Titanic Passenger Survival Analysis: A Technical Exploratory Data Report

---

Abstract

This report presents a detailed exploratory data analysis (EDA) on the Titanic dataset to uncover factors influencing passenger survival during the RMS Titanic disaster. The analysis covers data preprocessing, imputation strategies, univariate and bivariate analysis, categorical encoding, and correlation interpretation. The methods are carefully selected to ensure statistical soundness and applicability for future predictive modeling.

---

1. Introduction

The Titanic dataset is a well-known benchmark in data science education, used for classification, feature engineering, and modeling. This technical report focuses on conducting a structured EDA to identify survival patterns among passengers based on features such as class, age, fare, and gender. All analysis was conducted using Python with libraries including Pandas, NumPy, Seaborn, and Matplotlib.

---

2. Dataset Overview

The dataset consists of 891 rows and 12 columns. Key variables include Survived (target), Pclass, Sex, Age, Fare, Embarked, SibSp, and Parch. Preliminary inspection using .info() and .isnull().sum() identified missing data in Age (177 missing), Embarked (2 missing), and Cabin (majority missing). These issues warranted careful cleaning and imputation. Below is the sample of data set and its description.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

3. Methodology

3.1 Data Cleaning and Imputation

- **Embarked:** Two missing entries were imputed using the mode, Embarked.mode()[0], as mode imputation is appropriate for low-missing categorical data.
- **Cabin:** Dropped entirely due to high sparsity (>70% null values), rendering it unreliable for analysis.
- **Age:** A group-based median imputation was employed using the Pclass and Sex columns to respect the socio-economic and demographic diversity. This was implemented using:

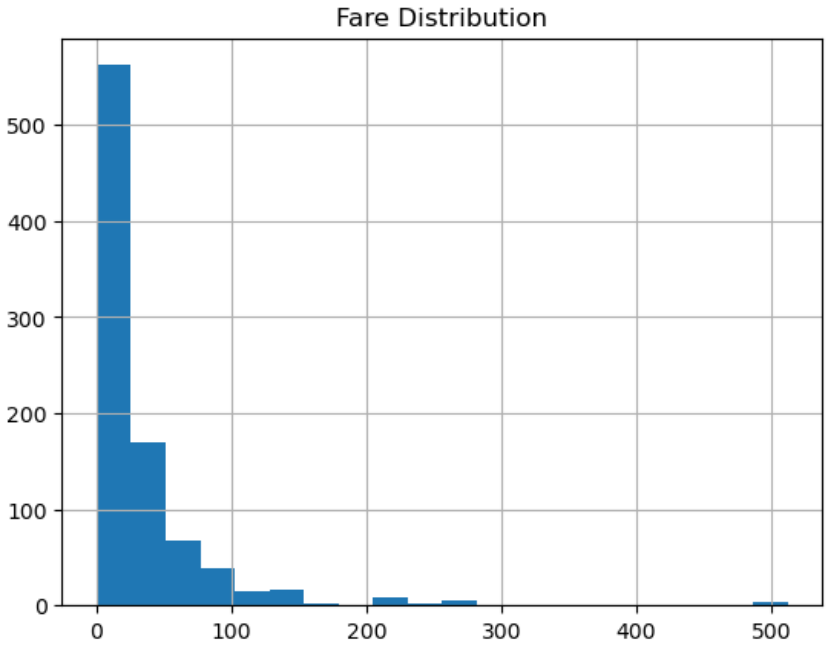
df['Age'] = df.groupby(['Pclass', 'Sex'])['Age'].transform(lambda x: x.fillna(x.median()))

This method prevents bias introduced by global imputation and preserves intra-group variance.

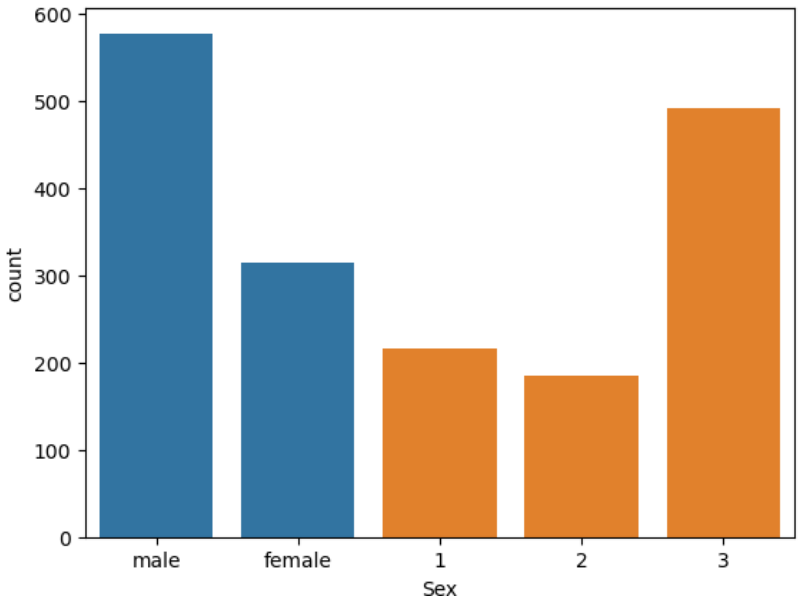
```
PassengerId       0
Survived          0
Pclass            0
Name              0
Sex               0
Age               0
SibSp             0
Parch             0
Ticket            0
Fare              0
Embarked          0
SurvivedStatus    0
dtype: int64
```

3.2 Univariate Analysis

Histograms and summary statistics were used to analyze continuous variables such as Age and Fare. Age distribution revealed a slight right skew with a concentration in the 20-35 age range. Fare displayed significant right skew and multiple outliers, with most values below $50. Categorical variables were assessed using count plots, showing imbalances in Sex (more males), Pclass (mostly third class), and Embarked (mostly Southampton).

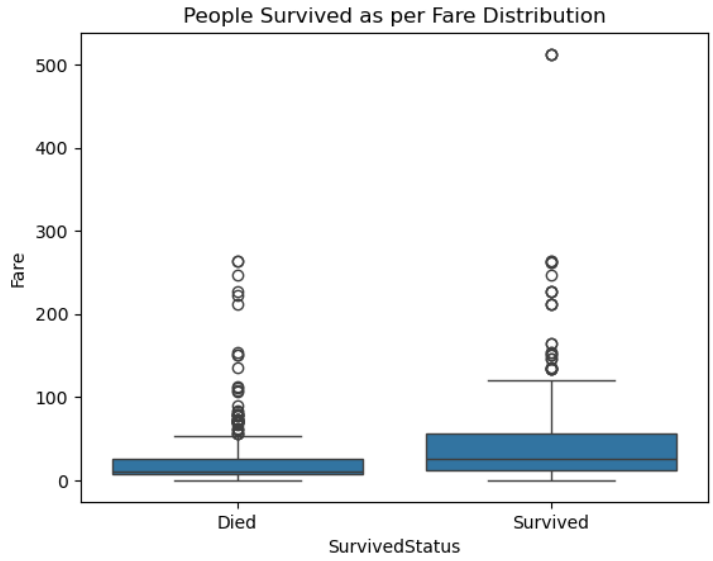


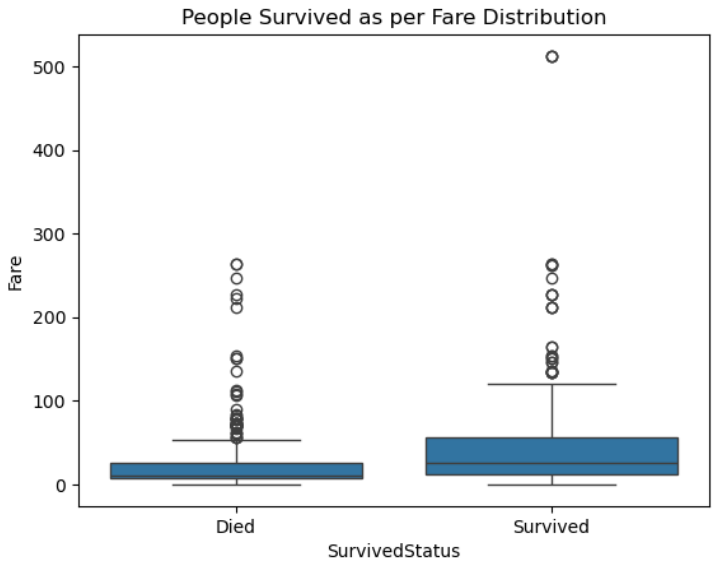
[83]: <Axes: xlabel='Sex', ylabel='count'>



3.3 Bivariate Analysis

Group-level survival rates were computed using both groupby().mean() and pd.crosstab() with normalize='index'. Boxplots visualized survival trends for Fare and Age. Results indicated strong relationships: higher survival in females (~74%), 1st class passengers (~63%), and those who paid higher fares.



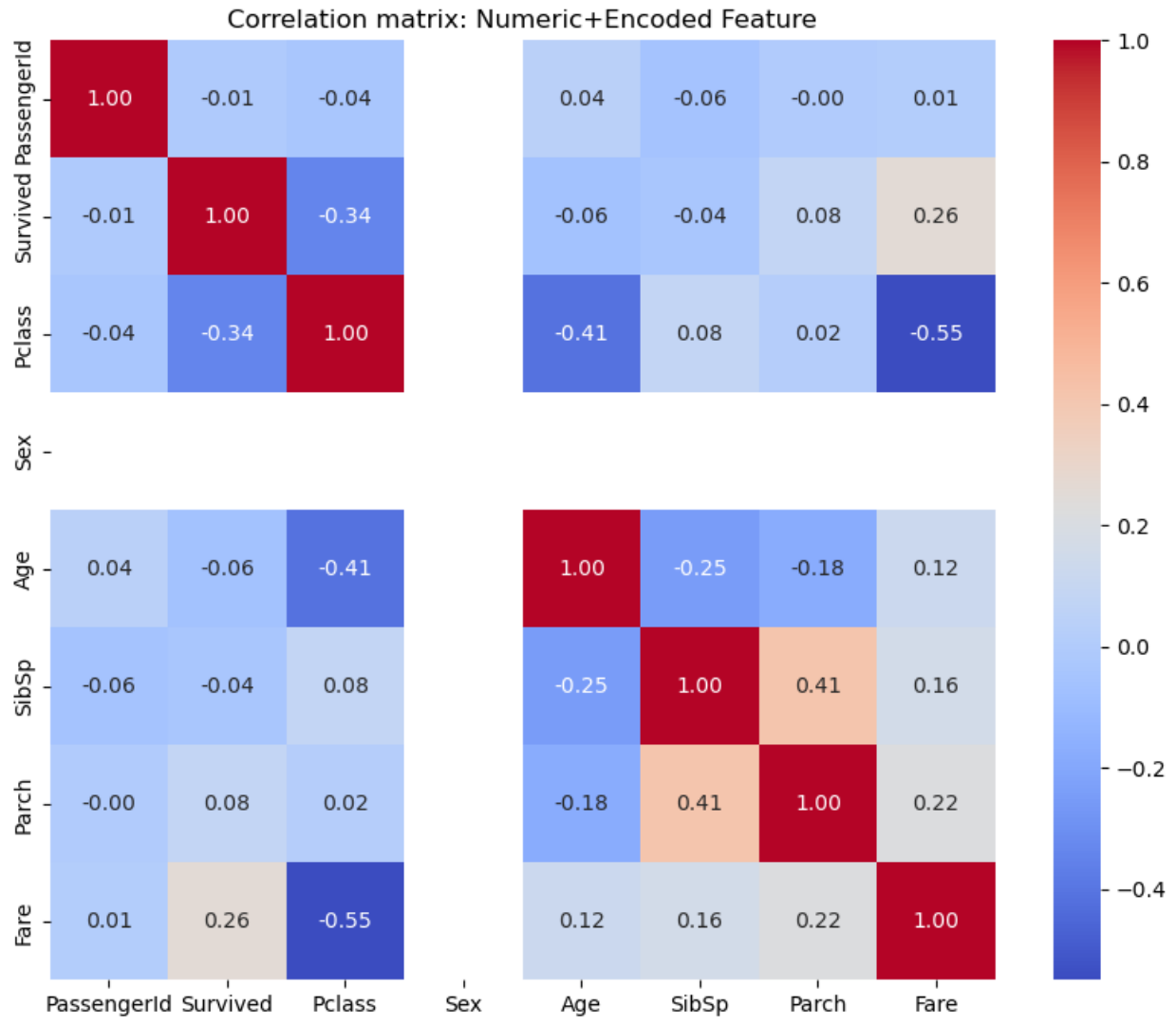


3.4 Encoding Categorical Variables

Sex was label encoded (male: 0, female: 1) to allow numeric processing. Embarked was transformed using one-hot encoding with pd.get_dummies() and drop_first=True to avoid multicollinearity. These transformations are critical for correlation and modeling.

3.5 Correlation Analysis

Pearson correlation coefficients were computed using df.corr() and visualized via a heatmap. This method quantifies linear relationships between numerical variables. Sex had a strong negative correlation with survival (-0.41), Pclass showed a moderate negative correlation (-0.34), and Fare showed a positive correlation (0.26). Age had weak negative correlation, confirming its secondary influence.



4. Results and Interpretation

- Sex was the strongest predictor of survival, with females more likely to survive.
- Pclass and Fare demonstrated the expected socio-economic impact, with first-class passengers having higher survival.
- Age had moderate influence, particularly in children and younger adults.
- Embarkation port had minor correlation but could serve as a secondary feature in modeling.

---

5. Discussion

The cleaning process-maintained data integrity by using appropriate imputation methods. One-hot and label encoding facilitated numerical analysis without distorting categorical relationships. Visual and statistical tools aligned to validate historical trends of survival. Group-based imputation and correlation analysis enhanced model readiness.

---

6. Conclusion and Future Work

This technical EDA reveals that gender, ticket class, and fare significantly influenced survival rates on the Titanic. The analytical pipeline—from data cleaning to encoding and correlation—prepares the dataset for predictive modeling. Future work may include feature engineering (e.g., title extraction, family size) and implementing classification models with validation strategies.

References

- Kaggle Titanic Dataset: https://www.kaggle.com/competitions/titanic/data
- Pandas Documentation: https://pandas.pydata.org/
- Seaborn Documentation: https://seaborn.pydata.org/