|  |  |
| --- | --- |
| Titanic Survival Report Analysis | |
| **By** **sanjay kumar** | |
| **Date: April 2025** |  |

Table of  
Contents

|  |  |
| --- | --- |
| 01 | Introduction |
| **02** | Project Objective |
| **03** | Dataset Overview |
| **04** | DESCRIPTIVE ANALYSIS |
| **05** | Outlier table |
| **06** | Descriptive Analysis Summary – Titanic Dataset |

Table of  
Contents

|  |  |
| --- | --- |
| 07 | Dashboard Insights - i |
| **08** | Dashboard Insights – ii |
| **09** | Dashboard Insights – iii |
| **10** | Dashboard Insights – iv |
| **11** | conclusion |
| **12** | appendix & tools |

1

|  |  |
| --- | --- |
| Introduction | |
| Understanding the Dataset and Objective | |
| An overview of the dataset and the aim behind analyzing Titanic passenger survival using data science techniques. | |
| This report presents an analysis of the **Titanic dataset**, aiming to understand the factors that influenced passenger *survival*. The dataset includes details such as **age, gender, ticket class, fare**, and more. | Our goal is to uncover patterns and insights through **descriptive analysis** and **data visualization**. This project is a part of my data analysis learning journey, with tools like *Python, Pandas, and PowerBI*. All steps were carefully guided and executed with a focus on learning and practical application. |

2

|  |  |
| --- | --- |
| Project Objective | |
| Goal of the Analysis | |
| The objective of this project is to identify patterns that influenced the survival chances of passengers aboard the Titanic. We aim to understand how factors such as gender, passenger class, embarkation port, fare, and age affected outcomes. | This analysis provides foundational insights using descriptive statistics and visualizations, serving as a base for further exploratory or predictive analysis. |

3

|  |
| --- |
| Dataset Overview |
| Understanding the Source and Structure |
| The dataset used in this analysis is the Titanic dataset, originally from the [Kaggle Titanic Competition.](https://www.kaggle.com/code/gusthema/titanic-competition-w-tensorflow-decision-forests) It contains information on 891 passengers, including demographic features (age, sex), socio-economic indicators (ticket class, fare), and survival status. The dataset was loaded and cleaned using Python and Pandas. After preprocessing, all missing values were handled, and irrelevant columns were removed for simplicity and clarity. |

| **Feature** | **Description** |
| --- | --- |
| Survived | 1 = Survived, 0 = Did not survive |
| Pclass | Ticket class (1st, 2nd, 3rd) |
| Sex | Gender of the passenger |
| Age | Age in years |
| SibSp | Siblings/spouses aboard |
| Parch | Parents/children aboard |
| Fare | Ticket fare |
| Embarked | Port of embarkation (C/Q/S) |

4

|  |
| --- |
| DESCRIPTIVE ANALYSIS |
| Understanding the Data with Statistics |
| In this section, descriptive statistics were used to understand the underlying structure of the Titanic dataset. Key measures such as **mean**, **median**, **standard deviation**, and **interquartile range (IQR)** were calculated for numerical features like Age, Fare, SibSp, and Parch. Outliers were identified using the IQR method and standard deviation checks. The analysis revealed that **Fare** had high variation and outliers, while **Age** showed a fairly normal distribution after imputation. |

A descriptive analysis was performed on key numerical columns: **Age**, **Fare**, and **SibSp**.  
We used measures like **mean**, **median**, **standard deviation**, and **interquartile range (IQR)** to understand data distribution and detect potential outliers.

* Age showed a relatively stable distribution after handling missing values.
* Fare had a high spread and several outliers.
* SibSp was moderately skewed with outliers at the higher end.  
  Outliers were identified using the **IQR rule** and **standard deviation thresholds**.

| **Column** | **Q1** | **Q3** | **IQR** | **Lower Bound** | **Upper Bound** |
| --- | --- | --- | --- | --- | --- |
| Age | 22.0 | 35.0 | 13.0 | 2.5 | 54.5 |
| SibSp | 0.0 | 1.0 | 1.0 | -1.5 | 2.5 |
| Fare | 7.91 | 31.0 | 23.09 | -26.72 | 65.63 |

**IQR Summary Table for Outlier Detection**

|  |
| --- |
| Outlier Table |
| Understanding the Data with Outliers |
| The previous table summarizes the IQR calculations used for detecting outliers. The table below shows the actual number of outliers identified in each numerical column. **Note:** The column *Parch* is excluded from this analysis as approximately 75% of its values are 0, offering limited variability for meaningful outlier detection. |

| **Column** | **Outlier Count** |
| --- | --- |
| Age | 66 |
| SibSp | 46 |
| Fare | 116 |

**Outlier Count by Column**

**The *Fare* column had the highest number of outliers, indicating heavy variability in passenger ticket prices.**

5

|  |
| --- |
| Descriptive Analysis Summary – Titanic Dataset |
| Objective  Performed descriptive analysis on the Titanic dataset to uncover survival trends and prepare the data for visualization. |

# 6

**Data Preparation**

* Removed unnecessary columns (PassengerId, Name, Ticket, Cabin)
* Handled missing values: Age (median), Embarked (mode)
* Treated Pclass and Survived as categorical variables

**Categorical Insights**

* **Sex:** 65% male, 35% female. Survival rate: females 74%, males 19%
* **Pclass:** 3rd class most frequent (55%) with lowest survival (24%)
* **Embarked:** Majority boarded at S (73%), but highest survival was from C (55%)

**Numerical Insights**

* **Age:** Mean ~29.4, stable distribution post-imputation
* **Fare:** Median = 14.45, mean = 32.2 → indicates skewness with many outliers
* **SibSp:** Skewed; Parch excluded due to lack of spread (mostly zero)

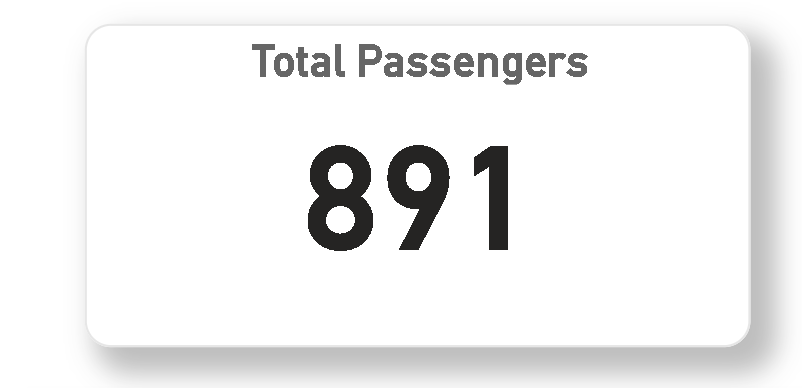
**Outlier Detection (IQR Method)**

* Outliers found in Fare, Age, and SibSp
* No rows were dropped to retain dataset integrity

**Key Patterns**

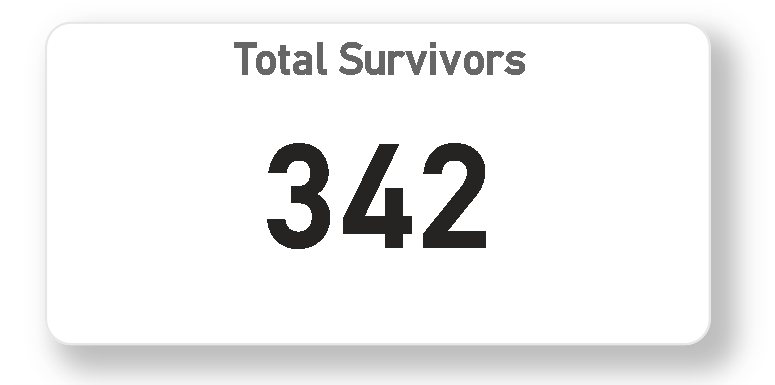
* Survival strongly associated with **gender** and **ticket class**
* **Embarkation port** had secondary influence
* **Age** showed minimal impact on survival outcome

|  |
| --- |
| Dashboard Insights |
| Key Visual Findings from Power BI |

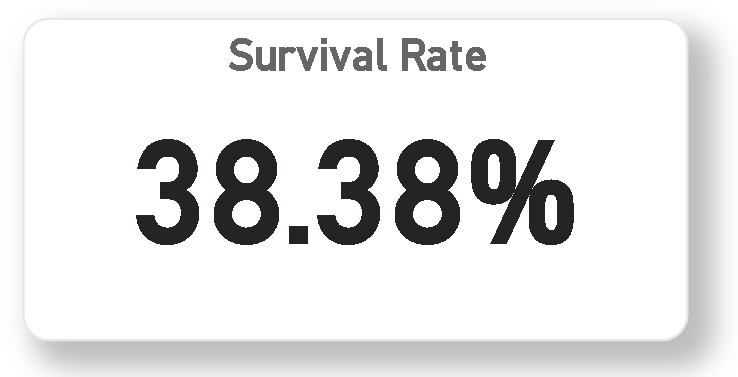


7

**Total Passengers:** Shows the full passenger count (**891**) used in the analysis after cleaning.



**Total Survivors:** Indicates how many passengers survived (342), forming the basis of the survival rate.

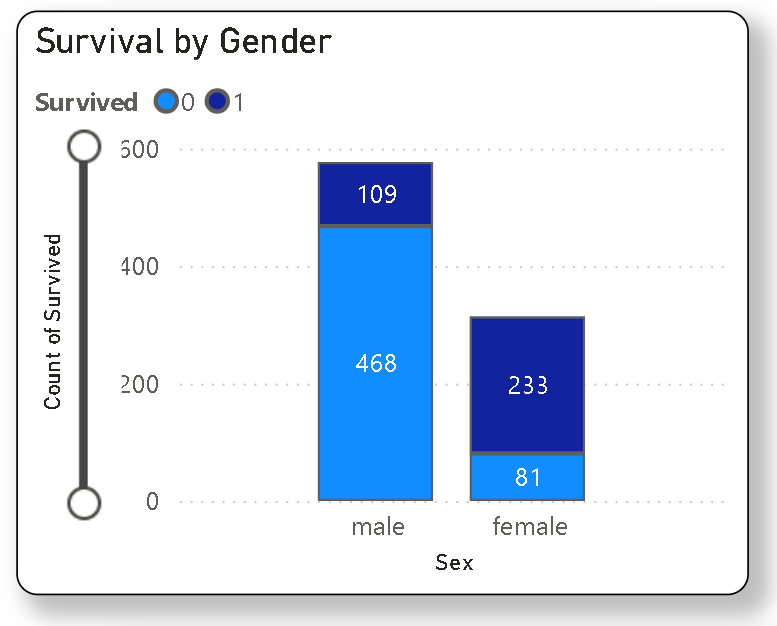


**Survival Rate:** Displays the overall survival percentage (38.4%), which is used as a benchmark across other visuals.

|  |
| --- |
| Dashboard Insights - ii |
| Key Visual Findings from Power BI |

**1. Survival by Gender**

8

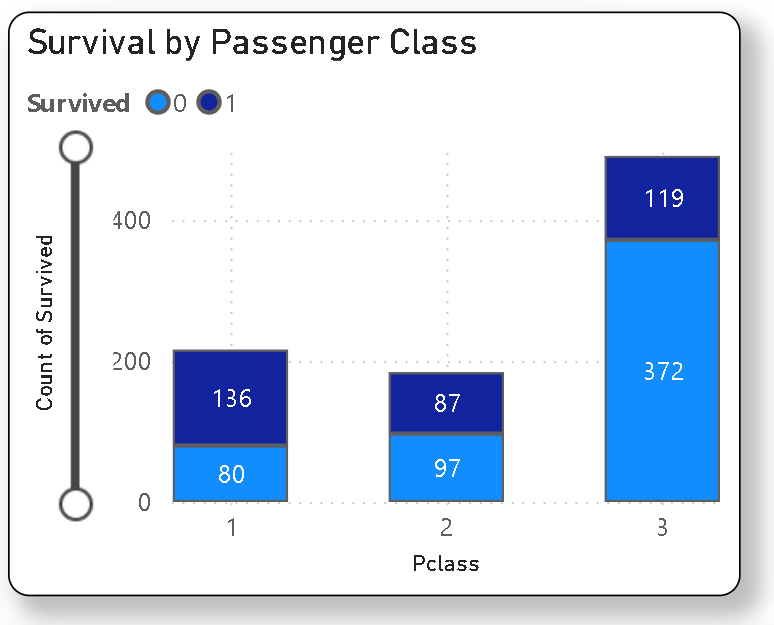


**Insight:**  
Female passengers had a **much higher survival rate** (~74%) compared to males (~19%).  
This shows a clear gender-based priority during evacuation.

|  |
| --- |
| Dashboard Insights – iii |
| Key Visual Findings from Power BI |

**2. Survival by Passenger Class (Pclass)**

9

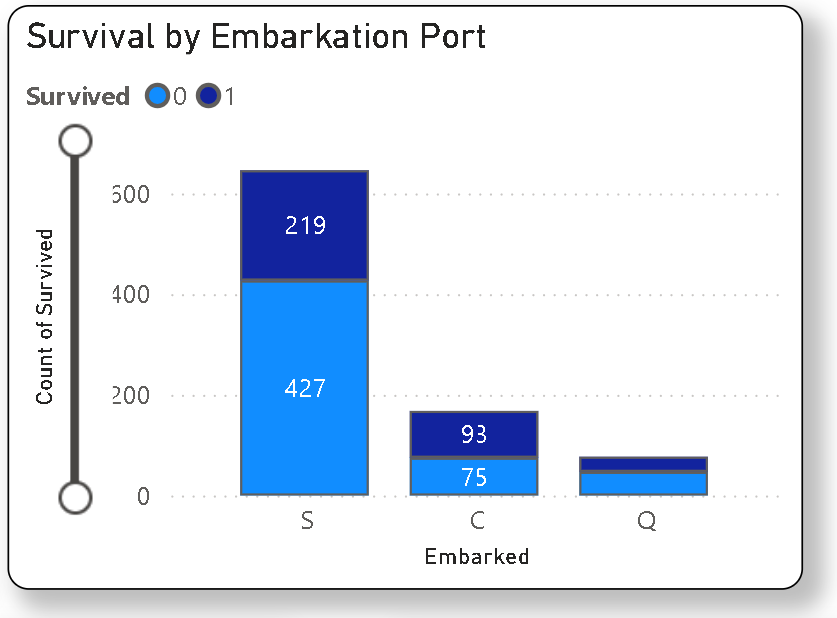


**Insight:**  
Passengers from **1st class** had the highest survival rate (~63%), followed by 2nd class (~47%), while **3rd class** had the lowest (~24%).  
Socio-economic status had a strong influence on survival.

|  |
| --- |
| Dashboard Insights - iv |
| Key Visual Findings from Power BI |

**3. Survival by Embarkation Port**

# 10



**Insight:**  
Passengers who embarked from **Cherbourg (C)** had the highest survival rate **(~55%)**, followed by **Queenstown (Q: ~39%)** and **Southampton (S: ~34%)**.  
This may reflect differences in passenger class or demographics per port.

|  |
| --- |
| Conclusion |
| Summary of Findings |

This analysis of the Titanic dataset reveals strong correlations between **passenger survival** and factors such as **gender, class, and embarkation port**.  
Female passengers and those in **1st class** had significantly higher survival rates. The **Fare** feature showed high variability and outliers, while the **Age** feature was more stable post-imputation.  
Through descriptive statistics and Power BI visualizations, clear survival patterns were uncovered that provide a foundation for more advanced exploration such as predictive modeling.

# 11

Limitations

|  |
| --- |
|  |

* The dataset contains **missing values** that were filled using imputation, which may slightly influence accuracy.
* The analysis is based on **historical data** with limited features — factors like crew status, health, or exact rescue location are not included.
* Only **descriptive analysis** was performed; no predictive modeling or hypothesis testing was applied in this report.
* **Outlier detection** was based on statistical rules (IQR, std dev), but contextual relevance was not further validated.

|  |
| --- |
| Appendix |
| **APPENDIX & TOOLS USED** |

**🛠️ Tools & Technologies:**

* **Python (Pandas) – data cleaning and analysis**
* **Jupyter Notebook – execution and documentation**
* **Power BI – visualization and dashboard**
* **Excel – dataset handling**
* **GitHub – project hosting**
* **Word – reporting**

**📦 Dataset Source:**

* **Titanic Dataset from Kaggle -** [Link](https://www.kaggle.com/code/gusthema/titanic-competition-w-tensorflow-decision-forests)

**🧩 Project Info:**

* **Duration: April 2025**
* **Platform: Windows 11, PyCharm IDE**
* **Author: *Sanjay Kumar***

# 12