In [1]:

```
#installing required libraries to perform Exploratory Data Analysis
pip install pandas numpy matplotlib seaborn
```

```
Defaulting to user installation because normal site-packages is not writeabl
e
Requirement already satisfied: pandas in c:\programdata\anaconda3\lib\site-p
ackages (2.0.3)
Requirement already satisfied: numpy in c:\programdata\anaconda3\lib\site-pa
ckages (1.24.3)
Requirement already satisfied: matplotlib in c:\programdata\anaconda3\lib\si
te-packages (3.7.2)
Requirement already satisfied: seaborn in c:\programdata\anaconda3\lib\site-
packages (0.12.2)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\programdata\anac
onda3\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\programdata\anaconda3\lib
\site-packages (from pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\programdata\anaconda3\li
b\site-packages (from pandas) (2023.3)
Requirement already satisfied: contourpy>=1.0.1 in c:\programdata\anaconda3
\lib\site-packages (from matplotlib) (1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\programdata\anaconda3\lib
\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\programdata\anaconda3
\lib\site-packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\programdata\anaconda3
\lib\site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\programdata\anaconda3\l
ib\site-packages (from matplotlib) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\programdata\anaconda3\lib
\site-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing<3.1,>=2.3.1 in c:\programdata\anaco
nda3\lib\site-packages (from matplotlib) (3.0.9)
Requirement already satisfied: six>=1.5 in c:\programdata\anaconda3\lib\site
-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:

```
#importing the downloaded libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```
#load the dataset from device to the code
#dataset downloaded from kaagal:
#https://www.kaggle.com/datasets/heptapod/titanic
df = pd.read_csv('C:/Users/anush/Downloads/archive/train.csv')
# Check the first few rows
print(df.head())
```

```
   Passengerid   Age     Fare  Sex  sibsp  zero  zero.1  zero.2  zero.3  \
0            1  22.0   7.2500    0      1     0       0       0       0   
1            2  38.0  71.2833    1      1     0       0       0       0   
2            3  26.0   7.9250    1      0     0       0       0       0   
3            4  35.0  53.1000    1      1     0       0       0       0   
4            5  35.0   8.0500    0      0     0       0       0       0   

   zero.4  ...  zero.12  zero.13  zero.14  Pclass  zero.15  zero.16  Embarke
d  \
0       0  ...        0        0        0       3        0        0       2.
0   
1       0  ...        0        0        0       1        0        0       0.
0   
2       0  ...        0        0        0       3        0        0       2.
0   
3       0  ...        0        0        0       1        0        0       2.
0   
4       0  ...        0        0        0       3        0        0       2.
0   

   zero.17  zero.18  2urvived  
0        0        0         0  
1        0        0         1  
2        0        0         1  
3        0        0         1  
4        0        0         0  

[5 rows x 28 columns]
```

In [5]:

```
#to understand the data in the dataset
#to check the structure and data types
print(df.info())
#to take summary statistics
print(df.describe())
```

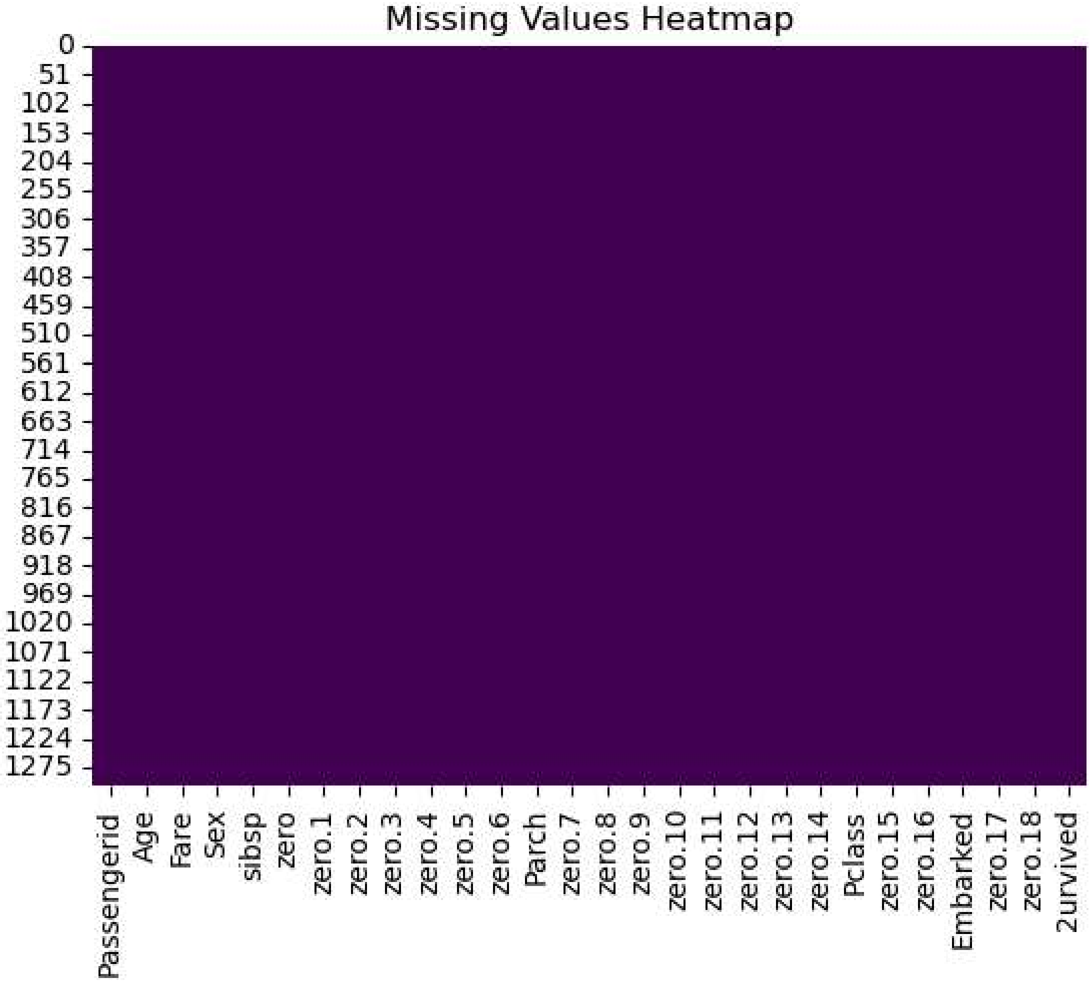
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 28 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Passengerid  1309 non-null   int64  
 1   Age          1309 non-null   float64
 2   Fare         1309 non-null   float64
 3   Sex          1309 non-null   int64  
 4   sibsp        1309 non-null   int64  
 5   zero         1309 non-null   int64  
 6   zero.1       1309 non-null   int64  
 7   zero.2       1309 non-null   int64  
 8   zero.3       1309 non-null   int64  
 9   zero.4       1309 non-null   int64  
 10  zero.5       1309 non-null   int64  
 11  zero.6       1309 non-null   int64  
 12  Parch        1309 non-null   int64  
 13  zero.7       1309 non-null   int64  
 14  zero.8       1309 non-null   int64  
 15  zero.9       1309 non-null   int64  
 16  zero.10      1309 non-null   int64  
 17  zero.11      1309 non-null   int64  
 18  zero.12      1309 non-null   int64  
 19  zero.13      1309 non-null   int64  
 20  zero.14      1309 non-null   int64  
 21  Pclass       1309 non-null   int64  
 22  zero.15      1309 non-null   int64  
 23  zero.16      1309 non-null   int64  
 24  Embarked     1307 non-null   float64
 25  zero.17      1309 non-null   int64  
 26  zero.18      1309 non-null   int64  
 27  2urvived     1309 non-null   int64  
dtypes: float64(3), int64(25)
memory usage: 286.5 KB
None
       Passengerid          Age         Fare          Sex        sibsp  \
count  1309.000000  1309.000000  1309.000000  1309.000000  1309.000000   
mean    655.000000    29.503186    33.281086     0.355997     0.498854   
std     378.020061    12.905241    51.741500     0.478997     1.041658   
min       1.000000     0.170000     0.000000     0.000000     0.000000   
25%     328.000000    22.000000     7.895800     0.000000     0.000000   
50%     655.000000    28.000000    14.454200     0.000000     0.000000   
75%     982.000000    35.000000    31.275000     1.000000     1.000000   
max    1309.000000    80.000000   512.329200     1.000000     8.000000   

         zero  zero.1  zero.2  zero.3  zero.4  ...  zero.12  zero.13  zero.1
4  \
count  1309.0  1309.0  1309.0  1309.0  1309.0  ...   1309.0   1309.0   1309.
0   
mean      0.0     0.0     0.0     0.0     0.0  ...      0.0      0.0      0.
0   
std       0.0     0.0     0.0     0.0     0.0  ...      0.0      0.0      0.
0   
min       0.0     0.0     0.0     0.0     0.0  ...      0.0      0.0      0.
0   
25%       0.0     0.0     0.0     0.0     0.0  ...      0.0      0.0      0.
```

```
0
50%       0.0     0.0     0.0     0.0     0.0  ...      0.0      0.0      0.
0
75%       0.0     0.0     0.0     0.0     0.0  ...      0.0      0.0      0.
0
max       0.0     0.0     0.0     0.0     0.0  ...      0.0      0.0      0.
0

            Pclass  zero.15  zero.16     Embarked  zero.17  zero.18  \
count  1309.000000   1309.0   1309.0  1307.000000   1309.0   1309.0   
mean      2.294882      0.0      0.0     1.492731      0.0      0.0   
std       0.837836      0.0      0.0     0.814626      0.0      0.0   
min       1.000000      0.0      0.0     0.000000      0.0      0.0   
25%       2.000000      0.0      0.0     1.000000      0.0      0.0   
50%       3.000000      0.0      0.0     2.000000      0.0      0.0   
75%       3.000000      0.0      0.0     2.000000      0.0      0.0   
max       3.000000      0.0      0.0     2.000000      0.0      0.0   

          2urvived  
count  1309.000000  
mean      0.261268  
std       0.439494  
min       0.000000  
25%       0.000000  
50%       0.000000  
75%       1.000000  
max       1.000000  

[8 rows x 28 columns]
```

In [6]:

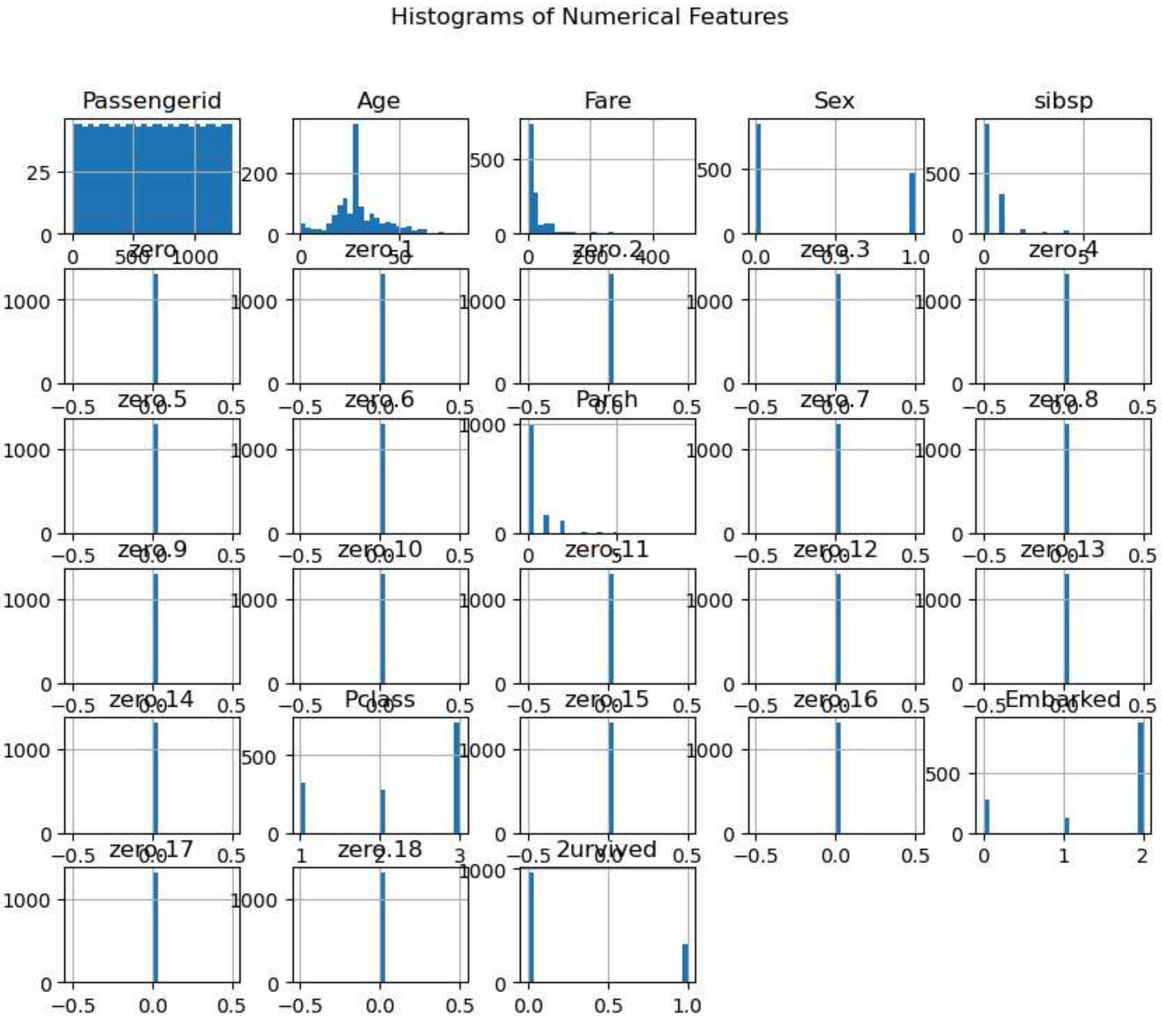
```
#to check for missing values
print(df.isnull().sum())

#to visualize missing values
sns.heatmap(df.isnull(), cbar=False, cmap="viridis")
plt.title("Missing Values Heatmap")
plt.show()
```

```
Passengerid    0
Age            0
Fare           0
Sex            0
sibsp          0
zero           0
zero.1         0
zero.2         0
zero.3         0
zero.4         0
zero.5         0
zero.6         0
Parch          0
zero.7         0
zero.8         0
zero.9         0
zero.10        0
zero.11        0
zero.12        0
zero.13        0
zero.14        0
Pclass         0
zero.15        0
zero.16        0
Embarked       2
zero.17        0
zero.18        0
2urvived       0
dtype: int64
```


Missing Values Heatmap
0
51
102
153
204
255
306
357
408
459
510
561
612
663
714
765
816
867
918
969
1020
1071
1122
1173
1224
1275
Passengerid
Age
Fare
Sex
sibsp
zero
zero.1
zero.2
zero.3
zero.4
zero.5
zero.6
Parch
zero.7
zero.8
zero.9
zero.10
zero.11
zero.12
zero.13
zero.14
Pclass
zero.15
zero.16
Embarked
zero.17
zero.18
2urvived

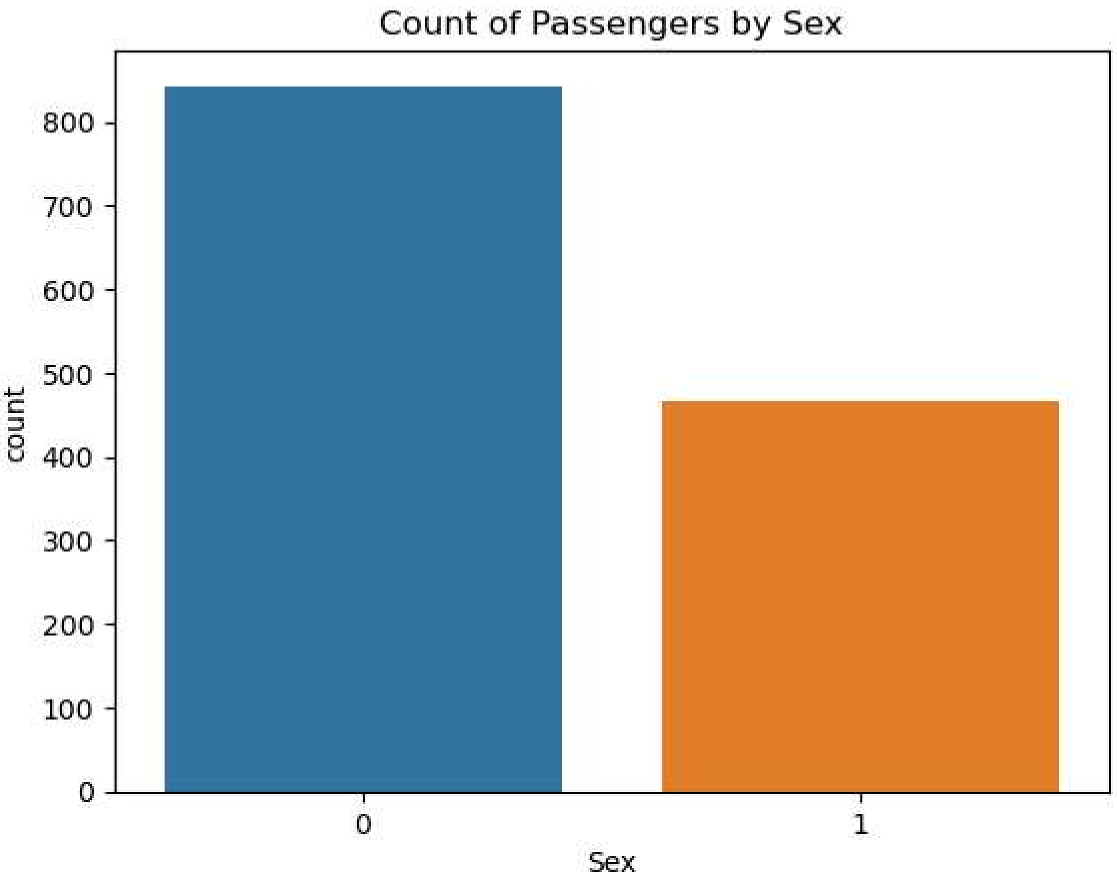
In [7]:

```
# Histograms for numerical features
df.hist(bins=30, figsize=(10, 8))
plt.suptitle("Histograms of Numerical Features")
plt.show()
```



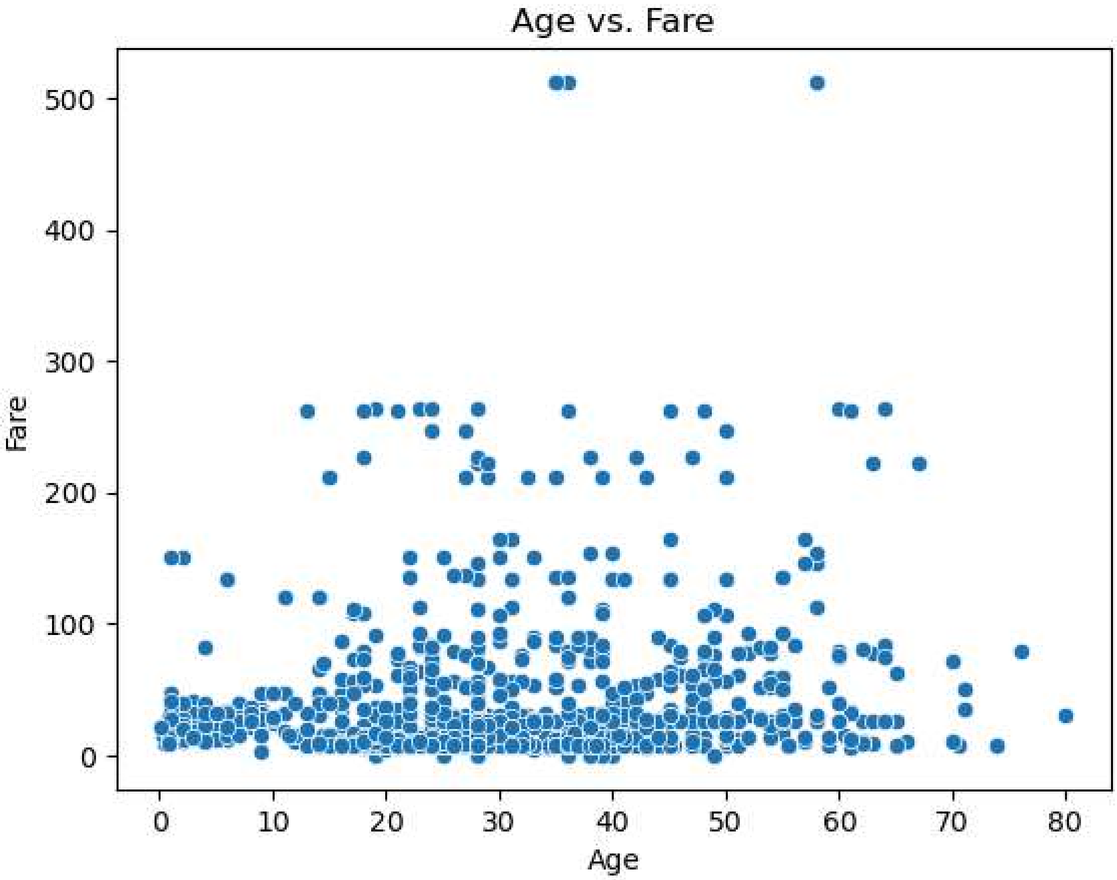
In [9]:

```
# Count plot for a categorical column (e.g., 'sex')
sns.countplot(x='Sex', data=df)
plt.title("Count of Passengers by Sex")
plt.show()
```



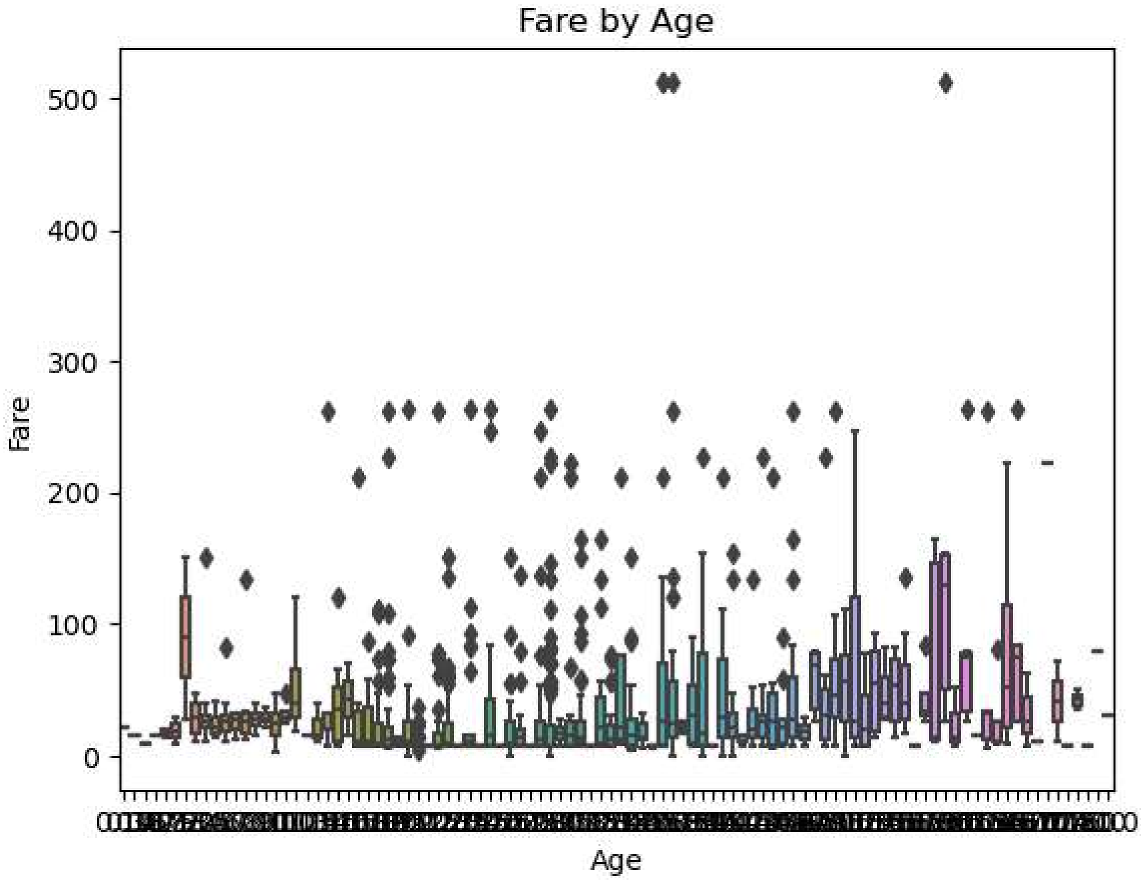
In [10]:

```
# Scatter plot between 'age' and 'fare'
sns.scatterplot(x='Age', y='Fare', data=df)
plt.title("Age vs. Fare")
plt.show()
```



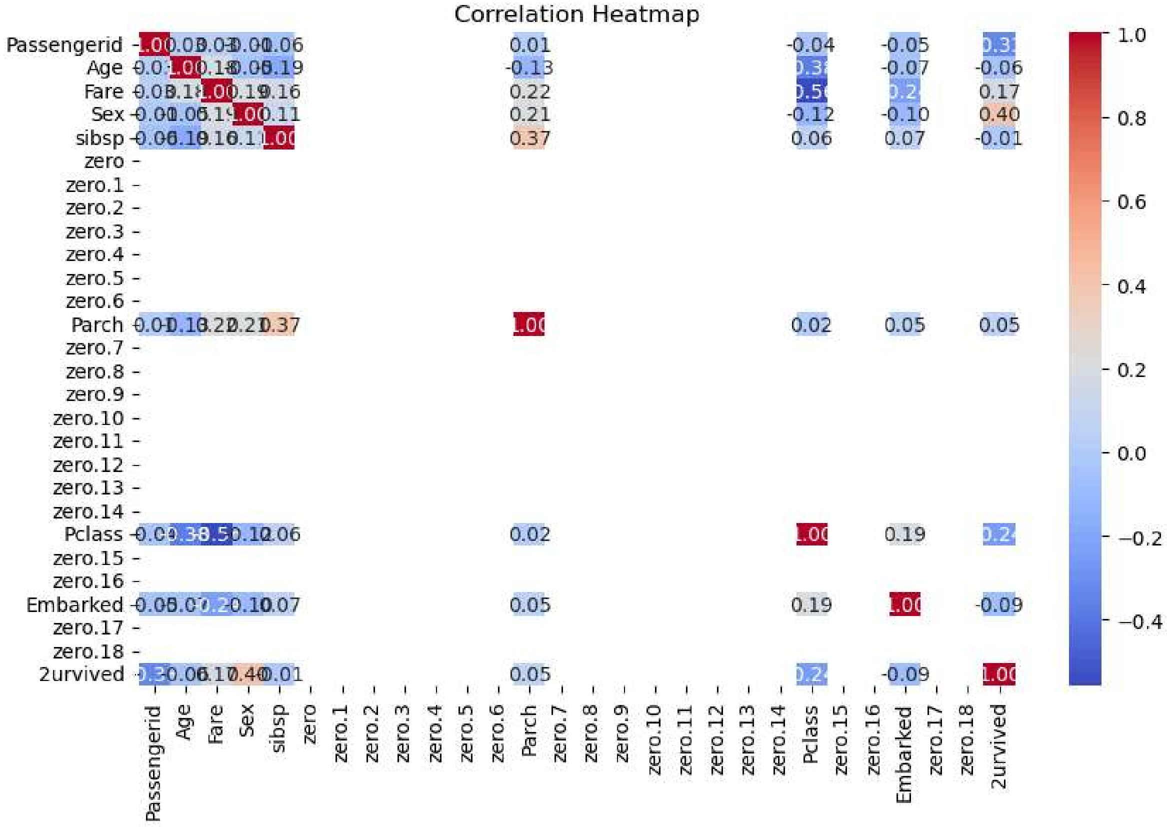
In [13]:

```
# Box plot for 'age' based on 'Fare'
sns.boxplot(x='Age', y='Fare', data=df)
plt.title("Fare by Age")
plt.show()
```



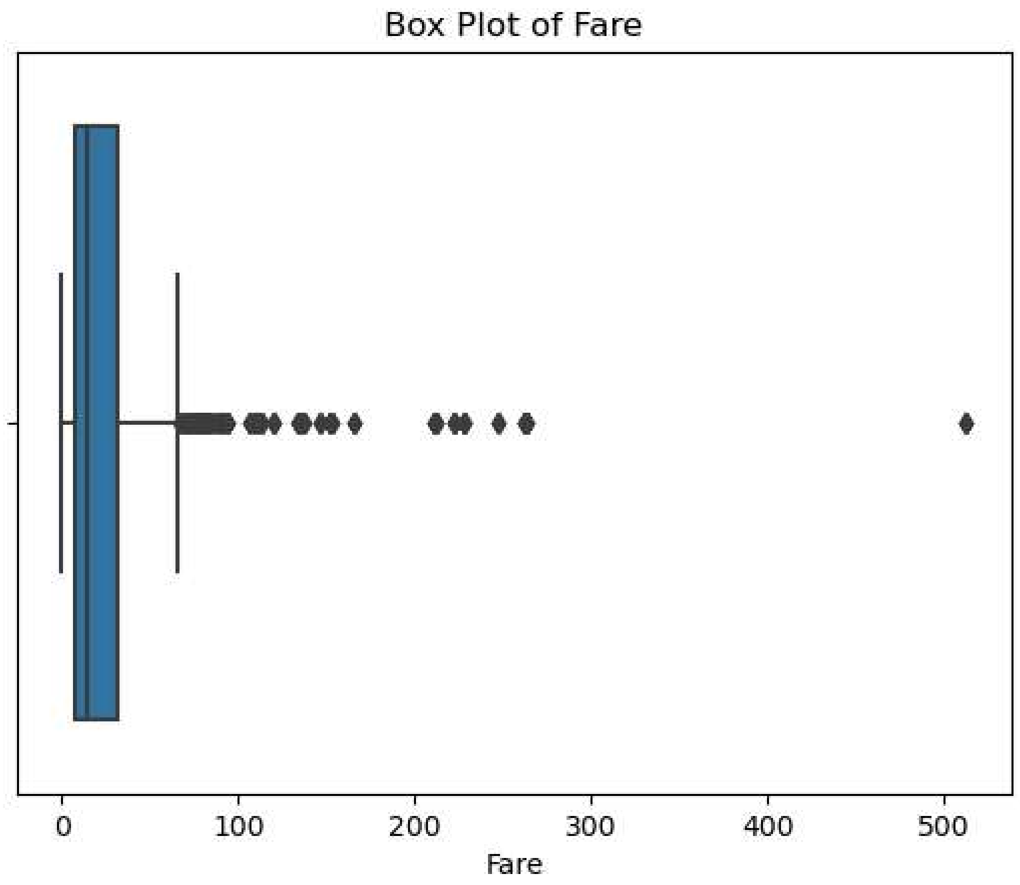
In [14]:

```
# Correlation heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title("Correlation Heatmap")
plt.show()
```



In [16]:

```
# Box plot for outliers
sns.boxplot(x=df['Fare'])
plt.title("Box Plot of Fare")
plt.show()

# Z-score to detect outliers
from scipy.stats import zscore
df['fare_zscore'] = zscore(df['Fare'])
outliers = df[np.abs(df['fare_zscore']) > 3]
print(outliers)
```



```
      Passengerid   Age      Fare  Sex  sibsp  zero  zero.1  zero.2  zero.3  
\
27             28  19.0  263.0000    0      3     0       0       0       0
88             89  23.0  263.0000    1      3     0       0       0       0
118           119  24.0  247.5208    0      0     0       0       0       0
258           259  35.0  512.3292    1      0     0       0       0       0
299           300  50.0  247.5208    1      0     0       0       0       0
311           312  18.0  262.3750    1      2     0       0       0       0
341           342  24.0  263.0000    1      3     0       0       0       0
377           378  27.0  211.5000    0      0     0       0       0       0
380           381  42.0  227.5250    1      0     0       0       0       0
438           439  64.0  263.0000    0      1     0       0       0       0
527           528  28.0  221.7792    0      0     0       0       0       0
557           558  28.0  227.5250    0      0     0       0       0       0
679           680  36.0  512.3292    0      0     0       0       0       0
689           690  15.0  211.3375    1      0     0       0       0       0
700           701  18.0  227.5250    1      1     0       0       0       0
716           717  38.0  227.5250    1      0     0       0       0       0
730           731  29.0  211.3375    1      0     0       0       0       0
737           738  35.0  512.3292    0      0     0       0       0       0
742           743  21.0  262.3750    1      2     0       0       0       0
779           780  43.0  211.3375    1      0     0       0       0       0
915           916  48.0  262.3750    1      1     0       0       0       0
944           945  28.0  263.0000    1      3     0       0       0       0
950           951  36.0  262.3750    1      0     0       0       0       0
955           956  13.0  262.3750    0      2     0       0       0       0
960           961  60.0  263.0000    1      1     0       0       0       0
965           966  35.0  211.5000    1      0     0       0       0       0
966           967  32.5  211.5000    0      0     0       0       0       0
972           973  67.0  221.7792    0      1     0       0       0       0
1005         1006  63.0  221.7792    1      1     0       0       0       0
1033         1034  61.0  262.3750    0      1     0       0       0       0
1047         1048  29.0  221.7792    1      0     0       0       0       0
1075         1076  27.0  247.5208    1      1     0       0       0       0
1093         1094  47.0  227.5250    0      1     0       0       0       0
1109         1110  50.0  211.5000    1      1     0       0       0       0
1215         1216  39.0  211.3375    1      0     0       0       0       0
1234         1235  58.0  512.3292    1      0     0       0       0       0
1266         1267  45.0  262.3750    1      0     0       0       0       0
1298         1299  50.0  211.5000    0      1     0       0       0       0

      zero.4  ...  zero.13  zero.14  Pclass  zero.15  zero.16  Embarked  \
27         0  ...        0        0       1        0        0       2.0   
88         0  ...        0        0       1        0        0       2.0   
118        0  ...        0        0       1        0        0       0.0   
258        0  ...        0        0       1        0        0       0.0   
299        0  ...        0        0       1        0        0       0.0   
311        0  ...        0        0       1        0        0       0.0   
341        0  ...        0        0       1        0        0       2.0   
377        0  ...        0        0       1        0        0       0.0   
380        0  ...        0        0       1        0        0       0.0   
438        0  ...        0        0       1        0        0       2.0   
527        0  ...        0        0       1        0        0       2.0   
557        0  ...        0        0       1        0        0       0.0   
679        0  ...        0        0       1        0        0       0.0   
689        0  ...        0        0       1        0        0       2.0   
700        0  ...        0        0       1        0        0       0.0   
```

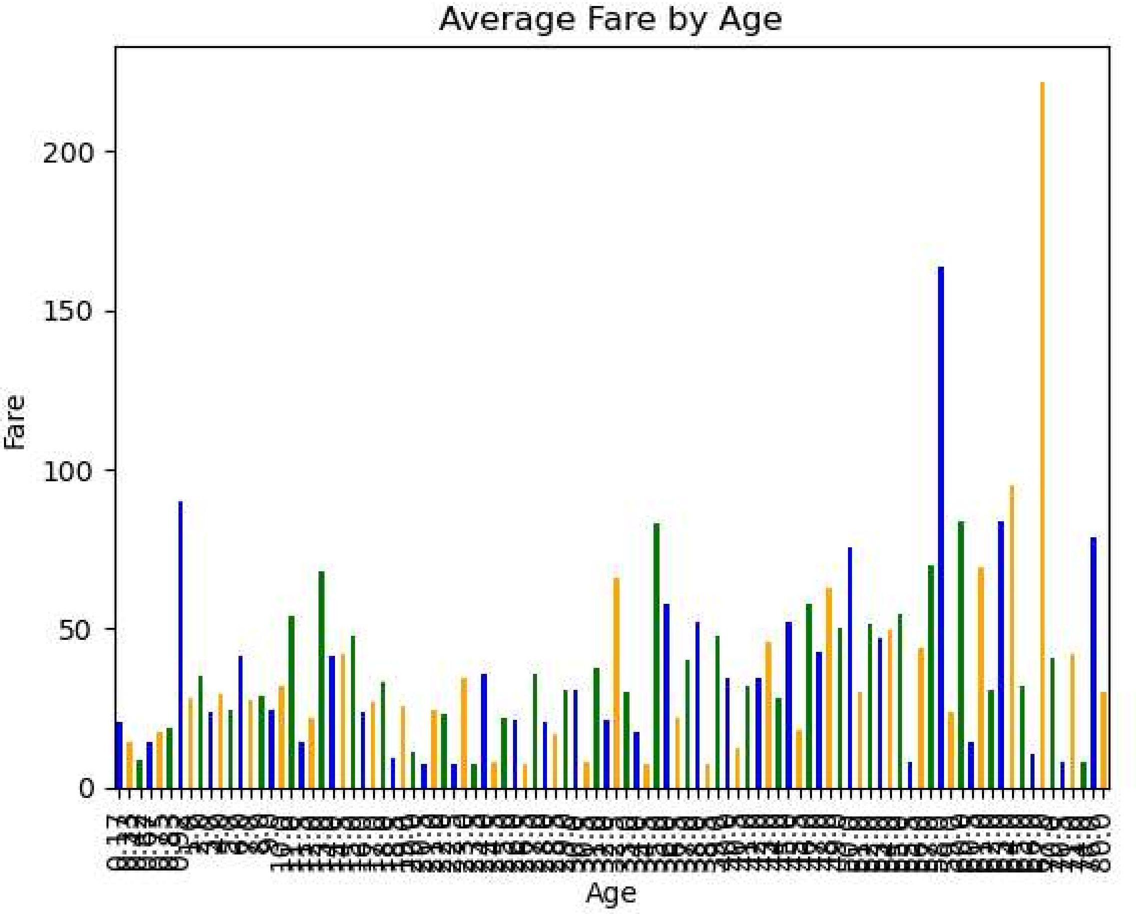
```
716        0  ...        0        0       1        0        0       0.0
730        0  ...        0        0       1        0        0       2.0
737        0  ...        0        0       1        0        0       0.0
742        0  ...        0        0       1        0        0       0.0
779        0  ...        0        0       1        0        0       2.0
915        0  ...        0        0       1        0        0       0.0
944        0  ...        0        0       1        0        0       2.0
950        0  ...        0        0       1        0        0       0.0
955        0  ...        0        0       1        0        0       0.0
960        0  ...        0        0       1        0        0       2.0
965        0  ...        0        0       1        0        0       0.0
966        0  ...        0        0       1        0        0       0.0
972        0  ...        0        0       1        0        0       2.0
1005       0  ...        0        0       1        0        0       2.0
1033       0  ...        0        0       1        0        0       0.0
1047       0  ...        0        0       1        0        0       2.0
1075       0  ...        0        0       1        0        0       0.0
1093       0  ...        0        0       1        0        0       0.0
1109       0  ...        0        0       1        0        0       0.0
1215       0  ...        0        0       1        0        0       2.0
1234       0  ...        0        0       1        0        0       0.0
1266       0  ...        0        0       1        0        0       0.0
1298       0  ...        0        0       1        0        0       0.0

      zero.17  zero.18  2urvived  fare_zscore
27          0        0         0     4.441439
88          0        0         1     4.441439
118         0        0         0     4.142160
258         0        0         1     9.262028
299         0        0         1     4.142160
311         0        0         1     4.429355
341         0        0         1     4.441439
377         0        0         0     3.445726
380         0        0         1     3.755557
438         0        0         0     4.441439
527         0        0         0     3.644466
557         0        0         0     3.755557
679         0        0         1     9.262028
689         0        0         1     3.442584
700         0        0         1     3.755557
716         0        0         1     3.755557
730         0        0         1     3.442584
737         0        0         1     9.262028
742         0        0         1     4.429355
779         0        0         1     3.442584
915         0        0         0     4.429355
944         0        0         0     4.441439
950         0        0         0     4.429355
955         0        0         0     4.429355
960         0        0         0     4.441439
965         0        0         0     3.445726
966         0        0         0     3.445726
972         0        0         0     3.644466
1005        0        0         0     3.644466
1033        0        0         0     4.429355
1047        0        0         0     3.644466
1075        0        0         0     4.142160
```

```
1093        0        0         0     3.755557
1109        0        0         0     3.445726
1215        0        0         0     3.442584
1234        0        0         0     9.262028
1266        0        0         0     4.429355
1298        0        0         0     3.445726

[38 rows x 29 columns]
```

In [20]:

```
# Group by 'age' and calculate mean fare
age_fare = df.groupby('Age')['Fare'].mean()
print(age_fare)

# Bar plot for grouped data
age_fare.plot(kind='bar', color=['blue', 'orange', 'green'])
plt.title("Average Fare by Age")
plt.ylabel("Fare")
plt.show()
```

```
Age
0.17     20.575000
0.33     14.400000
0.42      8.516700
0.67     14.500000
0.75     17.430533
           ...
70.50     7.750000
71.00    42.079200
74.00     7.775000
76.00    78.850000
80.00    30.000000
Name: Fare, Length: 98, dtype: float64
```



In [21]:

```
# Save cleaned dataset
df.to_csv('cleaned_dataset.csv', index=False)
```

In [ ]: