Exploratory Data Analysis Report

Prepared by: Vinay Raj Singh
Organization: Elevate Labs

Executive Summary

This report provides exploratory insights from the Titanic dataset, summarizing passenger demographics, survival distribution, and key statistical findings. The analysis includes visual representations extracted from the original notebook.

Exploratory Data Analysis on Titanic Dataset

Knowing your data:

Visualization:

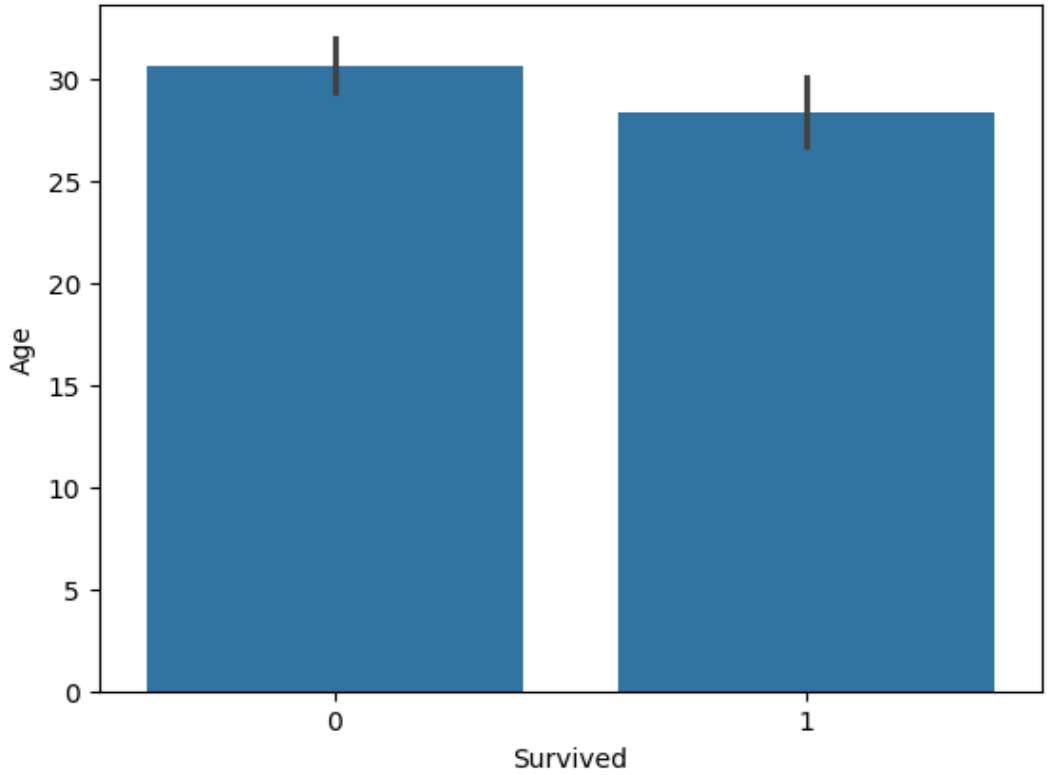
key findings: - **Higher average age among non-survivors (Survived = 0):** - The bar for those who did not survive is taller, indicating they were older on average. - **Lower average age among survivors (Survived = 1):** - Survivors were generally younger, suggesting age may have influenced survival chances. - **Error bars present:** - These likely represent standard deviation or standard error, showing variability in age within each group.

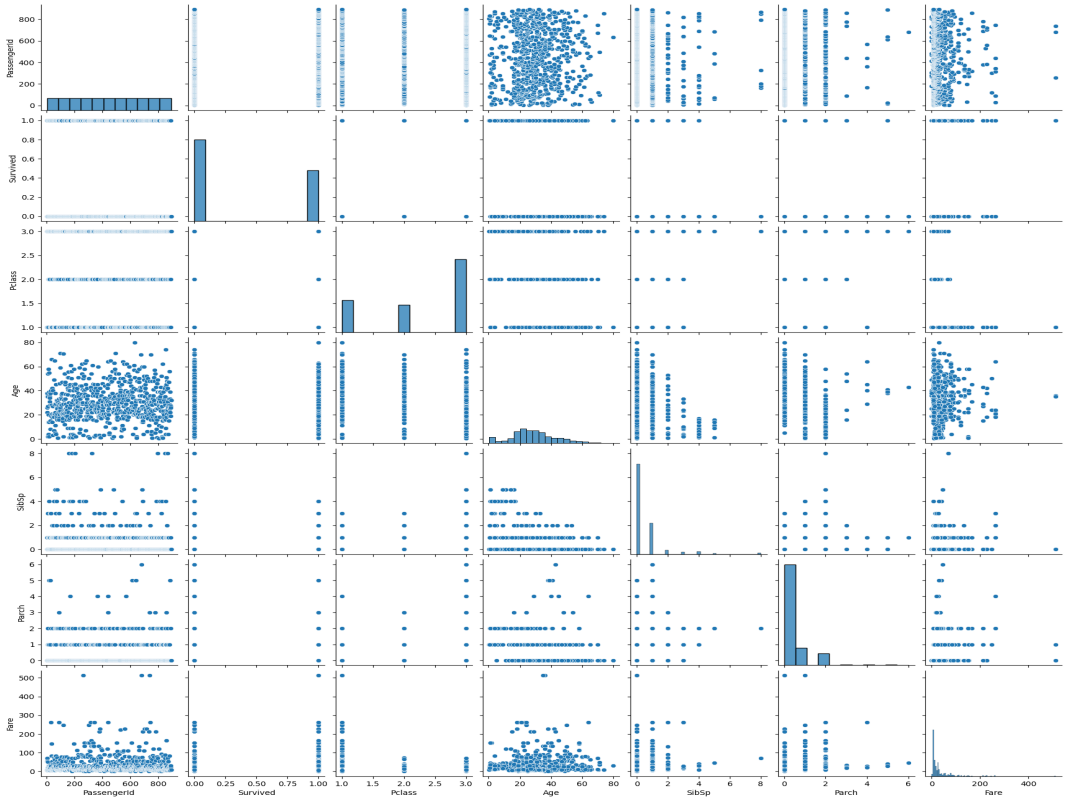
key findings: ■ Age Distribution Highlights - **Dominant Age Group:** The highest concentration of individuals falls between 20 and 30 years old, with the peak count around 90. This suggests a youthful population or sample. - **Right-Skewed Distribution:** The histogram shows a gradual decline in counts as age increases, indicating fewer older individuals. This skewness is typical in datasets like customer demographics, student populations, or early-career professionals. - **Sparse Older Population:** Very few individuals are above 70 years, which could imply limited representation of senior citizens or retirees in this dataset.

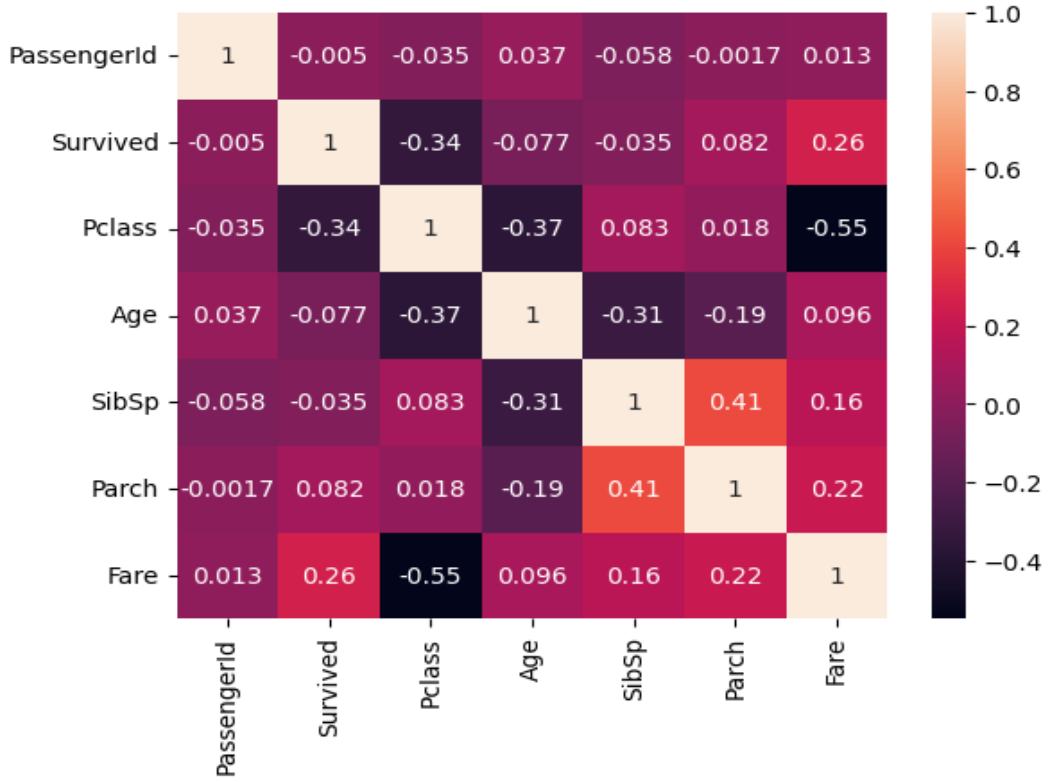
key findings: ■ Distribution Patterns - **Fare Concentration:** Most individuals paid fares below 100, regardless of age. This suggests a standard fare range for the majority, possibly due to fixed pricing or economic constraints. - **Outliers:** A few individuals—across various age groups—paid over 200, with some exceeding 500. These could represent premium services, VIP tickets, or special accommodations. - **Age Spread:** Fare values are scattered across all age groups, indicating no strong linear correlation between age and fare. Young and old individuals alike paid both low and high fares.

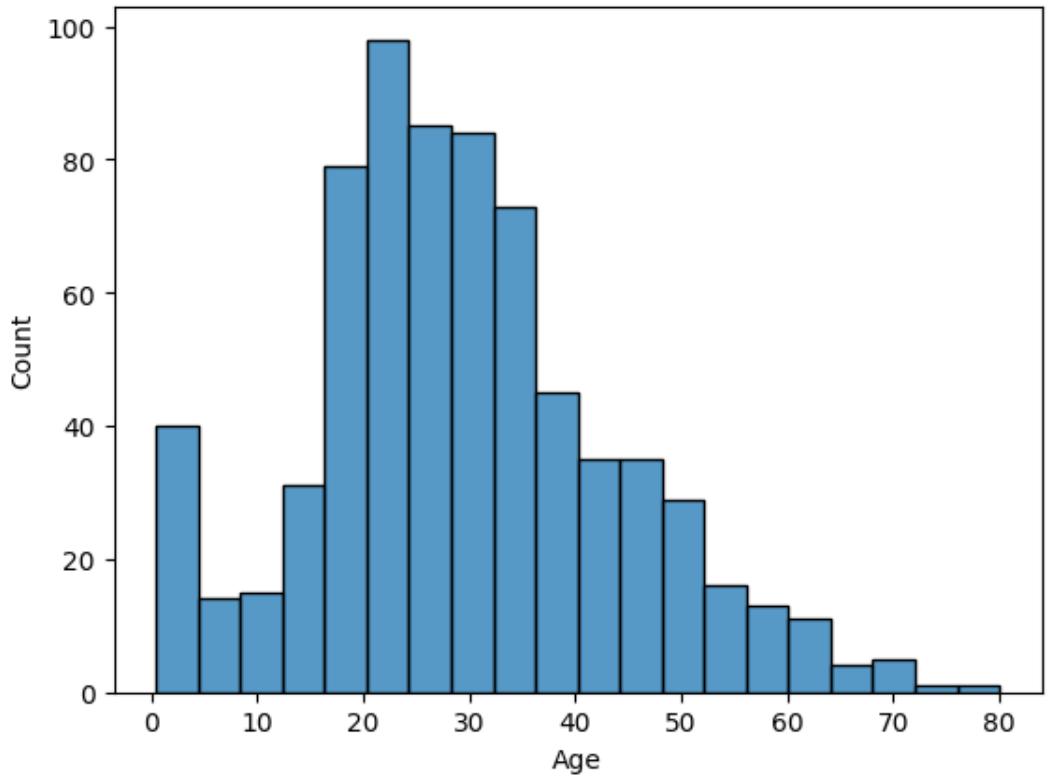
key findings: ■ Survival-Based Age Distribution - **Median Age Difference:** Survivors (label "1") have a slightly higher median age than non-survivors (label "0"). This challenges the assumption that younger individuals were more likely to survive. - **Spread & Variability:** - Both groups show a similar interquartile range (IQR), suggesting comparable age dispersion. - Non-survivors have a wider overall age range, with more extreme outliers—indicating that individuals of all ages were affected. - **Outliers:** Both categories include outliers, but non-survivors show more variability, especially at the lower and upper ends of the age spectrum.

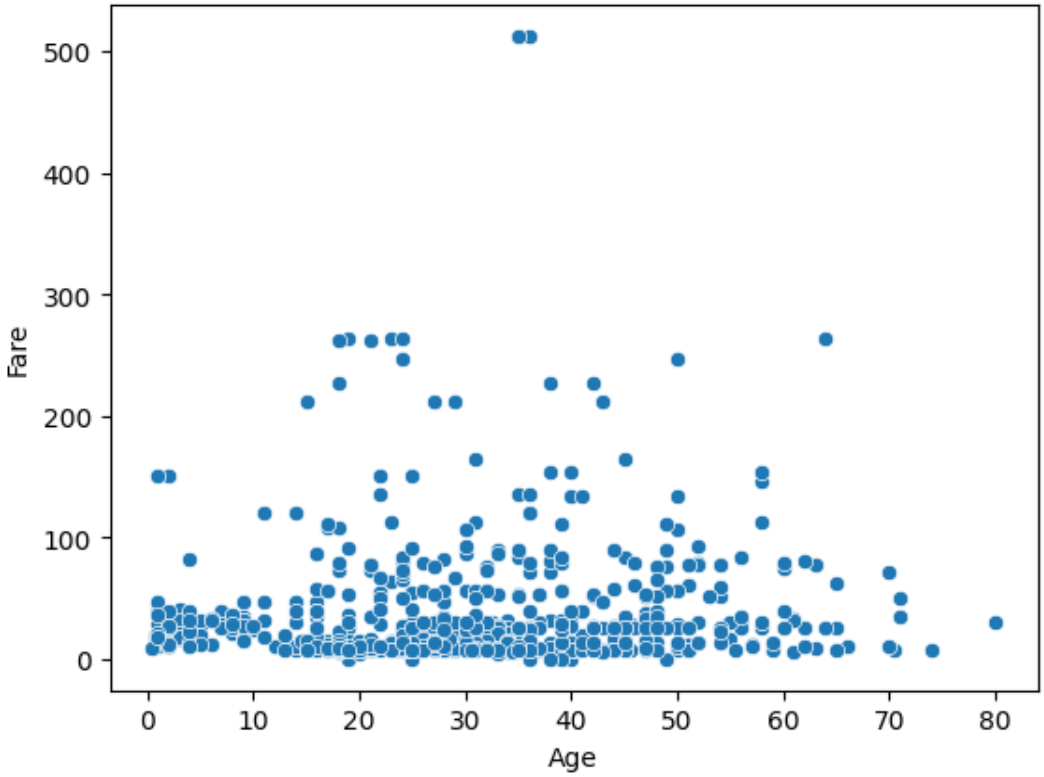
Visualizations

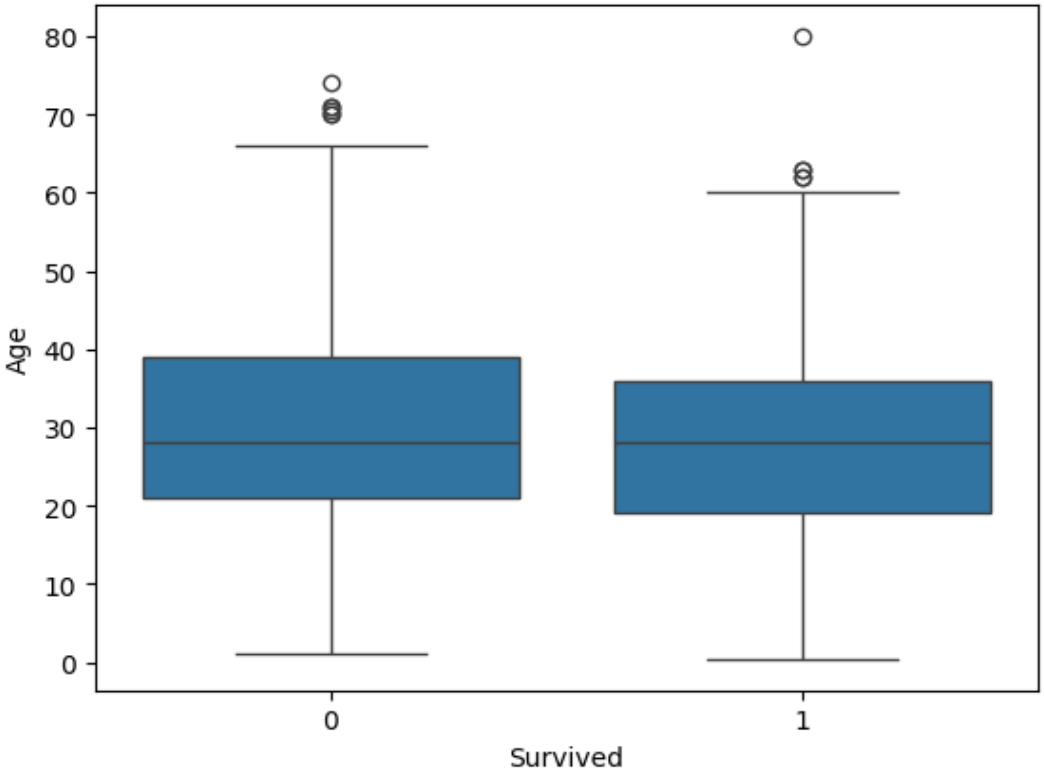





PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1 -0.005 -0.035 0.037 -0.058 -0.0017 0.013
-0.005 1 -0.34 -0.077 -0.035 0.082 0.26
-0.035 -0.34 1 -0.37 0.083 0.018 -0.55
0.037 -0.077 -0.37 1 -0.31 -0.19 0.096
-0.058 -0.035 0.083 -0.31 1 0.41 0.16
-0.0017 0.082 0.018 -0.19 0.41 1 0.22
0.013 0.26 -0.55 0.096 0.16 0.22 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Count
Age
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80


Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80


Age
Survived
0
10
20
30
40
50
60
70
80
0
1