Task - 2

Problem Statement : Perform a data cleaning and exploratory data analysis(EDA) on a dataset . Explore the relationships between variables and identify patterns and trends in the data.

Dataset Used : https://www.kaggle.com/c/titanic/data?select=train.csv (https://www.kaggle.com/c/titanic/data?select=train.csv)

About Dataset : The dataset is from Kaggle , titled "Titanic: Machine Learning from Disaster." The goal of this competition is to predict whether a passenger on the Titanic survived or not based on various attributes.The dataset contains several variables describing different aspects of passengers on the Titanic

In [45]:
```
#importing the dataset
import pandas as pd
import matplotlib.pyplot as plt
datas=pd.read_csv("D:\MSc Data Science\Semester 3\Extra Works\Prodigy InfoTech
```

In [46]:
```
print(datas.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None
```

In [47]:
```
#filter the specific columns
PassengerId = datas['PassengerId']
Survived=datas['Survived']
Name=datas['Name']
Sex=datas['Sex']
Age=datas['Age']
```

In [48]:
```
datas.head()
```

Out[48]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N

Data Cleaning

In [49]:
```
#removing the unwanted columns
columns_remove = ['SibSp','Parch','Ticket','Fare','Cabin','Embarked']
datas = datas.drop(columns=columns_remove, errors='ignore')
```

In [50]:
```
print(datas)
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age
0                             Braund, Mr. Owen Harris    male  22.0
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
2                              Heikkinen, Miss. Laina  female  26.0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
4                            Allen, Mr. William Henry    male  35.0
..                                                ...     ...   ...
886                            Montvila, Rev. Juozas    male  27.0
887                     Graham, Miss. Margaret Edith  female  19.0
888         Johnston, Miss. Catherine Helen "Carrie"  female   NaN
889                            Behr, Mr. Karl Howell    male  26.0
890                              Dooley, Mr. Patrick    male  32.0

[891 rows x 6 columns]
```

In [51]:
```
datas.head()
```

Out[51]:

	PassengerId	Survived	Pclass	Name	Sex	Age
0	1	0	3	Braund, Mr. Owen Harris	male	22.0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0
4	5	0	3	Allen, Mr. William Henry	male	35.0

In [52]:
```
#checking missing values
print('Missing values before cleaning:')
print(datas.isnull().sum())
```

```
Missing values before cleaning:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
dtype: int64
```

In [53]:
```
#dropping rows with missing value
datas.dropna(subset=['Age'],inplace=True)
```

In [55]:
```
#check missing values after cleaning
print('\nMissing values after cleaning:')
print(datas.isnull().sum())
print("\nCleaned dataframe")
print(datas)
```

```
Missing values after cleaning:
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
dtype: int64

Cleaned dataframe
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
885          886         0       3
886          887         0       2
887          888         1       1
889          890         1       1
890          891         0       3

                                                  Name     Sex   Age
0                              Braund, Mr. Owen Harris    male  22.0
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
2                               Heikkinen, Miss. Laina  female  26.0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
4                             Allen, Mr. William Henry    male  35.0
..                                                 ...     ...   ...
885               Rice, Mrs. William (Margaret Norton)  female  39.0
886                              Montvila, Rev. Juozas    male  27.0
887                       Graham, Miss. Margaret Edith  female  19.0
889                              Behr, Mr. Karl Howell    male  26.0
890                                Dooley, Mr. Patrick    male  32.0

[714 rows x 6 columns]
```

In [56]:
```
datas.head()
```

Out[56]:

	PassengerId	Survived	Pclass	Name	Sex	Age
0	1	0	3	Braund, Mr. Owen Harris	male	22.0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0
4	5	0	3	Allen, Mr. William Henry	male	35.0

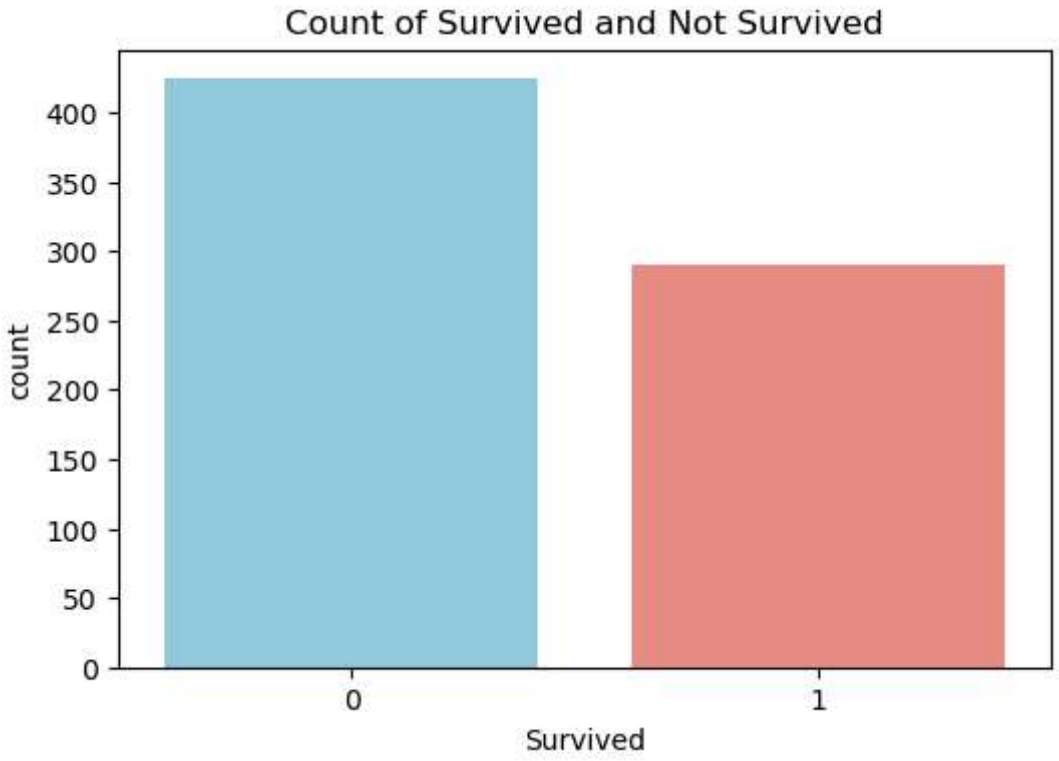
EDA

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

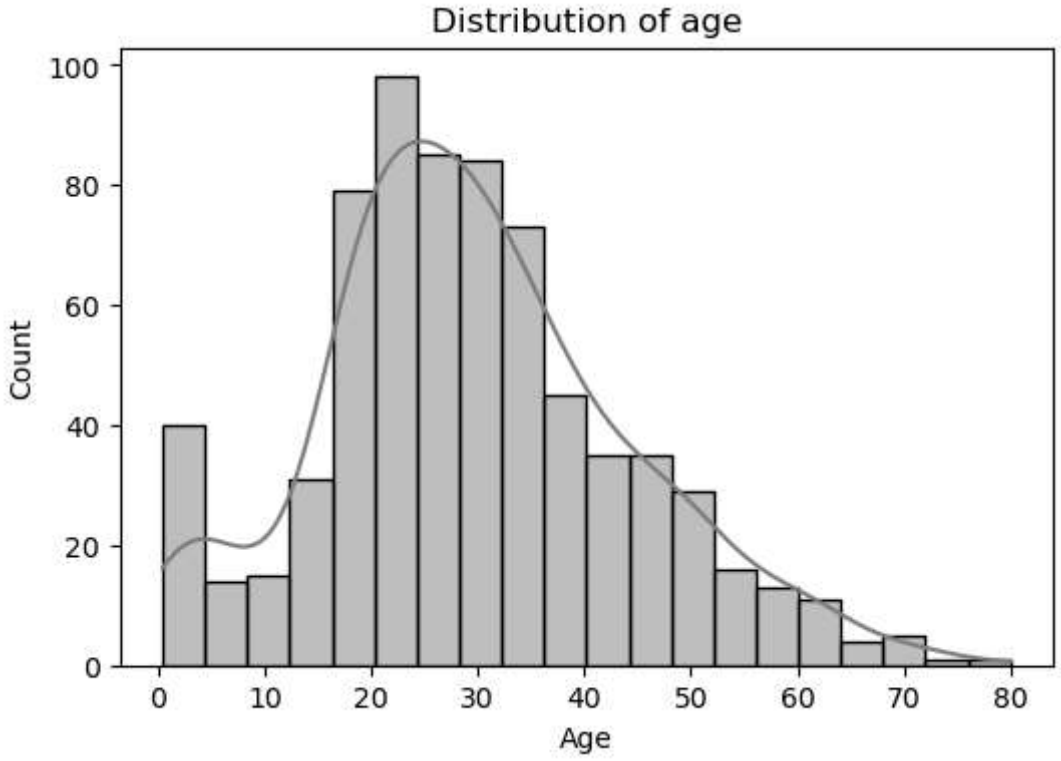
In [57]:
```
#descriptive statistics
print(datas.describe())
```

```
       PassengerId    Survived      Pclass         Age
count   714.000000  714.000000  714.000000  714.000000
mean    448.582633    0.406162    2.236695   29.699118
std     259.119524    0.491460    0.838250   14.526497
min       1.000000    0.000000    1.000000    0.420000
25%     222.250000    0.000000    1.000000   20.125000
50%     445.000000    0.000000    2.000000   28.000000
75%     677.750000    1.000000    3.000000   38.000000
max     891.000000    1.000000    3.000000   80.000000
```

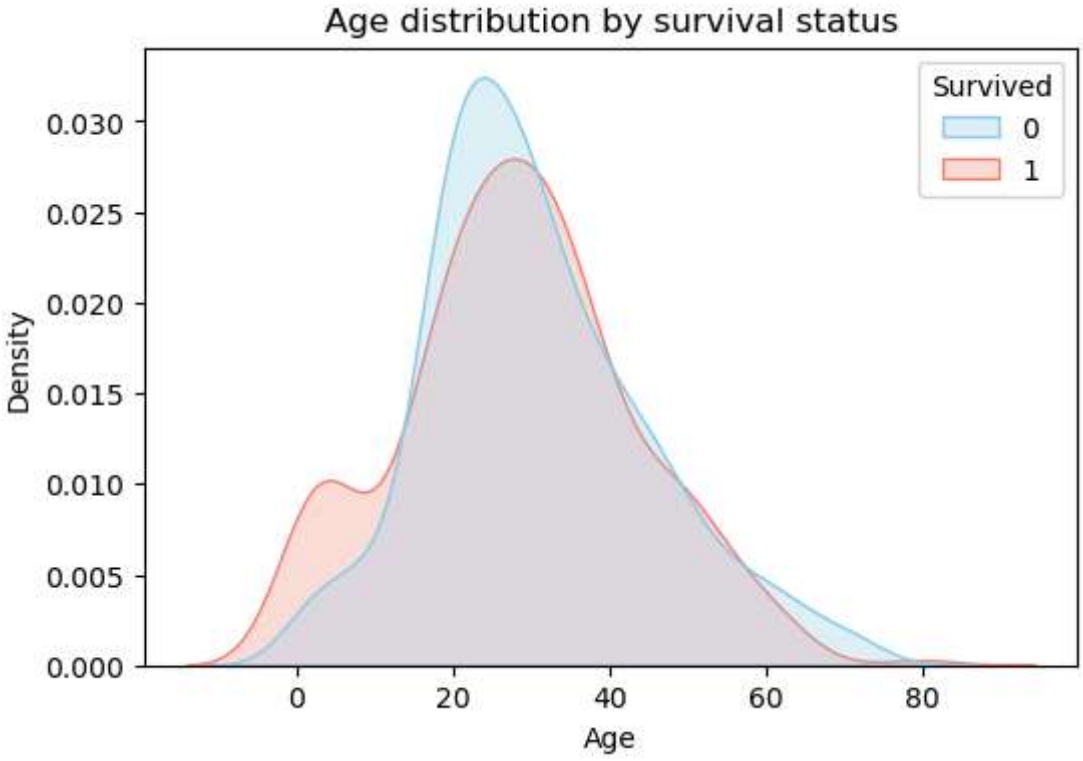
In [67]:
```
#countplot of survivors
plt.figure(figsize=(6,4))
colors = ["skyblue", "salmon"]
sns.countplot(x='Survived', data=datas , palette=colors)
plt.title('Count of Survived and Not Survived')
plt.show()
```



In [72]:
```
#distribution of age
plt.figure(figsize=(6,4))
sns.histplot(datas['Age'],bins=20, kde=True,color='grey')
plt.title('Distribution of age')
plt.show()
```



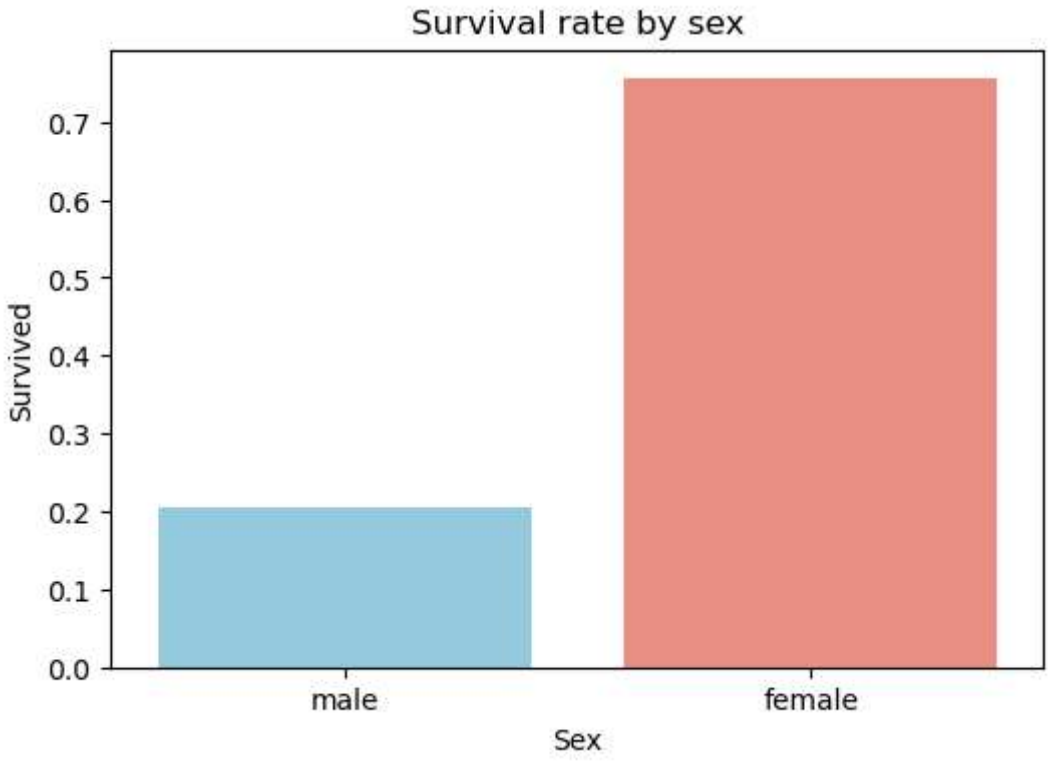
In [64]:
```
plt.figure(figsize=(6,4))
colors = ["skyblue", "salmon"]
sns.kdeplot(x='Age' ,hue='Survived', data=datas ,fill=True, common_norm=False,
plt.title('Age distribution by survival status')
plt.show()
```



Interpretation : Interpreting the density of age distribution among survivors as being high around younger individuals suggests that there may have been a prioritization of younger people during the evacuation process of the Titanic. It's possible that during the evacuation process, crew members or fellow passengers prioritized the safety of children, adolescents, and young adults, recognizing their vulnerability and potential to contribute to future generations.

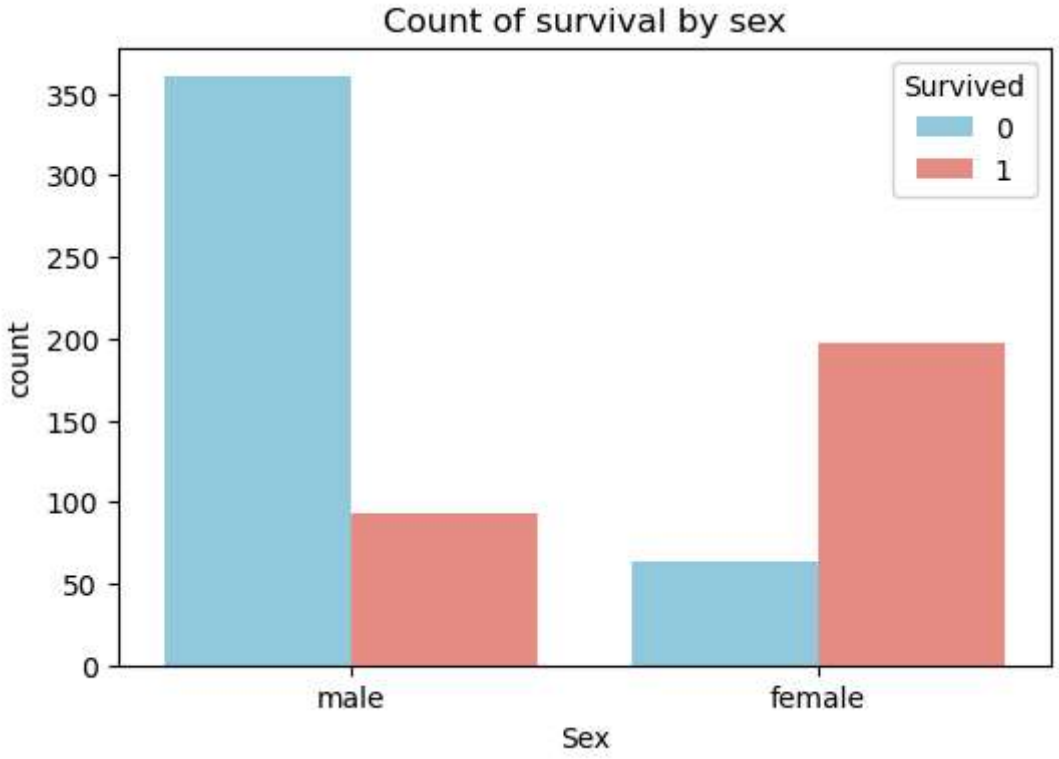
In [68]:
```
#survival rate by sex
plt.figure(figsize=(6,4))
colors = ["skyblue", "salmon"]
sns.barplot(x='Sex' ,y='Survived', data=datas ,ci=None, palette=colors)
plt.title('Survival rate by sex')
plt.show()
```

```
C:\Users\adith\AppData\Local\Temp\ipykernel_6960\4013713040.py:4: FutureWarni
ng:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.

  sns.barplot(x='Sex' ,y='Survived', data=datas ,ci=None, palette=colors)
```



Interpretation : If there is a significant difference in survival rates between males and females, with a higher survival rate among females, it suggests that priority may have been given to women during the evacuation of the Titanic.

In [69]:
```
#countplot of survival by sex
plt.figure(figsize=(6,4))
colors = ["skyblue", "salmon"]
sns.countplot(x='Sex' ,hue='Survived', data=datas,palette=colors)
plt.title('Count of survival by sex')
plt.show()
```



Interpretation : Gender could be a strong predictor of survival on the Titanic, with females having a higher likelihood of survival compared to males.It's possible that women, on average, exhibited different behavioral responses during the emergency compared to men.