Titanic DataSet Analysis Using Numpy and Pandas

NAME - SYED MIRAN, YASMIN SANIYA

CLASS - CSD 2ND YEAR

ROLL NUMBER - 160921750038, 160921750046

ASSIGNMENT - 3

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

Loading Data Set

In [ ]:
```
df = pd.read_csv('/content/train.csv')
df.head()
```

Out[ ]:

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

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

In [ ]:
```
df.shape
```

Out[ ]:
```
(891, 12)
```

In [ ]:
```
df.columns
```

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

Lets remove unnecessary columns which we are'nt gonna use

In [ ]:
```
df = df.drop(['Ticket', 'Fare', 'Cabin'], axis =1)
df.head()
```

Out[ ]:

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

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 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   Embarked     889 non-null    object
dtypes: float64(1), int64(5), object(3)
memory usage: 62.8+ KB
```

In [ ]:
```
df2 = df[["Age"]].mean()
print(df2)
```

```
Age    29.699118
dtype: float64
```

Average age of passengers boarding titanic was 29

In [ ]:
```
df.Age.isnull().sum()
```

Out[ ]:
```
177
```

In [ ]:
```
df['Age'].fillna(df['Age'].mean(), inplace = True)
```

In [ ]:
```
df['Age'].isnull().sum()
```

Out[ ]:
```
0
```

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 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          891 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Embarked     889 non-null    object
dtypes: float64(1), int64(5), object(3)
memory usage: 62.8+ KB
```

Finally we can conclude that this raw data has been cleaned for futher analysis

In [ ]:
```
df.head()
```

Out[ ]:

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

In [ ]:
```
df['Survived'].value_counts()
```

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

Almost 342 passengers on the board survived

In [ ]:
```
df['Pclass'].value_counts()
```

Out[ ]:
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

In [ ]:
```
af = pd.DataFrame(df['Pclass'].value_counts())
af
```

Out[ ]:

	Pclass
3	491
1	216
2	184

Here, 216 Passengers were on board in First class, followed by 184 passengers in second class and 491 passengers in third class.

In [ ]:
```
df['Sex'].value_counts()
```

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

In [ ]:
```
sf = pd.DataFrame(df['Sex'].value_counts())
sf.head()
```

Out[ ]:

	Sex
male	577
female	314

In [ ]:
```
df_males_death = 577
df_females_death = 314
```

Here we can conclude that there were more males than females on the board

In [ ]:
```
pd.pivot_table(df, index = 'Survived', values = ['Age','Sex'])
```

Out[ ]:

	Age
Survived	
0	30.415100
1	28.549778

In [ ]:
```
df_cat = df[['Survived','Pclass','Sex']]
df_cat
```

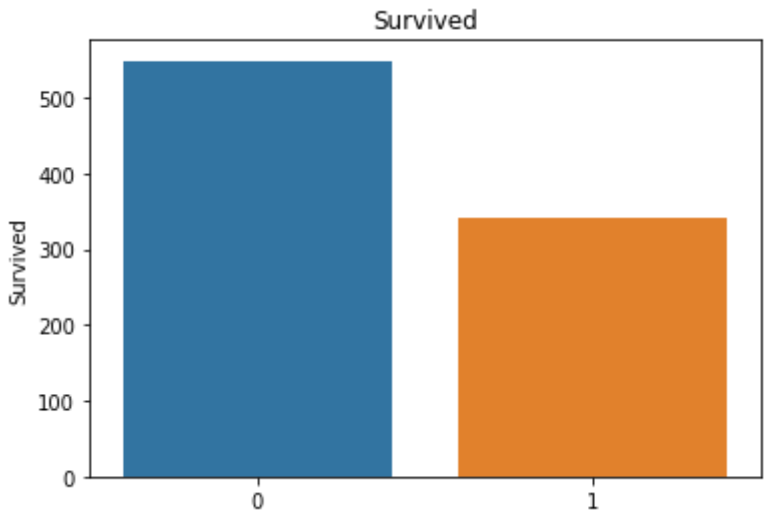
Out[ ]:

	Survived	Pclass	Sex
0	0	3	male
1	1	1	female
2	1	3	female
3	1	1	female
4	0	3	male
...	...	...	...
886	0	2	male
887	1	1	female
888	0	3	female
889	1	1	male
890	0	3	male

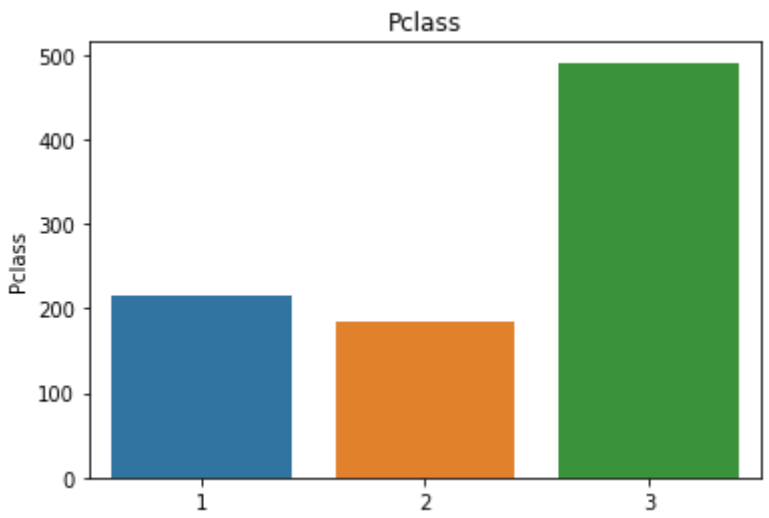
891 rows × 3 columns

In [ ]:
```
for i in df_cat.columns:
    sns.barplot(df_cat[i].value_counts().index,df_cat[i].value_counts()).set_title(i)
    plt.show()
```

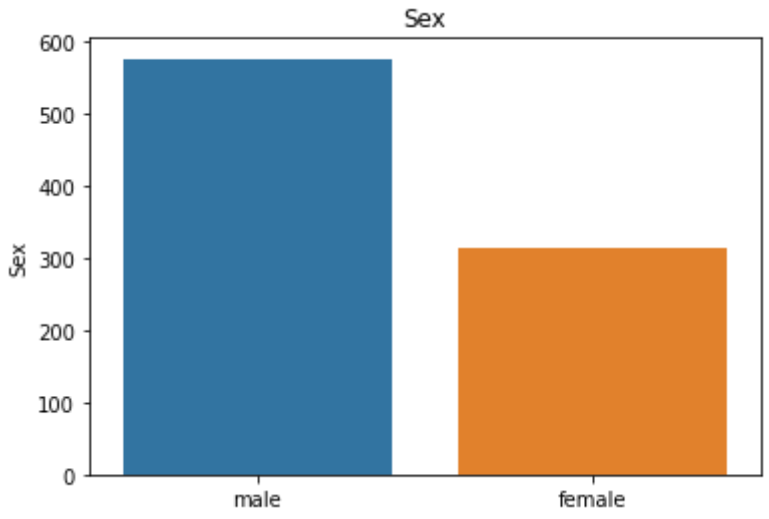
```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyw
ord args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```




```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyw
ord args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```




```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyw
ord args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



Survived: Most of the people died in the shipwreck, only around 300 people survived. Pclass: The majority of the people traveling, had tickets to the 3rd class. Sex: There were more males than females aboard the ship, roughly double the amount

In [ ]:
```
print(pd.pivot_table(df, index = 'Survived', columns = 'Sex',
                     aggfunc ='count'))
print()
```

```
            Age      Embarked        Name       Parch      PassengerId       \
Sex      female male   female male female male female male      female male
Survived
0            81  468       81  468     81  468     81  468          81  468
1           233  109      231  109    233  109    233  109         233  109

          Pclass       SibSp
Sex      female male female male
Survived
0            81  468     81  468
1           233  109    233  109
```

Sex: Most of the women survived, and the majority of the male died in the shipwreck. So it looks like the saying “Woman and children first” actually applied in this scenario

In [ ]:
```
df.head()
```

Out[ ]:

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

In [ ]:
```
df['Embarked'].value_counts()
```

Out[ ]:
```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

In [ ]:
```
emb = pd.DataFrame(df['Embarked'].value_counts())
emb
```

Out[ ]:

	Embarked
S	644
C	168
Q	77

In [ ]:
```
qmb = 891
```

In [ ]:
```
dict_ = {'Southampton' : 644, 'Cherbourg' : 168, 'Queenstown' : 77}
Place = list(dict_.keys())
No_of_pass = list(dict_.values())
```

In [ ]:
```
plt.bar(Place, No_of_pass, color = 'purple' )
plt.title('Embarkment of Passengers')
plt.xlabel('Passengers of Places')
plt.ylabel('Number of Passengers')
plt.grid(True)
plt.show()
```

As we can see majority of the passengers mounted from Southampton

THE END

In [ ]:
```
!jupyter nbconvert --to html /content/titanic__.ipynb
```

```
[NbConvertApp] Converting notebook /content/titanic__.ipynb to html
[NbConvertApp] Writing 394130 bytes to /content/titanic__.html
```