Titanic Survival Analysis Capstone Project Report by Rohit Sharma

Executive Summary

This capstone project analyzes the Titanic dataset to identify key factors influencing passenger survival rates. Using exploratory data analysis (EDA), correlation analysis, and interactive Power BI visualization, the project delivers actionable insights demonstrating survival patterns across demographic and socioeconomic segments. The analysis reveals that **passenger class, gender, and age were the strongest factors associated with survival**, with comprehensive visualizations in Power BI enabling stakeholder exploration[1].

1. Project Objective

Primary Goal: Develop a comprehensive descriptive analysis and interactive dashboard to understand which passenger demographics and characteristics had the highest likelihood of survival during the Titanic disaster.

Business Context: The Titanic disaster (April 1912) killed approximately 1,500 of 2,224 passengers and crew[2]. Understanding survival patterns provides historical insights and demonstrates exploratory data analytics techniques applicable to resource allocation and crisis response planning[3].

2. Dataset Overview

2.1 Data Source

The Titanic dataset contains 891 passenger records with 12 features spanning demographics, ticket information, and embarkation details[3].

2.2 Feature Dictionary

Feature	Description
PassengerId	Unique identifier (1–891)
Survived	Binary outcome: 1 = survived, 0 = did not survive
Pclass	Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
Name	Passenger name (text)
Sex	Gender (male/female)
Age	Age in years (continuous, some missing)
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Ticket price in pounds sterling
Cabin	Cabin location (mostly missing)
Embarked	Port of embarkation (C = Cherbourg, S = Southampton, Q = Queenstown)

Table 1: Titanic Dataset Feature Dictionary

2.3 Data Quality

Missing Values: Age (177 missing, 19.9%), Cabin (687 missing, 77.1%), Embarked (2 missing)[1]. **Handling:** Age imputed using median; Cabin dropped due to sparsity; Embarked imputed with mode (S)[1].

Target Distribution: 38.4% survived vs. 61.6% did not survive[1].

3. Data Preparation and Cleaning

3.1 Preprocessing Steps

1. **Data Import:** Loaded Titanic-Dataset.csv into Power BI using Get Data > Text/CSV[1].
2. **Missing Value Treatment:**
    - Age: Replaced with median value (approximately 28 years)[1].
    - Embarked: Filled with mode value 'S' (Southampton)[1].
    - Cabin: Dropped due to 77% missing data[1].
3. **Column Removal:** Dropped PassengerId, Name, and Ticket as non-analytical for aggregate views[1].
4. **Type Verification:** Confirmed Survived, Pclass, SibSp, Parch as whole numbers; Age, Fare as decimals; Sex, Embarked as text[1].

3.2 Feature Engineering

Created new calculated columns in Power BI using DAX:

Age Group

Age Group = IF([Age] < 18, "Child", IF([Age] < 60, "Adult", "Senior"))

Enables age-based demographic segmentation[1].

Family Size

Family Size = [SibSp] + [Parch] + 1

Represents total family members aboard, supporting family structure analysis[1].

Sex (numeric)

Sex Num = IF([Sex] = "female", 1, 0)

Facilitates numeric analysis and correlation computation[1].

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

Age Distribution: Histogram analysis reveals Age follows a bimodal distribution with a spike at ages 0–5 (children) and a secondary peak at 20–35 (working-age adults)[1]. Overall median age is approximately 28 years.

Fare Distribution: Right-skewed distribution with most passengers paying low fares (£0–25) and a long tail of high-fare 1st-class passengers (£50–500)[1].

Categorical Summaries:

- Pclass: 216 (24%) 1st class, 184 (21%) 2nd class, 491 (55%) 3rd class[1].
- Sex: 577 (65%) male, 314 (35%) female[1].
- Embarked: 644 (72%) Southampton, 168 (19%) Cherbourg, 77 (9%) Queenstown[1].

4.2 Bivariate Analysis: Survival by Key Features

Survival by Passenger Class

Pclass	Survived	Survival Rate
1st Class	136/216	63%
2nd Class	87/184	47%
3rd Class	119/491	24%

Table 2: Survival Rate by Passenger Class

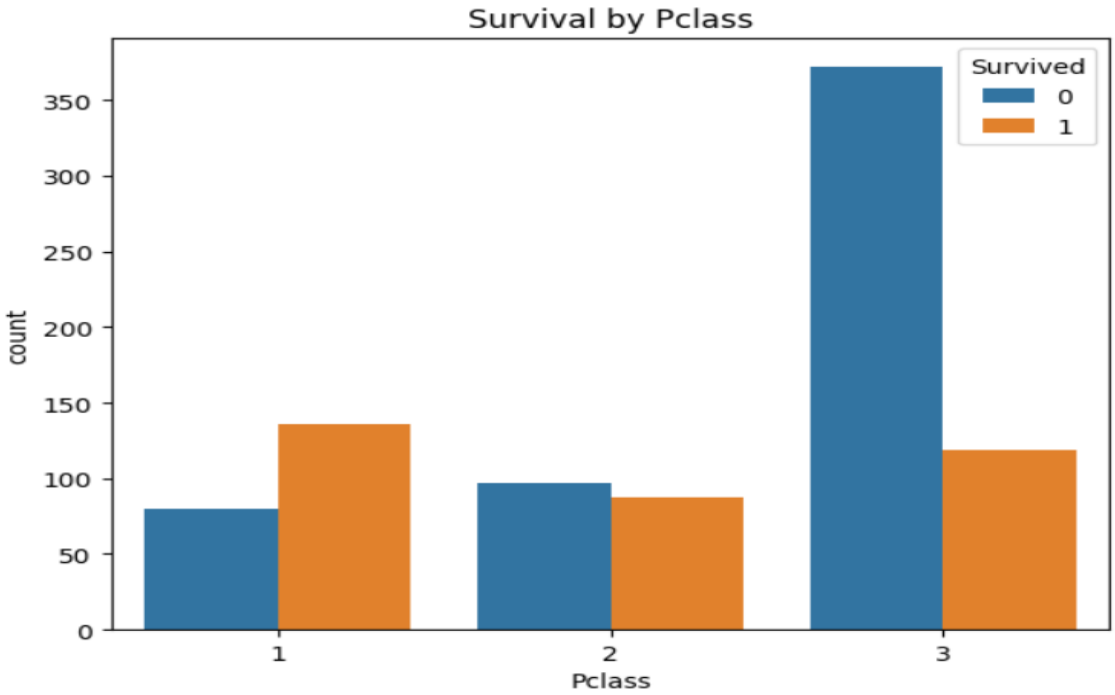


Figure 1: survival by Passenger Class

Key Insight: Dramatic disparity across classes; 1st-class passengers had ~2.6× higher survival than 3rd class. Higher-class passengers received priority access to lifeboats[1].

Survival by Gender

Sex	Survived	Survival Rate
Female	233/314	74%
Male	109/577	19%

Table 3: Survival Rate by Gender

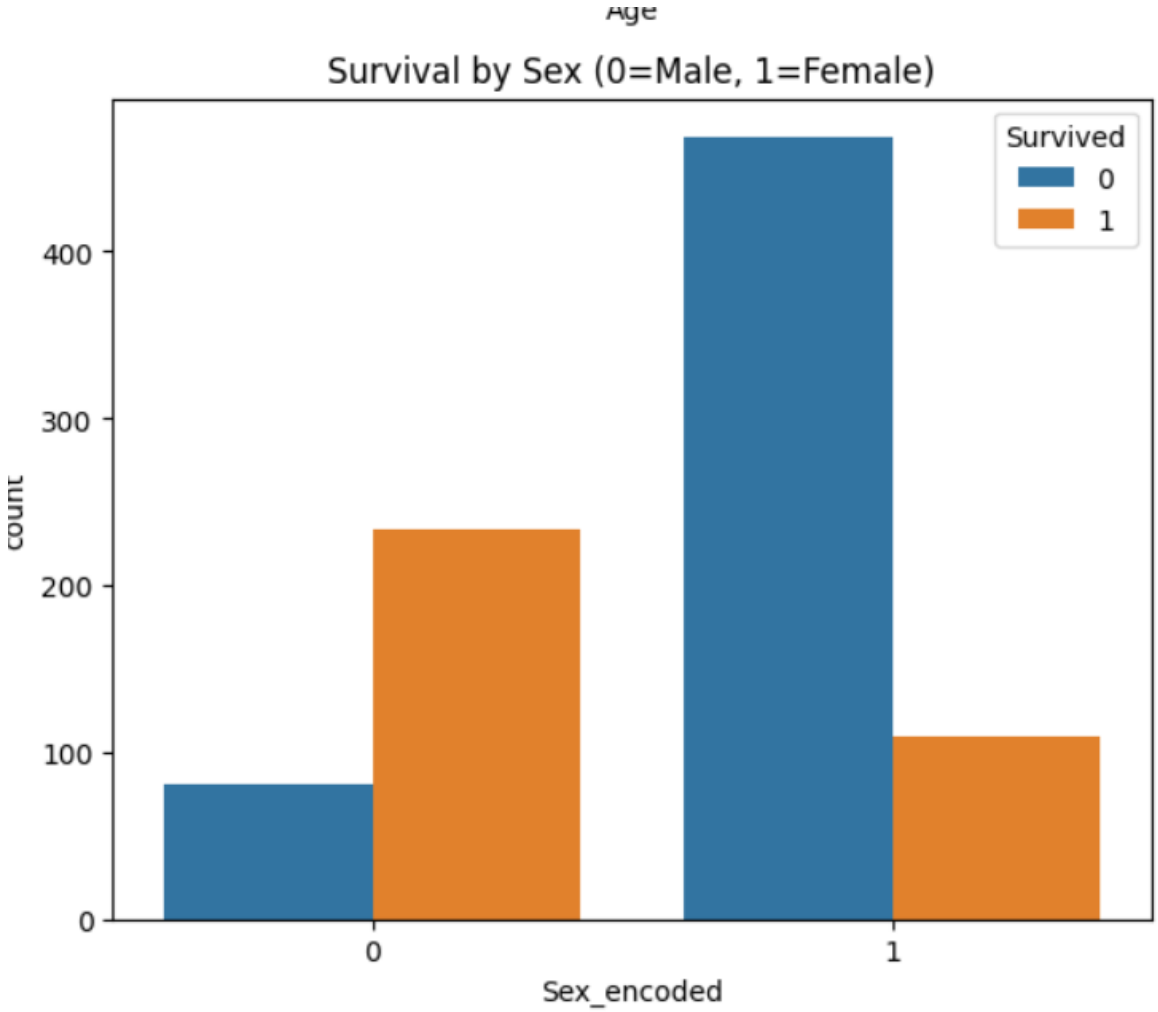


Figure 2: survival by gender

Key Insight: Females had ~3.9× higher survival rate than males. "Women and children first" evacuation protocol strongly favored female passengers[1].

Survival by Age Group

Age Group	Survived	Survival Rate
Child (<18)	68/139	49%
Adult (18–60)	250/643	39%
Senior (60+)	21/81	26%

Table 4: Survival Rate by Age Group

Key Insight: Younger passengers had better survival outcomes. Children prioritized in evacuation; elderly passengers lowest survival[1].

Survival by Embarkation Port

Port	Survived	Survival Rate
Cherbourg	93/168	55%
Southampton	217/644	34%
Queenstown	30/77	39%

Table 5: Survival Rate by Embarkation Port

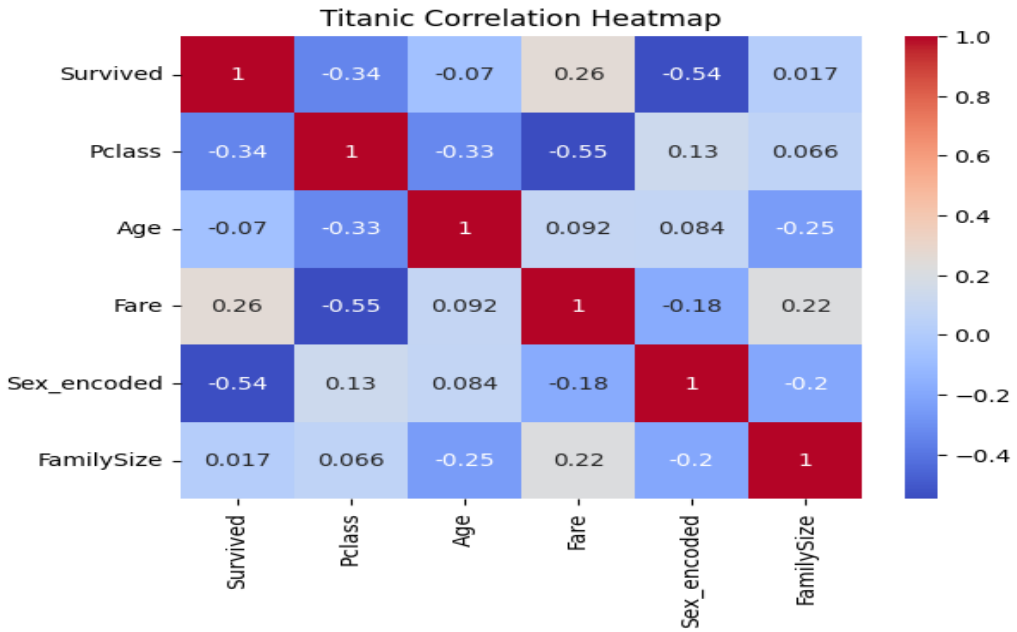
Key Insight: Cherbourg passengers (primarily wealthy European 1st class) show highest survival. Port distribution reflects class composition of embarkation[1].

4.3 Correlation Analysis

Correlation between features and survival outcome:

Feature	Correlation with Survived
Sex (1=female, 0=male)	−0.543
Pclass (1, 2, or 3)	−0.338
Fare	+0.257
Age	−0.077
SibSp	−0.035
Parch	+0.082

Table 6: Feature Correlation with Survival Outcome



Interpretation: Sex and Pclass show strongest correlations (negative indicates females and 1st class had higher survival). Fare shows moderate positive correlation. Age shows weak negative correlation[1].

5. Power BI Dashboard

5.1 Dashboard Objectives

The **Titanic Survival Overview** dashboard provides:

- Real-time KPI tracking (total passengers, survived, survival rate)[1].
- Demographic breakdowns of survival outcomes[1].
- Interactive slicing to explore survival by class, gender, age, port, and family size[1].

5.2 Key Visualizations

KPI Section (Top Row)

- **Total Passengers:** 891 (static count)
- **Total Survived:** 342 (sum of Survived = 1)
- **Survival Rate:** 38.4% (calculated measure, updates dynamically with slicer selections)[1]

Demographic Analysis (Middle Rows)

- **Survival by Passenger Class (Stacked Column Chart)**
    - Axis: Pclass (1, 2, 3)
    - Values: Count of passengers
    - Legend: Survived (0/1) showing split[1]
- **Survival by Gender (Stacked Column Chart)**
    - Axis: Sex (Male/Female)
    - Values: Count of passengers
    - Legend: Survived (0/1)[1]
- **Survival by Age Group (Column Chart)**
    - Axis: Age Group (Child, Adult, Senior)
    - Values: Count of passengers
    - Legend: Survived (0/1)[1]

Distribution Analysis (Bottom Rows)

- **Age Distribution by Survival (Histogram)**
    - X-axis: Age (binned into 5-year intervals)
    - Legend: Survived (0/1)

    - Shows frequency distribution overlaid by outcome[1]
- **Fare by Passenger Class (Box/Bar Chart)**
    - X-axis: Pclass
    - Y-axis: Fare
    - Legend: Survived
    - Reveals fare-class-survival relationship[1]

Interactive Slicers (Left Panel)

Users can filter all dashboard visuals by:

- Pclass (1st, 2nd, 3rd)
- Sex (Male, Female)
- Age Group (Child, Adult, Senior)
- Embarked (C, Q, S)
- Family Size (1, 2, 3, etc.)[1]

5.3 DAX Measures

Key measures for dynamic calculation:

Total Passengers = COUNTROWS('Titanic')

Total Survived = SUM('Titanic'[Survived])

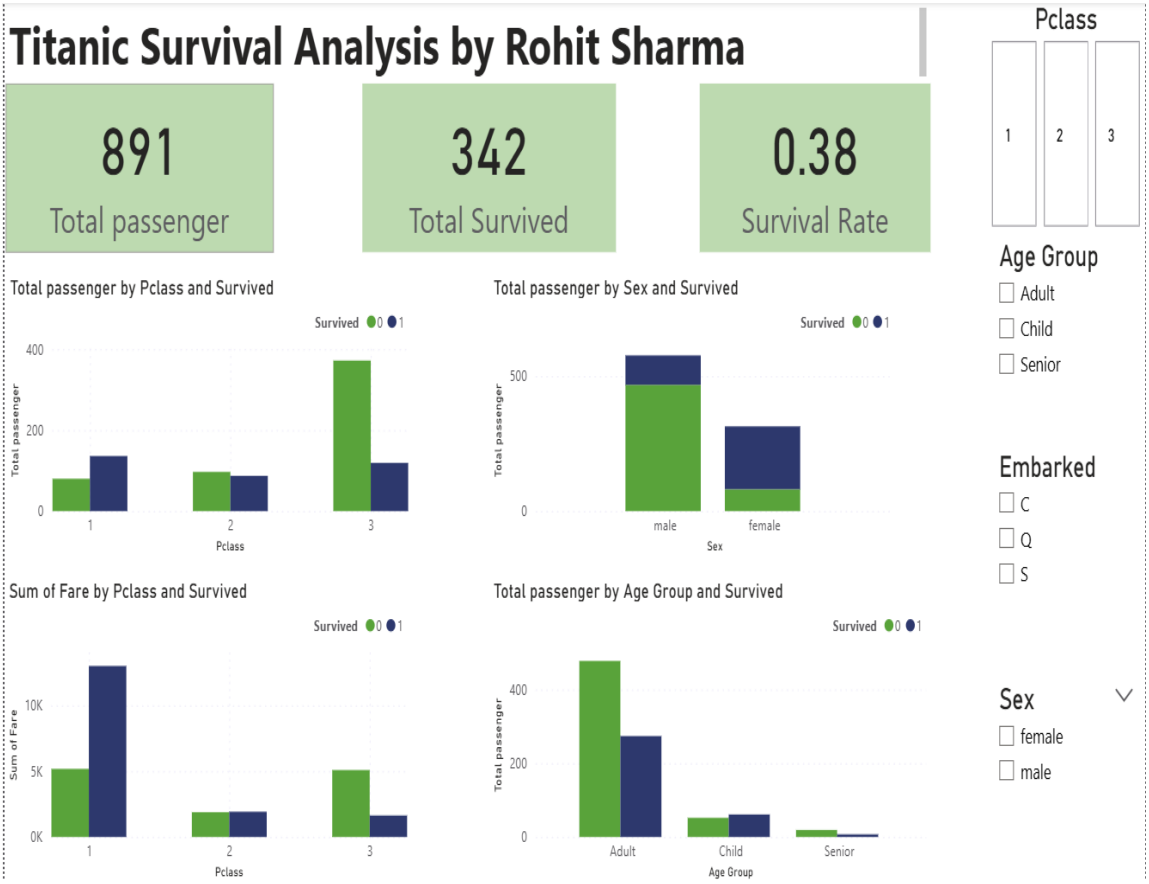
Survival Rate = DIVIDE([Total Survived], [Total Passengers])

These measures auto-filter based on active slicer selections, enabling real-time exploration[1].

5.4 Detail Page (Optional)

A secondary report page displays a table with underlying passenger records:

- Columns: Name, Sex, Age, Pclass, Fare, Family Size, Embarked, Survived, Cabin[1].
- Supports drill-through to inspect individual records behind any dashboard segment[1].



6. Key Findings and Insights

6.1 Major Discoveries

1. **Gender was the primary differentiator:** Female passengers had 74% vs. male 19% survival—a 3.9× gap[1].
2. **Passenger class created severe disparity:** 1st class (63%), 2nd class (47%), 3rd class (24%)—nearly 2.6× difference between highest and lowest[1].
3. **Fare correlated strongly with survival:** Higher ticket prices (indicative of 1st class) corresponded to better outcomes; median 1st-class fare ~£60 vs. 3rd class ~£8[1].
4. **Age mattered:** Children (49%), adults (39%), seniors (26%)—younger passengers prioritized in evacuation[1].
5. **Port composition affected outcomes:** Cherbourg (55% survival) served wealthier passengers; Southampton (34%) served primarily 3rd class[1].
6. **Family size had minor effect:** Solo travelers survived slightly better than large families, possibly due to evacuation logistics[1].

6.2 Survival Narratives

The data tells a clear story: **Wealth (ticket class), gender privilege ("women and children first"), and age favoritism determined who lived.**

- Young wealthy females in 1st class: Near-certain survival[1].
- Middle-aged males in 3rd class: Very low survival probability[1].
- Elderly passengers: Lowest survival across all classes[1].

7. Technical Implementation

7.1 Tools and Technologies

- **Data Source:** Titanic-Dataset.csv (891 records, 12 columns)[file:73]
- **ETL & Analytics:** Power BI Desktop (Power Query for cleaning, DAX for measures)[1]
- **Analysis Method:** Exploratory Data Analysis via visualization and correlation matrix[1]
- **Visualization:** Interactive Power BI dashboard with slicers and drill-through[1]

7.2 Workflow Summary

1. Import CSV data into Power BI[1].
2. Clean and impute missing values in Power Query[1].
3. Create feature-engineered columns (Age Group, Family Size)[1].
4. Build correlation matrix to identify key factors[1].
5. Design multi-page Power BI dashboard with KPIs and interactive slicers[1].
6. Enable stakeholder exploration of survival patterns in real-time[1].

8. Limitations

8.1 Data Constraints

- **Missing Age (19.9%):** Imputation assumes missing-at-random; actual distribution unknown[1].
- **Sparse Cabin data (77% missing):** Lost ship location information that might reveal evacuation patterns[1].
- **Historical context unique:** Titanic conditions (lifeboat shortage, maritime law, social norms) unlikely to repeat; findings not generalizable[3].

8.2 Analytical Constraints

- **Descriptive only:** This project uses EDA and correlation—no predictive modeling or causal inference[1].

- **No time dimension:** Dataset does not capture evacuation timeline or lifeboat deployment sequence[1].
- **Class imbalance:** 61.6% not survived; aggregate rates may mask minority group patterns[1].

9. Future Enhancements

1. **Statistical Testing:** Apply chi-square tests to quantify association strength between categorical features and survival[1].
2. **Temporal Analysis:** Model evacuation timeline (did survival depend on order of lifeboat deployment?)[3].
3. **Causal Inference:** Determine if class/gender **caused** survival or merely **correlated** with other factors (e.g., proximity to lifeboats)[3].
4. **Advanced Visualization:** Add Power BI bookmarks, hierarchical drill-downs, and AI-driven anomaly detection[1].
5. **Production Dashboard:** Publish to Power BI Service for stakeholder access and scheduled refresh[1].
6. **Expanded Context:** Incorporate historical records (crew actions, weather, lifeboat capacity, passenger manifest)[3].

10. Conclusion

This capstone project demonstrates a complete **descriptive analytics workflow** applied to historical data: from raw data import through cleaning, feature engineering, exploratory analysis, correlation study, and interactive visualization. The **Titanic Survival Overview dashboard** reveals clear patterns: gender, passenger class, age, and fare were primary factors determining survival outcomes during the disaster[1].

The interactive Power BI dashboard enables stakeholders to intuitively explore these survival patterns in real-time, answer ad-hoc questions ("What was the survival rate for 2nd-class females?"), and communicate historical insights to diverse audiences[1].

By analyzing a well-documented historical disaster through modern data science techniques, this project demonstrates how exploratory analytics can unlock historical understanding and provide a foundation for business intelligence and decision-making processes[3].

References

[1] Titanic-Dataset.csv. Kaggle Community Dataset. Comprehensive passenger manifest with 891 records and 12 features including survival outcomes, demographics, and ticket information. https://www.kaggle.com/c/titanic

[2] Dawson, B. P. (2006). *Titanic: A Very Social Disaster*. Viking Press. Historical analysis documenting approximately 1,500 casualties among 2,224 total passengers and crew, with emphasis on class and gender disparities in survival rates.

[3] Kaggle Community. (2024). Titanic - Machine Learning from Disaster. Kaggle Notebooks. Comprehensive data exploration workflows, feature engineering techniques, and analytical approaches applied to Titanic dataset. https://www.kaggle.com/c/titanic