

Project 1: Data Exploration, Cleaning, and Visualization

Student Name: Ahmed Elnoor Adam Ahmed

Course: Practical Machine Learning and Data Exploration (First Trimester 2026)

Date: 26 April 2026

1. Project Background

This project focuses on the fundamental steps of data preprocessing: handling missing information, identifying and managing outliers, and visualizing key patterns in the data. The Titanic Dataset from Kaggle was used for this analysis.

2. Data Cleaning process

2.1. Handling Missing Values

The dataset contained several missing values which were handled as follows:

- Age: Missing values were filled using the Median age of the passengers.
- Embarked: Missing values were filled with the Mode (most frequent docking port).
- Cabin: Due to the high percentage of missing values, these were labeled as "Unknown" to preserve the record without introducing bias.

2.2. Outlier Management (IQR Method)

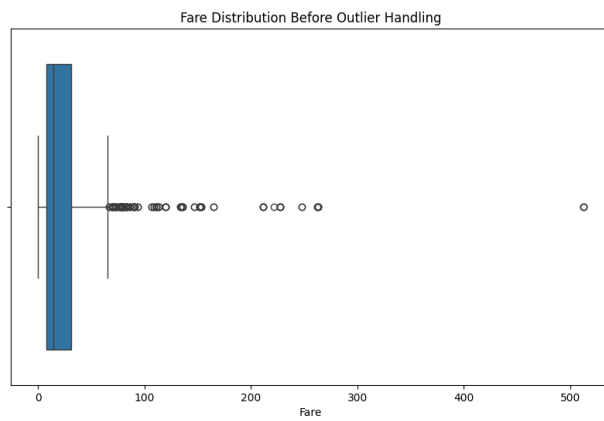
The 'Fare' column exhibited significant outliers. To prevent these from skewing the future model results, the Interquartile Range (IQR) method was applied:

- Upper Bound Identification: Any fare above
 - $Q3 + 1.5 \times IQR$
 - $Q3 + 1.5 \times IQR$ was considered an outlier.
 - Capping: Outliers were capped to the upper and lower bounds rather than being deleted, ensuring we retained the data records.
-

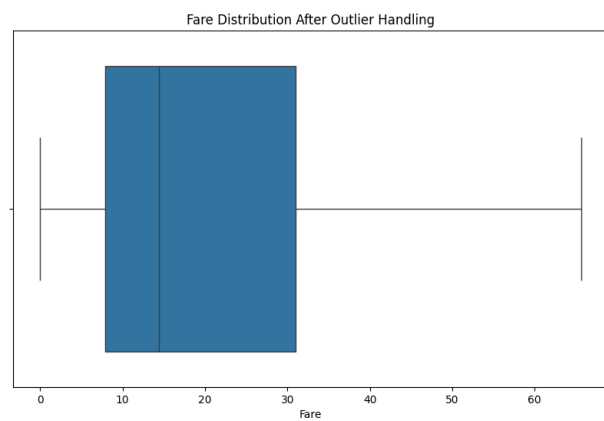
3. Data Visualizations

3.1. Outlier Detection (Before & After)

Before Outlier Handling

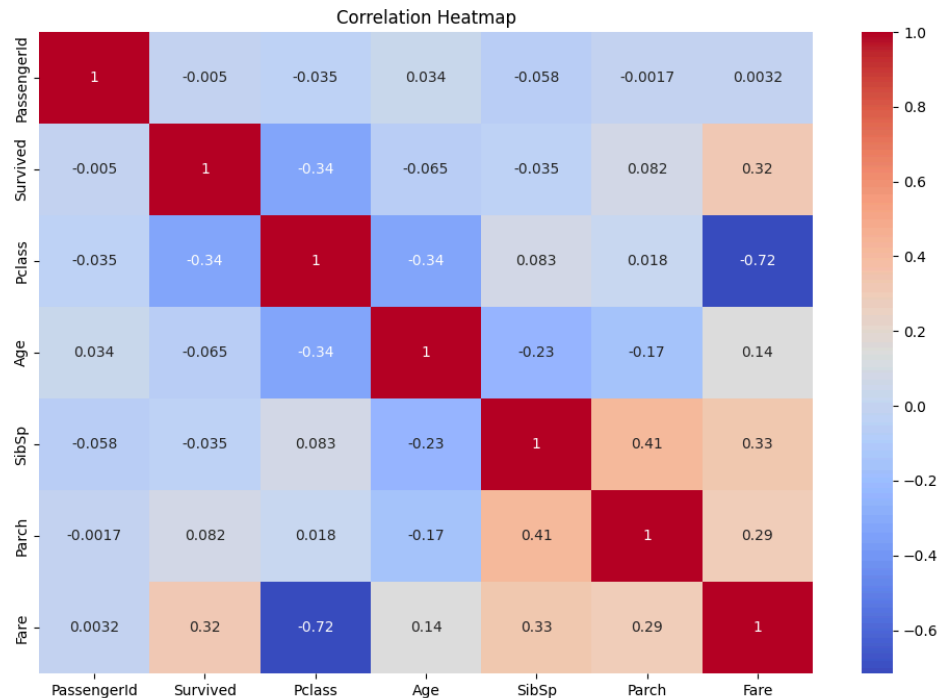


After Outlier Handling



3.2. Distribution and Correlation

- Survival by Class: Visualizes the survival rates across different passenger classes.
- Correlation Heatmap: Identifies relationships between numeric features.



4. Source Code

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Load Data
df = pd.read_csv('train.csv')

# Handling Missing Values
df['Age'] = df['Age'].fillna(df['Age'].median())
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
df['Cabin'] = df['Cabin'].fillna('Unknown')

# Handling Outliers (Fare)
Q1 = df['Fare'].quantile(0.25)
```

```
Q3 = df['Fare'].quantile(0.75)
IQR = Q3 - Q1
upper_bound = Q3 + 1.5 * IQR
df['Fare'] = np.where(df['Fare'] > upper_bound, upper_bound, df['Fare'])

# Visualizations
plt.figure(figsize=(10,6))
sns.boxplot(x=df['Fare']).set_title('Cleaned Fare Boxplot')
plt.savefig('cleaned_fare.png')

plt.figure(figsize=(12,8))
sns.heatmap(df.select_dtypes(include=[np.number]).corr(), annot=True)
plt.savefig('heatmap.png')
```

End of Project Report