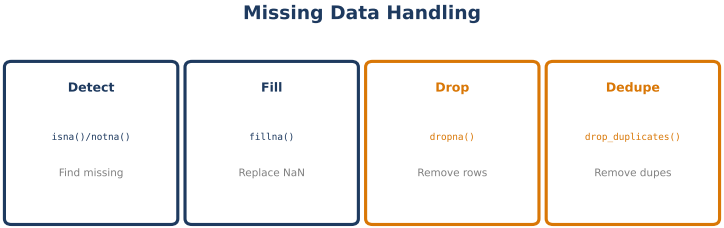
Lesson 07 Summary: Missing Data

Data Science with Python – Key Concepts

Data Science Program

Missing Data Overview



Missing data handling is critical for accurate analysis

Detecting Missing Data

Finding NaN values in your data:

- **df.isna():** Returns True where values are missing
- **df.notna():** Returns True where values exist
- **df.isna().sum():** Count missing per column
- **df.isna().sum().sum():** Total missing values

Quick check:

```
print(df.isna().sum()) # Missing per column
```

Always check for missing values before analysis

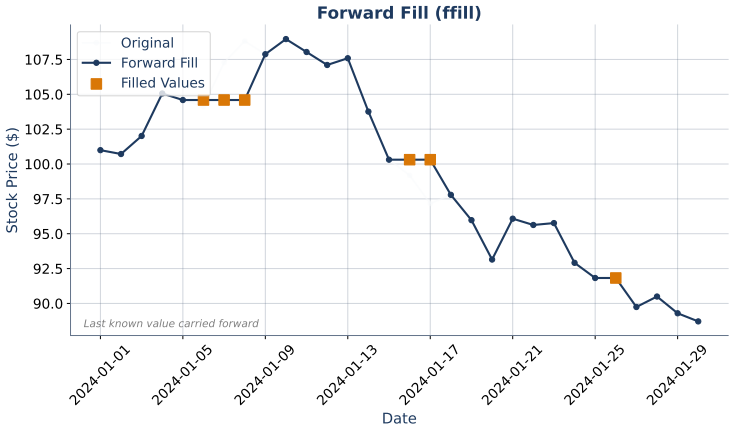
Fill Methods

Four ways to replace missing values:

- **Forward fill:** `df.fillna(method='ffill')`
  Use previous value – good for time series
- **Backward fill:** `df.fillna(method='bfill')`
  Use next value
- **Mean fill:** `df.fillna(df.mean())`
  Use column average
- **Interpolate:** `df.interpolate()`
  Linear interpolation between values

Choose fill method based on data characteristics

Forward Fill



Forward fill carries last known value forward

Dropping Missing Data

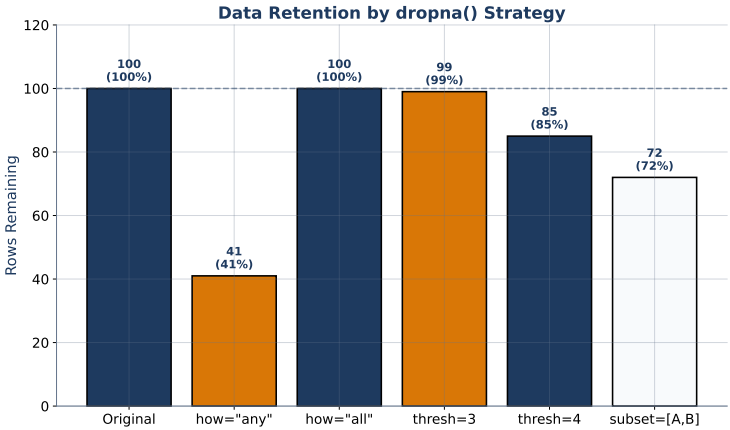
Remove rows or columns with NaN:

- **Drop any NaN:** `df.dropna()`
- **Drop if all NaN:** `df.dropna(how='all')`
- **Drop columns:** `df.dropna(axis=1)`
- **Threshold:** `df.dropna(thresh=3)`
  Keep rows with at least 3 non-null values

Caution: Dropping may lose valuable data!

Consider if dropping is appropriate for your analysis

Data Retention



Balance data quality with data quantity

Handling Duplicates

Finding and removing duplicate rows:

- **Find:** `df.duplicated()`
- **Count:** `df.duplicated().sum()`
- **Remove:** `df.drop_duplicates()`
- **Keep last:** `df.drop_duplicates(keep='last')`

Subset columns:

```
df.drop_duplicates(subset=['Date', 'Symbol'])
```

Duplicates can skew statistics and counts

Cleaning Workflow

Standard data cleaning process:

1. Load data: `pd.read_csv()`
2. Check missing: `df.isna().sum()`
3. Decide strategy: fill or drop
4. Handle duplicates: `drop_duplicates()`
5. Verify: check shape and statistics
6. Document decisions

Document all cleaning decisions for reproducibility

Quick Reference

Essential Missing Data Operations:

Operation	Syntax
Detect missing	`df.isna()`
Count missing	`df.isna().sum()`
Forward fill	`df.fillna(method='ffill')`
Mean fill	`df.fillna(df.mean())`
Interpolate	`df.interpolate()`
Drop rows	`df.dropna()`
Find duplicates	`df.duplicated()`
Remove duplicates	`df.drop_duplicates()`

Clean data is the foundation of good analysis