Exploring Survival Patterns on the Titanic

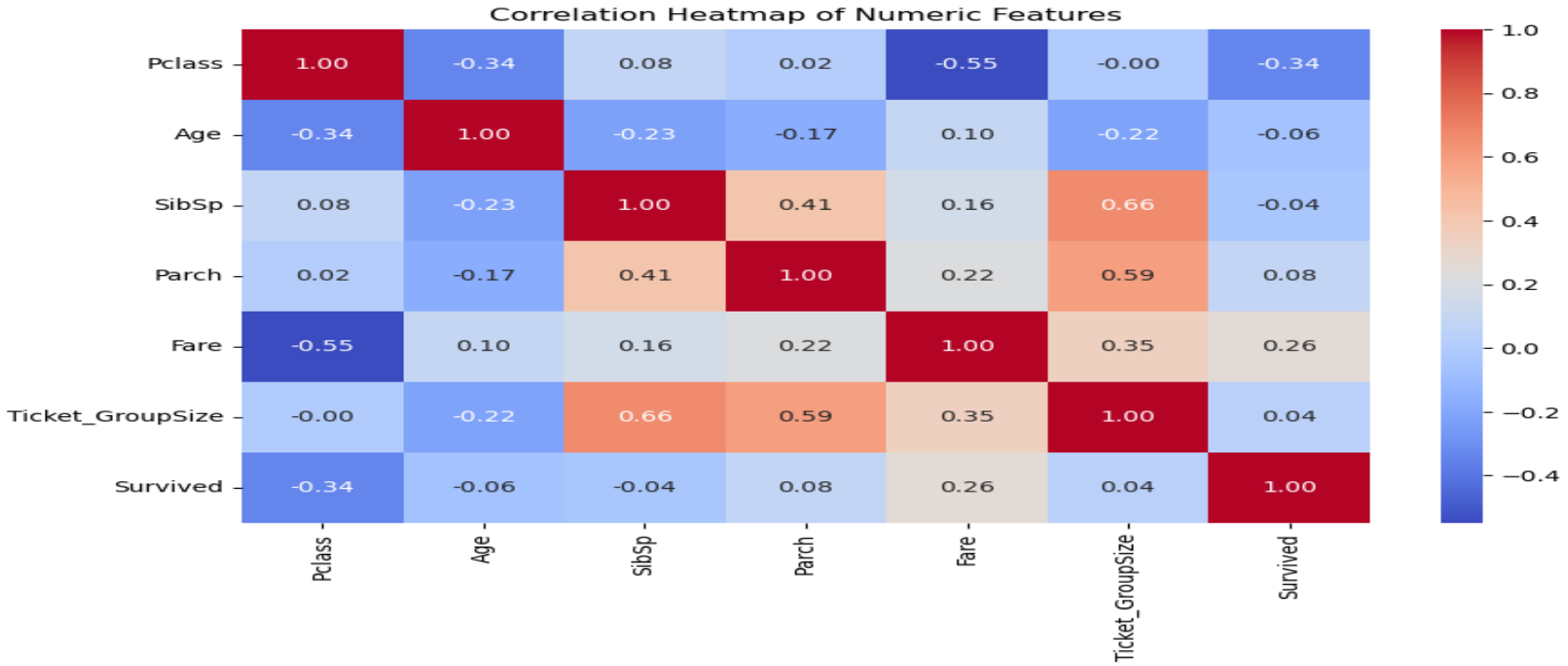
- **Pair Plot :**

- Socio-economic Status is Key for Survival (Pclass vs Fare):The diagonal plots and the scatter plots for Pclass and Fare show the strongest separation between survivors (blue) and non-survivors (red). Survival is highest among 1st class (Pclass=1) passengers and those who paid higher Fares, reinforcing that socio-economic status was a primary predictor.
- Fare Distribution is Driven by Class: The scatter plot of Fare vs Pclass (top-left plot) confirms that high fares are exclusively associated with Pclass 1 (the lowest Pclass number), while the vast majority of Pclass 3 passengers paid very low fares. This explains the strong correlation between Pclass and Fare.
- Age is Less Separating: The Age distribution plots show a large overlap between survivors and non-survivors, with the majority of both groups being young to middle-aged adults. While a slightly higher concentration of survivors appears among the youngest children, age alone is not a strong single differentiator across the whole population.
- Family Size Variables are Correlated but Complex: The plots involving SibSp (siblings/spouses), Parch (parents/children), and Ticket_GroupSize show that these family variables are highly correlated with each other. The relationship of these variables with survival is complex, with an intermediate group size appearing to have a slightly better survival chance than traveling alone or in very large groups.


Pairplot of Numeric Features by Survival
Pclass
Age
SibSp
Parch
Fare
Ticket_GroupSize
Survived
0
1

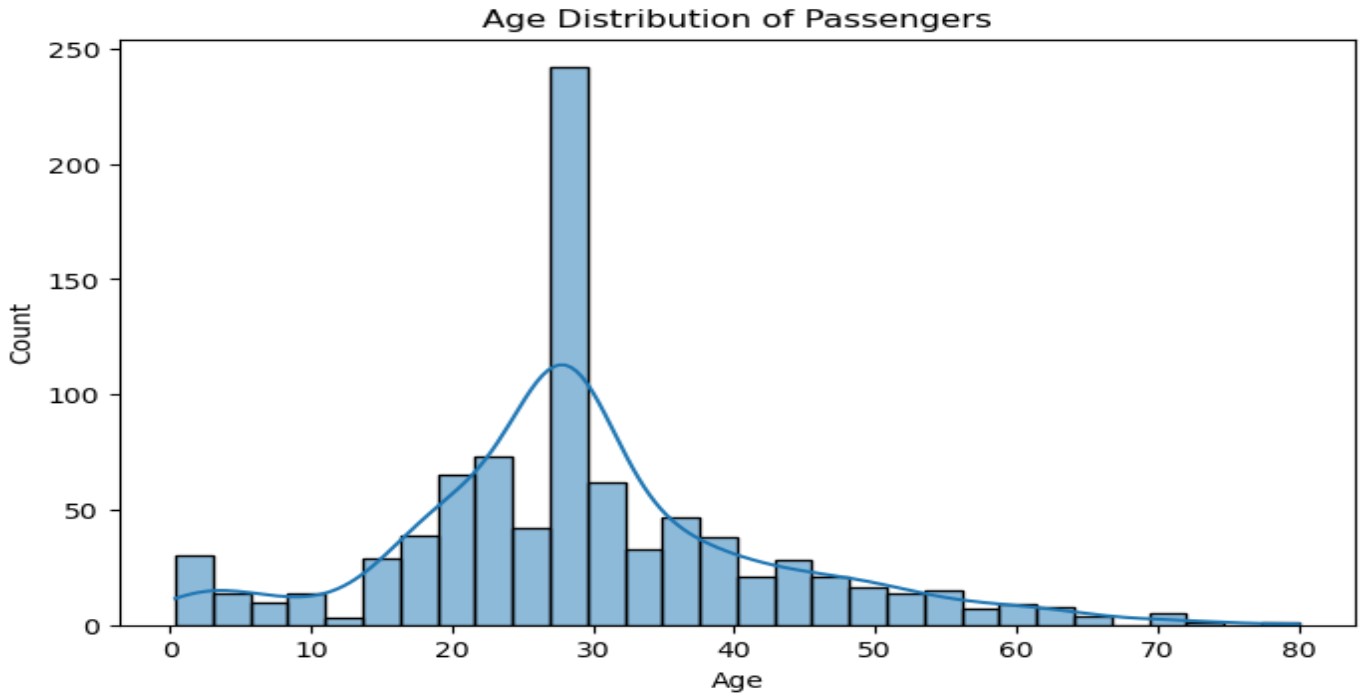
➢ Heatmap :

- The correlation heatmap quantifies the linear relationships between the numeric features. The most critical observation is the strong inverse correlation between Pclass and Survived (-0.34), meaning lower class numbers (i.e., 1st class) are associated with a higher survival rate. This is supported by a moderate positive correlation between Fare and Survived (0.26). Furthermore, the plot highlights a strong negative correlation between Pclass and Fare (-0.55), as expected, confirming that higher-class passengers paid significantly more. Age (-0.06) shows a negligible correlation with survival.



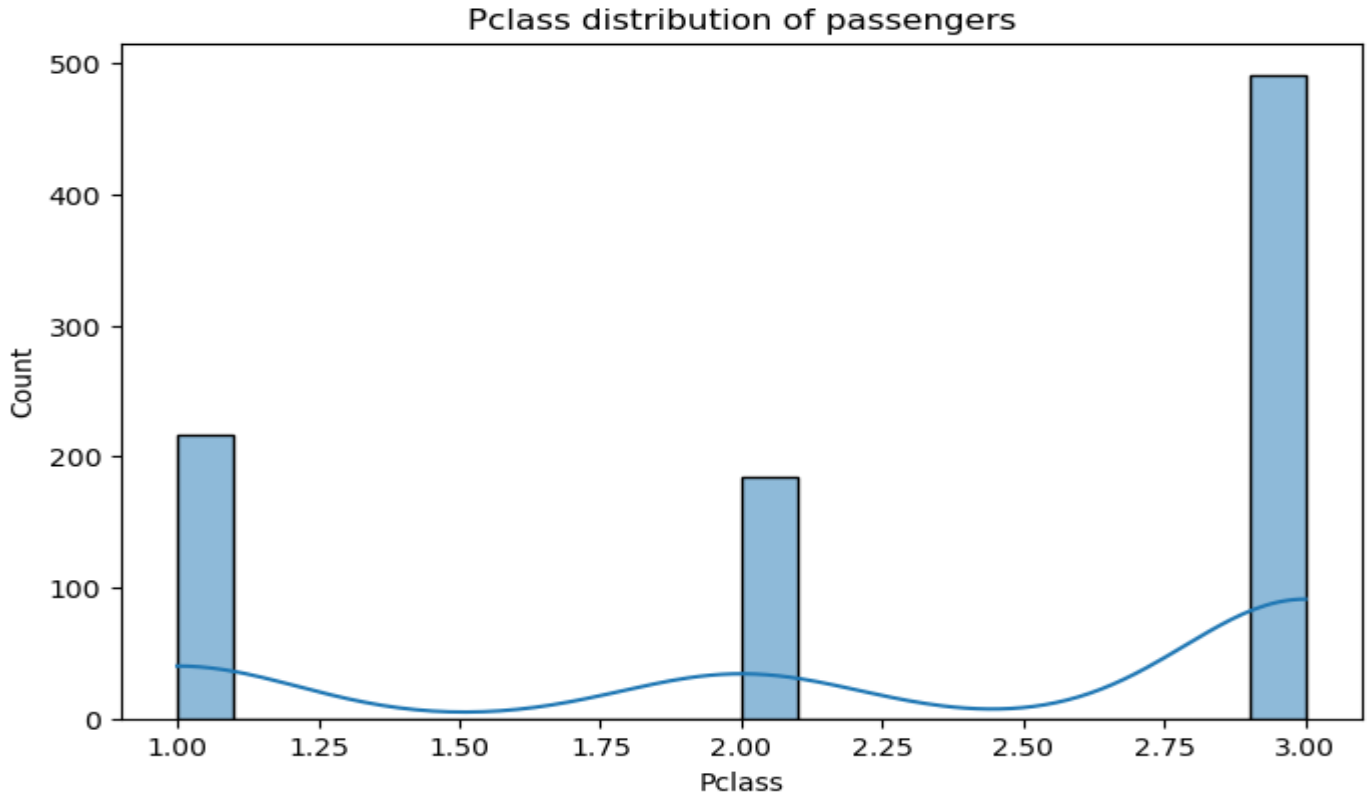
➢ Histogram(Age Distribution of Passengers):

- The Age Distribution of Passengers. The primary observation is that the passenger age data is concentrated around a strong peak between 25 and 30 years old, indicating the majority of individuals aboard were young adults. The distribution is generally bell-shaped but slightly skewed to the right, with a long tail extending toward older ages (up to 80). There is also a noticeable smaller cluster of very young children and infants (ages 0-5), confirming that the ship carried passengers across the entire age spectrum.



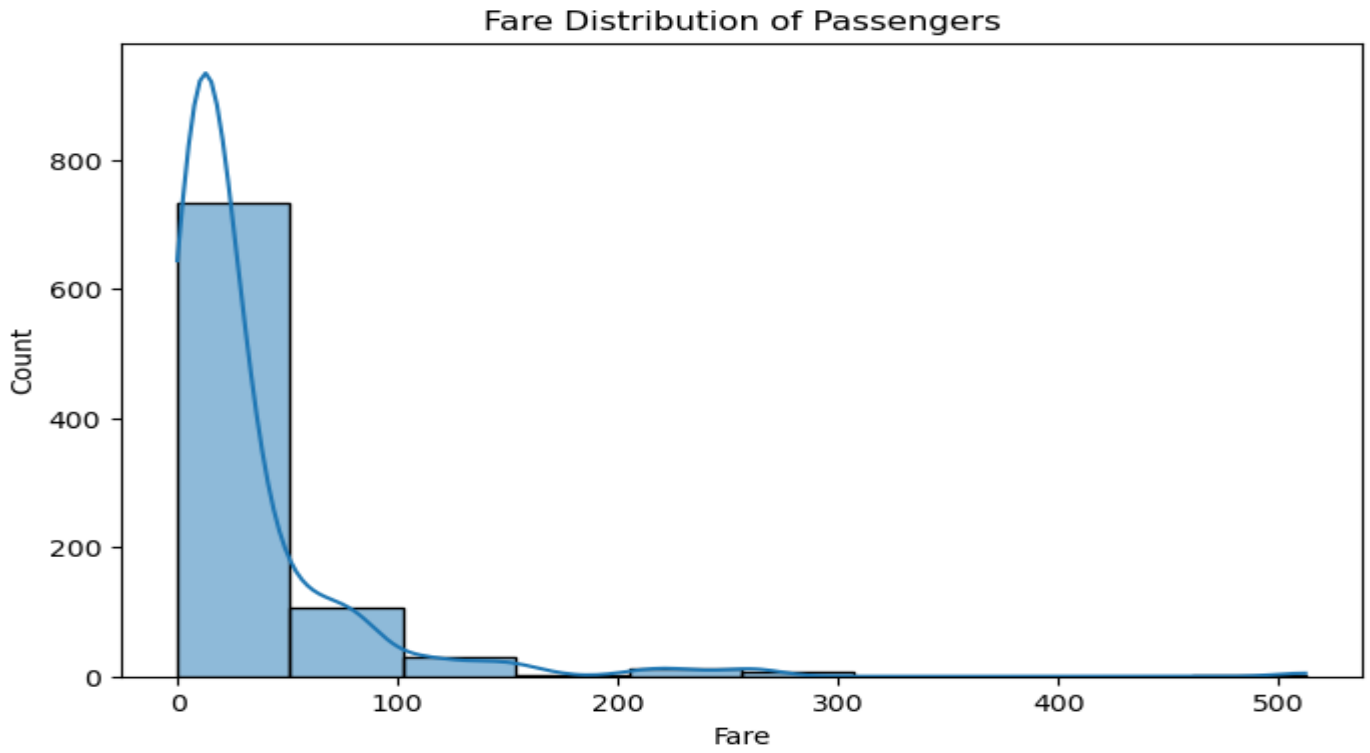
➢ Pclass Distribution of Passengers :

- The chart shows that the passenger distribution is highly uneven across classes. The overwhelming majority of passengers belonged to 3rd class (Pclass=3), with a count nearing 500. 1st class (Pclass=1) had the second-highest count, while 2nd class (Pclass=2) had the lowest, indicating that the dataset is dominated by the lowest socio-economic status group.



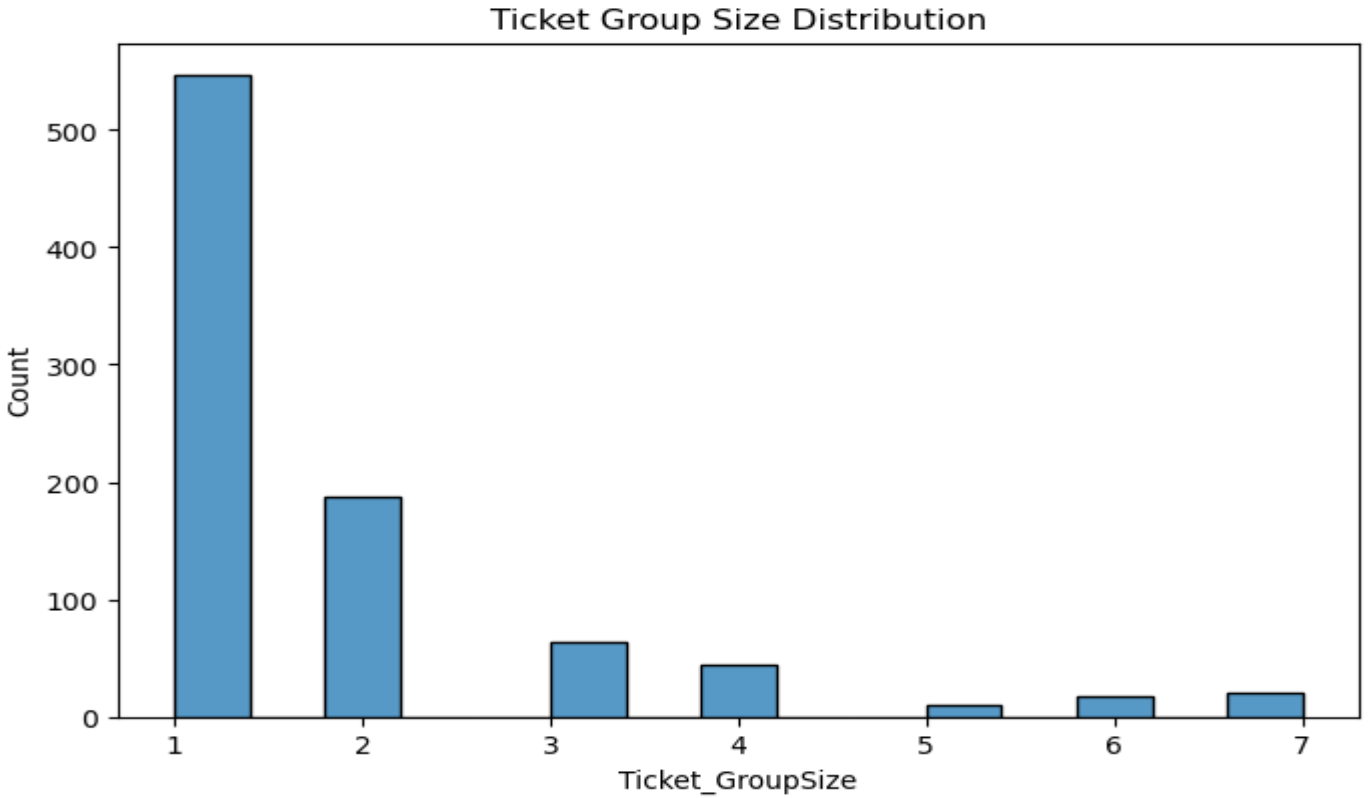
➢ Fare Distribution of Passengers :

- The primary observation from this histogram is that the fare distribution is extremely right-skewed, with the vast majority of passengers clustered in the lowest fare bracket (between 0 and approximately 50). The first bar dominates the plot, indicating a massive count of low-fare tickets. This suggests most passengers were traveling in the cheaper classes. The distribution has a very long, attenuated tail, showing that a small minority of passengers paid exceptionally high fares (reaching over 500).



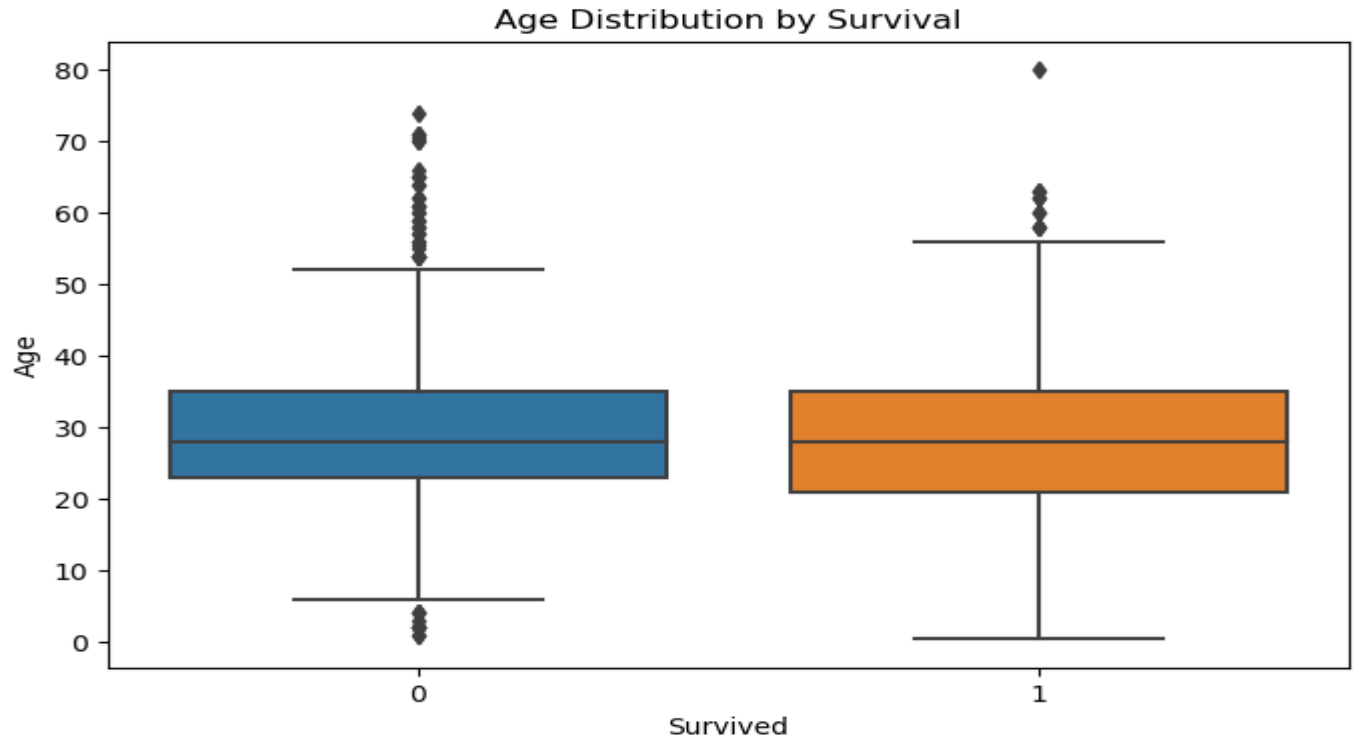
➢ Ticket Group Size Distribution:

- This bar chart clearly shows that the **vast majority of passengers traveled alone** (Ticket Group Size = 1), with this group having a count exceeding 500. The number of passengers drops sharply for groups of 2 and then decreases rapidly for larger group sizes. This confirms that solo travelers were the most common demographic aboard the ship.



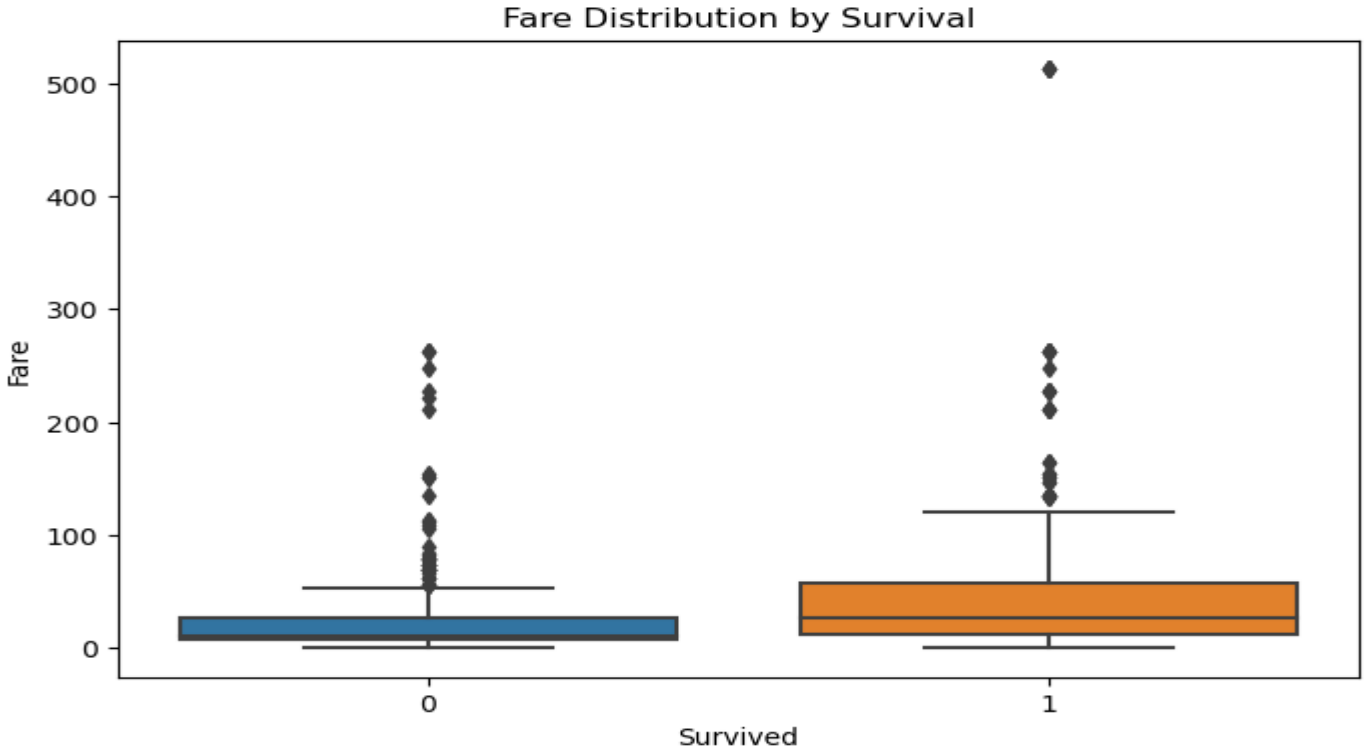
➢ Age Distribution by Survival:

- The box plot displays the age distribution for two groups: those who did not survive (Blue) and those who survived (Orange). We observe that the median age (the line inside the box) is very similar for both groups, around 28-29 years. However, the interquartile range (IQR), represented by the box's length, appears slightly wider for the non-survivors, suggesting a greater spread of ages in the middle 50%. The non-survivors also show a greater number of outliers at older ages, particularly above 65, though both groups have outliers at various ages. Overall, the visual suggests that age alone was not a dominant predictor of survival, as the central age tendencies are nearly identical.



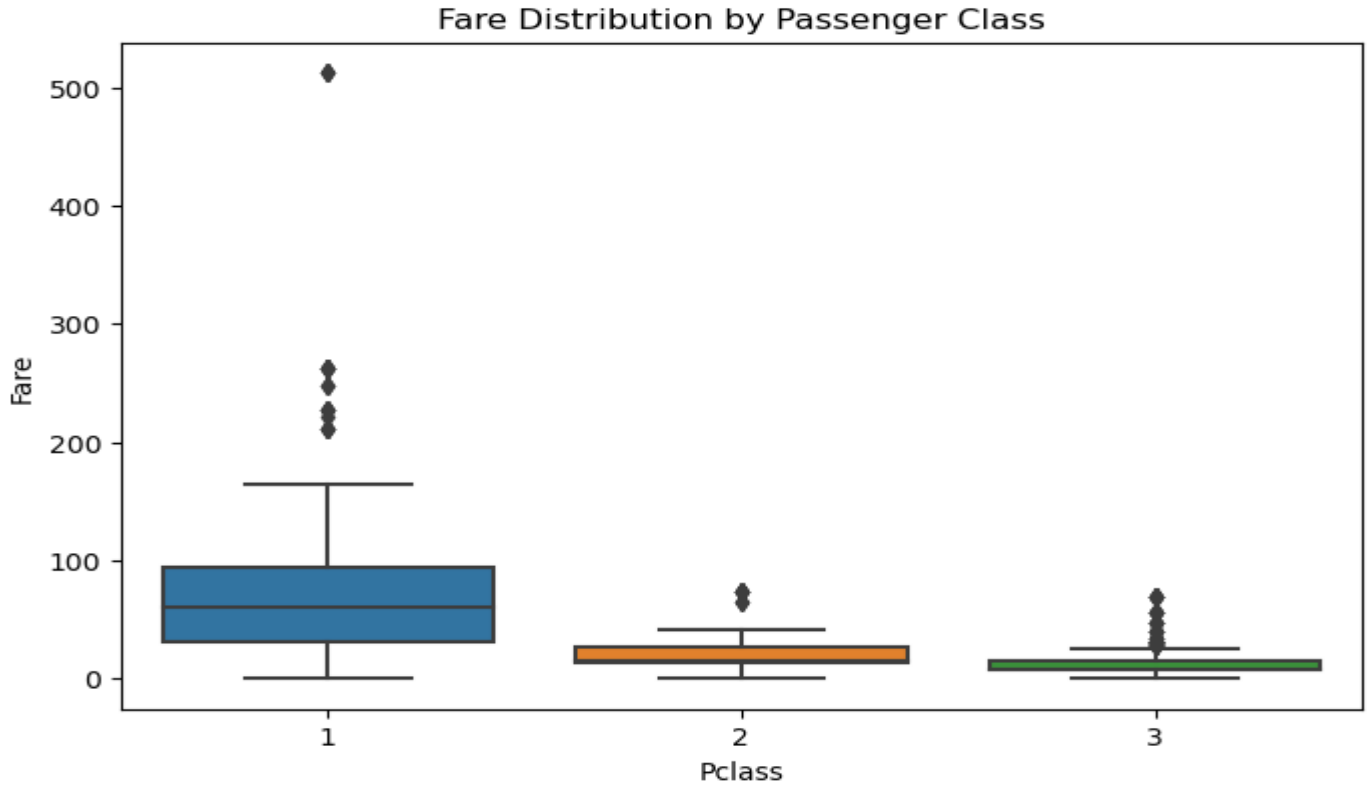
➢ Fare Distribution by Survival:

- The box plot compares the fare distribution for those who did not survive and those who survived . A key observation is that the median fare (the line inside the box) is significantly higher for the survivors (orange box), suggesting a positive correlation between fare paid and the likelihood of survival. Specifically, the median fare for survivors is around 30, while for non-survivors it's closer to 10-15. Both distributions are heavily right-skewed, as indicated by the many outliers (diamonds) showing very high fares, though the interquartile range (the box) is noticeably wider and higher for the surviving group. This visual evidence implies that passengers who paid higher fares, likely traveling in higher classes, had better survival odds.



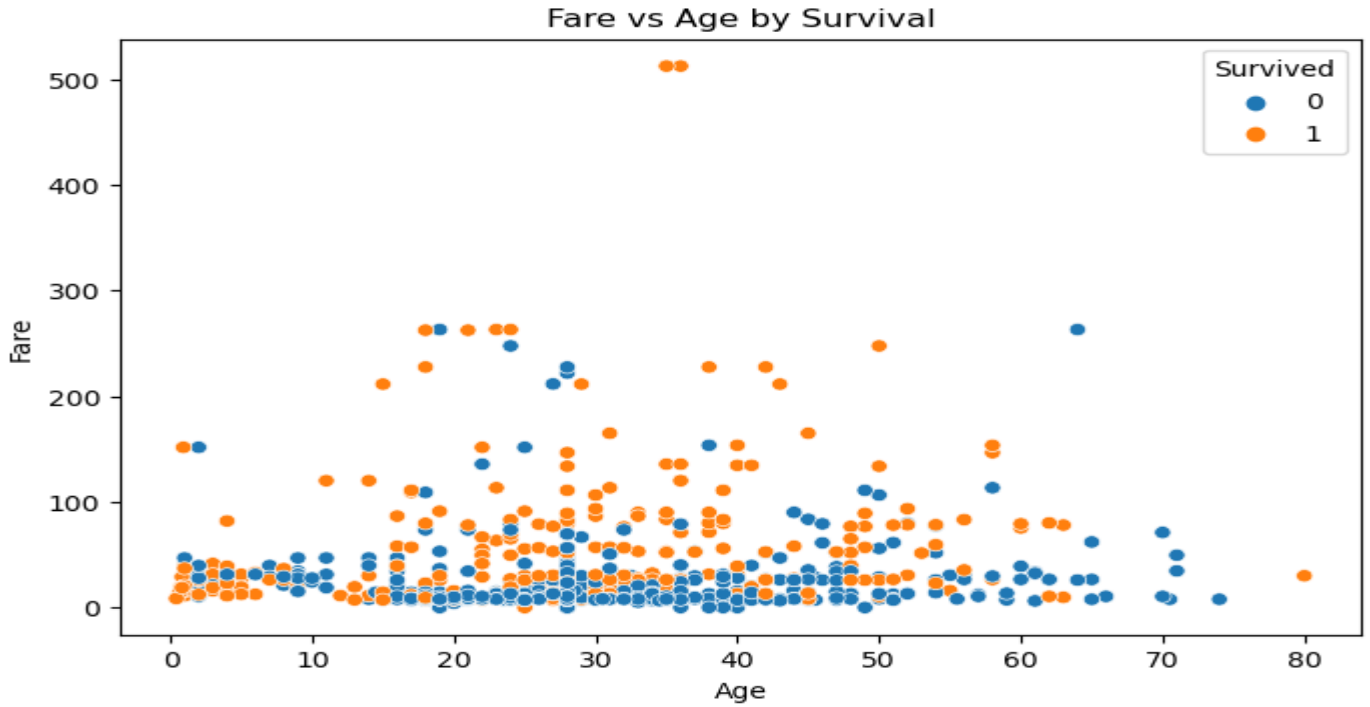
➢ Fare Distribution by Passenger Class:

- The Fare Distribution by Passenger Class (Pclass), which is categorized as 1st, 2nd, and 3rd class. The most significant observation is the direct relationship between passenger class and fare: 1st Class (1) passengers paid the highest fares, evidenced by the highest median (around 60), widest interquartile range (IQR), and numerous high-value outliers. 2nd Class (2) fares are distinctly lower with a median around 20, and a much smaller spread. Finally, 3rd Class (3) fares are the lowest, with the median and IQR centered near the 10-15 mark, indicating minimal variation in cost. This visual confirms that fare is a strong determinant of passenger class, with a clear hierarchy from 1st to 3rd class.



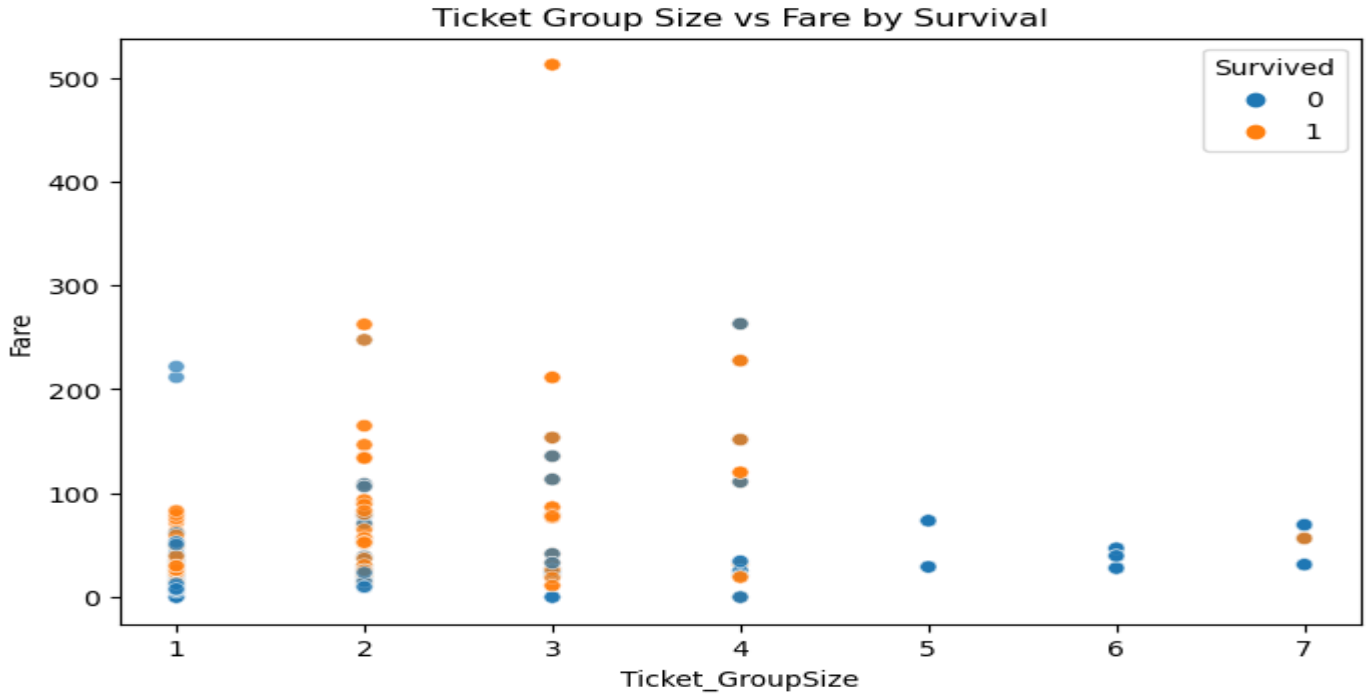
➢ Fare VS Age by Survival:

- The scatter plot visualizes Fare vs. Age, categorized by Survival (blue for non-survivors, orange for survivors). We observe a general cluster of passengers at lower fares (below 100) and younger ages (under 40), where both survivors and non-survivors are heavily mixed, indicating no clear trend in this dense region. However, at higher fares (above 100), there appears to be a noticeably higher proportion of survivors (orange points), suggesting that a higher fare was beneficial for survival across various ages. Similarly, among the oldest passengers (65 and above), the few survivors (orange) tended to be those who paid relatively higher fares, while non-survivors (blue) were more spread out in fare. This reinforces the idea that Fare was a stronger differentiator for survival than Age.



➢ Ticket Group VS Fare by Survival:

- The scatter plot illustrates the relationship between Ticket Group Size, Fare, and Survival. A distinct pattern is observed, particularly at a group size of 1 (solo travelers), where survivors (orange) and non-survivors (blue) are mixed across a wide range of fares, but the highest fare is associated with survival. For group sizes 2, 3, and 4, there's a clear trend where the individuals who survived (orange points) generally correspond to the higher fares within that group size, reinforcing the fare-survival link. As the Ticket Group Size increases to 5, 6, and 7, the number of data points significantly decreases, and the survival points become much scarcer. The overall distribution shows that survival is not solely dependent on group size but is strongly modulated by the fare paid within most group categories.



➢ Summary of Survival Factors:

The visualizations collectively suggest that financial status (Fare/Pclass) was a much stronger predictor of survival than age.

➢ The Role of Age and Survival

- Age was not a dominant predictor: The box plot of "Age Distribution by Survival" showed that the median age for both survivors and non-survivors was nearly identical (around 28-29 years).
- Outliers: While both groups had age outliers, there was a slightly wider range of ages and a greater number of older outliers among those who did not survive.

➢ The Dominance of Fare and Class

- Fare Correlates Strongly with Survival: The "Fare Distribution by Survival" box plot revealed a significant difference: the median fare paid by survivors (Group 1) was substantially higher (around 30) than the median fare paid by non-survivors (around 10-15).
- Fare vs. Age: The "Fare vs Age by Survival" scatter plot reinforced this, showing that at higher fares (above 100), the points representing survivors (orange) heavily outnumbered the non-survivors, regardless of the passenger's age.
- Fare and Class Hierarchy: The "Fare Distribution by Passenger Class" box plot confirmed the clear financial hierarchy: 1st Class paid the highest fares, followed by 2nd Class, with 3rd Class paying the lowest, which directly links fare and class to the previously observed survival advantage.

➢ Ticket Group Size and Fare

- Higher Fare = Higher Survival in Groups: The "Ticket Group Size vs Fare by Survival" scatter plot indicated that for passengers traveling in small groups (size 2, 3, and 4), those who survived consistently paid the highest fares within those group categories. This suggests that even when traveling together, the cost of the ticket (and thus class) was the crucial survival advantage.