Exploratory Data Analysis (EDA) Report

Dataset: Titanic Dataset (Kaggle)

# 1. Summary

This report conducts an in-depth Exploratory Data Analysis (EDA) on the Titanic dataset. The objective is to uncover trends, correlations, and anomalies to better understand the factors affecting passengers' survival chances aboard the Titanic.

# 2. Objective

Perform statistical and graphical exploration.

Identify influential features.

Detect patterns and anomalies.

Address missing data and skewed distributions.

# 3. Dataset Description

Source: Kaggle Titanic Competition

File: train.csv

Observations: 891 passengers

Features: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

# 4. Data Cleaning

Missing values in Age, Cabin, and Embarked detected.

Cabin column largely incomplete and dropped.

Age and embarked missing values imputed.

# 5. Exploratory Data Analysis (EDA)

5.1 Univariate Analysis:

Majority of passengers aged 20-40.

38% passengers survived.

Fare distribution highly skewed.

Most passengers were males.

5.2 Bivariate Analysis:

Females had a higher survival rate.

1st class passengers had higher survival rates.

Children had higher survival chances.

Cherbourg port passengers had slightly higher survival.

5.3 Multivariate Analysis:

Fare positively correlated with survival.

Pclass negatively correlated with survival.

Age showed weak correlation with survival.

# 6. Visualizations Used

- Histograms

- Countplots

- Boxplots

- Barplots

- Heatmaps

- Pairplots

# 7. Handling Missing Values

Age imputed with median value (28).

Embarked imputed with mode ('S').

Cabin column dropped due to heavy missingness.

# 8. Handling Skewness

Fare skewness (4.8) reduced using log transformation (np.log1p).

# 9. Key Insights

Females and children survived more.

1st class passengers had better survival rates.

Higher fare often indicated better survival probability.

Embarkation point had some effect on survival.

# 10. Conclusion

EDA revealed strong insights on key survival factors aboard the Titanic. The cleaned and processed dataset is now ready for predictive modeling steps.

# 11. Recommendations

Engineer features from Name and Family size.

Use more sophisticated age imputation (like regression).

Prioritize Sex, Pclass, and Fare during model building.