In [1]:

```
import pandas as pd
from pandas import DataFrame
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib
%matplotlib inline  #the aim of this command is to show figure in notebook without 'plt.show()'
#matplotlib.style.use('ggplot')
#from numpy.random import randn
```

In [2]:

```
titanic_data=pd.read_csv(r'titanic_data.csv')
titanic_data.head(10)
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	7
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	5
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	5
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	2
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	1
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	3

1.Preliminary Understanding

The first 10 records in our data set are showed above, indicating there are 12 parameters for each passenger, 'PassengerId' is the unique key for each person so it will be easy to count passenger. Next, I want to know **datatype of each columns and how many NaNs here.** Knowing these is important since **NaN** can tell me which columns needs further cleaning, and **datatype** can tell me what I can do using special columns.

In [3]:

```
columns_name=titanic_data.columns
NaN_count = titanic_data.ix[:,:].isnull().sum()
NaN_count_df = pd.DataFrame(NaN_count,columns=['NaN numbers'])
print 'there are %d rows,\nand %d columns in this data set ' % (len(titanic_data),titanic_data.s
hape[1])
datatype_df = pd.DataFrame(titanic_data.dtypes,columns=['Datatype'])
data_description = pd.merge(NaN_count_df,datatype_df,left_index =True,right_index = True)
data_description
```

```
there are 891 rows,
and 12 columns in this data set
```

Out[3]:

	NaN numbers	Datatype
PassengerId	0	int64
Survived	0	int64
Pclass	0	int64
Name	0	object
Sex	0	object
Age	177	float64
SibSp	0	int64
Parch	0	int64
Ticket	0	object
Fare	0	float64
Cabin	687	object
Embarked	2	object

Explanation: Note that there are **177** rows in 'age' and **687** rows in 'Cabin' NaNs in this data set with only 891 rows in total, actually maybe I will not use the data from Cabin as it is not an important factor. But age paramenter is significant, so that those 177 person will not make sense while analyzing data with respect to age.

How do I deal with those invalid values

In our dataset, we have three columns(Age Cabin Embarked) with invaild values, here shows how do I deal with those parameters

- **Age** : there are 177 NaNs in 'Age' columns with 891 records totally. In my report, I just dropped them, as pandas will ignore those NaN automatically while implementing statistical calculations(mean,max,etc), so I do not need to cancel them and create a new dataset, but when I drawed the distribution of age by boxplot, I droped those invalid age information by *dropna()* function.
- **Cabin** 687 records are invalid according to cabin columns, which indicates that 77 percent information of Cabin is unuseful, it will be much smaller if I drop them, so, here, I did not do anything because i did not use this data to analysis anything
- **Emarked** only 2 records in our dataset are lost, I dropped them, and since I just use this columns to draw some basic distribution aboard, and pandas will help me to drop them automaticlly, so I didn't do anything. Pandas is very useful for dealing with invalid data.

Questions

- 1. ***What is the distribution of ages,sibsp,sex,surived,parch,embarked,pclass and fare aboard?***
- 2. ***What is some common character of survivors and nonsurvivors? Does gender or social class or other parameters have significant impact on surivival***?

First question

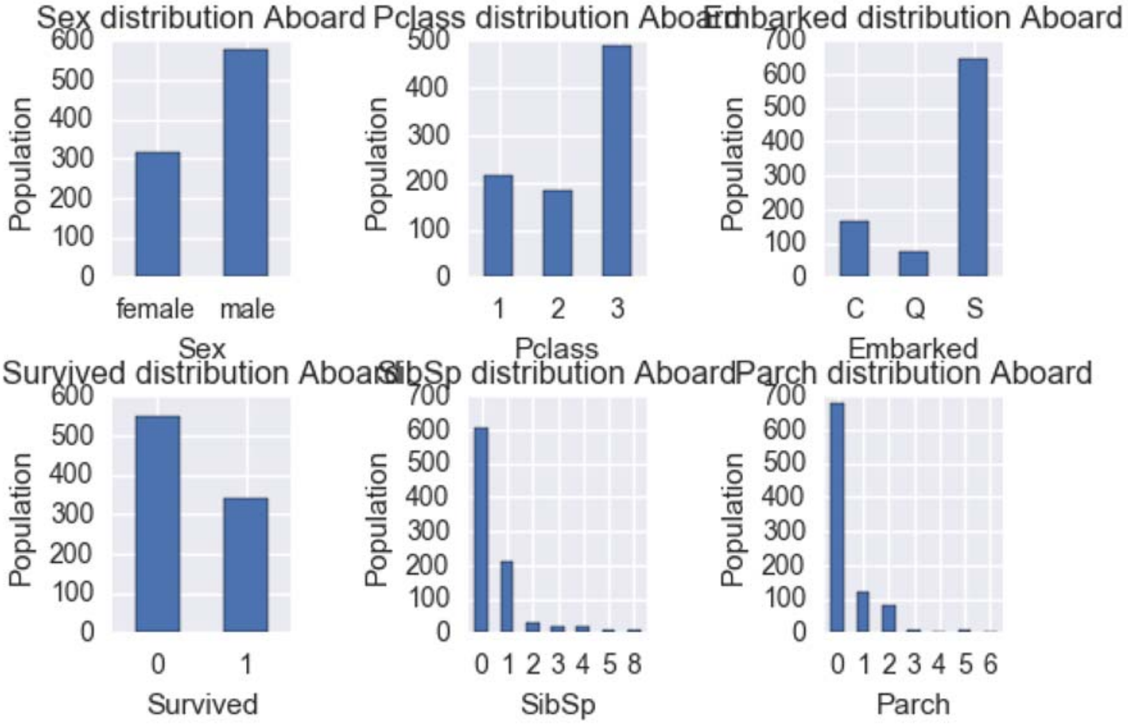
What is the distribution of ages,sibsp,sex,surived,parch,embarked,pclass and fare aboard?

In [4]:

```
#define a function to draw the basic plots of people characters
def distribution_plot(data,key='PassengerId',groupname= '',**kwargs):
    '''this fuction aims at drawing the distribtion of folks on the titanic board, function takes a lot of parameters for plot data on various
    variables
    data- means the data set that you want to draw
    key - means the columns name of the data set, this key will be a vital parameter for counting, so it must be exclusive, by defult,it is passengerID
    groupname - this function will show a distribution grouped by groupname
    plot type: this function will only provide a bar plot for each subplot,and all the obtained figures will be arranged in a figure.
    '''
    #check the key exists
    if not groupname:
        raise Exception("hi sorry,you did not provide a key to me, make sure which parameter that u wanna plot")
    if groupname not in data.columns.values:
        raise Exception("there is no '{}' in the data set,did u spell sth. wrong, be attention that all keys in the original data set with an upper case first letter".format(key))
    key_distribution = data[key].groupby(data[groupname]).count()
    fig = key_distribution.plot(kind = 'bar')
    fig.set_title('%s distribution Aboard'%(groupname))
    fig.set_ylabel("Population")
    plt.xticks(rotation = 0)
```

In [39]:

```
#draw six plots
fig=plt.figure()
groupnames = ['Sex','Pclass','Embarked','Survived','SibSp','Parch']
for i in xrange(len(groupnames)):
    ax_i = fig.add_subplot(2,3,i+1)
    distribution_plot(titanic_data,groupname=groupnames[i])
fig.subplots_adjust(wspace = 1,hspace = 0.5,top =0.9)
```



In [31]:

```
#titanic_data['Survived'].plot(kind='hist',bins = 2,)
#plt.hist(titanic_data['Sex'])
#plt.xticks([0,1])
```

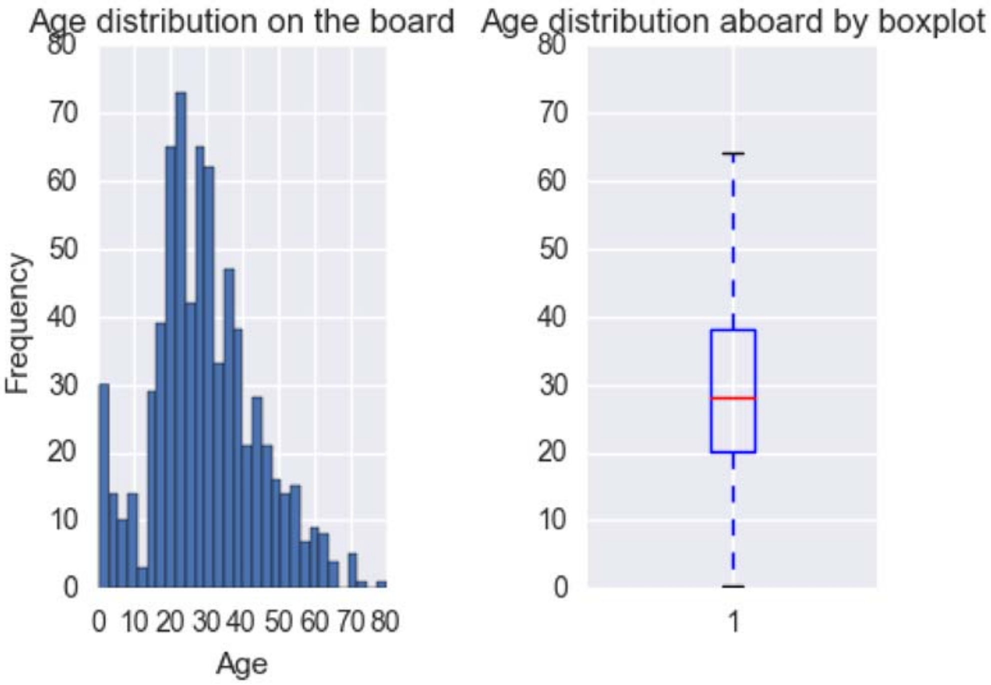
Sex,Age,Embarked,Class Siblings and Parch numbers distribution:

Explanation: we have 891 records in our data set which means the same number of people aboard, most of them died(about 61%), number of male is about 2 times than female's. Note that most of people were embarked in Southampton, and most are belong to third class which means poorer. sibsp distribution graph shows that most people do not have any siblings on the board, but it still have apprximate 200 people with 1 sibling and about 200 persons with 1 or 2 parents or children on the board.

In [49]:

```
#age distribution on the board
value_range = titanic_data[’Age’].max() - titanic_data[’Age’].min()
fig_age = plt.figure(figsize=(5,3.5))
bins = 30
ax1 = fig_age.add_subplot(121)
titanic_data[’Age’].plot(kind = ’hist’,bins = bins, width = value_range / bins,title = ’Age dist
ribution on the board’)
ax1.set_xlabel(’Age’)
ax2= fig_age.add_subplot(122)
plt.boxplot(titanic_data[’Age’].dropna())
ax2.set_title(’Age distribution aboard by boxplot’)
fig_age.subplots_adjust(wspace = 0.7,hspace = 0.5)
mean_age = titanic_data[’Age’].mean()
min_age = titanic_data[’Age’].min()
max_age = titanic_data[’Age’].max()
print ’The average age of people aboard is only %s years old \n the youngest is %s years old \n
the oldest is %s years old’ % (mean_age, min_age, max_age)
#plt.savefig(’age_distribution_by_bar_and_boxplot.png’,dpi=300,transparent = True)
```

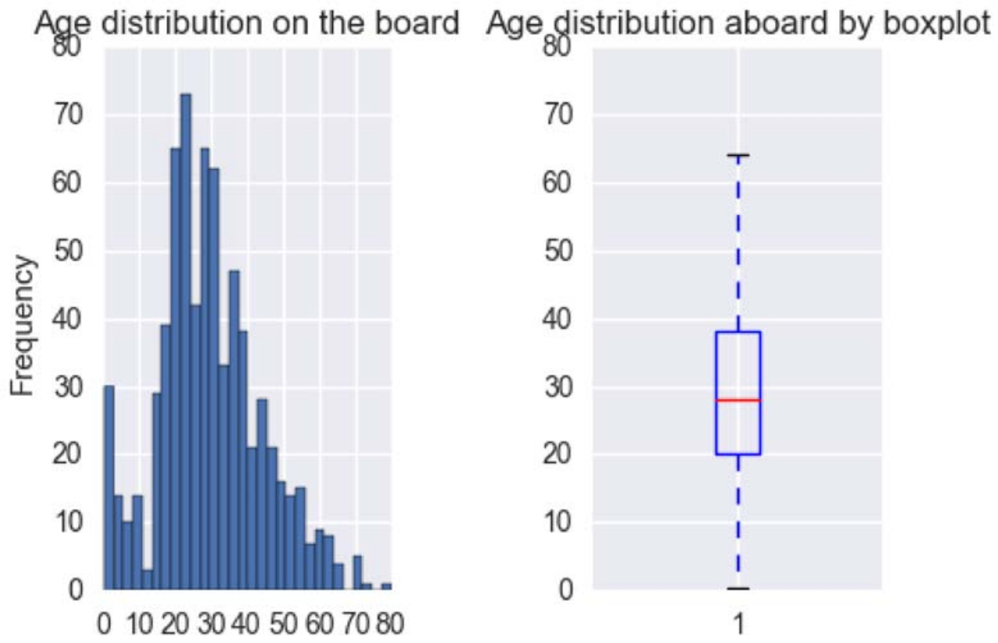
```
The average age of people aboard is only 29.6991176471 years old
 the youngest is 0.42 years old
 the oldest is 80.0 years old
```



In [43]:

```
#age distribution on the board
value_range = titanic_data['Age'].max() - titanic_data['Age'].min()
fig_age = plt.figure(figsize=(5,3.5))
ax1 = fig_age.add_subplot(121)
titanic_data['Age'].plot(kind = 'hist',bins = 30, width = value_range / bins, x='Age',ax = ax1)
ax1.set_title('Age distribution on the board')
ax2= fig_age.add_subplot(122)
plt.boxplot(titanic_data['Age'].dropna())
ax2.set_title('Age distribution aboard by boxplot')
fig_age.subplots_adjust(wspace = 0.7,hspace = 0.5)
mean_age = titanic_data['Age'].mean()
min_age = titanic_data['Age'].min()
max_age = titanic_data['Age'].max()
print 'The average age of people aboard is only %s years old \n the youngest is %s years old \n
the oldest is %s years old' % (mean_age, min_age, max_age)
#plt.savefig('age_distribution_by_bar_and_boxplot.png',dpi=300,transparent = True)
```

```
The average age of people aboard is only 29.6991176471 years old
 the youngest is 0.42 years old
 the oldest is 80.0 years old
```

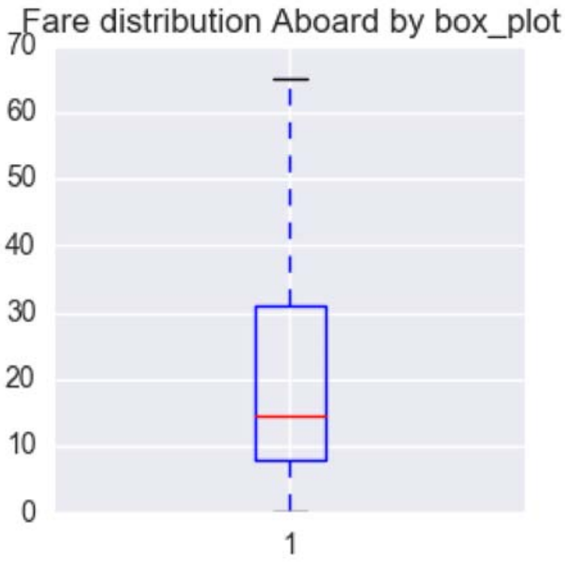


Age distribution aboard

Explanation: From age distribution drawing, we can clearly see that **most people are youngs aged from 20 to 38, and the averge age is only 29.7**, and the oldest person is 80 years old, the youngest is only 0.42 years old.

In [7]:

```
#passenger fare distribution aboard by box plot
box_fig = plt.figure(figsize = (3,3))
ax = box_fig.add_subplot(111)
plt.boxplot(titanic_data['Fare'])
ax.set_title('Fare distribution Aboard by box_plot')
ax.set_ylim([0,70])
#plt.savefig('fare_distribution_by_boxplot.png',dpi = 200,transparent = True)
plt.show()
mean_fare = titanic_data['Fare'].mean()
min_fare = titanic_data['Fare'].min()
max_fare = titanic_data['Fare'].max()
print 'The average fare of each one aboard is %s dollars \n the cheapest is %s \n the most expen
sive fare is %s dollars' % (mean_fare, min_fare, max_fare)
```



```
The average fare of each one aboard is 32.2042079686 dollars
 the cheapest is 0.0
 the most expensive fare is 512.3292 dollars
```

Fare distribution

Explanation: The price of ticket is mostly between 9 to 30 for different classes.The average fare of each one aboard is 32.2042079686 dollars,the cheapest is 0.0, the most expensive fare is 512.3292 dollars this is also outliers of this boxplot.

2 Plots and Exploration:

we have finished some individual plots above, and we now know the population conditions on the board. But the aim of this report is to make sure which kind of people is much more easily survived in this disaster. Let's firstly take a look at what the average persons looks based on the survived or not. since the status of life is only 0 or 1, so we divide all data into two groups and calculate its average or distribution.

In [8]:

```
col_names = [’Pclass’,’SibSp’,’Parch’,’Fare’,’Age’]
groupby_survived = titanic_data.groupby(’Survived’)[col_names].mean()
df_sex_surv = titanic_data.pivot_table(’PassengerId’,index = [’Survived’,’Sex’],aggfunc=’count’)
df_embarked = titanic_data.pivot_table(’PassengerId’,index = [’Survived’,’Embarked’],aggfunc=’count’)
print df_sex_surv,df_embarked
groupby_survived
```

```
Survived  Sex
0         female     81
          male      468
1         female    233
          male      109
Name: PassengerId, dtype: int64 Survived  Embarked
0         C            75
          Q            47
          S           427
1         C            93
          Q            30
          S           217
Name: PassengerId, dtype: int64
```

Out[8]:

	Pclass	SibSp	Parch	Fare	Age
Survived					
0	2.531876	0.553734	0.329690	22.117887	30.626179
1	1.950292	0.473684	0.464912	48.395408	28.343690

A few observations between we dig deeper:

Explanation: 0 represents died, 1 represents survived. It seems that Sex,Fare,Pclass have much impact on the surviral consequence.Male and person who had payed lower fare have low proportion survived during this disaster.

Second questions:

***What is some common character of survivors and nonsurvivors? Does gender or social class or other parameters have significant impact on surivival*?**

I am so interested in the relationship between Pclass and Fare and how they affect passenger's surviral results. So Let's take a look at Pclass versus fare.

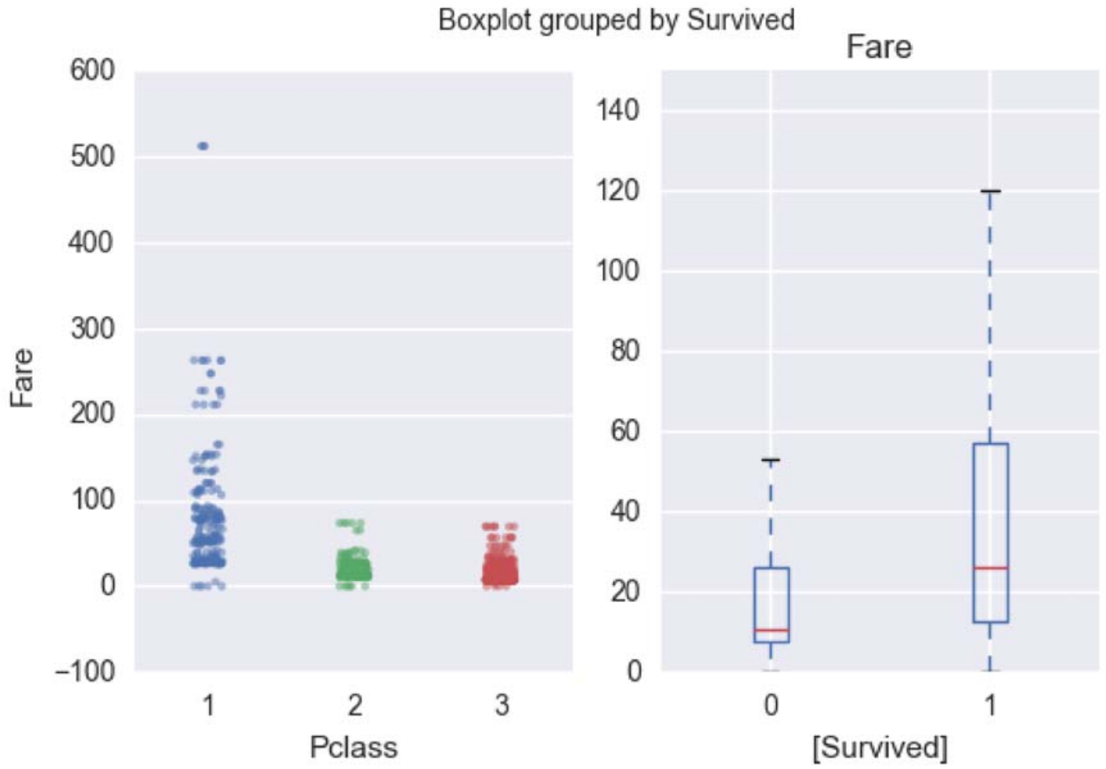
In [32]:

```
fig_pclass = plt.figure()
ax_1 = fig_pclass.add_subplot(121)
sns.stripplot(x='Pclass', y='Fare', data=titanic_data,
              size=3, alpha=.5, jitter=True, edgecolor='none')
sns.despine()
ax_2 = fig_pclass.add_subplot(122)
ax = plt.gca()
titanic_data[['Survived','Fare']].boxplot(by = 'Survived',figsize = (5,5),ax=ax)
plt.ylim([0,150])

#http://stackoverflow.com/questions/15067668/how-to-get-a-matplotlib-axes-instance-to-plot-to
 link to gca()
# the link introducing jitter, http://dataviztalk.blogspot.com/2016/02/how-to-add-jitter-to-plot
-using-pythons.html
```

Out[32]:

```
(0, 150)
```



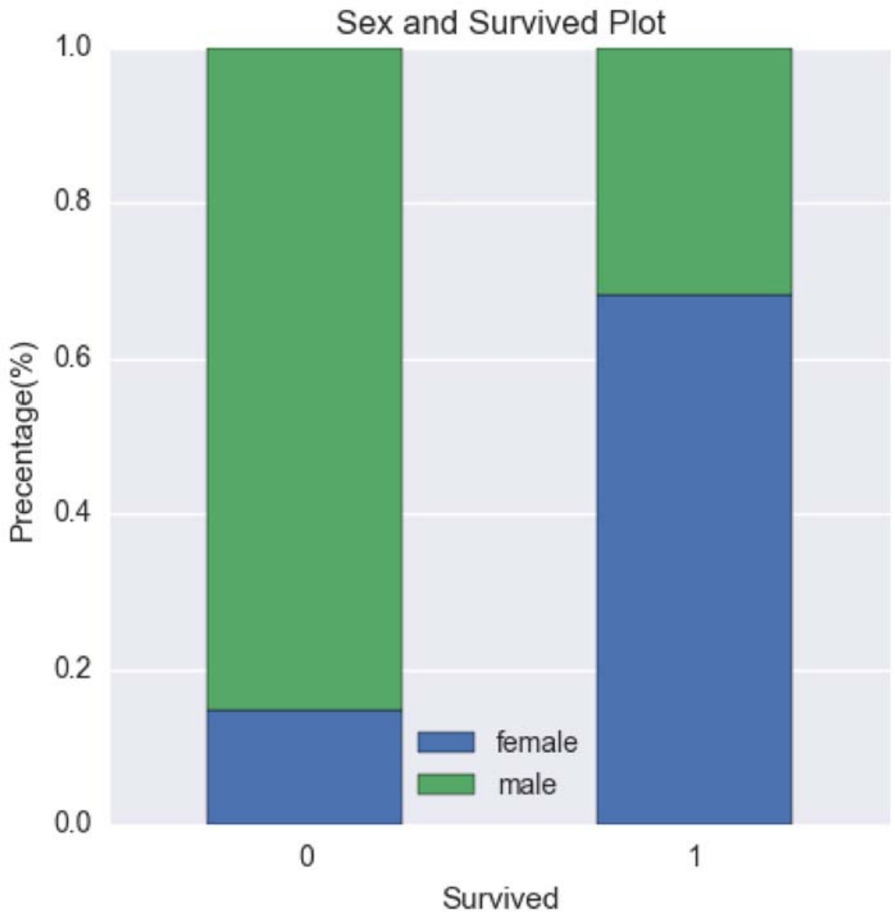
Explanation: From two graphs above, we can see that people in higher level class pays higher fare than folks in low class, it may comes to a conclusion that pclass and fare have a weak correlation. This is really easy to understand. From our boxplot of fare groupped by survived, it seems that most of deaths spend less for their tickets, note that the correlation between pclass and fare, this fact indicate that high class person always have high probability survive whereas low class person with low probility, and this result is also consistent with our data in the table showing the average pclass of survied persons is 1.95(high class), however for died person is 2.53(low class) Next I want to take a look at sex and age distribution, and wanna konw if there some common features of survivors or non-survivors about gender and age.

Age, sex versus survival

In [37]:

```
#firstly I want to know gender distribution of survivors and nonsurvivors, now i use stackd bar
 plot.
def stacked_plot(data,aggfunc,index,columns,values,title):
    data_pivot = data.pivot_table(values,columns=columns,index = index ,aggfunc= aggfunc)
    data_pivot_pct = data_pivot.div(data_pivot.sum(1).astype(float),axis=0)
    data_pivot_pct.plot(kind = 'bar',stacked = True,figsize = (5,5))
    plt.xticks(rotation = False)
    plt.ylabel('Precentage(%)')
    plt.legend(loc='best')
    plt.title(title)
    print data_pivot
#fig_sex_stack,axes = plt.subplots()
stacked_plot(data=titanic_data,aggfunc='count',index='Survived',columns='Sex',values='PassengerI
d',title='Sex and Survived Plot')
```

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



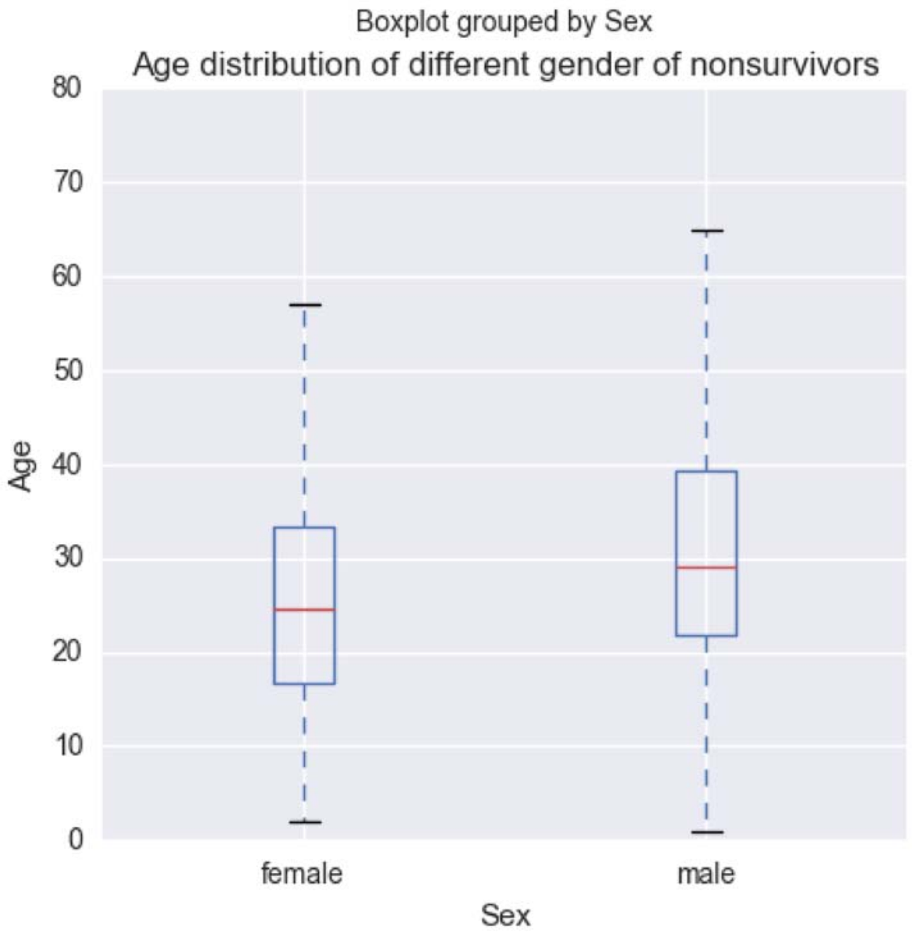
Explanation: The fact is that about 85% nonsurvivors are male,and actually aboout 80% male persons on the board are died however only 24% female person died.This is maybe in this disaster women and childern are always allowed going ahead.

In [36]:

```
titanic_data[titanic_data.Survived==0][['Sex','Age']].boxplot(by = 'Sex',figsize = (5,5))
plt.title('Age distribution of different gender of nonsurvivors')
plt.xlabel('Sex')
plt.ylabel('Age')
```

Out[36]:

```
<matplotlib.text.Text at 0x1152cfd0>
```

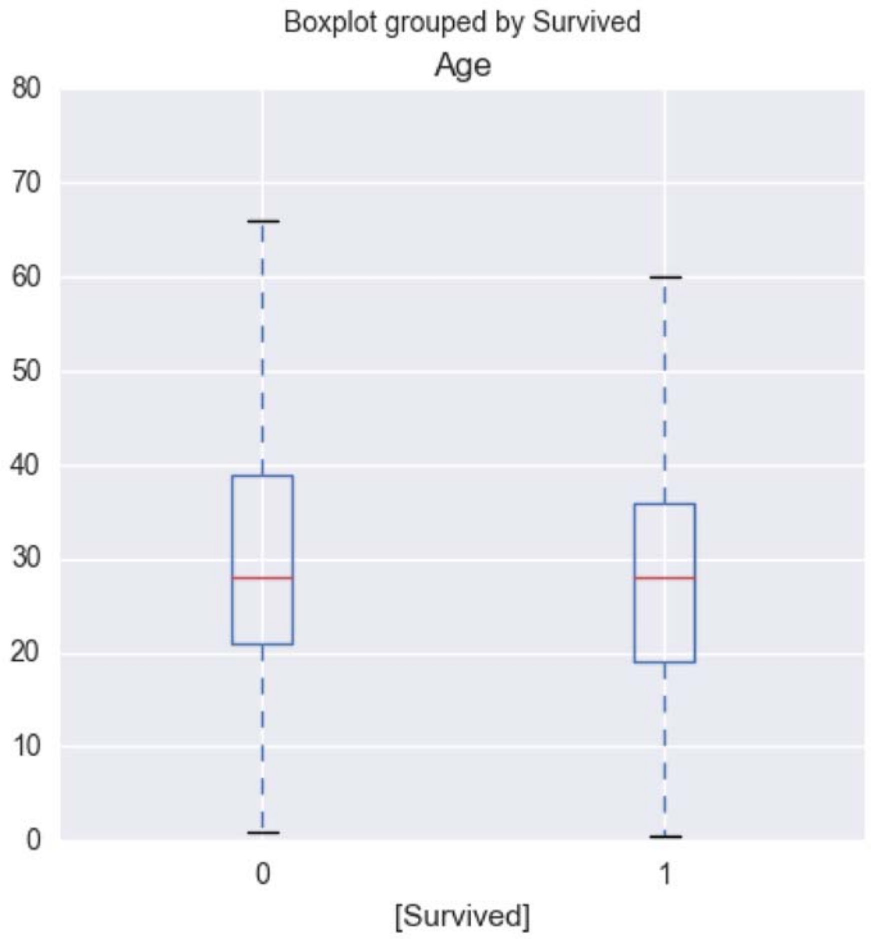


In [35]:

```
titanic_data[['Survived','Age']].boxplot(by = 'Survived',figsize = (5,5))
```

Out[35]:

```
<matplotlib.axes._subplots.AxesSubplot at 0xe0c8470>
```



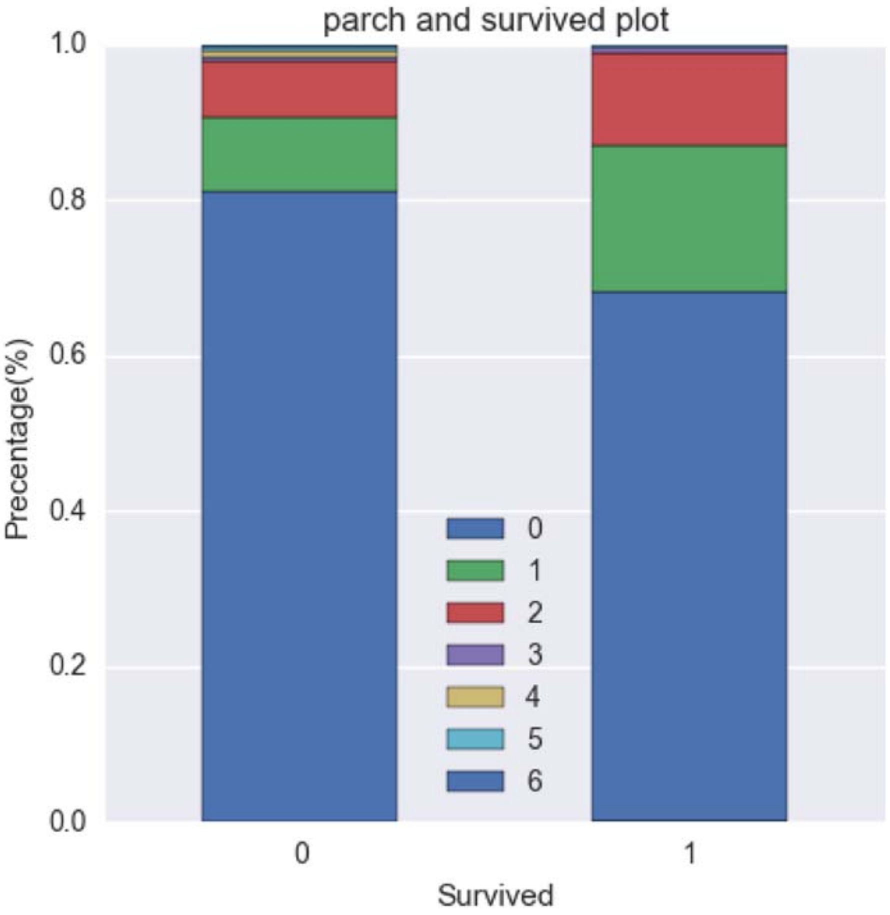
Explanation: So let's take a look at the age distribution of survivors and nonsurvivors. the conclusion is that age of survivors is little younger than non-survivors', whish originally means youngs are easy to survive but without obvious features. then I want to know that how siblings or parents or childern on the board will affect the results.

SibSp and parch versus survival

```
In  [13]:
```

```
stacked_plot(data=titanic_data,aggfunc='count',index='Survived',columns='Parch',values='Passenge
rId',title = 'parch and survived plot')
```

```
Parch         0     1     2    3    4    5    6
Survived
0         445.0  53.0  40.0  2.0  4.0  4.0  1.0
1         233.0  65.0  40.0  3.0  NaN  1.0  NaN
```

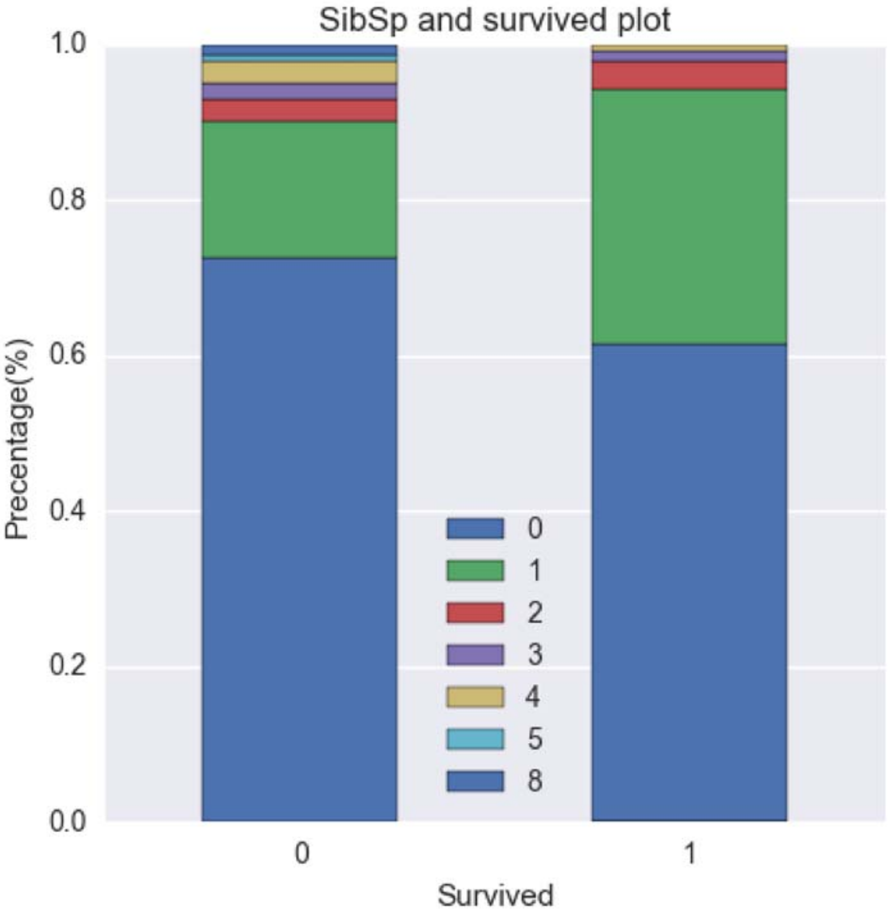


Explanation: 'Parch and survived plot' graph shows that 80% nonsurvivors do not have parent or children on board,and the proportion of survivors is 70%.

```
In  [14]:
```

```
stacked_plot(data=titanic_data,aggfunc='count',index='Survived',columns='SibSp',values='Passenge
rId',title = 'SibSp and survived plot')
```

```
SibSp         0      1     2     3     4    5    8
Survived
0         398.0   97.0  15.0  12.0  15.0  5.0  7.0
1         210.0  112.0  13.0   4.0   3.0  NaN  NaN
```



Explanation: The effect of silbings number aboard to survivals is similar to Parch's. About 40% survivors have at least one silbings or spouses, however only 30% nonsurvivors have at least one silbings or spouses. **It seems that people with siblings or parchs or spouses have more possibility survived.**

3. Conclusions:

It is too early to make any substaintial conclusions but there is a summary information obtained from this data set.

3.1 Descriptive conclusions:

we have 891 records in our data set which means the same number of people aboard, most of them died(about 61%), number of male is about 2 times than female's. Note that most of people were embarked in Southampton, and most are belong to third class which means poorer. sibsp distribution graph shows that most people do not have any siblings on the board, but it still have apprximate 200 people with 1 sibling and about 200 persons with 1 or 2 parents or children on the board.From age distribution drawing, we can clearly see that most people are youngs aged from 20 to 38, and the averge age is only 29.7.The price of ticket is mostly between 9 to 30 for different classes.The average fare of each one aboard is 32.2042079686 dollars

3.2 Further conclusions:

- Features of survivors:
  - most of them are from high class holding expensive tickets
  - most of females are survived
  - 32% of them have at least one parchs and 40% have at least one sibsps aboard
- Features of non-survivors:
  - most of them are in low class holding cheap tickets
  - most of them(85%) are males
  - 20% of them have at least one parchs and 25% have at least one sibsps aboard

3.3 Comparisions

- Males have lower survival rate than females
- People in higher social class with higher price ticket have more possibility survived in the disaster
- The average age of survivors is little lower than nonsurvivors.

4. Limitations

1.By my data analysis, I conclude that male is more easily died than females, and giving a reason that women always being allowed to escape first. It seems reasonable, but let me think forther, can males still be so gentle during this disaster, are there any other factors will affect this results. Here are some guesses:

- There is no information about how many males aboard are stuffs,since stuffs and polices are organiser and maybe have more possibility died, and if that,those results are casused by their work instead of gender.
- Is that possible that males struggled to go to lifeboat, but most of them dropped into water eventually.
- Saple error is another possibility to cause this results.

2.Another limitation in my report is I merely apply some statistical calculations and basic plots, this maybe cause some wrong conclusions because of sample error. It will be more precise if I can carry out some hypothesis testing.