Missing Data Strategies

Missing Patterns

- **MCAR**: Missing Completely At Random
- **MAR**: Missing At Random
- **MNAR**: Missing Not At Random

Visualize patterns with missing data heatmap

Imputation Methods

- **Mean/Median** Imputation
- **KNN** Imputation
- **MICE**: Multiple Imputation
- **Deep Learning** Based

Impact Analysis

Compare model performance before and after imputation, sensitivity analysis, impact assessment by missing rate

Missing Patterns in Detail

MCAR (Missing Completely At Random)

Missingness occurs completely randomly, independent of other variables. Data loss exists but no bias.

Example: Random non-responses in a survey

MAR (Missing At Random)

Missingness depends on observed variables but not on the missing value itself. Most common pattern.

Example: Older people are more likely to omit income information

MNAR (Missing Not At Random)

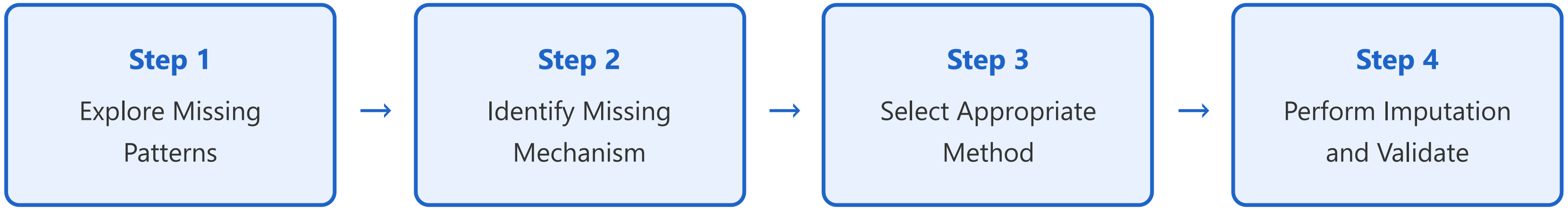
Missingness is related to the missing value itself. Most difficult pattern to handle.

Example: People with very high or low income omit income information

Imputation Methods Comparison

Method	Advantages	Disadvantages	Suitable Situations
Mean/Median	Fast and simple	Reduces variance, distorts relationships	MCAR, low missing rate
KNN	Utilizes similar cases	High computational cost	MAR, moderate missing rate
MICE	Reflects uncertainty	Complex and slow	MAR, high missing rate
Deep Learning	Learns complex patterns	Requires large data	Large-scale datasets

Missing Data Handling Process



Step 1
Explore Missing Patterns

→

Step 2
Identify Missing Mechanism

→

Step 3
Select Appropriate Method

→

Step 4
Perform Imputation and Validate

Performance Comparison Example

Original Accuracy
94.2%

Mean Imputation
89.5%

KNN Imputation
92.8%

MICE Imputation
93.6%

Key Considerations

⚠️ **Cautions**

- Consider variable removal over imputation if missing rate exceeds 40%
- Check impact of imputation method on target variable
- Compare multiple methods for optimal selection

✓ **Recommendations**

- Derive insights through missing pattern visualization
- Evaluate imputation methods with cross-validation
- Leverage domain knowledge for imputation

Python Implementation Example

```
# Check missing patterns

import missingno as msno
msno.matrix(df)

# Mean imputation
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='mean')
df_filled = imputer.fit_transform(df)

# KNN imputation
from sklearn.impute import KNNImputer
imputer = KNNImputer(n_neighbors=5)
df_filled = imputer.fit_transform(df)

# MICE imputation
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
imputer = IterativeImputer()
df_filled = imputer.fit_transform(df)
```

Recommended Strategy by Missing Rate

≤ 5%	5-20%	20-40%	> 40%

Simple Imputation
(Mean/Median)

KNN or
Regression Imputation

MICE or
Advanced Methods

Consider Variable
Removal