

CASE STUDY ANALYSIS

Data Engineering with Pandas

End-to-End Pipeline for the Titanic Dataset



Phase 1: Data Profiling

Inspecting the raw signals and statistical
distribution

Titanic Statistical Overview

Using `df.describe()` allows us to identify count mismatches and potential skewness in numerical variables.

Titanic Statistical Overview					
Survived	891	0.38	0.48	0.00	1.00
Age	714	29.69	14.52	0.42	80.00
Fare	891	32.20	49.69	0.00	512.32

Observation: Age requires 177 imputations; Fare shows massive outlier potential (Max vs Mean).

| Ingesting Raw Sources

Pandas facilitates direct streaming of datasets from cloud repositories, ensuring a single source of truth for the engineering pipeline.

```
import pandas as pd url = "https://bit.ly/titanic-data" df = pd.read_csv(url) # Previewing structural health
```

Treating Missing Values



Dropping

Features with >70% missingness (e.g. 'Cabin') are removed to prevent noise injection.



Imputing

Numerical nulls ('Age') are replaced with the **Median** to preserve central tendency.



Mode Fill

Categorical nulls ('Embarked') are filled with the most frequent value (Mode).

Implementation: Imputation Logic

Numerical Strategy

Filling 177 missing Age values using the median to avoid skewing the distribution.

```
# Median Imputation age_median = df['Age'].median() df['Age'].fillna(age_median, inplace=True)
```

Theoretical Framework: IQR

The Interquartile Range (IQR) method is a robust statistical tool for defining "Normal" ranges without being affected by extreme outliers.

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

Data points falling outside $[Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR}]$ are considered anomalies.

 Boxplot Outliers

Pipeline Logic: IQR Filtering Code

Applying the IQR filter to the 'Fare' column to remove extreme premium ticket holders that skew analysis.

```
# 1. Calculating Quantiles and IQR Q1 = df['Fare'].quantile(0.25) Q3 = df['Fare'].quantile(0.75) IQR = Q3 -
```

*This transformation ensures the model focuses on representative data trends.

Visualizing the Baseline

Removing Fare outliers and imputing Age results in a "Clean Baseline Matrix."

This dataset is now statistically normalized, reducing the noise-to-signal ratio significantly.

- ✓ Robust Variance
- ✓ No Missing Observations

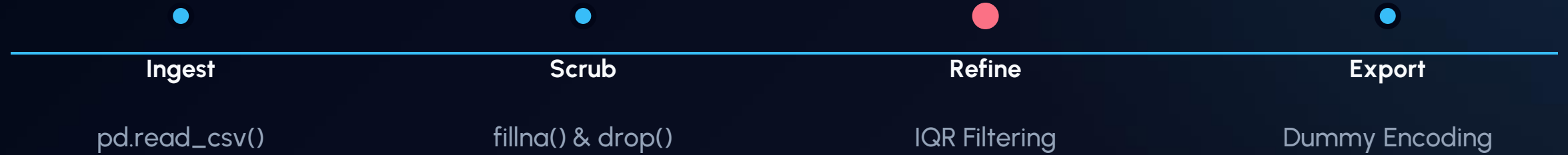


| Encoding Categorical Data

Machine learning algorithms require numerical inputs. Categorical text like 'Sex' must be converted into numerical dummy variables.

```
# One-Hot Encoding df_final = pd.get_dummies(df_clean, columns=['Sex', 'Embarked'], drop_first=True) # Result
```

The Full Pipeline Flow



From raw messy CSV to a computational-ready Baseline Matrix.

Questions?

Module 2: Data Engineering & Exploration with
Pandas



technical-
lead@pandas.edu



www.pandas-
engineering.io

Image Sources



<https://img.icons8.com/color/512/pandas.png>

Source: [icons8.com](https://img.icons8.com/)



https://miro.medium.com/O*XG2sFucPoFMg6NeV.png

Source: [medium.com](https://miro.medium.com/)



for

[unsplash.com](https://plus.unsplash.com/premium_photo-1764695511875-5f939ab9a031?fm=jpg&q=60&w=3000&auto=format&fit=crop&ixlib=rb-4.1.0&ixid=M3wxMjA3fDB8MHxwaG90bylwYWdlfHx8fGVufDB8fHx8fA%3D%3D)

https://plus.unsplash.com/premium_photo-1764695511875-5f939ab9a031?fm=jpg&q=60&w=3000&auto=format&fit=crop&ixlib=rb-4.1.0&ixid=M3wxMjA3fDB8MHxwaG90bylwYWdlfHx8fGVufDB8fHx8fA%3D%3D

Source: [unsplash.com](https://plus.unsplash.com/)



https://img.freepik.com/free-vector/big-data-background-technology-visualization-artificial_598544-127.jpg?semt=ais_hybrid&w=740&q=80

Source: [www.freepik.com](https://img.freepik.com/)
