In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid", palette="pastel")
```

In [2]:
```
df=pd.read_csv("train.csv")
```

In [3]:
```
df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 66.2+ KB
```

In [4]:
```
print(df.describe(include='all'))
```

```
        PassengerId    Survived      Pclass  \
count    891.000000  891.000000  891.000000
unique          NaN         NaN         NaN
top             NaN         NaN         NaN
freq            NaN         NaN         NaN
mean     446.000000    0.383838    2.308642
std      257.353842    0.486592    0.836071
min        1.000000    0.000000    1.000000
25%      223.500000    0.000000    2.000000
50%      446.000000    0.000000    3.000000
75%      668.500000    1.000000    3.000000
max      891.000000    1.000000    3.000000

                                                 Name   Sex         Age  \
count                                             891   891  714.000000
unique                                            891     2         NaN
top     de Messemaeker, Mrs. Guillaume Joseph (Emma)  male         NaN
freq                                                1   577         NaN
mean                                              NaN   NaN   29.699118
std                                               NaN   NaN   14.526497
min                                               NaN   NaN    0.420000
25%                                               NaN   NaN   20.125000
50%                                               NaN   NaN   28.000000
75%                                               NaN   NaN   38.000000
max                                               NaN   NaN   80.000000

             SibSp       Parch  Ticket        Fare Cabin Embarked
count   891.000000  891.000000     891  891.000000   204      889
unique         NaN         NaN     681         NaN   147        3
top            NaN         NaN  347082         NaN    G6        S
freq           NaN         NaN       7         NaN     4      644
mean      0.523008    0.381594     NaN   32.204208   NaN      NaN
std       1.102743    0.806057     NaN   49.693429   NaN      NaN
min       0.000000    0.000000     NaN    0.000000   NaN      NaN
25%       0.000000    0.000000     NaN    7.910400   NaN      NaN
50%       0.000000    0.000000     NaN   14.454200   NaN      NaN
75%       1.000000    0.000000     NaN   31.000000   NaN      NaN
max       8.000000    6.000000     NaN  512.329200   NaN      NaN
```

In [5]:
```
print("\nMissing Values:\n", df.isnull().sum())
```

```
Missing Values:
 PassengerId      0
Survived          0
Pclass            0
Name              0
Sex               0
Age             177
SibSp             0
Parch             0
Ticket            0
Fare              0
Cabin           687
Embarked          2
dtype: int64
```

In [6]:
```
for col in df.select_dtypes(include='object').columns:
    print(f"\nValue counts for {col}:\n", df[col].value_counts())
```

```
Value counts for Name:
 de Messemaeker, Mrs. Guillaume Joseph (Emma)          1
Lewy, Mr. Ervin G                                      1
Bradley, Mr. George ("George Arthur Brayton")          1
Hassab, Mr. Hammad                                     1
Coutts, Master. Eden Leslie "Neville"                  1
                                                      ..
Laitinen, Miss. Kristina Sofia                         1
Longley, Miss. Gretchen Fiske                          1
Frauenthal, Mrs. Henry William (Clara Heinsheimer)     1
Arnold-Franchi, Mr. Josef                              1
Quick, Miss. Phyllis May                               1
Name: Name, Length: 891, dtype: int64

Value counts for Sex:
 male      577
female    314
Name: Sex, dtype: int64

Value counts for Ticket:
 347082              7
CA. 2343             7
1601                 7
3101295              6
CA 2144              6
                    ..
28551                1
240929               1
STON/O2. 3101282     1
36864                1
234604               1
Name: Ticket, Length: 681, dtype: int64

Value counts for Cabin:
 G6             4
C23 C25 C27     4
B96 B98         4
F33             3
D               3
               ..
E10             1
D6              1
B73             1
C82             1
D56             1
Name: Cabin, Length: 147, dtype: int64

Value counts for Embarked:
 S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

In [7]:
```
df['Age'].fillna(df['Age'].median(), inplace=True)
```

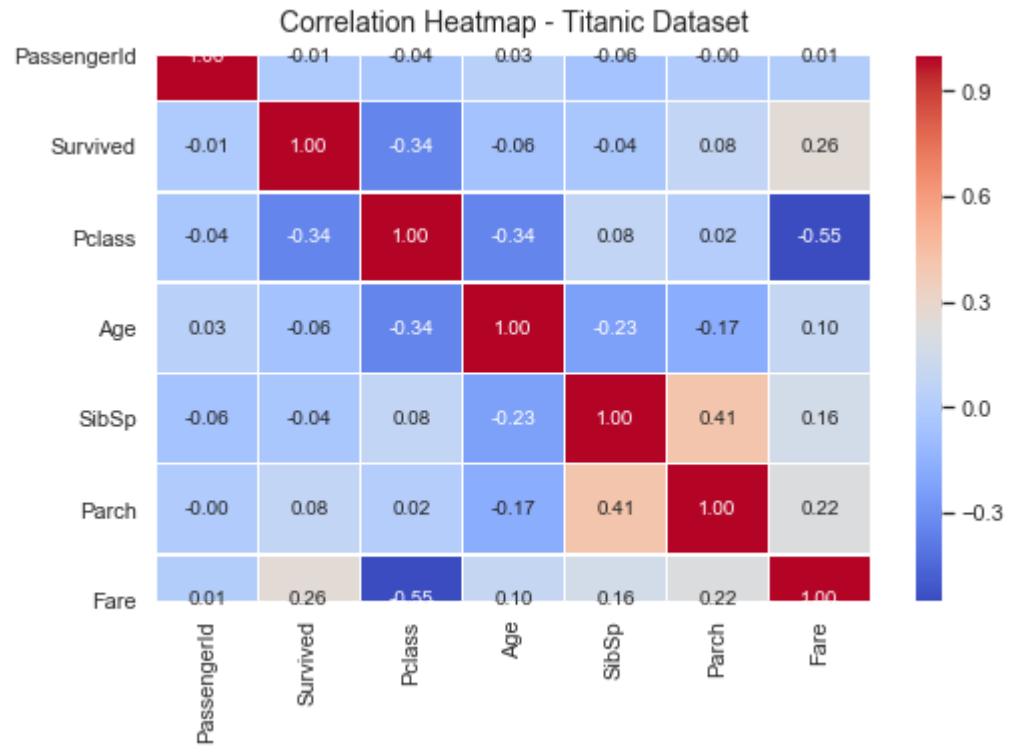
In [8]:
```
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)


df['Cabin'].fillna('Unknown', inplace=True)


df.isnull().sum()
```

Out[8]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

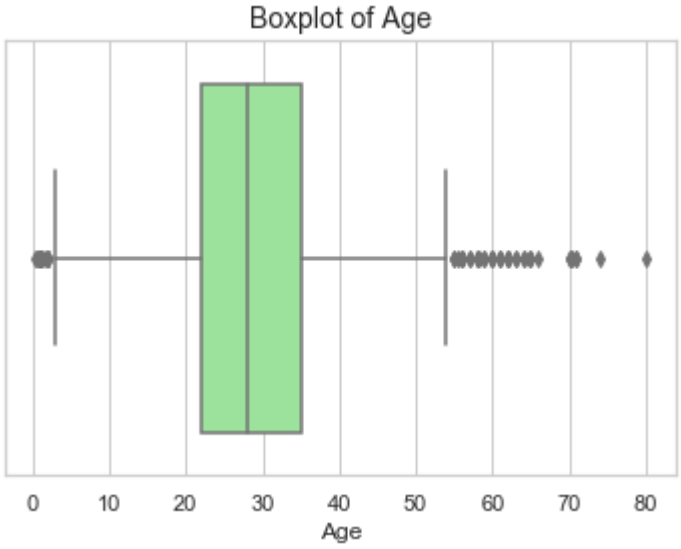
In [9]:
```
import seaborn as sns
import matplotlib.pyplot as plt


pairplot_cols = ['Survived', 'Age', 'Fare', 'Pclass', 'SibSp', 'Parch']


sns.pairplot(df[pairplot_cols], hue='Survived', palette='Set2', diag_kind
='kde')


plt.suptitle("Pair Plot - Titanic Dataset", y=1.02)
plt.show()
```

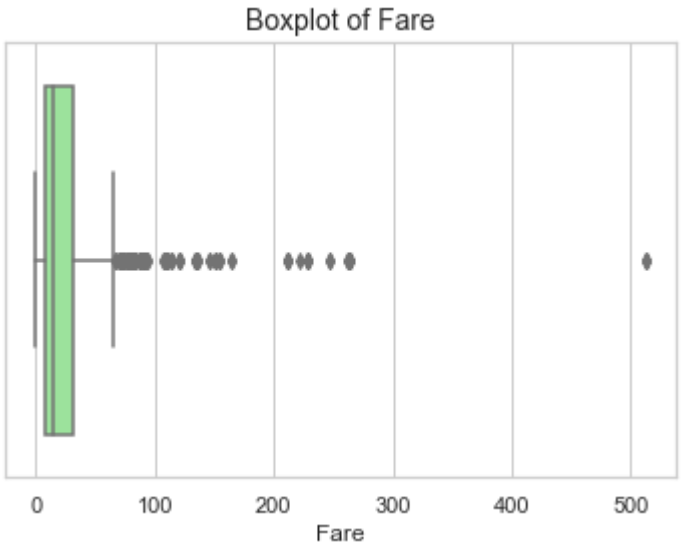
```
C:\Users\Realme\Anaconda3\lib\site-packages\statsmodels\nonparametric\kde.
py:487: RuntimeWarning: invalid value encountered in true_divide
  binned = fast_linbin(X, a, b, gridsize) / (delta * nobs)
C:\Users\Realme\Anaconda3\lib\site-packages\statsmodels\nonparametric\kdet
ools.py:34: RuntimeWarning: invalid value encountered in double_scalars
  FAC1 = 2*(np.pi*bw/RANGE)**2
```

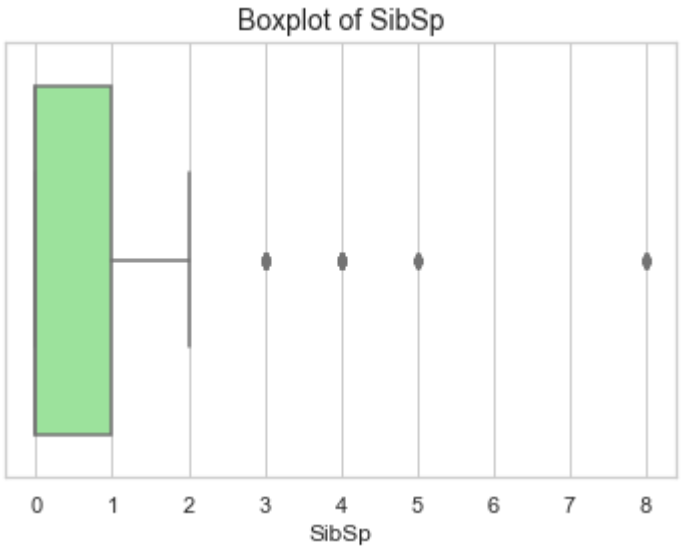
In [10]:
```
numeric_df = df.select_dtypes(include=['int64', 'float64'])

corr = numeric_df.corr()

plt.figure(figsize=(8, 5))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Correlation Heatmap - Titanic Dataset", fontsize=14, pad=12)

plt.show()
```

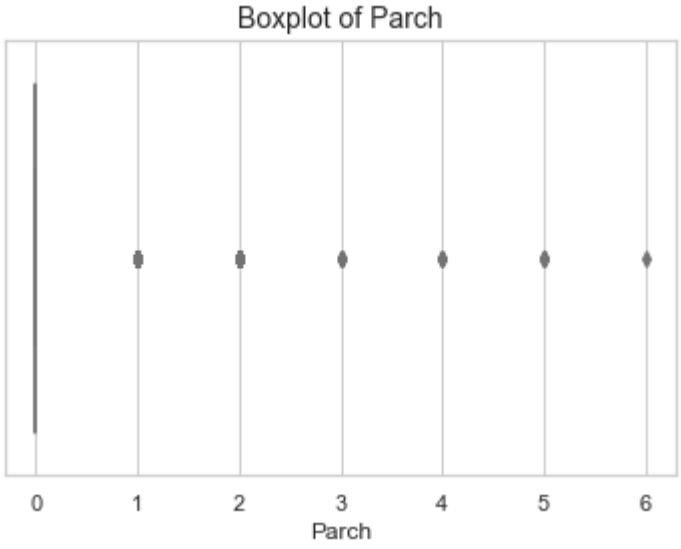


In [15]:
```
# Plot boxplots for numerical columns
num_cols = ['Age', 'Fare', 'SibSp', 'Parch']

for col in num_cols:
    plt.figure(figsize=(6,4))
    sns.boxplot(x=df[col], color='lightgreen')
    plt.title(f"Boxplot of {col}", fontsize=14)
    plt.xlabel(col)
    plt.show()
```


Boxplot of Age
0
10
20
30
40
50
60
70
80
Age


Boxplot of Fare
0
100
200
300
400
500
Fare

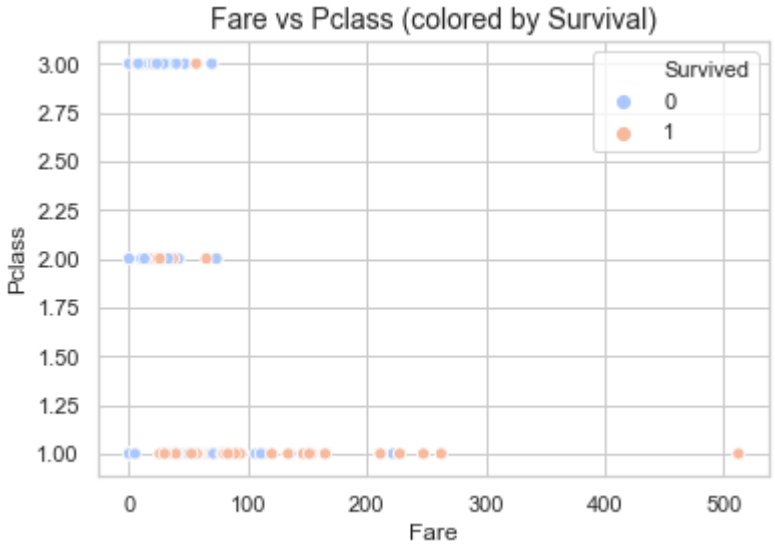

Boxplot of SibSp
0
1
2
3
4
5
6
7
8
SibSp



In [16]:
```
plt.figure(figsize=(6,4))
sns.scatterplot(x='Age', y='Pclass', hue='Survived', data=df, palette='hus
l')
plt.title("Age vs Pclass (colored by Survival)", fontsize=14)
plt.show()
```

In [17]:
```
plt.figure(figsize=(6,4))
sns.scatterplot(x='Age', y='Pclass', hue='Survived', data=df, palette='husl')
plt.title("Age vs Pclass (colored by Survival)", fontsize=14)
plt.show()
```

In [18]:
```
plt.figure(figsize=(6,4))
sns.scatterplot(x='Fare', y='Pclass', hue='Survived', data=df, palette='coolwarm')
plt.title("Fare vs Pclass (colored by Survival)", fontsize=14)
plt.show()
```



In [ ]: