The Challenge

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

Data Dictionary

Variable Definition Key survival Survival 0 = No, 1 = Yes,

pclass Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd

sex Sex

Age Age in years sibsp # of siblings / spouses aboard the Titanic parch # of parents / children aboard the Titanic ticket Ticket number fare Passenger fare cabin Cabin number embarked Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton Variable Notes pclass: A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import warnings
warnings.filterwarnings('ignore')

train=pd.read_csv(r"C:\Users\ASUS\Downloads\train (1).csv")
test=pd.read_csv(r"C:\Users\ASUS\Downloads\test (1).csv")

train.head()
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

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

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S

test.head()

   PassengerId  Pclass                                          Name
Sex  \
0          892       3                              Kelly, Mr. James
male
1          893       3              Wilkes, Mrs. James (Ellen Needs)
female
2          894       2                     Myles, Mr. Thomas Francis
male
3          895       3                              Wirz, Mr. Albert
male
4          896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
female

    Age  SibSp  Parch   Ticket     Fare Cabin Embarked
0  34.5      0      0   330911   7.8292   NaN        Q
1  47.0      1      0   363272   7.0000   NaN        S
2  62.0      0      0   240276   9.6875   NaN        Q
3  27.0      0      0   315154   8.6625   NaN        S
4  22.0      1      1  3101298  12.2875   NaN        S

train.shape,test.shape
```

```
((891, 12), (418, 11))
```

```
combine=pd.concat([train,test],ignore_index=True)
```

```
combine.tail()
```

```
      PassengerId  Survived  Pclass                          Name
Sex  \
1304         1305       NaN       3            Spector, Mr. Woolf
male
1305         1306       NaN       1  Oliva y Ocana, Dona. Fermina
female
1306         1307       NaN       3  Saether, Mr. Simon Sivertsen
male
1307         1308       NaN       3           Ware, Mr. Frederick
male
1308         1309       NaN       3      Peter, Master. Michael J
male

       Age  SibSp  Parch              Ticket      Fare Cabin Embarked

1304   NaN      0      0           A.5. 3236    8.0500   NaN        S

1305  39.0      0      0            PC 17758  108.9000  C105        C

1306  38.5      0      0  SOTON/O.Q. 3101262    7.2500   NaN        S

1307   NaN      0      0              359309    8.0500   NaN        S

1308   NaN      1      1                2668   22.3583   NaN        C
```

Univariate Analysis

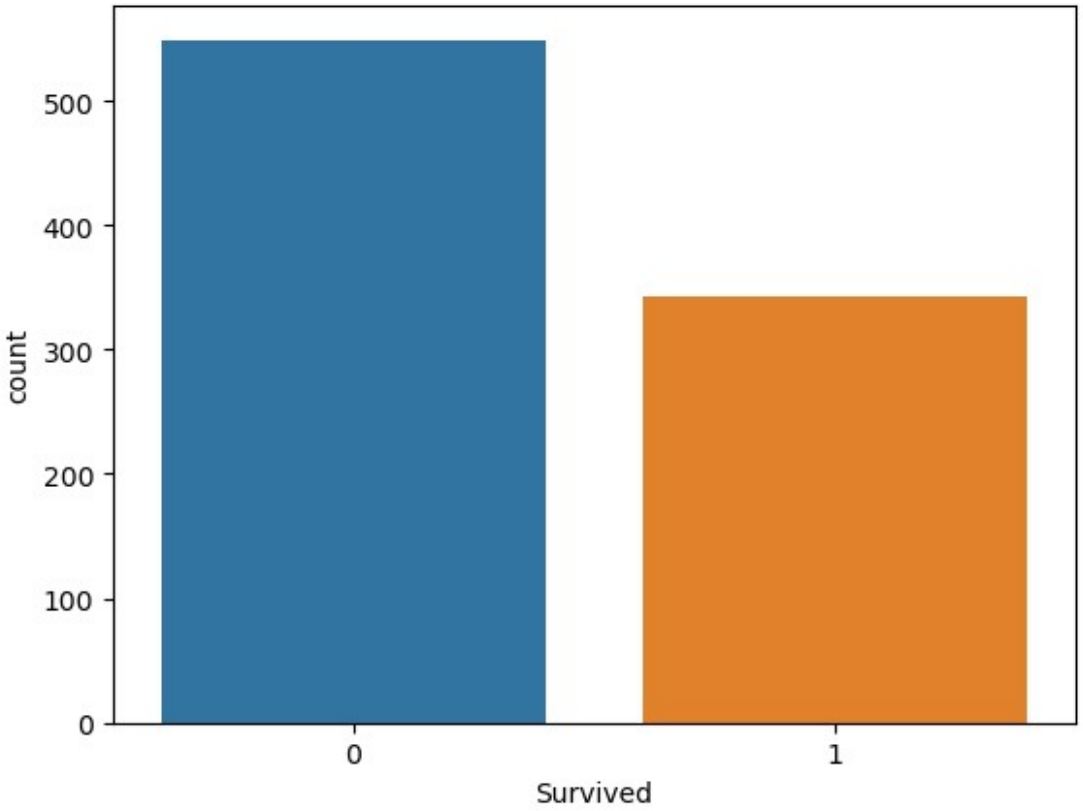
```
combine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  1309 non-null   int64
 1   Survived     891 non-null    float64
 2   Pclass       1309 non-null   int64
 3   Name         1309 non-null   object
 4   Sex          1309 non-null   object
 5   Age          1046 non-null   float64
 6   SibSp        1309 non-null   int64
 7   Parch        1309 non-null   int64
 8   Ticket       1309 non-null   object
 9   Fare         1308 non-null   float64
```

```
 10  Cabin        295 non-null    object
 11  Embarked     1307 non-null   object
dtypes: float64(3), int64(4), object(5)
memory usage: 122.8+ KB
```

```
# Survived plot
sns.countplot(train.Survived)
#Inferenece:The Number people who died more than the Number of people survied
```

```
<AxesSubplot:xlabel='Survived', ylabel='count'>
```

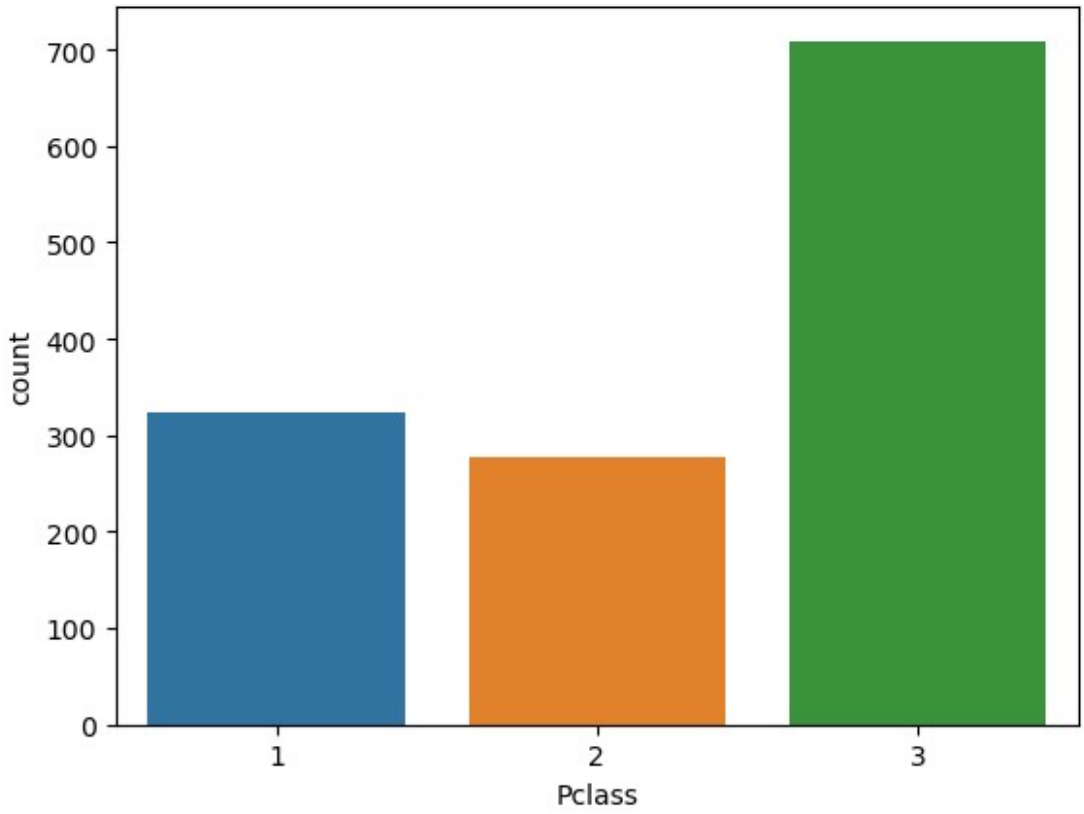



```
#Proportion  of people died or Survived
train.Survived.value_counts(normalize=True)
# 61% people Died,39% people are Survived
```

```
0    0.616162
1    0.383838
Name: Survived, dtype: float64
```

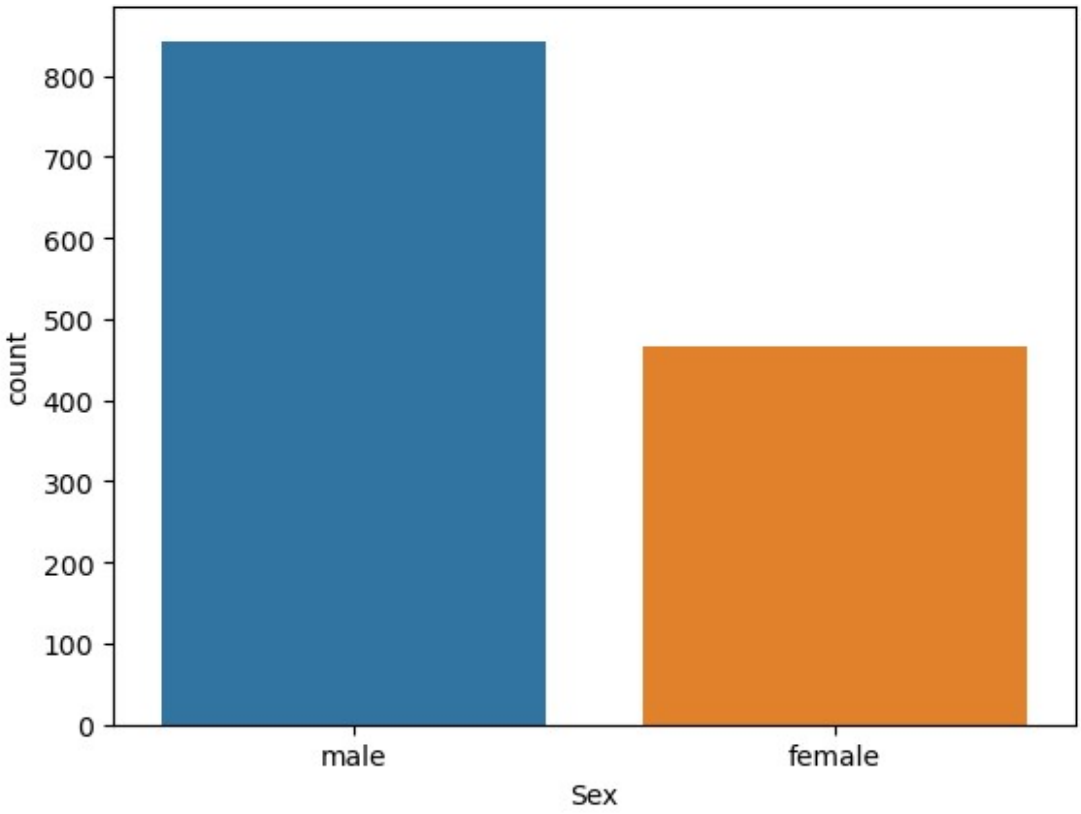
```
#Pclass
sns.countplot(combine.Pclass)
plt.show()
```

```
#Inference:Majority of the Passengers belong to Class3
```



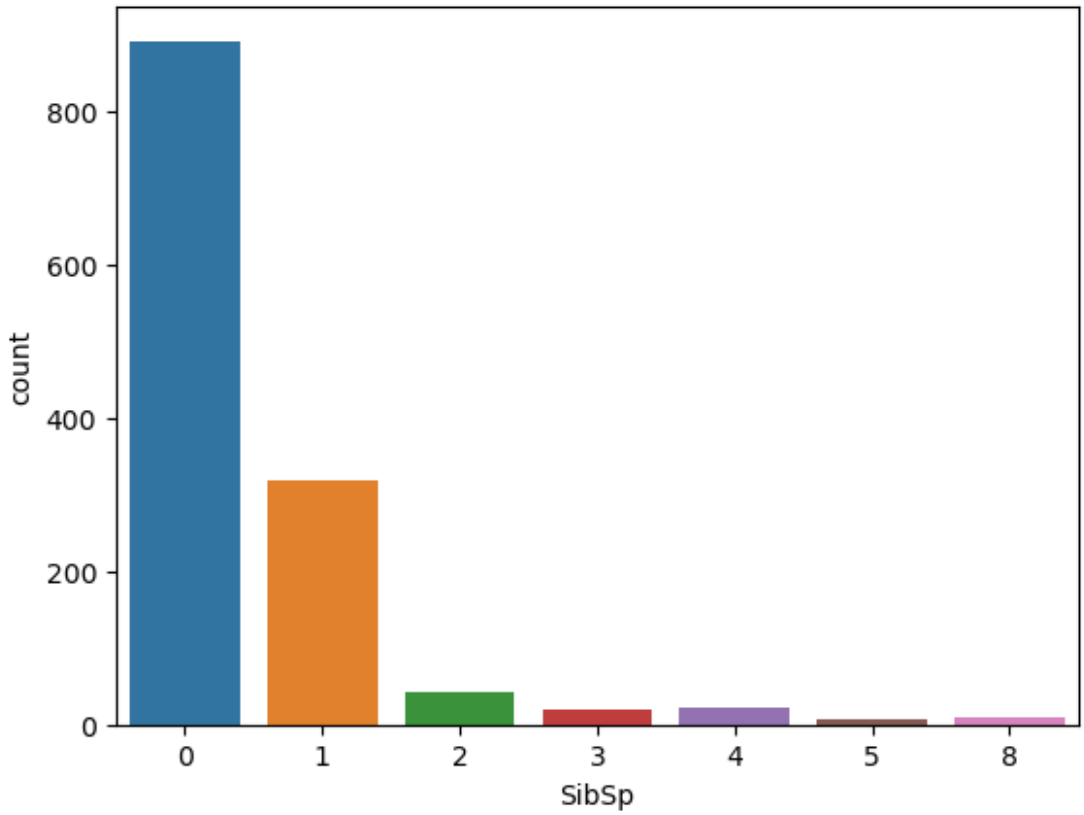

```
#Gender
sns.countplot(combine.Sex)
# count of males are higher than females...
```

```
<AxesSubplot:xlabel='Sex', ylabel='count'>
```



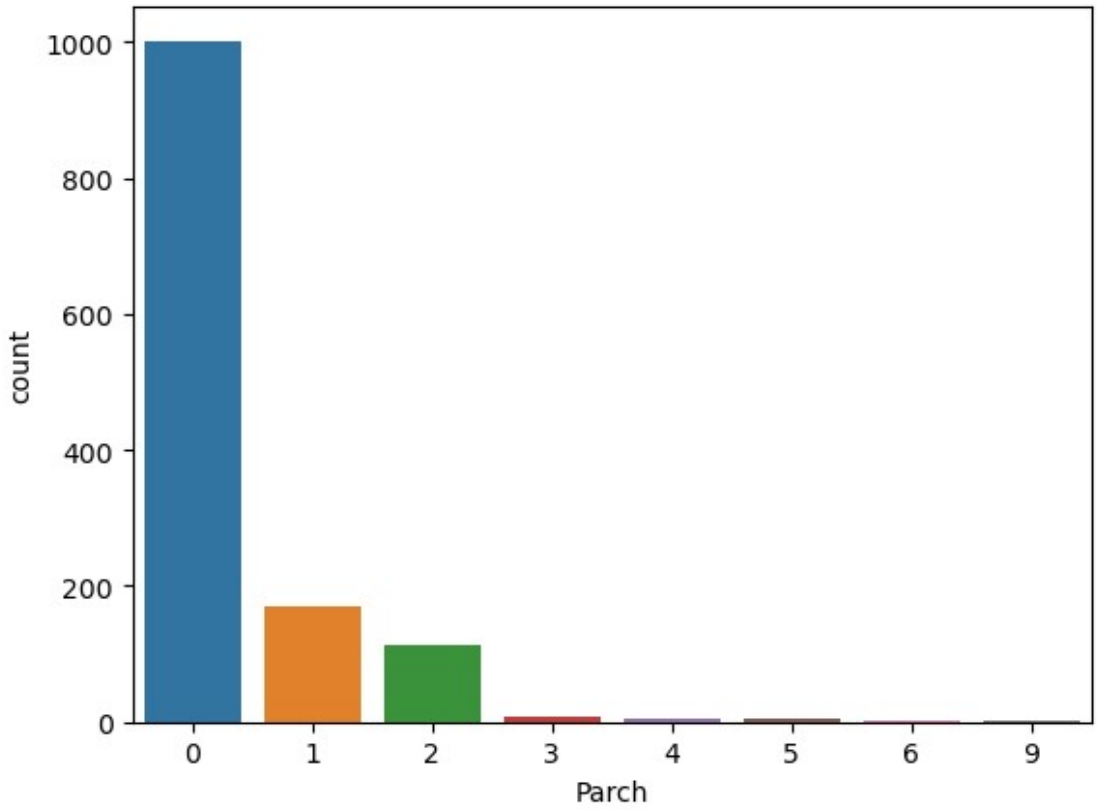

```
#Sibsp
sns.countplot(combine.SibSp)
# 0---menas solo traveller,so most of the traveller are solo traveller
# Maxmium number of people together travelling  are 8
# Hypothesis:Large families may/may not have survived....
```

```
<AxesSubplot:xlabel='SibSp', ylabel='count'>
```



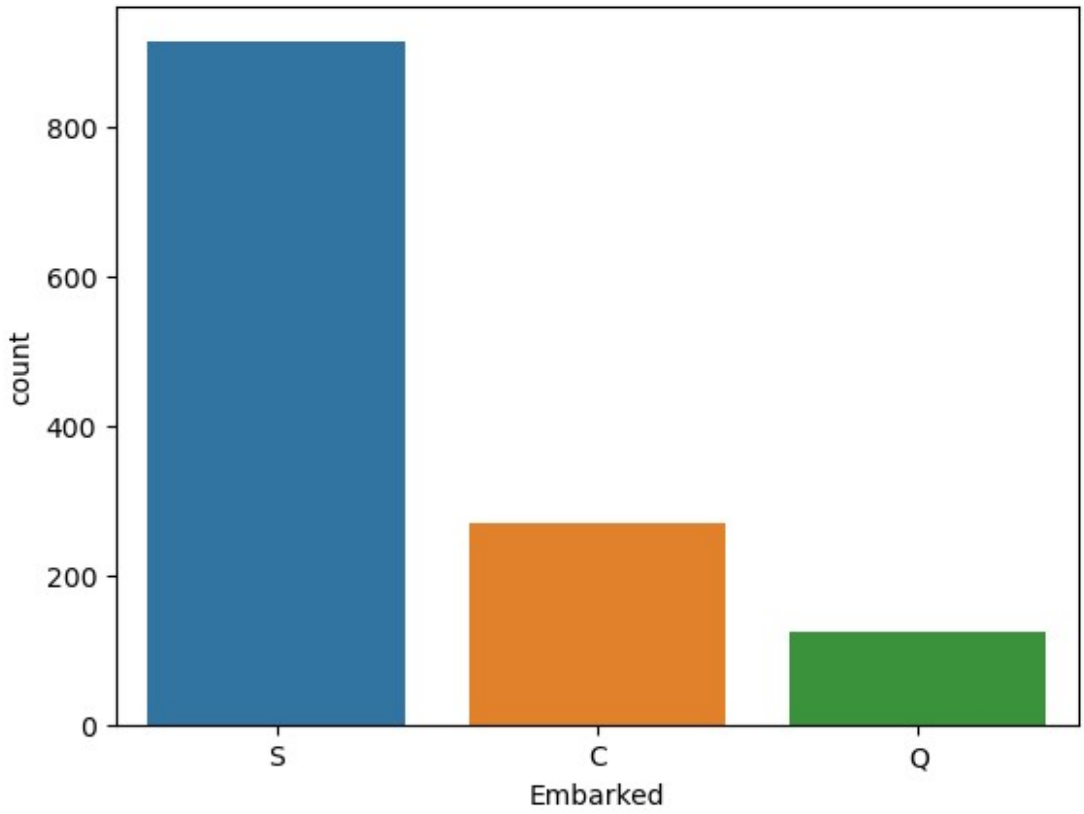

```
#Parch---number of parent and children
sns.countplot(combine.Parch)
#Hypothesis:Large families may/may not have survived....
```

```
<AxesSubplot:xlabel='Parch', ylabel='count'>
```




```
#Embarked:embarked Port of Embarkation C = Cherbourg, Q = Queenstown,
S = Southampton
#Infernce:Most of the traveller boarded from southhampton followed by
Cherbourg
sns.countplot(combine.Embarked)
```

```
<AxesSubplot:xlabel='Embarked', ylabel='count'>
```

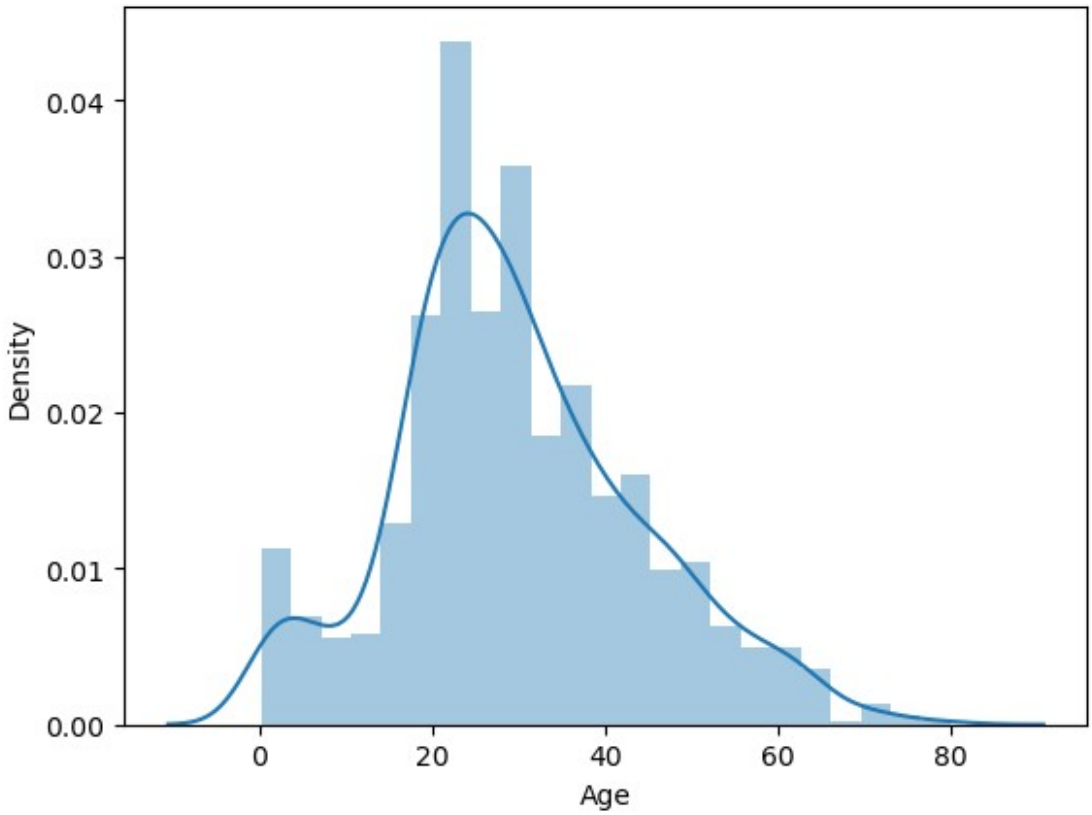



```
# Age
combine.Age.describe()
```

```
count    1046.000000
mean       29.881138
std        14.413493
min         0.170000
25%        21.000000
50%        28.000000
75%        39.000000
max        80.000000
Name: Age, dtype: float64
```

```
sns.distplot(combine.Age)
# Inferences:-Most of the Passengers are in the age range of 20yr to
60yr
```

```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```

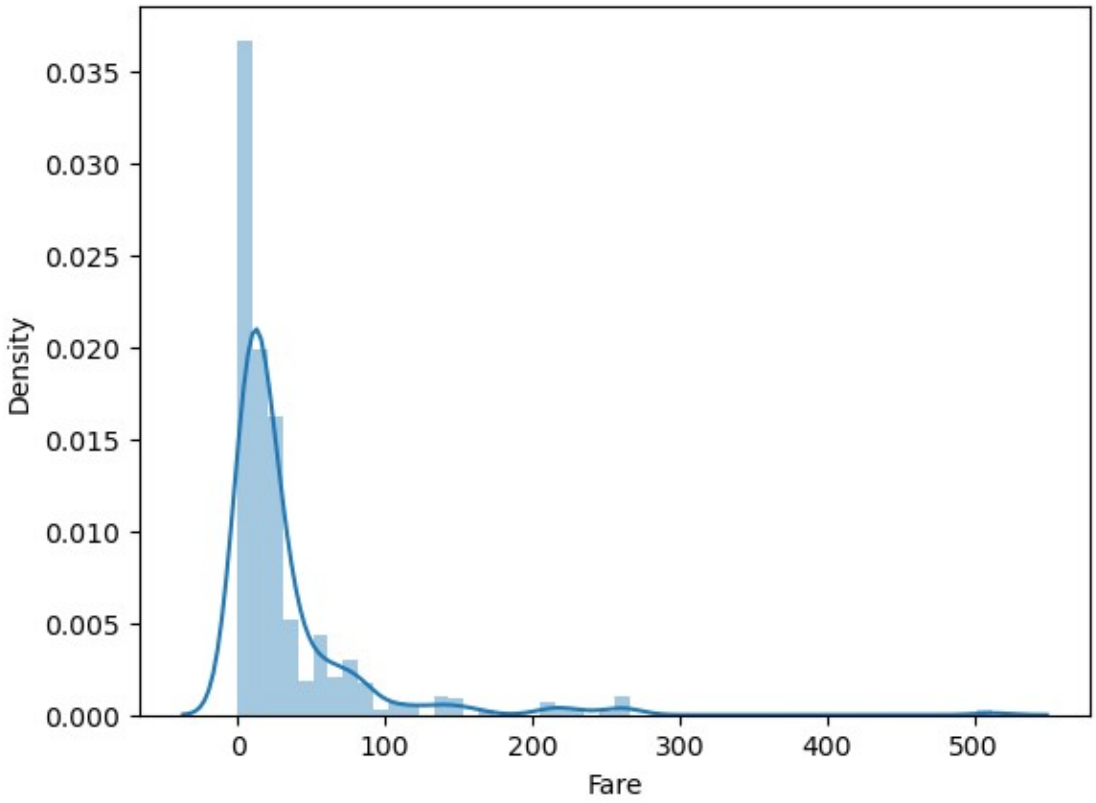



```
combine.Age.skew(),combine.Age.kurtosis()
```

```
(0.40767455974362266, 0.1469476357378139)
```

```
# Fare
sns.distplot(combine.Fare)
# Some of the passenger who travels are paid high Fare
```

```
<AxesSubplot:xlabel='Fare', ylabel='Density'>
```



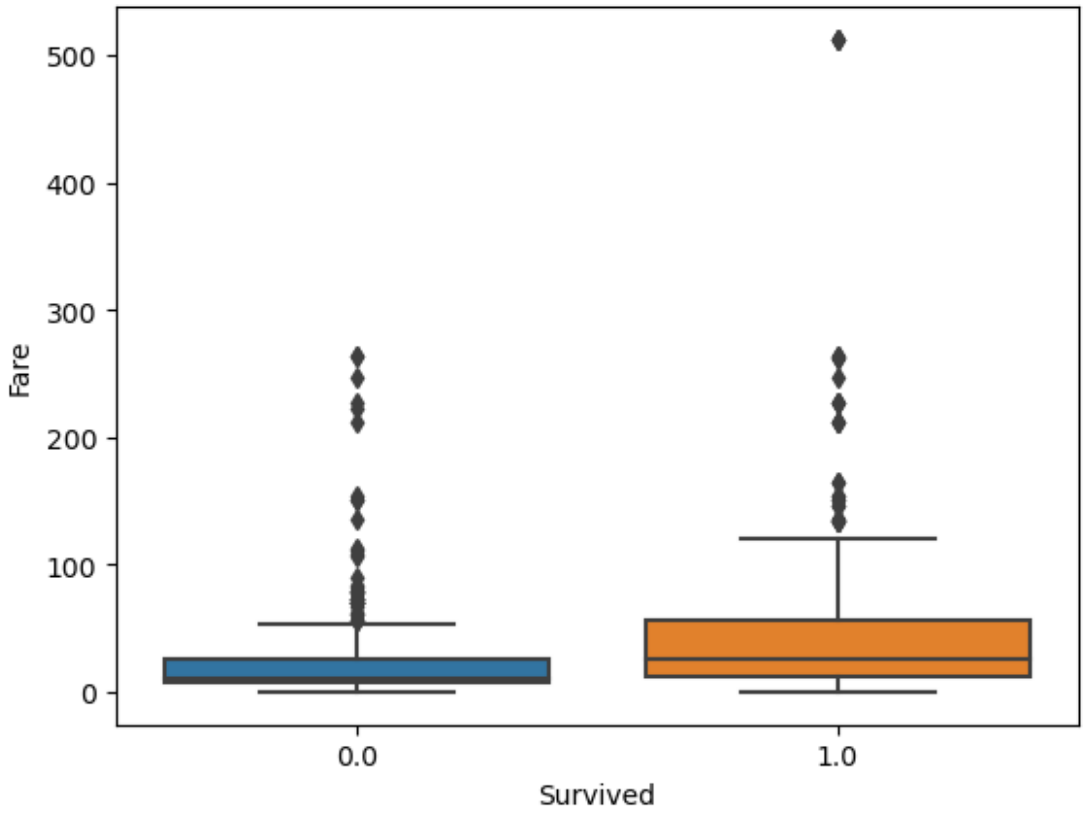

```
combine.Fare.describe()
# Here we didnot remove the Outliers as its a past data
```

```
count    1308.000000
mean       33.295479
std        51.758668
min         0.000000
25%         7.895800
50%        14.454200
75%        31.275000
max       512.329200
Name: Fare, dtype: float64
```

Bivariate Analysis

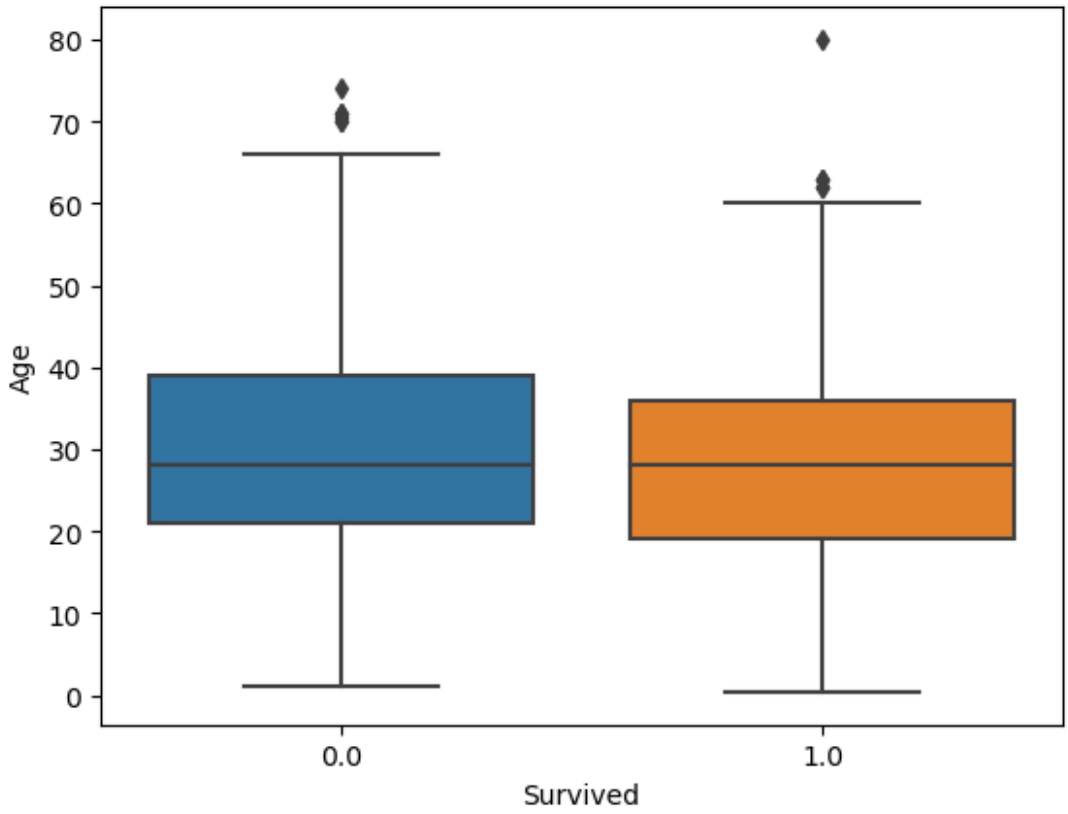
```
#cat vs Num
# Survived Vs Fare
sns.boxplot(combine.Survived,combine.Fare)
# The people who paid more fare had a higher chance of survival...
```

```
<AxesSubplot:xlabel='Survived', ylabel='Fare'>
```




```
# Survived vs Age
sns.boxplot(combine.Survived,combine.Age)
# Passenger who younger had high chance of survival as compare to Old
Passenger
```

```
<AxesSubplot:xlabel='Survived', ylabel='Age'>
```

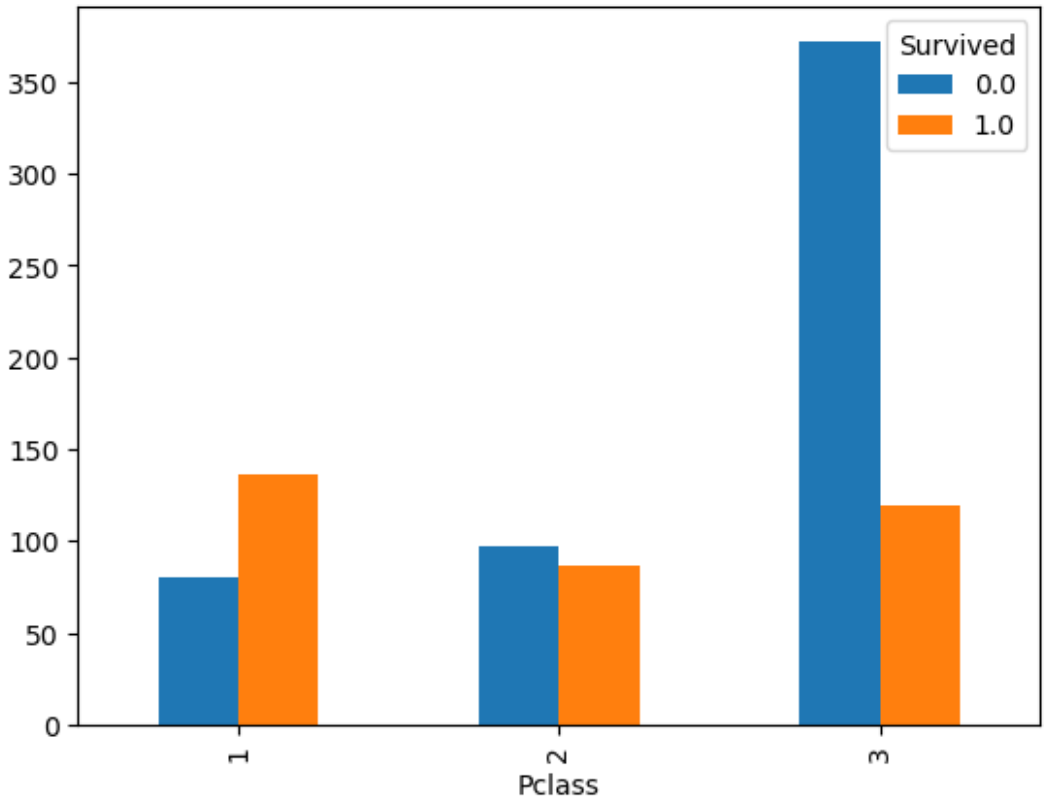



```
# Category vs Category
#Pclass vs Survived
pd.crosstab(combine.Pclass,combine.Survived)
```

```
Survived  0.0  1.0
Pclass
1          80  136
2          97   87
3         372  119
```

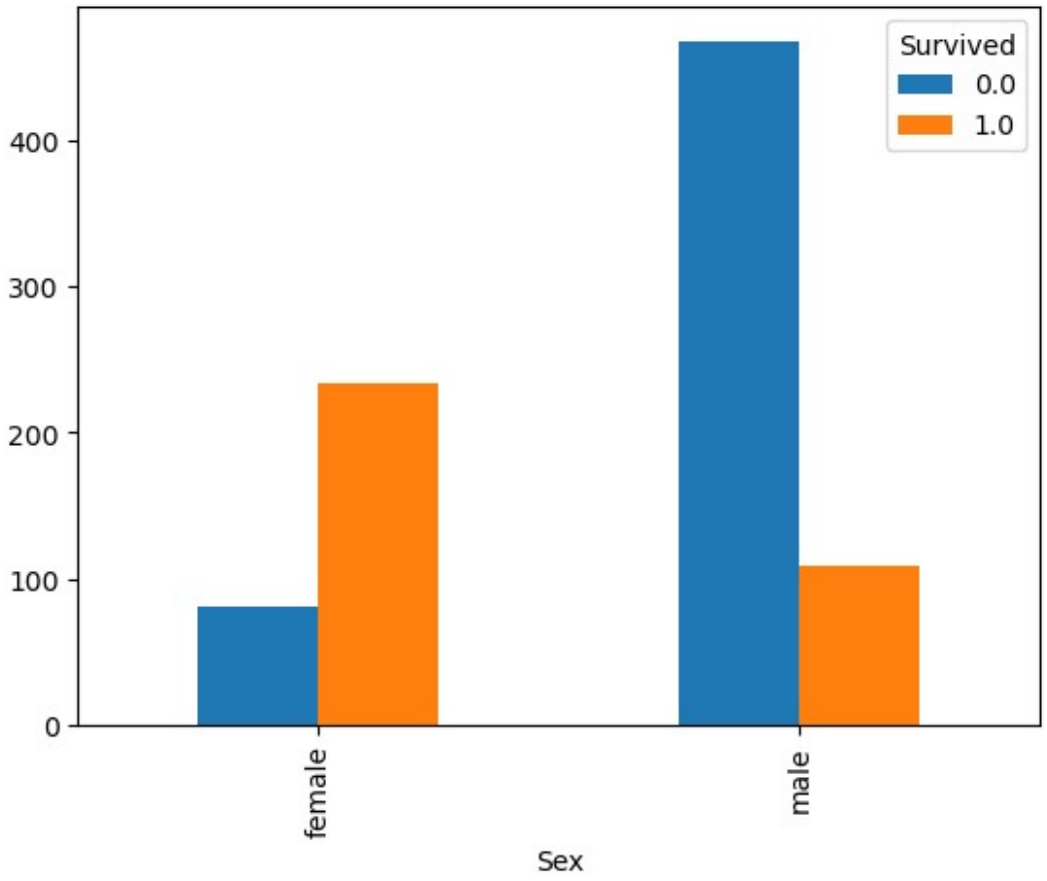
```
pd.crosstab(combine.Pclass,combine.Survived).plot(kind='bar')
#class 3 passenger are the most who did not survived as compare to
others class passenger and class 1 passenger are
#higher survival rates...
```

```
<AxesSubplot:xlabel='Pclass'>
```



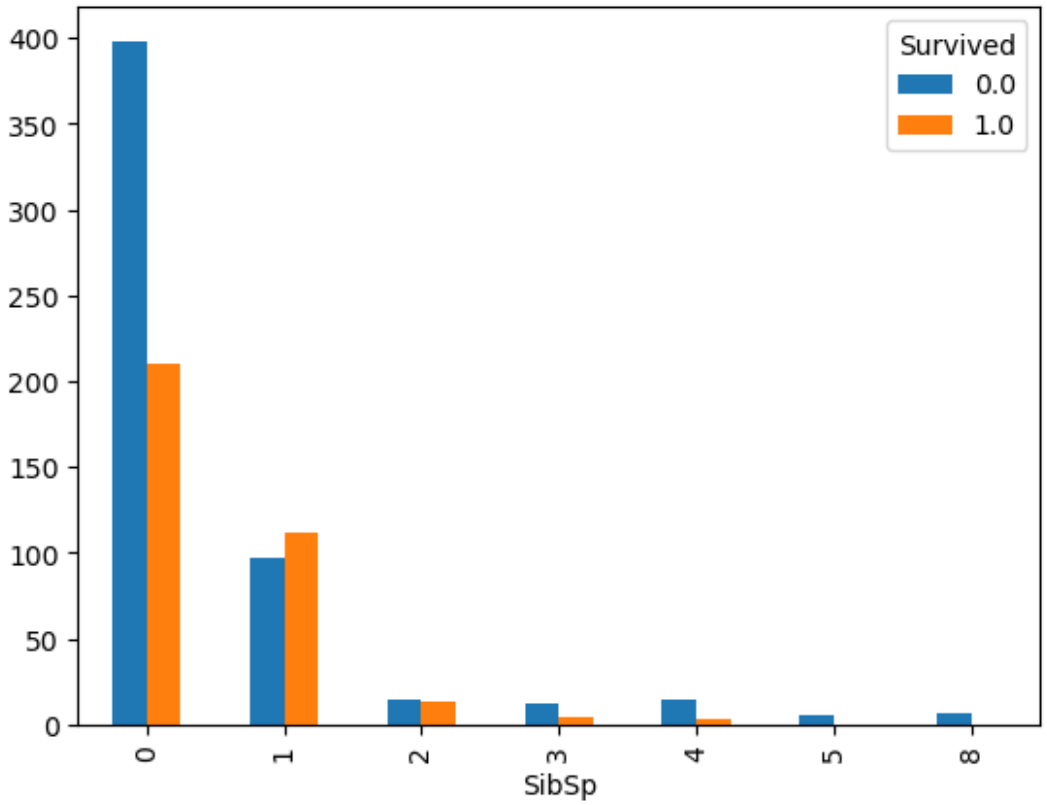

```
#Gender vs Survived
pd.crosstab(combine.Sex,combine.Survived).plot(kind='bar')
# Males have no chance of Survival on Titanic....
# Survival rates of females are high
```

```
<AxesSubplot:xlabel='Sex'>
```



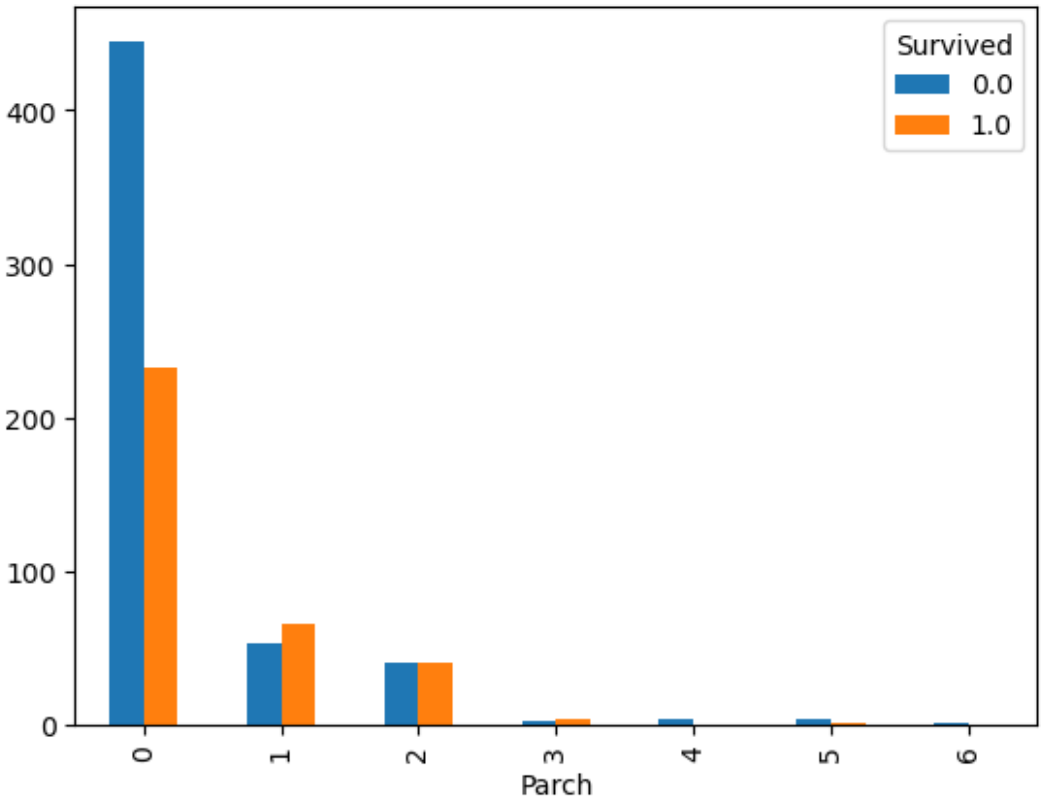

```
# Sibsp vs Survived
pd.crosstab(combine.SibSp,combine.Survived).plot(kind='bar')
#Singles and couples are high chance of survived
```

```
<AxesSubplot:xlabel='SibSp'>
```



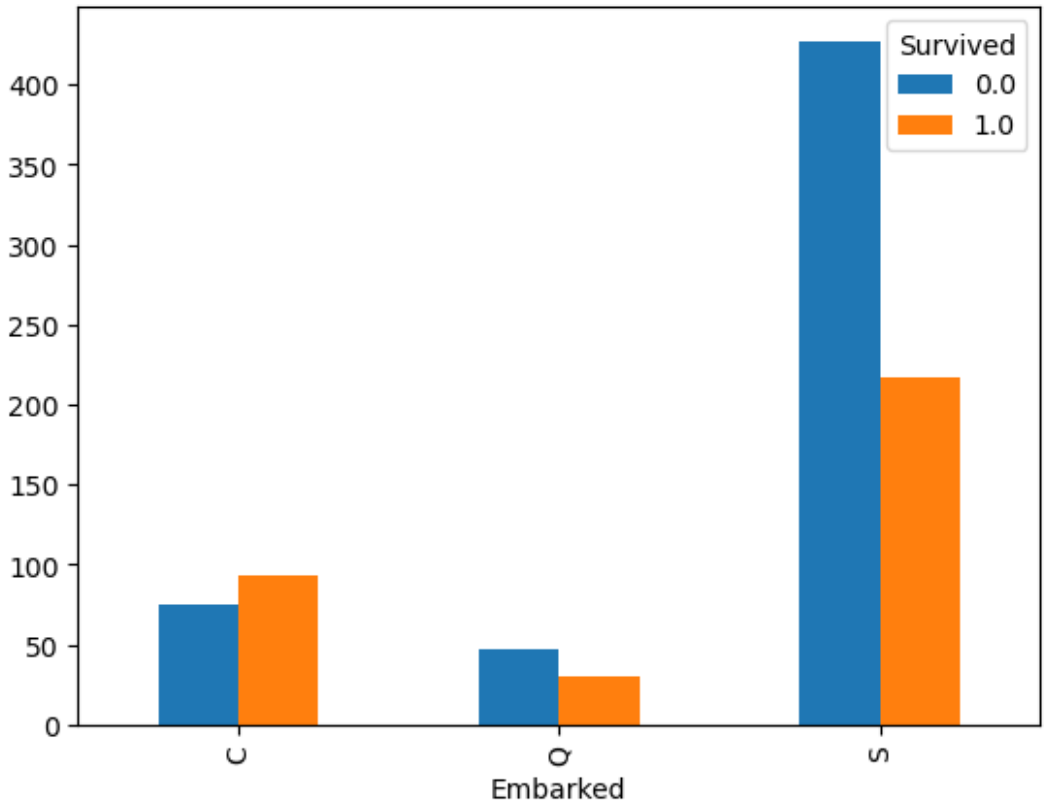

```
# Parch vs Survived
pd.crosstab(combine.Parch,combine.Survived).plot(kind='bar')
# Solo travellers and two family members are able to survived
most.....
```

```
<AxesSubplot:xlabel='Parch'>
```




```
# Embarked vs Survived
pd.crosstab(combine.Embarked,combine.Survived).plot(kind='bar')
#chebourgs have high chances of survival...
```

```
<AxesSubplot:xlabel='Embarked'>
```




```
combine.groupby('Embarked')['Survived'].value_counts(normalize=True)
# people from chebourgs are survived more in terms of percentage
approx 55%
```

```
Embarked  Survived
C         1.0         0.553571
          0.0         0.446429
Q         0.0         0.610390
          1.0         0.389610
S         0.0         0.663043
          1.0         0.336957
Name: Survived, dtype: float64
```

```
combine.groupby(['Embarked','Pclass'])
['Survived'].value_counts(normalize=True)
```

```
Embarked  Pclass  Survived
C         1       1.0         0.694118
                  0.0         0.305882
          2       1.0         0.529412
                  0.0         0.470588
          3       0.0         0.621212
                  1.0         0.378788
Q         1       0.0         0.500000
                  1.0         0.500000
          2       1.0         0.666667
```

```
                  0.0         0.333333
          3       0.0         0.625000
                  1.0         0.375000
S         1       1.0         0.582677
                  0.0         0.417323
          2       0.0         0.536585
                  1.0         0.463415
          3       0.0         0.810198
                  1.0         0.189802
Name: Survived, dtype: float64
```

```
combine.isnull().sum()
```

```
PassengerId       0
Survived        418
Pclass            0
Name              0
Sex               0
Age             263
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          2
dtype: int64
```

```
#Missing Value
```

```
combine.groupby('Sex')['Age'].describe()
```

```
        count       mean        std   min   25%   50%   75%   max
Sex
female  388.0  28.687088  14.576962  0.17  19.0  27.0  38.0  76.0
male    658.0  30.585228  14.280581  0.33  21.0  28.0  39.0  80.0
```

```
combine.groupby(['Pclass','Sex'])['Age'].describe()
```

```
               count       mean        std   min   25%   50%    75%
max
Pclass Sex

1      female  133.0  37.037594  14.272460  2.00  24.0  36.0  48.00
76.0
       male    151.0  41.029272  14.578529  0.92  30.0  42.0  50.00
80.0
2      female  103.0  27.499223  12.911747  0.92  20.0  28.0  34.00
60.0
       male    158.0  30.815380  13.977400  0.67  23.0  29.5  38.75
70.0
3      female  152.0  22.185329  12.205254  0.17  16.0  22.0  30.00
63.0
```

```
       male    349.0  25.962264  11.682415  0.33  20.0  25.0  32.00
74.0
```

```
combine.Name[0].split(", ")[1].split(". ")[0]
```

```
'Mr'
```

```
combine.Name[0].split(', ')[1].split('.')[0]
```

```
'Mr'
```

```
title=[]
for i in combine.Name:
    title.append(i.split(', ')[1].split('.')[0])


combine['Title']=pd.Series(title)
```

```
combine
```

```
      PassengerId  Survived  Pclass  \
0               1       0.0       3
1               2       1.0       1
2               3       1.0       3
3               4       1.0       1
4               5       0.0       3
...           ...       ...     ...
1304         1305       NaN       3
1305         1306       NaN       1
1306         1307       NaN       3
1307         1308       NaN       3
1308         1309       NaN       3

                                                 Name     Sex   Age
SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0
1
1     Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                              Heikkinen, Miss. Laina  female  26.0
0
3          Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                            Allen, Mr. William Henry    male  35.0
0
...                                               ...     ...   ...
...
1304                                Spector, Mr. Woolf    male   NaN
0
1305                      Oliva y Ocana, Dona. Fermina  female  39.0
0
1306                      Saether, Mr. Simon Sivertsen    male  38.5
```

```
0
1307                                Ware, Mr. Frederick    male   NaN
0
1308                            Peter, Master. Michael J    male   NaN
1

      Parch             Ticket      Fare Cabin Embarked   Title
0         0          A/5 21171    7.2500   NaN        S      Mr
1         0           PC 17599   71.2833   C85        C     Mrs
2         0  STON/O2. 3101282    7.9250   NaN        S    Miss
3         0             113803   53.1000  C123        S     Mrs
4         0             373450    8.0500   NaN        S      Mr
...     ...                ...       ...   ...      ...     ...
1304      0          A.5. 3236    8.0500   NaN        S      Mr
1305      0           PC 17758  108.9000  C105        C    Dona
1306      0  SOTON/O.Q. 3101262    7.2500   NaN        S      Mr
1307      0             359309    8.0500   NaN        S      Mr
1308      1               2668   22.3583   NaN        C  Master

[1309 rows x 13 columns]
```

```
combine.Title.value_counts()
```

```
Mr              757
Miss            260
Mrs             197
Master           61
Rev               8
Dr                8
Col               4
Mlle              2
Major             2
Ms                2
Lady              1
Sir               1
Mme               1
Don               1
Capt              1
the Countess      1
Jonkheer          1
Dona              1
Name: Title, dtype: int64
```

```
combine.groupby('Title')['Age'].describe()
```

```
              count       mean        std    min    25%   50%    75%
max
Title

Capt            1.0  70.000000        NaN  70.00  70.00  70.0  70.00
70.0
```

```
Col             4.0  54.000000   5.477226  47.00  51.50  54.5  57.00
60.0
Don             1.0  40.000000        NaN  40.00  40.00  40.0  40.00
40.0
Dona            1.0  39.000000        NaN  39.00  39.00  39.0  39.00
39.0
Dr              7.0  43.571429  11.731115  23.00  38.00  49.0  51.50
54.0
Jonkheer        1.0  38.000000        NaN  38.00  38.00  38.0  38.00
38.0
Lady            1.0  48.000000        NaN  48.00  48.00  48.0  48.00
48.0
Major           2.0  48.500000   4.949747  45.00  46.75  48.5  50.25
52.0
Master         53.0   5.482642   4.161554   0.33   2.00   4.0   9.00
14.5
Miss          210.0  21.774238  12.249077   0.17  15.00  22.0  30.00
63.0
Mlle            2.0  24.000000   0.000000  24.00  24.00  24.0  24.00
24.0
Mme             1.0  24.000000        NaN  24.00  24.00  24.0  24.00
24.0
Mr            581.0  32.252151  12.422089  11.00  23.00  29.0  39.00
80.0
Mrs           170.0  36.994118  12.901767  14.00  27.00  35.5  46.50
76.0
Ms              1.0  28.000000        NaN  28.00  28.00  28.0  28.00
28.0
Rev             8.0  41.250000  12.020815  27.00  29.50  41.5  51.75
57.0
Sir             1.0  49.000000        NaN  49.00  49.00  49.0  49.00
49.0
the Countess    1.0  33.000000        NaN  33.00  33.00  33.0  33.00
33.0

combine

      PassengerId  Survived  Pclass  \
0               1       0.0       3
1               2       1.0       1
2               3       1.0       3
3               4       1.0       1
4               5       0.0       3
...           ...       ...     ...
1304         1305       NaN       3
1305         1306       NaN       1
1306         1307       NaN       3
1307         1308       NaN       3
1308         1309       NaN       3
```

```
                                                 Name     Sex   Age
SibSp  \
0                               Braund, Mr. Owen Harris    male  22.0
1
1     Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                                Heikkinen, Miss. Laina  female  26.0
0
3          Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                              Allen, Mr. William Henry    male  35.0
0
...                                                 ...     ...   ...
...
1304                                 Spector, Mr. Woolf    male   NaN
0
1305                       Oliva y Ocana, Dona. Fermina  female  39.0
0
1306                       Saether, Mr. Simon Sivertsen    male  38.5
0
1307                                Ware, Mr. Frederick    male   NaN
0
1308                           Peter, Master. Michael J    male   NaN
1

      Parch              Ticket      Fare Cabin Embarked   Title
0         0           A/5 21171    7.2500   NaN        S      Mr
1         0            PC 17599   71.2833   C85        C     Mrs
2         0    STON/O2. 3101282    7.9250   NaN        S    Miss
3         0              113803   53.1000  C123        S     Mrs
4         0              373450    8.0500   NaN        S      Mr
...     ...                 ...       ...   ...      ...     ...
1304      0           A.5. 3236    8.0500   NaN        S      Mr
1305      0            PC 17758  108.9000  C105        C    Dona
1306      0  SOTON/O.Q. 3101262    7.2500   NaN        S      Mr
1307      0              359309    8.0500   NaN        S      Mr
1308      1                2668   22.3583   NaN        C  Master

[1309 rows x 13 columns]
```

```
combine.Title.unique()
```

```
array(['Mr', 'Mrs', 'Miss', 'Master', 'Don', 'Rev', 'Dr', 'Mme', 'Ms',
       'Major', 'Lady', 'Sir', 'Mlle', 'Col', 'Capt', 'the Countess',
       'Jonkheer', 'Dona'], dtype=object)
```

```
title_ignore=['Don', 'Rev', 'Dr', 'Mme',
       'Major', 'Lady', 'Sir', 'Mlle', 'Col', 'Capt', 'the Countess',
       'Jonkheer', 'Dona']
```

```
def ignore(x):
    if x in title_ignore:
        return ('Others')
    else:
        return (x)

combine['Titles']=combine.Title.apply(ignore)

combine
```

```
      PassengerId  Survived  Pclass  \
0               1       0.0       3
1               2       1.0       1
2               3       1.0       3
3               4       1.0       1
4               5       0.0       3
...           ...       ...     ...
1304         1305       NaN       3
1305         1306       NaN       1
1306         1307       NaN       3
1307         1308       NaN       3
1308         1309       NaN       3

                                                 Name     Sex   Age
SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0
1
1     Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                              Heikkinen, Miss. Laina  female  26.0
0
3          Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                            Allen, Mr. William Henry    male  35.0
0
...                                               ...     ...   ...
...
1304                                 Spector, Mr. Woolf    male   NaN
0
1305                       Oliva y Ocana, Dona. Fermina  female  39.0
0
1306                       Saether, Mr. Simon Sivertsen    male  38.5
0
1307                                Ware, Mr. Frederick    male   NaN
0
1308                            Peter, Master. Michael J    male   NaN
1

      Parch              Ticket      Fare Cabin Embarked   Title
Titles
0         0           A/5 21171    7.2500   NaN        S      Mr
```

```
Mr
1         0            PC 17599   71.2833   C85        C     Mrs
Mrs
2         0    STON/O2. 3101282    7.9250   NaN        S    Miss
Miss
3         0              113803   53.1000  C123        S     Mrs
Mrs
4         0              373450    8.0500   NaN        S      Mr
Mr
...     ...                 ...       ...   ...      ...     ...     .
..
1304      0           A.5. 3236    8.0500   NaN        S      Mr
Mr
1305      0            PC 17758  108.9000  C105        C    Dona
Others
1306      0  SOTON/O.Q. 3101262    7.2500   NaN        S      Mr
Mr
1307      0              359309    8.0500   NaN        S      Mr
Mr
1308      1                2668   22.3583   NaN        C  Master
Master

[1309 rows x 14 columns]
```

```
combine.groupby('Titles')['Age'].describe()
```

```
         count       mean        std    min   25%   50%   75%   max
Titles
Master   53.0   5.482642   4.161554   0.33   2.0   4.0   9.0  14.5
Miss    210.0  21.774238  12.249077   0.17  15.0  22.0  30.0  63.0
Mr      581.0  32.252151  12.422089  11.00  23.0  29.0  39.0  80.0
Mrs     170.0  36.994118  12.901767  14.00  27.0  35.5  46.5  76.0
Ms        1.0  28.000000        NaN  28.00  28.0  28.0  28.0  28.0
Others   31.0  43.129032  12.309189  23.00  32.5  45.0  52.5  70.0
```

```
combine.groupby('Title')['Age'].describe()
```

```
              count       mean        std    min    25%   50%    75%
max
Title

Capt            1.0  70.000000        NaN  70.00  70.00  70.0  70.00
70.0
Col             4.0  54.000000   5.477226  47.00  51.50  54.5  57.00
60.0
Don             1.0  40.000000        NaN  40.00  40.00  40.0  40.00
40.0
Dona            1.0  39.000000        NaN  39.00  39.00  39.0  39.00
39.0
Dr              7.0  43.571429  11.731115  23.00  38.00  49.0  51.50
54.0
```

```
Jonkheer        1.0  38.000000        NaN  38.00  38.00  38.0  38.00
38.0
Lady            1.0  48.000000        NaN  48.00  48.00  48.0  48.00
48.0
Major           2.0  48.500000   4.949747  45.00  46.75  48.5  50.25
52.0
Master         53.0   5.482642   4.161554   0.33   2.00   4.0   9.00
14.5
Miss          210.0  21.774238  12.249077   0.17  15.00  22.0  30.00
63.0
Mlle            2.0  24.000000   0.000000  24.00  24.00  24.0  24.00
24.0
Mme             1.0  24.000000        NaN  24.00  24.00  24.0  24.00
24.0
Mr            581.0  32.252151  12.422089  11.00  23.00  29.0  39.00
80.0
Mrs           170.0  36.994118  12.901767  14.00  27.00  35.5  46.50
76.0
Ms              1.0  28.000000        NaN  28.00  28.00  28.0  28.00
28.0
Rev             8.0  41.250000  12.020815  27.00  29.50  41.5  51.75
57.0
Sir             1.0  49.000000        NaN  49.00  49.00  49.0  49.00
49.0
the Countess    1.0  33.000000        NaN  33.00  33.00  33.0  33.00
33.0
```

```
combine['Age']=combine.groupby('Titles')['Age'].apply(lambda
x:x.fillna(x.median()))
```

```
combine.head()
```

```
   PassengerId  Survived  Pclass  \
0            1       0.0       3
1            2       1.0       1
2            3       1.0       3
3            4       1.0       1
4            5       0.0       3

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

```
   Parch            Ticket     Fare Cabin Embarked Title Titles
0      0         A/5 21171   7.2500   NaN        S    Mr     Mr
1      0          PC 17599  71.2833   C85        C   Mrs    Mrs
2      0  STON/O2. 3101282   7.9250   NaN        S  Miss   Miss
3      0            113803  53.1000  C123        S   Mrs    Mrs
4      0            373450   8.0500   NaN        S    Mr     Mr
```

```
combine.groupby('Titles')['Age'].describe()
```

```
        count       mean        std    min    25%   50%    75%   max
Titles
Master   61.0   5.288197   3.906924   0.33   2.00   4.0   8.00  14.5
Miss    260.0  21.817654  11.003754   0.17  17.00  22.0  27.00  63.0
Mr      757.0  31.496037  10.966971  11.00  25.00  29.0  35.00  80.0
Mrs     197.0  36.789340  11.991282  14.00  28.00  35.5  45.00  76.0
Ms        2.0  28.000000   0.000000  28.00  28.00  28.0  28.00  28.0
Others   32.0  43.187500  12.113542  23.00  32.75  45.0  52.25  70.0
```

```
combine.isnull().sum()
```

```
PassengerId       0
Survived        418
Pclass            0
Name              0
Sex               0
Age               0
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          2
Title             0
Titles            0
dtype: int64
```

```
combine.loc[combine.Fare.isnull(),
['Fare']]=combine.loc[(combine.Titles=='Mr')&(combine.Pclass==3)&(comb
ine.Embarked=='S')]['Fare'].median()
```

```
combine.loc[(combine.Titles=='Mr')&(combine.Pclass==3)&(combine.Embark
ed=='S')]['Fare'].median()
```

```
7.925
```

```
combine.isnull().sum()
```

```
PassengerId       0
Survived        418
```

```
Pclass            0
Name              0
Sex               0
Age               0
SibSp             0
Parch             0
Ticket            0
Fare              0
Cabin          1014
Embarked          2
Title             0
Titles            0
dtype: int64
```

```
combine[combine.Fare.isnull()]
```

```
Empty DataFrame
Columns: [PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch,
Ticket, Fare, Cabin, Embarked, Title, Titles]
Index: []
```

```
combine[combine.Embarked.isnull()]
```

```
     PassengerId  Survived  Pclass
Name  \
61            62       1.0       1                         Icard, Miss.
Amelie
829          830       1.0       1  Stone, Mrs. George Nelson (Martha
Evelyn)

        Sex   Age  SibSp  Parch  Ticket  Fare Cabin Embarked Title
Titles
61   female  38.0      0      0  113572  80.0   B28      NaN  Miss
Miss
829  female  62.0      0      0  113572  80.0   B28      NaN   Mrs
Mrs
```

```
combine.Cabin.value_counts()
```

```
C23 C25 C27        6
G6                 5
B57 B59 B63 B66    5
C22 C26            4
F33                4
                  ..
A14                1
E63                1
E12                1
E38                1
C105               1
Name: Cabin, Length: 186, dtype: int64
```

```
combine.corr()
```

```
             PassengerId  Survived    Pclass       Age     SibSp
Parch  \
PassengerId     1.000000 -0.005007 -0.038354  0.024678 -0.055224
0.008942
Survived       -0.005007  1.000000 -0.338481 -0.071235 -0.035322
0.081629
Pclass         -0.038354 -0.338481  1.000000 -0.391580  0.060832
0.018322
Age             0.024678 -0.071235 -0.391580  1.000000 -0.214428 -
0.129649
SibSp          -0.055224 -0.035322  0.060832 -0.214428  1.000000
0.373587
Parch           0.008942  0.081629  0.018322 -0.129649  0.373587
1.000000
Fare            0.031027  0.257307 -0.558742  0.179632  0.160389
0.221668

                 Fare
PassengerId  0.031027
Survived     0.257307
Pclass      -0.558742
Age          0.179632
SibSp        0.160389
Parch        0.221668
Fare         1.000000
```

```
combine.Cabin.unique()
```

```
array([nan, 'C85', 'C123', 'E46', 'G6', 'C103', 'D56', 'A6',
       'C23 C25 C27', 'B78', 'D33', 'B30', 'C52', 'B28', 'C83', 'F33',
       'F G73', 'E31', 'A5', 'D10 D12', 'D26', 'C110', 'B58 B60',
'E101',
       'F E69', 'D47', 'B86', 'F2', 'C2', 'E33', 'B19', 'A7', 'C49',
'F4',
       'A32', 'B4', 'B80', 'A31', 'D36', 'D15', 'C93', 'C78', 'D35',
       'C87', 'B77', 'E67', 'B94', 'C125', 'C99', 'C118', 'D7', 'A19',
       'B49', 'D', 'C22 C26', 'C106', 'C65', 'E36', 'C54',
       'B57 B59 B63 B66', 'C7', 'E34', 'C32', 'B18', 'C124', 'C91',
'E40',
       'T', 'C128', 'D37', 'B35', 'E50', 'C82', 'B96 B98', 'E10',
'E44',
       'A34', 'C104', 'C111', 'C92', 'E38', 'D21', 'E12', 'E63',
'A14',
       'B37', 'C30', 'D20', 'B79', 'E25', 'D46', 'B73', 'C95', 'B38',
       'B39', 'B22', 'C86', 'C70', 'A16', 'C101', 'C68', 'A10', 'E68',
       'B41', 'A20', 'D19', 'D50', 'D9', 'A23', 'B50', 'A26', 'D48',
       'E58', 'C126', 'B71', 'B51 B53 B55', 'D49', 'B5', 'B20', 'F
G63',
       'C62 C64', 'E24', 'C90', 'C45', 'E8', 'B101', 'D45', 'C46',
```

```
'D30',
       'E121', 'D11', 'E77', 'F38', 'B3', 'D6', 'B82 B84', 'D17',
'A36',
       'B102', 'B69', 'E49', 'C47', 'D28', 'E17', 'A24', 'C50', 'B42',
       'C148', 'B45', 'B36', 'A21', 'D34', 'A9', 'C31', 'B61', 'C53',
       'D43', 'C130', 'C132', 'C55 C57', 'C116', 'F', 'A29', 'C6',
'C28',
       'C51', 'C97', 'D22', 'B10', 'E45', 'E52', 'A11', 'B11', 'C80',
       'C89', 'F E46', 'B26', 'F E57', 'A18', 'E60', 'E39 E41',
       'B52 B54 B56', 'C39', 'B24', 'D40', 'D38', 'C105'],
dtype=object)
```

```
cabin_avbl=['C85', 'C123', 'E46', 'G6', 'C103', 'D56', 'A6',
       'C23 C25 C27', 'B78', 'D33', 'B30', 'C52', 'B28', 'C83', 'F33',
       'F G73', 'E31', 'A5', 'D10 D12', 'D26', 'C110', 'B58 B60',
'E101',
       'F E69', 'D47', 'B86', 'F2', 'C2', 'E33', 'B19', 'A7', 'C49',
'F4',
       'A32', 'B4', 'B80', 'A31', 'D36', 'D15', 'C93', 'C78', 'D35',
       'C87', 'B77', 'E67', 'B94', 'C125', 'C99', 'C118', 'D7', 'A19',
       'B49', 'D', 'C22 C26', 'C106', 'C65', 'E36', 'C54',
       'B57 B59 B63 B66', 'C7', 'E34', 'C32', 'B18', 'C124', 'C91',
'E40',
       'T', 'C128', 'D37', 'B35', 'E50', 'C82', 'B96 B98', 'E10',
'E44',
       'A34', 'C104', 'C111', 'C92', 'E38', 'D21', 'E12', 'E63',
'A14',
       'B37', 'C30', 'D20', 'B79', 'E25', 'D46', 'B73', 'C95', 'B38',
       'B39', 'B22', 'C86', 'C70', 'A16', 'C101', 'C68', 'A10', 'E68',
       'B41', 'A20', 'D19', 'D50', 'D9', 'A23', 'B50', 'A26', 'D48',
       'E58', 'C126', 'B71', 'B51 B53 B55', 'D49', 'B5', 'B20', 'F
G63',
       'C62 C64', 'E24', 'C90', 'C45', 'E8', 'B101', 'D45', 'C46',
'D30',
       'E121', 'D11', 'E77', 'F38', 'B3', 'D6', 'B82 B84', 'D17',
'A36',
       'B102', 'B69', 'E49', 'C47', 'D28', 'E17', 'A24', 'C50', 'B42',
       'C148', 'B45', 'B36', 'A21', 'D34', 'A9', 'C31', 'B61', 'C53',
       'D43', 'C130', 'C132', 'C55 C57', 'C116', 'F', 'A29', 'C6',
'C28',
       'C51', 'C97', 'D22', 'B10', 'E45', 'E52', 'A11', 'B11', 'C80',
       'C89', 'F E46', 'B26', 'F E57', 'A18', 'E60', 'E39 E41',
       'B52 B54 B56', 'C39', 'B24', 'D40', 'D38', 'C105']
```

```
len(cabin_avbl)
#Total 187 cabin are availabels
```

```
186
```

```
def available(x):
    if x in cabin_avbl:
```

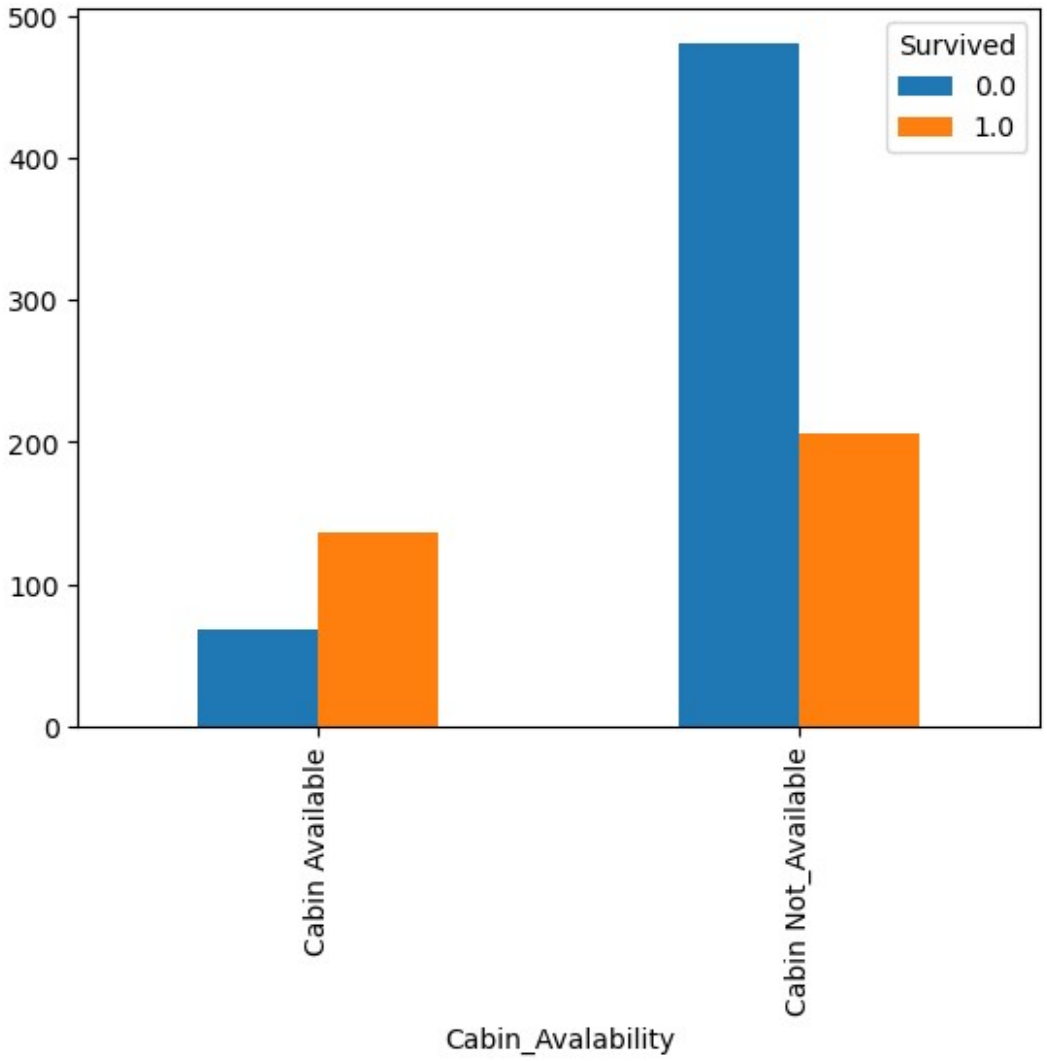
```
        return ('Cabin Available')
    else:
        return('Cabin Not_Available')

combine['Cabin_Avalability']=combine.Cabin.apply(available)

combine
```

```
      PassengerId  Survived  Pclass  \
0               1       0.0       3
1               2       1.0       1
2               3       1.0       3
3               4       1.0       1
4               5       0.0       3
...           ...       ...     ...
1304         1305       NaN       3
1305         1306       NaN       1
1306         1307       NaN       3
1307         1308       NaN       3
1308         1309       NaN       3

                                                 Name     Sex   Age
SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0
1
1     Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                              Heikkinen, Miss. Laina  female  26.0
0
3          Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                            Allen, Mr. William Henry    male  35.0
0
...                                               ...     ...   ...
...
1304                               Spector, Mr. Woolf    male  29.0
0
1305                     Oliva y Ocana, Dona. Fermina  female  39.0
0
1306                     Saether, Mr. Simon Sivertsen    male  38.5
0
1307                              Ware, Mr. Frederick    male  29.0
0
1308                         Peter, Master. Michael J    male   4.0
1

      Parch              Ticket      Fare Cabin Embarked   Title
Titles  \
0         0           A/5 21171    7.2500   NaN        S      Mr
Mr
1         0            PC 17599   71.2833   C85        C     Mrs
```

```
Mrs
2         0    STON/O2. 3101282    7.9250   NaN        S    Miss
Miss
3         0              113803   53.1000  C123        S     Mrs
Mrs
4         0              373450    8.0500   NaN        S      Mr
Mr
...     ...                 ...       ...   ...      ...     ...     .
..
1304      0           A.5. 3236    8.0500   NaN        S      Mr
Mr
1305      0            PC 17758  108.9000  C105        C    Dona
Others
1306      0  SOTON/O.Q. 3101262    7.2500   NaN        S      Mr
Mr
1307      0              359309    8.0500   NaN        S      Mr
Mr
1308      1                2668   22.3583   NaN        C  Master
Master

       Cabin_Avalability
0     Cabin Not_Available
1         Cabin Available
2     Cabin Not_Available
3         Cabin Available
4     Cabin Not_Available
...                   ...
1304  Cabin Not_Available
1305      Cabin Available
1306  Cabin Not_Available
1307  Cabin Not_Available
1308  Cabin Not_Available

[1309 rows x 15 columns]
```

```
pd.crosstab(combine.Cabin_Avalability,combine.Survived).plot(kind='bar
')
#Inference: Cabin_Available passengers are survived more as compare to
Cabin not available people
```

```
<AxesSubplot:xlabel='Cabin_Avalability'>
```




```
combine.Cabin_Avalability.value_counts()

Cabin Not_Available    1014
Cabin Available         295
Name: Cabin_Avalability, dtype: int64

combine.head()

   PassengerId  Survived  Pclass  \
0            1       0.0       3
1            2       1.0       1
2            3       1.0       3
3            4       1.0       1
4            5       0.0       3

                                                Name     Sex   Age
SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0
```

```
1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                             Heikkinen, Miss. Laina  female  26.0
0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                           Allen, Mr. William Henry    male  35.0
0

   Parch            Ticket     Fare Cabin Embarked Title Titles  \
0      0         A/5 21171   7.2500   NaN        S    Mr     Mr
1      0          PC 17599  71.2833   C85        C   Mrs    Mrs
2      0  STON/O2. 3101282   7.9250   NaN        S  Miss   Miss
3      0            113803  53.1000  C123        S   Mrs    Mrs
4      0            373450   8.0500   NaN        S    Mr     Mr

     Cabin_Avalability
0  Cabin Not_Available
1      Cabin Available
2  Cabin Not_Available
3      Cabin Available
4  Cabin Not_Available
```

```
new_data=combine.drop(['Name','PassengerId','Ticket','Cabin'],axis=1)
```

```
new_data.head()
```

```
   Survived  Pclass     Sex   Age  SibSp  Parch     Fare Embarked
Title  \
0       0.0       3    male  22.0      1      0   7.2500        S
Mr
1       1.0       1  female  38.0      1      0  71.2833        C
Mrs
2       1.0       3  female  26.0      0      0   7.9250        S
Miss
3       1.0       1  female  35.0      1      0  53.1000        S
Mrs
4       0.0       3    male  35.0      0      0   8.0500        S
Mr

  Titles    Cabin_Avalability
0     Mr  Cabin Not_Available
1    Mrs      Cabin Available
2   Miss  Cabin Not_Available
3    Mrs      Cabin Available
4     Mr  Cabin Not_Available
```

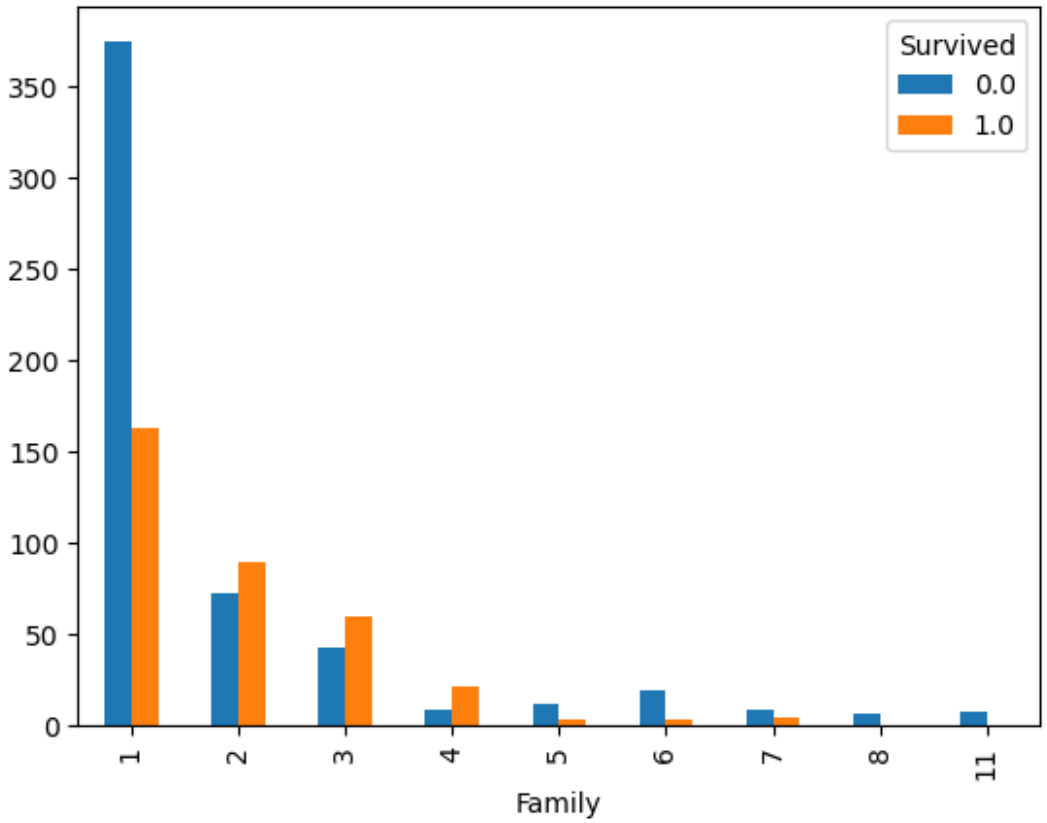
```
# Family
new_data['Family']=new_data.SibSp+new_data.Parch+1
```

```
new_data
```

```
      Survived  Pclass     Sex   Age  SibSp  Parch      Fare Embarked
Title  \
0          0.0       3    male  22.0      1      0    7.2500        S
Mr
1          1.0       1  female  38.0      1      0   71.2833        C
Mrs
2          1.0       3  female  26.0      0      0    7.9250        S
Miss
3          1.0       1  female  35.0      1      0   53.1000        S
Mrs
4          0.0       3    male  35.0      0      0    8.0500        S
Mr
...        ...     ...     ...   ...    ...    ...       ...      ...
...
1304       NaN       3    male  29.0      0      0    8.0500        S
Mr
1305       NaN       1  female  39.0      0      0  108.9000        C
Dona
1306       NaN       3    male  38.5      0      0    7.2500        S
Mr
1307       NaN       3    male  29.0      0      0    8.0500        S
Mr
1308       NaN       3    male   4.0      1      1   22.3583        C
Master

      Titles    Cabin_Avalability  Family
0         Mr  Cabin Not_Available       2
1        Mrs      Cabin Available       2
2       Miss  Cabin Not_Available       1
3        Mrs      Cabin Available       2
4         Mr  Cabin Not_Available       1
...      ...                  ...     ...
1304      Mr  Cabin Not_Available       1
1305  Others      Cabin Available       1
1306      Mr  Cabin Not_Available       1
1307      Mr  Cabin Not_Available       1
1308  Master  Cabin Not_Available       3

[1309 rows x 12 columns]
```

```
pd.crosstab(new_data.Family,new_data.Survived).plot(kind='bar')
#People who are travelling alone are high chance of survival
```

```
<AxesSubplot:xlabel='Family'>
```



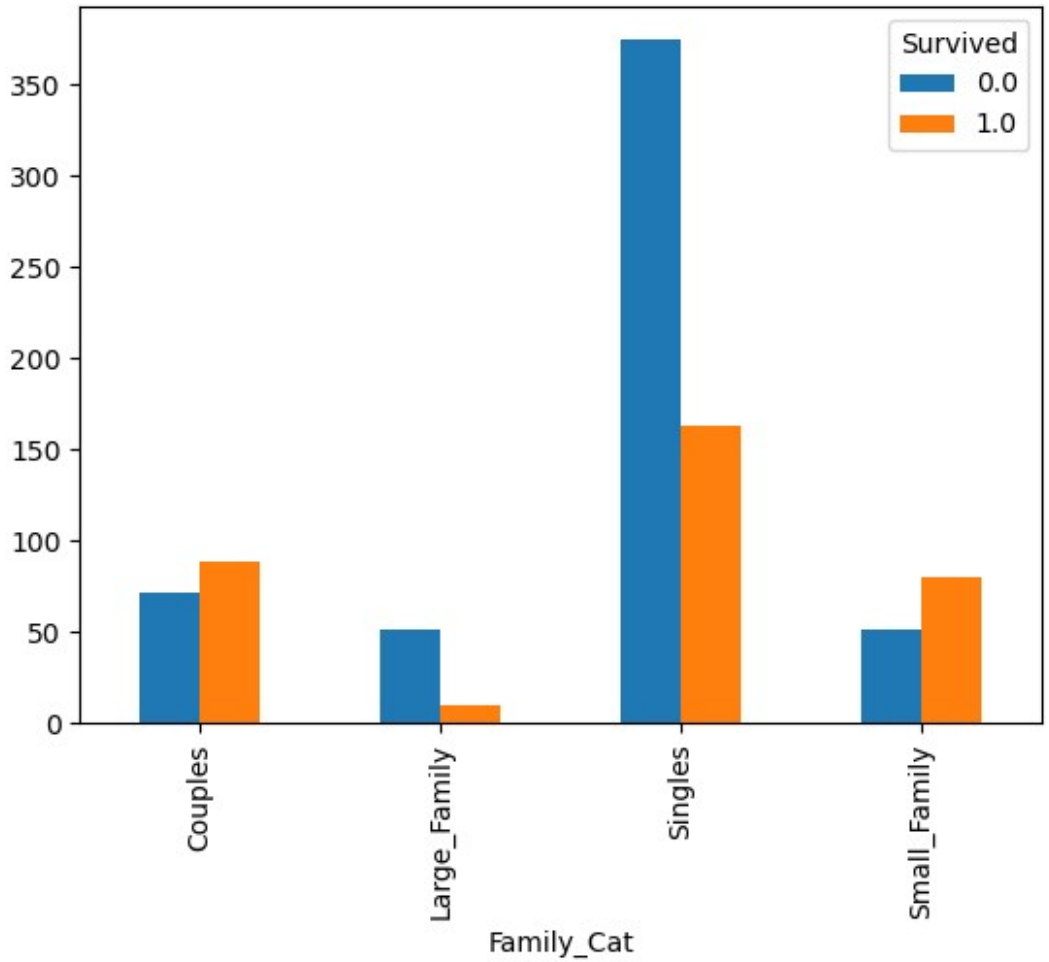

```
# When your Tagetvariable is Category in nature then its refers that
we can create more category

# Binnig Family
new_data
def fam(x):
    if x>=5:
        return ('Large_Family')
    elif (x>=3):
        return ('Small_Family')
    elif (x==2):
        return ('Couples')
    else:
        return ('Singles')

new_data['Family_Cat']=new_data.Family.apply(fam)

new_data
```

```
      Survived  Pclass     Sex   Age  SibSp  Parch      Fare Embarked
Title  \
0          0.0       3    male  22.0      1      0    7.2500        S
Mr
1          1.0       1  female  38.0      1      0   71.2833        C
Mrs
2          1.0       3  female  26.0      0      0    7.9250        S
```

```
Miss
3          1.0       1  female  35.0      1      0   53.1000        S
Mrs
4          0.0       3    male  35.0      0      0    8.0500        S
Mr
...        ...     ...     ...   ...    ...    ...       ...      ...
...
1304       NaN       3    male  29.0      0      0    8.0500        S
Mr
1305       NaN       1  female  39.0      0      0  108.9000        C
Dona
1306       NaN       3    male  38.5      0      0    7.2500        S
Mr
1307       NaN       3    male  29.0      0      0    8.0500        S
Mr
1308       NaN       3    male   4.0      1      1   22.3583        C
Master

      Titles    Cabin_Avalability  Family    Family_Cat
0         Mr  Cabin Not_Available       2       Couples
1        Mrs      Cabin Available       2       Couples
2       Miss  Cabin Not_Available       1       Singles
3        Mrs      Cabin Available       2       Couples
4         Mr  Cabin Not_Available       1       Singles
...      ...                  ...     ...           ...
1304      Mr  Cabin Not_Available       1       Singles
1305  Others      Cabin Available       1       Singles
1306      Mr  Cabin Not_Available       1       Singles
1307      Mr  Cabin Not_Available       1       Singles
1308  Master  Cabin Not_Available       3  Small_Family

[1309 rows x 13 columns]
```

```
pd.crosstab(new_data.Family_Cat,new_data.Survived).plot(kind='bar')
# Couples,Small_Family passengers are had high chance of survival
```

```
<AxesSubplot:xlabel='Family_Cat'>
```




```
new_data.head()
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Title
0	0.0	3	male	22.0	1	0	7.2500	S	Mr
1	1.0	1	female	38.0	1	0	71.2833	C	Mrs
2	1.0	3	female	26.0	0	0	7.9250	S	Miss
3	1.0	1	female	35.0	1	0	53.1000	S	Mrs
4	0.0	3	male	35.0	0	0	8.0500	S	Mr

	Titles	Cabin_Avalability	Family	Family_Cat
0	Mr	Cabin Not_Available	2	Couples
1	Mrs	Cabin Available	2	Couples
2	Miss	Cabin Not_Available	1	Singles
3	Mrs	Cabin Available	2	Couples
4	Mr	Cabin Not_Available	1	Singles

```
# Fare Per Person
new_data['Fare_Per_Head']=new_data.Fare/new_data.Family

new_data.corr()
```

```
              Survived    Pclass       Age     SibSp     Parch
Fare  \
Survived       1.000000 -0.338481 -0.071235 -0.035322  0.081629
0.257307
Pclass        -0.338481  1.000000 -0.391580  0.060832  0.018322 -
0.558742
Age           -0.071235 -0.391580  1.000000 -0.214428 -0.129649
0.179632
SibSp         -0.035322  0.060832 -0.214428  1.000000  0.373587
0.160389
Parch          0.081629  0.018322 -0.129649  0.373587  1.000000
0.221668
Fare           0.257307 -0.558742  0.179632  0.160389  0.221668
1.000000
Family         0.016639  0.050027 -0.211904  0.861952  0.792296
0.226654
Fare_Per_Head  0.221600 -0.504336  0.193300 -0.089666 -0.065370
0.832045

                Family  Fare_Per_Head
Survived       0.016639       0.221600
Pclass         0.050027      -0.504336
Age           -0.211904       0.193300
SibSp          0.861952      -0.089666
Parch          0.792296      -0.065370
Fare           0.226654       0.832045
Family         1.000000      -0.094708
Fare_Per_Head -0.094708       1.000000
```

```
new_data[new_data.Fare==0]
```

```
      Survived  Pclass   Sex   Age  SibSp  Parch  Fare Embarked
Title  \
179        0.0       3  male  36.0      0      0   0.0        S
Mr
263        0.0       1  male  40.0      0      0   0.0        S
Mr
271        1.0       3  male  25.0      0      0   0.0        S
Mr
277        0.0       2  male  29.0      0      0   0.0        S
Mr
302        0.0       3  male  19.0      0      0   0.0        S
Mr
413        0.0       2  male  29.0      0      0   0.0        S
Mr
466        0.0       2  male  29.0      0      0   0.0        S
```

```
Mr
481        0.0       2  male  29.0      0      0   0.0        S
Mr
597        0.0       3  male  49.0      0      0   0.0        S
Mr
633        0.0       1  male  29.0      0      0   0.0        S
Mr
674        0.0       2  male  29.0      0      0   0.0        S
Mr
732        0.0       2  male  29.0      0      0   0.0        S
Mr
806        0.0       1  male  39.0      0      0   0.0        S
Mr
815        0.0       1  male  29.0      0      0   0.0        S
Mr
822        0.0       1  male  38.0      0      0   0.0        S
Jonkheer
1157       NaN       1  male  29.0      0      0   0.0        S
Mr
1263       NaN       1  male  49.0      0      0   0.0        S
Mr

      Titles    Cabin_Avalability  Family Family_Cat  Fare_Per_Head
179       Mr  Cabin Not_Available       1    Singles            0.0
263       Mr      Cabin Available       1    Singles            0.0
271       Mr  Cabin Not_Available       1    Singles            0.0
277       Mr  Cabin Not_Available       1    Singles            0.0
302       Mr  Cabin Not_Available       1    Singles            0.0
413       Mr  Cabin Not_Available       1    Singles            0.0
466       Mr  Cabin Not_Available       1    Singles            0.0
481       Mr  Cabin Not_Available       1    Singles            0.0
597       Mr  Cabin Not_Available       1    Singles            0.0
633       Mr  Cabin Not_Available       1    Singles            0.0
674       Mr  Cabin Not_Available       1    Singles            0.0
732       Mr  Cabin Not_Available       1    Singles            0.0
806       Mr      Cabin Available       1    Singles            0.0
815       Mr      Cabin Available       1    Singles            0.0
822   Others  Cabin Not_Available       1    Singles            0.0
1157      Mr  Cabin Not_Available       1    Singles            0.0
1263      Mr      Cabin Available       1    Singles            0.0

new_data

      Survived  Pclass     Sex   Age  SibSp  Parch      Fare Embarked
Title  \
0          0.0       3    male  22.0      1      0    7.2500        S
Mr
1          1.0       1  female  38.0      1      0   71.2833        C
Mrs
2          1.0       3  female  26.0      0      0    7.9250        S
```

```
Miss
3          1.0       1  female  35.0      1      0   53.1000        S
Mrs
4          0.0       3    male  35.0      0      0    8.0500        S
Mr
...        ...     ...     ...   ...    ...    ...       ...      ...
...
1304       NaN       3    male  29.0      0      0    8.0500        S
Mr
1305       NaN       1  female  39.0      0      0  108.9000        C
Dona
1306       NaN       3    male  38.5      0      0    7.2500        S
Mr
1307       NaN       3    male  29.0      0      0    8.0500        S
Mr
1308       NaN       3    male   4.0      1      1   22.3583        C
Master
```

	Titles	Cabin_Avalability	Family	Family_Cat	Fare_Per_Head
0	Mr	Cabin Not_Available	2	Couples	3.625000
1	Mrs	Cabin Available	2	Couples	35.641650
2	Miss	Cabin Not_Available	1	Singles	7.925000
3	Mrs	Cabin Available	2	Couples	26.550000
4	Mr	Cabin Not_Available	1	Singles	8.050000
...	...	...	...	...	...
1304	Mr	Cabin Not_Available	1	Singles	8.050000
1305	Others	Cabin Available	1	Singles	108.900000
1306	Mr	Cabin Not_Available	1	Singles	7.250000
1307	Mr	Cabin Not_Available	1	Singles	8.050000
1308	Master	Cabin Not_Available	3	Small_Family	7.452767

[1309 rows x 14 columns]

```
new_data.isnull().sum()
```

```
Survived             418
Pclass                 0
```

```
Sex                    0
Age                    0
SibSp                  0
Parch                  0
Fare                   0
Embarked               2
Title                  0
Titles                 0
Cabin_Avalability      0
Family                 0
Family_Cat             0
Fare_Per_Head          0
dtype: int64
```

```
new_data.loc[(new_data.Pclass==3)&(new_data.Titles=='Mr')&(new_data.Ca
bin_Avalability=='Cabin Not_Available')&
       (new_data.Family_Cat=='Singles'),'Fare'].median()
```

```
7.8958
```

```
new_data.loc[new_data.Fare.isnull(),'Fare']=new_data.loc[(new_data.Pcl
ass==3)&(new_data.Titles=='Mr')&(new_data.Cabin_Avalability=='Cabin
Not_Available')&
       (new_data.Family_Cat=='Singles'),'Fare'].median()
```

```
new_data.isnull().sum()
```

```
Survived             418
Pclass                 0
Sex                    0
Age                    0
SibSp                  0
Parch                  0
Fare                   0
Embarked               2
Title                  0
Titles                 0
Cabin_Avalability      0
Family                 0
Family_Cat             0
Fare_Per_Head          0
dtype: int64
```

```
new_data[new_data.Embarked.isnull()]
```

```
     Survived  Pclass     Sex   Age  SibSp  Parch  Fare Embarked Title
Titles  \
61        1.0       1  female  38.0      0      0  80.0      NaN  Miss
Miss
829       1.0       1  female  62.0      0      0  80.0      NaN   Mrs
Mrs
```

```
    Cabin_Avalability  Family Family_Cat  Fare_Per_Head
61    Cabin Available       1    Singles           80.0
829   Cabin Available       1    Singles           80.0
```

```
new_data.loc[

(new_data.Sex=='female')&(new_data.Family_Cat=='Singles')&(new_data.Pc
lass==1),'Embarked'].mode()[0]
```

```
'C'
```

```
new_data.loc[new_data.Embarked.isnull(),'Embarked']='C'
```

```
new_data.Survived.value_counts()
```

```
0.0    549
1.0    342
Name: Survived, dtype: int64
```

```
new_data[new_data.Embarked.isnull()]
```

```
Empty DataFrame
Columns: [Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked,
Title, Titles, Cabin_Avalability, Family, Family_Cat, Fare_Per_Head]
Index: []
```

```
new_data.groupby(['Sex','Embarked','Titles','Cabin_Avalability'])
['Pclass'].transform('count')
```

```
0       480
1        32
2       123
3        43
4       480
       ...
1304    480
1305      5
1306    480
1307    480
1308      9
Name: Pclass, Length: 1309, dtype: int64
```

```
new_data['Magic_1']=new_data.groupby(['Sex','Embarked','Titles','Cabin
_Avalability'])['Pclass'].transform('count')
```

```
new_data['Magic_2']=new_data.groupby(['Pclass','Embarked','Titles','Ca
bin_Avalability','Family_Cat'])['Fare'].transform('median')
```

```
new_data.head()
```

```
   Survived  Pclass     Sex   Age  SibSp  Parch     Fare Embarked
Title  \
0       0.0       3    male  22.0      1      0   7.2500        S
```

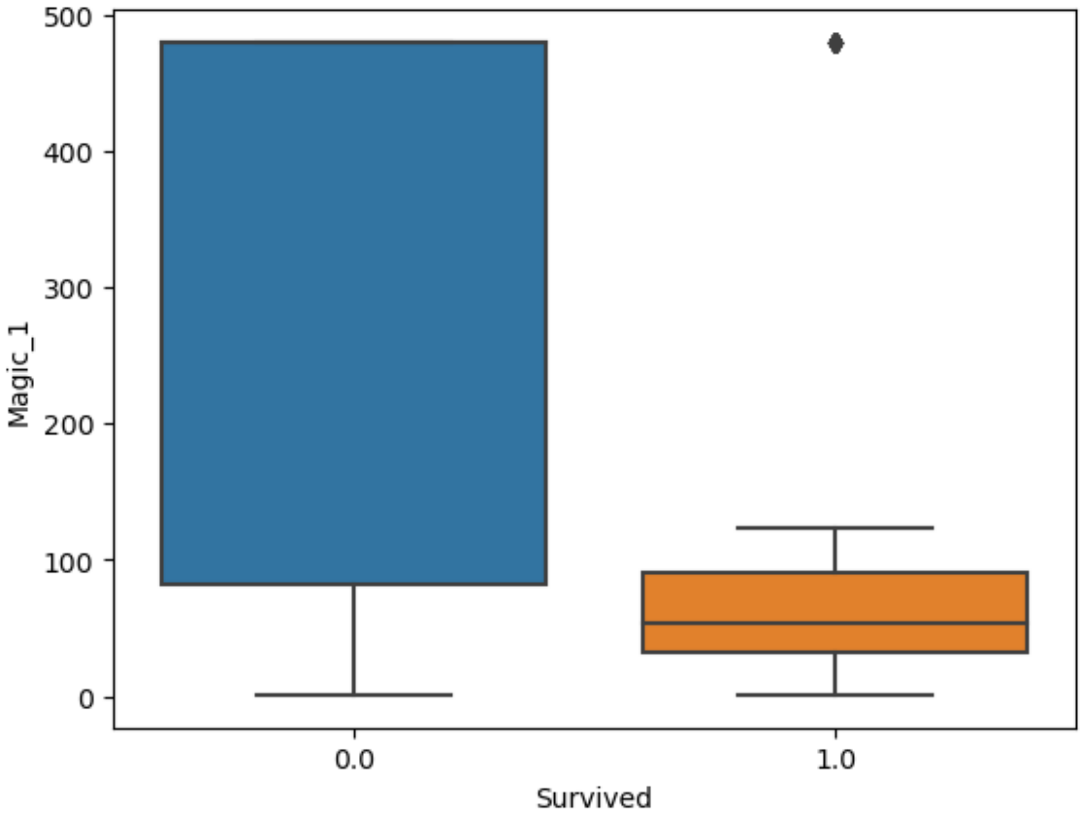
```
Mr
1       1.0       1  female  38.0      1      0  71.2833        C
Mrs
2       1.0       3  female  26.0      0      0   7.9250        S
Miss
3       1.0       1  female  35.0      1      0  53.1000        S
Mrs
4       0.0       3    male  35.0      0      0   8.0500        S
Mr

  Titles    Cabin_Avalability  Family Family_Cat  Fare_Per_Head
Magic_1  \
0     Mr  Cabin Not_Available       2    Couples        3.62500
480
1    Mrs      Cabin Available       2    Couples       35.64165
32
2   Miss  Cabin Not_Available       1    Singles        7.92500
123
3    Mrs      Cabin Available       2    Couples       26.55000
43
4     Mr  Cabin Not_Available       1    Singles        8.05000
480

   Magic_2
0  15.0250
1  83.1583
2   7.9250
3  60.0000
4   7.8958

new_data.corr()

              Survived    Pclass       Age     SibSp     Parch
Fare  \
Survived       1.000000 -0.338481 -0.071235 -0.035322  0.081629
0.257307
Pclass        -0.338481  1.000000 -0.391580  0.060832  0.018322 -
0.558742
Age           -0.071235 -0.391580  1.000000 -0.214428 -0.129649
0.179632
SibSp         -0.035322  0.060832 -0.214428  1.000000  0.373587
0.160389
Parch          0.081629  0.018322 -0.129649  0.373587  1.000000
0.221668
Fare           0.257307 -0.558742  0.179632  0.160389  0.221668
1.000000
Family         0.016639  0.050027 -0.211904  0.861952  0.792296
0.226654
Fare_Per_Head  0.221600 -0.504336  0.193300 -0.089666 -0.065370
0.832045
```

```
Magic_1       -0.456498  0.336720  0.016422 -0.149930 -0.204599 -
0.296807
Magic_2        0.318082 -0.646534  0.192402  0.209397  0.288079
0.763642
```

	Family	Fare_Per_Head	Magic_1	Magic_2
Survived	0.016639	0.221600	-0.456498	0.318082
Pclass	0.050027	-0.504336	0.336720	-0.646534
Age	-0.211904	0.193300	0.016422	0.192402
SibSp	0.861952	-0.089666	-0.149930	0.209397
Parch	0.792296	-0.065370	-0.204599	0.288079
Fare	0.226654	0.832045	-0.296807	0.763642
Family	1.000000	-0.094708	-0.210445	0.295187
Fare_Per_Head	-0.094708	1.000000	-0.213136	0.492650
Magic_1	-0.210445	-0.213136	1.000000	-0.352488
Magic_2	0.295187	0.492650	-0.352488	1.000000

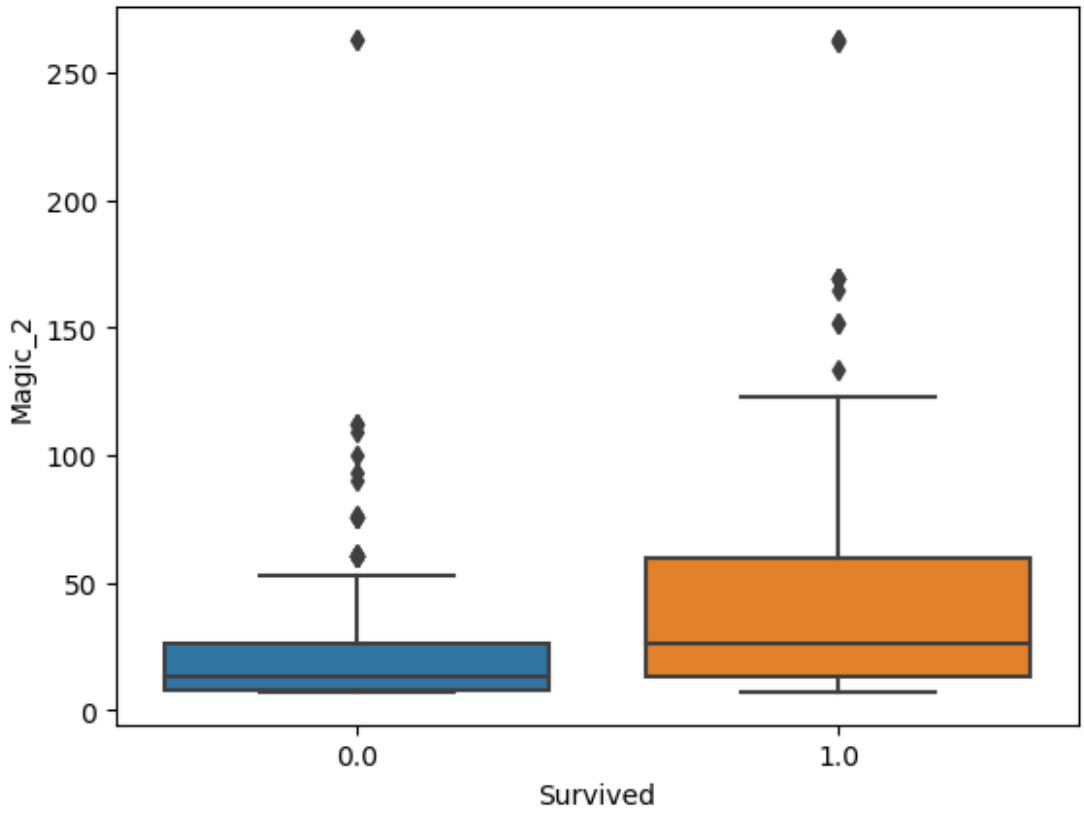
```
sns.boxplot(new_data.Survived,new_data.Magic_1) # Hidden features
```

```
<AxesSubplot:xlabel='Survived', ylabel='Magic_1'>
```




```
sns.boxplot(new_data.Survived,new_data.Magic_2)
```

```
<AxesSubplot:xlabel='Survived', ylabel='Magic_2'>
```




```
new_data.std()
```

```
Survived           0.486592
Pclass             0.837836
Age               13.159685
SibSp              1.041658
Parch              0.865560
Fare              51.743631
Family             1.583639
Fare_Per_Head     35.762353
Magic_1          202.028178
Magic_2           35.744779
dtype: float64
```

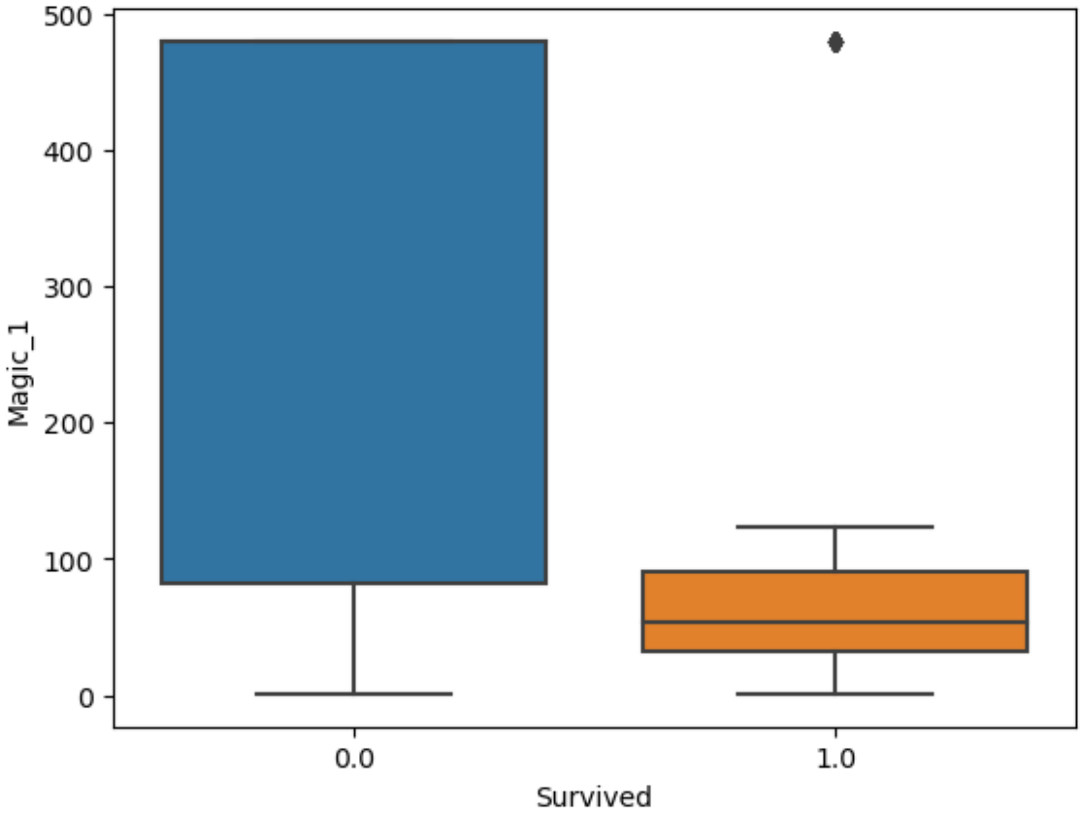
```
new_data.Magic_1.value_counts() # doubt
```

```
480    480
123    123
90      90
87      87
82      82
32      64
55      55
53      53
52      52
43      43
```

```
38      38
28      28
26      26
24      24
5       15
12      12
9        9
2        8
7        7
1        6
4        4
3        3
Name: Magic_1, dtype: int64
```

```
sns.boxplot(y=new_data.Magic_1,x=new_data.Survived)
```

```
<AxesSubplot:xlabel='Survived', ylabel='Magic_1'>
```




```
new_data.shape
```

```
(1309, 16)
```

```
#Spilt train,test data set
train.shape,test.shape
```

```
((891, 12), (418, 11))
```

```
new_data.loc[0:train.shape[0]-1].shape
```

```
(891, 16)
```

```
newtrain=new_data.loc[0:train.shape[0]-1,:]
```

```
new_test.drop('Title',axis=1,inplace=True)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_21780\3578656877.py in <module>
----> 1 new_test.drop('Title',axis=1,inplace=True)

NameError: name 'new_test' is not defined
```

```
newtrain.drop('Title',axis=1,inplace=True)
```

```
newtrain.head()
```

```
new_data.loc[train.shape[0]:,:]
```

```
new_test=new_data.loc[train.shape[0]:,:]
```

```
new_test
```

```
newtrain.shape,new_test.shape
```

```
new_test.drop(columns='Survived',inplace=True)
```

```
new_test
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_21780\2344501716.py in <module>
----> 1 new_test

NameError: name 'new_test' is not defined
```

```
newtrain.shape,new_test.shape
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_21780\4090994966.py in <module>
----> 1 newtrain.shape,new_test.shape

NameError: name 'new_test' is not defined
```

Scaling

```
newtrain.head()

# Convert the Tgt into Int
newtrain['Survived']=newtrain.Survived.astype('int')

newtrain.head(2)

# Scale the Age,Fare and Fare_per_Head
cols=['Age','Fare','Fare_Per_Head']
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
for i in cols:
    newtrain.loc[:,cols]=sc.fit_transform(newtrain.loc[:,cols])
    new_test.loc[:,cols]=sc.transform(new_test.loc[:,cols])

sns.distplot(combine.Fare)

sns.distplot(sc.fit_transform(pd.DataFrame(new_test['Fare'])))

combine.describe()

newtrain.describe()

new_test.describe()

dummytrain=pd.get_dummies(newtrain,drop_first=True)
dummytest=pd.get_dummies(new_test,drop_first=True)

dummytrain.shape,dummytest.shape

newtrain.head()

new_test.head()

dummytrain
```

Random Forest Model

```
from sklearn.ensemble import RandomForestClassifier
X=dummytrain.drop(columns='Survived')
y=dummytrain.Survived

rf=RandomForestClassifier(criterion='entropy')

pred=rf.fit(X, y).predict(dummytest)

# Submission File
soln=pd.DataFrame({"PassengerId":test.PassengerId,
                   "Survived":pred})

# Export to outside
soln.to_csv('Titanic_submission_File',index=False)
```

```
cd

tbl = pd.crosstab(new_data.Cabin_Avalability, new_data.Survived)
# Family Cat has any relation with the Survival...

# Ho: Family Cat has no Relation with the Survival
# Ha: Family Cat is Related with the Target Variable

#Note: Since bogth the var are categorical in nature therefore, we
will apply Chi Square Test

import scipy.stats as stats
teststats, pvalue, df, exp_freq = stats.chi2_contingency(tbl)
pvalue # Since the Pvalue is < 0.05, We Reject the Ho meaning
Family_Cat is related to tgt.



newtrain

     Survived  Pclass     Sex   Age  SibSp  Parch     Fare Embarked
Title  \
0         0.0       3    male  22.0      1      0   7.2500        S
Mr
1         1.0       1  female  38.0      1      0  71.2833        C
Mrs
2         1.0       3  female  26.0      0      0   7.9250        S
Miss
3         1.0       1  female  35.0      1      0  53.1000        S
Mrs
4         0.0       3    male  35.0      0      0   8.0500        S
Mr
..        ...     ...     ...   ...    ...    ...      ...      ...
...
886       0.0       2    male  27.0      0      0  13.0000        S
Rev
887       1.0       1  female  19.0      0      0  30.0000        S
Miss
888       0.0       3  female  22.0      1      2  23.4500        S
Miss
889       1.0       1    male  26.0      0      0  30.0000        C
Mr
890       0.0       3    male  32.0      0      0   7.7500        Q
Mr

     Titles    Cabin_Avalability  Family    Family_Cat  Fare_Per_Head
\
0        Mr  Cabin Not_Available       2       Couples        3.62500

1       Mrs      Cabin Available       2       Couples       35.64165
```

```
2      Miss  Cabin Not_Available       1       Singles        7.92500

3       Mrs      Cabin Available       2       Couples       26.55000

4        Mr  Cabin Not_Available       1       Singles        8.05000

..      ...                  ...     ...           ...            ...

886  Others  Cabin Not_Available       1       Singles       13.00000

887    Miss      Cabin Available       1       Singles       30.00000

888    Miss  Cabin Not_Available       4  Small_Family        5.86250

889      Mr      Cabin Available       1       Singles       30.00000

890      Mr  Cabin Not_Available       1       Singles        7.75000


     Magic_1   Magic_2
0        480  15.02500
1         32  83.15830
2        123   7.92500
3         43  60.00000
4        480   7.89580
..       ...       ...
886       12  13.00000
887       32  93.50000
888      123  19.10625
889       52  32.82710
890       55   7.75000

[891 rows x 16 columns]
```

HYPOTHESIS TESTING

```
import scipy.stats as stats

cat_cols=['Pclass','Sex','Embarked','Titles','Cabin_Avalability','Fami
ly_Cat']

for i in cat_cols:
    tbl=pd.crosstab(newtrain.loc[:,i],newtrain.Survived)
    teststats,pvalue,df,exp=stats.chi2_contingency(tbl)
    print('Pvalue For',i,'is',pvalue)

    print('TestStatistic',i,'is',teststats)
    print()
```

```
Pvalue For Pclass is 4.549251711298793e-23
TestStatistic Pclass is 102.88898875696056

Pvalue For Sex is 1.1973570627755645e-58
TestStatistic Sex is 260.71702016732104

Pvalue For Embarked is 8.294156968447598e-07
TestStatistic Embarked is 28.005088727541892

Pvalue For Titles is 1.9783487591671835e-59
TestStatistic Titles is 284.6667057498796

Pvalue For Cabin_Avalability is 6.7419704360811776e-21
TestStatistic Cabin_Avalability is 87.94148561238097

Pvalue For Family_Cat is 2.747307908074899e-16
TestStatistic Family_Cat is 75.56079716081948


tbl

Survived      0.0  1.0
Family_Cat
Couples        72   89
Large_Family   52   10
Singles       374  163
Small_Family   51   80

# Fare is different for the people who died vs survived

# Ho: Fare has no relation with survived
# Ha: Fare has relation with survived
zero_fare=newtrain.loc[newtrain.Survived==0,'Fare']
one_fare=newtrain.loc[newtrain.Survived==1,'Fare']
stats.shapiro(zero_fare)
stats.shapiro(one_fare)

ShapiroResult(statistic=0.5967273712158203,
pvalue=1.8337799743381398e-27)

stats.mannwhitneyu(zero_fare,one_fare)

MannwhitneyuResult(statistic=57806.5, pvalue=4.553477179250237e-22)

# Age
zero_age=newtrain.loc[newtrain.Survived==0,'Age']
one_age=newtrain.loc[newtrain.Survived==1,'Age']

stats.shapiro(zero_age),stats.shapiro(one_age)
```

```
(ShapiroResult(statistic=0.9440184831619263, pvalue=1.55326245572171e-
13),
 ShapiroResult(statistic=0.9793311953544617,
pvalue=7.951998122734949e-05))

stats.mannwhitneyu(zero_age,one_age)

MannwhitneyuResult(statistic=99183.5, pvalue=0.15476166755486367)
```