In [2]:

```
#Q.5)Perform EDA to show outliers and anomalies from given data set
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df = pd.read_csv("titanic.csv")
df.head()
```

Out[2]:

	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
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [5]:

```
numerical_cols = ['Age', 'Fare']
df[numerical_cols].describe()
```

Out[5]:

	Age	Fare
count	891.000000	891.000000
mean	29.361582	32.204208
std	13.019697	49.693429
min	0.420000	0.000000
25%	22.000000	7.910400
50%	28.000000	14.454200
75%	35.000000	31.000000
max	80.000000	512.329200

In [6]:

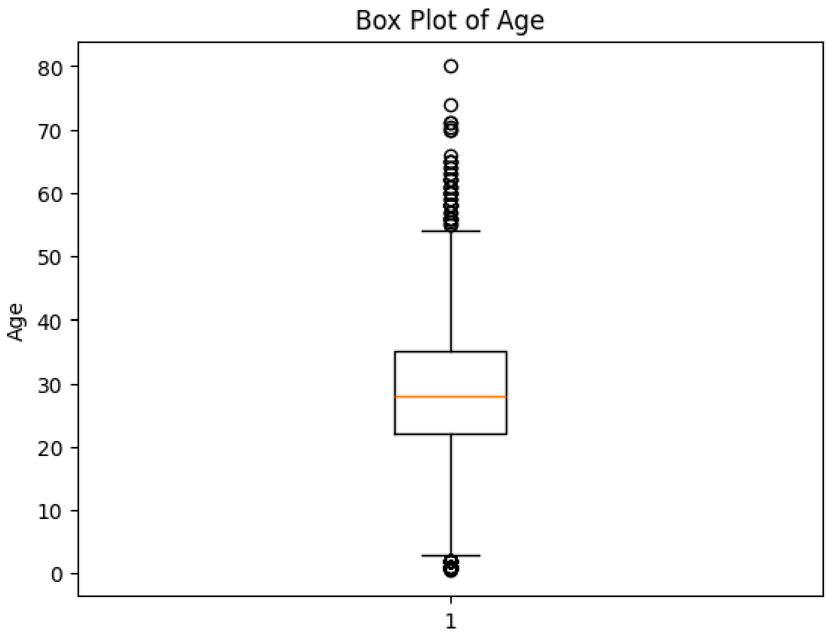
```
numerical_cols = ['Age', 'Fare']
df[numerical_cols].describe()
```

Out[6]:

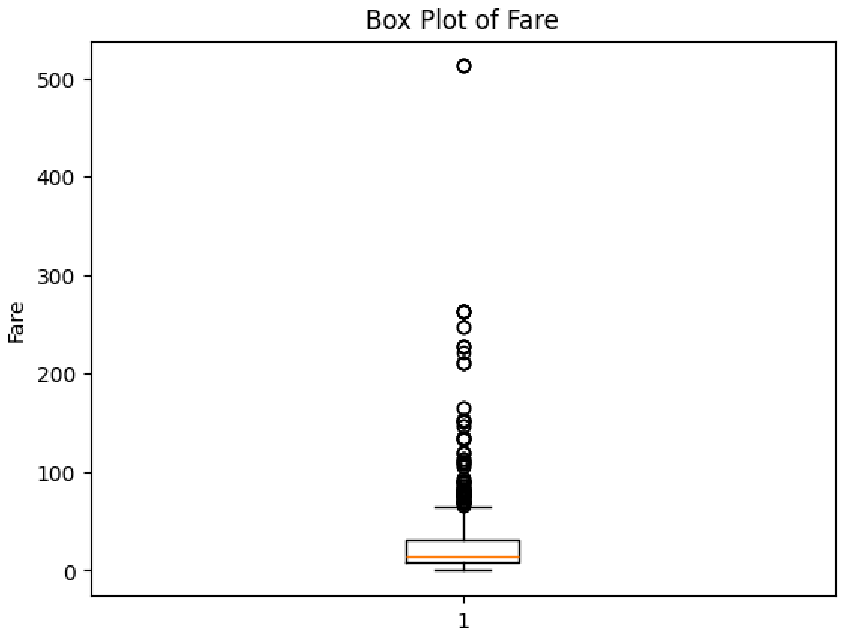
	Age	Fare
count	891.000000	891.000000
mean	29.361582	32.204208
std	13.019697	49.693429
min	0.420000	0.000000
25%	22.000000	7.910400
50%	28.000000	14.454200
75%	35.000000	31.000000
max	80.000000	512.329200

In [7]:

```
plt.figure()
plt.boxplot(df['Age'])
plt.title("Box Plot of Age")
plt.ylabel("Age")
plt.show()
```



In [8]:
```
plt.figure()
plt.boxplot(df['Fare'])
plt.title("Box Plot of Fare")
plt.ylabel("Fare")
plt.show()
```

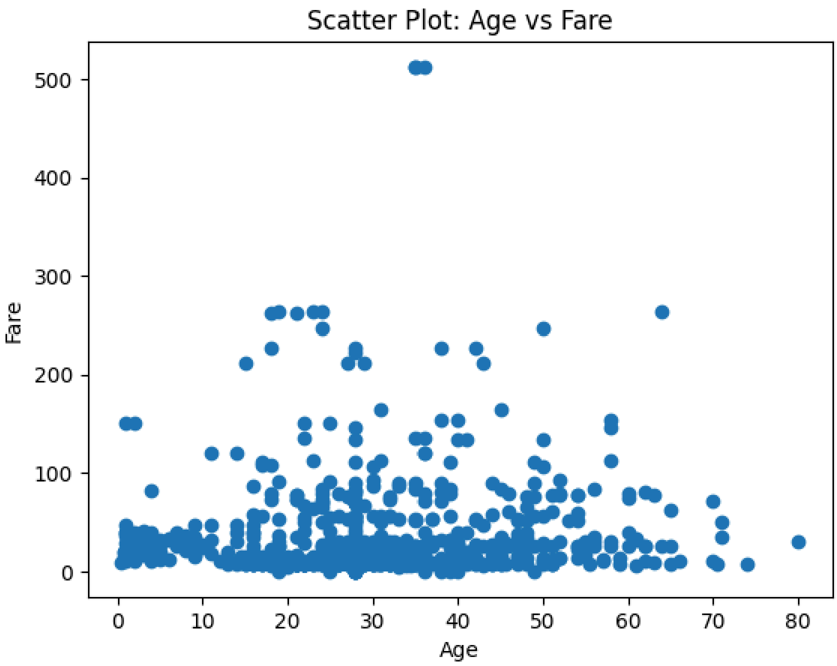


In [9]:
```
def detect_outliers_iqr(column):
    Q1 = column.quantile(0.25)
    Q3 = column.quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    return column[(column < lower) | (column > upper)]
```

In [10]:
```
fare_outliers = detect_outliers_iqr(df['Fare'])
fare_outliers.head()
```

Out[10]:
```
1      71.2833
27    263.0000
31    146.5208
34     82.1708
52     76.7292
Name: Fare, dtype: float64
```

In [11]:
```
plt.figure()
plt.scatter(df['Age'], df['Fare'])
plt.title("Scatter Plot: Age vs Fare")
plt.xlabel("Age")
plt.ylabel("Fare")
plt.show()
```



In [12]:

```
from scipy import stats

z_scores = np.abs(stats.zscore(df['Fare']))
anomalies = df[z_scores > 3]

anomalies[['Age', 'Fare']].head()
```

Out[12]:

	Age	Fare
27	19.0	263.0000
88	23.0	263.0000
118	24.0	247.5208
258	35.0	512.3292
299	50.0	247.5208

In [ ]: