In [35]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
pd.options.display.max_columns = 15
```

In [36]:
```
data = pd.read_csv(r"X:\Data Science\UofT Data Sciencce and AI\Sem 1\Assig

```

In [37]:
```
1 data
```

Out[37]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500

891 rows × 12 columns

In [38]:
```
1 data = data.drop(["Ticket","Cabin", "Embarked"], axis = 1)
```

In [39]:
```
1 data.head()
```

Out[39]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	8.0500

In [40]:
```
1 data.isnull().sum()
```

Out[40]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Fare             0
dtype: int64
```

In [41]:
```
1 data["Age"].fillna(data["Age"].mean(), inplace = True)
```

In [42]:
```
1 data.isnull().sum()
```

Out[42]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Fare           0
dtype: int64
```

In [ ]:
```
1
```

1. Total Survided and Deaths counts

In [43]:
```
1 data["Survived"].count()
```

Out[43]:
```
891
```

In [44]:
```
1 data["Survived"].value_counts()
```

Out[44]:
```
0    549
1    342
Name: Survived, dtype: int64
```

A total of 342 People Survived out of 891 people.

2. Analysis of the survival of people based on class

In [45]:
```
1 Pclass_survived = data.groupby("Survived")["Pclass"].value_counts()[1].to_
2 Pclass_survived = Pclass_survived.rename(columns = {"Pclass": "Alive"})
3 Pclass_survived
```

Out[45]:

Pclass	Alive
1	136
3	119
2	87

In [46]:
```
1 Pclass_Total = data["Pclass"].value_counts().to_frame()
2 Pclass_Total
```

Out[46]:

	Pclass
3	491
1	216
2	184

In [47]:
```
1 Pclass_data = Pclass_survived.merge(Pclass_Total, how = "inner", left_inde
2 Pclass_data = Pclass_data.rename(columns = {"Pclass": "Total_People"}).loc
3 Pclass_data
```

Out[47]:

	Total_People	Alive
1	216	136
3	491	119
2	184	87

In [48]:
```
1 Pclass_data["Alive_Percentage"] = (Pclass_data["Alive"] / Pclass_data["Tot
```

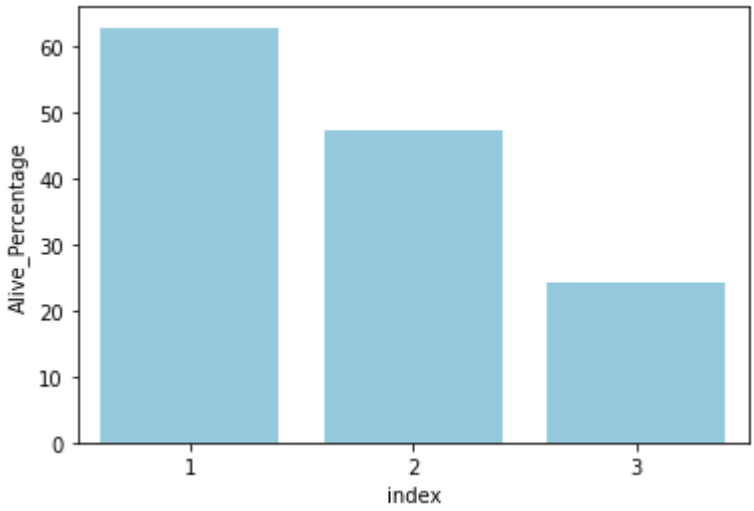
In [49]:
```
Pclass_data = Pclass_data.sort_index()
Pclass_data
```

Out[49]:

	Total_People	Alive	Alive_Percentage
1	216	136	62.962963
2	184	87	47.282609
3	491	119	24.236253

In [50]:
```
sns.barplot(data = Pclass_data.reset_index(), x = "index", y = "Alive_Perc
```

Out[50]: <AxesSubplot:xlabel='index', ylabel='Alive_Percentage'>



The highest People Alive were among the First Class holders which are around 62.96%. Only 24% of people survived who had booked 3rd class.

3. Analysis on the basis of Gender

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

Out[51]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	8.0500

In [52]:
```
total_sex = data["Sex"].value_counts().to_frame().rename(columns= {"Sex":
total_sex
```

Out[52]:

	Total
male	577
female	314

In [53]:
```
Alive_sex = data.groupby("Survived")["Sex"].value_counts()[1].to_frame().r
Alive_sex
```

Out[53]:

	Alive
Sex	
female	233
male	109

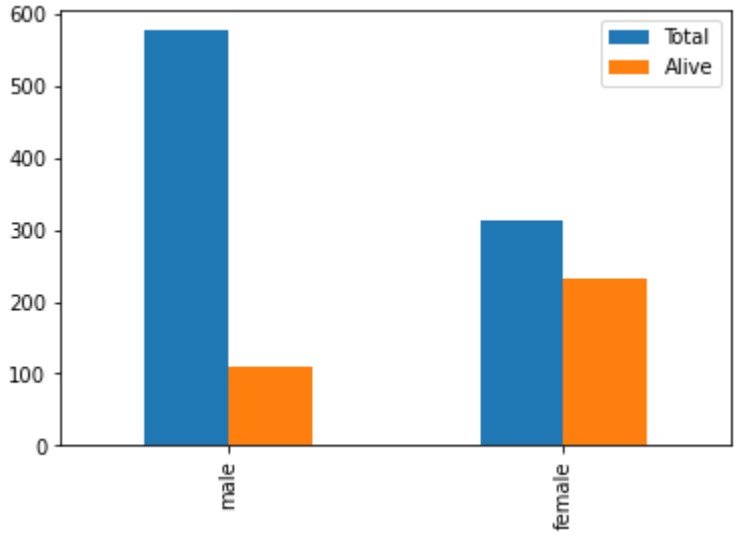
In [54]:
```
sex_data = total_sex.merge(Alive_sex, how = "inner", left_index = True, ri
sex_data
```

Out[54]:

	Total	Alive
male	577	109
female	314	233

In [55]:
```
sex_data.plot(kind = 'bar')
```

Out[55]: <AxesSubplot:>

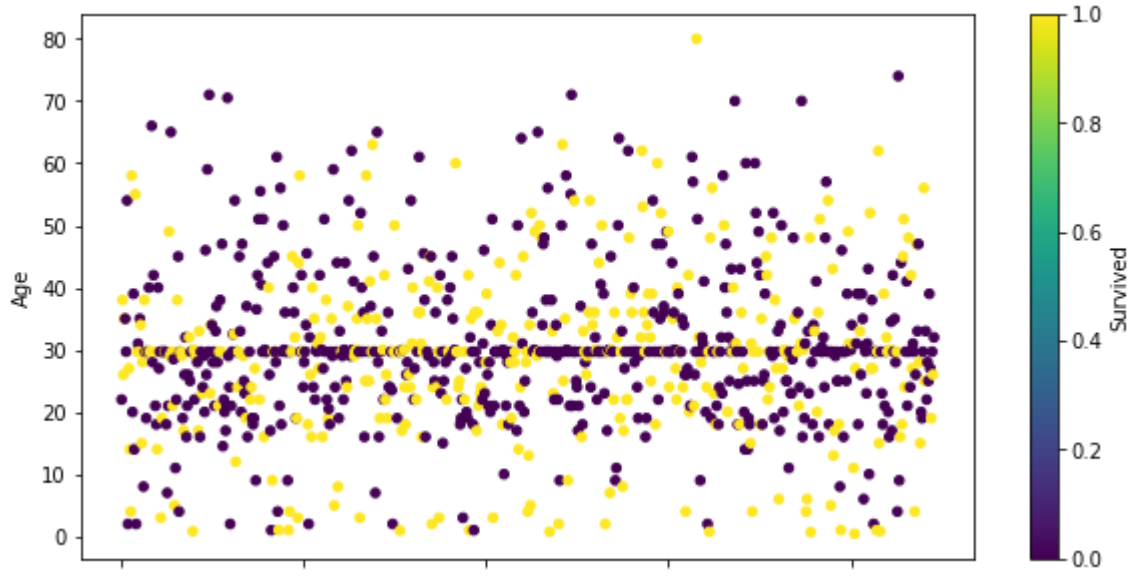


Most of the females had survived than men.

4. Analysis on the basis of Age

In [56]:
```
1 data.plot.scatter(x = "PassengerId", y = "Age",c = "Survived", colormap =
```

Out[56]:
```
<AxesSubplot:xlabel='PassengerId', ylabel='Age'>
```



most of the childeren below 10 years of age has survived and most of the old age people above age 65 has died.

In [57]:
```
1 data[data["Age"] > 75]
```

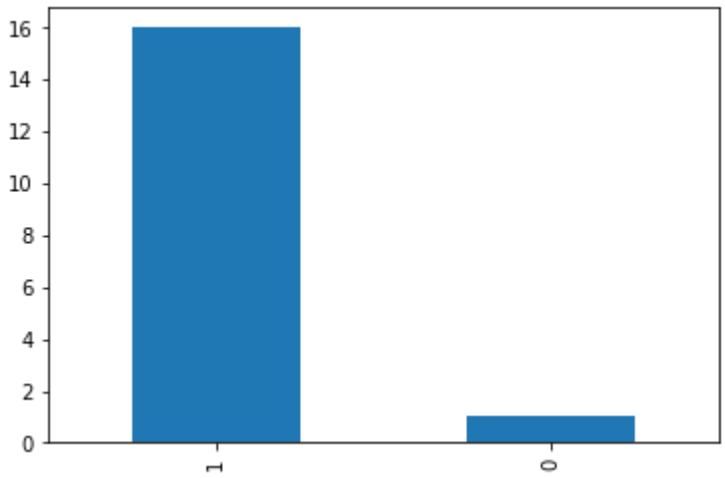
Out[57]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Fare
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	30.0

The interesting fact here is that the oldest person on the ship was Barkworth, Mr. Algernon Henry Wilson had survived.😎

In [58]:
```
1 data[(data["Age"] > 50) & (data["Sex"] == "female")]["Survived"].value_cou
```

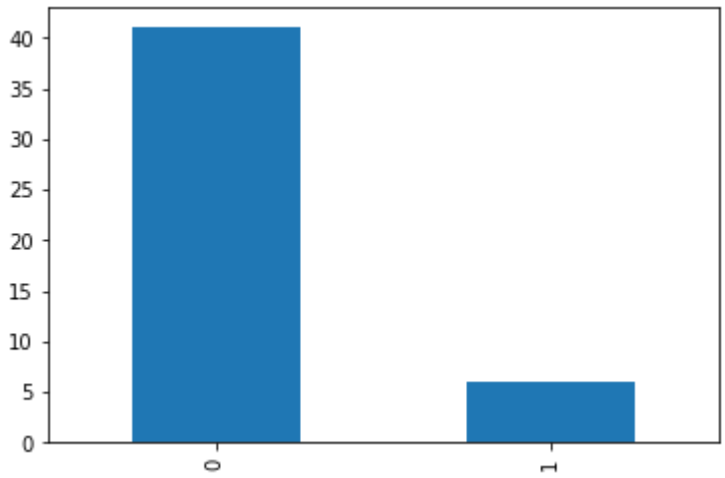
Out[58]: <AxesSubplot:>



There are 17 Women above age 50. All of them has survived except 1. 🧐

In [59]:
```
1 data[(data["Age"] > 50) & (data["Sex"] == "male")]["Survived"].value_count
```

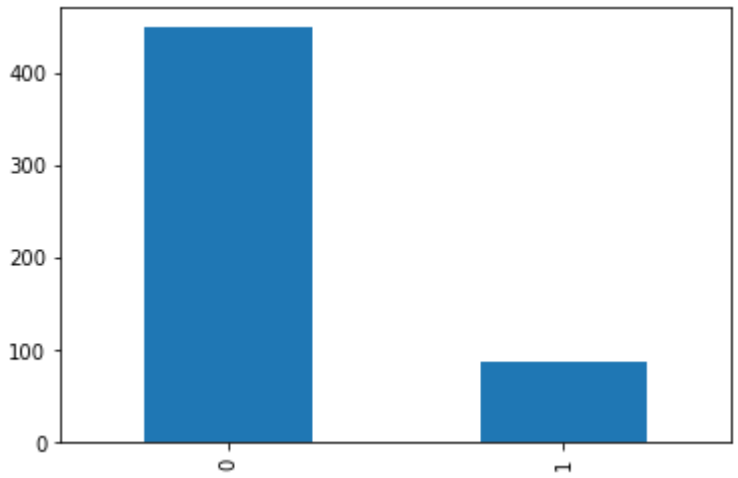
Out[59]: <AxesSubplot:>



There are 47 male above the age 50. Only 6 of them survived.

In [60]:
```
data[(data["Age"] > 15) & (data["Sex"] == "male")]["Survived"].value_count
```

Out[60]: <AxesSubplot:>



In [61]:
```
data["Sex_Num"] = data['Sex']
data['Sex_Num'].replace("female",1, inplace = True)
data["Sex_Num"].replace('male',0, inplace = True)
data.head()
```

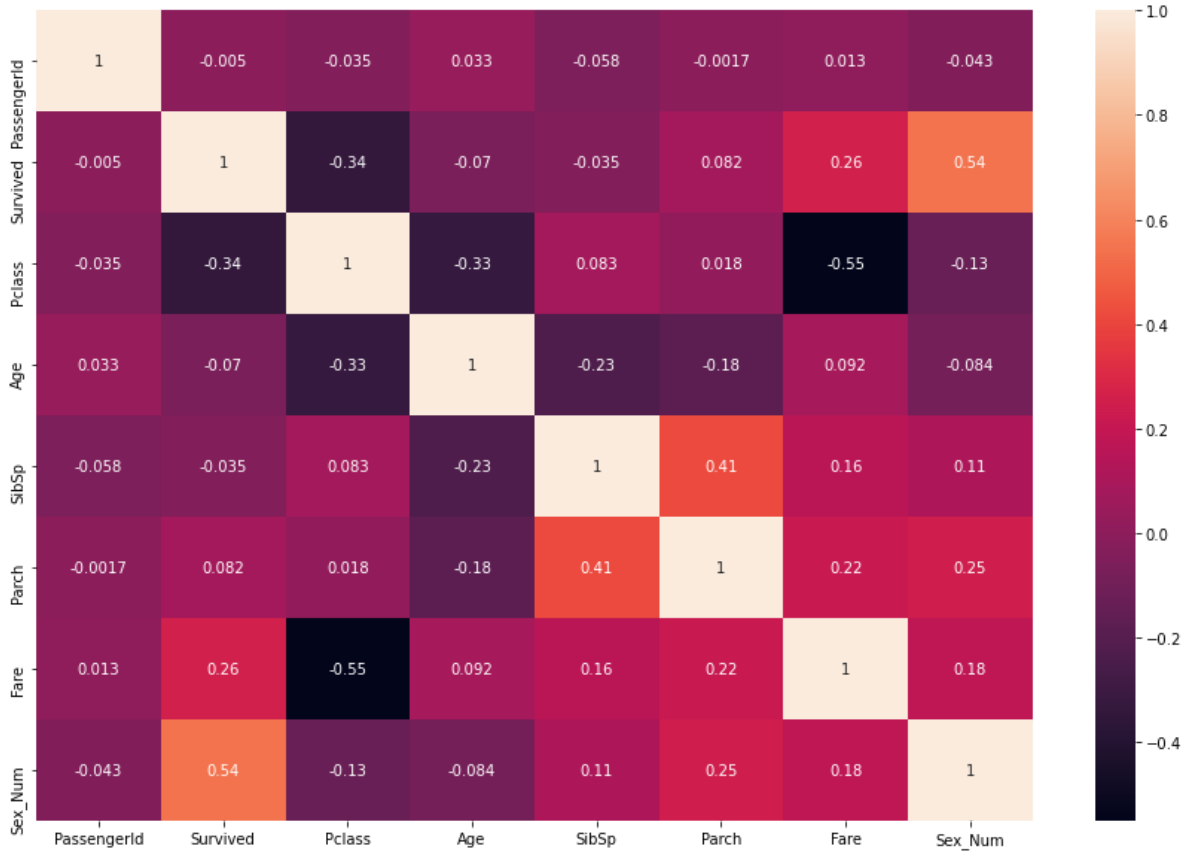
Out[61]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Fare	Sex_Num
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	7.2500	0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	71.2833	1
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	7.9250	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	53.1000	1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	8.0500	0

Converted Sex data into female into 1 and male into 0 to find the correlation.

In [62]:
```
plt.figure(figsize=(15,10))
sns.heatmap(data.corr(), annot = True,)
```

Out[62]: <AxesSubplot:>



1. There is positive high corration between Gender(Sex_Num) and Survived means Female has high chanses of survival than men.
2. There is negative correlationship between Survived and Pclass. If you are in lower class there are high chanses that passenger will die.

In [63]:
```
1 data.head()
```

Out[63]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Fare	Sex_Num
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	7.2500	0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	71.2833	1
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	7.9250	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	53.1000	1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	8.0500	0

In [64]:
```
1 data.groupby("Pclass")["Fare"].mean()
```

Out[64]:
```
Pclass
1    84.154687
2    20.662183
3    13.675550
Name: Fare, dtype: float64
```

QUESTION 1:

What categories of passengers were most likely to survive the Titanic disaster?

ANSWER:

1. Females were more likely to survive in Titanic Disaster because there is a high positive correlation between females and survived columns.
2. Among the females, old age Women had higher chances of survival as only one woman died out of 17 above the age of 60. On the other hand, only one man has survived above the age of 60.

Prediction: -It shows that Titanic crew members have given first preference to transfer old age women and other females on lifeboats.

3. People with the 1st class had more chances to survive as 63 % of people has survived who had traveled with first class on the opposite, only 22% of people had survived who traveled with 3rd class.
4. People, who had paid higher Fares for the voyage had higher chances of survival as there is a positive correlation between Fare and survival of travelers.

Prediction: #There are more chances that people who had paid higher fares had traveled first class. so, they are rich. There is the possibility that they have paid a bribe to crew members to give them a place in a lifeboat. #There are chances that the first class might be located on the top floor on Titanic, so they sank last than the bottom floor. So, they had enough time to get into a lifeboat. #There is the possibility that the Titanic had reserved lifeboats for fluent class.

QUESTION 2:

What other attributes did you use for the analysis? Explain how you used them. Provide a complete list of all attributes used.

ANSWER:

1. Fare and Pclass: There is a positive correlation between Fare and the Survival of passengers. The passengers paid an average of 84 dollars for the first-class ticket in 1912. Which could only afford by rich people. So, fluent-class people had a higher chance of survival.
2. Sex: There is a positive correlation between gender and the survival of the passenger. The female passenger had a higher chance of survival.

QUESTION 3

Did you engineer any attributes? If yes, explain the rationale and how the new attributes were used in the analysis? If you have excluded any attributes from the analysis, provide an explanation why you believe they have to be excluded

ANSWER

1. New Attribute: I created the "Sex_Num" attribute, where I have converted the "Sex" attribute into numeric column male = "0" and Female = "1". to find a correlation between gender and survival. There is a positive correlation (0.54) between Sex_num and survived column. So there was a high chance that Female passengers had a higher survival rate than a male passengers.
2. Droped attribute:I have dropped the "Ticket", "Cabin" and "Embarked" attributes as it is not leading to any analysis.

QUESTION 4:

How did you treat missing values? Provide a detailed explanation in the comments.

ANSWER

1. Age Attribute: There are 177 missing values in the Age column. I have filled the null value with an average age if I delete all values it leads to losing 177 rows which comes to around 20% of the data. Which might lead us to the wrong analysis.
2. Embarked attribute: I have deleted the column as it was not helpful for any analysis.

In [ ]:
```
1
```