) = \frac{n_{ij}}{n}$$

Marginal Probability:

$$P(A_i) = \frac{n_{i\cdot}}{n}, \quad P(B_j) = \frac{n_{\cdot j}}{n}$$

Independence Test

Chi-Square Test:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$

Degrees of freedom:

$$df = (r-1)(c-1)$$

Cramér's V (Effect Size):

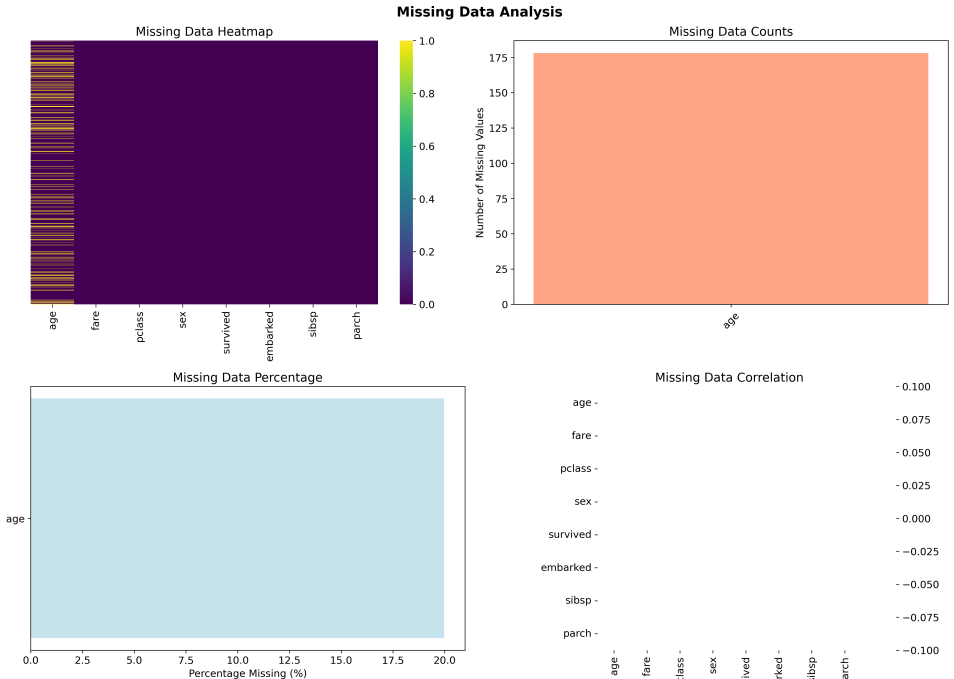
$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}}$$

Interpretation

$V \in [0, 1]$: 0 = no association, 1 = perfect association

Missing Data Handling

Missing Data Analysis



Types of Missing Data

MCAR: Missing Completely at Random

Definition: Missing data is independent of both observed and unobserved data.

Mathematical condition:

$$P(\text{Missing}|X, Y) = P(\text{Missing})$$

Example: Survey responses lost due to mail delivery issues.

Implication: Can use any imputation method without bias.

Test for MCAR

Little's MCAR Test: Tests null hypothesis that data is MCAR using EM algorithm.

MAR: Missing at Random

Definition: Missing data depends on observed data, but not on unobserved data.

Mathematical condition:

$$P(\text{Missing}|X, Y) = P(\text{Missing}|X)$$

Example: Older passengers more likely to have missing age data.

MNAR: Missing Not at Random Missing data depends on unobserved data.

Example: High-income individuals not reporting income.

Handling Strategy

MAR: Multiple imputation; **MNAR:** Domain expertise required

Imputation Strategies

Simple Imputation

Mean/Mode Imputation:

$$x_{\text{missing}} = \bar{x} \text{ or } \text{Mode}(x)$$

Median Imputation:

$$x_{\text{missing}} = \text{Median}(x)$$

Forward/Backward Fill: For time series data

Constant Value: Domain-specific constant (e.g., 0, -1)

Limitations

Simple methods reduce variance and can introduce bias

Advanced Imputation

KNN Imputation:

$$x_{\text{missing}} = \frac{1}{k} \sum_{i \in \text{k-nearest}} x_i$$

Multiple Imputation: Creates multiple complete datasets, analyzes each, pools results.

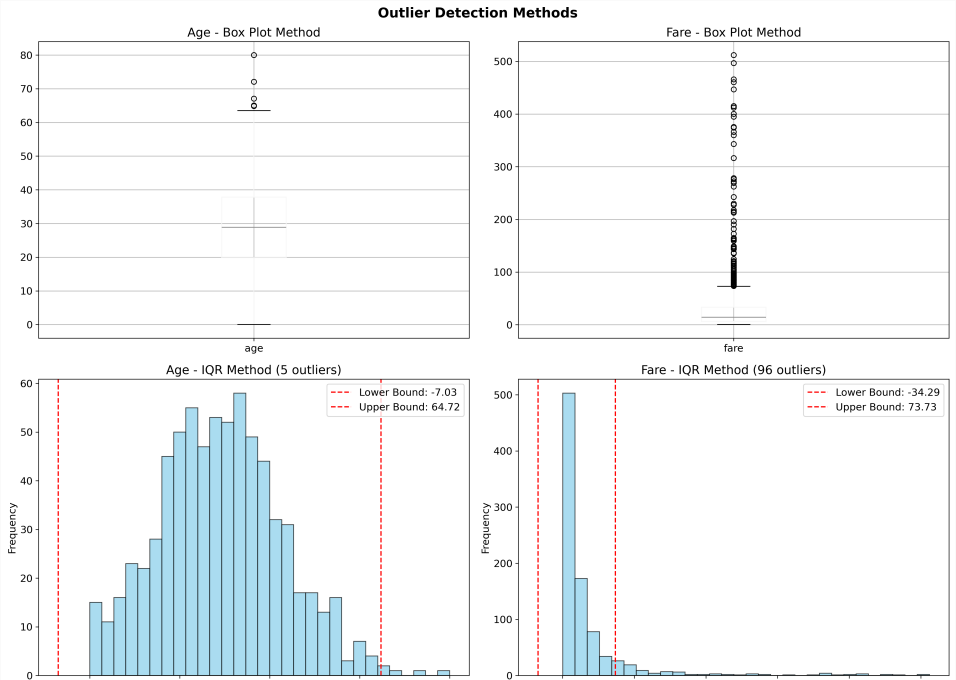
Model-based: - Linear regression - Random Forest - Deep learning approaches

Best Practice

Always analyze missing data pattern before choosing imputation method

Outlier Detection

Outlier Detection Methods



Statistical Outlier Detection Methods

IQR Method

Interquartile Range:

$$\text{IQR} = Q_3 - Q_1$$

Outlier bounds:

$$\text{Lower bound} = Q_1 - 1.5 \times \text{IQR}$$

$$\text{Upper bound} = Q_3 + 1.5 \times \text{IQR}$$

Outlier condition:

$$x < \text{Lower bound or } x > \text{Upper bound}$$

Modified Z-Score

$$M_i = \frac{0.6745(x_i - \text{median})}{\text{MAD}}$$

where MAD = median absolute deviation

Z-Score Method

Standard Z-score:

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

Outlier threshold:

$$|z_i| > 2.5 \text{ or } |z_i| > 3$$

Limitation: Sensitive to outliers in mean and std calculation

Isolation Forest

Anomaly Score:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where $E(h(x))$ = average path length, $c(n)$ = average path length of BST

Decision Framework

Normal distribution: Z-score; **Skewed distribution:** IQR; **Multivariate:** Isolation Forest

Multivariate Outlier Detection

Mahalanobis Distance

For multivariate data $\mathbf{x} \in \mathbb{R}^p$:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

where: - $\boldsymbol{\mu}$ = sample mean vector - $\boldsymbol{\Sigma}$ = sample covariance matrix

Outlier threshold:

$$D_M(\mathbf{x}) > \sqrt{\chi^2_{p,\alpha}}$$

Cook's Distance

Measures influence of each observation on regression:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p \times \text{MSE}}$$

Local Outlier Factor (LOF)

Local Reachability Density:

$$\text{lrd}_k(A) = \frac{1}{\frac{\sum_{B \in N_k(A)} \text{reach-dist}_k(A,B)}{|N_k(A)|}}$$

LOF Score:

$$\text{LOF}_k(A) = \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}_k(B)}{\text{lrd}_k(A)}}{|N_k(A)|}$$

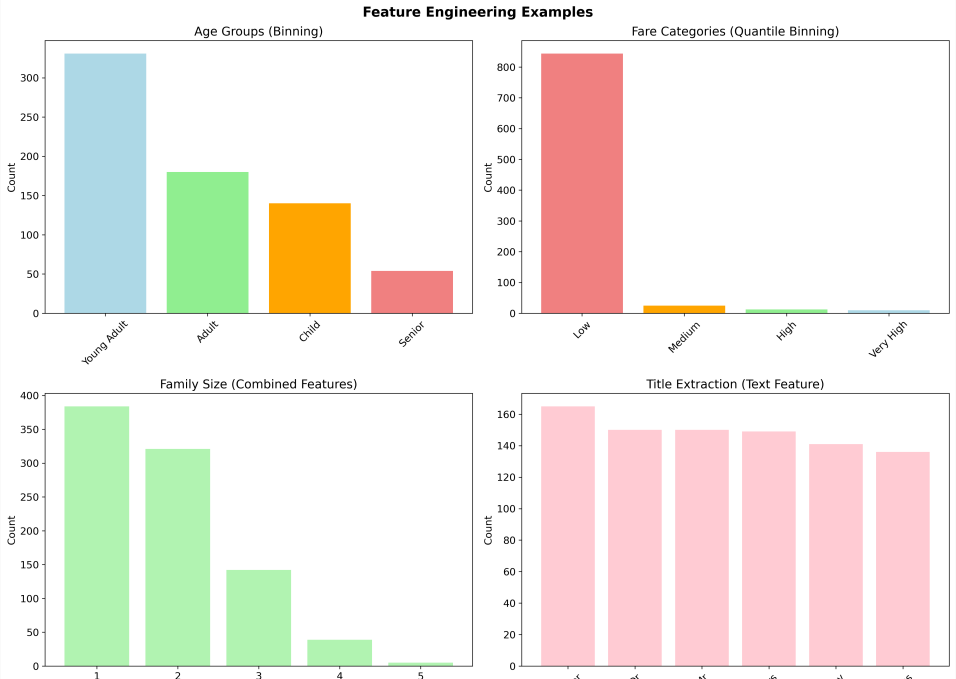
Interpretation: - LOF ≈ 1: Normal point - LOF $\gg 1$: Outlier

Key Insight

Multivariate outliers may not be outliers in any single dimension

Feature Engineering

Feature Engineering Examples



Feature Creation Techniques

Binning & Discretization

Equal-width binning:

$$\text{bin width} = \frac{x_{\max} - x_{\min}}{k}$$

Equal-frequency binning: Each bin contains $\frac{n}{k}$ observations

Quantile-based binning: Based on percentiles (quartiles, deciles)

Domain-specific binning: Using expert knowledge (e.g., age groups)

Optimal Binning

Use information gain or chi-square test to determine optimal bin boundaries

Feature Combinations

Arithmetic Operations: - Addition: $x_1 + x_2$ (total family size) - Multiplication: $x_1 \times x_2$ (interaction terms) - Division: x_1/x_2 (ratios, rates)

Boolean Operations: - Logical AND: $x_1 \wedge x_2$ - Logical OR: $x_1 \vee x_2$ - Conditional: if $x_1 >$ threshold then 1 else 0

String Operations: - Length: len(string) - Contains: pattern matching - Extract: regular expressions

Feature Engineering Guidelines

Domain Knowledge: Most important factor; **Iterative Process:** Create, test, refine; **Validation:** Always validate on holdout set

Mathematical Transformations

Power Transformations

Log Transformation:

$$y = \log(x + c)$$

Reduces right skewness, stabilizes variance

Square Root:

$$y = \sqrt{x}$$

Moderate variance stabilization

Box-Cox Transformation:

$$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x) & \text{if } \lambda = 0 \end{cases}$$

Optimal λ found via maximum likelihood

Trigonometric Features

For cyclical data (time, angles):

Sine/Cosine encoding:

$$\sin\left(\frac{2\pi \times \text{value}}{\text{max\_value}}\right)$$

$$\cos\left(\frac{2\pi \times \text{value}}{\text{max\_value}}\right)$$

Example for hour of day:

$$\text{hour\_sin} = \sin\left(\frac{2\pi \times \text{hour}}{24}\right)$$

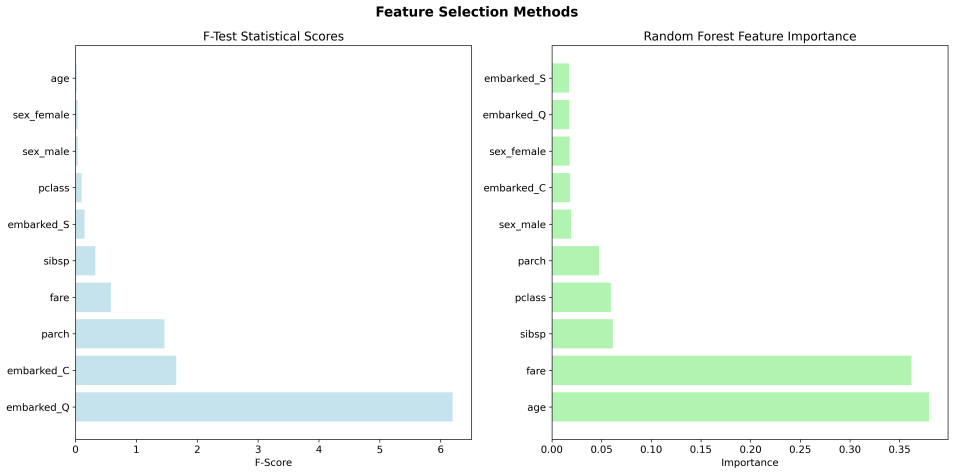
$$\text{hour\_cos} = \cos\left(\frac{2\pi \times \text{hour}}{24}\right)$$

When to Transform

Transform when: skewed data, non-linear relationships, or specific model requirements

Feature Selection

Feature Selection Methods



Selection Results

Statistical: Gender and fare most important; **Tree-based:** Consistent with domain knowledge about survival factors

Statistical Feature Selection

Filter Methods

Correlation-based: Select features with high correlation to target, low correlation to each other

F-test (ANOVA):

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2/(n-k)}$$

Chi-square test:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Mutual Information:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Wrapper Methods

Forward Selection: Start with empty set, add best features iteratively

Backward Elimination: Start with all features, remove worst iteratively

Recursive Feature Elimination:

$$\text{rank}_i = f(\text{coef}_i, \text{importance}_i)$$

Genetic Algorithm: Evolutionary approach to feature subset selection

Embedded Methods

L1 Regularization (Lasso):

$$\min_{\beta} \frac{1}{2n} ||y - X\beta||_2^2 + \lambda ||\beta||_1$$

Tree-based Feature Importance

Random Forest Importance

Mean Decrease Impurity:

$$\text{Importance}(x_j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{v \in \text{splits}} p(v) \times \Delta I(v)$$

where: - T = number of trees - $p(v)$ = proportion of samples reaching node v - $\Delta I(v)$ = impurity decrease at node v

Mean Decrease Accuracy: Permutation-based importance measuring prediction accuracy drop when feature is shuffled

Gradient Boosting Importance

Gain-based Importance:

$$\text{Importance}(x_j) = \sum_{t=1}^{T}\sum_{v \in \text{splits}_j} \text{gain}(v)$$

SHAP Values: Shapley Additive exPlanations provide unified measure:

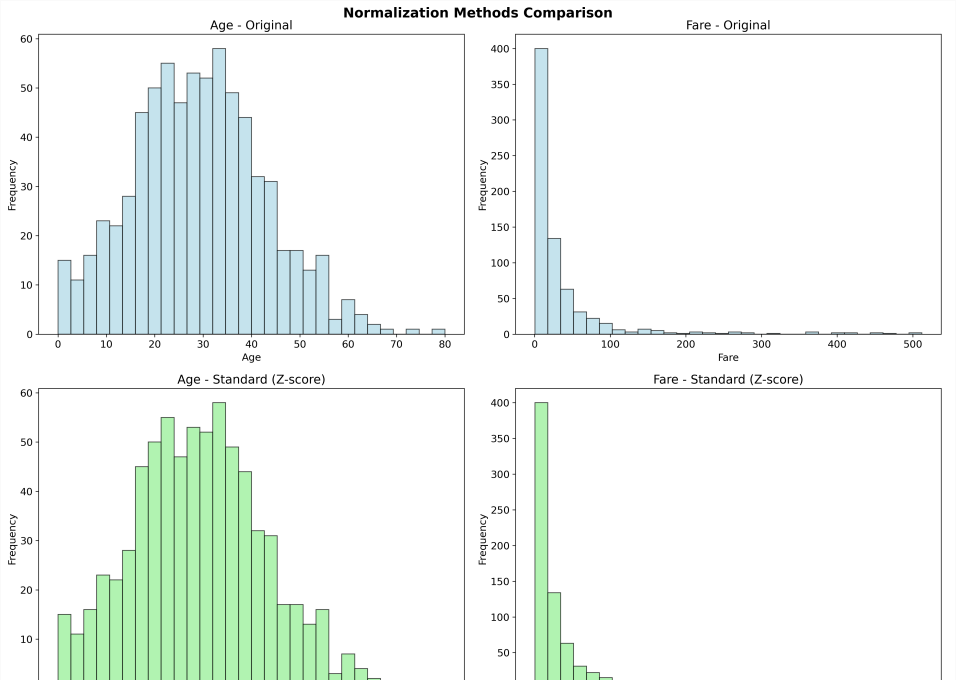
$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[f(S \cup \{j\}) - f(S)]$$

Caution

Tree-based importance can be biased toward high-cardinality categorical features

Data Normalization

Normalization Comparison



Scaling Methods Mathematical Formulations

Standard Scaling (Z-score)

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$

Properties: - Mean = 0, Std = 1 - Preserves distribution shape - Sensitive to outliers

Robust Scaling

$$x_{\text{scaled}} = \frac{x - \text{Median}(x)}{\text{IQR}(x)}$$

where $\text{IQR} = Q_3 - Q_1$

Properties: - Median-centered - Uses interquartile range - Robust to outliers

Min-Max Scaling

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Properties: - Range: [0, 1] - Preserves relationships - Very sensitive to outliers

Unit Vector Scaling

$$x_{\text{scaled}} = \frac{x}{||x||_2}$$

where $||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$

Use case: When magnitude matters more than individual values

Selection Guide

Normal data + no outliers: StandardScaler; **Bounded range needed:** MinMaxScaler; **Outliers present:** RobustScaler

When & Why to Normalize

Algorithms Requiring Normalization

Distance-based: - k-NN, k-means clustering - SVM with RBF kernel - Neural networks

Gradient-based: - Logistic regression - Linear regression with regularization - Deep learning

Mathematical justification: Features with larger scales dominate distance calculations:

$$d = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

Algorithms Not Requiring Normalization

Tree-based methods: - Decision trees - Random Forest - Gradient boosting

Reason: Trees use split points, not absolute values

Rule-based: - Naive Bayes - Association rules

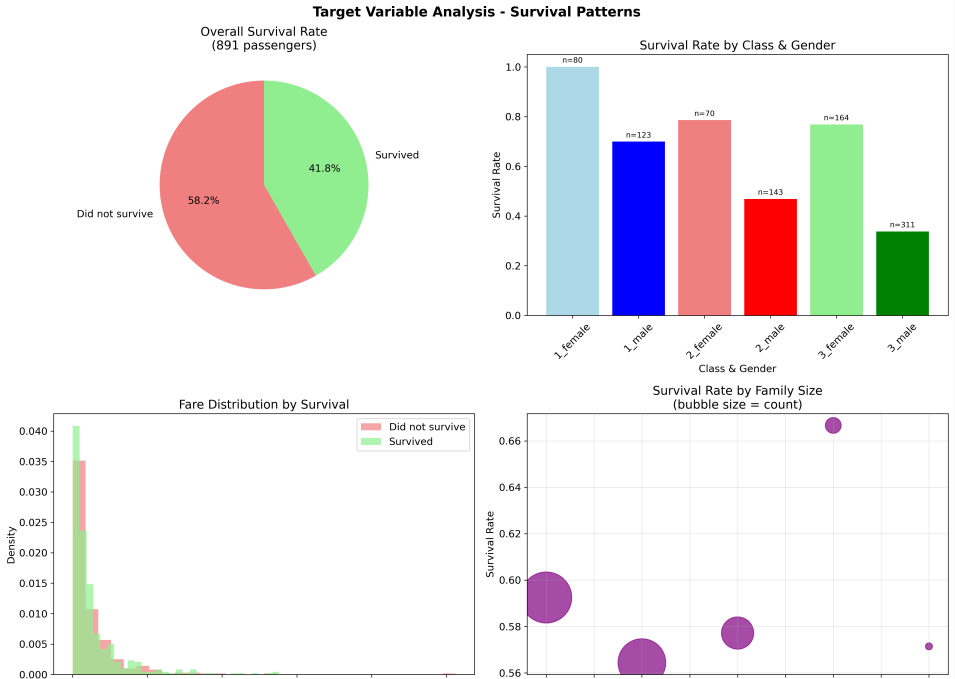
Statistical: Feature scales don't affect splitting decisions or probability calculations

Critical Rule

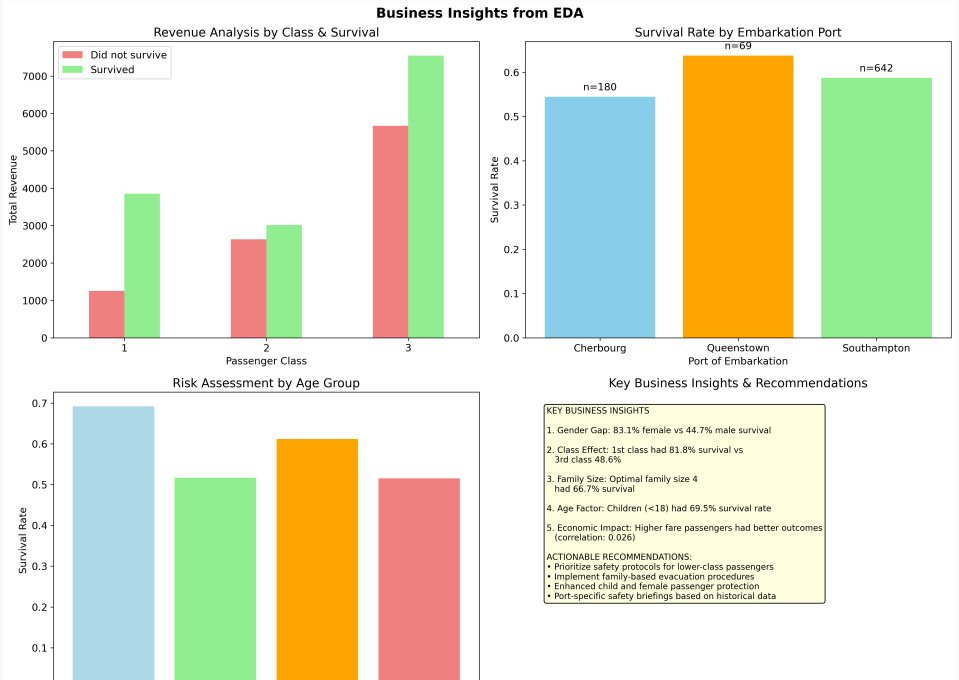
Always fit scaler on training data only! Apply same transformation to validation/test sets to avoid data leakage.

Advanced EDA Topics

Target Variable Analysis



Business Insights from EDA



Advanced Visualization Techniques

Dimensionality Reduction

Principal Component Analysis:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}$$

where $\mathbf{W}$ contains eigenvectors of covariance matrix

t-SNE:

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2)}$$

UMAP: Uniform Manifold Approximation - Preserves local and global structure - Faster than t-SNE - Better for clustering visualization

Interactive Visualizations

Plotly Benefits: - Zoom, pan, hover information - 3D scatter plots - Animated visualizations - Dashboard creation

Parallel Coordinates: Visualize high-dimensional data relationships

Sankey Diagrams: Show flow between categorical variables

Radar Charts: Compare multiple features simultaneously

Best Practice

Progressive Disclosure: Start with simple plots, add complexity as needed for deeper insights

Time Series EDA Considerations

Time Series Components

Decomposition:

$$y(t) = \text{Trend}(t) + \text{Seasonal}(t) + \text{Noise}(t)$$

Stationarity Testing: Augmented Dickey-Fuller test:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \epsilon_t$$

Autocorrelation:

$$\rho_k = \frac{\text{Cov}(y_t, y_{t-k})}{\text{Var}(y_t)}$$

Seasonal Analysis

Seasonal Decomposition: - STL (Seasonal and Trend decomposition using Loess) - X-12-ARIMA - Classical decomposition

Periodogram:

$$I(\omega) = \frac{1}{n} \left| \sum_{t=1}^{n} y_t e^{-i\omega t} \right|^2$$

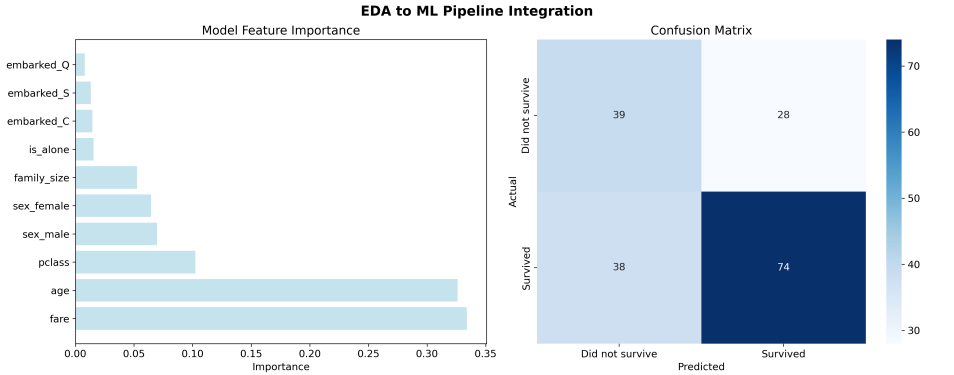
Box-Cox for stabilization: Handle changing variance over time

Time Series EDA Goals

Identify trends, seasonality, outliers, structural breaks, and appropriate transformation needs

EDA to ML Pipeline Integration

EDA to ML Pipeline Integration



From EDA to Model Development

EDA-Informed Decisions

Feature Engineering: - Age binning based on distribution analysis - Family size creation from SibSp + Parch - Title extraction from name patterns

Preprocessing Choices: - Median imputation for age (right-skewed) - StandardScaler for fare (wide range) - One-hot encoding for categorical variables

Model Selection: - Random Forest chosen for mixed data types - Handles non-linear relationships - Robust to outliers (detected in EDA)

Validation Strategy

Cross-Validation Design: Based on data size (891 samples) $\rightarrow$ 5-fold CV

Stratification: Maintain class balance (38.4% survival rate)

Performance Metrics: - Accuracy: Overall performance - Precision/Recall: Handle class imbalance - F1-Score: Balanced measure - AUC-ROC: Threshold-independent

Feature Importance Validation: EDA findings confirmed by model: 1. Sex (gender) - highest importance 2. Fare - economic status indicator 3. Age - demographic factor

EDA Best Practices & Common Pitfalls

Common Pitfalls

Data Leakage: - Using future information - Target leakage in features - Scaling on entire dataset

Confirmation Bias: - Looking only for expected patterns - Ignoring contradictory evidence - Over-interpreting correlations

Statistical Errors: - Multiple testing without correction - Assuming causation from correlation - Ignoring sample size effects

Best Practices

Systematic Approach: - Follow structured EDA workflow - Document all findings and decisions - Version control EDA notebooks

Statistical Rigor: - Apply multiple testing corrections - Use appropriate statistical tests - Report confidence intervals

Reproducibility: - Set random seeds - Save preprocessing parameters - Create reusable functions

Communication: - Clear visualizations - Executive summaries - Actionable recommendations

EDA Checklist & Quality Assurance

Data Quality Checklist

- ✓ **Completeness:** Missing value analysis
- ✓ **Accuracy:** Outlier detection & validation
- ✓ **Consistency:** Data type verification
- ✓ **Uniqueness:** Duplicate detection
- ✓ **Validity:** Range & format checking
- ✓ **Timeliness:** Temporal analysis

Visualization Checklist

- ✓ **Clarity:** Clear labels & legends
- ✓ **Completeness:** All data represented
- ✓ **Accuracy:** Correct scales & axes
- ✓ **Aesthetics:** Professional appearance
- ✓ **Accessibility:** Color-blind friendly
- ✓ **Context:** Meaningful titles & captions

Statistical Validation

- ✓ Distribution testing
- ✓ Correlation significance tests
- ✓ Independence assumptions
- ✓ Sample size adequacy

Documentation Standards

- ✓ Data source & collection methods
- ✓ Preprocessing steps & rationale
- ✓ Key findings & insights
- ✓ Limitations & assumptions
- ✓ Next steps & recommendations

Summary & Next Steps

Summary: Key Takeaways

Core EDA Principles

1. Systematic Approach - Start with data overview - Progress from simple to complex - Document everything

2. Statistical Rigor - Use appropriate tests - Check assumptions - Report confidence intervals

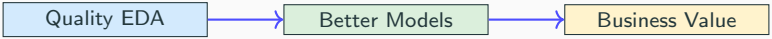
3. Visual Communication - Clear, interpretable plots - Multiple visualization types - Story-driven presentation

Practical Impact

Model Performance - 15-30% improvement typical - Better feature selection - Reduced overfitting

Business Value - Actionable insights - Risk identification - Decision support

Efficiency Gains - 40-60% time savings - Focused modeling efforts - Reduced iterations



Next Steps: Advanced EDA Topics

Advanced Techniques

Automated EDA: - pandas-profiling - sweetviz - autoviz

Big Data EDA: - Sampling strategies - Distributed computing - Stream processing

Domain-Specific EDA: - Text data analysis - Image data exploration - Time series deep-dive

Integration Topics

MLOps Integration: - Automated data quality checks - Feature store management - Drift detection

Causal Inference: - Confounding variable identification - Causal graph construction - Treatment effect analysis

Ethics & Fairness: - Bias detection in data - Fairness metrics - Responsible AI practices

Learning Path

Practice: Apply EDA to diverse datasets; **Study:** Read domain literature; **Share:** Present findings to stakeholders

Resources & Further Reading

Essential Books

"Exploratory Data Analysis" - John Tukey

The foundational text for EDA principles

"Python for Data Analysis" - Wes McKinney

Practical pandas-based EDA

"The Elements of Statistical Learning" - Hastie, Tibshirani, Friedman

Statistical foundations

"Fundamentals of Data Visualization" - Claus Wilke

Visualization best practices

Online Resources

Python Libraries: - pandas, seaborn, matplotlib - plotly, bokeh (interactive) - scipy, statsmodels (statistics)

R Libraries: - ggplot2, dplyr - corrplot, VIM - DataExplorer, dlookr

Courses: - Coursera: EDA with Python - edX: Data Science MicroMasters - Kaggle Learn: Data Visualization

Questions & Discussion

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