Exploratory Data Analysis (EDA) Report

Dataset Context: Titanic Dataset (Passenger data)

1. 🎟️ Fare Distribution by Passenger Class

(Box Plot - box plot.png)

Key Observations:

- **Pclass 1** passengers have the **highest fare distribution** with a median much higher than Pclass 2 and 3.
- **Presence of outliers**: Especially in Pclass 1, where fares exceed 500.
- Fares **decrease** from Pclass 1 → Pclass 3, indicating a **strong relationship between class and fare**.

Insights:

- Passenger class is a **strong indicator of fare**.
- Outliers suggest **premium ticket holders** in 1st class.

---

2. 🧊 Correlation Heatmap

(Heatmap - heatmap.png)

Key Observations:

- **Fare and Pclass**: High negative correlation (-0.55), confirms higher-class passengers paid more.
- **Survived and Pclass**: Negative correlation (-0.34), suggesting survival rates were better in higher classes.
- **Fare and Survived**: Positive correlation (0.26), indicating higher fare passengers had higher chances of survival.
- **SibSp & Parch**: Positive correlation (0.41), meaning families often traveled together.

Insights:

- **Pclass, Fare, and Survived** are interlinked and **impact survival rate**.
- Heatmaps help spot potential **feature dependencies** or multicollinearity issues.

3. 📈 Distribution of Age

(Histogram - histogram plots.png)

Key Observations:

- **Right-skewed distribution**: More young passengers; fewer elderly ones.
- Peak age group is **20–30 years**.
- Distribution includes infants and children under 10.

Insights:

- Skewness suggests a **young majority**, useful when segmenting by age groups.
- Tail on the right might affect models sensitive to normal distribution (e.g., linear regression).

🔍 **Final Summary:**

Feature	Insight
Fare	Strongly influenced by passenger class; outliers exist
Pclass	Strong indicator of both fare paid and survival probability
Age	Right-skewed; mostly young passengers
Survived	Influenced by Pclass, Fare, and Age
Correlations	Pclass & Fare (-ve), Fare & Survived (+ve), SibSp & Parch (+ve)

Conclusion:

- Visuals reveal **valuable feature relationships**, trends, and anomalies (e.g., outliers, skewness).
- This analysis can guide **feature selection**, **data transformation**, and **model choice** in ML.