

Project-1 Titanic dataset EDA

Name: Suheal Ahmad

Objective:

To perform exploratory data analysis on the Titanic dataset by inspecting missing values and data types, analysing survival rates by gender, passenger class, and age groups, and visualizing key insights using charts.

Codes:

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

# Load Titanic dataset (Seaborn built-in)
df = sns.load_dataset("titanic")

# -----
# 1. Inspect dataset
# -----
print(df.head())
print("\nMissing Values:\n", df.isnull().sum())
print("\nData Types:\n", df.dtypes)

# -----
# 2. Survival rate by Sex
# -----
survival_sex = df.groupby("sex")["survived"].mean()

plt.figure(figsize=(6,4))
survival_sex.plot(kind="bar")
plt.title("Survival Rate by Gender")
plt.ylabel("Survival Rate")
plt.show()

# -----
# 3. Survival rate by Passenger Class
# -----
survival_class = df.groupby("pclass")["survived"].mean()

plt.figure(figsize=(6,4))
survival_class.plot(kind="bar")
plt.title("Survival Rate by Passenger Class")
plt.xlabel("Passenger Class")
plt.ylabel("Survival Rate")
plt.show()

"
```

```

# 4. Age Buckets
# -----
df["age_group"] = pd.cut(
    df["age"],
    bins=[0, 12, 18, 35, 60, 100],
    labels=["Child", "Teen", "Young Adult", "Adult", "Senior"]
)

survival_age = df.groupby("age_group")["survived"].mean()

plt.figure(figsize=(7,4))
survival_age.plot(kind="bar")
plt.title("Survival Rate by Age Group")
plt.ylabel("Survival Rate")
plt.show()

# -----
# 5. Boxplot & Violin Plot
# -----
plt.figure(figsize=(6,4))
sns.boxplot(x="survived", y="age", data=df)
plt.title("Age Distribution by Survival")
plt.show()

plt.figure(figsize=(6,4))
sns.violinplot(x="sex", y="age", hue="survived", data=df, split=True)
plt.title("Age Distribution by Gender & Survival")
plt.show()

```

Results:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	\
0	0	3	male	22.0	1	0	7.2500	S	Third	
1	1	1	female	38.0	1	0	71.2833	C	First	
2	1	3	female	26.0	0	0	7.9250	S	Third	
3	1	1	female	35.0	1	0	53.1000	S	First	
4	0	3	male	35.0	0	0	8.0500	S	Third	

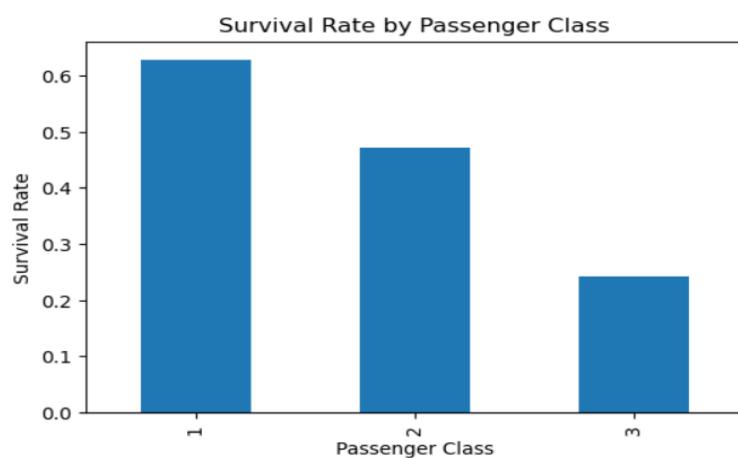
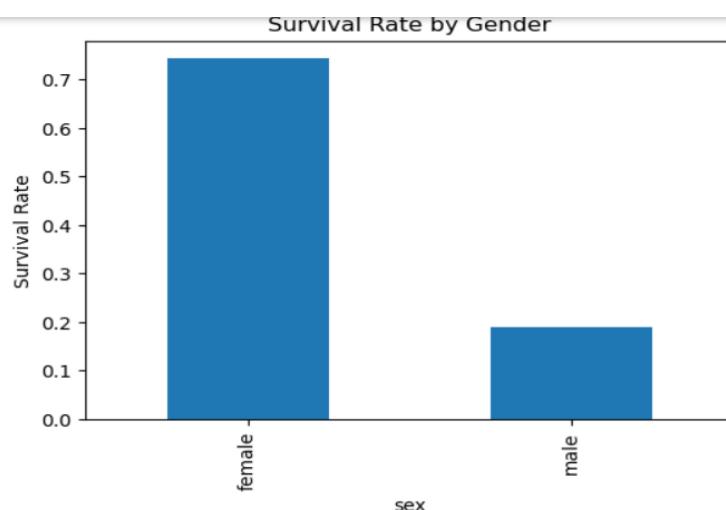
	who	adult_male	deck	embark_town	alive	alone
0	man	True	NaN	Southampton	no	False
1	woman	False	C	Cherbourg	yes	False
2	woman	False	NaN	Southampton	yes	True
3	woman	False	C	Southampton	yes	False
4	man	True	NaN	Southampton	no	True

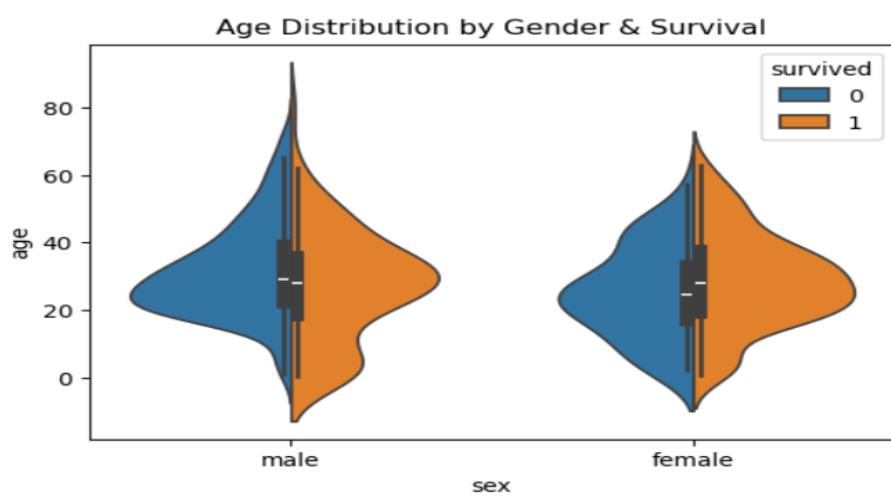
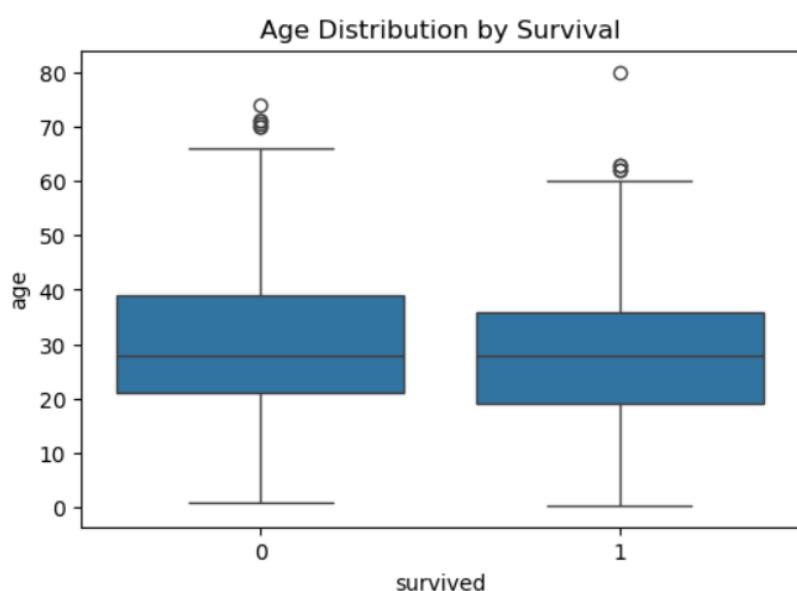
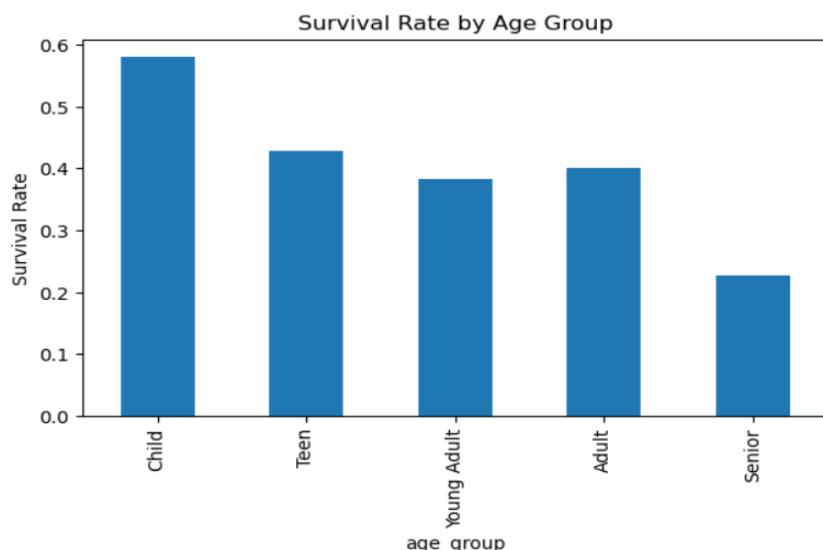
Missing Values:

survived	0
pclass	0
sex	0
age	177
sibsp	0
parch	0
fare	0
embarked	2
class	0
who	0
adult_male	0
deck	688
embark_town	2
alive	0
alone	0
dtype: int64	

Data Types:

```
survived           int64
pclass            int64
sex              object
age             float64
sibsp            int64
parch            int64
fare             float64
embarked        object
class            category
who              object
adult_male       bool
deck            category
embark_town     object
alive            object
alone            bool
dtype: object
```





Explanation of Titanic EDA Results:

The exploratory data analysis of the Titanic dataset was conducted to understand how different demographic and socio-economic factors influenced passenger survival. Visual analysis using bar charts, boxplots, and violin plots helped uncover meaningful patterns that are not immediately visible through raw numbers alone.

The gender-based survival analysis clearly shows a significant disparity between male and female passengers. Female passengers had a much higher survival rate, indicating that rescue efforts followed a gender-prioritized evacuation strategy. This pattern remained consistent across different age ranges, reinforcing the importance of gender as a decisive factor.

The passenger class analysis revealed that survival rates decreased as passenger class declined. First-class passengers had better access to lifeboats and evacuation routes, while third-class passengers faced structural and logistical disadvantages. This highlights how socio-economic status directly influenced survival outcomes during the disaster.

The age-group analysis shows that children had the highest survival rate, reflecting evacuation priorities. Younger adults and adults had moderate survival chances, while senior passengers were the most vulnerable group, likely due to limited physical mobility and delayed access to safety measures.

Boxplots and violin plots further revealed differences in age distribution between survivors and non-survivors. Survivors tended to be slightly younger on average, although some older passengers did survive. These visualizations confirmed that while age influenced survival, it was not as dominant as gender or class.

Conclusion:

The Titanic dataset EDA clearly demonstrates that survival during the disaster was not random but strongly influenced by structured social and demographic factors. Gender emerged as the most significant determinant of survival, followed closely by passenger class, while age played a secondary yet important role. The analysis confirms that women and children were prioritized during evacuation, and passengers from higher socio-economic classes benefited from better access to safety measures.

This project highlights the importance of exploratory data analysis in uncovering real-world patterns and biases hidden within data. Through effective use of visualizations, the analysis transformed raw data into meaningful insights that explain human behaviour, policy decisions, and structural inequalities during crisis situations. Overall, this EDA provides a strong foundation for predictive modelling and reinforces the value of data-driven reasoning in understanding historical events.