In [4]:
```
# import necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [6]:
```
# load database
df=pd.read_csv("test.csv")
```

In [7]:
```
# basic info
print("dataset info")
print(df.info(),"\n")
```

```
dataset info
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Pclass       418 non-null    int64  
 2   Name         418 non-null    object 
 3   Sex          418 non-null    object 
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64  
 6   Parch        418 non-null    int64  
 7   Ticket       418 non-null    object 
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object 
 10  Embarked     418 non-null    object 
dtypes: float64(2), int64(4), object(5)
memory usage: 36.1+ KB
None
```

In [9]:
```
# display the 1st 5 rows
print("first 5 rows")
print(df.head(),"\n")
```

```
first 5 rows
   PassengerId  Pclass                                          Name     Sex  \
0          892       3                              Kelly, Mr. James    male   
1          893       3              Wilkes, Mrs. James (Ellen Needs)  female   
2          894       2                     Myles, Mr. Thomas Francis    male   
3          895       3                              Wirz, Mr. Albert    male   
4          896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female   

    Age  SibSp  Parch   Ticket     Fare Cabin Embarked  
0  34.5      0      0   330911   7.8292   NaN        Q  
1  47.0      1      0   363272   7.0000   NaN        S  
2  62.0      0      0   240276   9.6875   NaN        Q  
3  27.0      0      0   315154   8.6625   NaN        S  
4  22.0      1      1  3101298  12.2875   NaN        S  
```

In [12]:
```
# display the statitics
print("statistical summary")
print(df.describe(include="all"))
```

```
statistical summary
        PassengerId      Pclass                      Name   Sex         Age  \
count    418.000000  418.000000                       418   418  332.000000
unique          NaN         NaN                       418     2         NaN
top             NaN         NaN  Peter, Master. Michael J  male         NaN
freq            NaN         NaN                         1   266         NaN
mean    1100.500000    2.265550                       NaN   NaN   30.272590
std      120.810458    0.841838                       NaN   NaN   14.181209
min      892.000000    1.000000                       NaN   NaN    0.170000
25%      996.250000    1.000000                       NaN   NaN   21.000000
50%     1100.500000    3.000000                       NaN   NaN   27.000000
75%     1204.750000    3.000000                       NaN   NaN   39.000000
max     1309.000000    3.000000                       NaN   NaN   76.000000

             SibSp       Parch    Ticket        Fare            Cabin Embarked
count   418.000000  418.000000       418  417.000000               91      418
unique         NaN         NaN       363         NaN               76        3
top            NaN         NaN  PC 17608         NaN  B57 B59 B63 B66        S
freq           NaN         NaN         5         NaN                3      270
mean      0.447368    0.392344       NaN   35.627188              NaN      NaN
std       0.896760    0.981429       NaN   55.907576              NaN      NaN
min       0.000000    0.000000       NaN    0.000000              NaN      NaN
25%       0.000000    0.000000       NaN    7.895800              NaN      NaN
50%       0.000000    0.000000       NaN   14.454200              NaN      NaN
75%       1.000000    0.000000       NaN   31.500000              NaN      NaN
max       8.000000    9.000000       NaN  512.329200              NaN      NaN
```

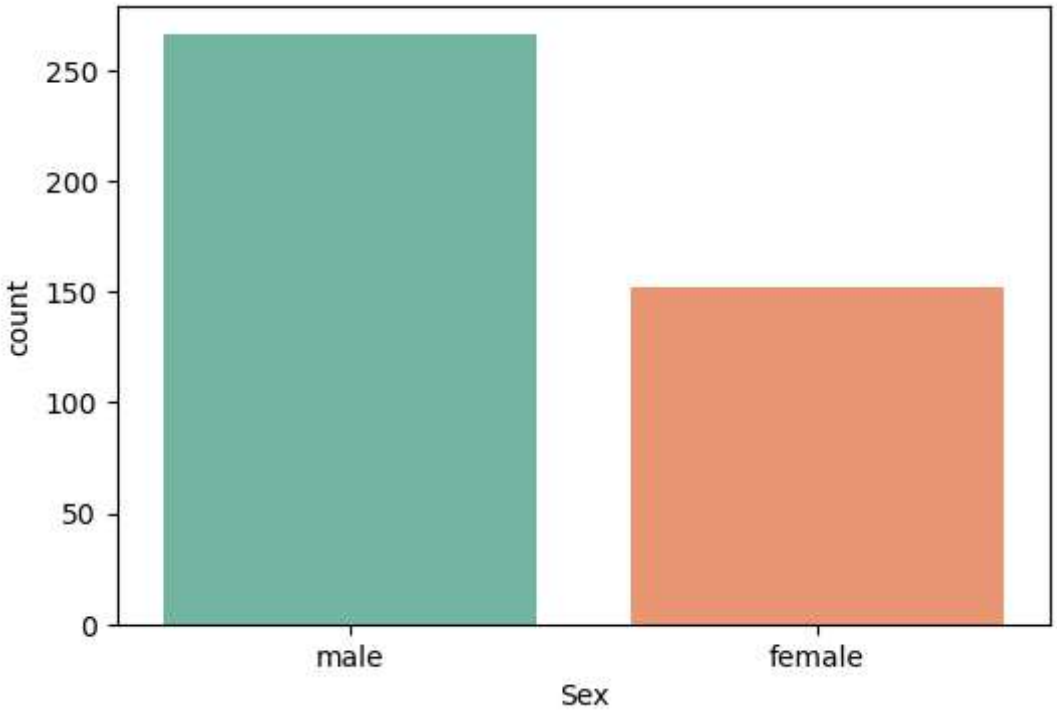
In [13]:
```
# missing values
print("\n ? missingvalue")
print(df.isnull().sum())
```

```
 ? missingvalue
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

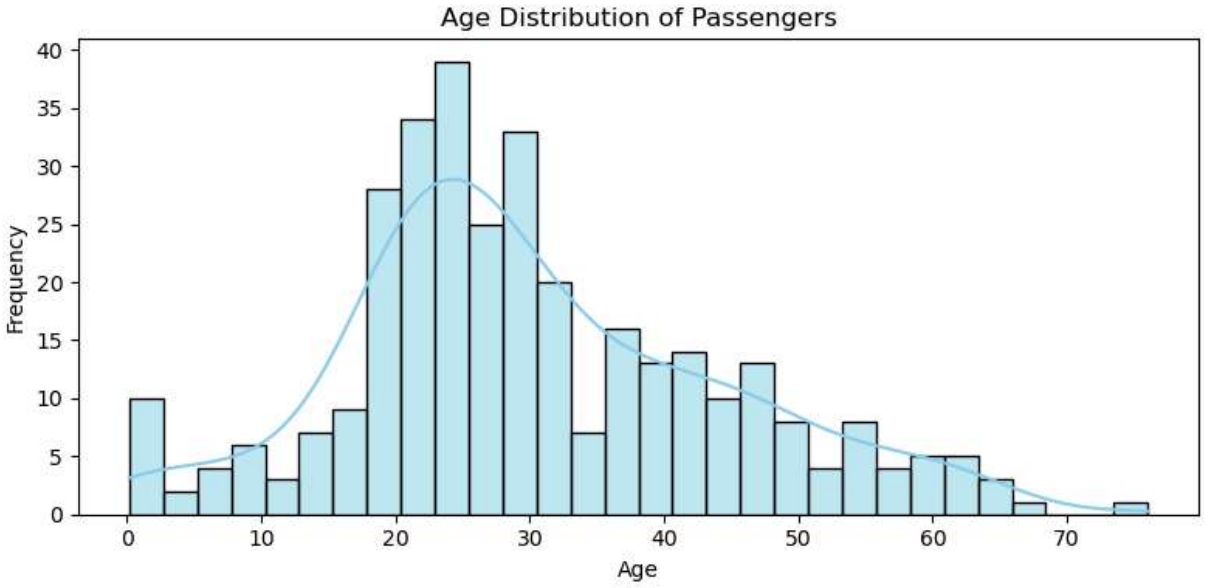
In [36]:
```
# count of the passengers with sex
plt.figure(figsize=(6,4))
sns.countplot(data=df,x='Sex', palette='Set2')
plt.tittle("passengers count with sex")
plt.xlabel('sex')
plt.ylabel('count')
```

```
plt.taight_layout()
plt.show()
```

```
C:\Users\Admins\AppData\Local\Temp\ipykernel_13276\3186592205.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.1
4.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(data=df,x='Sex', palette='Set2')
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
Cell In[36], line 4
      2 plt.figure(figsize=(6,4))
      3 sns.countplot(data=df,x='Sex', palette='Set2')
----> 4 plt.tittle("passengers count with sex")
      5 plt.xlabel('sex')
      6 plt.ylabel('count')

AttributeError: module 'matplotlib.pyplot' has no attribute 'tittle'
```

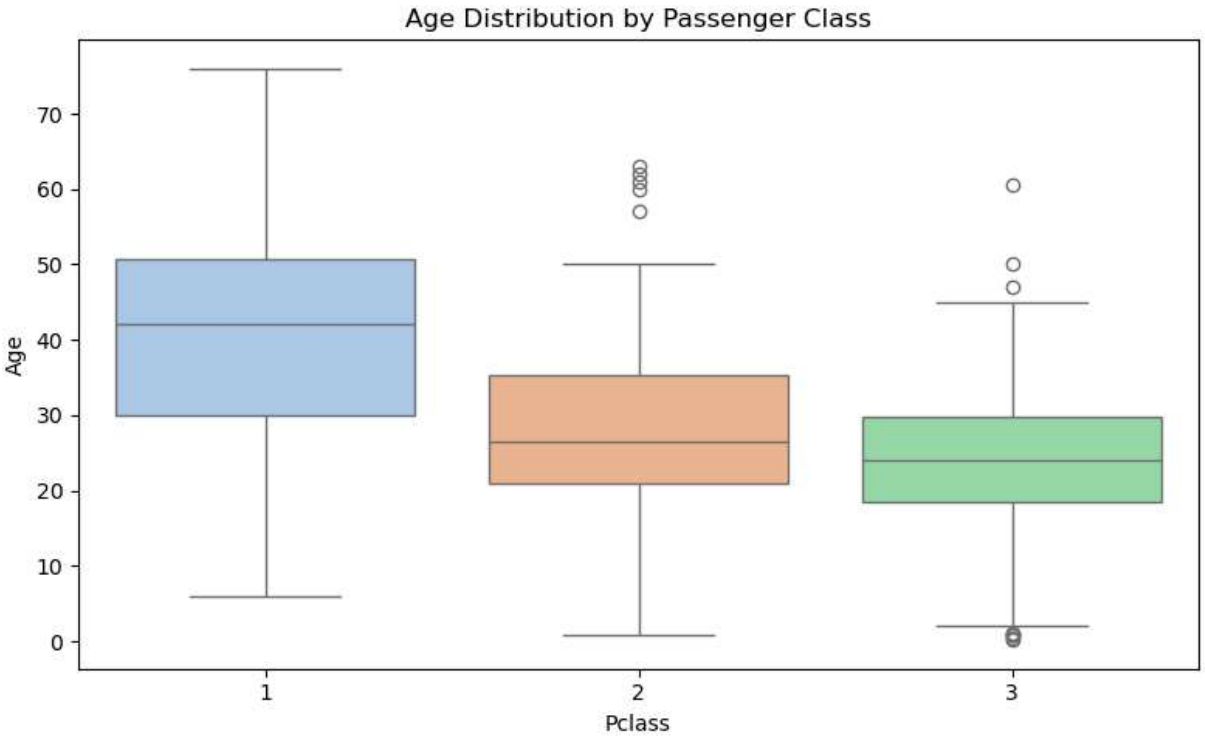


In [35]:
```
# visual age distribution
plt.figure(figsize=(8, 4))
sns.histplot(df['Age'].dropna(), bins=30, kde=True, color='skyblue')
plt.title('Age Distribution of Passengers')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

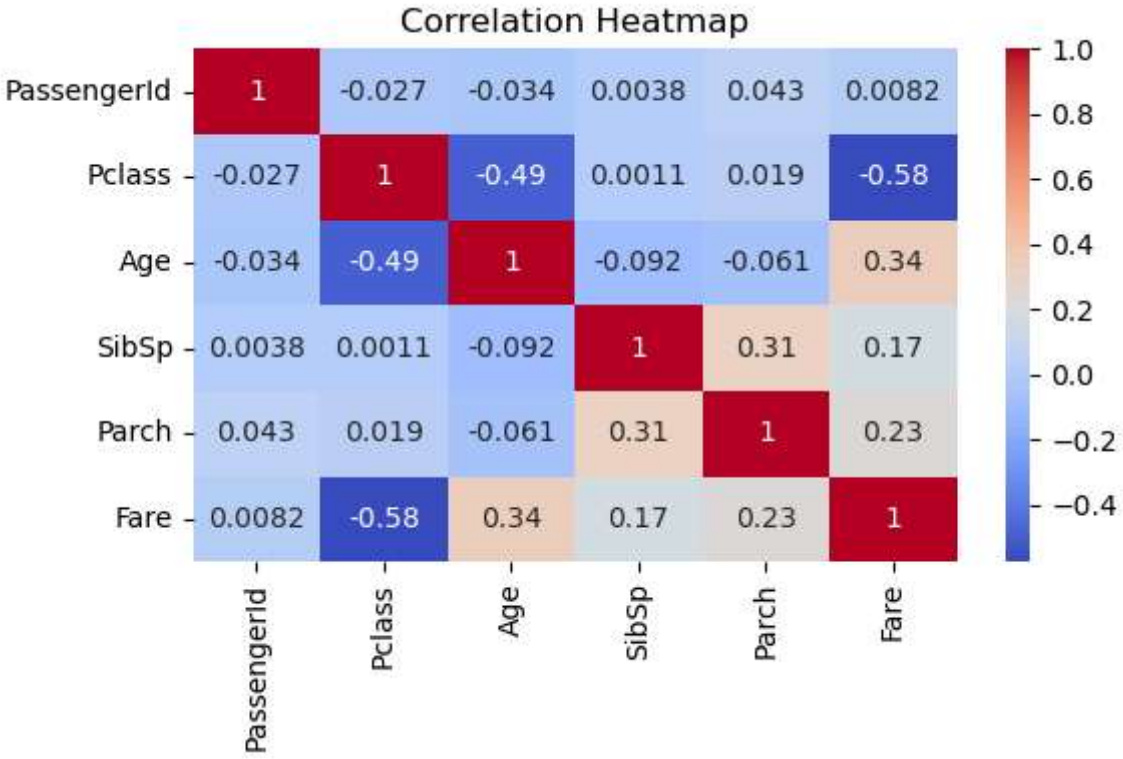


In [40]:
```
# pclass vs age
plt.figure(figsize=(8, 5))
sns.boxplot(x='Pclass', y='Age', data=df, palette='pastel')
plt.title('Age Distribution by Passenger Class')
plt.tight_layout()
plt.show()
```

```
C:\Users\Admins\AppData\Local\Temp\ipykernel_13276\1970821199.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.1
4.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x='Pclass', y='Age', data=df, palette='pastel')
```



In [43]:
```
# heat map
plt.figure(figsize=(6, 4))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.tight_layout()
plt.show()
```



In [44]:
```
# Summary of insights
print("\n EDA Summary:")
print("""
- The dataset contains test set data for Titanic survivors.
- Missing values commonly found in 'Age', 'Fare', and 'Cabin'.
- Most passengers in this test set are in Pclass 3.
- Majority of passengers are male.
- There's some correlation between Fare and Pclass.
""")
```

```
 EDA Summary:

- The dataset contains test set data for Titanic survivors.
- Missing values commonly found in 'Age', 'Fare', and 'Cabin'.
- Most passengers in this test set are in Pclass 3.
- Majority of passengers are male.
- There's some correlation between Fare and Pclass.
```

In [ ]: