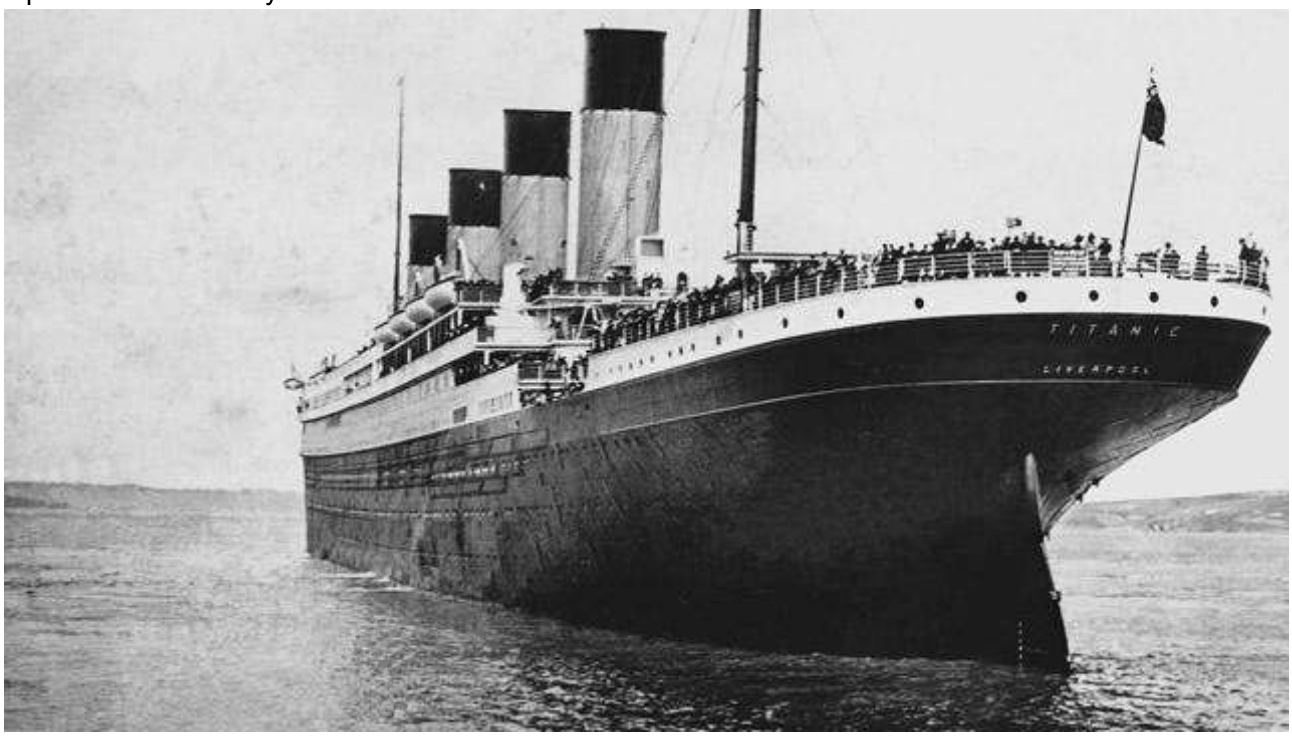
The **RMS TITANIC** was a British passenger ship .The sinking of this ship is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, it sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.



Now lets analyse the people who were more likely to survive .Here we can arise some questions

- What is the percentage of male and female survived ?
- Did Age play role in their survival ?
- Which class passengers were more likely to survive?
- What was the ticket fare for survivers and victims?

In [1]:
```
#imports
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:
```
# Loading data set

titanic=pd.read_csv('https://d17h27t6h515a5.cloudfront.net/topher/2016/September/57e9a84c_titanic-data/titanic-data.csv')
```

In [3]:
```
#Showing first 5 rows in the dataset

titanic.head()
```

Out[3]:

	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	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.92
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05

In [4]:
```
titanic.shape
```

Out[4]:
```
(891, 12)
```

As we see we have data about 891 passengers .However according to wikipedia (https://en.wikipedia.org/wiki/RMS_Titanic) the no of passengers were 2224 .But we will consider this amount of dataset .So lets come to the first question i.e

What is the percentage of male and female survived ?

In [5]:
```
#Categorising male and female we can get their ratios respectively

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

Out[5]:
```
Sex
female    314
male      577
dtype: int64
```

As we can see the no of *male* passengers is considerably higher than that of *females*. Now lets see **"how much male and female survived?"**

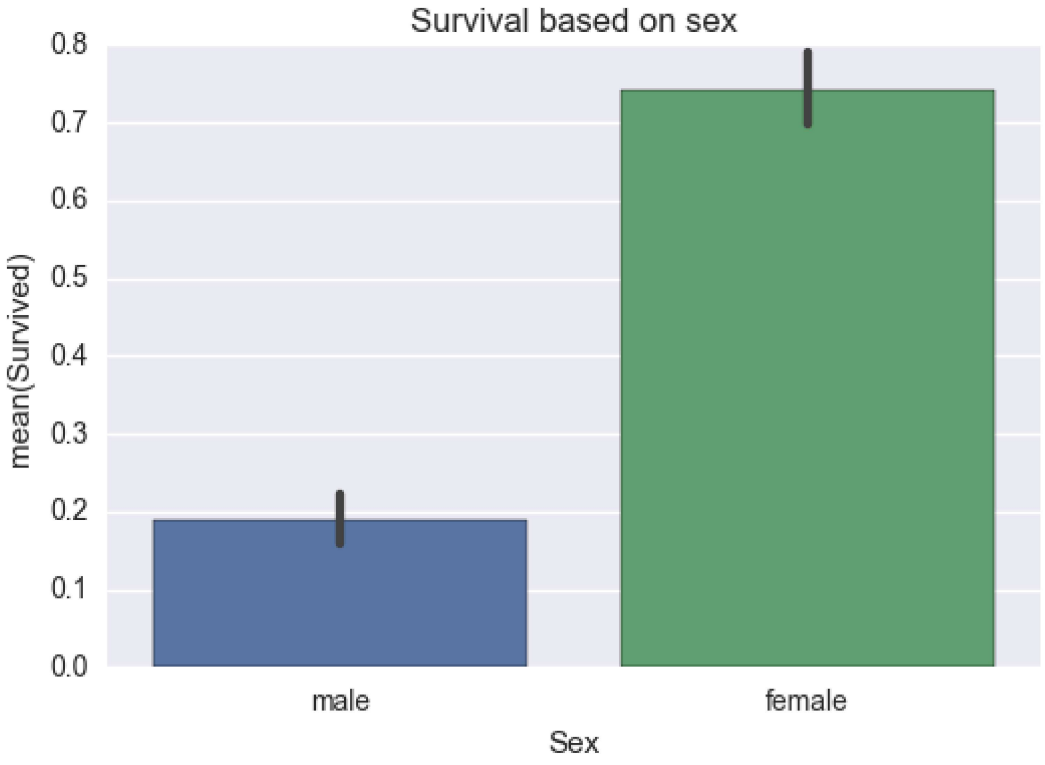
In [6]:
```
titanic.groupby('Sex').Survived.sum()
```

Out[6]:
```
Sex
female    233
male      109
Name: Survived, dtype: int64
```

As we can see the percentage of male survived is only 19% and that of female is 74%. From this we can expect that the females were rescued with higher priority than males thats why the male survivers are very fewer than that of males. Lets plot to get a better view .

In [7]:
```
ax=sns.barplot(titanic.Sex,titanic.Survived)
ax.set_title('Survival based on sex')
```

Out[7]:
```
<matplotlib.text.Text at 0x1f87c0fe438>
```



Now lets come to the next question i.e **Did Age play role in their survival ?**

Lets first see if there is any null values

In [8]:
```
titanic.isnull().sum()
```

Out[8]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Handling missing values

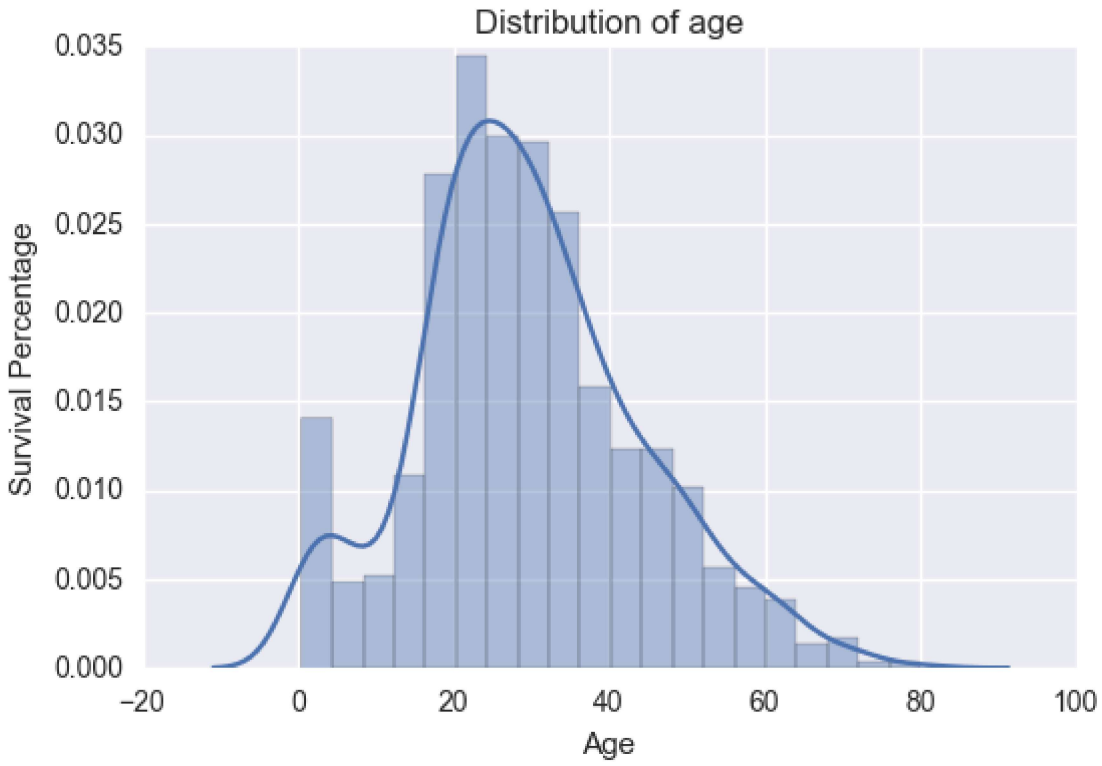
In [9]:
```
#lets first exclude the rows with null Age

Removed_nullAge=titanic.dropna(subset=['Age'],axis=0)
```

In [10]:
```
ax=sns.distplot(Removed_nullAge.Age)
ax.set_title('Distribution of age')
ax.set(ylabel='Survival Percentage',xlabel='Age')
```

```
C:\Users\kamal\Anaconda3\lib\site-packages\statsmodels\nonparametric\kdetool
s.py:20: VisibleDeprecationWarning: using a non-integer number instead of an
 integer will result in an error in the future
  y = X[:m/2+1] + np.r_[0,X[m/2+1:],0]*1j
```

Out[10]:
```
[<matplotlib.text.Text at 0x1f87c022860>,
 <matplotlib.text.Text at 0x1f87c92f550>]
```



Now we can say most passengers are under age 20 to 40 .Now lets categorise the passengers into age groups so that we can analyse the survival according to different age groups.

In [11]:
```
age_bins = [1, 15, 30, 45, 60, 80]
age_labels = ["Child","Young", "Middle-aged", "Senior", "Old"]
Removed_nullAge['Age_cat']=pd.cut(Removed_nullAge.Age,age_bins, labels=age_lab
els,
    right=True, include_lowest=True)
```

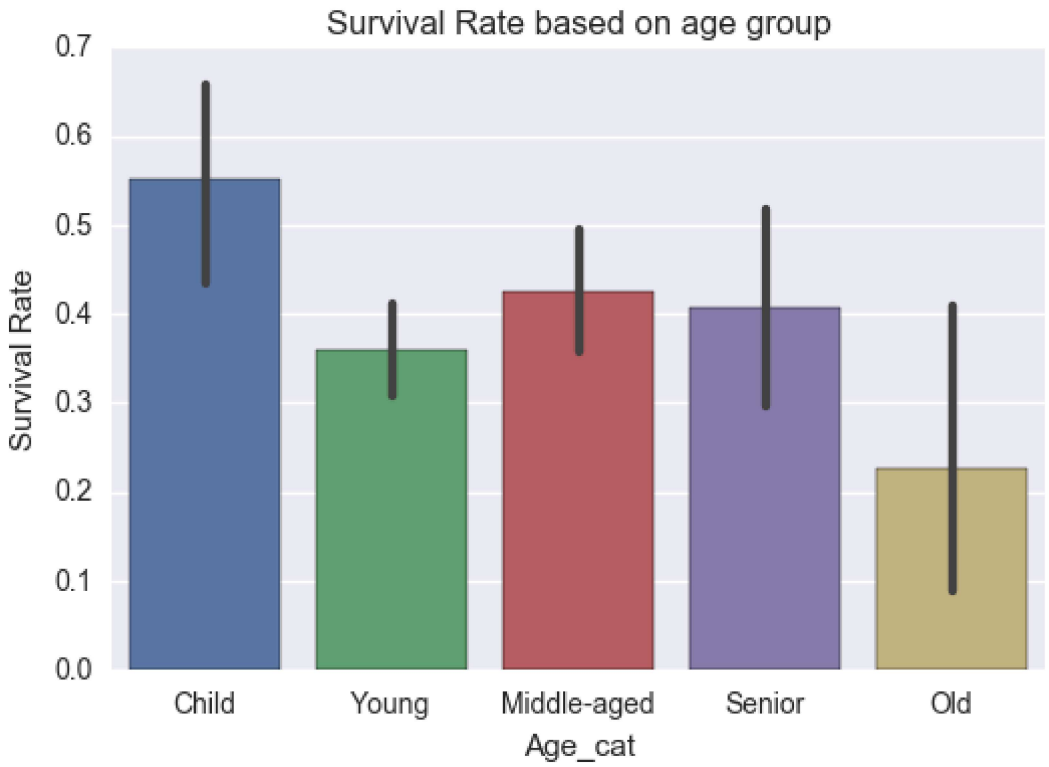
```
C:\Users\kamal\Anaconda3\lib\site-packages\ipykernel\__main__.py:4: SettingWi
thCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/st
able/indexing.html#indexing-view-versus-copy
```

In [12]:
```
#Calculating the no of people survived
Removed_nullAge.groupby('Age_cat').Survived.sum()
```

Out[12]:
```
Age_cat
Child           42
Young          117
Middle-aged     86
Senior          33
Old              5
Name: Survived, dtype: int64
```

In [13]:
```
ax=sns.barplot(data=Removed_nullAge,x='Age_cat',y='Survived')
ax.set(ylabel='Survival Rate')
ax.set_title('Survival Rate based on age group')
sns.plt.show()
```



The intresting thing is we can see the survival nuber of children is very high i.e greater than 60% where as the survival no of other groups are less 35% . Here we can say that both children and women were rescued first .

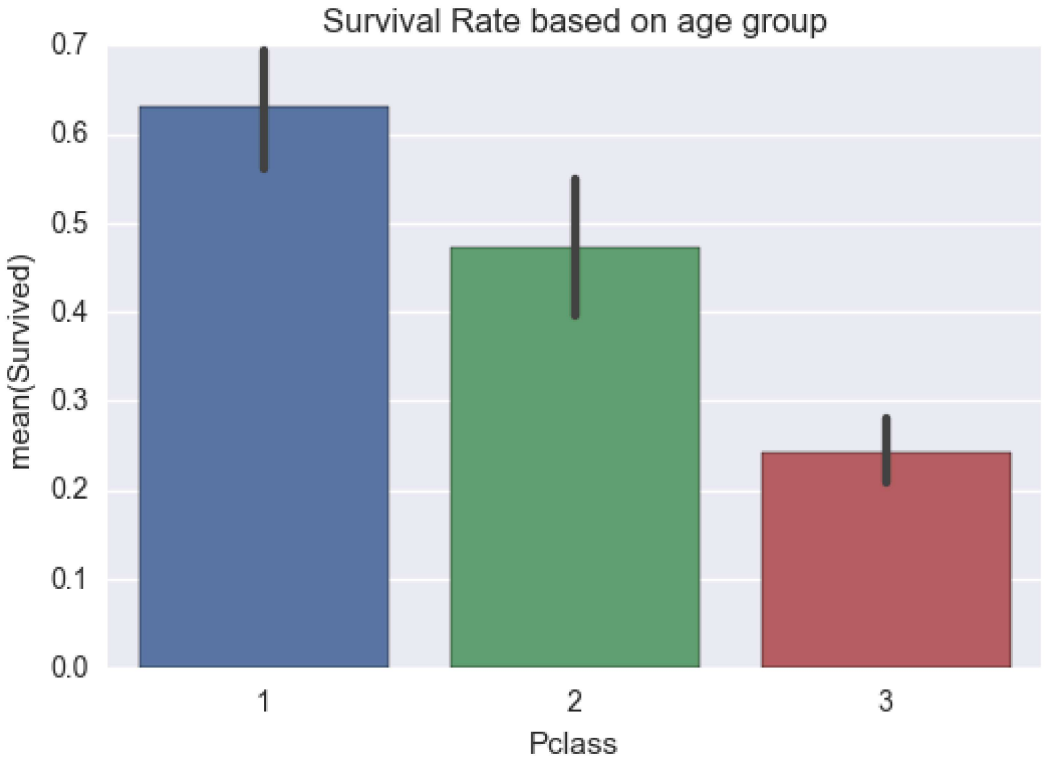
Now let move to the next question **"Which class passengers were more likely to survive?"**

In [14]:
```
#grouping the dataset according to Pclass

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

Out[14]:
```
Pclass
1    216
2    184
3    491
dtype: int64
```

In [15]:
```
#t=titanic.groupby('Pclass').Survived.sum()
#plotting the mean of survival according to Pclass
ax=sns.barplot(titanic.Pclass,titanic.Survived)
ax.set_title('Survival Rate based on age group')
sns.plt.show()
```



Here we can see the survival of first class passengers is very high .whereas the number of 2nd class passengers is fewer and the survival no. of 3rd class passengers is the lowest .So we can expect the passengers of higher class were rescued with higher prioity than that of the lower ones which causes such a result .

Now lets move to the next question i.e **"What was the ticket fare for survivers and victims?"** .Lets first see how the ticket price was varrying with Pclass

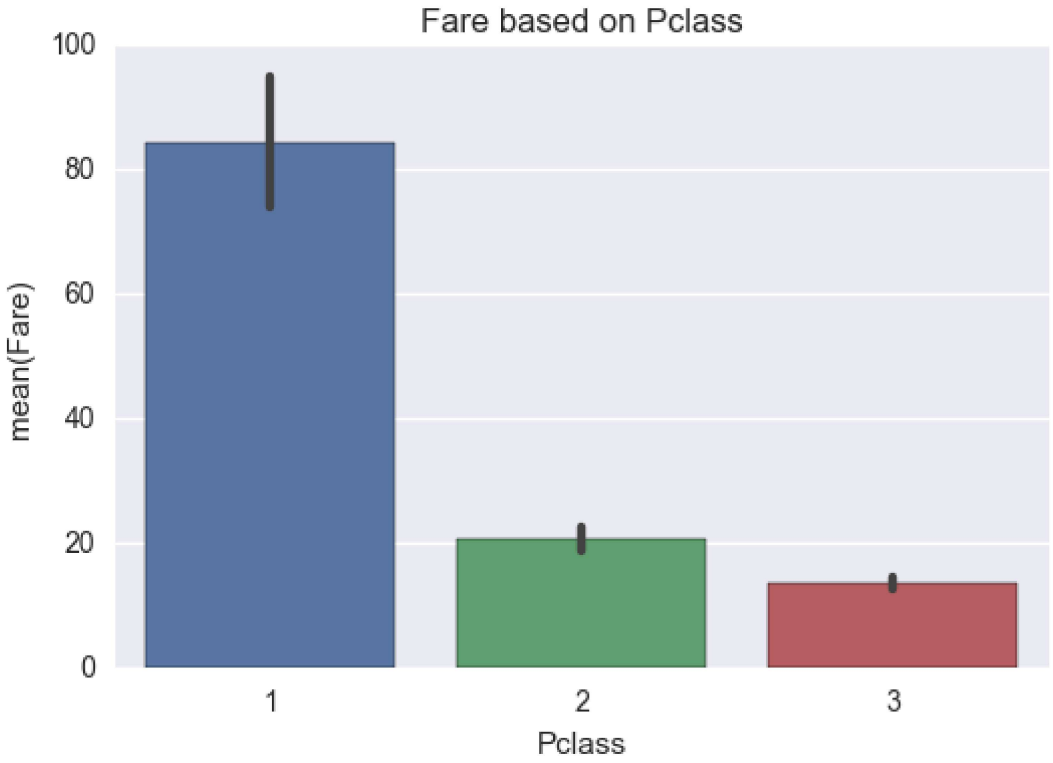
In [16]:
```
#calculating mean of fares according to Pclass

titanic.groupby('Pclass').Fare.mean()
```

Out[16]:
```
Pclass
1    84.154687
2    20.662183
3    13.675550
Name: Fare, dtype: float64
```

In [17]:
```
#plotting the mean fare according to Pclass

ax=sns.barplot(titanic.Pclass,titanic.Fare)
ax.set_title('Fare based on Pclass')
sns.plt.show()
```



As we can see the mean fare rate of first class passengers was extremely high than that of the others .It shows they must be rich family or businessmen .That is why their survival rate was high as they were rescued with high priority .

Now lets see if there were people having journey with no fare --

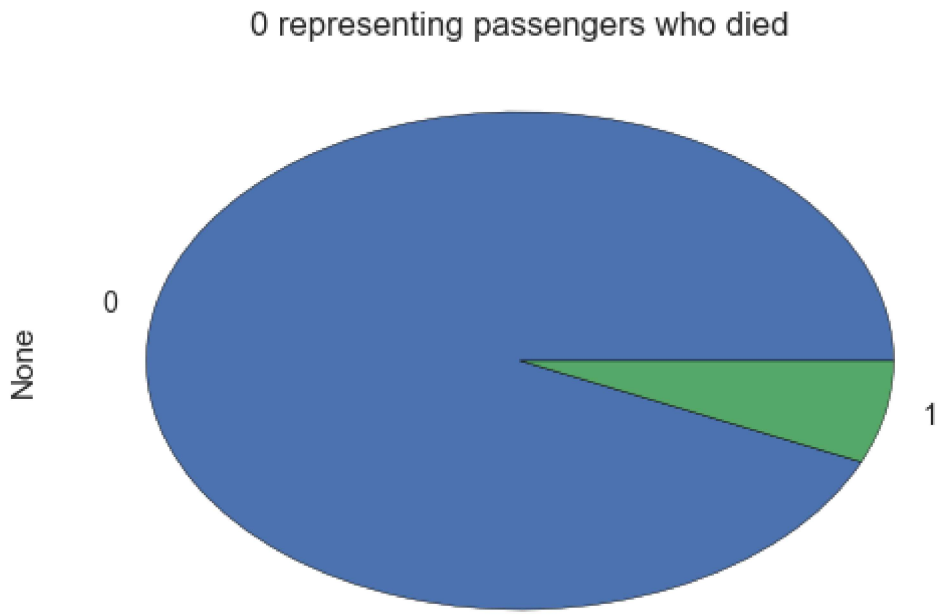
In [18]:
```
# people going free of cost
titanic.loc[titanic.Fare==0,:].Name
```

Out[18]:
```
179                 Leonard, Mr. Lionel
263               Harrison, Mr. William
271         Tornquist, Mr. William Henry
277          Parkes, Mr. Francis "Frank"
302      Johnson, Mr. William Cahoone Jr
413       Cunningham, Mr. Alfred Fleming
466                Campbell, Mr. William
481     Frost, Mr. Anthony Wood "Archie"
597                 Johnson, Mr. Alfred
633        Parr, Mr. William Henry Marsh
674           Watson, Mr. Ennis Hastings
732                 Knight, Mr. Robert J
806               Andrews, Mr. Thomas Jr
815                     Fry, Mr. Richard
822      Reuchlin, Jonkheer. John George
Name: Name, dtype: object
```

When I googled all these names I found most of them are staffs or got pass to travel .Now lets see How many of them were lucky enough to get survived.

In [19]:
```
ax=titanic.loc[titanic.Fare==0,:].groupby('Survived').size().plot(kind='pie')
ax.set_title('0 representing passengers who died')
```

Out[19]:
```
<matplotlib.text.Text at 0x1f87cb89dd8>
```



So only a single person survived out of all others who were travelling free .So we can say their journey was free of cost but the consequence was overpriced.

Conclusion :

After all the analysis we can say that the survival rate of female was significantly higher than the male passengers . And also we can say the Child aged passengers also had a higher survival rate than others . Also we found that the passengers in Pclass had a very high survival rate.

In the last observation we found that the passenger with free ticket had a less chance of survival which can be considered as a correlation but not a causation .

The main conclusion about the analysis is that the data set is not a good sample to perform some analysis and made conclusions about the population. Various issue and limitation with the dataset or report are :

- We did our analysis only taking 40% people into account which could have an effect in our analysis
- The dataset has alot of missing value related to Age which may produce different trend.Its just that we dont know how it will affect
- We haven't analysed the passenger survival based on cabin in which they stayed.