Titanic

Analysing the Titanic.Csv starts with questions.

Questions:

- What factors made people more likely to survive?
- Are there any missing data?
- Are there more survivors than deceased?.
- Which age group had most passengers?
- which age group had highest survivors?
- Which Passengerclass had most passengers?
- which Passengerclass had most survivors?
- Which Embarking station boarded more passengers?
- Were all the children travelling with a relative ?.
- Which gender survived the most?

There are various factors to consider, to analyze this data . Analysing data through various factors like Age, Sex ,Passenger class and Embarkment station helps in finding conclusions.Relatives taken into account are Parent,children,siblings and spouse(as per dataset)

In [97]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline



Filename ='/users/ambilurama/Downloads/Nano/data analysis/survival - Sheet1.csv'

Raw_survival_df = pd.read_csv(Filename ,index_col='PassengerId' )
```

To begin with all necessary libraries are imported . The given csv file is read and copied to Raw_survival_df. Each PassengerId is unique and hence used as index.

Question: Are there any missing data?

In [98]:

```
Raw_survival_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 891 entries, 1 to 891
Data columns (total 11 columns):
Survived    891 non-null int64
Pclass      891 non-null int64
Name        891 non-null object
Sex         891 non-null object
Age         714 non-null float64
SibSp       891 non-null int64
Parch       891 non-null int64
Ticket      891 non-null object
Fare        891 non-null float64
Cabin       204 non-null object
Embarked    889 non-null object
dtypes: float64(2), int64(4), object(5)
memory usage: 83.5+ KB
```

on analysing the data through all the variables ,it is found that two columns have missing datas.

Age contains 714 given values and lacks 177 values .Age is one of our independent variable so 'Nan' gets filled in for missing values and analysis is continued with known values.

Cabin contains 204 given values and lacks 685 values.Cabin has more missing values and is not our independent variable.

In [75]:

```
Raw_survival_df.groupby('Survived').size()
```

Out[75]:

```
Survived
0    549
1    342
dtype: int64
```

In the above output the survivors are read from index 1 and deceased are read from the index 0 in the Survived column. It is convenient to use inferred data at some places. A new column named Survival was created to give a proper reference for the survivors and the deceased,using the map()

In [76]:

```
Raw_survival_df['Survival'] = Raw_survival_df.Survived.map({0 : 'Deceased', 1 : 'Su
```

A method was created to give the ages in group with a interval of 10. A new column named Age_category was created to give the age in group values ,using apply().It helps in analysing the data in various perspectives.

In [77]:

```
def age_categorize(age):

    if age>=90:
        return '90+'
    elif age>=80:
        return '80-90'
    elif age>=70:
        return '70-80'
    elif age>=60:
        return '60-70'
    elif age>=50:
        return '50-60'
    elif age>=40:
        return '40-50'
    elif age>=30:
        return '30-40'
    elif age>=20:
        return '20-30'
    elif age>=10:
        return '10-20'
    elif age<10:
        return '0-10'
```

In [78]:

```
age_categorize(48)
```

Out[78]:

```
'40-50'
```

In [79]:

```
Raw_survival_df['Age_category'] = Raw_survival_df.Age.apply(age_categorize)
Raw_survival_df.head()
```

Out[79]:

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

Question: Are there more survivors than deceased?

In [80]:

```
survival_numbers = pd.Series(Raw_survival_df.groupby('Survival').size() ,name ='Surv
print survival_numbers
```

```
Survival
Deceased    549
Survived    342
Name: Survival, dtype: int64
```

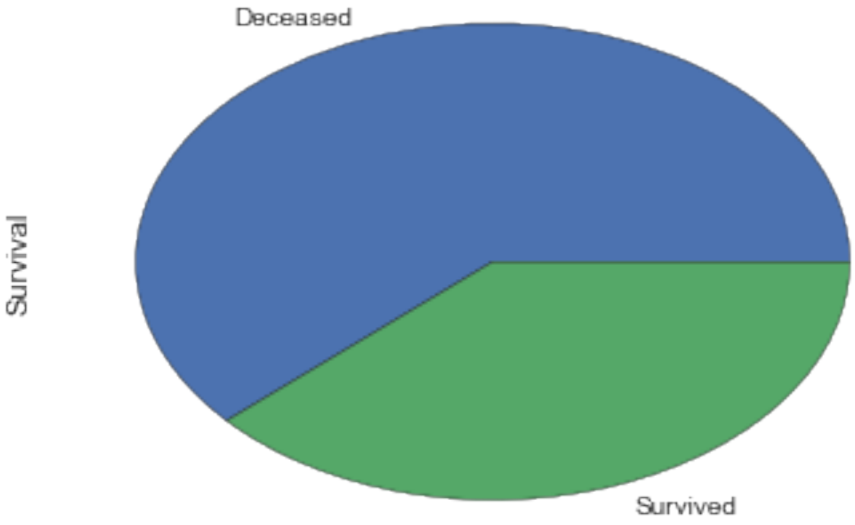
The resulting data shows that the number of survivors are less than the number of deceased.

```
In [81]:
```

```
survival_numbers.plot.pie()
```

```
Out[81]:
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x11d9f1d10>
```



The data is visualized using a pie plot. The visualized data depicts the fact that deceased people are more than the survivors.

Question: Which gender survived the most?

one dimensional exploration:

```
In [82]:
```

```
Raw_survival_df.groupby('Sex').size()
```

```
Out[82]:
```

```
Sex
female    314
male      577
dtype: int64
```

Two dimensional exploration:

In [83]:

```
Survival_by_sex =Raw_survival_df.groupby(['Sex','Survival']).size().unstack()
print Survival_by_sex
```

```
Survival  Deceased  Survived
Sex
female          81       233
male           468       109
```

The resulting datas from 1d and 2d exploration shows that

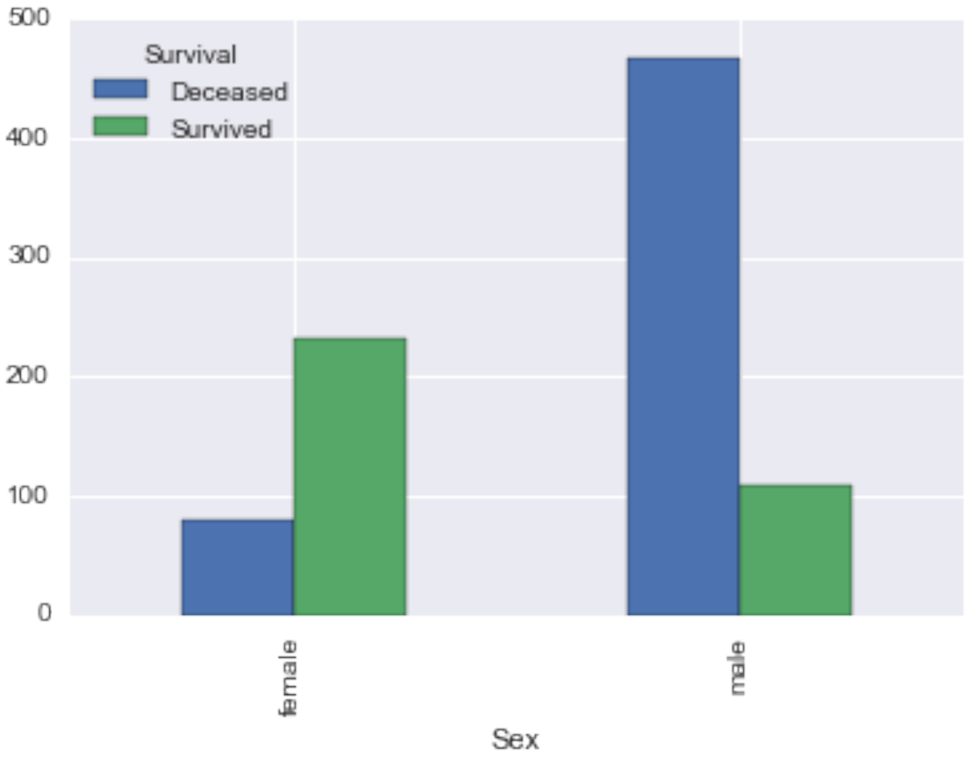
-There were more female survivors than male survivors. -Eventhough there were more male passengers ,male survivors were less than the female survivors.This is a interesting point to consider.It is likely that women were rescued first leaving behind the men.

In [84]:

```
Survival_by_sex.plot.bar()
```

Out[84]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11da0f450>
```



The data visualization is done using a bar plot.It clearly shows that number of female survivors are more than the male survivors.

Questions: Which age group had most passengers? which age group had highest survivors?

one dimensional exploration:

In [85]:

```
Raw_survival_df.groupby('Age_category').size()
```

Out[85]:

```
Age_category
0-10      62
10-20    102
20-30    220
30-40    167
40-50     89
50-60     48
60-70     19
70-80      6
80-90      1
dtype: int64
```

In [ ]:

```

```

Two dimensional exploration:

In [86]:

```
Aboard_Age = Raw_survival_df.groupby(['Age_category','Survival']).size().unstack()
```

In [87]:

```
print Aboard_Age
print "\nsum is...\n",Aboard_Age.sum()
```

```
Survival      Deceased  Survived
Age_category
0-10              24.0      38.0
10-20             61.0      41.0
20-30            143.0      77.0
30-40             94.0      73.0
40-50             55.0      34.0
50-60             28.0      20.0
60-70             13.0       6.0
70-80              6.0       NaN
80-90              NaN       1.0

sum is...
Survival
Deceased    424.0
Survived    290.0
dtype: float64
```

The results from 1d and 2d explorations show that there were more passengers in the age group of 20-30.Also,the highest number of survivors are from the age group of 20-30.

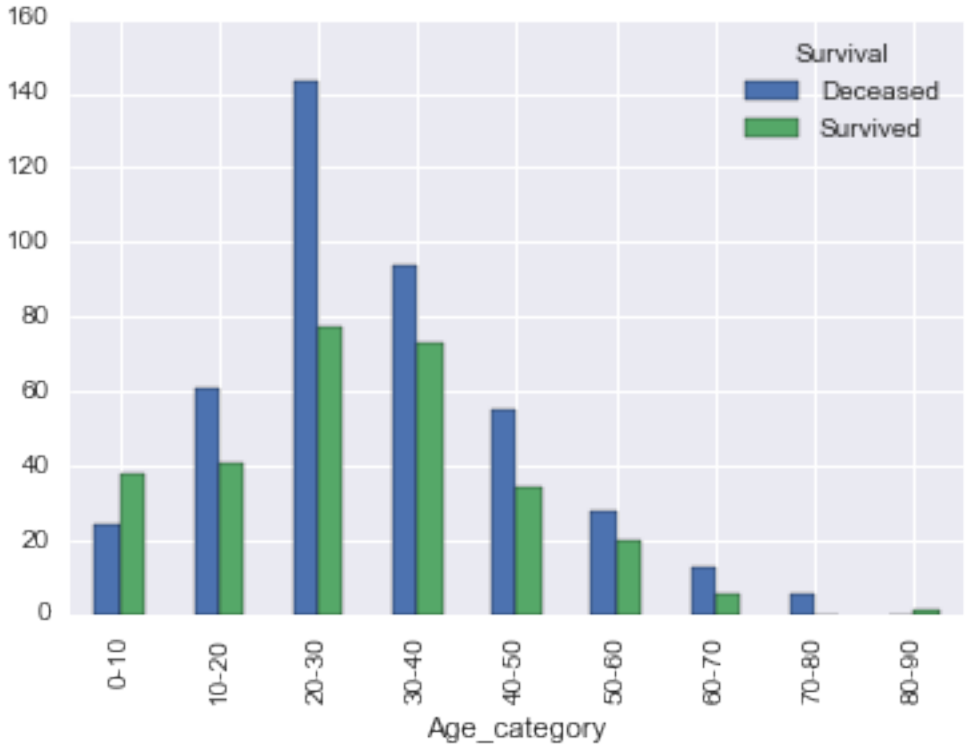
A point to be noted is that the total number of people in all age groups found to be less than the number of people aboard. It is because ,age is unknown for some passengers.

In [88]:

```
Aboard_Age.plot.bar()
```

Out[88]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11de52450>
```



The data is visualized using a bar plot.The highest number of passengers and survivors falls in the age group 20-30.

Question: Were all the children travelling with a relative ?. (Relatives taken into account are Parent,children,siblings and spouse(as per dataset))

In [89]:

```
Age_group_under10 = Raw_survival_df.groupby('Age_category').get_group('0-10')
```

In [90]:

```
Aboard_under10 = pd.DataFrame.from_dict(Age_group_under10)
Aboard_under10.groupby(['SibSp','Parch']).get_group((0,0))
```

Out[90]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
PassengerId										
778	1	3	Emanuel, Miss. Virginia Ethel	female	5.0	0	0	364516	12.475	NaN

The passenger with passengerid 778 was travelling without any of the described relative.

It is likely that this passenger was aboard with someone who is not in the relatives list

Question: Which Passengerclass had most passengers? which Passengerclass had most survivors?

one dimensional exploration:

In [91]:

```
Raw_survival_df.groupby('Pclass').size()
```

Out[91]:

```
Pclass
1    216
2    184
3    491
dtype: int64
```

In [ ]:

Two dimensional exploration:

```
In [92]:
```

```
Aboard_pclass = Raw_survival_df.groupby(['Pclass','Survival']).size().unstack()
print Aboard_pclass
```

```
Survival  Deceased  Survived
Pclass
1               80       136
2               97        87
3              372       119
```

The results from 1d and 2d exploration shows that

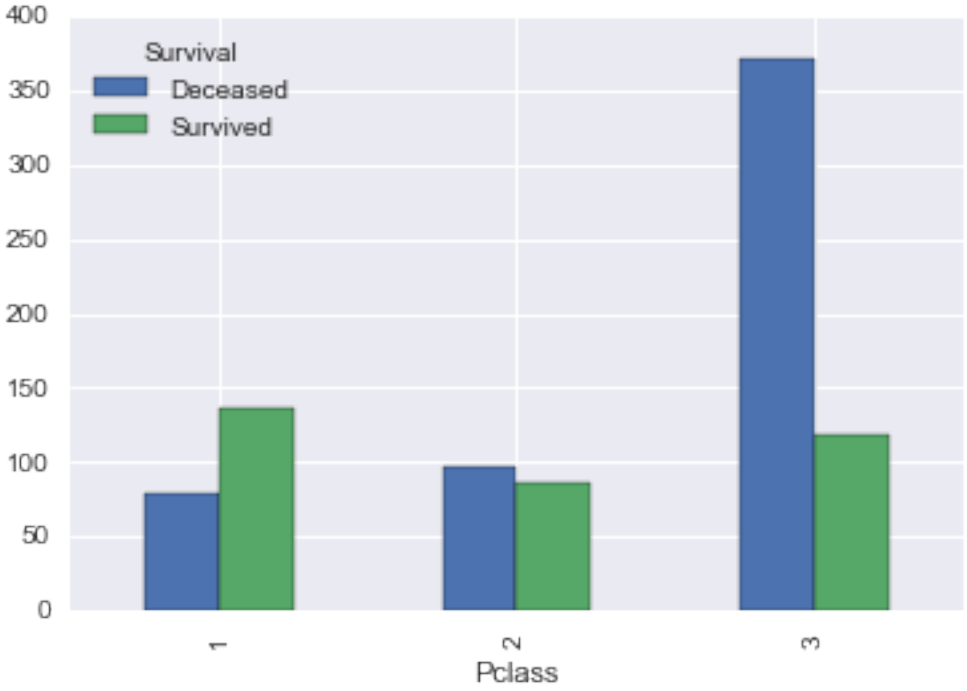
- the highest number of passengers are from Pclass 3.
- the highest number of survivors are from Pclass 1.

```
In [93]:
```

```
Aboard_pclass.plot.bar()
```

```
Out[93]:

<matplotlib.axes._subplots.AxesSubplot at 0x11de74a50>
```



The data is visualized using a bar plot. The highest number of passengers are from Pclass 3 and the highest number of survivors are from Pclass 1.

Question: Which Embarking station boarded more passengers?

one dimensional exploration:

In [94]:

```
Raw_survival_df.groupby('Embarked').size()
```

Out[94]:

```
Embarked
C    168
Q     77
S    644
dtype: int64
```

Two dimeniosnal explortaion:

In [95]:

```
Embarked_station = pd.Series(Raw_survival_df.groupby(['Embarked','Survival']).size(
Embarked_station
```

Out[95]:

Survival	**Deceased**	**Survived**
Embarked		
C	75	93
Q	47	30
S	427	217

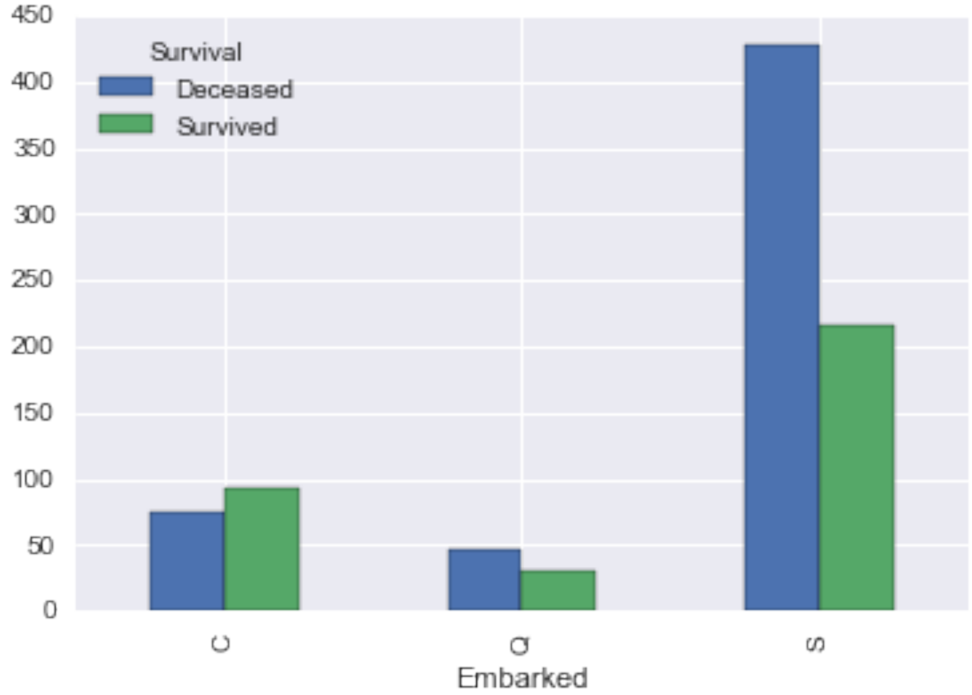
In [96]:

```

Embarked_station.plot.bar()
```

Out[96]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11e156090>
```



The results from 1d and 2d exploration shows that

- Embarking station S(Southampton) had most passengers and
- Embarking station S(Southampton) had more survivors too. The visualization is created using a bar.

Conclusion: The Titanic.csv Contains demographics and passenger information from 891 of the 2224 passengers and crew on board the Titanic.

The given dataset was analysed on factors like age,sex,passenger class and Embarked station. The dependent variable was Survived column. The independent variables were Age,sex,passenger class and Embarked station.

on Analysis with Sex and survived ,the results showed that -There were more male passengers than female. -Male survivors were less than the Female survivors. -It is likely that female were rescued first.

on Analysis with Age and survived ,the results showed that -There were more passengers from the age group of 20-30 -There were more survivors from the age group of 20-30 -people who had better health and stamina might have helped them to survive.

on Analysis with Passengerclass and survived ,the results showed that -The highest number of passengers are from pclass 3 and -The highest number of survivors are from pclass 1 -easy access to rescue team and rescue area likely helped the survivors on Analysis with Embarked and survived ,the results showed that -The embarkation at station S had more passengers and more survivors too. -designated area for station S passengers might had easy access to rescue team.

After analysing the data in all these factors ,one can conclude that the survival happened only based on correlation and not on causation.

In [ ]: