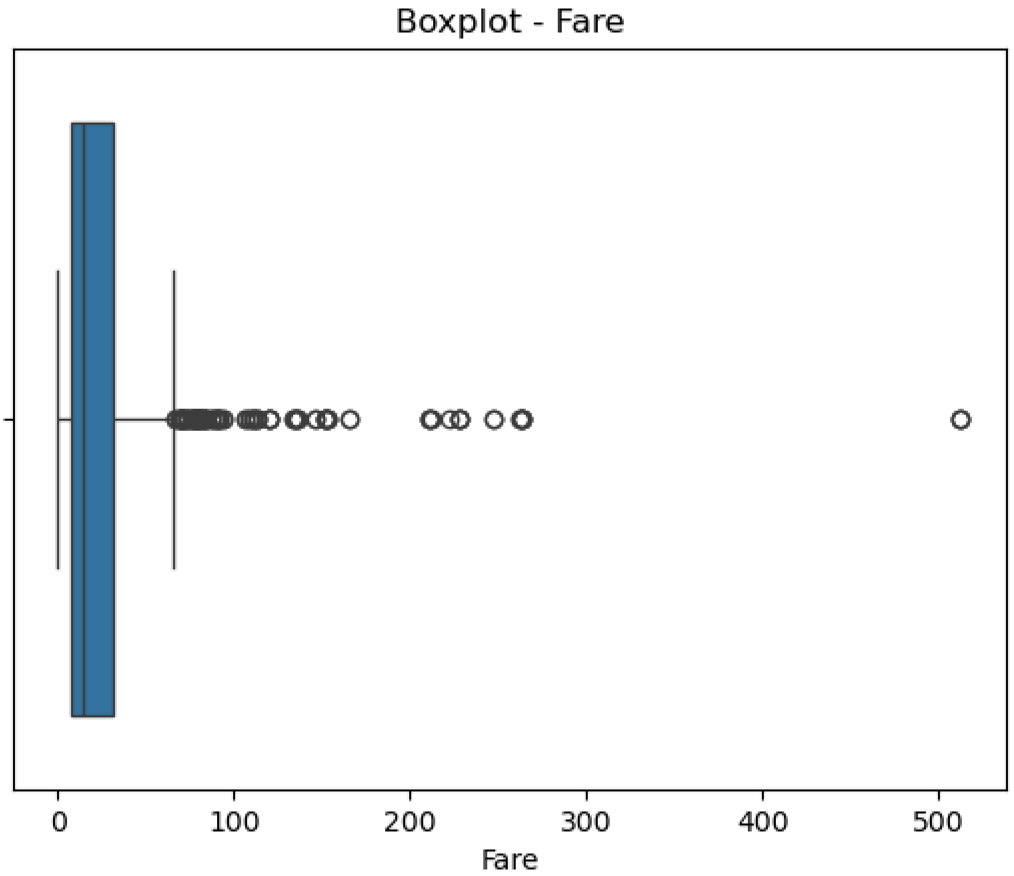
In [3]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder, StandardScaler

# Load the dataset
df = pd.read_csv("https://raw.githubusercontent.com/datasciencedojo/datasets/mas

# View metadata
meta = df.describe(include="all")

# Drop unused columns
df.drop(columns=["Cabin", "Ticket", "Name"], inplace=True)

# Fill missing values
df["Age"] = df["Age"].fillna(df["Age"].median())
df["Embarked"] = df["Embarked"].fillna(df["Embarked"].mode()[0])

# Encode categorical features
enc = LabelEncoder()
df["Sex"] = enc.fit_transform(df["Sex"])
df = pd.get_dummies(df, columns=["Embarked"], drop_first=True)

# Define IQR bounds and clip function BEFORE scaling
# This way we handle outliers on the original data
bounds = lambda x: (x.quantile(0.25), x.quantile(0.75))
iqr_clip = lambda s: s[(s >= bounds(s)[0] - 1.5 * (bounds(s)[1] - bounds(s)[0]))
                        (s <= bounds(s)[1] + 1.5 * (bounds(s)[1] - bounds(s)[0])

# Show boxplot before removing outliers
sns.boxplot(x=df["Fare"])
plt.title("Boxplot - Fare")
plt.show()

# Filter DataFrame based on clipped Fare values
df = df[df["Fare"].isin(iqr_clip(df["Fare"]))].reset_index(drop=True)

# Scale numerical features AFTER handling outliers
scaler = StandardScaler()
for col in ["Age", "Fare"]:
    # Use DataFrame indexing to ensure we get back a DataFrame, not an array
    df[col] = scaler.fit_transform(df[[col]])

# Output the shape and first few rows
print(df.shape)
print(df.head())
```



```
(775, 10)
   PassengerId  Survived  Pclass  Sex       Age  SibSp  Parch      Fare  \
0            1         0       3    1 -0.528321      1      0 -0.779117   
1            3         1       3    0 -0.215182      0      0 -0.729373   
2            4         1       1    0  0.489381      1      0  2.599828   
3            5         0       3    1  0.489381      0      0 -0.720161   
4            6         0       3    1 -0.058613      0      0 -0.690071   

   Embarked_Q  Embarked_S  
0       False        True  
1       False        True  
2       False        True  
3       False        True  
4        True       False  
```

```
In [ ]:
```