Project Goal

This is a comprehensive Exploratory Data Analysis of the Titanic dataset along with visualization.

My aim is to analyse the factors that contribute to the survival of the passengers of RMS Titanic.

I will use visualizations and statistical techniques to unravel the insights within the data set.

It will have Observational summary, statistical summary, missing values and it's potential impact on the data.

The aim is to answer questions such as "Did women and children have higher survival rates?" and "How did the passenger class affect the survival chances?"

A particular focus is given to the titles of passengers and their corresponding survival rates which tells about any potential social distinctions that played a crutial role in survival outcomes.

By the end, we will have a detailed visual and quantitative understanding of the Titanic's passenger data.

In [38]:

```
# Imports

# pandas
import pandas as pd
from pandas import Series,DataFrame

# numpy, matplotlib, seaborn
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('whitegrid')
%matplotlib inline
```

In [15]:

```
# Loading the file

titanic_df = pd.read_csv(r"C:\Users\Lenovo\Downloads\train.csv")
```

In [16]:

```
# Displaying the data

titanic_df.head()
```

Out[16]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [17]:

```
# Ensuring, random data is flawless

titanic_df.sample(5)
```

Out[17]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
286	287	1	3	de Mulder, Mr. Theodore	male	30.0	0	0	345774	9.5000	NaN	S
881	882	0	3	Markun, Mr. Johann	male	33.0	0	0	349257	7.8958	NaN	S
396	397	0	3	Olsson, Miss. Elina	female	31.0	0	0	350407	7.8542	NaN	S
676	677	0	3	Sawyer, Mr. Frederick Charles	male	24.5	0	0	342826	8.0500	NaN	S
444	445	1	3	Johannesen-Bratthammer, Mr. Bernt	male	NaN	0	0	65306	8.1125	NaN	S

In [18]:

```
titanic_df.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
```

Observation Summary

The titanic_df dataset contains a total of 12 columns, of which there are 7 numerical columns (PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare) and 5 categorical columns (Name, Sex, Ticket, Cabin, Embarked).

The shape of the titanic_df DataFrame is 891 rows and 12 columns, indicating that there are 891 entries, each with 12 attributes.

In [19]:

```
titanic_df.describe()
```

Out[19]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Statistical Summary

Survived is an indicator where 1 represents survival and 0 represents non-survival.

The mean survival rate is 0.383838 (approximately 38.38%), suggesting that less than half of the passengers survived.

Pclass represents the class of travel with a lower number indicating a higher class.

The passengers are spread across three classes, with a mean Pclass of 2.308642, implying that most passengers are in the second and third classes.

The Age of passengers has a mean of 29.699118 years, with the youngest being 0.42 years old (likely a few months old) and the oldest at 80 years.

Looking for null values in the dataset.

In [22]:

```
titanic_df.isnull()
```

Out[22]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

891 rows × 12 columns

In [23]:

```
titanic_df.isnull().sum()
```

Out[23]:

```
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 [24]:

```
# I want to know the total number of null values in my dataset.

titanic_df.isnull().sum().sum()
```

Out[24]:

```
866
```

Observation

In [ ]:

```
In the titanic_df dataset of 891 entries, several columns have missing values:

Age has 177 missing entries.

Cabin has 687 missing entries.

Embarked has 2 missing entries.
```

Numerical Analysis

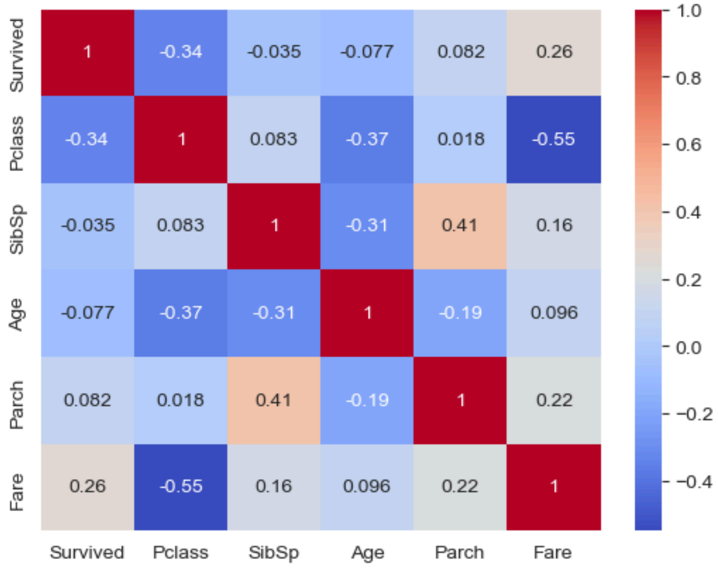
In [30]:

```
titanic_df.columns
```

Out[30]:

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [44]:
```
heatmap = sns.heatmap(titanic_df[['Survived', 'Pclass','SibSp','Age','Parch', 'Fare']].corr(), annot = True, cmap = '
```



Survived column has a little relation with Fare column, the relation is positive (0.26) meaning, more the fare, higher is the chances of survival.

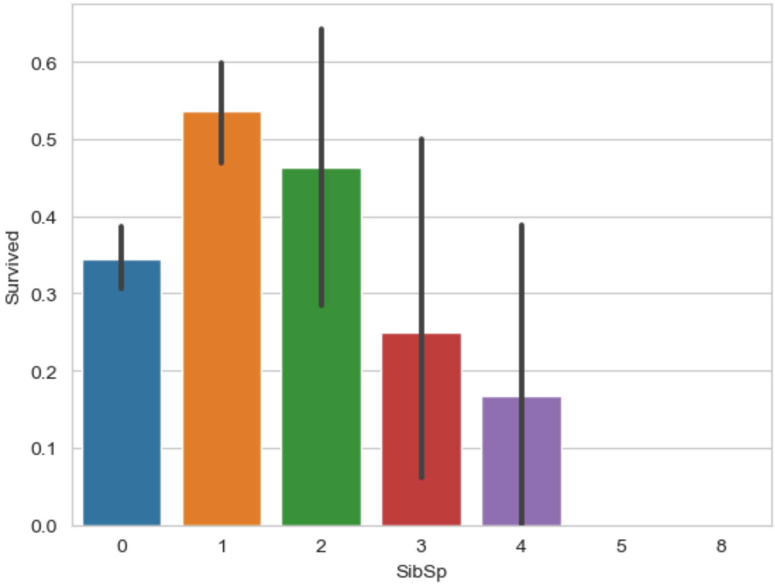
This does not mean that other variables are not worthy. Need to look at the other variables too for insights.

In [34]:
```
titanic_df['SibSp'].unique()
```

Out[34]:
```
array([1, 0, 3, 4, 2, 5, 8], dtype=int64)
```

This means that a passenger has either of these number of siblings, minimum being none to maximum being 8

In [41]:
```
bargraphsibsp = sns.barplot(data = titanic_df, x = 'SibSp', y = 'Survived')

plt.show()
```

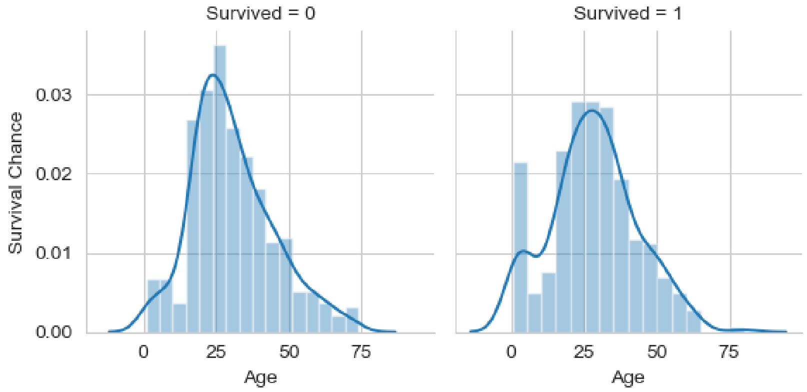


Passengers having more number of siblings are less likely to survive.

Passengers who are single or have 1 or 2 siblings are more likely to survive.

In [66]:

```
ageplot = sns.FacetGrid(titanic_df, col = 'Survived')
plt.figure(figsize = (10,8))
ageplot = ageplot.map(sns.distplot,'Age')
ageplot = ageplot.set_ylabels('Survival Chance')
```



```
<Figure size 1000x800 with 0 Axes>
```

We can see that there is a peak to the corresponding young passengers that have survived.

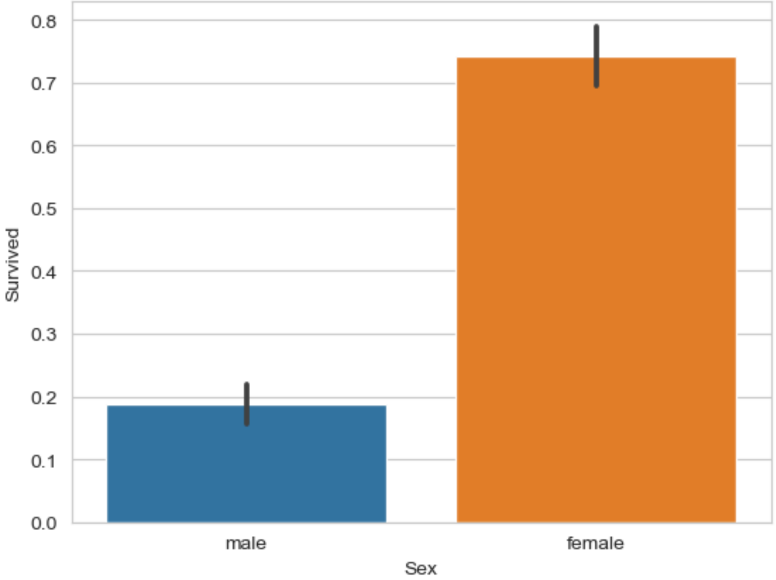
It can be seen that if age is 60 or 80, chances of survival decreases.

Even though age column is not related to survival column, it is seen that age categories of passengers show survival less/more.

Young passengers are more probable to survive aged between 25-40, may be because of their physical agilty.

In [69]:

```
sexplot = sns.barplot(x = 'Sex',y = 'Survived',data = titanic_df)
plt.figure(figsize = (10,2))
plt.show()
```



```
<Figure size 1000x200 with 0 Axes>
```

In [73]:

```
titanic_df[["Sex", "Survived"]].groupby("Sex").mean()
```

Out[73]:

	Survived
Sex	
female	0.742038
male	0.188908

It is clearly showing that females have more chances of survival than males as female were rescued first in the lifeboats.

So Sex played an important role during the evacuation process and hence women had more chances of survival.

In [74]:

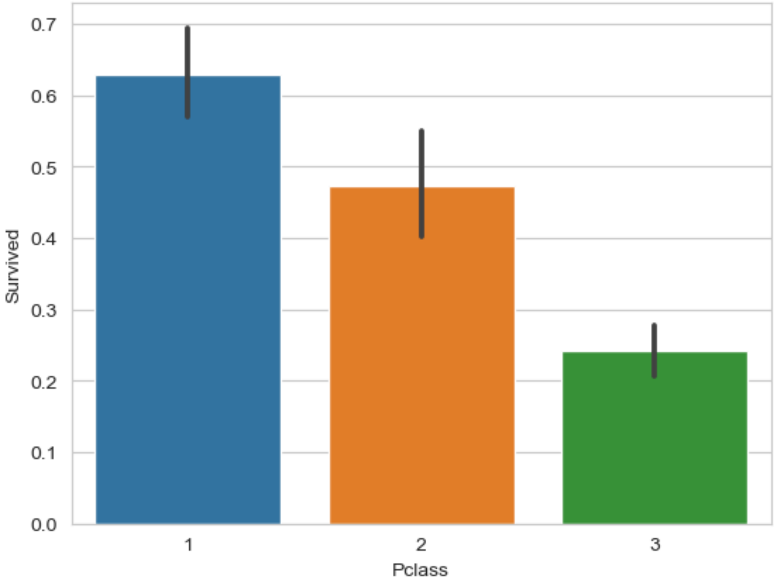
```
titanic_df.columns
```

Out[74]:

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [75]:

```
pclassplot = sns.barplot(x = "Pclass",y = "Survived",data = titanic_df)
plt.show()
```

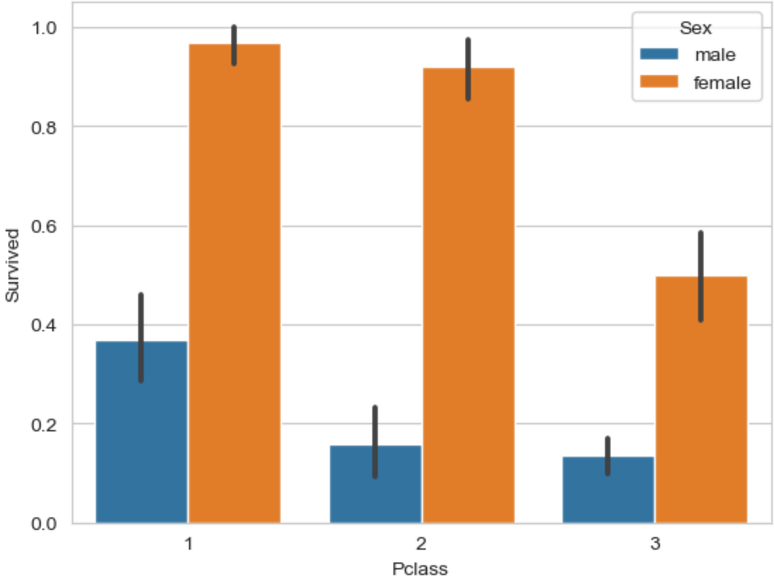


It is evident that as the Passenger class increases, the chances of survival increases.

Meaning passengers with higher class has higher chances of survival compared to passengers of lower class.

In [77]:

```
x = sns.barplot(x = "Pclass",y = "Survived",data = titanic_df, hue = "Sex")
plt.show()
```



From this visualization it is clear that, overall female has the higher survival chances compared to males.

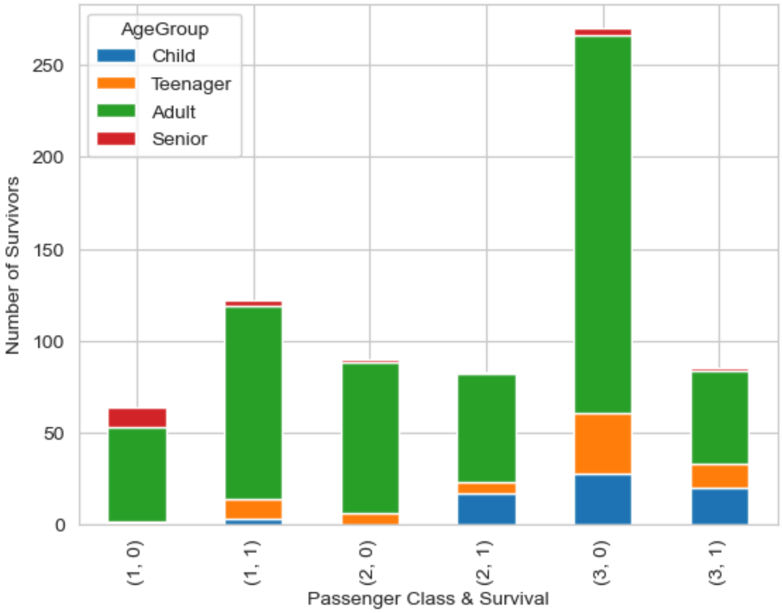
Drilling down, with passenger class taken into consideration, female in higher class has the most chances of survival and male of the lower class has the least chance of survival.

In [90]:

```
bins = [0, 12, 18, 60, np.inf]

labels = ['Child', 'Teenager', 'Adult', 'Senior']

titanic_df['AgeGroup'] = pd.cut(titanic_df['Age'], bins = bins, labels=labels)

pd.crosstab([titanic_df['Pclass'], titanic_df['Survived']], titanic_df['AgeGroup']).plot(kind='bar', stacked=True)

plt.xlabel("Passenger Class & Survival")
plt.ylabel("Number of Survivors")
```

Out[90]:

```
Text(0, 0.5, 'Number of Survivors')
```



This visualizaion shows that adults in 3rd class has the highest number of non-survivals.

However it must be noted that, out of total passengers maximum may be of 3rd class and seat distribustion between the classes are never same.

The survival chances are great if the passenger belongs to 1st class and that too an adult.

We can see clear discrimination based on class factor.

First Class Survival: Higher survival rates for all age groups, particularly adults.

Teenagers: Similar survival and fatality rates in the second class.

Third Class Outcomes: Significantly higher fatalities across all age groups, with children being notably affected.

Adults: Majority of the fatalities, especially in third class.

Seniors: Lowest survival rates across all classes.

Conclusion

The findings suggest that survival on the RMS Titanic was not random, but rather significantly influenced by socio-economic factors such as passenger class, in addition to demographic factors like age, sex, and family relationships.

Key points includes:

1. Higher survival rates for women and children.
2. There is a negative correlation between passenger class and survival, with first-class passengers more likely to survive.

Henceforth, the analysis highlights the impact of social status and class and demographics on survival chances during maritime disasters.