Laporan Praktikum 4

Exploratory Data Analysis (EDA)

Peno

221220095

September 7, 2025

Identitas Praktikum

Mata Kuliah : Data Science
Pertemuan : 4
Judul : Exploratory Data Analysis (EDA)
Nama : Peno
NIM : 221220095
Tanggal : 24 Oktober 2025

Tujuan

- Mahasiswa memahami konsep EDA.
- Mahasiswa mampu menghitung statistik deskriptif (numerik dan kategorikal).
- Mahasiswa mampu membuat visualisasi sederhana (histogram, bar chart, boxplot, scatter).
- Mahasiswa dapat menemukan insight awal dari dataset Titanic.

Dasar Teori

Exploratory Data Analysis (EDA) adalah proses investigasi awal pada data untuk menemukan pola, anomali, menguji hipotesis, dan memeriksa asumsi dengan bantuan statistik ringkas dan representasi grafis (visualisasi). EDA berfokus pada pemahaman data sebelum melakukan pemodelan.

Statistik deskriptif digunakan untuk memberikan ringkasan kuantitatif tentang data, baik untuk data numerik (seperti mean, median, std) maupun data kategorikal (seperti frekuensi atau modus). Visualisasi (seperti histogram, boxplot, dan bar chart) membantu dalam mengidentifikasi distribusi data, outlier, dan hubungan antar variabel secara visual. Dataset Titanic (Kaggle) adalah dataset klasik yang sering digunakan untuk latihan klasifikasi biner dan EDA.

Alat dan Bahan

1. Python 3
2. Library: pandas, matplotlib, seaborn
3. Dataset: titanic.csv (Kaggle)

Langkah Kerja

1. Membaca dataset dengan pd.read csv().
2. Menghitung statistik deskriptif kolom numerik (Age, Fare).
3. Menghitung distribusi kategorikal (Sex, Pclass).
4. Membuat visualisasi (histogram, bar chart, boxplot, scatter).
5. Menarik insight awal dari hasil eksplorasi.

Kode Program

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

# membaca dataset
df = pd.read_csv("titanic.csv")

# statistik deskriptif numerik
print(df["Age"].describe())
print(df["Fare"].describe())

# statistik kategorikal
print(df["Sex"].value_counts())
print(df["Pclass"].value_counts())

# visualisasi sederhana
# Histogram Age
df["Age"].hist(bins=20)
plt.title("Distribusi Umur")
plt.show()
# Countplot Sex
sns.countplot(x="Sex", data=df)
plt.title("Distribusi Jenis Kelamin")
```

```
plt.show()
# Boxplot Fare vs Pclass
sns.boxplot(x="Pclass", y="Fare", data=df)
plt.title("Distribusi Harga Tiket per Kelas")
plt.show()
# Scatterplot Age vs Fare
sns.scatterplot(x="Age", y="Fare", hue="Survived", data=df)
plt.title("Age vs Fare dengan Survival")
plt.show()
```

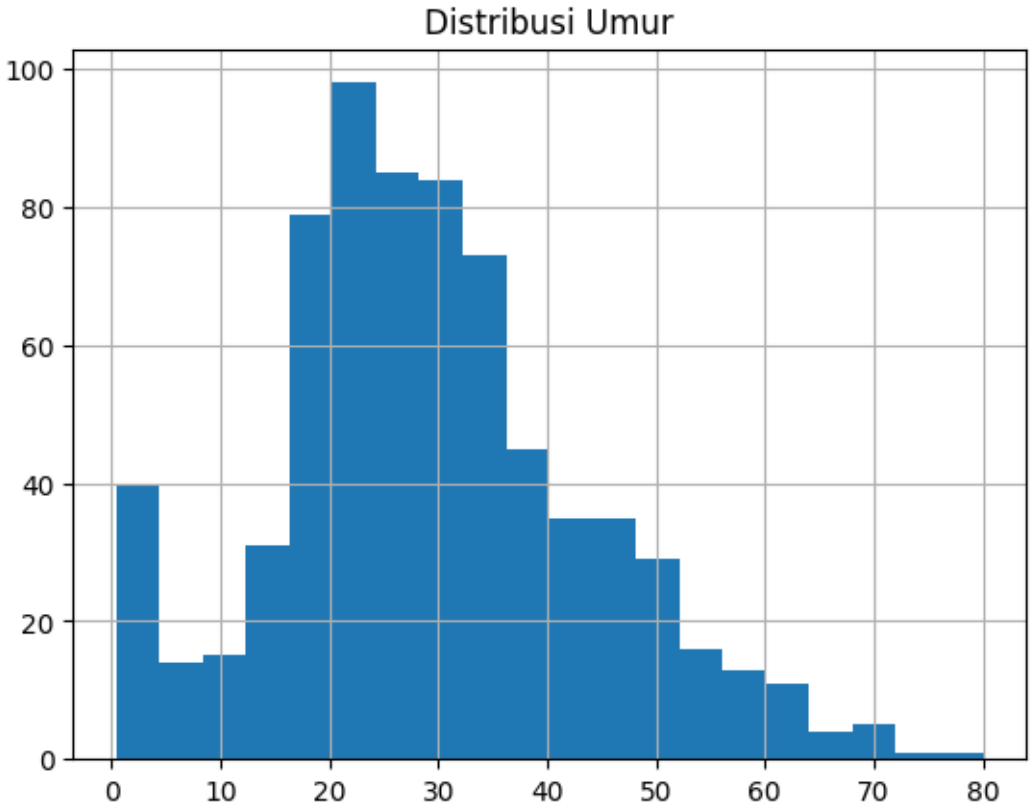
Hasil dan Output

Statistik Deskriptif

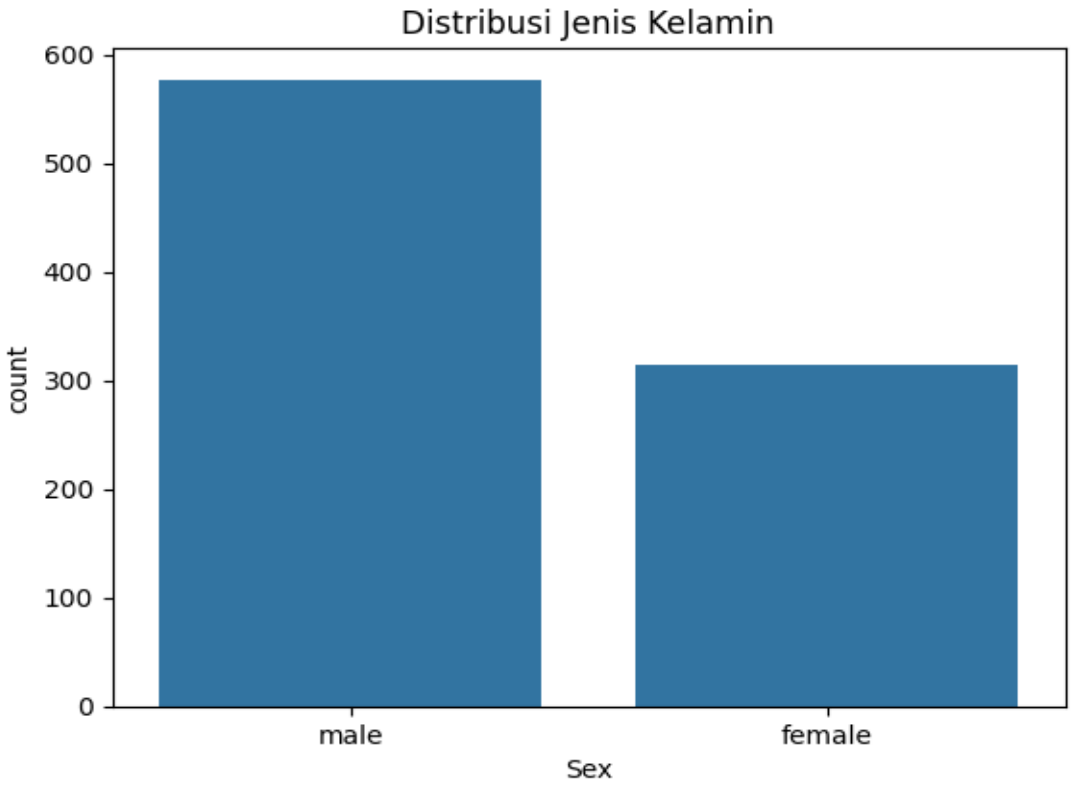
```
count   714.000000
mean     29.699118
std      14.526497
min       0.420000
25%      20.125000
50%      28.000000
75%      38.000000
max      80.000000
Name: Age, dtype: float64
count   891.000000
mean     32.204208
std      49.693429
min       0.000000
25%       7.910400
50%      14.454200
75%      31.000000
max     512.329200
Name: Fare, dtype: float64
Sex
male     577
female   314
Name: count, dtype: int64
Pclass
3   491
1   216
2   184
Name: count, dtype: int64
```

Visualisasi

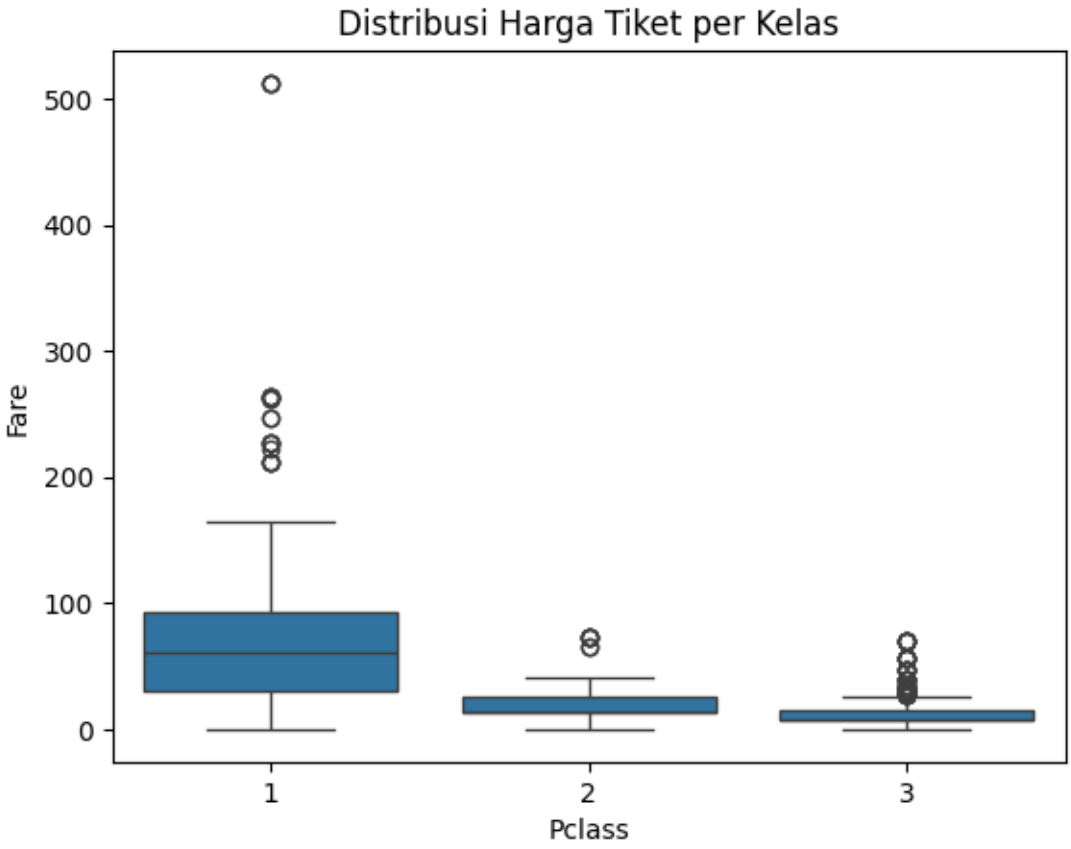
- Histogram distribusi umur:



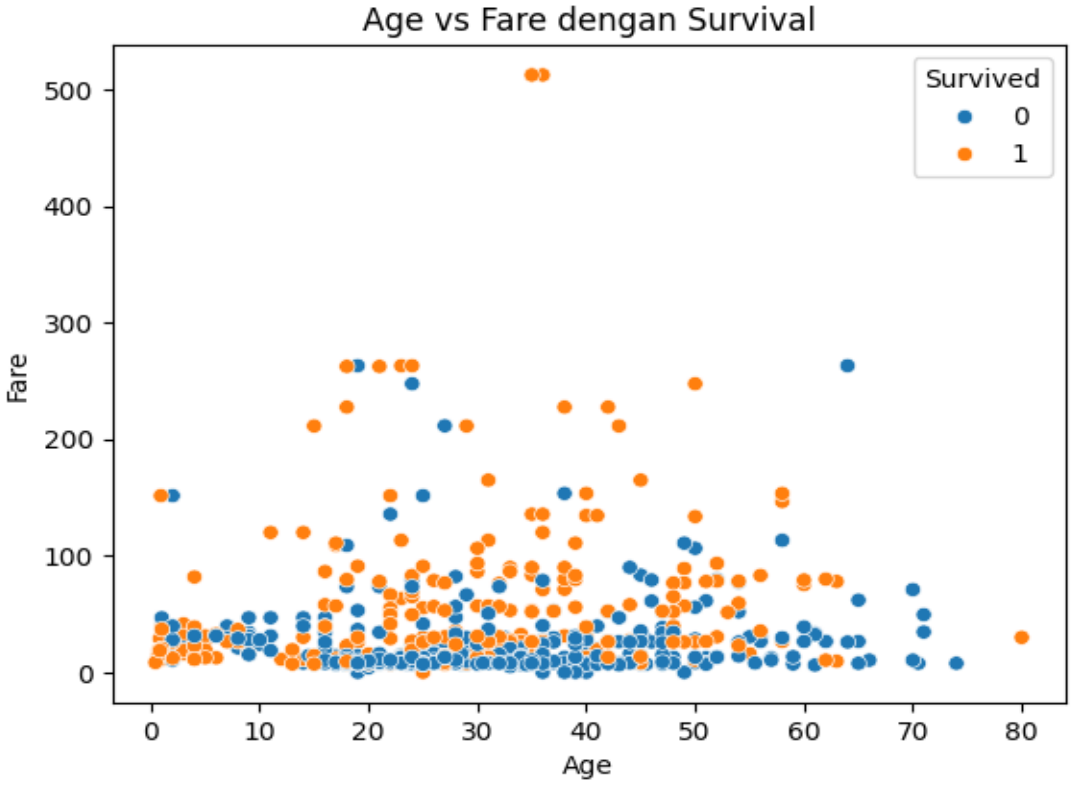
- Bar chart jenis kelamin:



- Boxplot Fare per kelas:



- Scatter Age vs Fare:



Pembahasan

Berdasarkan hasil statistik dan visualisasi, didapatkan beberapa insight awal:

- Terdapat *missing values* (data yang hilang) pada kolom Age, karena *count* (714) lebih sedikit dari total data.
- Distribusi umur penumpang mendekati distribusi normal, dengan konsentrasi penumpang terbanyak berada di rentang usia 20-40 tahun.
- Mayoritas penumpang adalah laki-laki.
- Penumpang mayoritas berada di kelas 3 (Pclass = 3).
- Harga tiket kelas 1 (Pclass = 1) jauh lebih tinggi dan memiliki variasi harga yang lebih besar dibandingkan kelas 2 dan 3, seperti yang terlihat jelas pada boxplot.
- Dari scatterplot, terlihat bahwa penumpang kelas 1 (yang membayar tiket mahal) cenderung lebih banyak yang selamat (Survived = 1).

Kesimpulan

1. EDA memberikan gambaran awal yang komprehensif mengenai karakteristik dataset Titanic.
2. Statistik deskriptif dan visualisasi berhasil digunakan untuk mengidentifikasi pola, distribusi data, outlier (seperti harga tiket yang sangat mahal), dan adanya data yang hilang.
3. Informasi yang diperoleh (seperti pentingnya Pclass dan Sex) sangat penting sebagai dasar untuk langkah analisis data dan feature engineering pada tahap modeling berikutnya..