Outcome: Gain skill in finding patterns, trends, and anomalies.

```
from google.colab import files
import pandas as pd

# Step 1 – Upload CSV file
uploaded = files.upload()  # Choose train.csv from your computer

# Step 2 – Read CSV into DataFrame
df = pd.read_csv("train.csv")  # name must match the uploaded file
df.head()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving train.csv to train (2).csv
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
[illegible]	[illegible]	[illegible]	[illegible]	Futrelle, Mrs. Jacques Heath (Lily May	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

```
# Shape (rows, columns)
df.shape

# Info about data types & missing values
df.info()

# Summary statistics (numeric columns)
df.describe()

# Missing values count
df.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

	0
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

dtype: int64

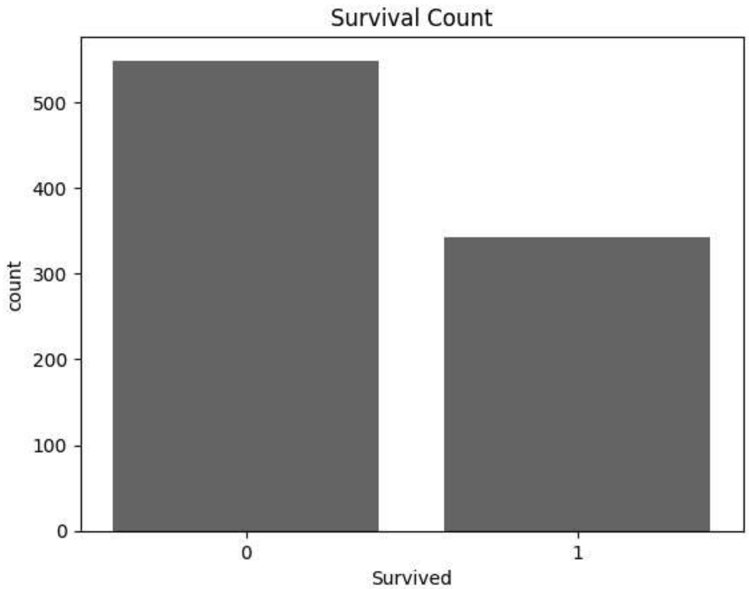
```
# Fill missing Embarked with most common value
most_common_embarked = df['Embarked'].mode().iloc[0]
df['Embarked'] = df['Embarked'].fillna(most_common_embarked)

# Fill missing Age with median based on Pclass & Sex
df['Age'] = df.groupby(['Pclass','Sex'])['Age'] \
              .transform(lambda x: x.fillna(x.median()))

# Drop Cabin (too many missing)
df = df.drop(columns=['Cabin'])

# Confirm no more missing
df.isnull().sum()
```

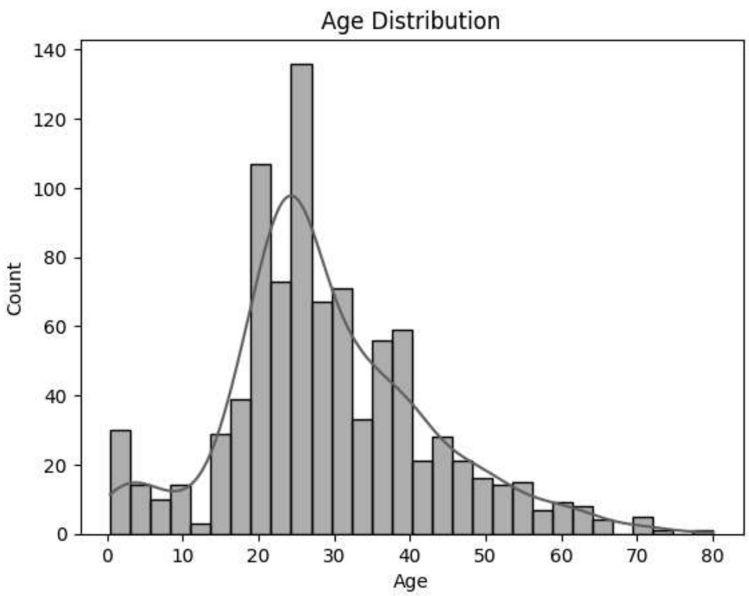
	0
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	0
SibSp	0
Parch	0
Ticket	0
Fare	0
Embarked	0

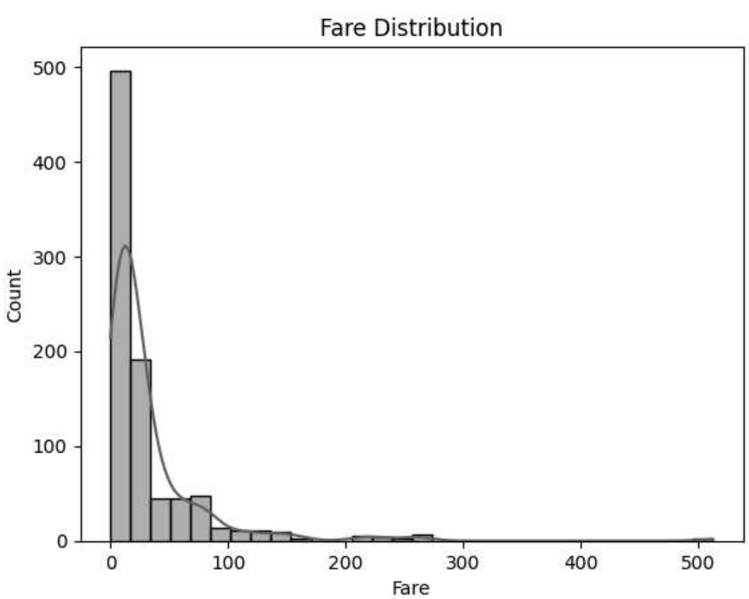
dtype: int64

```
# Family size
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1

# Age group
df['AgeGroup'] = pd.cut(df['Age'],
                        bins=[0,12,20,40,60,100],
                        labels=['Child','Teen','Adult','MidAge','Senior'])
```

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

# Survival count
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.show()

# Age distribution
sns.histplot(df['Age'], bins=30, kde=True)
plt.title('Age Distribution')
plt.show()

# Fare distribution
sns.histplot(df['Fare'], bins=30, kde=True)
plt.title('Fare Distribution')
plt.show()
```







Observation 1 - Survival Count

There were more passengers who did not survive compared to those who did. This shows the disaster had a high mortality rate.

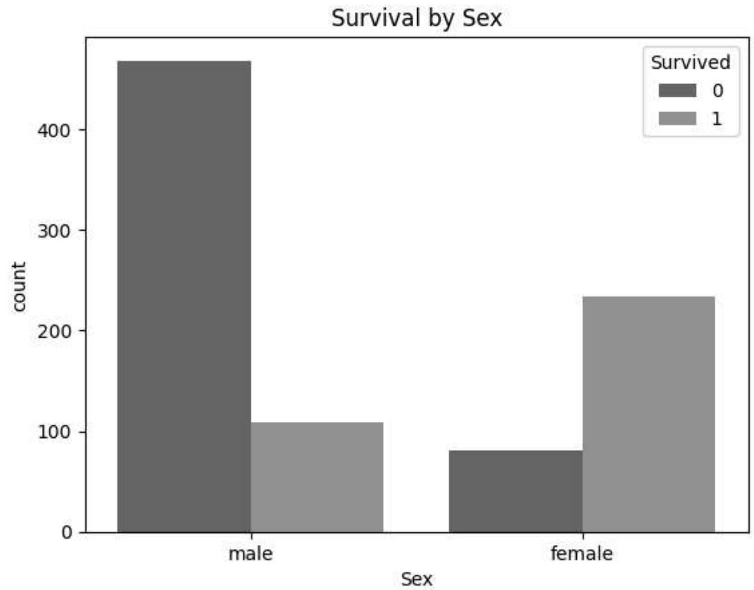
Observation 2 - Age Distribution

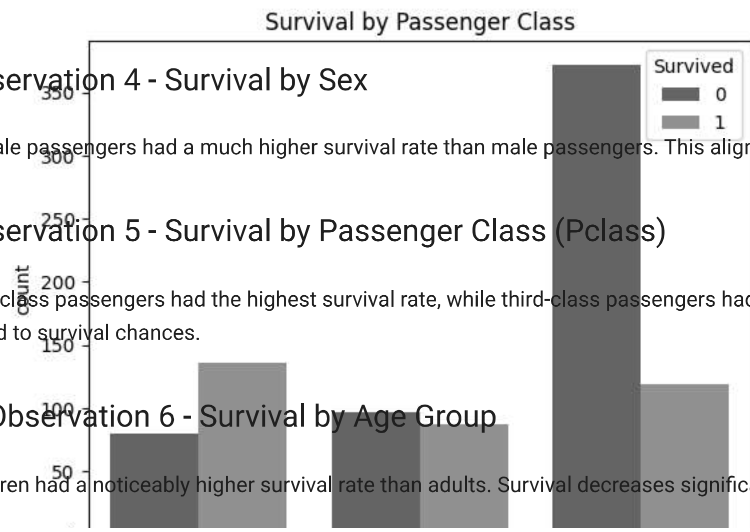
Most passengers were between 20–40 years old. There were fewer children (under 12) and fewer seniors (over 60).

Observation 3 - Fare Distribution

Most passengers paid relatively low fares (under $50), but there are some outliers who paid very high fares.

```
# Survival rate by Sex
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival by Sex')
plt.show()

# Survival rate by Pclass
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Survival by Passenger Class')
plt.show()

# Survival rate by AgeGroup
sns.countplot(x='AgeGroup', hue='Survived', data=df)
plt.title('Survival by Age Group')
plt.show()
```





Observation 4 - Survival by Sex

Female passengers had a much higher survival rate than male passengers. This aligns with the 'women and children first' evacuation policy.

Observation 5 - Survival by Passenger Class (Pclass)

First-class passengers had the highest survival rate, while third-class passengers had the lowest. Passenger class appears to be strongly linked to survival chances.

Observation 6 - Survival by Age Group

Children had a noticeably higher survival rate than adults. Survival decreases significantly for teens and adults.

```
# Select only numeric columns
numeric_df = df.select_dtypes(include=['int64', 'float64'])

plt.figure(figsize=(8,6))
```