Task 5 — Exploratory Data Analysis (EDA)

1. Introduction

This report demonstrates a complete Exploratory Data Analysis (EDA) workflow using Python. It contains data loading, cleaning, summary statistics, visualizations, and interpretations. A sample synthetic dataset (Titanic-like) is used for demonstration; replace it with your own dataset in the notebook.

2. Dataset Overview

The synthetic dataset used: rows = 200, columns = 6.
Columns: survived (0/1), class (First/Second/Third), sex, age, fare, embarked.
Replace the data-loading cell in the notebook with: pd.read_csv('yourfile.csv') and rerun the cells.

3. Data Cleaning & Preprocessing

Cleaning steps applied in the notebook:
• Inspect missing values and percent missing per column.
• Imputed 'age' with median for summary/statistics and plotting.
• Left other categorical variables as-is for value_counts; encoding guidance provided.
• Removed negative fares by taking absolute value when generating synthetic data; for real data check for errors.

4. Summary Statistics

Summary statistics (table truncated in PDF; full in notebook):

	survived	class	sex	age	fare	embarked
count	200.000000	200	200	180.000000	200.000000	200
unique	NaN	3	2	NaN	NaN	3
top	NaN	Third	male	NaN	NaN	S
freq	NaN	95	119	NaN	NaN	128
mean	0.365000	NaN	NaN	27.652222	70.992900	NaN
std	0.482638	NaN	NaN	14.627278	55.235444	NaN
min	0.000000	NaN	NaN	-12.600000	1.950000	NaN
25%	0.000000	NaN	NaN	18.425000	23.402500	NaN
50%	0.000000	NaN	NaN	27.450000	57.575000	NaN
75%	1.000000	NaN	NaN	36.625000	102.112500	NaN
max	1.000000	NaN	NaN	67.700000	249.780000	NaN

5. Visualizations & Key Findings

Visualizations generated (see figures). Interpretations:
• Age Distribution: Shows central tendency and spread; missing ages were imputed with median in analysis.
• Fare Boxplot: Reveals skew and outliers — consider log-transform for modeling.
• Survival Rate by Class: Differences indicate class-related survival disparities.
• Age vs Fare Scatter: Checks correlation; in this synthetic data there's no strong linear relation.

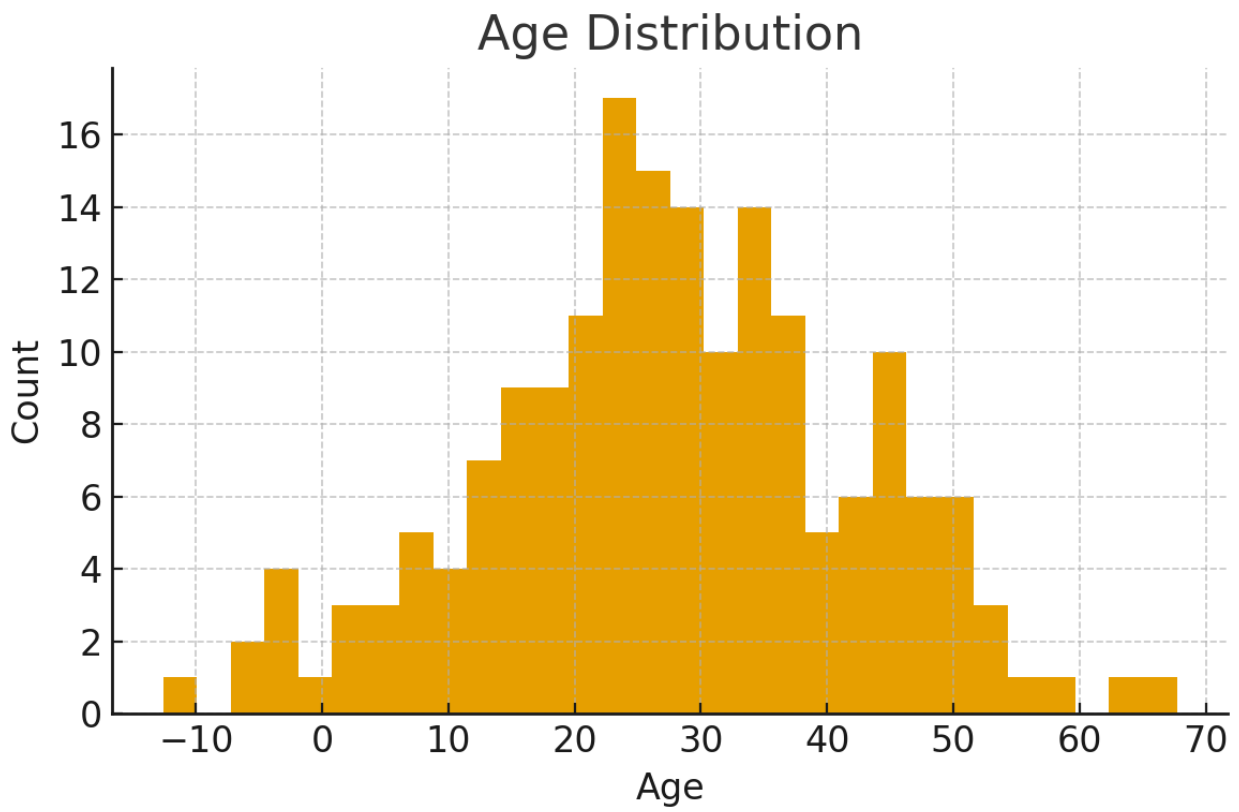

Figure 1 — Age Distribution

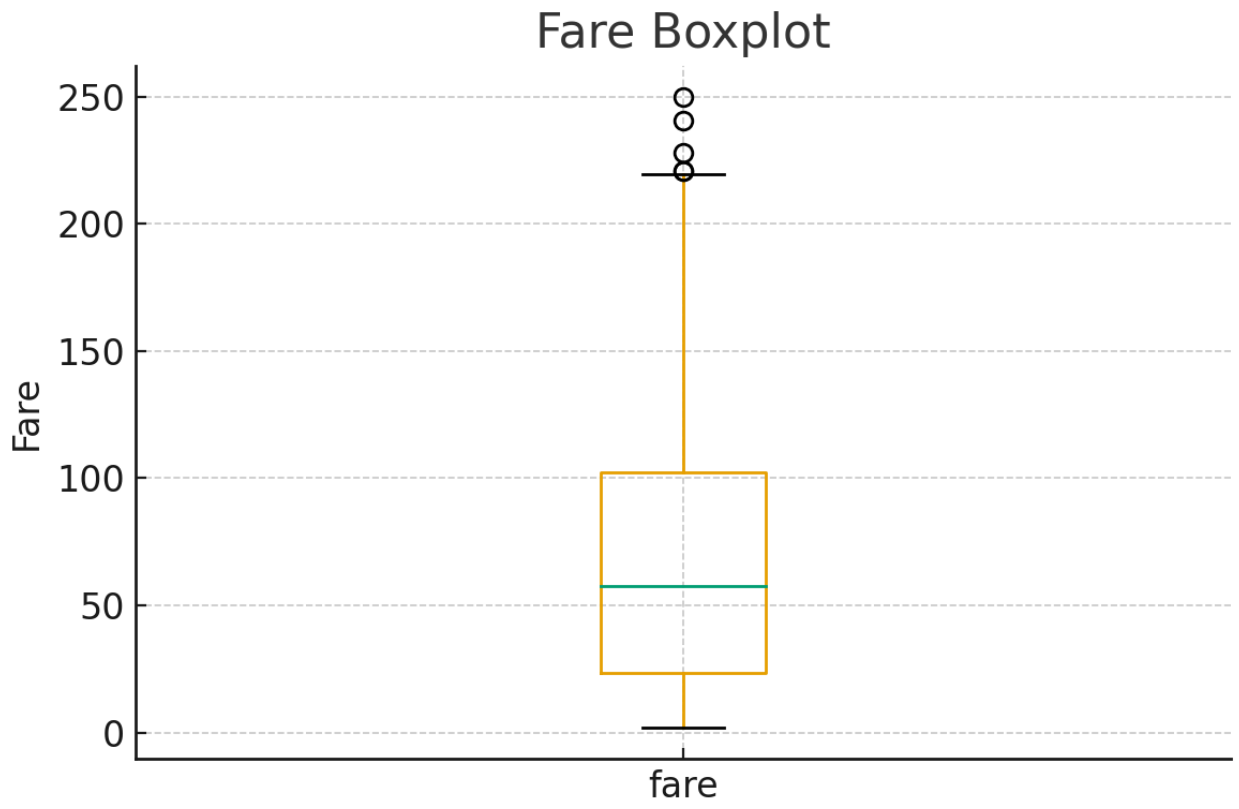

Figure 2 — Fare Boxplot

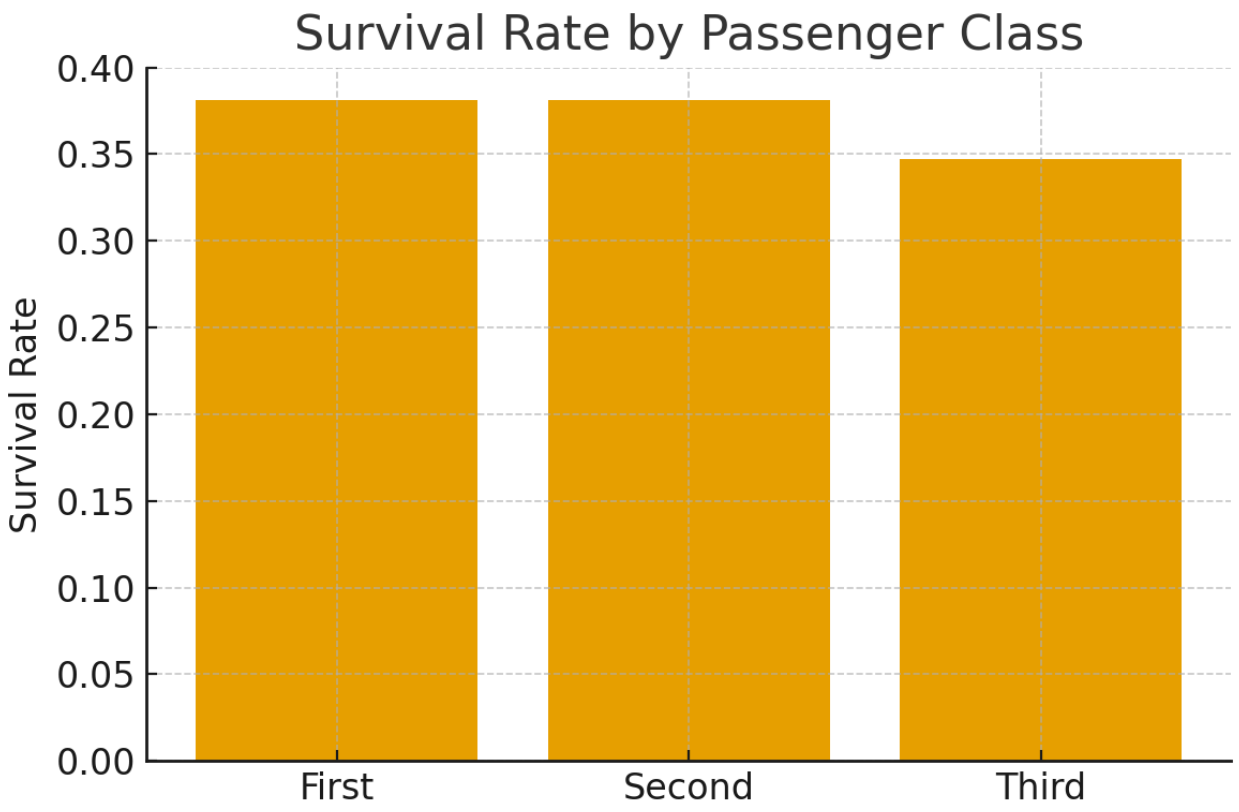

Figure 3 — Survival Rate by Class

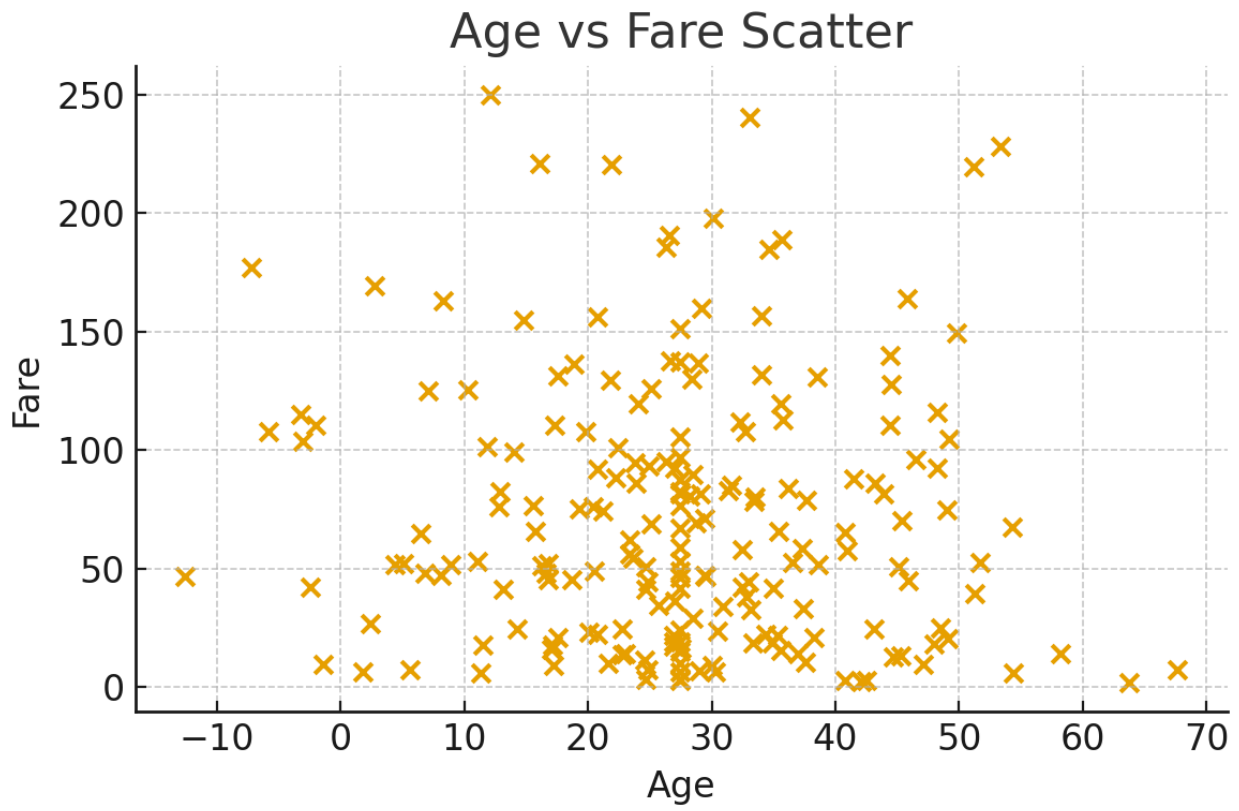

Figure 4 — Age vs Fare

6. Conclusions & Recommendations

Example conclusions from the synthetic dataset:
• Differences in survival rate by class suggest the model should include class as an important predictor.
• Missing age values should be handled carefully; consider models that can handle missingness or use informed imputation.