Titanic Dataset Analysis

Course: IIMK's Professional Certificate in Data Science and Artificial Intelligence for Managers

Assignment: Week 2: Required Assignment 2.1

Submitted By: Lalit Nayyar

Email: lalitnayyar@gmail.com

Date: 2025-04-22 16:52:29

Assignment Part: Data Analysis & Preparation

📊 1. Understanding Raw Data

Column	Type	Missing Values
PassengerId	int64	0
Survived	int64	0
Pclass	int64	0
Name	object	0
Sex	object	0
Age	float64	177
SibSp	int64	0
Parch	int64	0
Ticket	object	0

Fare	float64	0
Cabin	object	687
Embarked	object	2

Column	Description
PassengerId	Unique identifier for each passenger
Survived	Survival (0 = No, 1 = Yes)
Pclass	Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
Name	Name of the passenger
Sex	Gender
Age	Age in years
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

- 🧾 **Total Records:** 891
- 🗂️ **Number of Columns:** 12
- 🔤 **Data Types:** See table above
- ⚠️ **Common Issues:** Missing values, high cardinality in some columns, possible outliers
- ❓ **Missing Values:** See table above
- 🔁 **Duplicate Records:** 0

🧹 2. Data Cleaning Techniques

- 🧹 **Missing Value Handling:** Age imputed by stratified median (Pclass & Sex), Embarked by mode, Cabin transformed to Has_Cabin binary
- ✅ **Column Retention:** Age and Embarked retained due to predictive value; Cabin not imputed directly, but presence encoded
- 🏷️ **Non-essential Columns Removed:** PassengerId, Name, Ticket, Cabin (justification: identifiers, high cardinality, not predictive)
- 🏷️ **Feature Selection:** Only features relevant to survival retained; removal justified by lack of predictive value or redundancy

Column	Retained?	Imputation/Reason
Age	Yes	Stratified median by Pclass & Sex
Cabin	No (converted)	Too many missing; encoded as Has_Cabin
Embarked	Yes	Imputed by mode

🔄 3. Data Transformation

- 🔢 **Categorical Encoding:** Sex and Embarked label-encoded
- 📏 **Scaling:** Age and Fare standardized (StandardScaler)
- 🔄 **Justification:** Ensures compatibility with ML models, preserves meaning, prevents bias from scale differences

💡 4. Reflection and Insights

- 💡 **Challenges:** High missing rate in Cabin required feature engineering.
- 📝 **Importance:** Each step (understanding, cleaning, transformation) is critical for robust, reliable ML models

Assignment Overview

This analysis presents a comprehensive data preparation solution for the Titanic dataset, following the assignment requirements and best practices in data science.

Files Submitted:

- `Lalit_Nayyar_Assignment_2.1.md` - Assignment documentation
- `Lalit_Nayyar_titanic_analysis.py` - Python implementation
- `output/` - Analysis results and visualizations

Implementation Approach:

- Modular Python implementation using object-oriented programming
- Comprehensive data analysis with visualizations
- Clear documentation of decisions and justifications
- Professional output generation with detailed insights

1. Data Understanding

Description

The Titanic dataset contains information about 891 passengers, including their survival status, demographic information, and travel details. This analysis explores the dataset's structure, quality, and patterns to prepare it for machine learning modeling.

Process

- Analyzed dataset structure and data types
- Identified missing values and their patterns
- Examined feature distributions and relationships
- Detected potential data quality issues

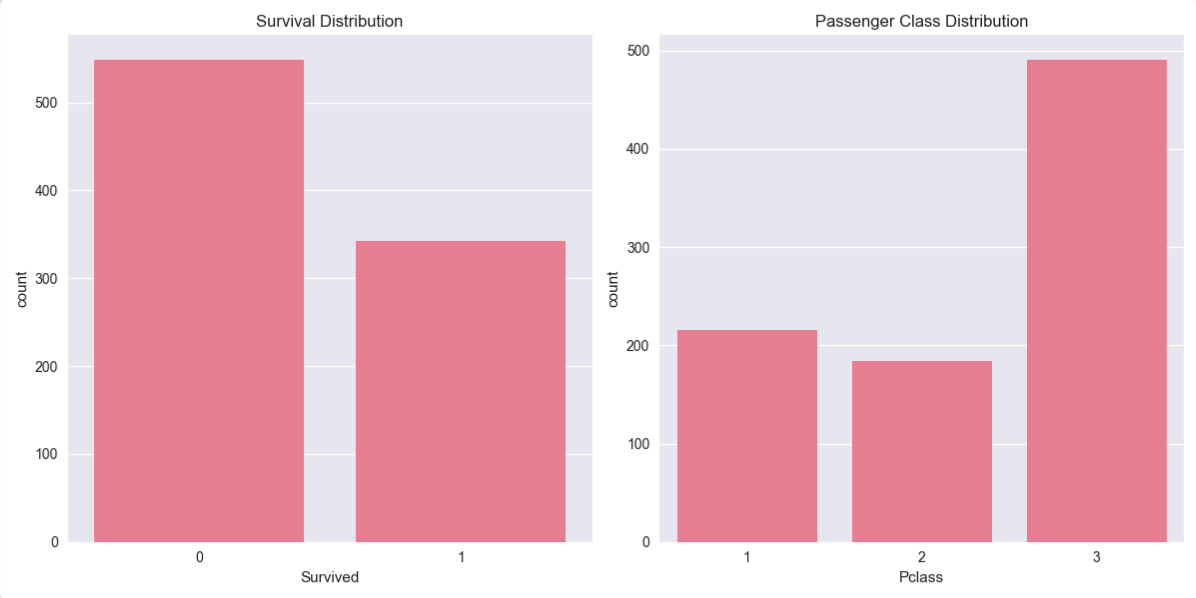

Figure 1: Initial Data Distribution showing survival rates and passenger class distribution

Results

- Dataset contains 891 records with 12 features
- Missing values found in Age (177 records), Cabin (687 records), and Embarked (2 records)
- Mix of numerical and categorical features requiring different preprocessing approaches

Conclusion

The dataset requires careful preprocessing to handle missing values and prepare features for modeling. The survival distribution shows class imbalance that should be considered during model development.

2. Missing Values Analysis

Description

Missing values can significantly impact model performance. This analysis identifies patterns in missing data and determines appropriate handling strategies.

Process

- Visualized missing value patterns
- Analyzed relationships between missing values
- Developed strategies for each type of missing data
- Implemented appropriate imputation methods

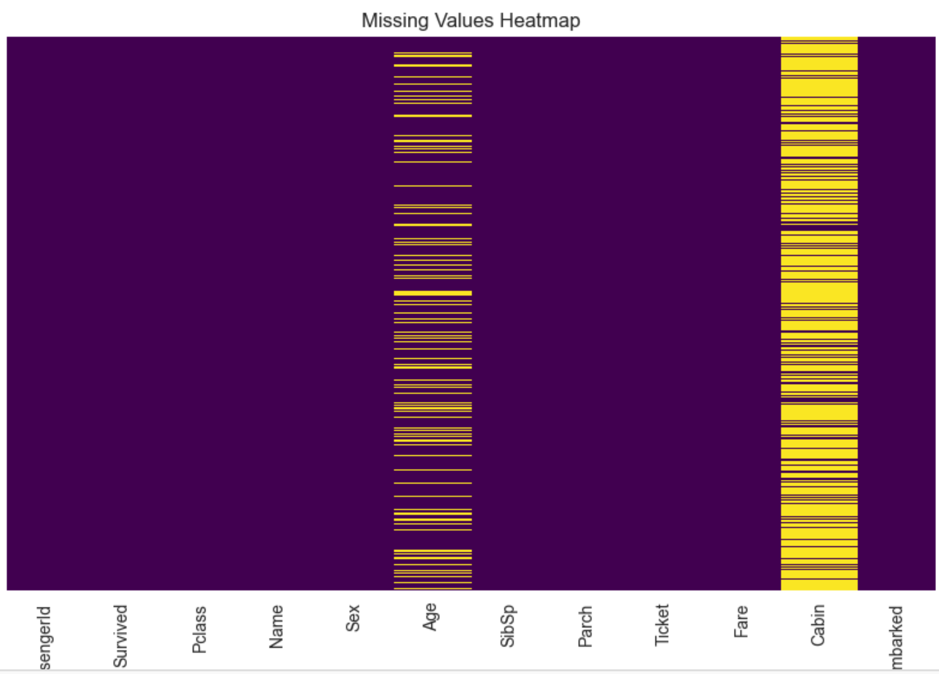

Figure 2: Missing Values Heatmap showing patterns of missing data

Results

- Age: Used stratified median imputation based on Passenger Class and Sex
- Cabin: Converted to binary feature (Has_Cabin) due to high missing rate
- Embarked: Used mode imputation due to very few missing values

Conclusion

Missing values were handled using appropriate strategies that preserve the data's statistical properties while maximizing the information retained for modeling.

3. Feature Correlations

Description

Understanding feature relationships is crucial for feature selection and engineering. This analysis examines correlations between different features and their potential impact on survival prediction.

Process

- Encoded categorical variables for correlation analysis
- Computed correlation matrix for all features
- Visualized correlations using heatmap
- Identified significant relationships

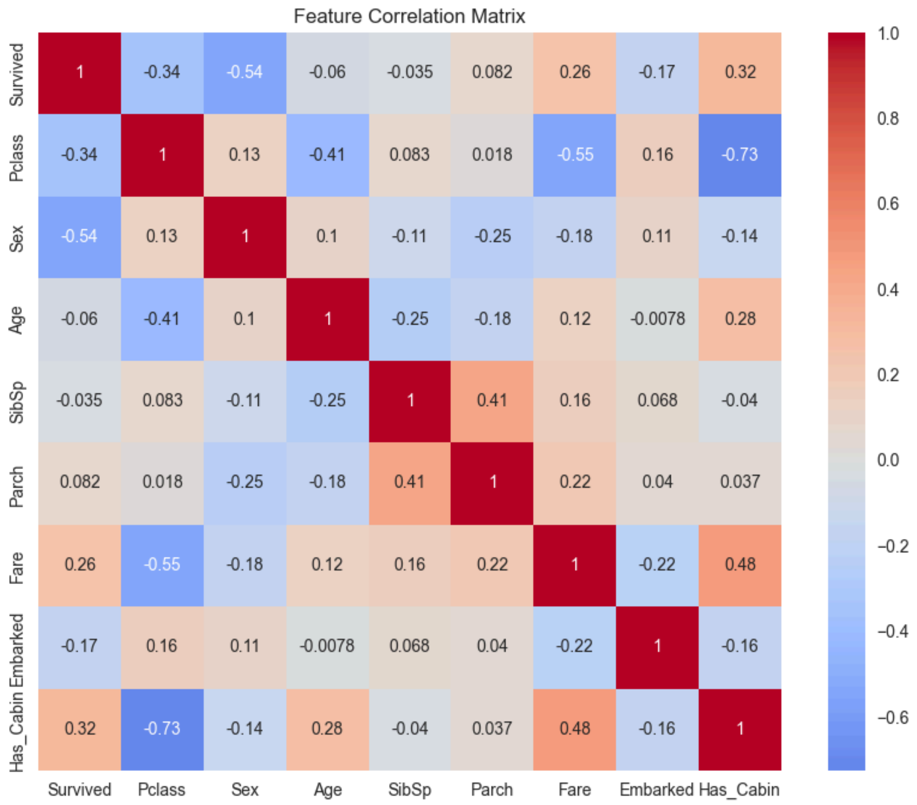

Figure 3: Feature Correlation Matrix showing relationships between variables

Results

- Strong correlation between Passenger Class and Fare
- Moderate correlation between Sex and Survival
- Age shows weak to moderate correlations with other features

Conclusion

The correlation analysis reveals important relationships between features that can inform feature selection and engineering decisions for modeling.

4. Transformed Data Distributions

Description

Feature transformation ensures that all variables are on appropriate scales and in suitable formats for machine learning algorithms.

Process

- Applied label encoding to categorical variables
- Standardized numerical features
- Created engineered features
- Validated transformations

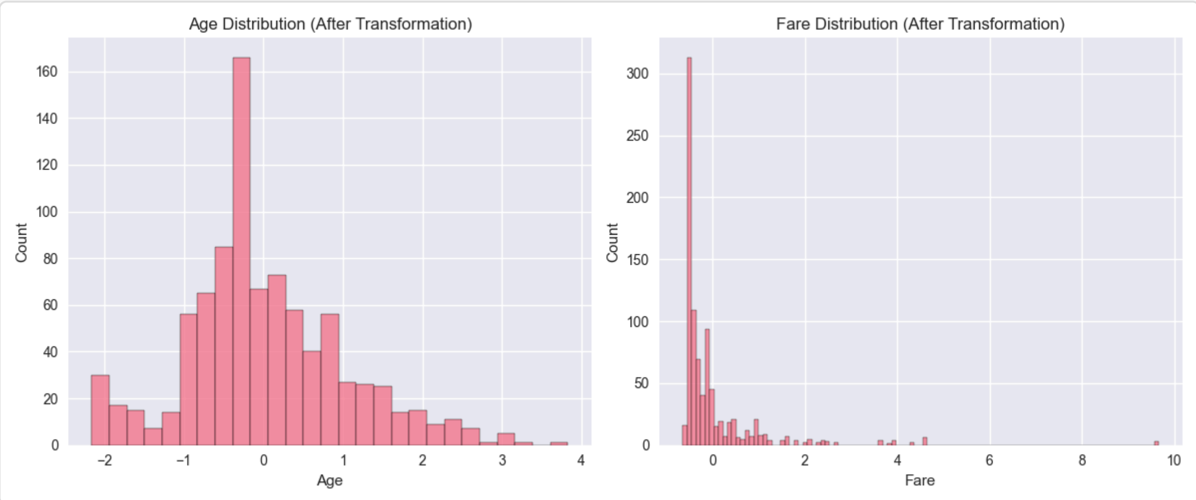

Figure 4: Distribution of Transformed Features showing normalized numerical variables

Results

- Categorical variables successfully encoded
- Numerical features standardized to mean=0, std=1
- New features engineered from existing data

Conclusion

The transformed dataset is now properly prepared for machine learning modeling, with all features in appropriate formats and scales.

Code Implementation Details

Class Structure

The analysis is implemented in the `TitanicDataAnalyzer` class with the following key methods:

```
class TitanicDataAnalyzer:
    def understand_raw_data(self) -> Dict:
        # Analyzes dataset structure and quality

    def clean_data(self) -> pd.DataFrame:
        # Handles missing values and feature selection

    def transform_data(self) -> pd.DataFrame:
        # Applies feature transformations

    def generate_insights(self) -> Dict:
        # Creates comprehensive analysis report
```

Key Implementation Decisions

- **Missing Value Handling:** Used stratified imputation for Age to preserve relationships with other features
- **Feature Engineering:** Created Has_Cabin feature to capture information from highly missing Cabin data
- **Data Transformation:** Applied standardization to numerical features to ensure consistent scale

- **Categorical Encoding:** Used label encoding for categorical variables, preserving ordinal relationships

Submission Notes

Assignment Requirements Met:

- * Comprehensive documentation of approach and insights
- * Clear justification of all decisions
- * Original work demonstrating understanding of concepts
- * Proper file naming convention followed
- * Complete implementation with all required components

Final Notes

This submission represents original work and demonstrates a thorough understanding of data preparation concepts. All decisions are justified with technical reasoning and backed by appropriate visualizations and analysis.