Exploratory Data Analysis (EDA) Summary

4. Identify patterns, trends, or anomalies in the data.

Patterns

Looking at the Titanic dataset, several consistent patterns emerge. For instance, most passengers belonged to the lower class (Pclass 3), pointing to a large portion of the ship's population traveling economically. The average age was around 29.7 years, suggesting that the passenger list was made up largely of young adults. Additionally, the majority of passengers traveled alone or with very few family members, as shown by the low average values for both SibSp (siblings/spouses aboard) and Parch (parents/children aboard). These patterns give us a good sense of the general makeup of the Titanic's passengers—mostly young individuals in lower classes traveling with few companions.

Trends

One of the most noticeable trends is the survival rate—only about 38% of the passengers survived the disaster. This indicates a clear imbalance in the target variable, which is important to consider during model training. Another trend is seen in the Fare data. Most passengers paid relatively low fares, but the average fare is pushed upward due to a few very expensive tickets, suggesting a highly skewed distribution. This aligns with historical context, where a small number of wealthy individuals paid premium prices for first-class cabins, while the majority paid significantly less.

Anomalies

The dataset also includes a few problems that need to be addressed before modeling. For one, there are 177 missing values in the Age column, which means age imputation will be necessary. The Fare column includes extreme values—some passengers paid over 500, while others paid nothing at all. These outliers and the heavy right skew in Fare distribution could distort model training if not handled carefully. Similarly, the presence of passengers with unusually high numbers in SibSp and Parch (up to 8 and 6 respectively) could represent rare cases or data errors that may need to be reviewed.

5. Make basic feature-level inferences from visuals.

1. Age Distribution (Histogram)

The age histogram shows a moderately right-skewed distribution, where most passengers fall between 20 and 40 years. There's a visible peak around 20–30 years, suggesting that a large number of passengers were young adults. Children and elderly passengers were fewer, with some outliers above 70.

2. Fare Distribution (Histogram)

The fare distribution is highly right-skewed, indicating that most passengers paid low fares, likely in third class. However, a few passengers paid very high fares (over 500), creating a long tail and confirming the presence of outliers.

3. Age vs. Survival (Boxplot)

The boxplot comparing age and survival status shows that median ages are similar between survivors and non-survivors. However, survivors show a slightly wider range among younger ages. This suggests that age may not have a strong influence on survival overall, though younger passengers had slightly better chances.

4. Fare vs. Passenger Class (Boxplot)

Fare clearly varies by class. First-class passengers paid significantly higher fares, with wide variability and outliers. Second- and third-class passengers paid much less with less variance. This indicates that Fare is a strong proxy for Pclass, and both can be important for predicting survival.

5. Correlation Matrix

The correlation matrix shows:

- Fare and Survived have a positive correlation (0.26), indicating higher fare-paying passengers were more likely to survive.
- Age and Survived show a weak negative correlation (-0.08).
- SibSp and Parch are positively correlated (0.41), which makes sense as people with more siblings often traveled with parents too.
- Most features are weakly correlated with each other and with survival, but Fare has the strongest relation among them.

6. Pairplot (Scatter + Distribution Plots)

The pairplot reveals:

- Survivors (orange) cluster more around lower SibSp and Parch values, but there are a few exceptions.
- Survivors appear to be spread more evenly across age and fare, but higher fares have more survivors.
- Again, we see Fare’s influence on survival more visibly, confirming that wealthy passengers had better chances.