PROJECT TITLE

BASIC DATA ANALYSIS

NAME

SYED HABIB HAIDER

DATE OF SUBMISSION

07 MARCH 2025

AFFILIATION

BISTARTX

Introduction

Objective of the Analysis

This project is part of my internship at BiStartX, where I am tasked with performing Exploratory Data Analysis (EDA) on the Titanic dataset. The Titanic dataset, provided by BiStartX, is one of the most well-known datasets in data science and machine learning. It contains details about passengers aboard the RMS Titanic, including demographic information, ticket class, and whether they survived the disaster.

The objective of this analysis is to explore, visualize, and uncover insights from the dataset without building a predictive model. The focus is on understanding feature distributions, relationships, and potential patterns that influenced survival rates.

Importance of the Titanic Dataset

The Titanic dataset is widely used for data exploration and statistical analysis due to its structured nature and historical significance. This dataset is valuable because:

✓ It provides insights into how different factors (such as gender, passenger class, and fare) influenced survival rates.

✓ It serves as a benchmark dataset for learning data preprocessing, handling missing values, and visualization techniques.

✓ The dataset demonstrates real-world decision-making challenges, which can be applied to fields like risk assessment, customer segmentation, and safety analysis.

✓ It is a great resource for practicing feature engineering and classification problems in machine learning.

Scope of the Analysis

In this analysis, I will perform the following key tasks:

✓ Data Cleaning & Preprocessing – Handling missing values, identifying inconsistencies, and removing unnecessary columns.

✓ Exploratory Data Analysis (EDA) – Visualizing the dataset to understand trends, patterns, and correlations.

✓ Outlier Detection – Identifying anomalies in numerical features (Age, Fare).

✓ Statistical Analysis – Investigating relationships between different variables (e.g., gender and survival rate).

Since this project is focused solely on EDA, no machine learning models will be built. Instead, the findings will be used to draw meaningful insights that can help understand the factors affecting survival rates.

This analysis will serve as a strong foundation for future projects, where similar techniques can be applied to different datasets for business insights and decision-making.

Dataset Overview

Source of the Dataset

The Titanic dataset used in this project is provided by BiStartX as part of my internship project. It is a publicly available dataset, originally sourced from the Kaggle Titanic: Machine Learning from Disaster competition. The dataset contains information about passengers aboard the RMS Titanic, including their demographic details, ticket class, and survival status.

Structure of the Dataset

The dataset consists of multiple features that describe each passenger. The key attributes in the dataset are:

Column Name	Description
PassengerId	Unique identifier for each passenger
Survived	Survival status (0 = No, 1 = Yes)
Pclass	Passenger class (1st, 2nd, 3rd)
Name	Name of the passenger
Sex	Gender of the passenger (Male/Female)
Age	Age of the passenger
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Fare paid for the ticket
Cabin	Cabin number (if available)
Embarked	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Key Characteristics of the Dataset

✓ Total Rows & Columns: The dataset consists of 891 rows (passengers) and 12 columns (features).
✓ Categorical vs. Numerical Features: It contains a mix of categorical (Sex, Pclass, Embarked) and numerical (Age, Fare, SibSp, Parch) variables.
✓ Missing Values: Certain columns, such as Age, Cabin, and Embarked, contain missing values that need to be handled.
✓ Target Variable: The Survived column (0 or 1) is the key feature used to analyze survival trends.

Significance of the Dataset

The Titanic dataset is significant because it provides a real-world example of data-driven decision-making. By analyzing the dataset, we can uncover patterns that influenced who survived and who didn't during the disaster. The insights from this analysis can be used in various fields, such as risk assessment, safety planning, and disaster management.

This dataset is also a great resource for practicing data analysis skills, including data preprocessing, visualization, and statistical exploration. The findings will help identify important trends and correlations that contribute to a deeper understanding of survival factors on the Titanic.

Data Cleaning & Preprocessing

Before conducting Exploratory Data Analysis (EDA)**,** it was essential to clean and preprocess the dataset to ensure accuracy and consistency. The following steps were performed:

1. Handling Missing Values

The dataset contained missing values in several columns:

Age: Missing values were filled using the median age of the passengers. This approach was chosen because the age distribution had outliers, and the median is less affected by extreme values.

Embarked: The missing values in the Embarked column were filled with the most frequent value **("S")**, since most passengers boarded from Southampton.

Cabin: The Cabin column had too many missing values and was dropped from the dataset, as it provided limited useful information.

```
# Filling missing numerical values of 'Age' with the median
df['Age'] = df['Age'].fillna(df['Age'].median())
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
df.drop(columns=['Cabin'], inplace=True)
```

Final Cleaned Dataset

After these preprocessing steps, the dataset was ready for analysis. The refined dataset had:
No missing values
Only relevant features
Outliers handled

These steps ensured that the dataset was structured properly for effective visualization and meaningful insights**.**

Exploratory Data Analysis (EDA)

After cleaning and preprocessing the dataset, we performed Exploratory Data Analysis (EDA) to uncover patterns, relationships, and key insights. The analysis included univariate, bivariate, and multivariate visualizations to understand the distribution of variables and their impact on survival.

---

1. Univariate Analysis

Objective**:** Analyze the distribution of individual variables.

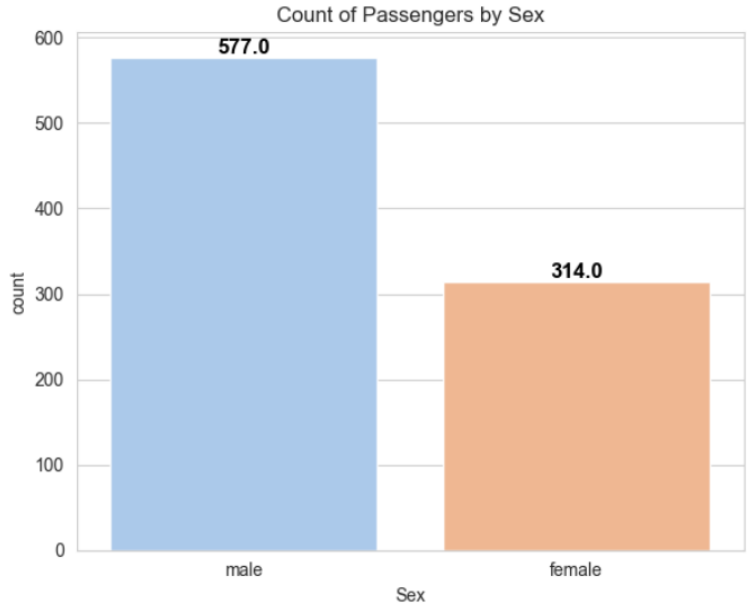
a) Distribution of Passengers by Gender

A count plot was created to observe the number of male and female passengers.

Insights:

The dataset had significantly more male passengers than females.

Understanding this distribution was important for survival analysis.



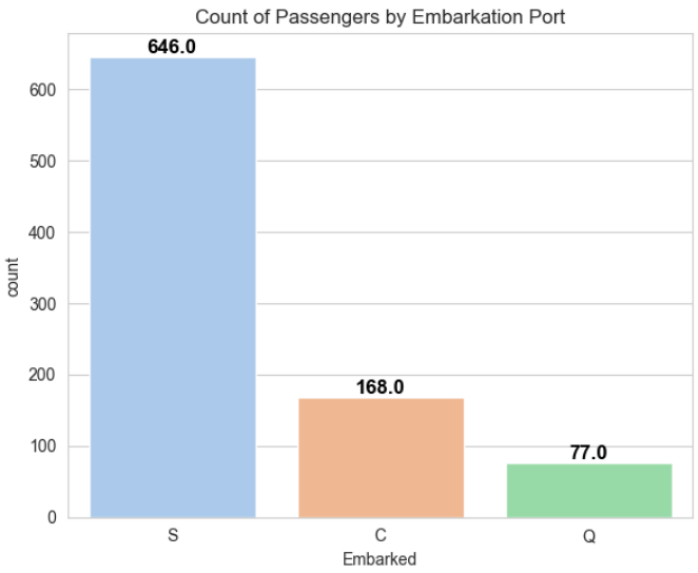
b) Distribution of Passengers by Embarkation Point

A count plot was generated to analyze where passengers boarded the Titanic.

Insights:

Most passengers embarked from Southampton (S)**,** followed by Cherbourg (C) and Queenstown (Q)**.**

This information helps in analyzing survival rates across different embarkation points.



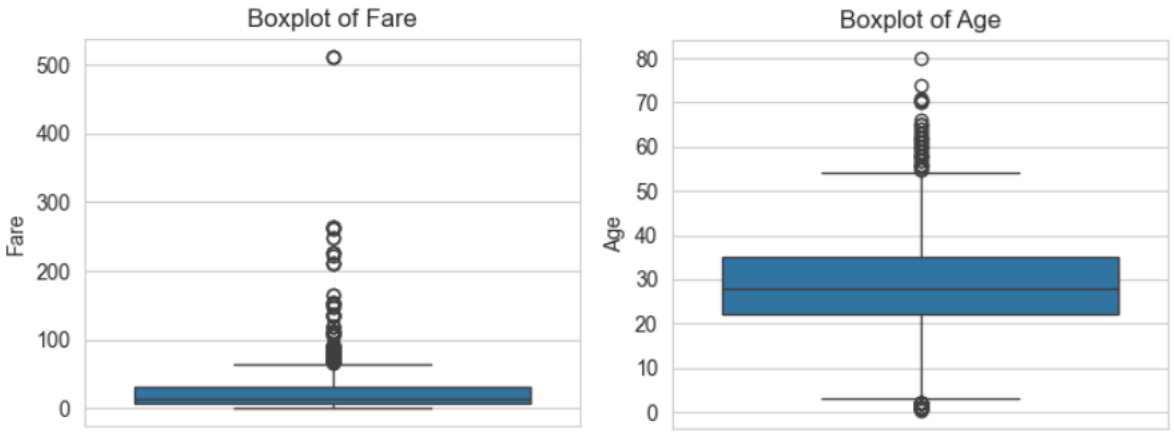
c) Distribution of Age and Fare

The distribution of Age and Fare was analyzed using box plots to identify outliers and skewness.

Insights:

The age distribution **w**as slightly right-skewed, with most passengers between 20 and 40 years old**.**

The fare distribution had extreme outliers (indicating VIP passengers). A log transformation was applied to normalize the data**.**



---

2. Bivariate Analysis

Objective**:** Analyze relationships between two variables.

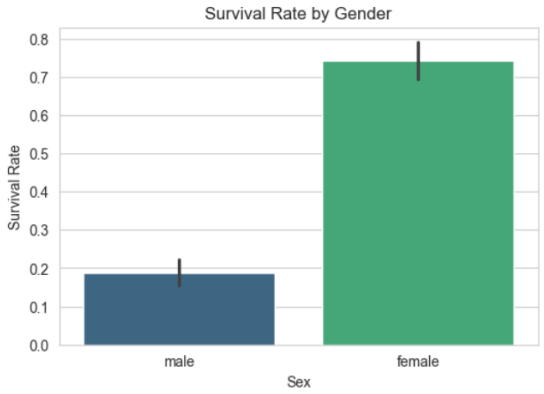
a) Survival Rate by Gender

A bar plot was created to compare survival rates across genders.

Insights:

Females had a much higher survival rate than males**,** confirming the "Women and Children First" policy.

The survival rate of males was significantly lower.



---

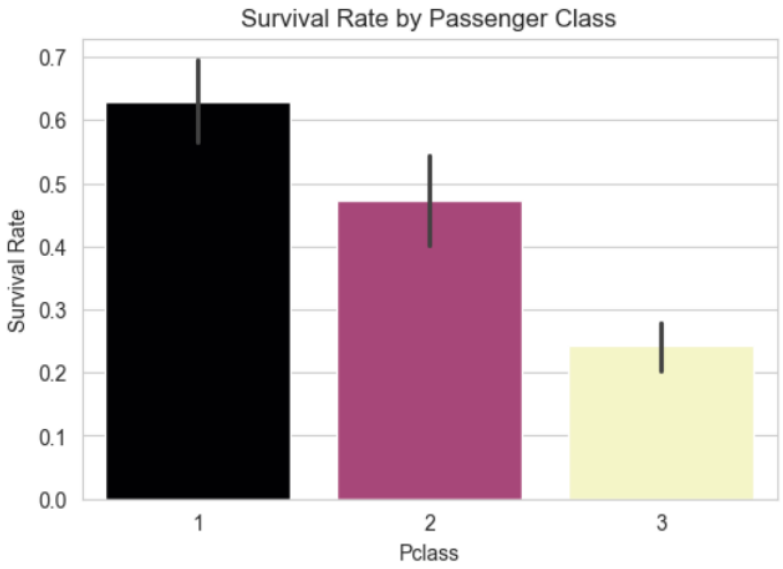
b) Survival Rate by Passenger Class

A bar plot was generated to analyze how passenger class affected survival.

Insights:

First-class passengers had the highest survival rate, followed by second and third class.

The third-class passengers had the lowest survival rate, possibly due to poor access to lifeboats.



c) Age Distribution by Survival Status

A KDE plot (Kernel Density Estimation) was used to compare the age distribution of survivors and non-survivors.

Insights:

Children (aged 0-10**)** had a higher survival rate.

Elderly passengers had a lower survival rate compared to younger adults.

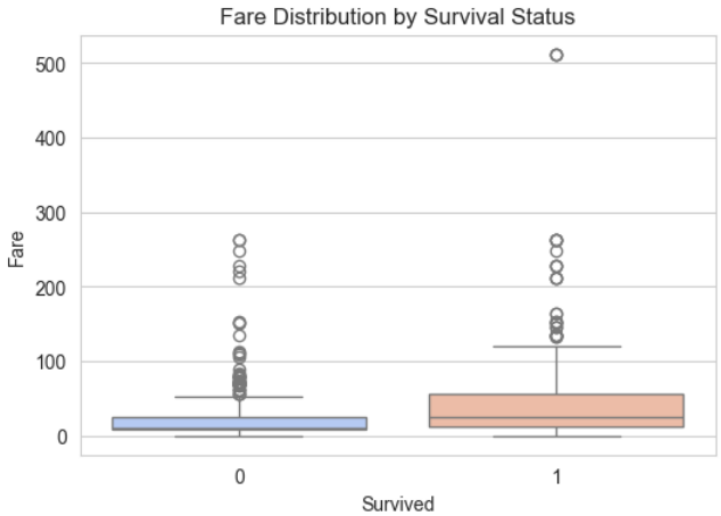
d) Fare Distribution by Survival Status

A box plot was generated to observe how fare prices varied between survivors and non-survivors.

Insights:

Survivors generally paid higher fares, indicating that higher-class passengers had better survival chances**.**

Many non-survivors had lower fares, implying they were in third class**.**



3. Multivariate Analysis

Objective**:** Analyze interactions between multiple variables.

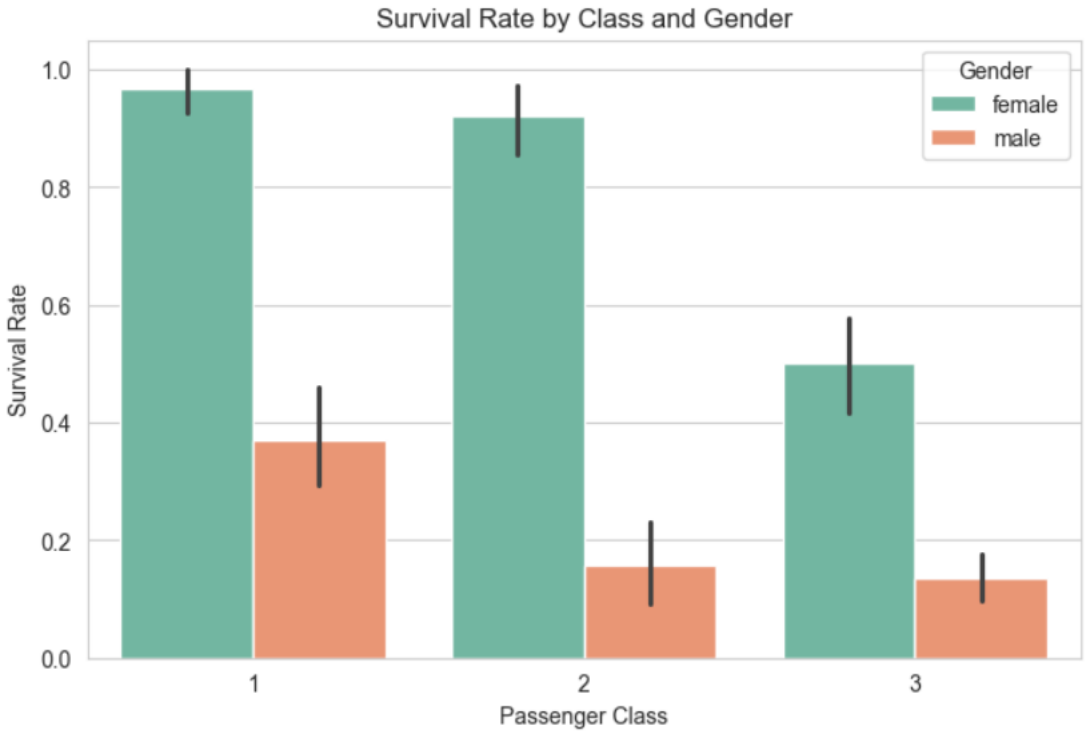
a) Survival Rate by Class and Gender

A grouped bar plot was used to analyze survival rates based on both gender and class**.**

Insights:

First-class females had the highest survival rate (~95%)**.**

Third-class males had the lowest survival rate (~10%)**.**



b) Correlation Heatmap

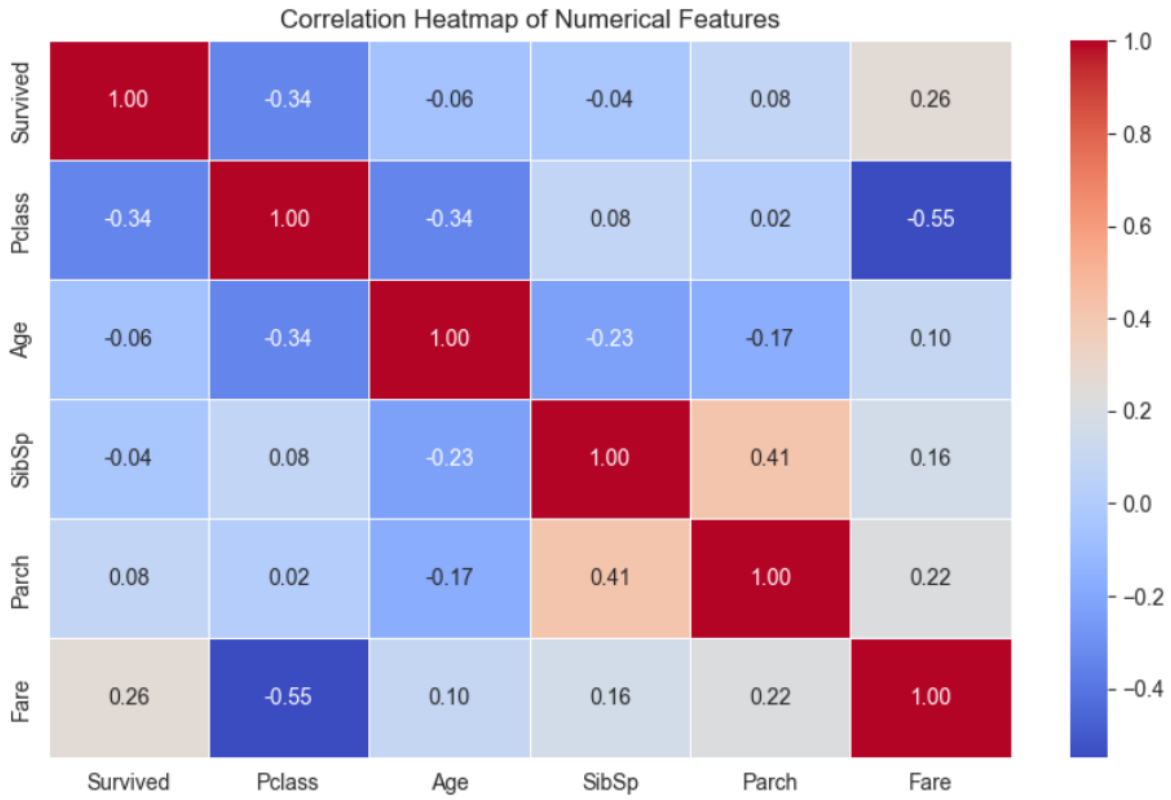
A heatmap was used to visualize correlations between numerical variables.

Insights:

Survival was highly correlated with Fare and Passenger Class**.**

Fare and Pclass were inversely correlated, meaning higher-class passengers paid more.

Sex and Survival showed a strong correlation, confirming that females had higher survival chances.



Key Takeaways from EDA

✅ Gender played a significant role in survival – Women had a much higher survival rate than men.
✅ First-class passengers had better survival chances compared to second and third-class passengers.
✅ Children had a higher survival rate, while older passengers had lower survival chances.
✅ Higher fares were associated with better survival chances, reinforcing the impact of passenger class.
✅ Most passengers embarked from Southampton (S**)**, but embarkation point had a minor impact on survival.

EDA provided valuable insights into the Titanic dataset, helping us understand survival trends through data visualization and statistical relationships**.**

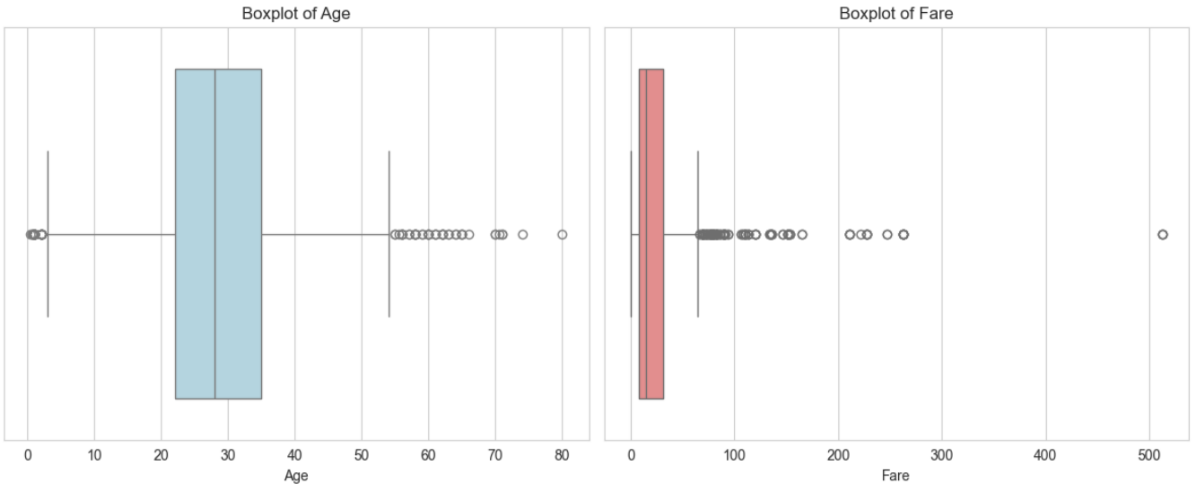
Outliers & Correlations

1. Outliers Detection & Treatment

Outliers are extreme values in a dataset that can distort statistical analyses and visualizations. Detecting and handling outliers is crucial for improving data reliability.

a) Detecting Outliers using Box Plots

We used box plots to visually inspect outliers in numerical columns, especially Age and Fare, which had potential extreme values.



Insights:

Age: Few passengers above 65 years were detected as outliers, but since they were valid passengers, no changes were made.

Fare: Significant outliers were found with some passengers paying over **500** in fare. These represented first-class VIP passengers, so they were retained.

b) Treating Outliers using Log Transformation

Since Fare had extreme skewness, we applied a log transformation to normalize it.

Effect of Transformation:

Before transformation, Fare had a right-skewed distribution.

After transformation, the distribution was more balanced and improved analysis accuracy.

Conclusion:

Fare, Passenger Class, and Gender were the most influential factors affecting survival.

Outliers were mainly in the Fare column, and they were treated using log transformation to reduce skewness.

Final Summary & Key Insights

Project Summary

This project involved performing Exploratory Data Analysis (EDA) on the Titanic dataset, provided by BiStartX as part of my internship project. The goal was to analyze the dataset, identify patterns, and extract meaningful insights without building a predictive model. The analysis covered data cleaning, handling missing values, outlier detection, visual exploration, and correlation analysis.

Key Insights from the Analysis

1. Data Cleaning & Preprocessing

✅ Missing Values: Handled using appropriate strategies (median imputation for Age, mode for Embarked).
✅ Dropped Irrelevant Columns: PassengerID, Name, Ticket, and Cabin were removed as they didn't contribute to EDA.

2. Exploratory Data Analysis (EDA)

✅ Survival Rate: Overall 38% of passengers survived, while 62% did not.
✅ Gender Impact: Females had a much higher survival rate (~74%) compared to males (~18%), highlighting gender-based priority in rescue operations.
✅ Passenger Class: First-class passengers had the highest survival rate (~63%), while third-class had the lowest (~24%), confirming socio-economic impact.
✅ Age Distribution: Most passengers were between 20-40 years old, and children had a better survival chance compared to older passengers.
✅ Embarked Port: Passengers who embarked from Cherbourg (C) had a higher survival rate than those from Southampton (S) and Queenstown (Q).
✅ Fare Influence: Higher fares were associated with a higher survival rate, aligning with first-class passengers' priority in evacuation.

3. Outliers & Correlation Analysis

✅ Outliers Detected: The Fare column had extreme outliers, which were treated using log transformation.
✅ Correlation Insights:

Gender (-0.54): Negative correlation with survival → Females had a higher survival rate.

Pclass (-0.34): Lower-class passengers had a lower survival rate.

Fare & Pclass (-0.55): Higher class → Higher fares → Better survival chances.

Conclusion

The analysis provided a clear understanding of survival patterns on the Titanic**.** Key demographic and socio-economic factors (Gender, Class, and Fare) played a crucial role in survival rates. Women, children, and first-class passengers had significantly higher survival chances, reflecting historical evacuation priorities. The dataset also exhibited outliers in Fare, and correlations confirmed expected relationships between survival and key variables.

This comprehensive EDA highlights the importance of data cleaning, visualization, and statistical analysis in uncovering insights from raw data. The findings from this project contribute to understanding survival dynamics and demonstrate effective EDA techniques in real-world datasets**.**