In [30]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from matplotlib import style
style.use('ggplot')

import warnings
warnings.filterwarnings('ignore')
```

Extraction of Data

In [32]:
```
titanic = pd.read_csv('train.csv')
titanic.head(10)
```

Out[32]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0

Data Cleaning

In [7]:
```
titanic.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
```

Checking for unique values

In [9]:
```
print('Survived   :',titanic.Survived.unique())
print('Pclass   :',titanic.Pclass.unique())
print('Gender   :',titanic.Sex.unique())
print('Embarked   :',titanic.Embarked.dropna().unique())
```

```
Survived   : [0 1]
Pclass   : [3 1 2]
Gender   : ['male' 'female']
Embarked   : ['S' 'C' 'Q']
```

In [10]:
```
titanic.duplicated().sum()                    # no duplicate values
```

Out[10]:
```
0
```

In [11]:
```
titanic.isnull().sum()
```

Out[11]:
```
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
```

1. Numerical column

In [13]:
```
mean_age = titanic.Age.mean()
titanic.Age.replace(np.nan,mean_age,inplace = True)
```

In [14]:
```
titanic.Age.isnull().sum()
```

Out[14]: 0

2. Categorical column

In [16]:
```
mode_emb = titanic.Embarked.dropna().mode()[0]
```

In [17]:
```
titanic.Embarked.replace(np.nan,mode_emb,inplace = True)
```

In [18]:
```
titanic.Embarked.isnull().sum()
```

Out[18]: 0

3. If number of missing values is large wrt to the number of rows then we can drop the column

In [20]:
```
titanic.drop('Cabin', axis = 1, inplace = True)
```

In [21]:
```
titanic.head()
```

Out[21]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0

4. Combining two columns

In [112…
```
titanic['family_members'] = titanic.SibSp + titanic.Parch
```

In [114…
```
titanic.head()
```

Out[114…

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	I
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0

5. Dropping Extra unwanted columns

In [116…
```
titanic.drop(['SibSp','Parch'],axis = 1,inplace = True)
```

In [118…
```
titanic.head()
```

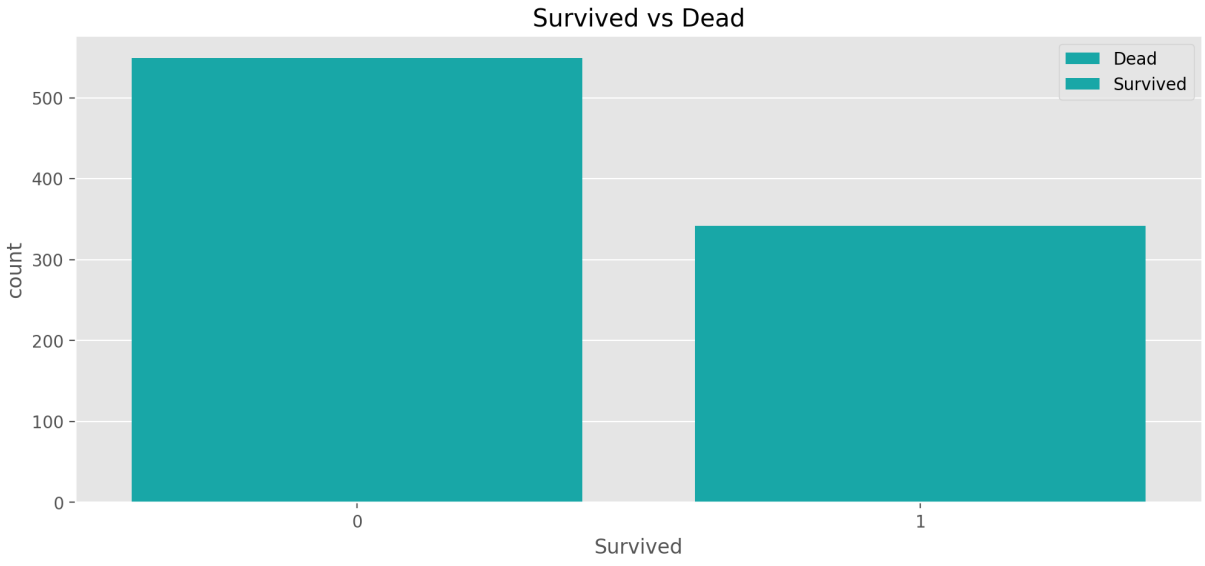
Out[118...

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

Data Analysis

In [44]:
```
plt.figure(figsize = (12,5),dpi = 200)
sns.countplot(x = 'Survived',label = ['Dead','Survived'],data = titanic,colo
plt.title('Survived vs Dead')
```

Out[44]:
```
Text(0.5, 1.0, 'Survived vs Dead')
```



In [46]:
```
titanic.Survived.value_counts()
```

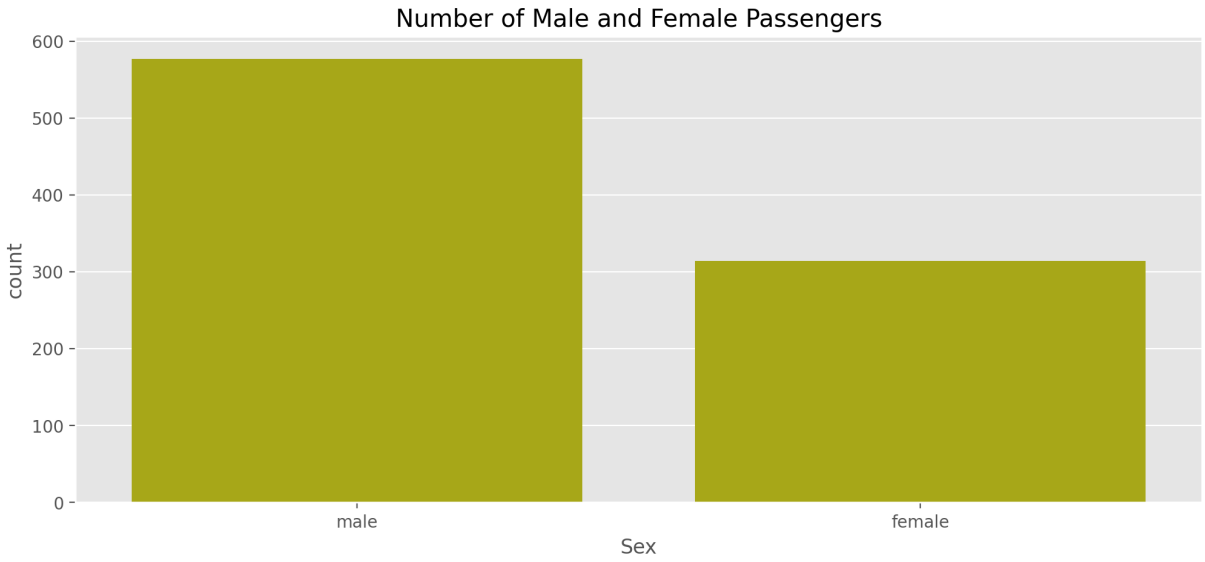
Out[46]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

In [48]:
```
Survival = titanic.Survived.value_counts(normalize = True)
print('Dead       :',round(Survival[0]*100),'%')
print('Survived   :',round(Survival[1]*100),'%')
```

```
Dead       : 62 %
Survived   : 38 %
```

In [56]:
```
plt.figure(figsize = (12,5),dpi = 200)
sns.countplot(x = 'Sex',data = titanic, color = 'y')
plt.title('Number of Male and Female Passengers')
```

Out[56]:
```
Text(0.5, 1.0, 'Number of Male and Female Passengers')
```

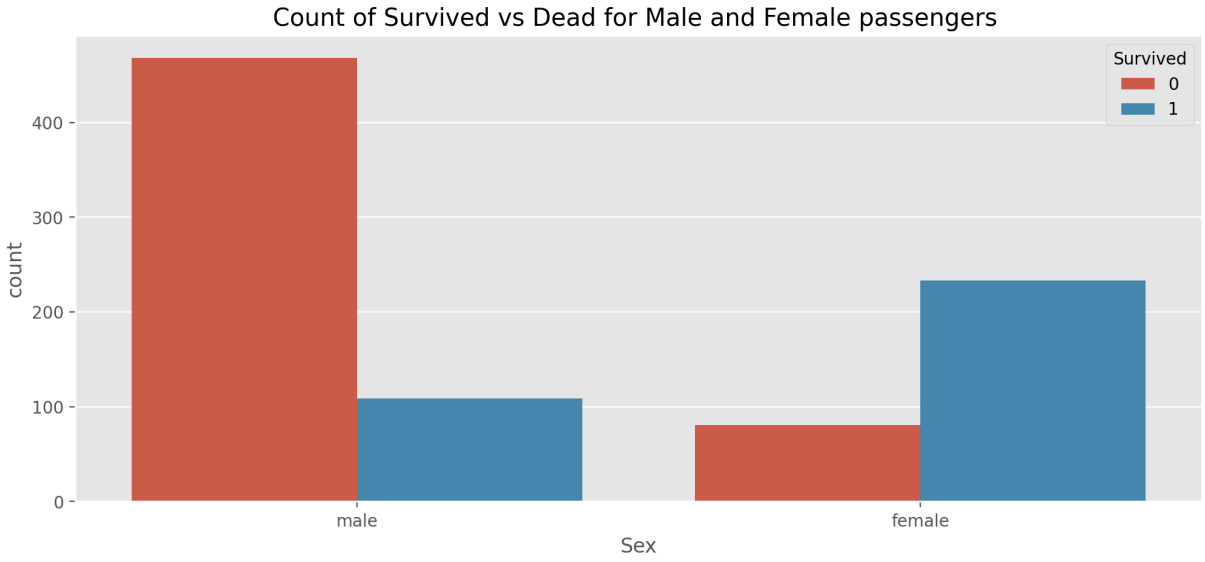


In [58]:
```
gender = titanic.Sex.value_counts()
gender
```

Out[58]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [68]:
```
plt.figure(figsize = (12,5),dpi = 200)
sns.countplot(x = 'Sex',hue = 'Survived',data = titanic )
plt.title('Count of Survived vs Dead for Male and Female passengers ')
```

Out[68]:
```
Text(0.5, 1.0, 'Count of Survived vs Dead for Male and Female passengers ')
```



Observation: Female passengers had a higher rate of survival than Male passengers

In [73]:
```
gender_svl_rate= titanic.groupby('Sex').Survived.value_counts(normalize = Tr
gender_svl_rate * 100
```

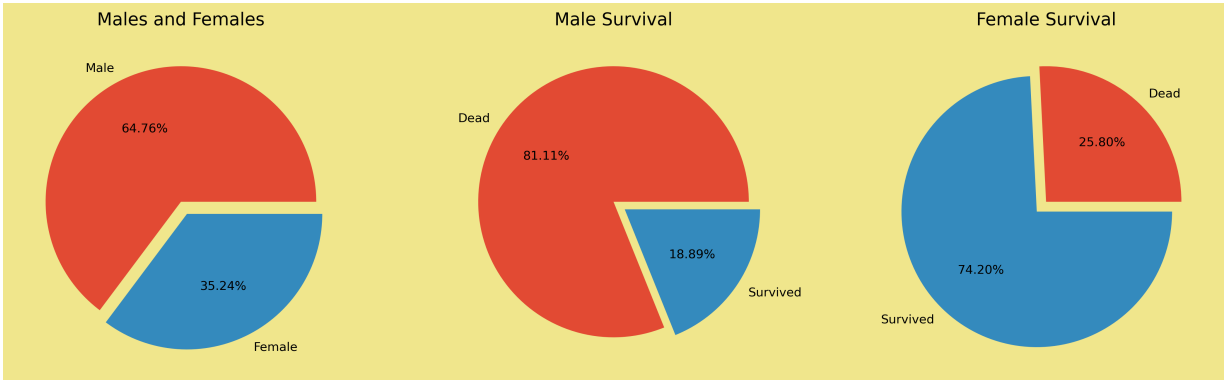
Out[73]:
```
Sex     Survived
female  0           25.796178
        1           74.203822
male    0           81.109185
        1           18.890815
Name: proportion, dtype: float64
```

In [77]:
```
print('Female Survivors :',round(gender_svl_rate['female'][1]*100,2),'%')
print('Females Dead :',round(gender_svl_rate['female'][0]*100,2),'%')
print('Male Survivors :',round(gender_svl_rate['male'][1]*100,2),'%')
print('Males Dead:',round(gender_svl_rate['male'][0]*100,2),'%')
```

```
Female Survivors : 74.2 %
Females Dead : 25.8 %
Male Survivors : 18.89 %
Males Dead: 81.11 %
```

In [87]:
```
# Creating Pie charts
gender = titanic.Sex.value_counts()
gender_label = ['Male','Female']

gender_svl = titanic.groupby('Sex').Survived.value_counts().sort_index()
male_svl = titanic.groupby('Sex').Survived.value_counts().sort_index()['male
female_svl = titanic.groupby('Sex').Survived.value_counts().sort_index()['fe

svl_labels = ['Dead','Survived']

plt.figure(figsize = (18,5),dpi = 300, facecolor = 'khaki')

plt.subplot(1,3,1)
plt.pie(gender,labels = gender_label,autopct = '%1.2f%%',explode = [0,0.1])
plt.title('Males and Females')
```

```
plt.subplot(1,3,2)
plt.pie(male_svl,labels = svl_labels,autopct = '%1.2f%%',explode = [0,0.1])
plt.title('Male Survival')

plt.subplot(1,3,3)
plt.pie(female_svl,labels = svl_labels,autopct = '%1.2f%%',explode = [0,0.1]
plt.title('Female Survival')

plt.show()
```

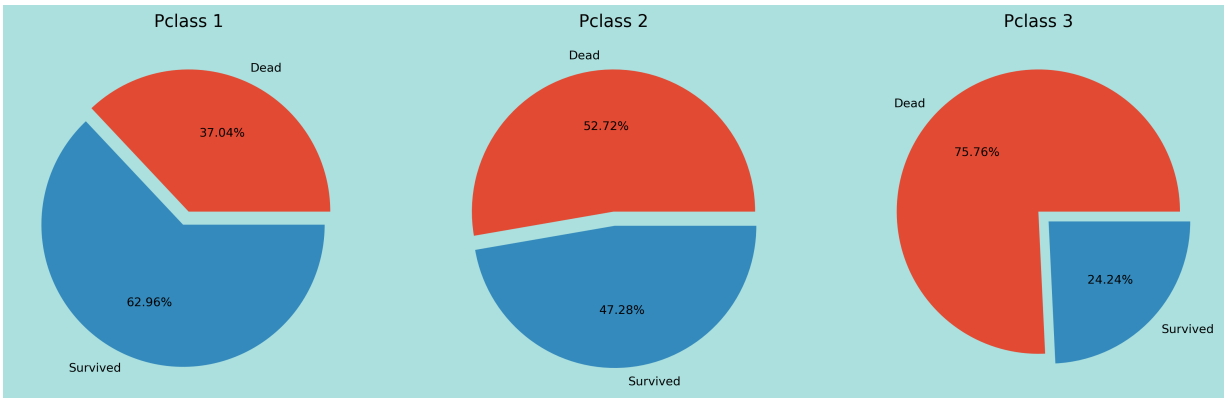


Observation: Of the total passengers 64.76% were male and 35.24% were female. However The rate of survival of female passenger was higher at 74.20% then that of male passenger at 18.89%.

In [90]:
```
pclass_svl= titanic.groupby('Pclass').Survived.value_counts().sort_index()
pclass_svl
```

Out[90]:
```
Pclass  Survived
1       0            80
        1           136
2       0            97
        1            87
3       0           372
        1           119
Name: count, dtype: int64
```

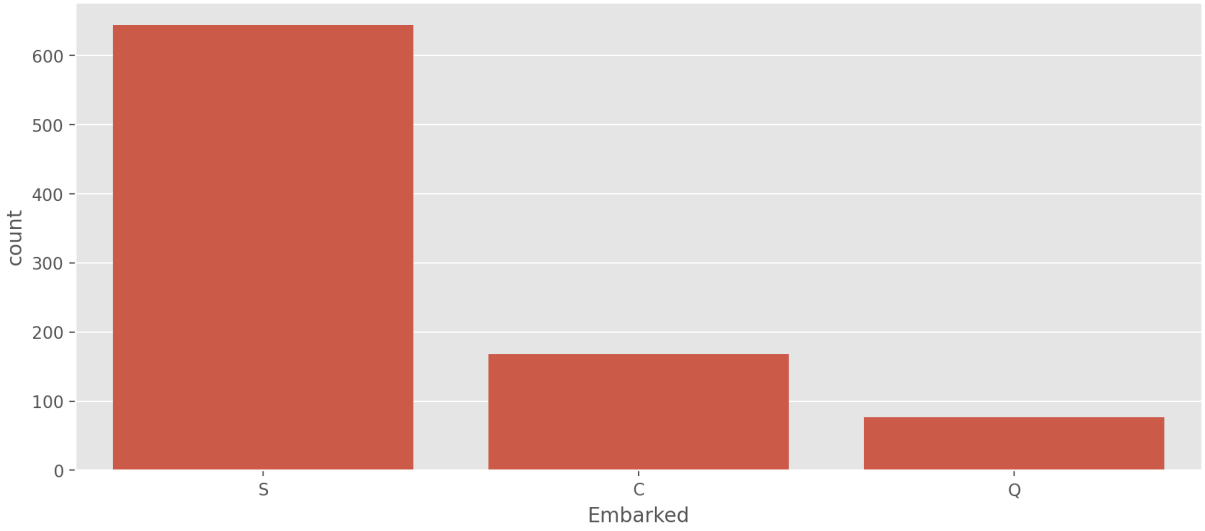
In [92]:
```
plt.figure(figsize = (18,6),dpi = 300, facecolor = '#abe0de')

plt.subplot(1,3,1)
plt.pie(pclass_svl[1],labels = ['Dead','Survived'],autopct = '%1.2f%%',explo
plt.title('Pclass 1')

plt.subplot(1,3,2)
plt.pie(pclass_svl[2],labels = ['Dead','Survived'],autopct = '%1.2f%%',explo
plt.title('Pclass 2')

plt.subplot(1,3,3)
plt.pie(pclass_svl[3],labels = ['Dead','Survived'],autopct = '%1.2f%%',explo
plt.title('Pclass 3')

plt.show()
```



The Rate of Survival of passengers from Class 1 was much higher than that of passengers from class 2 and 3.

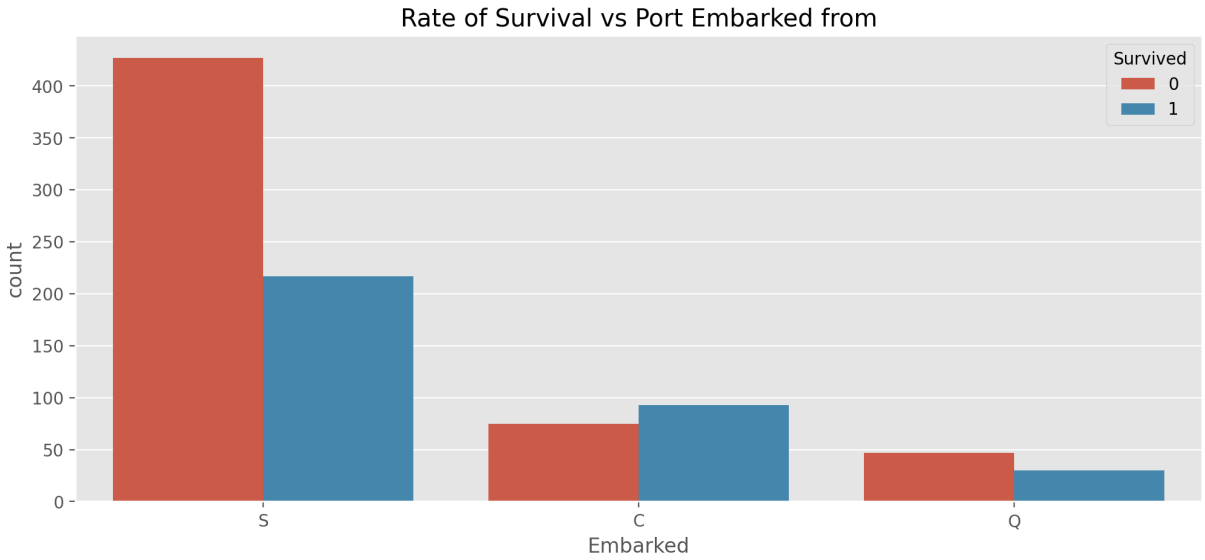
In [94]:
```
# Embarked
plt.figure(figsize = (12,5),dpi = 200)
sns.countplot(x = 'Embarked',data=titanic)
```

Out[94]:
```
<Axes: xlabel='Embarked', ylabel='count'>
```



In [98]:
```
# Embarked
plt.figure(figsize = (12,5),dpi = 200)
sns.countplot(x = 'Embarked',hue = 'Survived',data=titanic)
plt.title('Rate of Survival vs Port Embarked from')
```

Out[98]:
```
Text(0.5, 1.0, 'Rate of Survival vs Port Embarked from')
```



Observation: Passenger who boarded from port C had a much higher rate of survival than that from port S and Q

In [101…
```
titanic.Embarked.value_counts()
```

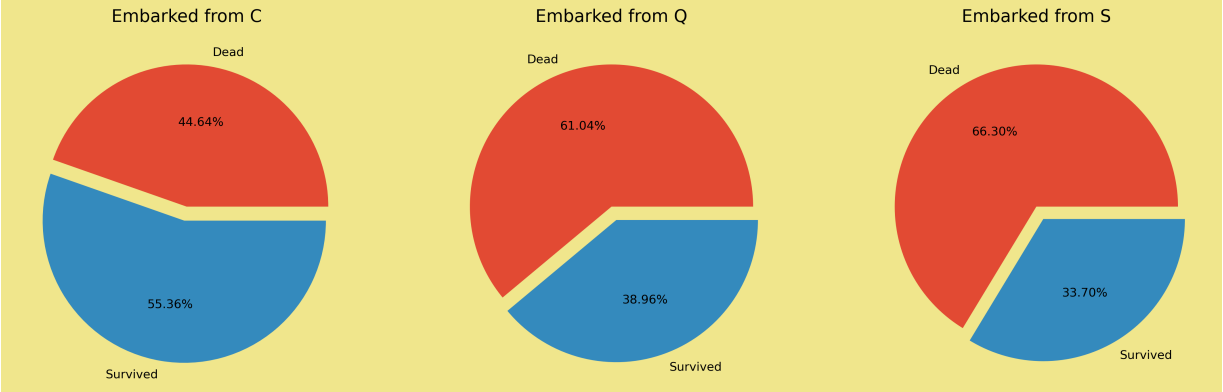
Out[101…
```
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

In [103…
```
embarked_svl = titanic.groupby('Embarked').Survived.value_counts().sort_inde
embarked_svl
```

Out[103…
```
Embarked  Survived
C         0            75
          1            93
Q         0            47
          1            30
S         0           427
          1           217
Name: count, dtype: int64
```

In [107…
```
plt.figure(figsize = (18,6),dpi = 300, facecolor = 'khaki')

plt.subplot(1,3,1)
plt.pie(embarked_svl['C'],labels = ['Dead','Survived'],autopct = '%1.2f%%',e
plt.title('Embarked from C')

plt.subplot(1,3,2)
plt.pie(embarked_svl['Q'],labels = ['Dead','Survived'],autopct = '%1.2f%%',e
plt.title('Embarked from Q')

plt.subplot(1,3,3)
plt.pie(embarked_svl['S'],labels = ['Dead','Survived'],autopct = '%1.2f%%',e
plt.title('Embarked from S')
```

```
plt.show()
```



Observation: The chances of survival of passengers travelling from Port C was 55.36% while that of port Q and C was 38.96% and 33.90% respectively.

In [120…
```
titanic.family_members.unique()
```

Out[120…
```
array([ 1,  0,  4,  2,  6,  5,  3,  7, 10], dtype=int64)
```

In [126…
```
plt.figure(figsize = (12,5),dpi = 200)
sns.countplot(x = 'family_members',data = titanic)
plt.title('Count Number of Family members Passenger travelling with')
```

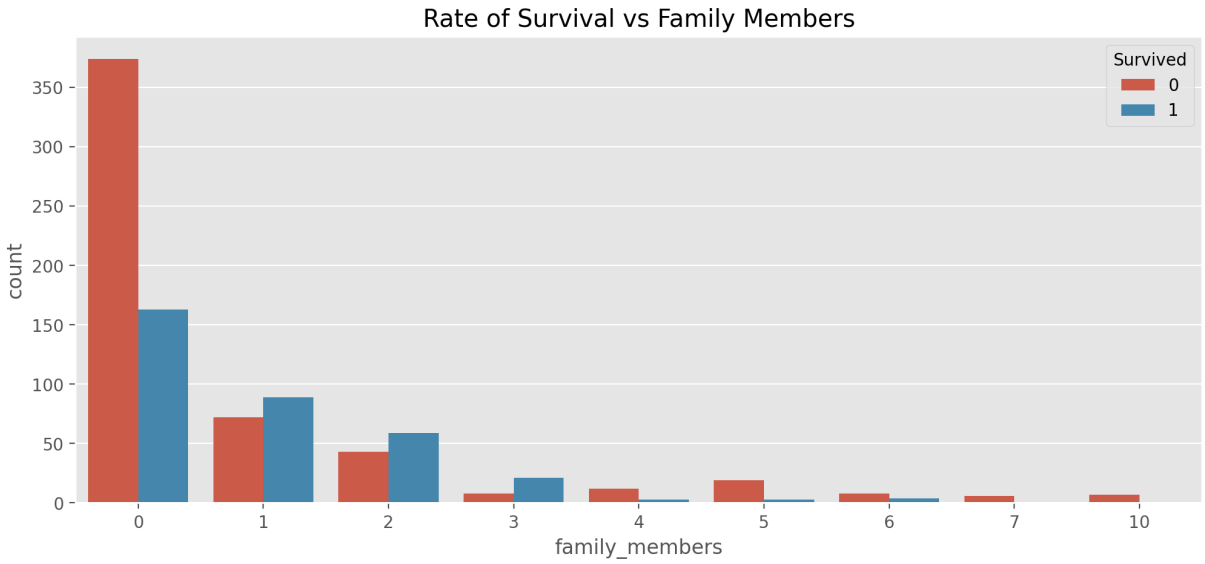
Out[126…
```
Text(0.5, 1.0, 'Count Number of Family members Passenger travelling with')
```

In [128…
```
titanic['family_members'].value_counts()
```

Out[128…
```
family_members
0      537
1      161
2      102
3       29
5       22
4       15
6       12
10       7
7        6
Name: count, dtype: int64
```

In [132…
```
plt.figure(figsize = (12,5),dpi = 200)
sns.countplot(x = 'family_members', hue = 'Survived',data = titanic)
plt.title('Rate of Survival vs Family Members')
```

Out[132…
```
Text(0.5, 1.0, 'Rate of Survival vs Family Members')
```



Observation: The Rate of Survival of passengers with 1,2 and 3 family members was much higher than that of passengers travelling with 0 and more than 4 family members.

In [135…
```
#Find the survival rate of passengers who were traveling with 0 family mambe
fm_svl_ttl = titanic.family_members.value_counts().sort_index()
print(fm_svl_ttl[0])
```

```
537
```

In [137…
```
fm_svl = titanic.groupby('family_members').Survived.value_counts().sort_inde
fm_svl
```

Out[137…

```
family_members  Survived
0               0           374
                1           163
1               0            72
                1            89
2               0            43
                1            59
3               0             8
                1            21
4               0            12
                1             3
5               0            19
                1             3
6               0             8
                1             4
7               0             6
10              0             7
Name: count, dtype: int64
```

In [139…

```
print('Survival rate of passengers travelling with 0 family members is :',ro
```

```
Survival rate of passengers travelling with 0 family members is : 30.35 %
```

In [141…

```
# Survival of passengers travelling with 1,2, 3 family members

fm_svl_123 = fm_svl[1][1] + fm_svl[2][1] + fm_svl[3][1]
fm_123 = fm_svl_ttl[1] + fm_svl_ttl[2]+fm_svl_ttl[3]
print('Survival rate of passengers travelling with 1,2,3 family members is :
```

```
Survival rate of passengers travelling with 1,2,3 family members is : 57.88 %
```

In [143…

```
# Survival of passengers travelling with 4 or more family members passengers

print('Survival rate of passengers travelling with 4 or more than 4 family m
```

```
Survival rate of passengers travelling with 4 or more than 4 family members i
s : 16.13 %
```

In [145…

```
titanic[(titanic.Fare>=500)]
```

Out[145…

	PassengerId	Survived	Pclass	Name	Sex	Age	Ticket	Fare	Cabin	Em
258	259	1	1	Ward, Miss. Anna	female	35.0	PC 17755	512.3292	NaN	
679	680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	PC 17755	512.3292	B51 B53 B55	
737	738	1	1	Lesurer, Mr. Gustave J	male	35.0	PC 17755	512.3292	B101	

In [147…
```
per_fare = len(titanic[(titanic.Fare>=200) & (titanic.Fare < 300)])
per_fare
```

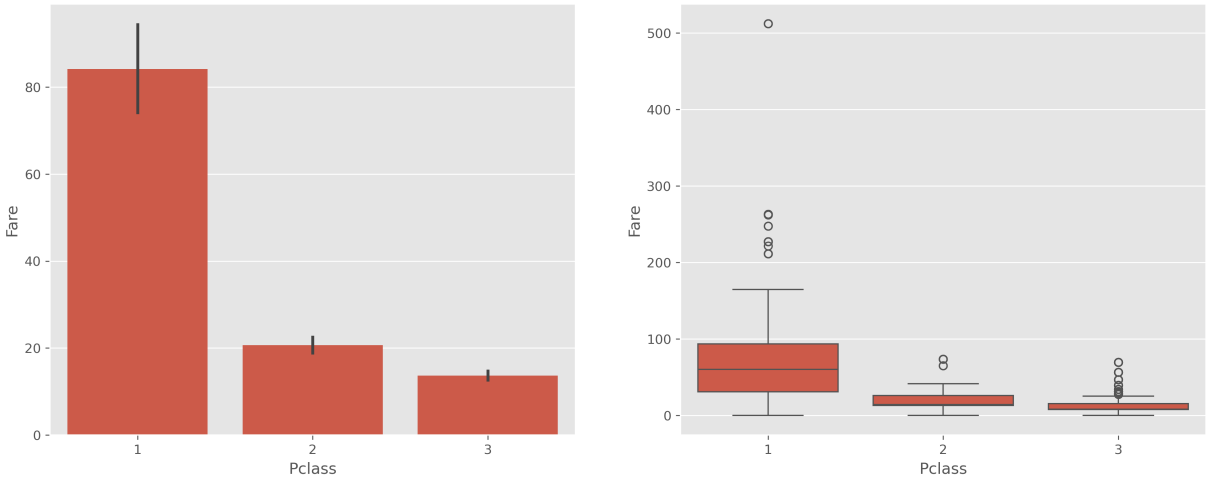
Out[147… 17

In [149…
```
total_fare = per_fare/len(titanic) * 100
total_fare
```

Out[149… 1.9079685746352413

In [151…
```
# Fare and Passnger Class
plt.figure(figsize = (16,6),dpi = 300)

plt.subplot(1,2,1)
sns.barplot(x = 'Pclass',y = 'Fare',data = titanic)

plt.subplot(1,2,2)
sns.boxplot(x = 'Pclass',y = 'Fare',data = titanic)
```

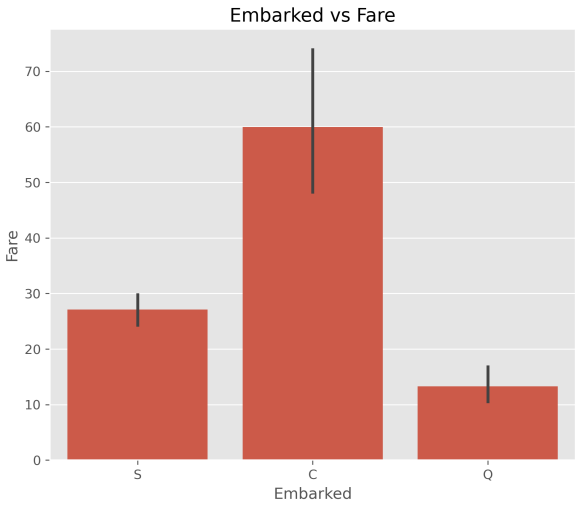
Out[151… <Axes: xlabel='Pclass', ylabel='Fare'>

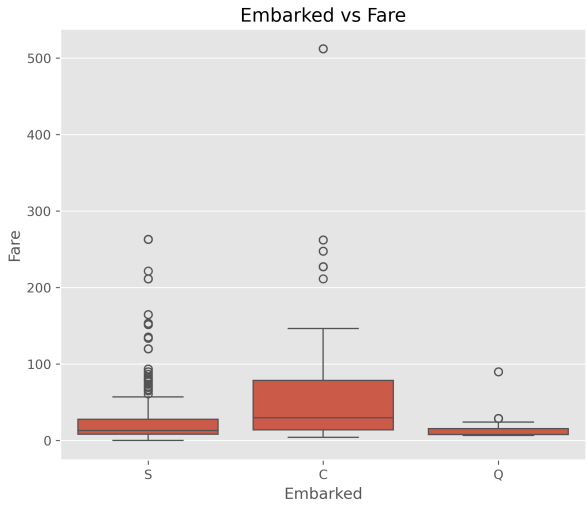


Observation: The fare for Class 1 is hgher than that of class 2 and 3

In [158…
```
# Fare and Embarked
plt.figure(figsize = (16,6),dpi = 300)

plt.subplot(1,2,1)
sns.barplot(x = 'Embarked',y = 'Fare',data = titanic)
plt.title('Embarked vs Fare')

plt.subplot(1,2,2)
sns.boxplot(x = 'Embarked',y = 'Fare',data = titanic)
plt.title('Embarked vs Fare')
```

Out[158… Text(0.5, 1.0, 'Embarked vs Fare')





Observation: Passengers from port C paid a higher fare as compared to passengers from port S and Q

In [ ]:
```
# From which port the ratio of travellers in passenger class 1 was the highe
```

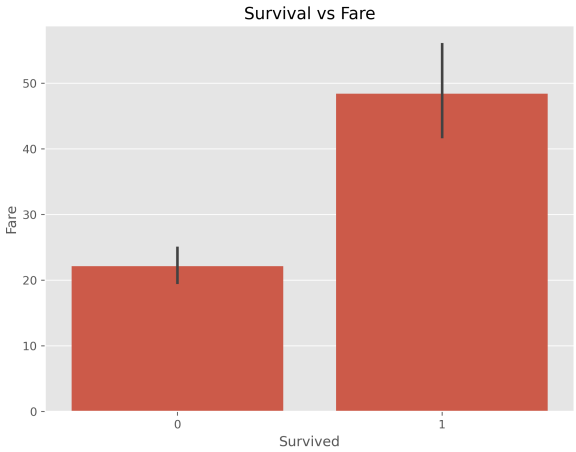
In [161…
```
titanic.groupby('Embarked').Pclass.value_counts(normalize = True)
```

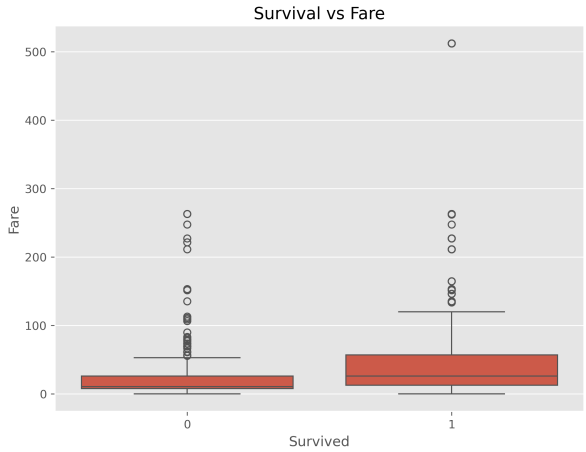
Out[161…
```
Embarked  Pclass
C         1         0.505952
          3         0.392857
          2         0.101190
Q         3         0.935065
          2         0.038961
          1         0.025974
S         3         0.548137
          2         0.254658
          1         0.197205
Name: proportion, dtype: float64
```

Passengers from port C have a higher ratio for Class 1 then port C and S

In [167…
```
# Fare and Survived
plt.figure(figsize = (18,6),dpi = 300)

plt.subplot(1,2,1)
sns.barplot(x = 'Survived',y = 'Fare',data = titanic)
plt.title('Survival vs Fare')

plt.subplot(1,2,2)
sns.boxplot(x = 'Survived',y  = 'Fare',data = titanic)
plt.title('Survival vs Fare')
```

Out[167…
```
Text(0.5, 1.0, 'Survival vs Fare')
```

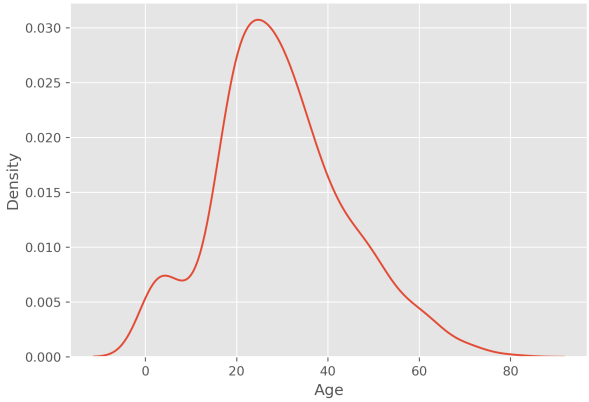


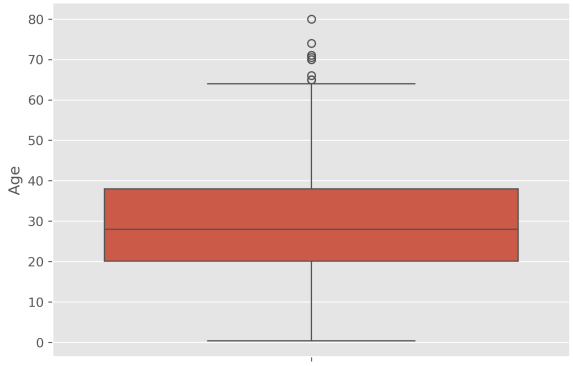


Observation: Passengers who paid higher fare had a higher rate of Survival

In [170…
```
# Age
plt.figure(figsize = (16,5),dpi = 300)
plt.subplot(1,2,1)
sns.distplot(titanic.Age,hist = False)
plt.subplot(1,2,2)
sns.boxplot(y = titanic.Age,data = titanic)
```

Out[170… <Axes: ylabel='Age'>



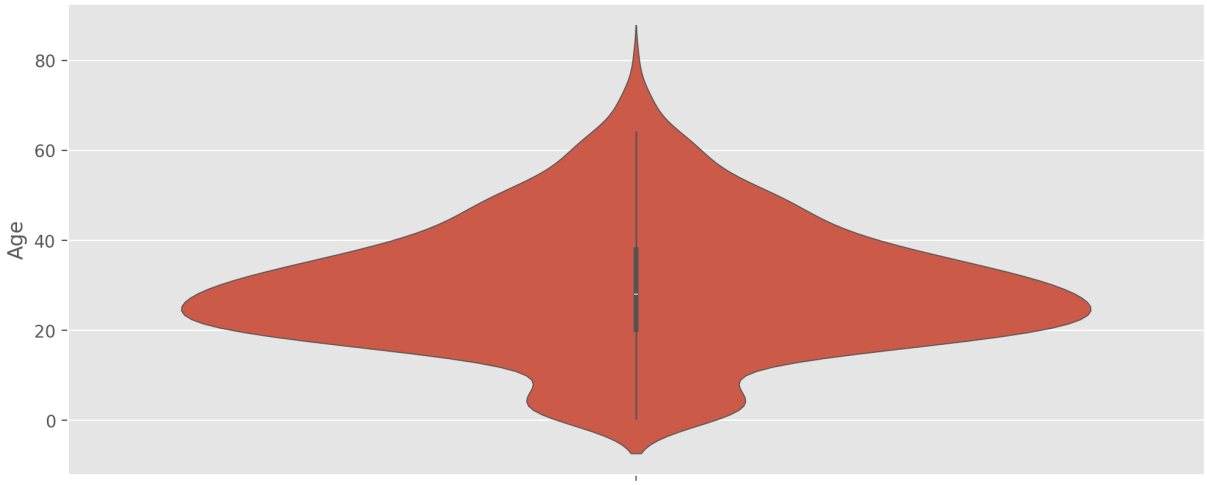


In [172…
```
# What percentage of travellers have age between 20 to 40?
a = len(titanic[(titanic.Age >= 20)*(titanic.Age<=40)])
a/len(titanic) * 100
```

Out[172… 44.89337822671156

In [174…
```
plt.figure(figsize = (12,5),dpi = 200)
sns.violinplot(y = 'Age',data = titanic)
```

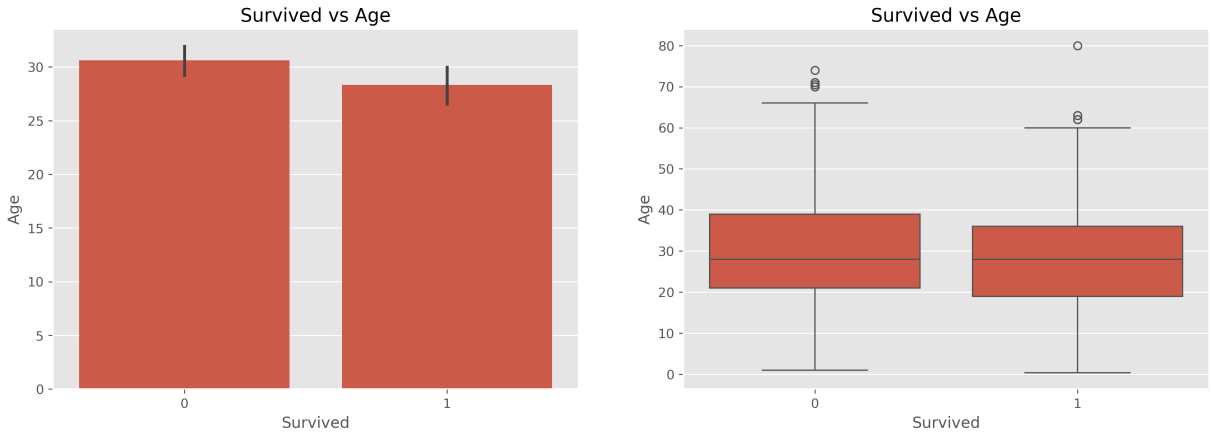
Out[174… <Axes: ylabel='Age'>



Observation: Maximum number of passenger travelling were in the age group of 20 to 40

In [188…
```
# Age and survived
plt.figure(figsize = (16,5),dpi = 300)
plt.subplot(1,2,1)
sns.barplot(x = 'Survived',y = 'Age',data = titanic)
plt.title('Survived vs Age')
plt.subplot(1,2,2)
sns.boxplot(x = 'Survived',y = 'Age',data = titanic)
plt.title('Survived vs Age')
```

Out[188… `Text(0.5, 1.0, 'Survived vs Age')`

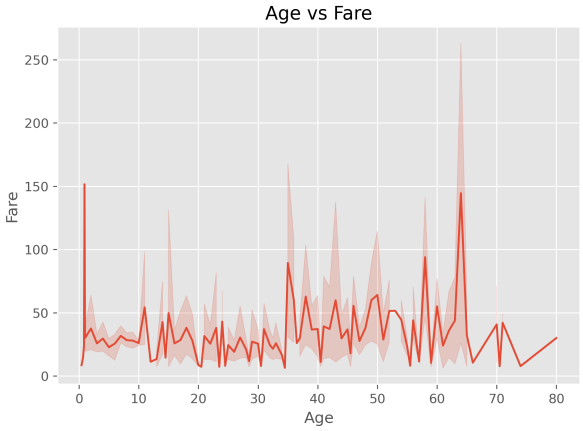


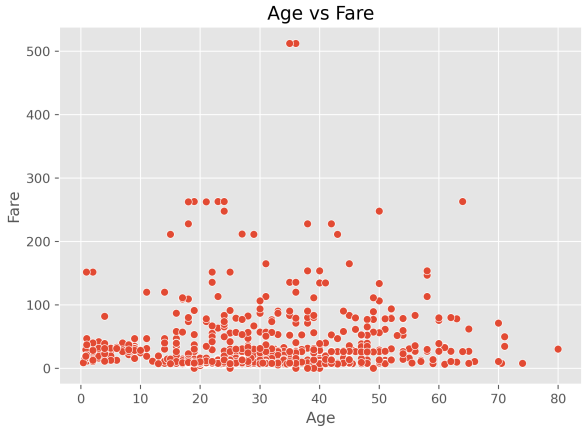
Observation: As apparent from the plots Age has no effect upon the Survival

In [182…
```
# Age and Fare

plt.figure(figsize = (16,5),dpi = 300)
plt.subplot(1,2,1)
sns.lineplot(x = 'Age',y = 'Fare',data = titanic)
plt.title('Age vs Fare')

plt.subplot(1,2,2)
```

```
sns.scatterplot(x = 'Age',y = 'Fare',data = titanic)
plt.title('Age vs Fare')
```

Out[182… Text(0.5, 1.0, 'Age vs Fare')

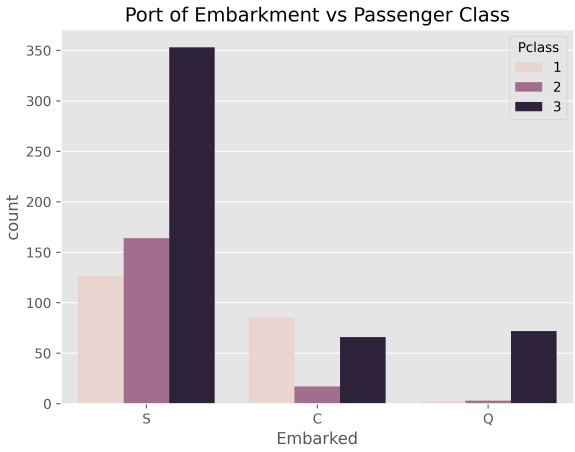


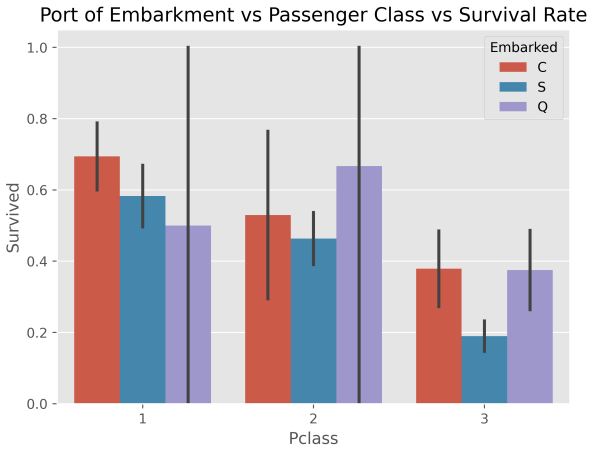


As can be seen the trend is flat hence there is no effect of Age upon the Fare of the ticket.

In [190…
```
plt.figure(figsize = (15,5),dpi = 500)
plt.subplot(1,2,1)
sns.countplot(x = 'Embarked',hue = 'Pclass',data = titanic)
plt.title('Port of Embarkment vs Passenger Class')

plt.subplot(1,2,2)
sns.barplot(x = 'Pclass',y = 'Survived',hue = 'Embarked',data = titanic)
plt.title('Port of Embarkment vs Passenger Class vs Survival Rate')
```

Out[190… Text(0.5, 1.0, 'Port of Embarkment vs Passenger Class vs Survival Rate')





In [192…
```
a = round(titanic.groupby(titanic.Embarked).Survived.value_counts(normalize
b= titanic.Embarked.value_counts().sort_index()
print(a,'\n')
print(b)
```

```
Embarked  Survived
C         0           44.64
          1           55.36
Q         0           61.04
          1           38.96
S         0           66.30
          1           33.70
Name: proportion, dtype: float64

Embarked
C    168
Q     77
S    644
Name: count, dtype: int64
```

Conclusion:

We conclude from the above analysis that Female passengers had a higher survival rate then Male passenger. Also, Passengers from Class 1 and passengers travelling with 1, 2 or 3 family members had a higher Ratio of Survival than those travelling in class 2 and 3 or passengers travelling alone or with 4 or higher number of family members. As the fare for class 1 was higher then that of class 2 and 3, it can also be said that passengers who paid higher Fare had a higher Rate of Survival then others.

In [ ]: