Module 02: Data Prep, EDA, and Automated Pipelines

Comprehensive Terms and Concepts Reference

Statistical Terms

Parameter

A characteristic of a population (e.g., population mean μ).

Statistic

A characteristic of a sample (e.g., sample mean x■).

Variable

Any characteristic or attribute that can be measured or counted.

Inference

Drawing conclusions about a population based on sample data.

Measures of Central Tendency

Mean

The average value calculated by summing all values and dividing by the number of values.

Median

The middle value when data is arranged in order. More robust to outliers than mean.

Mode

The most frequently occurring value in the dataset.

Measures of Spread

Variance

Average squared deviation from the mean. Measures how spread out the data is.

Standard Deviation

Square root of variance. Measures spread in original units, making it more interpretable.

Range

Difference between maximum and minimum values.

IQR (Interquartile Range)

Q3 - Q1. Contains the middle 50% of data. Useful for outlier detection.

Measures of Shape

Skewness

Measure of asymmetry in a distribution.

Positive Skew (Right-Skewed)

Tail on the right side. Mean > Median > Mode. Common in income data.

Negative Skew (Left-Skewed)

Tail on the left side. Mean < Median < Mode. Common in test scores.

Kurtosis

Measure of 'tailedness' or presence of extreme values in a distribution.

Mesokurtic

Normal distribution. Kurtosis ≈ 3 (excess kurtosis ≈ 0).

Leptokurtic

Heavy tails with more extreme values. Kurtosis > 3. Higher peak.

Platykurtic

Light tails with fewer extreme values. Kurtosis < 3. Flatter distribution.

Distribution Terms

Quartiles

Values dividing data into four equal parts: Q1 (25th percentile), Q2/Median (50th percentile), Q3 (75th percentile).

Quantiles

Generalization of quartiles that can divide data into any number of equal parts.

Percentiles

Values that divide data into 100 equal parts. The 95th percentile means 95% of data is below this value.

Outliers

Data points that differ significantly from other observations. Can indicate errors or genuine extreme values.

Box Plot

Graphical representation showing median, quartiles, and outliers. Useful for comparing distributions.

Fence

Cutoff value for identifying outliers. Lower fence = Q1 - 1.5 × IQR, Upper fence = Q3 + 1.5 × IQR.

Whiskers

Lines extending from the box in a box plot. Show the range of data within the fences.

Data Types

Numerical Data:

Discrete: Countable values with no intermediate values (e.g., number of children, counts).

Continuous: Measurable values that can take any value in a range (e.g., height, temperature, time).

Categorical Data:

Nominal: No inherent order. Categories are just labels (e.g., colors, names, cities).

Ordinal: Meaningful order but unequal intervals (e.g., rankings, education levels, satisfaction ratings).

Special Numerical Types:

Interval: Ordered with equal distances but no true zero (e.g., temperature in Celsius).

Ratio: Has a true zero point allowing for meaningful ratios (e.g., height, weight, income).

Data Quality Terms

Cardinality

Number of unique values in a dataset or column. High cardinality (many unique values) affects encoding strategies.

Reliability

Consistency and reproducibility of measurements. Can results be replicated?

Validity

Accuracy of measurements in representing true values. Does it measure what it claims to measure?

Precision

Level of detail or exactness in measurement. How specific is the measurement?

Accuracy

How close measurements are to the true value. A measurement can be precise but not accurate.

Missing Data Mechanisms

MCAR (Missing Completely at Random)

Missingness is unrelated to any observed or unobserved variables. Occurs entirely by chance. Safe to delete or impute without bias.

MAR (Missing at Random)

Missingness depends on observed variables but not on the missing values themselves. Can be handled with appropriate imputation methods.

MNAR (Missing Not at Random)

Missingness depends on the missing values themselves. Most complex type. Requires careful handling to avoid bias.

Data Preparation Concepts

Constants

Features with constant values (same value for all rows) should be deleted because they provide zero information for machine learning models. They cannot help distinguish between data points and may cause mathematical errors (e.g., division by zero in standardization).

Quasi-Constants

Features where one value dominates (e.g., 95-99% of observations). These provide minimal predictive value and should typically be removed. Common thresholds: 95%, 98%, or 99% of values being the same.

Imputation

Process of replacing missing data with substituted values rather than deleting rows. Preserves sample size and can maintain statistical power.

Common Imputation Techniques:

Method	How It Works	Best Used For
Mean	Replace with column average	Numerical data without outliers
Median	Replace with middle value	Numerical data with outliers (robust)
Mode	Replace with most frequent value	Categorical data
KNN	Use similar rows to predict	Complex datasets with feature relationships
LOCF	Use previous known value	Time-series data

Outliers

Data points that differ significantly from other observations. Can result from measurement errors, data entry errors, or genuine extreme values.

Detection Methods:

Z-Score Method: Points beyond ±3 standard deviations from mean are typically considered outliers.

IQR Method: Points beyond Q1 - 1.5×IQR or Q3 + 1.5×IQR are flagged as outliers.

Handling Strategies:

Trimming: Remove outlier rows entirely. Best for clear data errors.

Capping (Winsorization): Cap outliers at a threshold rather than deleting them.

Transformation: Apply mathematical functions (log, square root) to reduce outlier impact.

Imputation: Treat outliers as missing values and replace using imputation techniques.

Feature Engineering

Discretization (Binning)

Converting continuous numerical data into discrete bins or intervals. Reduces noise, improves interpretability, and manages outliers.

Method	Description	Example
Equal Width	Intervals of equal size	Ages: 0-20, 21-40, 41-60
Equal Frequency	Same number of observations per bin	10 students per bin
Custom/Domain	Business logic rules	Minor (<18) vs Adult (≥18)

Scaling

Transforming numerical data so all features occupy similar ranges. Essential for distance-based algorithms and gradient descent optimization.

Method	Range	Formula	Best For	Outliers
Min-Max	[0, 1]	(x - min) / (max - min)	Neural nets, Images	Sensitive
Standardization	No fixed	(x - μ) / σ	Regression, PCA	Decent
Robust	No fixed	(x - median) / IQR	Data with many outliers	Excellent

Categorical Encoding

Process of converting text-based categories into numerical format for machine learning algorithms.

Label/Ordinal Encoding

Assigns unique integers to each category (1, 2, 3). Best for ordinal data with natural ranking (Small, Medium, Large). Risk: May introduce false mathematical relationships for nominal data.

One-Hot Encoding

Creates binary column for each category. Best for nominal data without order (City, Color, Department). Risk: High-cardinality features can cause dimensionality explosion (Dummy Variable Trap).

Frequency Encoding

Replaces categories with their frequency count or percentage. Best for high-cardinality features (100+ categories). Captures importance of categories without creating many columns.

Target Encoding

Replaces categories with the mean of the target variable for that category. Risk: Can cause data leakage if not done properly within cross-validation folds.

Exploratory Data Analysis (EDA)

Univariate Analysis

Examining one variable at a time using histograms, box plots, summary statistics.

Bivariate Analysis

Examining relationships between two variables using scatter plots, correlation, contingency tables.

Multivariate Analysis

Examining relationships among three or more variables using correlation matrices, pair plots, heatmaps.

Correlation

Statistical measure of linear relationship between variables. Ranges from -1 to +1.

Pearson Correlation

Measures linear relationships. Sensitive to outliers and assumes normality.

Spearman Correlation

Measures monotonic relationships using ranks. More robust to outliers.

Interpreting Correlation Values:

0.0 - 0.3: Weak correlation

0.3 - 0.7: Moderate correlation

0.7 - 1.0: Strong correlation

Near +1: Strong positive relationship

Near -1: Strong negative (inverse) relationship

Near 0: No linear relationship

Common Statistical Tests in EDA

T-Test

Compares means between two groups. Tests if difference is statistically significant.

ANOVA

Compares means across multiple groups. Extension of t-test for 3+ groups.

Chi-Square Test

Tests independence between categorical variables. Used for contingency tables.

Cramér's V

Effect size measure for chi-square. Indicates strength of association (0 to 1).

Automated Pipelines

Pipeline

sklearn object that chains preprocessing steps and estimators. Ensures correct order of operations.

ColumnTransformer

Applies different transformations to different column subsets (e.g., numerical vs categorical).

Transformer

Object that implements fit() and transform() methods. Used for preprocessing steps.

BaseEstimator

Base class providing get_params() and set_params() for sklearn compatibility.

TransformerMixin

Mixin class providing fit_transform() by combining fit() and transform().

Why Use Pipelines?

Reproducibility: Same preprocessing applied consistently across train and test sets.

Prevent Data Leakage: Fitting only on training data within cross-validation folds.

Clean Code: Encapsulates preprocessing logic in reusable components.

Easy Deployment: Single object can be serialized and deployed to production.

Proper Cross-Validation: Ensures preprocessing happens within each CV fold.

Common Pipeline Components

Component	Purpose	Example
SimpleImputer	Handle missing values	strategy="mean" or "median"
StandardScaler	Standardize features	Mean=0, StdDev=1
MinMaxScaler	Scale to range	Scale to [0, 1]
OneHotEncoder	Encode categories	Create dummy variables
LabelEncoder	Encode ordinal data	Convert to integers
PolynomialFeatures	Create interactions	x■, x■ → x■, x■, x■², x■x■, x■²

Advanced Concepts

Data Leakage

When information from test set influences training. Common causes: fitting scalers on full data, target leakage in features.

Multicollinearity

High correlation between independent variables. Can cause unstable coefficients and difficult interpretation.

VIF (Variance Inflation Factor)

Measure of multicollinearity. VIF > 10 indicates problematic correlation.

Feature Importance

Measure of how much each feature contributes to model predictions. Varies by algorithm.

Cross-Validation

Technique for assessing model performance by splitting data into multiple train-test folds.

Stratified Sampling

Sampling that preserves the proportion of classes in classification problems.

Best Practices

Data Preparation Order

1. Remove duplicates and constant features
2. Handle missing data
3. Detect and handle outliers
4. Feature engineering (create new features)
5. Encode categorical variables
6. Scale numerical features
7. Feature selection (if needed)

EDA Best Practices

- Start with univariate analysis before moving to bivariate and multivariate
- Visualize everything - plots reveal patterns statistics might miss
- Check for missing data patterns and understand why data is missing
- Always investigate outliers before removing them
- Test assumptions (normality, independence, homoscedasticity)
- Document all insights and decisions for reproducibility
- Consider domain knowledge when interpreting results

Pipeline Best Practices

- Always fit preprocessing on training data only
- Use ColumnTransformer for mixed data types
- Handle missing values before scaling
- Encode categorical variables before modeling
- Create custom transformers for complex preprocessing
- Save pipelines using joblib for deployment
- Test pipelines thoroughly on unseen data
- Version control your pipelines along with code

Common Pitfalls to Avoid

- Fitting scalers on entire dataset before train-test split (data leakage)
- Removing outliers without investigation (may lose important information)
- Using mean imputation on skewed data (use median instead)
- One-hot encoding high-cardinality features (use frequency/target encoding)
- Ignoring the dummy variable trap (always drop first category)

- Not handling unseen categories in test set (use handle_unknown='ignore')
- Scaling before imputation (will create NaN in scaled output)
- Using correlation to infer causation (correlation ≠ causation)

Interview Tips

- Always explain your reasoning for choosing specific techniques
- Mention trade-offs (e.g., mean vs median imputation)
- Discuss how to handle edge cases (new categories, extreme outliers)
- Show awareness of data leakage and how to prevent it
- Demonstrate knowledge of when to use which encoding method
- Connect preprocessing choices to model selection
- Be ready to explain pipeline benefits for production deployment
- Know how to detect and handle class imbalance