

Titanic Survival Analysis

Exploratory Data Analysis & Interactive Dashboarding

By – Sanjay Dutt Mishra
Date: January 19, 2026

1. SUMMARY

Performed exploratory data analysis on the Titanic dataset using Python and built a modern executive-style Power BI dashboard to communicate survival insights.

This project demonstrates an end-to-end data analytics workflow, moving from raw data processing to insight generation. Key highlights include:

- Cleaned and pre-processed 891 passenger records, using statistical imputation for missing age values.
- Identified gender and passenger class as the primary drivers of survival, with females surviving at a significantly higher rate.
- Designed a single-page executive Power BI dashboard to visualize key metrics, overcoming aggregation challenges with custom DAX measures.

2. PROJECT BACKGROUND & OBJECTIVES

The Titanic disaster is a historical maritime tragedy with a rich dataset available for analysis. The objective of this project was not predictive modeling, but rather **Exploratory Data Analysis (EDA)** and **Visual Storytelling**. The goal was to transform raw passenger data into an accessible, interactive format for non-technical stakeholders.

Tools & Technologies:

- **Python (Pandas, NumPy, Matplotlib, Seaborn):** Data cleaning, manipulation, and initial statistical analysis.
- **Power BI Desktop:** Data modeling, DAX measure creation, and interactive dashboard design.

3. DATASET OVERVIEW

The analysis is based on the standard Titanic passenger dataset. The raw data structure is as follows:

Total Records	Total Columns	Target Variable
891 rows	12 columns	Survived (0 = No, 1 = Yes)

Missing Values identified during initial inspection:

- **Age:** 177 missing values (~20% of data).
- **Cabin:** 687 missing values (~77% of data).
- **Embarked:** 2 missing values.

4. DATA CLEANING (PYTHON)

Data quality is critical for accurate reporting. The following cleaning steps were executed in Python:

1. **Irrelevant Features:** Dropped `PassengerId` and `Ticket` as they provided no analytical value.
2. **Age Imputation:** Missing ages were filled using the median age grouped by *Passenger Class* and *Sex* to maintain statistical distribution.
3. **Embarked Imputation:** Filled missing port values with the mode (most common value).
4. **Feature Engineering:** The `Cabin` column had too many missing values to be used directly. It was converted into a binary feature `Has_Cabin` (1 if cabin known, 0 otherwise), and the original column was dropped.

```
# Python Data Cleaning Pseudo-code df.drop(columns=["PassengerId", "Ticket"], inplace=True) # Impute Age
using
Median by Group df['Age'] = df.groupby(['Pclass', 'Sex'])['Age'].transform( lambda x: x.fillna(x.median()) ) #
Feature Engineering Cabin df['Has_Cabin'] = df['Cabin'].apply(lambda x: 0 if pd.isna(x) else 1)
df.drop(columns='Cabin', inplace=True) df.to_csv("titanic_cleaned.csv", index=False)
```

5. EXPLORATORY DATA ANALYSIS (EDA) FINDINGS

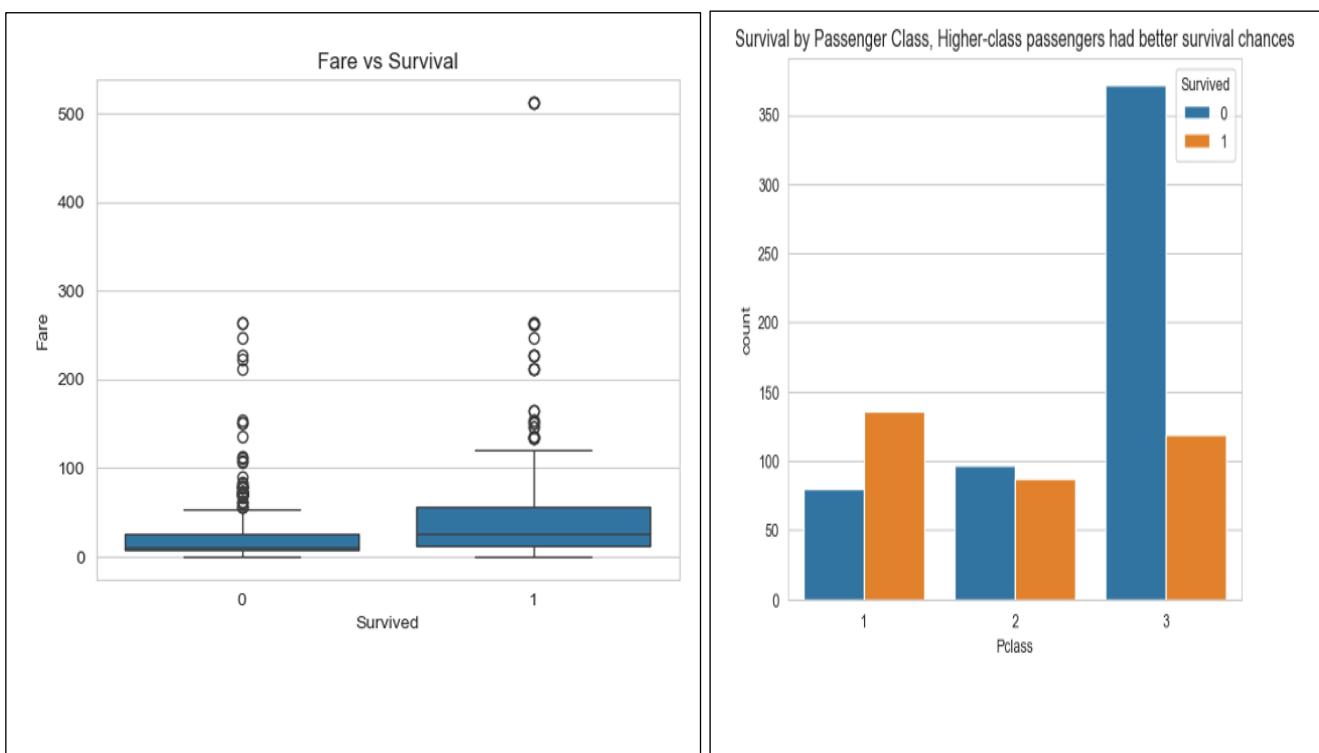
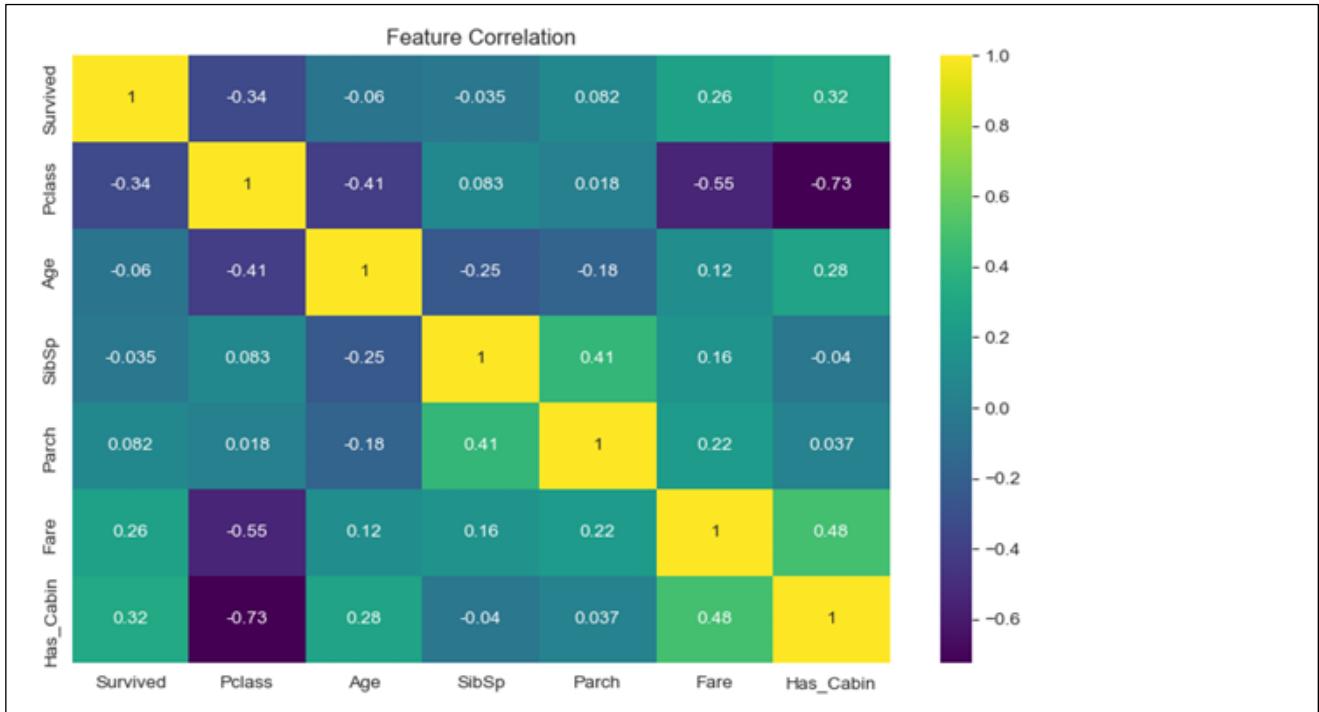
Before building the dashboard, statistical analysis in Python revealed several key patterns:

- **Overall Survival:** Only 342 passengers survived (38.4%), while 549 died.

- **Gender Dynamics:** Female passengers had a significantly higher survival probability compared to males.
- **Class Hierarchy:** First-class passengers had the highest survival rate, indicating socio-economic status played a major role.
- **Fare Distribution:** Survivors generally paid higher fares. The average fare for survivors was notably higher than for non-survivors.
- **Cabin Availability:** Passengers with recorded cabin numbers (often First Class) had better outcomes.

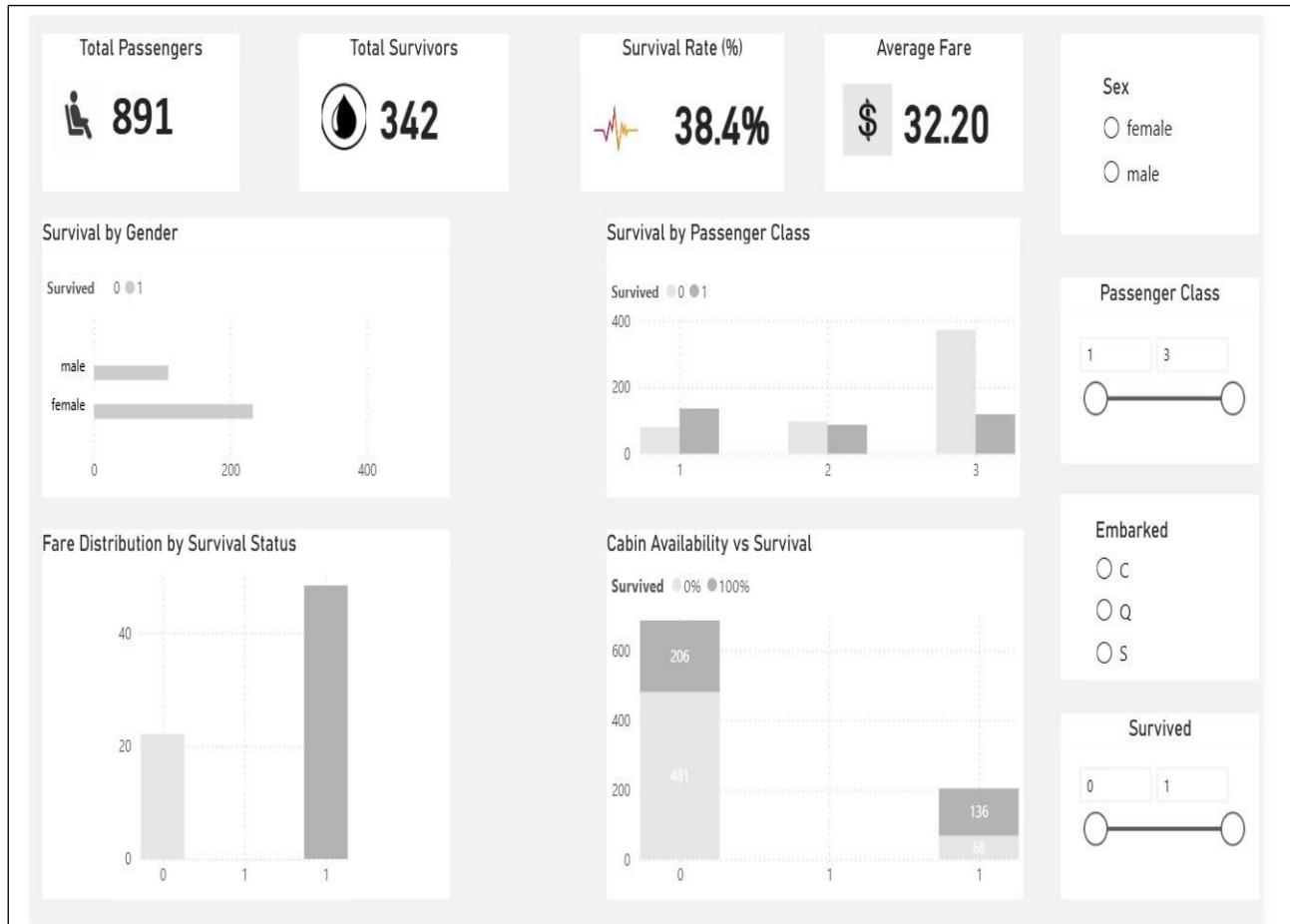
Correlation Heatmap (Numeric Features)

A correlation heatmap was used to understand relationships among numeric features (e.g., *Survived*, *Pclass*, *Age*, *SibSp*, *Parch*, *Fare*, and *Has_Cabin*). This helps identify features that move together and supports interpretation of survival drivers.



6. POWER BI DASHBOARD

The cleaned dataset was imported into Power BI to create an interactive executive view. The design focuses on clarity, utilizing a clean layout with high-level KPIs at the top and detailed breakdown visuals below.



Key EDA Insights

- Female passengers had significantly higher survival rates
- First-class passengers were more likely to survive
- Higher fare and cabin availability positively influenced survival
- Socio-economic status played a major role in survival outcomes

Key Performance Indicators (KPIs)

The dashboard highlights four headline metrics designed to give an instant overview

891	342	38.4%	\$32.20
TOTAL PASSENGERS	TOTAL SURVIVORS	SURVIVAL RATE	AVG FARE

Visualizations & Slicers

- **Survival by Gender:** A horizontal bar chart clearly comparing male vs. female survival counts.
- **Survival by Passenger Class:** Clustered column chart showing survival rates across classes 1, 2, and 3.
- **Fare Distribution:** Compares the average fare paid by survivors vs. non-survivors.
- **Slicers:** Interactive filters for *Sex*, *Passenger Class*, and *Embarked* allow users to drill down into specific segments (e.g., "Survival rate of females in 3rd class").

DAX Measures

To ensure accuracy, explicit DAX measures were created instead of using implicit sums. Examples include:

```
Total Passengers = COUNT(Titanic[PassengerId])
Total Survivors = CALCULATE(COUNTROWS(Titanic),
    Titanic[Survived]
)
= 1) Survival Rate % = DIVIDE([Total Survivors], [Total Passengers], 0)
```

7. CHALLENGES & SOLUTIONS

Challenge: Aggregation Issues. Power BI initially defaulted to summing numerical fields like "Passenger Class" (e.g., summing 1+2+3).

Solution: Changed summarization settings to "Don't Summarize" for categorical numerical fields and created explicit DAX measures for counts.

Challenge: Slicer Interactions. Slicers initially filtered each other, making it difficult to see the full context options.

Solution: Edited interactions to ensure a smooth user experience where context remains clear.

8. CONCLUSION

This project successfully translated raw historical data into actionable insights. By leveraging Python for robust data cleaning and Power BI for interactive visualization, we established that **gender, class, and fare** were the strongest predictors of survival on the Titanic. The final dashboard serves as a concise tool for stakeholders to explore these variables dynamically.

9. APPENDIX: PROJECT FILES

- [01_data_understanding.pdf](#) - Initial data inspection and stats.
- [02_data_cleaning.pdf](#) - Python cleaning scripts.
- [03_eda.pdf](#) - Exploratory analysis visualizations.
- [Titanic Survival Analysis – Project Report.pdf](#) - Full original report.

