Name : Tanishq Thuse

Year : SY

Div : CS(AI)-B

Roll no : 60

PRN : 12310237

Assignment -7

Problem Statement : Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram

Dataset link : https://github.com/datasciencedojo/datasets/blob/master/titanic.csv

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

#We are going to use titanic dataset
path = "/content/titanic.csv"

df = pd.read_csv(path)
```

Data preprocessing

```
df.head()
```

	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

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df.tail()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.00	Unknown	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.45	Unknown	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.00	C148	C

```
# Check for null values
print(df.isnull().sum())
```

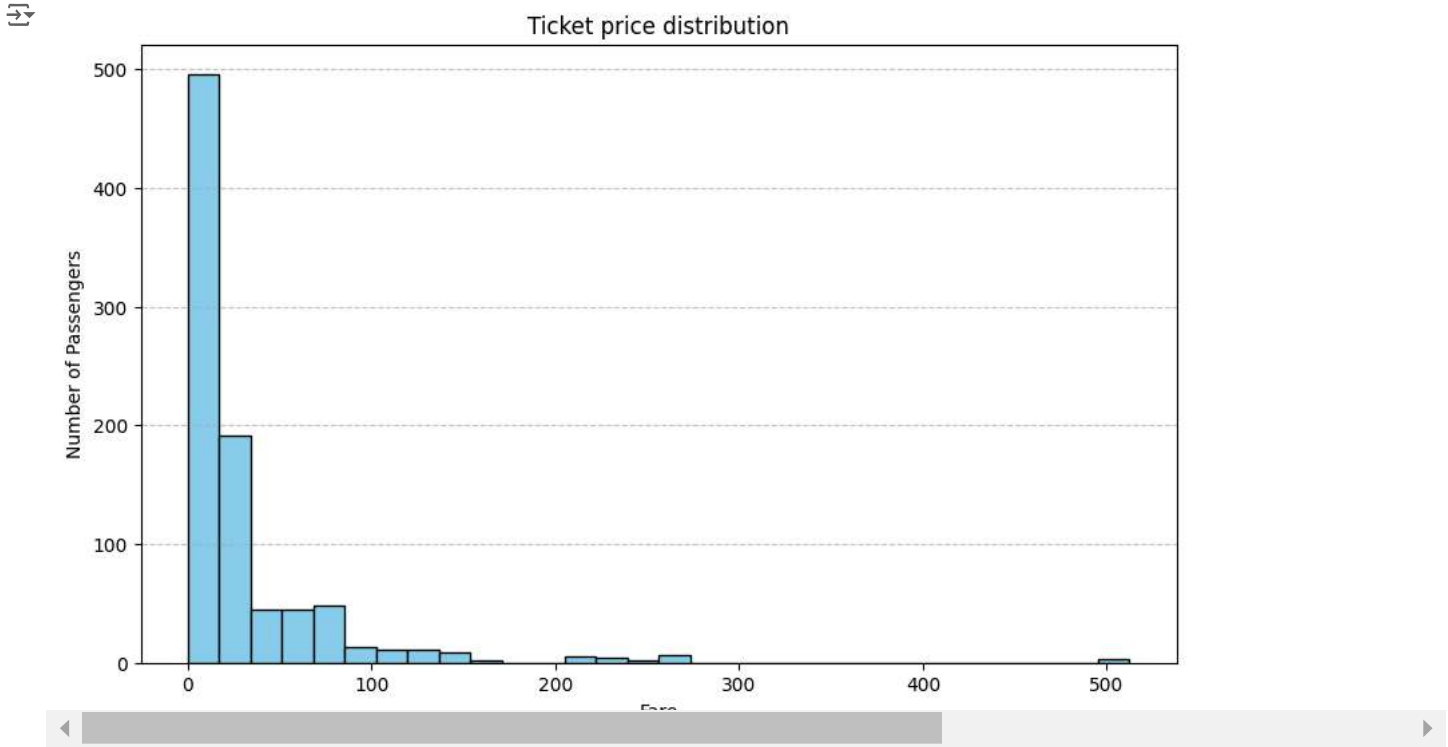
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

```
# Handle missing values (example: fill with mean for numerical columns)
df['Age'].fillna(df['Age'].mean(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df['Cabin'].fillna('Unknown', inplace=True)


# Check for null values after handling
print("\nAfter handling missing values:")
print(df.isnull().sum())
```

```
After handling missing values:
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

```
data = df
plt.figure(figsize=(10, 6))
plt.hist(data['Fare'], bins=30, color='skyblue', edgecolor='black')
plt.title('Ticket price distribution')
plt.xlabel('Fare')
plt.ylabel('Number of Passengers')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```




```
plt.figure(figsize=(12, 6))
# Creating a boxplot of 'Age' grouped by 'Sex' and 'Survived' status
data.boxplot(column='Age', by=['Sex', 'Survived'], grid=False,
patch_artist=True, showfliers=True, notch=True)
plt.title('Distribution of Age by Gender and Survival Status')
plt.suptitle('') # Removing the default suptitle
plt.xlabel('Gender and Survival Status (0 = Did Not Survive, 1 = Survived)')
plt.ylabel('Age')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

```
<Figure size 1200x600 with 0 Axes>
```