Exploratory Data Analysis (EDA) of a Data Set Report

Description of EDA process

1. Objective

The primary goal of this project is to perform a comprehensive exploratory data analysis on the dataset provided (train.csv). The aim is to:

- Understand the structure and composition of the data.
- Identify important patterns, trends, and anomalies.
- Detect relationships between different variables.
- Prepare the dataset for subsequent machine learning or statistical modeling tasks.

2. Dataset Overview

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	

Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q =Queenstown, S =Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...

Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

3. Data Cleaning and Preprocessing

3.1 Handling Missing Values

- A missing value analysis was conducted.
- Columns with missing entries were identified.
- Potential strategies for handling missing data were considered, such as:
    - Filling with mean/median/mode (for numerical features).
    - Imputation with the most frequent category (for categorical features).
    - Dropping columns/rows if missing values were substantial.

3.2 Data Types Verification

- It was ensured that each feature had an appropriate data type.
- For example:
    - Categorical variables like Gender, Category, or Class were confirmed as object.
    - Continuous variables like Age, Salary, or Fare were confirmed as numerical (float64, int64).

3.3 Consistency Checks

- Duplicate rows were checked.
- Unique values per column were counted to identify:
    - Potential encoding needs.
    - Redundant features with very little variability.

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

The distribution of individual features was studied.

- **Numerical Features**:
    - Measures such as mean, median, standard deviation, min, and max were computed (describe() function).
    - Histograms plotted for features like Age, Fare, etc., to visualize their distribution.

    - Outlier detection through visualization (e.g., boxplots).
- **Categorical Features**:
    - Frequency tables (value counts) were generated.
    - Bar plots created to visualize the proportion of each category.
    - Analysis revealed dominant categories and minority classes.

Example:
If the feature is Gender, the bar plot may show a 65%-35% split between Male and Female.

4.2 Bivariate and Multivariate Analysis

Exploring relationships between two or more variables:

- **Categorical vs Numerical**:
    - Grouped the data based on categorical variables and computed aggregate statistics on numerical variables.
    - Example: Mean Fare by Passenger Class, average Salary by Department, etc.
- **Categorical vs Categorical**:
    - Cross-tabulation (pd.crosstab) to observe interaction between two categorical variables.
    - Stacked bar charts to visualize joint distributions.
- **Numerical vs Numerical**:
    - Correlation matrix computed using Pearson correlation.
    - Scatter plots between important numerical features to visually inspect linear/non-linear relationships.

4.3 Correlation Analysis

- A correlation heatmap was generated using seaborn.
- Key observations included:
    - Highly correlated feature pairs (positive or negative).
    - Features with little to no correlation with others, indicating possible irrelevance.

Interpretation of correlation coefficients:

- Close to +1: Strong positive correlation

- Close to -1: Strong negative correlation
- Close to 0: Little or no linear correlation

5. Data Visualization

Several visualization techniques were used to enhance understanding:

Type of Plot	Purpose	Features Visualized
Histogram	View distribution of numerical features	Age, Fare, etc.
Bar Plot	Compare category frequencies	Embarked, Sex, etc.
Heatmap	Visualize feature correlations	All numerical features
Box Plot	Detect outliers and spread	Fare, Age vs Class
Scatter Plot	Explore pairwise relationships	Age vs Fare, etc.

Visualizations revealed skewed distributions, outliers, and potential feature relationships important for modeling.

6. Key Findings

Based on the EDA, several important insights were gathered:

- **Data Imbalance**: Some categorical features were highly imbalanced, with one class dominating.
- **Missing Values**: Certain columns had significant missing values, which must be handled before modeling.
- **Outliers**: Some numerical features showed outliers, possibly influencing model performance if not treated.
- **Important Relationships**: Strong correlations between certain features suggest multicollinearity (requiring attention during model building).

- **Feature Importance Indication**: Preliminary grouping and aggregation suggested which features may be most influential in predicting outcomes.

7. Conclusion

The EDA provided a thorough understanding of the dataset's structure, quality, and underlying patterns. It laid a strong foundation for subsequent phases such as:

- Feature engineering (creating new features, transforming existing ones).
- Selection of appropriate machine learning models.
- Dealing with class imbalance, missing values, and outliers.
- Performing further advanced analysis like feature selection and dimensionality reduction.

Findings after EDA on the dataset

1. Data Overview

- Dataset loaded from train.csv.
- Basic inspections like head() showed the dataset has typical Titanic features such as Survived, Pclass, Sex, Age, Fare, etc.

2. Data Information

- **train_df.info()** revealed:
    - Total entries: 891 rows.
    - Some columns have missing values.
    - Mixture of data types: integer, float, and object (categorical).

3. Summary Statistics

- **train_df.describe()** showed:
    - Age range: ~0.42 to 80 years.
    - Fare had a wide spread (min=0, max=512).

- Mean Age was around 29.7 years.
- Ticket prices were highly skewed, with most fares lower and few extremely high.

4. Missing Values Analysis

- **train_df.isnull().sum()** revealed:
  - **Age** has **177 missing values**.
  - **Cabin** has **687 missing values** (high percentage, around 77% missing).
  - **Embarked** has **2 missing values**.
- **Total Missing Values**: 866 across the entire dataset.

5. Target Variable (Survived) Distribution

- **train_df.Survived.value_counts()**:
  - 0: 549 passengers (did not survive).
  - 1: 342 passengers (survived).
- Conclusion: The dataset is **imbalanced** with more non-survivors than survivors.

6. Feature Distributions (Univariate Analysis)

- **Histograms** for Age and Fare:
  - Age is approximately normally distributed but slightly skewed to the right (more young people).
  - Fare distribution is highly right-skewed (majority paid low fare, few paid very high fares).
- **Countplots** for categorical variables (Survived, Pclass, Sex, Embarked):
  - **Pclass**:
    - Class 3 had the highest number of passengers.
  - **Sex**:
    - More males than females in the dataset.
  - **Embarked**:
    - Most passengers embarked from port 'S' (Southampton).

7. Bivariate Analysis (Relationships with Survived)

- **Pclass vs Survived**:
    - Survival rate decreases as the class number increases.
    - 1st class passengers had the highest survival rate.
- **Sex vs Survived**:
    - Females had a significantly higher survival rate than males.
- **Embarked vs Survived**:
    - Passengers who embarked at 'C' (Cherbourg) had a higher survival rate compared to 'S' and 'Q'.

8. Multivariate Visualization

- **Heatmap of Correlations** (not shown fully but implied):
    - Strong correlations noted:
        - Fare positively correlated with Pclass.
        - Survived correlated with Sex and Pclass.