Titanic Dataset Analysis Report

Author: Jassmesh Singh Kochhar
Date: 28 April 2025

1. Introduction

This report summarizes insights from an exploratory analysis of the Titanic dataset, which contains information about passengers aboard the RMS Titanic, including survival status, demographics, and travel details. The goal is to uncover patterns and prepare the data for further analysis.

2. Dataset Overview

Size: 891 rows × 12 columns

Variables:

PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. '	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunderco	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, M	female	55	0	0	248706	16		S
18	17	0	3	Rice, Mast	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, M	male		0	0	244373	13		S
20	19	0	3	Vander Pla	female	31	1	0	345763	18		S
21	20	1	3	Masselmai	female		0	0	2649	7.225		C
22	21	0	2	Fynney, Mr	male	35	0	0	239865	26		S
23	22	1	2	Beesley, M	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan	female	15	0	0	330923	8.0292		Q
25	24	1	1	Sloper, Mr	male	28	0	0	113788	35.5	A6	S
26	25	0	3	Palsson, M	female	8	3	1	349909	21.075		S
27	26	1	3	Asplund, M	female	38	1	5	347077	31.3875		S
28	27	0	3	Emir, Mr. F	male		0	0	2631	7.225		C
29	28	0	1	Fortune, M	male	19	3	2	19950	263	C23 C25 C	S

Key Observations:

Missing Values:

- Age: 177 missing
- Cabin: 687 missing
- Embarked: 2 missing

3. Data Cleaning

Handling Missing Values

- **Age**: Filled missing values with the median age (28 years).
- **Embarked**: Replaced 2 missing values with the mode ("S").
- **Cabin**: Dropped the column due to excessive missing data (77% missing).

Result

No missing values remained after cleaning.

4. Key Findings

A. Categorical Variable Distribution

Sex:

- Male: 577 (64.8%)
- Female: 314 (35.2%)

Survival:

- Survived: 342 (38.4%)
- Did not survive: 549 (61.6%)

Passenger Class (Pclass):

- 3: 491
- 1: 216
- 2: 184

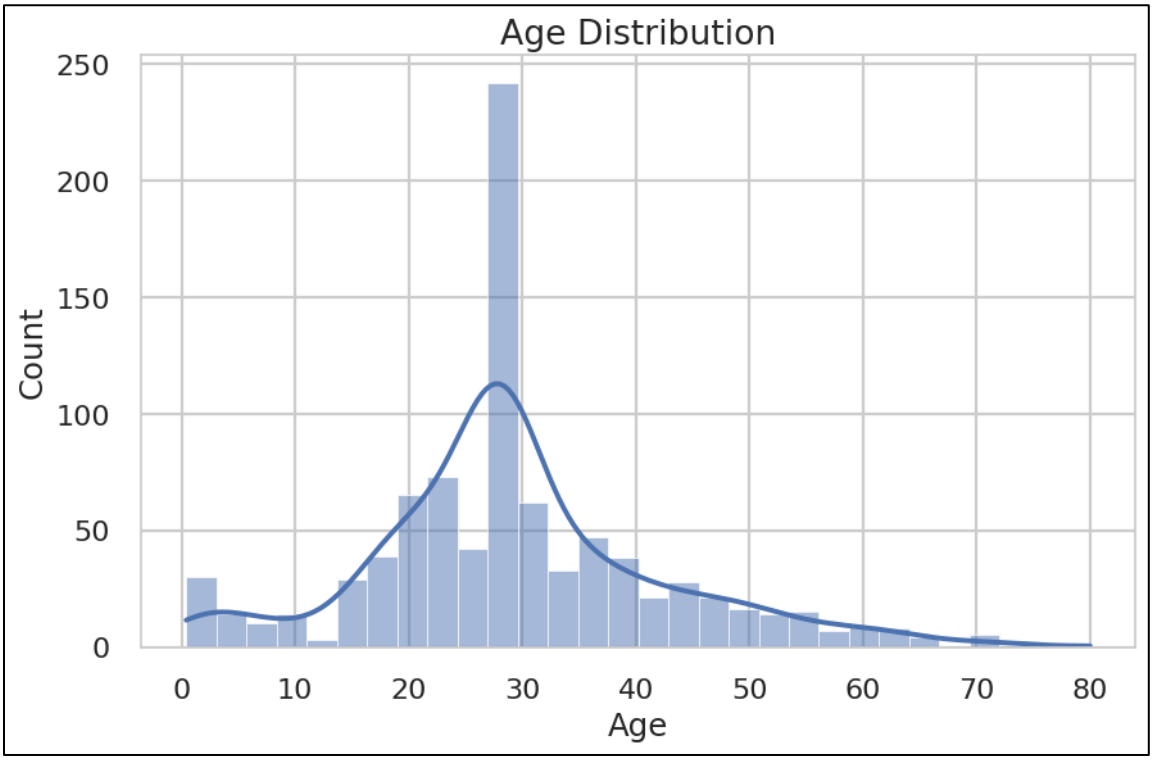
Embarkation Port:

- Southampton (S): 646 (72.7%)
- Cherbourg (C): 168 (18.9%)
- Queenstown (Q): 77 (8.7%)

B. Univariate Analysis

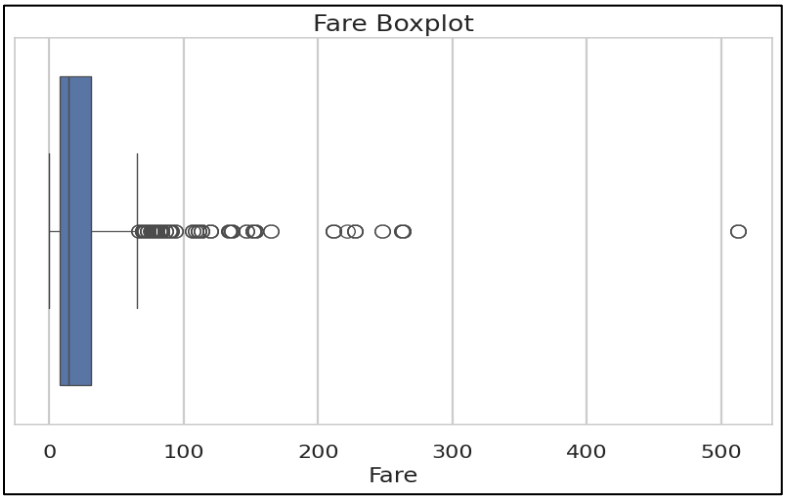
1. Age Distribution:

- Majority of passengers were between 20–40 years old.
- Median age: 28 years.



2. Fare Analysis:

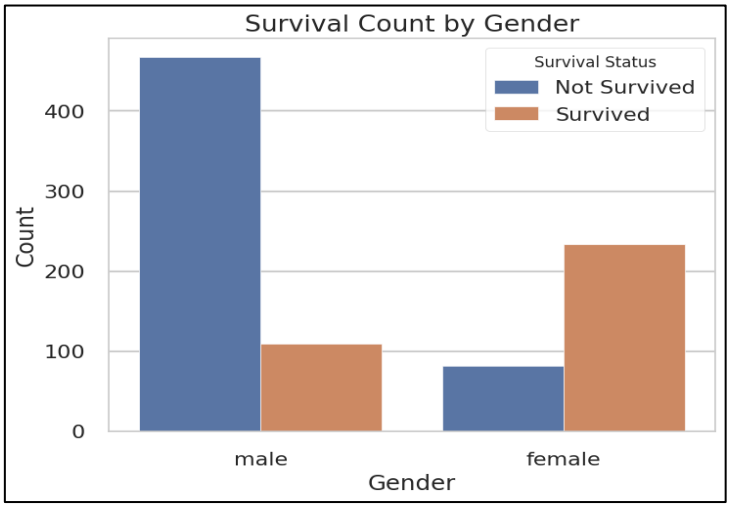
- Median fare: £14.45.
- Significant outliers (e.g., £512 maximum fare).



C. Bivariate Analysis

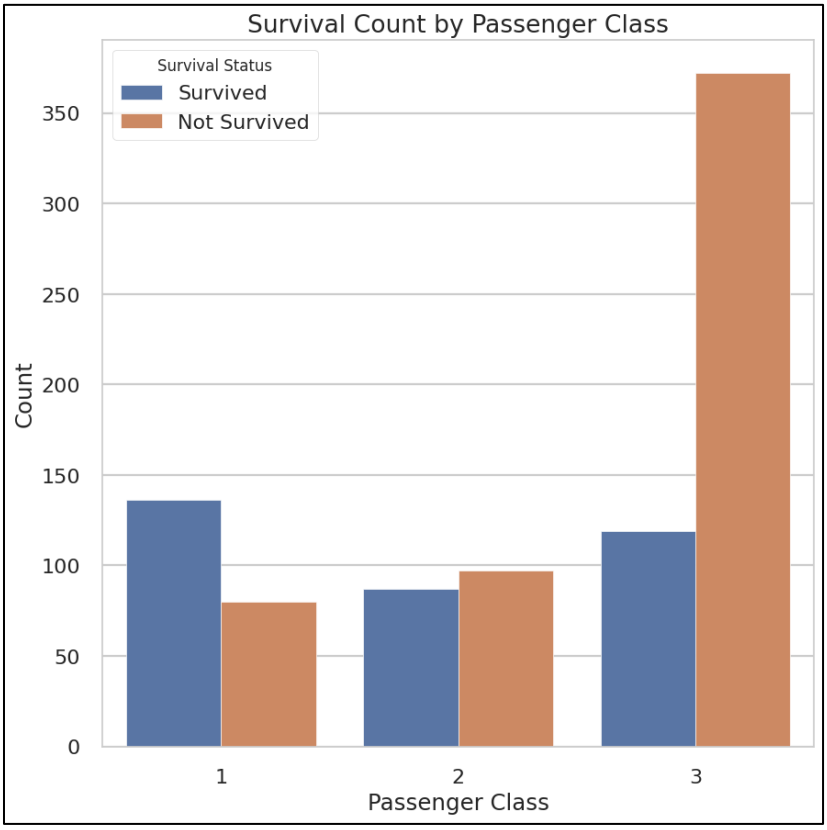
1. Survival Count by Gender

- Majority of Not survived were Male and Majority of Survived were Women.



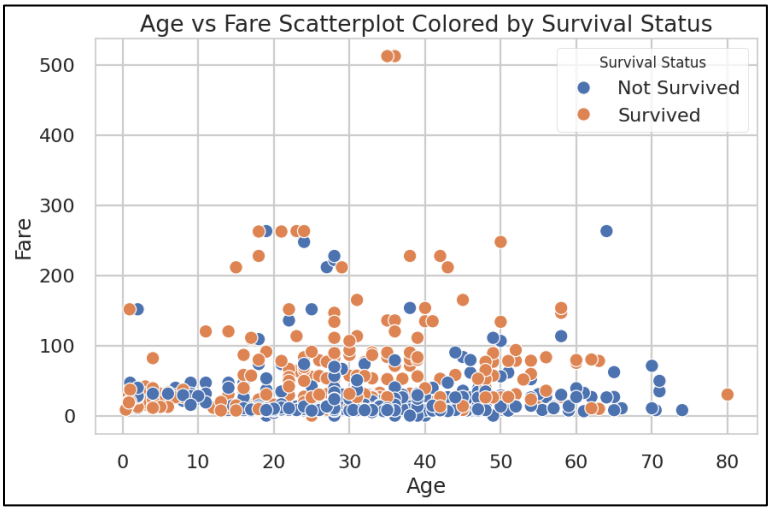
2. Survival Count by Passenger Class

- Majority of Not survived were from 3rd Passenger Class.



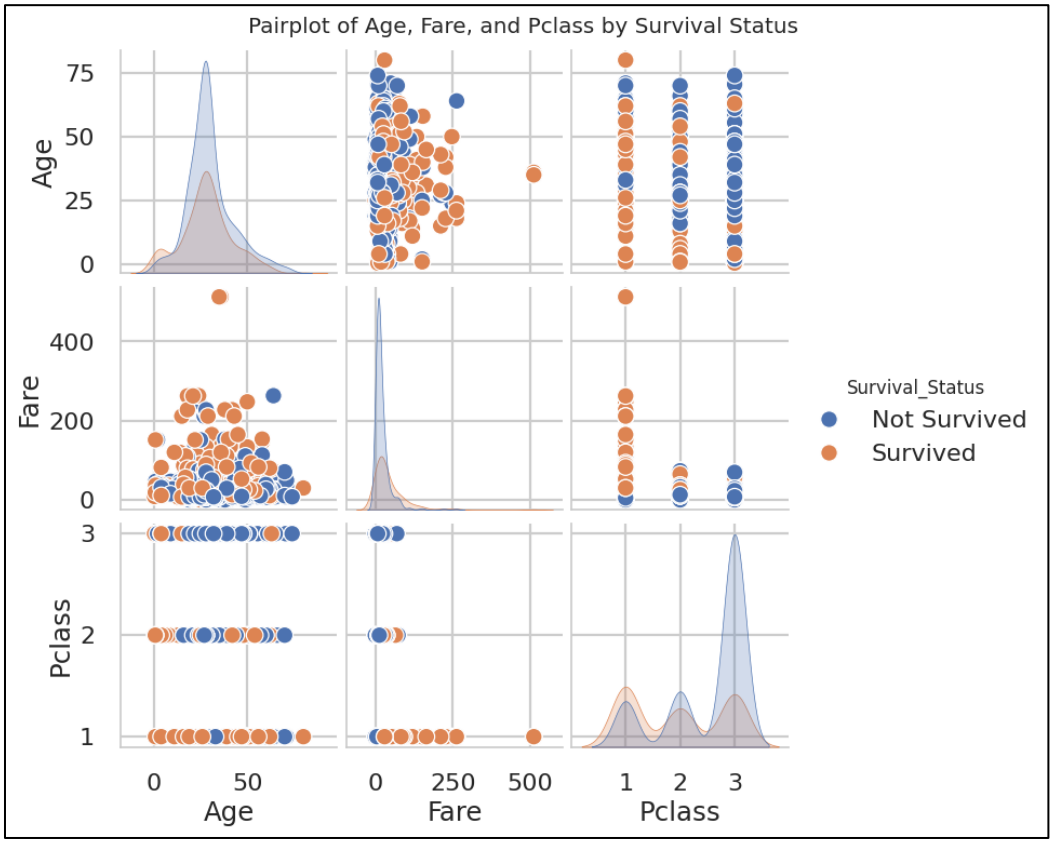
3. Survival Plot by Age Vs Fare

- Lower fare passengers had reduced survival rates compared to those who paid higher fares.



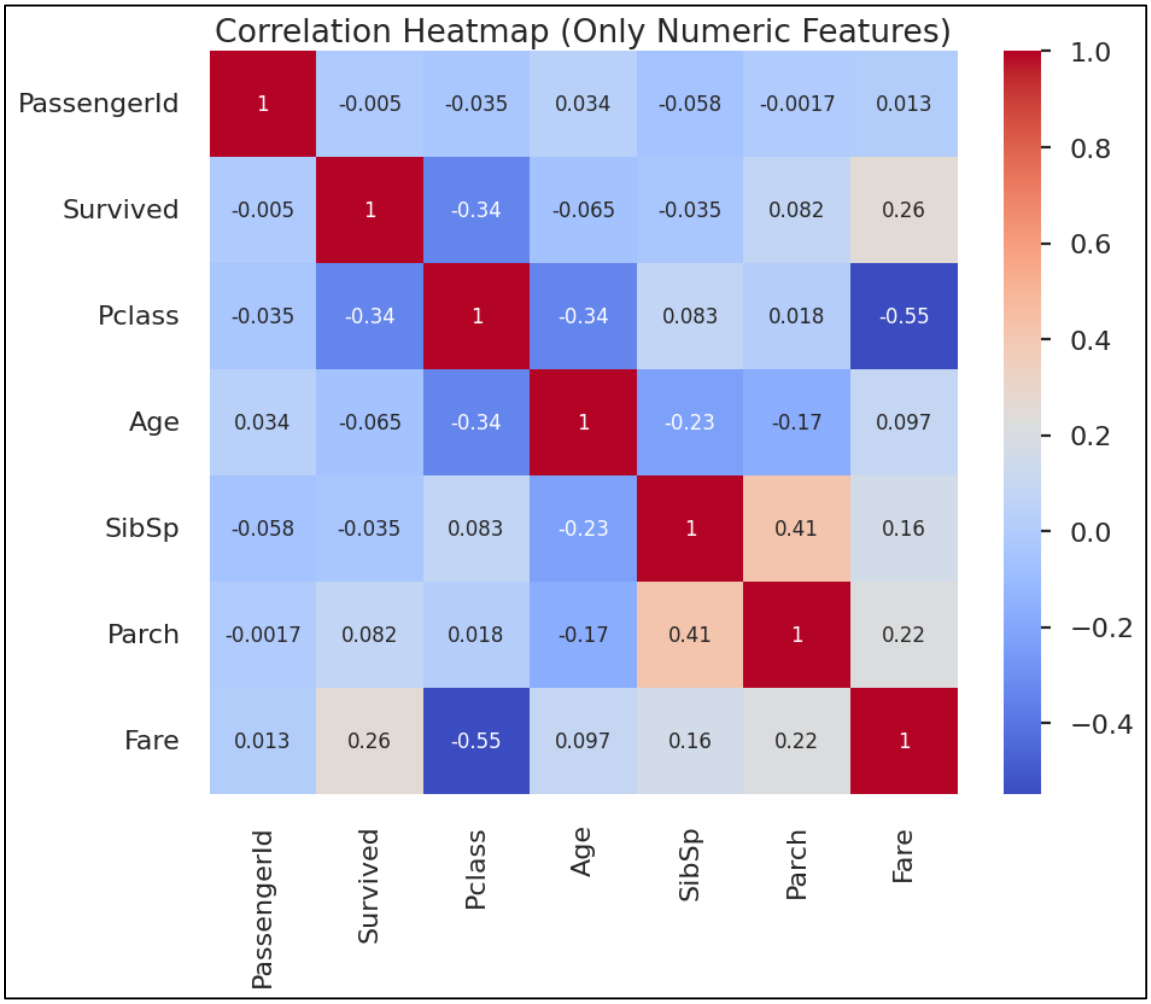
D. Multivariate Analysis

- Showed relationships across multiple features(Age, Fare, Pclass, By Survival status) together.



E. Correlation Matrix and Heatmap

- Find which features(only numeric ones) are correlated.



F. Detect Skewness

- Fare is highly skewed so applied transformation like fare_log.

Summary of Findings:

- Most passengers were aged between 20–40.
- Females had a higher survival rate compared to males.
- 1st class passengers had much better chances of survival.
- Fare distribution is highly skewed; many passengers paid low fare, few paid very high fare.
- Pclass is negatively correlated with survival (higher class → better chance).
- Missing values were handled properly (Age filled with median, Embarked with mode).
- Cabin feature was dropped due to excessive missing values.