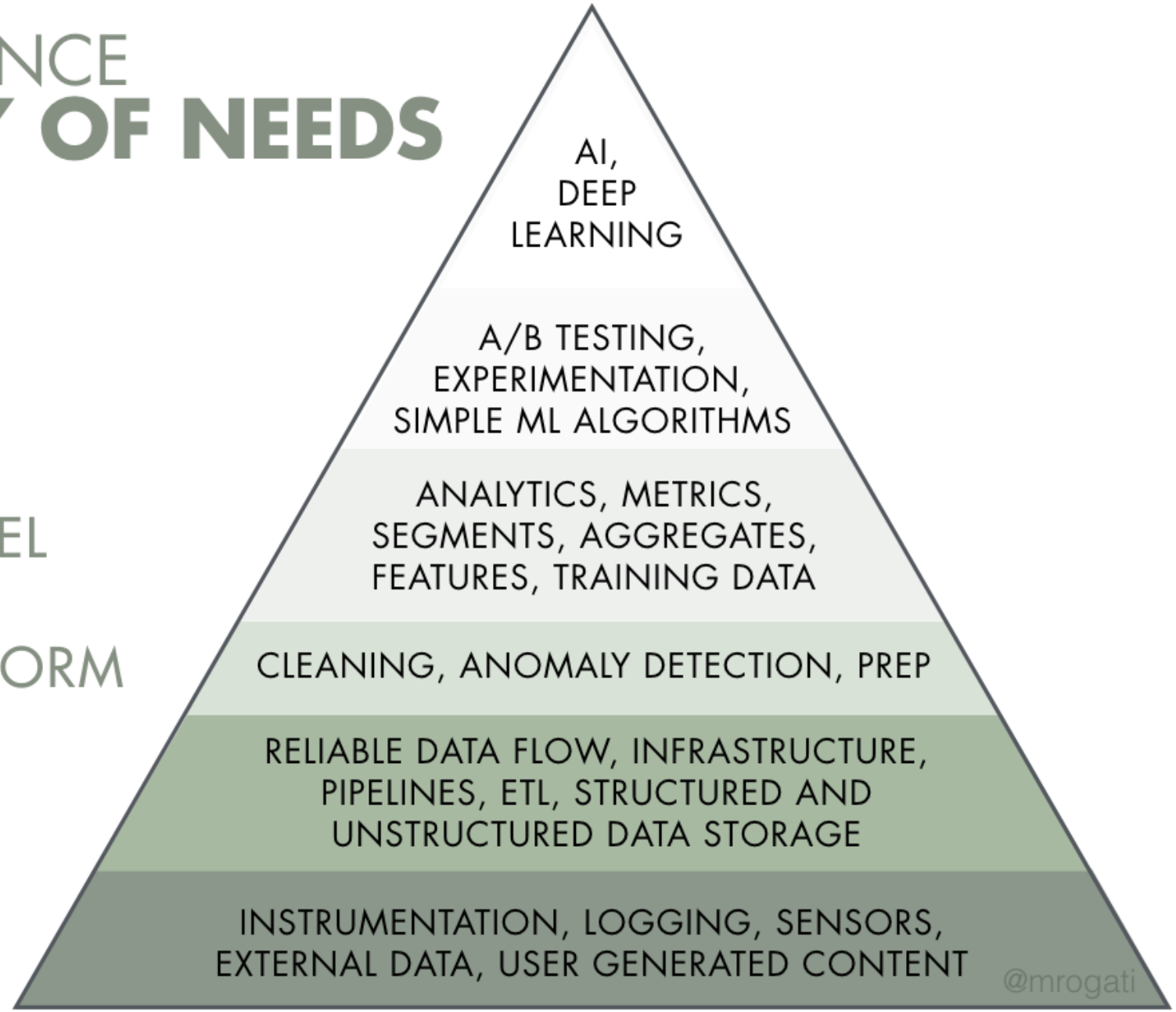
Data Set Cards

CSC 2621 Introduction to Data Science

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Exploratory Data Analysis

- EDA is the process of visually and statistically exploring your dataset to understand its structure, patterns, and any underlying relationships. It's the detective work of data science.

- Benefits:
  - Build familiarity with the data
  - Identify relationships between features and response
  - Understand what data might be missing, incomplete, or “wrong”

EDA

- Goals:
  - Uncover Patterns
  - Identify Outliers
  - Understand Distributions
  - Feature Selection and Opportunities for Feature Engineering
  - Assumption Checking
  - Summarize 1000’s of Observations Visually

Data Cards

- Adapted from hugging face
  - Each dataset may be documented by a readme.md
- Concise documents designed to describe and track data sets
- Provides a comprehensive snapshot of the datasets
  - Origin
  - Collection methods
  - Features
  - Statistical Insights

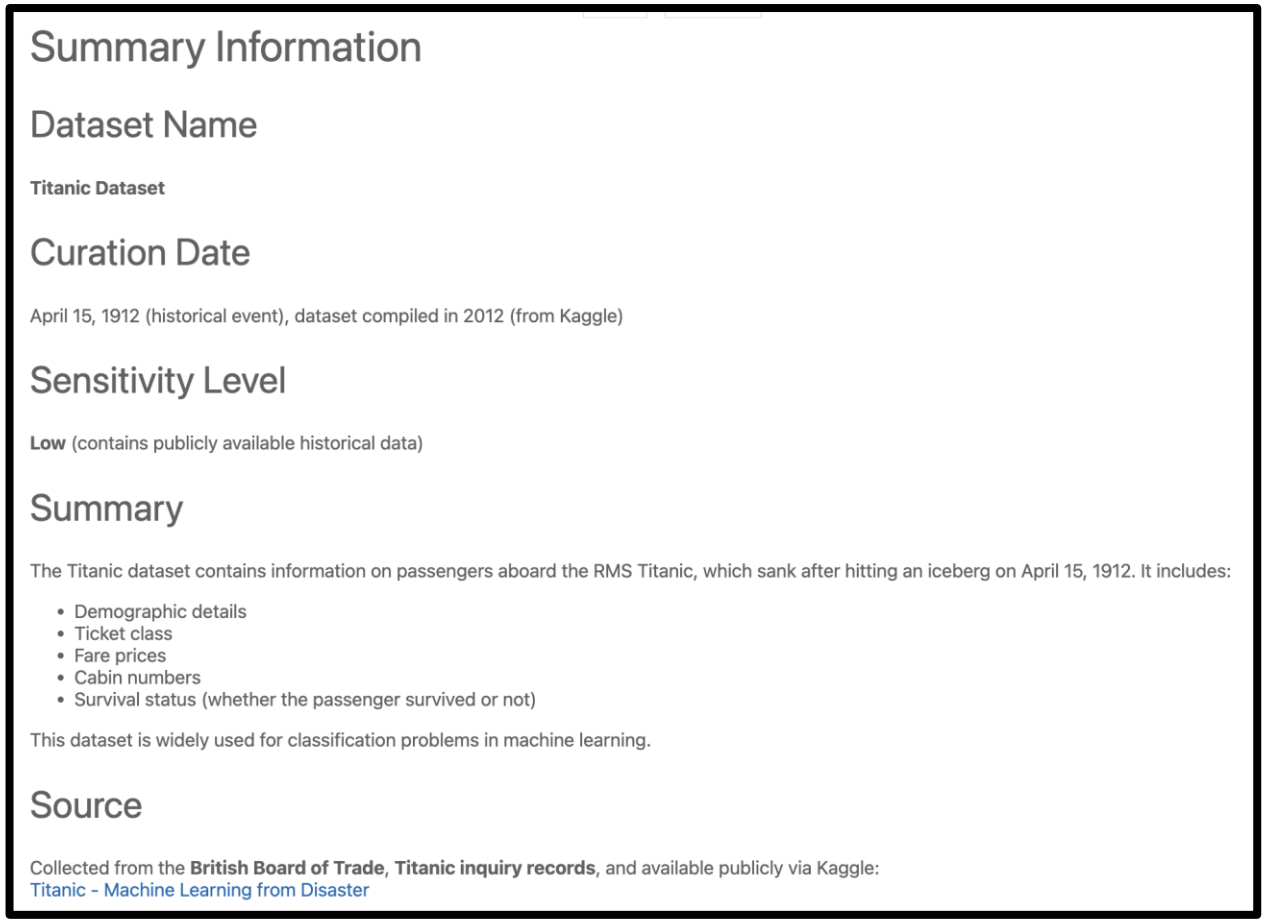
Importance of Data Cards

- Why do we need them?
  - In the pursuit of data science, clarity about your data is as vital as the data itself. Data Set Cards serve as a compass, guiding you through the intricacies of your dataset.

Summary Information

- Describes the dataset and its purpose.

- Key fields:
  - Name
  - Curation Date
  - Sensitivity Level
  - Summary
  - Source



Summary Information

Dataset Name

Titanic Dataset

Curation Date

April 15, 1912 (historical event), dataset compiled in 2012 (from Kaggle)

Sensitivity Level

Low (contains publicly available historical data)

Summary

The Titanic dataset contains information on passengers aboard the RMS Titanic, which sank after hitting an iceberg on April 15, 1912. It includes:

- Demographic details
- Ticket class
- Fare prices
- Cabin numbers
- Survival status (whether the passenger survived or not)

This dataset is widely used for classification problems in machine learning.

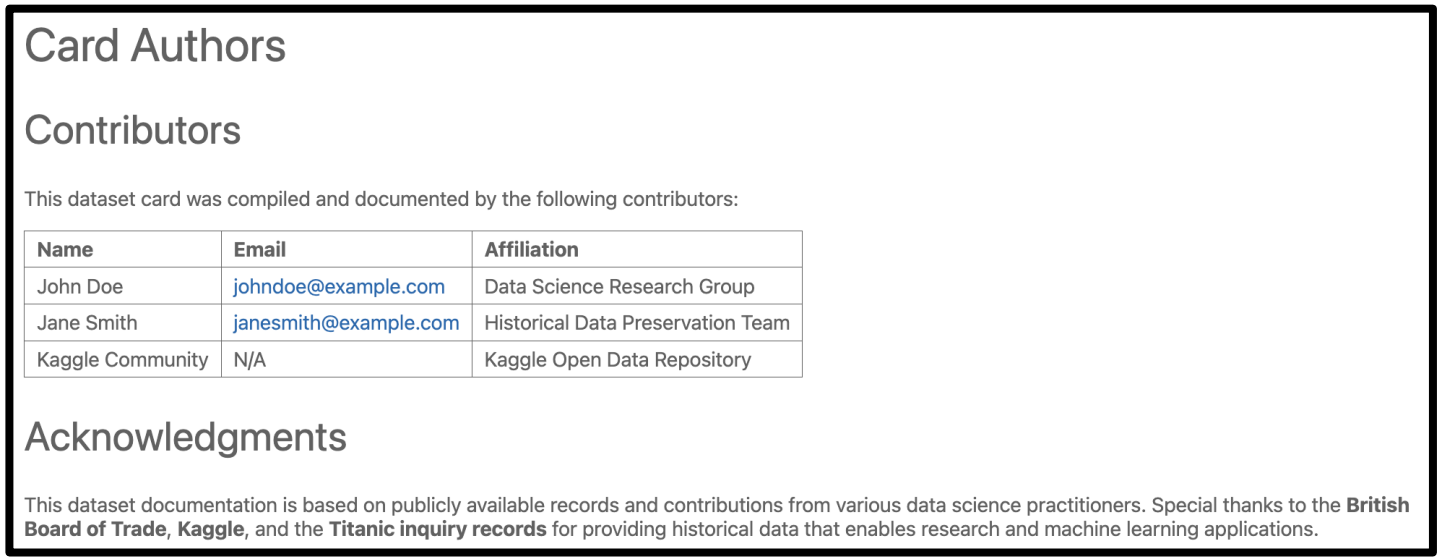
Source

Collected from the **British Board of Trade**, **Titanic inquiry records**, and available publicly via Kaggle:
Titanic - Machine Learning from Disaster

Card Authors

- Documents the contributors who created the dataset card.

- Why it matters:
  - Tracks authorship and accountability
  - Enables collaboration



Card Authors

Contributors

This dataset card was compiled and documented by the following contributors:

Name	Email	Affiliation
John Doe	johndoe@example.com	Data Science Research Group
Jane Smith	janesmith@example.com	Historical Data Preservation Team
Kaggle Community	N/A	Kaggle Open Data Repository

Acknowledgments

This dataset documentation is based on publicly available records and contributions from various data science practitioners. Special thanks to the **British Board of Trade**, **Kaggle**, and the **Titanic inquiry records** for providing historical data that enables research and machine learning applications.

Known Sensitive Features

- Highlights potentially sensitive attributes in the dataset.

- Risk Reduction:
  - Anonymization
  - Aggregation

Sensitive Features

Feature Name	Sensitive Type	Risk Reduction Method
Name	Personally Identifiable Information (PII)	Anonymization or removal for privacy
Ticket Number	Potentially Identifiable	Partial obfuscation if needed
Cabin Number	Potentially Identifiable	Removal or generalization
Age	Demographic Data	Used responsibly to avoid bias
Fare	Socioeconomic Indicator	Awareness of potential bias

Considerations

- **Ethical Use**: While the dataset does not contain explicit personal identifiers, it does include demographic and socioeconomic indicators that should be handled responsibly in analyses.
- **Bias & Fairness**: Be mindful of potential biases in survival prediction models, as factors like gender, class, and age played significant roles in survival outcomes.
- **Historical Context**: This dataset reflects societal structures of the early 1900s and should not be used to make generalized predictions about modern scenarios.

Data Overview

- Provides a high-level summary of the dataset characteristics.

- Key fields:
  - Storage Size
  - Number of Rows
  - Number of Features
  - Data Format

Data Overview

General Description

The Titanic dataset contains structured information about passengers aboard the RMS Titanic. It includes demographic attributes, ticket and fare details, and survival outcomes.

Dataset Characteristics

Field	Value
Storage Size	~60 KB (CSV format)
Number of Rows	891 (train set) / 418 (test set)
Number of Features	12
Data Format	CSV, XLSX

Feature Summary

Feature Name	Data Type	Description
PassengerId	Integer	Unique ID for each passenger
Survived	Integer	1 = Survived, 0 = Did not survive
Pclass	Integer	Ticket class (1st, 2nd, 3rd)
Name	String	Passenger's name
Sex	String	Male/Female
Age	Float	Age in years

Numerical Features Summary

- Summarizes key statistics for numerical variables.

- Metrics:
  - Count
  - Mean
  - Standard Deviation
  - Min/Max values
  - Quartiles (25%, 50%, 75%)

Numerical Features Summary

Overview

The Titanic dataset contains several numerical features that describe passenger attributes such as **age, fare, and family size**. Below is a summary of key statistics for each numerical feature.

Summary Statistics

Feature	Count	Mean	Std Dev	Min	25%	Median	75%	Max
Age	714	29.70	14.53	0.42	20.12	28.00	38.00	80.00
SibSp	891	0.52	1.10	0	0	0	1	8
Parch	891	0.38	0.81	0	0	0	0	6
Fare	891	32.20	49.69	0	7.91	14.45	31.00	512.33

Key Insights

- **Age**: The median age of passengers is **28 years**, with a wide range from **0.42 to 80 years**.
- **SibSp** (Siblings/Spouses aboard): Most passengers traveled **alone or with one companion**.
- **Parch** (Parents/Children aboard): The majority of passengers had **no family members onboard**.
- **Fare**: Fares vary significantly, with a **maximum value of 512.33 GBP**, indicating some **high-paying first-class passengers**.

Considerations

- **Missing Data**: The `Age` column has **177 missing values** that need to be addressed.
- **Skewed Fare Distribution**: The **high standard deviation (49.69 GBP)** suggests that fare prices are **right-skewed**, likely due to first-class passengers paying much higher fares.
- **Feature Engineering**: New features such as **Family Size (SibSp + Parch + 1)** could provide additional insights.

Categorical Features Summary

- Summarizes key statistics for categorical variables.

- Metrics:
  - Unique Values
  - Most Common Value

Categorical Features Summary

Overview

The Titanic dataset includes several categorical features that describe **passenger demographics, ticket class, and embarkation points**. Below is a summary of unique values and most common categories.

Summary Statistics

Feature	Unique Values	Most Common Value (Mode)
Sex	2	Male
Pclass	3	3rd Class
Embarked	3	Southampton (S)
Cabin	147	Missing (NaN)
Ticket	681	Unique Per Passenger

Key Insights

- **Sex**: The dataset has a nearly even split between Male and Female passengers, with **Male** being the most common.
- **Pclass (Ticket Class)**: Most passengers traveled in **3rd class**, reflecting affordability.
- **Embarked**: The majority of passengers boarded at **Southampton (S)**, which was Titanic's main departure port.
- **Cabin**: There are **147 unique cabin numbers**, but **many values are missing** (NaN).
- **Ticket**: There are **681 unique ticket numbers**, suggesting that many passengers had individual ticket purchases.

Considerations

- **Missing Data**: The `Cabin` column has a **large number of missing values** and may not be useful without imputation.
- **Ordinal Encoding**: `Pclass` can be treated as an **ordinal feature** (1st class > 2nd class > 3rd class).
- **One-Hot Encoding**: `Embarked` and `Sex` should be **one-hot encoded** for machine learning models.
- **Ticket Uniqueness**: Since many ticket numbers are unique, this feature may not contribute meaningfully to prediction models.

Field Information

- Documents each feature in the dataset.

- Includes:
- - Feature Name
- - Data Type
- - Statistical Type
- - Description

Field Information

Overview

The Titanic dataset consists of various features that describe passenger demographics, travel details, and survival outcomes. Below is a structured breakdown of each feature, its data type, and its role in analysis.

Feature Details

Feature Name	Data Type	Statistical Type	Description
PassengerId	Integer	Identifier	Unique ID assigned to each passenger
Survived	Integer	Binary	Survival status (1 = Survived, 0 = Did not survive)
Pclass	Integer	Ordinal	Ticket class (1st = Luxury, 2nd = Mid-tier, 3rd = Economy)
Name	String	Nominal	Full name of the passenger
Sex	String	Nominal	Gender (Male/Female)
Age	Float	Continuous	Passenger's age in years
SibSp	Integer	Discrete	Number of siblings/spouses aboard the Titanic
Parch	Integer	Discrete	Number of parents/children aboard the Titanic
Ticket	String	Nominal	Ticket number
Fare	Float	Continuous	Ticket fare in British pounds (£)
Cabin	String	Nominal	Cabin number (if assigned)
Embarked	String	Nominal	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Considerations

- **PassengerId**: Serves as a unique identifier but does not contribute to survival prediction.
- **Pclass**: This is an **ordinal feature**, meaning it has an inherent order (1st > 2nd > 3rd class).
- **Age & Fare**: These **continuous variables** may need normalization for machine learning models.
- **Sex, Embarked**: These are **categorical features** and should be one-hot encoded when used in models.
- **Cabin**: Has many missing values and might need imputation or exclusion.
- **SibSp & Parch**: Can be combined to create a **"Family Size"** feature for better predictive modeling.

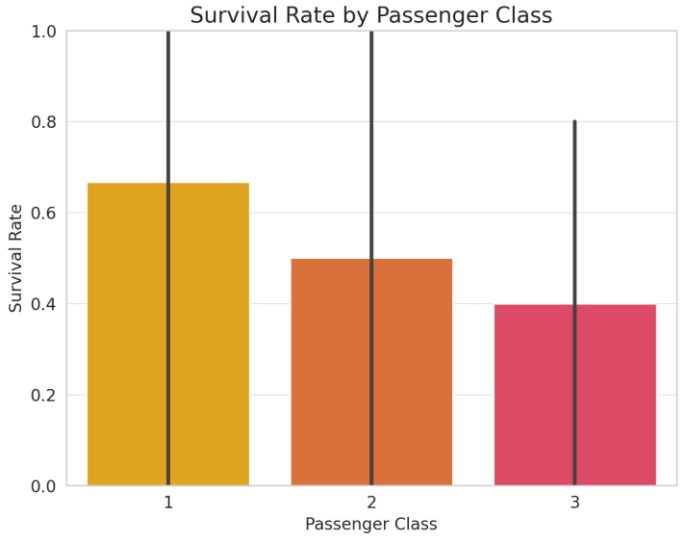
Example Entry

- Provides representative dataset examples.

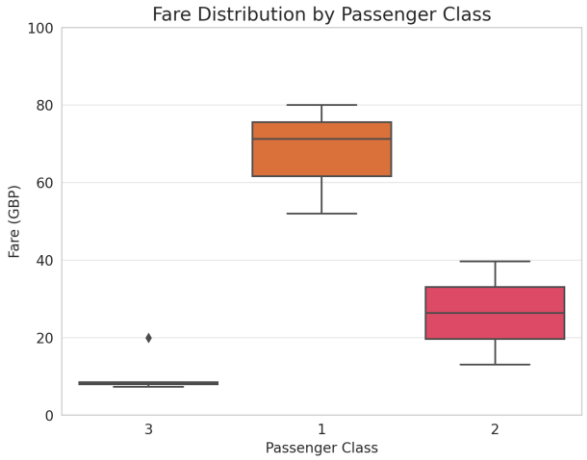
- Why?
  - Helps users understand typical and edge-case data points
  - Aids in debugging and model development

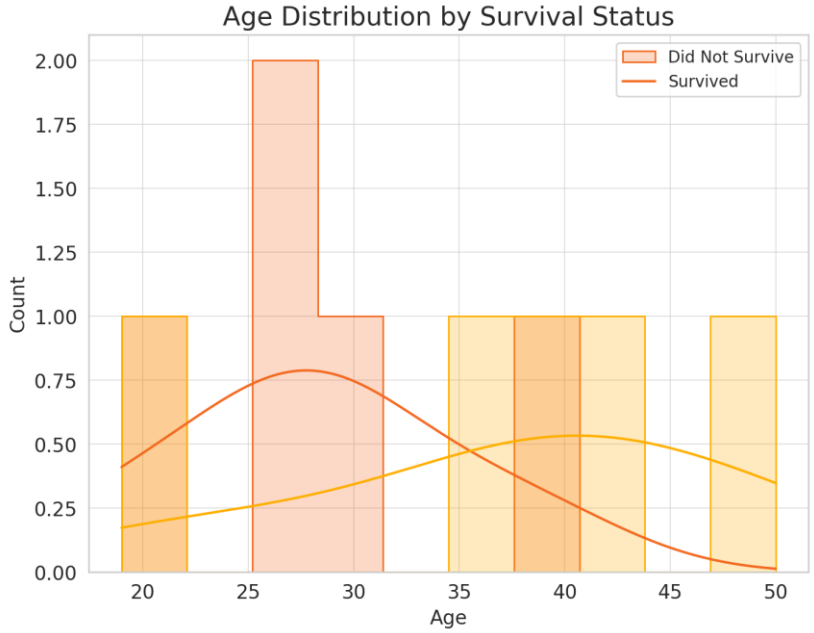
Exploratory Charts

- Visualizes dataset characteristics.



- Common visualizations:
  - Scatter plot matrices for numerical features
  - Bar graphs for categorical features
  - Principal Component Analysis (PCA) plots



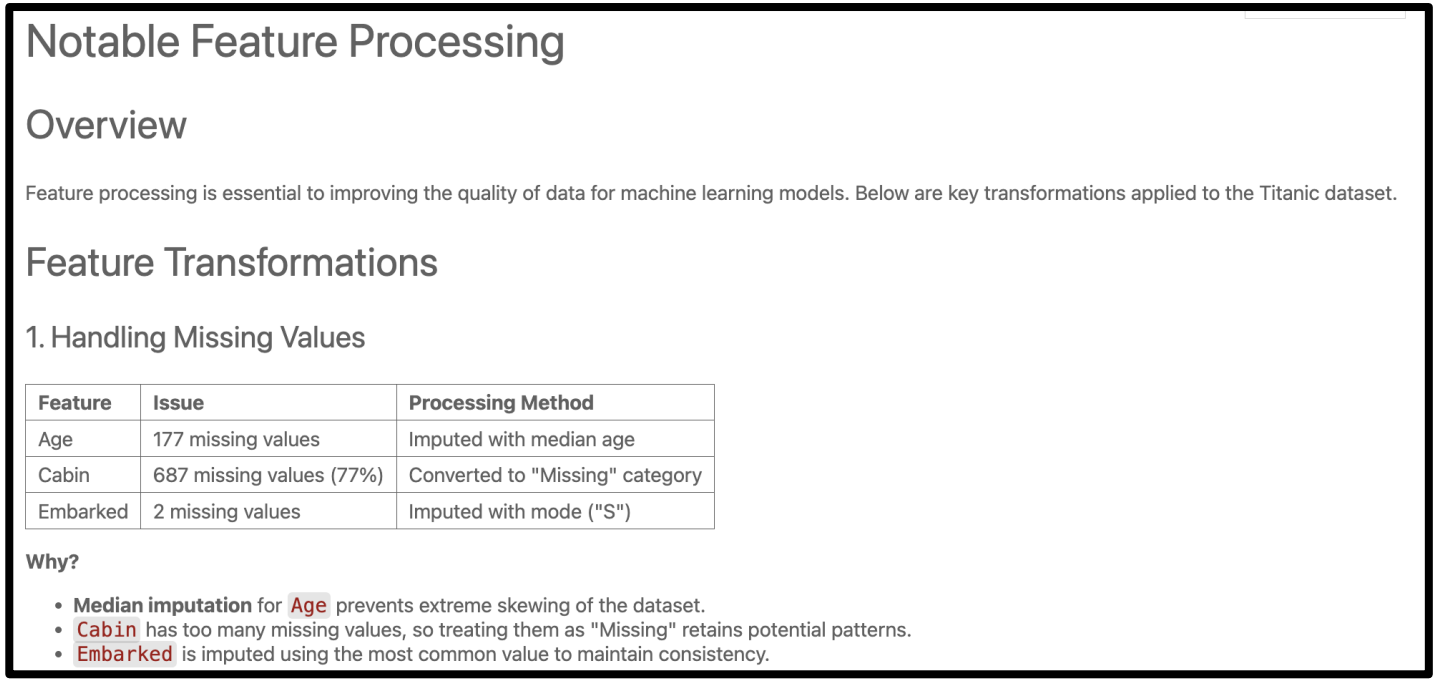


Notable Feature Processing

- Records any transformations applied to features.

- Examples:
  - Feature scaling
  - Feature engineering

- Why it matters:
  - Ensures reproducibility
  - Facilitates model application on future data



Notable Feature Processing

Overview

Feature processing is essential to improving the quality of data for machine learning models. Below are key transformations applied to the Titanic dataset.

Feature Transformations

1. Handling Missing Values

Feature	Issue	Processing Method
Age	177 missing values	Imputed with median age
Cabin	687 missing values (77%)	Converted to "Missing" category
Embarked	2 missing values	Imputed with mode ("S")

Why?

- **Median imputation** for `Age` prevents extreme skewing of the dataset.
- `Cabin` has too many missing values, so treating them as "Missing" retains potential patterns.
- `Embarked` is imputed using the most common value to maintain consistency.

Notes

- Captures additional considerations and dataset nuances.

- Examples:
  - Limitations of the data
  - Special handling requirements
  - Observations during data exploration

Notes

General Considerations

- The **Titanic dataset** is widely used for **classification problems**, specifically **binary survival prediction**.
- It contains **missing values** in some features (`Age`, `Cabin`, `Embarked`), requiring preprocessing.
- **Class imbalance** exists—more **passengers did not survive** than survived.
- Historical **biases** (e.g., social class, gender survival disparities) should be considered when analyzing results.

Data Limitations

Limitation	Description
Missing Data	`Age` (177 values), `Cabin` (687 values), `Embarked` (2 values) are missing and require imputation.
Class Imbalance	About 38% of passengers survived, which may impact model performance.
Historical Bias	Women and children had higher survival rates due to evacuation policies.
Small Dataset	Only 891 training samples, which may limit model generalization.

Ethical Considerations

- This dataset is **historical** and should not be used to **generalize survival predictions** beyond the Titanic event.
- Some features like **gender and socioeconomic class** had a strong influence on survival, which can introduce **bias**.
- While this dataset does **not contain personally identifiable information (PII)**, it still reflects real historical data.