1. Set Up Jupyter Notebook and Import Libraries

In [3]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib
import os

# Load the cleaned Titanic dataset
path = r"C:\Users\Asus\Music\achievement 6 project"
data = pd.read_csv(os.path.join(path, "Data", "tested.csv"))

# Display the first few rows of the dataset
data.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

2. Pick Variables for Exploratory Analysis

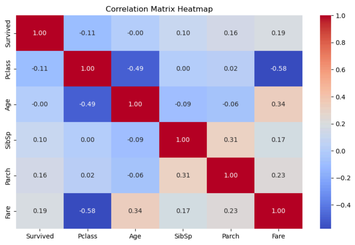
In [6]:
```
# Drop non-essential columns
data = data.drop(columns=['PassengerId', 'Name'])
```

3. Create a Correlation Matrix Heatmap

In [11]:
```
# Select only the numeric columns for the correlation matrix
numeric_data = data.select_dtypes(include=[np.number])

# Compute the correlation matrix
corr_matrix = numeric_data.corr()

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



Interpretation:

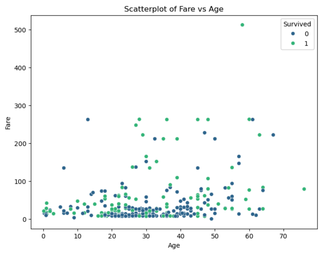
Each cell in the heatmap shows the correlation coefficient between two variables.

A correlation coefficient close to +1 or -1 indicates a strong relationship, while a coefficient close to 0 indicates a weak relationship.

For example, if Fare and Pclass show a negative correlation, it implies that higher-class passengers (lower Pclass values) paid more for their tickets (Fare).

4. Create Scatterplots for Strong Correlations

In [15]:
```
# Scatterplot for Fare vs Age
plt.figure(figsize=(8, 6))
sns.scatterplot(data=data, x='Age', y='Fare', hue='Survived', palette='viridis')
plt.title('Scatterplot of Fare vs Age')
plt.show()
```



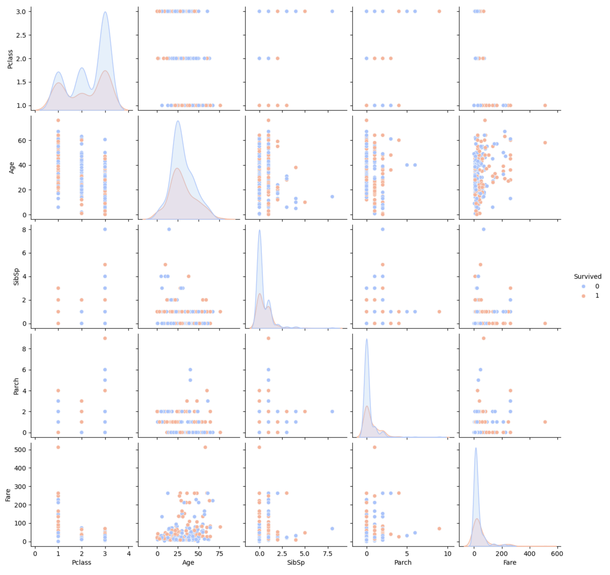
Interpretation:

This scatterplot shows the relationship between Age and Fare.

The hue for Survived helps us see how survival rates may vary across different ages and fares.

5. Create a Pair Plot of the Entire Dataset

In [19]:
```
# Pair plot of the dataset
sns.pairplot(data, hue='Survived', palette='coolwarm')
plt.show()
```



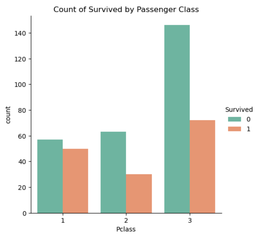
Interpretation:

Pair plots allow us to see the distribution of single variables and relationships between two variables.

Look for clusters or patterns that might indicate correlations or groupings.

6. Create a Categorical Plot

In [23]:
```
# Categorical plot for Pclass and Survived
sns.catplot(x='Pclass', hue='Survived', kind='count', data=data, palette='Set2')
plt.title('Count of Survived by Passenger Class')
plt.show()
```



Interpretation:

This plot shows the survival rate across different passenger classes (Pclass).

Notice which classes had higher survival rates and speculate why this might be the case (e.g., wealthier passengers in higher classes had better access to lifeboats).

7. Revisit and Answer Questions

Based on the visual exploration:

Which classes had the highest survival rates?

Did age or fare affect survival chances?

Were there any significant correlations?

8. Define Hypotheses

Based on my visual exploration, you might hypothesize:

"Passengers in higher classes (Pclass = 1) had higher survival rates."

"Women and children had higher survival rates than men."

You can later test these hypotheses with statistical analysis.