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

In [2]:
```
df = pd.read_csv(r"C:\Users\vallu\Downloads\Titanic_Data.csv")
df
```

Out[2]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1	1	Allen, Miss. Elisabeth Walton	female	29.0000	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1304	3	0	Zabour, Miss. Hileni	female	14.5000	1	0	2665	14.4542	NaN	C	NaN	328.0	NaN
1305	3	0	Zabour, Miss. Thamine	female	NaN	1	0	2665	14.4542	NaN	C	NaN	NaN	NaN
1306	3	0	Zakarian, Mr. Mapriededer	male	26.5000	0	0	2656	7.2250	NaN	C	NaN	304.0	NaN
1307	3	0	Zakarian, Mr. Ortin	male	27.0000	0	0	2670	7.2250	NaN	C	NaN	NaN	NaN
1308	3	0	Zimmerman, Mr. Leo	male	29.0000	0	0	315082	7.8750	NaN	S	NaN	NaN	NaN

1309 rows × 14 columns

In [3]:
```
# Display the first few rows
print(df.head())
```

```
   pclass  survived                                             name     sex  \
0       1         1                    Allen, Miss. Elisabeth Walton  female   
1       1         1                   Allison, Master. Hudson Trevor    male   
2       1         0                     Allison, Miss. Helen Loraine  female   
3       1         0             Allison, Mr. Hudson Joshua Creighton    male   
4       1         0  Allison, Mrs. Hudson J C (Bessie Waldo Daniels)  female   

       age  sibsp  parch  ticket      fare    cabin embarked boat   body  \
0  29.0000      0      0   24160  211.3375       B5        S    2    NaN   
1   0.9167      1      2  113781  151.5500  C22 C26        S   11    NaN   
2   2.0000      1      2  113781  151.5500  C22 C26        S  NaN    NaN   
3  30.0000      1      2  113781  151.5500  C22 C26        S  NaN  135.0   
4  25.0000      1      2  113781  151.5500  C22 C26        S  NaN    NaN   

                         home.dest  
0                     St Louis, MO  
1  Montreal, PQ / Chesterville, ON  
2  Montreal, PQ / Chesterville, ON  
3  Montreal, PQ / Chesterville, ON  
4  Montreal, PQ / Chesterville, ON  
```

In [4]:
```
# Display summary statistics
print(df.describe())
```

```
            pclass     survived          age        sibsp        parch  \
count  1309.000000  1309.000000  1046.000000  1309.000000  1309.000000   
mean      2.294882     0.381971    29.881135     0.498854     0.385027   
std       0.837836     0.486055    14.413500     1.041658     0.865560   
min       1.000000     0.000000     0.166700     0.000000     0.000000   
25%       2.000000     0.000000    21.000000     0.000000     0.000000   
50%       3.000000     0.000000    28.000000     0.000000     0.000000   
75%       3.000000     1.000000    39.000000     1.000000     0.000000   
max       3.000000     1.000000    80.000000     8.000000     9.000000   

              fare        body  
count  1308.000000  121.000000  
mean     33.295479  160.809917  
std      51.758668   97.696922  
min       0.000000    1.000000  
25%       7.895800   72.000000  
50%      14.454200  155.000000  
75%      31.275000  256.000000  
max     512.329200  328.000000  
```

In [5]:
```
# Display information about the dataset
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   pclass     1309 non-null   int64  
 1   survived   1309 non-null   int64  
 2   name       1309 non-null   object 
 3   sex        1309 non-null   object 
 4   age        1046 non-null   float64
 5   sibsp      1309 non-null   int64  
 6   parch      1309 non-null   int64  
 7   ticket     1309 non-null   object 
 8   fare       1308 non-null   float64
 9   cabin      295 non-null    object 
 10  embarked   1307 non-null   object 
 11  boat       486 non-null    object 
 12  body       121 non-null    float64
 13  home.dest  745 non-null    object 
dtypes: float64(3), int64(4), object(7)
memory usage: 143.3+ KB
None
```

In [6]:
```
# Check for missing values
print(df.isnull().sum())
```

```
pclass          0
survived        0
name            0
sex             0
age           263
sibsp           0
parch           0
ticket          0
fare            1
cabin        1014
embarked        2
boat          823
body         1188
home.dest     564
dtype: int64
```

```
In [7]: # Distribution of passengers by class
        print(df['pclass'].value_counts())
```

```
pclass
3    709
1    323
2    277
Name: count, dtype: int64
```

```
In [8]: # Distribution of passengers by gender
        print(df['sex'].value_counts())
```

```
sex
male      843
female    466
Name: count, dtype: int64
```

```
In [9]: # Distribution of passengers by age
        print(df['age'].describe())
```

```
count    1046.000000
mean       29.881135
std        14.413500
min         0.166700
25%        21.000000
50%        28.000000
75%        39.000000
max        80.000000
Name: age, dtype: float64
```

```
In [10]: # Survival rate by class
         print(df.groupby('pclass')['survived'].mean())
```

```
pclass
1    0.619195
2    0.429603
3    0.255289
Name: survived, dtype: float64
```

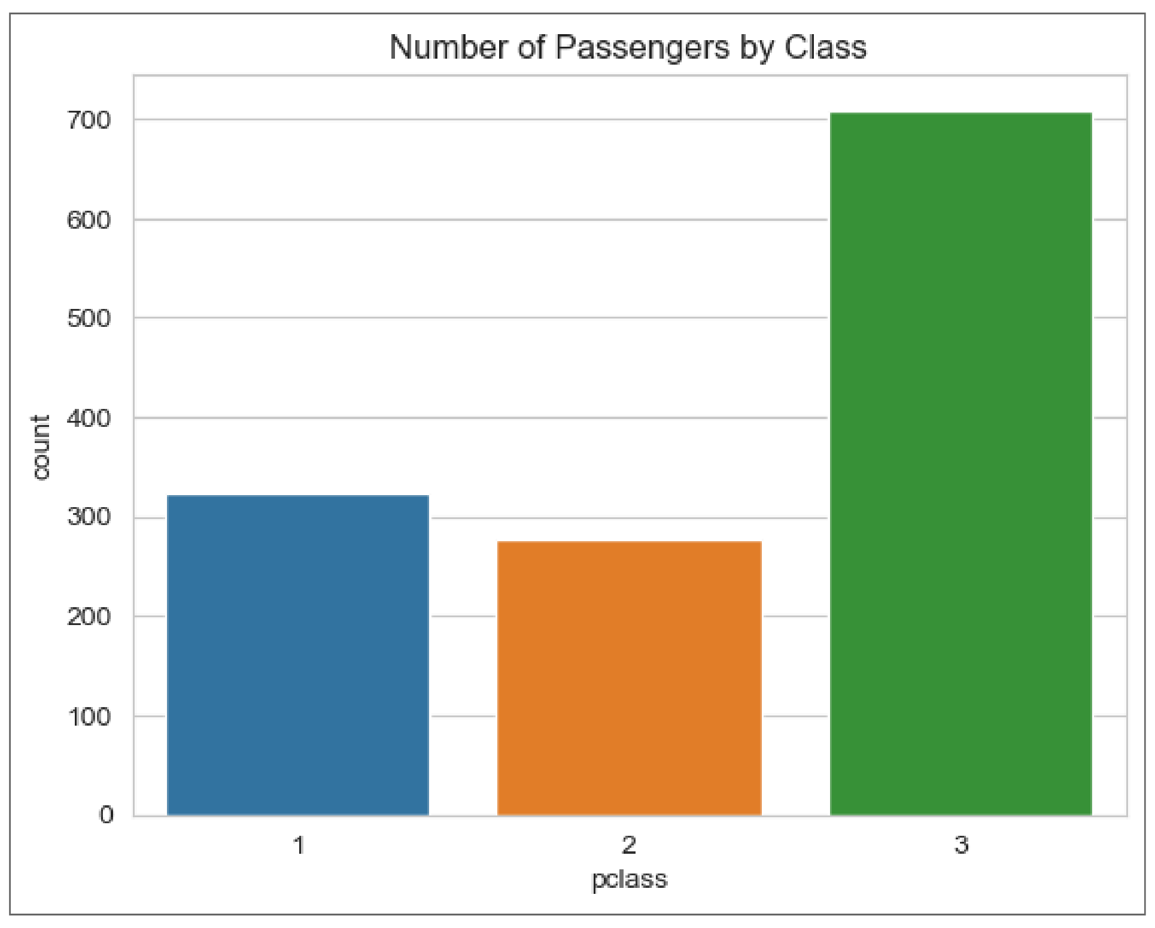
```
In [11]: # Survival rate by gender
         print(df.groupby('sex')['survived'].mean())
```

```
sex
female    0.727468
male      0.190985
Name: survived, dtype: float64
```

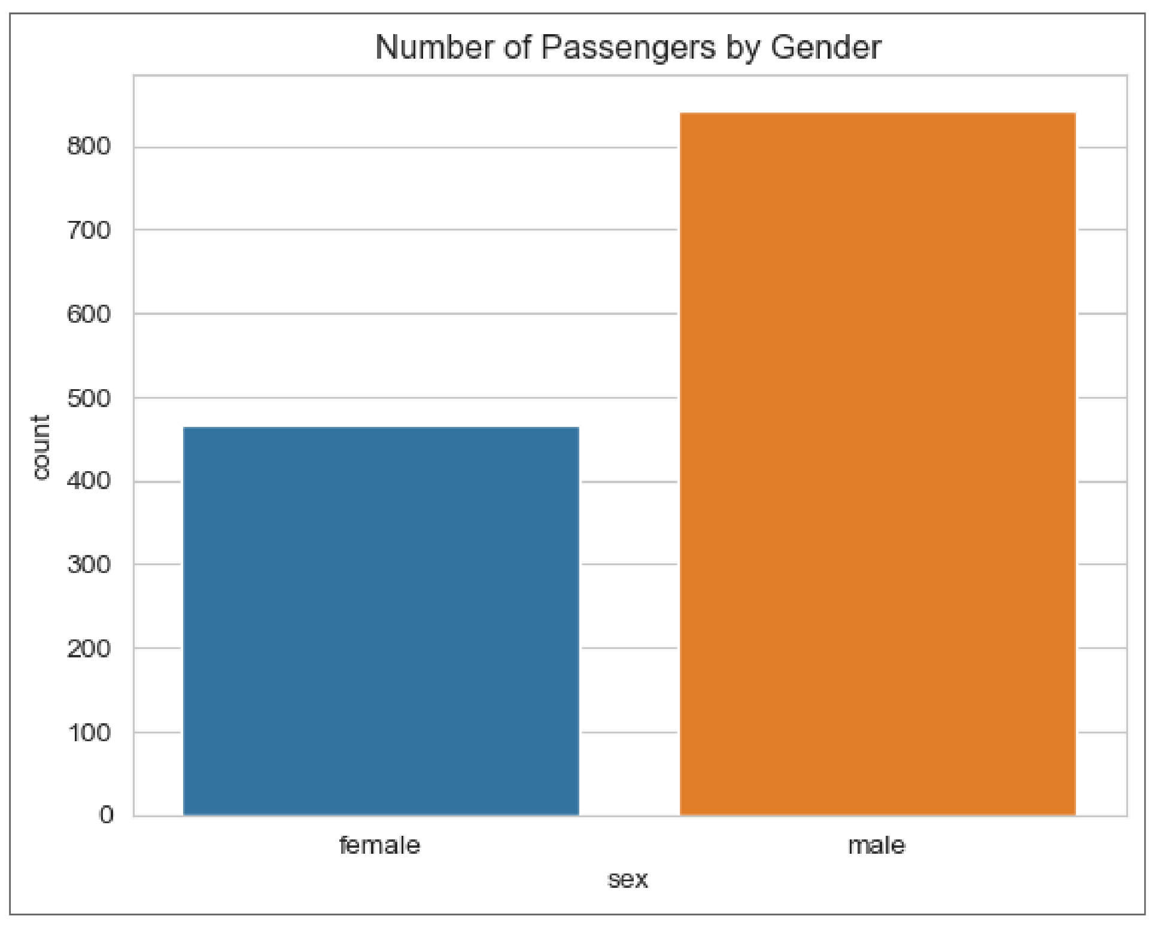
```
In [12]: import matplotlib.pyplot as plt
         import seaborn as sns

# Set the aesthetic style of the plots
sns.set_style('whitegrid')
```

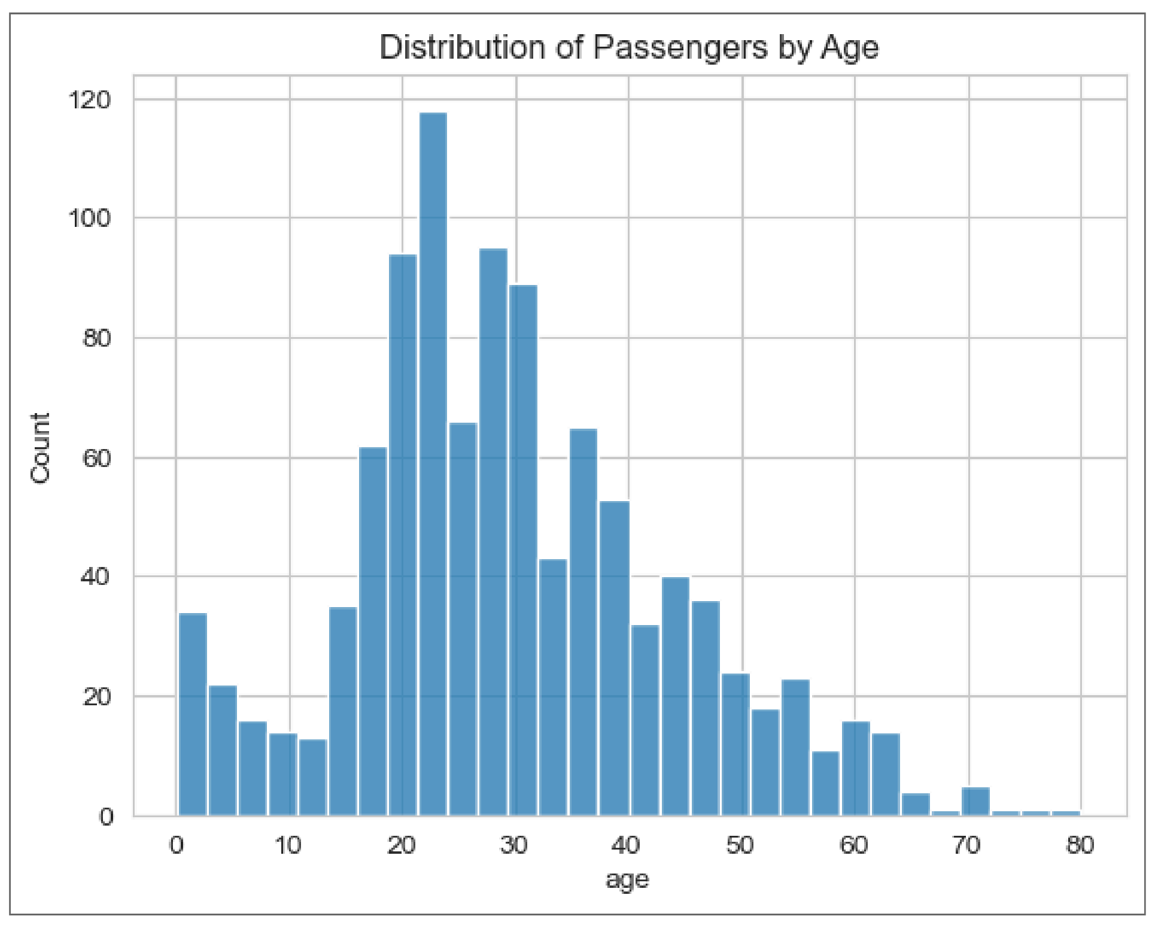
```
In [13]: # Distribution of passengers by class
         sns.countplot(x='pclass', data=df)
plt.title('Number of Passengers by Class')
plt.show()
```



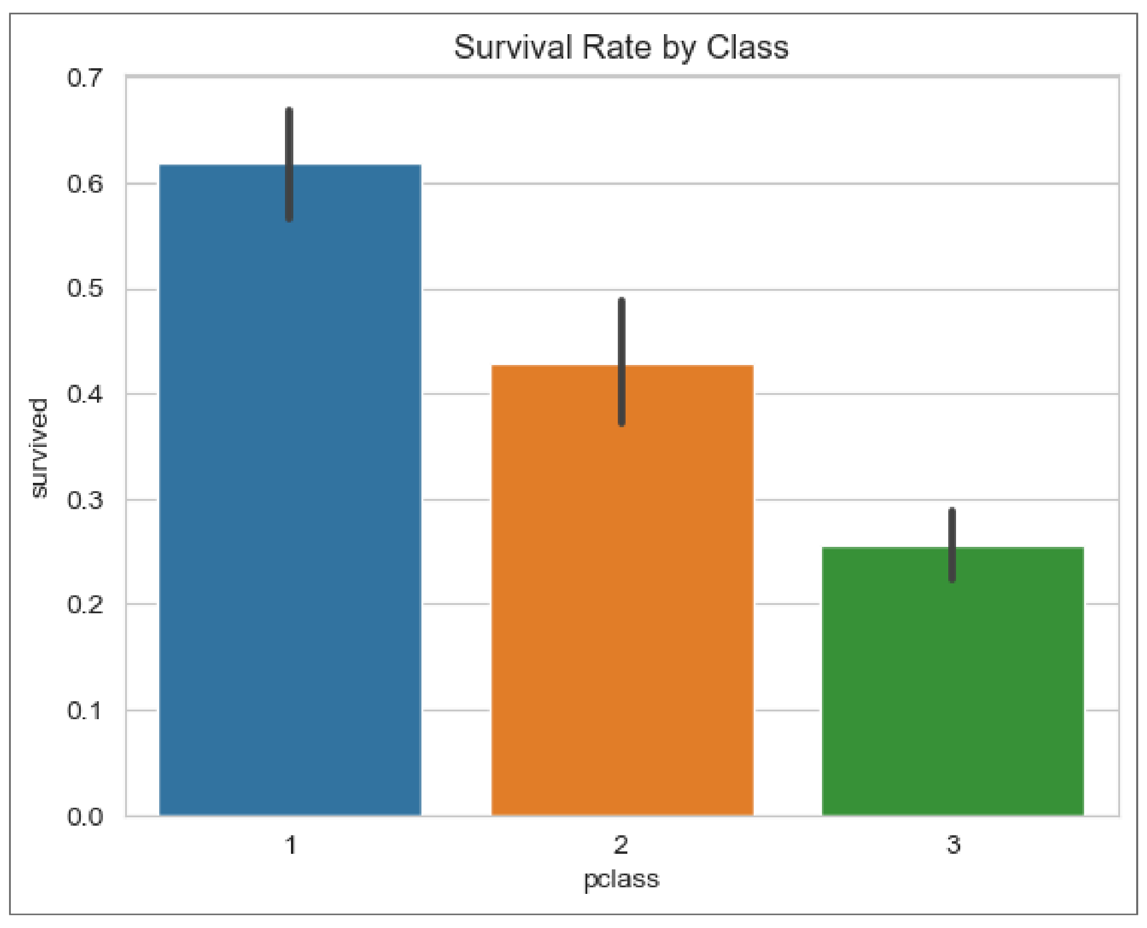
```
In [14]: # Distribution of passengers by gender
         sns.countplot(x='sex', data=df)
plt.title('Number of Passengers by Gender')
plt.show()
```



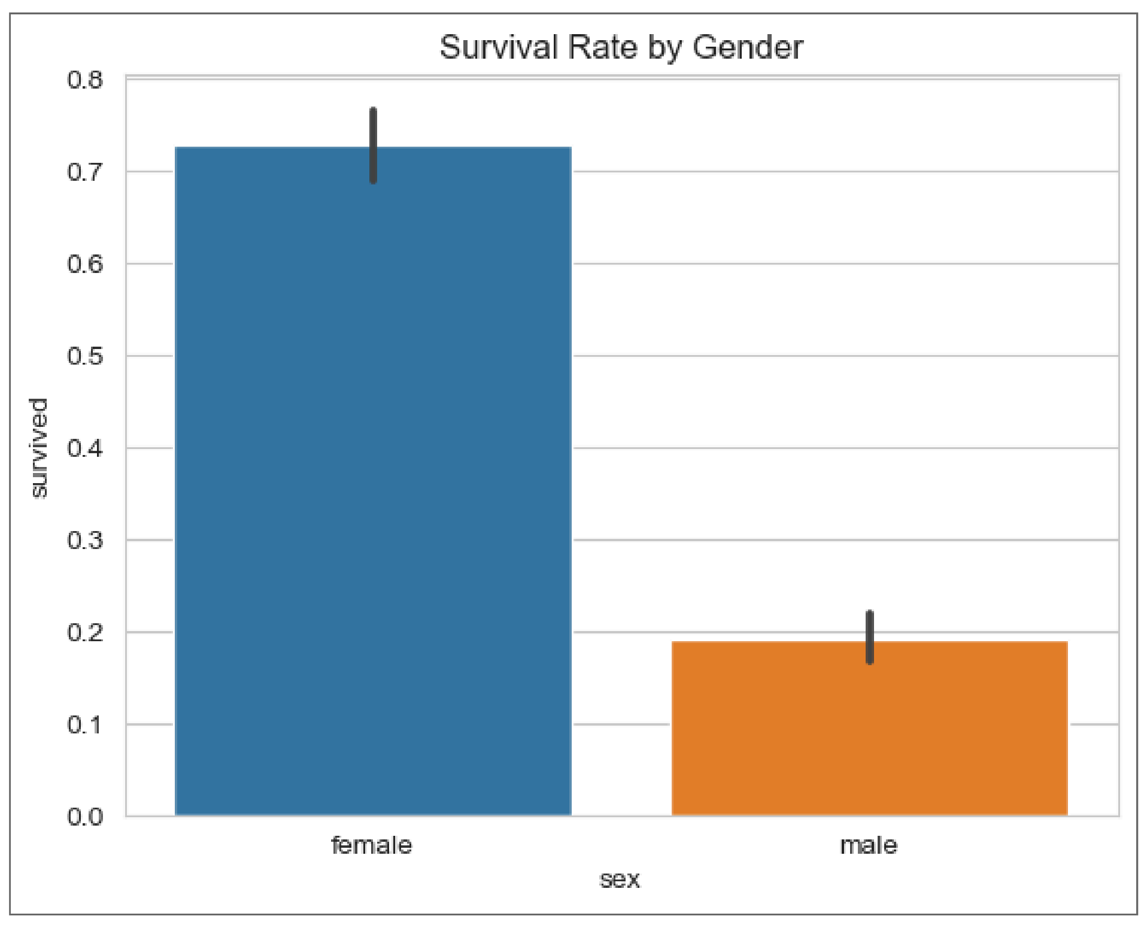
```
In [15]: # Distribution of passengers by age
         sns.histplot(df['age'].dropna(), kde=False, bins=30)
plt.title('Distribution of Passengers by Age')
plt.show()
```



```
In [16]: # Survival rate by class
         sns.barplot(x='pclass', y='survived', data=df)
plt.title('Survival Rate by Class')
plt.show()
```



```
In [17]: # Survival rate by gender
         sns.barplot(x='sex', y='survived', data=df)
plt.title('Survival Rate by Gender')
plt.show()
```



```
In [18]: from sklearn.impute import SimpleImputer

# Impute missing values in 'Age' using the mean
imputer = SimpleImputer(strategy='mean')
df['age'] = imputer.fit_transform(df[['age']])
```

```
In [19]: # Drop rows where 'Embarked' is missing
         df = df.dropna(subset=['embarked'])
```

```
In [20]: # Fill missing values in 'Cabin' with 'Unknown' using .loc
         df.loc[df['cabin'].isnull(), 'cabin'] = 'Unknown'
```

```
In [21]: # Fill missing value in 'Fare' with the median fare
         fare_median = df['fare'].median()
df['fare'].fillna(fare_median, inplace=True)
```

```
C:\Users\vallu\AppData\Local\Temp\ipykernel_16440\3906223535.py:3: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
In [22]: # Fill missing values in 'Boat' with 'Unknown'
         df.loc[df['boat'].isnull(), 'boat'] = 'Unknown'

# Fill missing values in 'Body' with 'Not Recovered'
df.loc[df['body'].isnull(), 'body'] = 'Not Recovered'

# Fill missing values in 'Home.dest' with 'Unknown'
df.loc[df['home.dest'].isnull(), 'home.dest'] = 'Unknown'
```

```
In [29]: df['body'] = df['body'].astype(str)
```

```
C:\Users\vallu\AppData\Local\Temp\ipykernel_16440\3447641020.py:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
In [30]: # Verify that there are no more missing values
         print(df.isnull().sum())
```

```
pclass       0
survived     0
name         0
sex          0
age          0
sibsp        0
parch        0
ticket       0
fare         0
cabin        0
embarked     0
boat         0
body         0
home.dest    0
dtype: int64
```

```
In [31]: titanic_report=sv.analyze(df)
         titanic_report.show_html('titanic.html')
```

```
                                               |                                              | [  0%]   00:00 ->…
```

```
Report titanic.html was generated! NOTEBOOK/COLAB USERS: the web browser MAY not pop up, regardless, the report IS saved in your notebook/colab files.
```

```
In [ ]: d = dtale.show(df)
        d.open_browser()
```

```
In [ ]:
```