泰坦尼克号乘客生还影响因素的分析

提出问题

关于泰坦尼克号沉没时的自救以及乘客生还情况已经有很多的报道。一方面，船员和乘客在危急时刻优先让妇女、小孩和老人登上救生艇，体现了人类的善良的品质。另一方面，根据资料显示，由于一等舱的位置比其他舱要靠上，一等舱乘客有着更好的逃生机会，生还的可能性也更高。

利用泰坦尼克号乘客数据，可以掌握登船乘客的主要特征（年龄、性别、客舱等级），并通过数据与图表来揭示哪些因素与乘客生还有着明显的联系。

这篇报告尝试回答以下三个问题：
小孩和老人的生还率是否比成年人要高?
女性是不是有着更高的生还可能性?
一等舱和二等舱的乘客是否有着更好的生还机会?
因此，这篇报告中主要分析不同年龄、性别以及客舱等级的人群在生还率上的差异。

In [32]:
```
import numpy as np
import pandas as pd

filename = '/Users/BoHai/Desktop/Desktop/Udacity/p2-探索数据集/titanic/titanic_data.csv'
titanic_df = pd.read_csv(filename)
titanic_df.head()
```

Out[32]:

	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	7.250(
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.283
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925(
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10(
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050(

数据总体描述

In [33]:
```
titanic_df.describe()
```

Out[33]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.2042
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.6934
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.00000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.91040
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.4542
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.0000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329

从总体上来看，数据中包括的891名乘客中有38%的乘客生还。乘客中最小年龄不到1岁，最大年龄为80岁，平均年龄为29.7岁。船上的兄弟姐妹及配偶，最少的0人，最多的8人，平均为0.5人。乘客的父母/子女数，最少为0人，最多为6人，平均为0.38人。乘客的票价，最低为0，最高为512，平均为32。

In [34]:
```
titanic_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
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(5), object(5)
memory usage: 83.6+ KB
```

从缺失值报告来看，12个变量中只有年龄和仓位两个变量有缺失。其中年龄这一变量有714个样本不为缺失值，这意味着有12.8%的样本年龄缺失，对于样本的分析会造成一定的影响。

In [35]:
```
passengerid = titanic_df['PassengerId']
survived = titanic_df['Survived']
pclass = titanic_df['Pclass']
sex = titanic_df['Sex']
age = titanic_df['Age']
sibsp = titanic_df['SibSp']
parch = titanic_df['Parch']
fare = titanic_df['Fare']
carbin = titanic_df['Cabin']
embarked = titanic_df['Embarked']
```

年龄（年龄组）与生还率

由于年龄这一变量的缺失较多，在使用年龄变量分析时利用均值进行插补（fillna函数）。

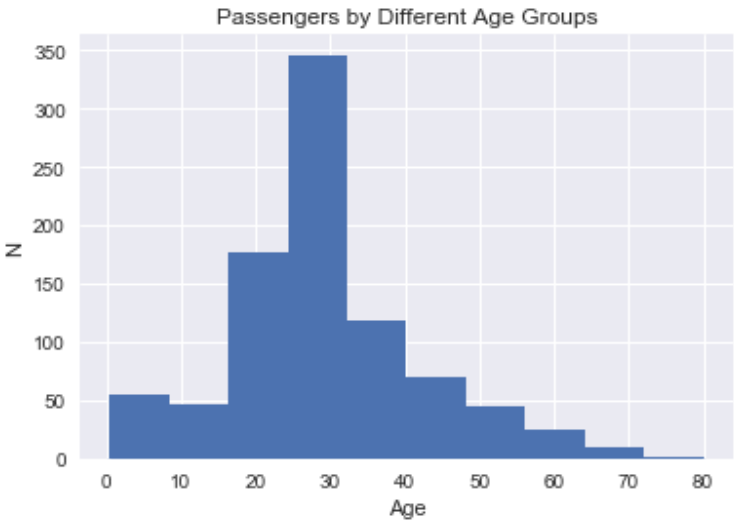
In [36]:
```
age_fix=age.fillna(age.mean())
print age_fix.count()
print age_fix.mean()
```

```
891
29.6991176471
```

In [37]:
```
import matplotlib
import seaborn
%pylab inline
plt.hist(age_fix)
plt.xlabel('Age')
plt.ylabel('N')
plt.title('Passengers by Different Age Groups')
```

```
Populating the interactive namespace from numpy and matplotlib
```

Out[37]: <matplotlib.text.Text at 0x118cdb990>



乘客年龄组频数分布图显示了乘客年龄大多集中在20-40岁，20岁以下的人群大约有100人，而50岁以上的人群则要更少。

下面将年龄进行分组，考察不同年龄组人群的生还率
未成年组(Nonage):0-16岁；
成年组(Adult):16-50岁；
老年组(Elder):50岁-；

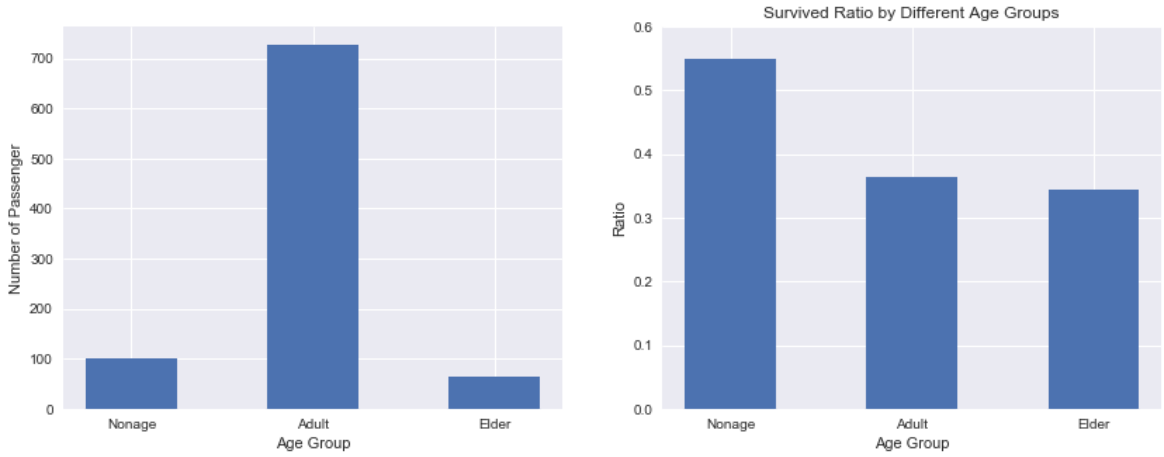
In [38]:
```
print age_fix.min()
print age_fix.max()
bins = [0, 16, 50, 80]
cats_age = pd.cut(age_fix, bins, labels=['Nonage','Adult','Elder'])
titanic_df['Age_group']=cats_age
print survived.groupby(cats_age).mean()
```

```
0.42
80.0
Age
Nonage    0.550000
Adult     0.364512
Elder     0.343750
Name: Survived, dtype: float64
```

In [39]:
```
%pylab inline
plt.figure(figsize = (14,5))
plt.subplot(1,2,1)
plt.bar(np.arange(3),survived.groupby(cats_age).count().dropna(),0.5)
plt.ylabel('Number of Passenger')
plt.xlabel('Age Group')
plt.xticks(np.arange(3), ('Nonage','Adult','Elder'))
plt.subplot(1,2,2)
plt.bar(np.arange(3),survived.groupby(cats_age).mean().dropna(),0.5)
plt.xlabel('Age Group')
plt.ylabel('Ratio')
plt.xticks(np.arange(3), ('Nonage','Adult','Elder'))
plt.yticks(np.arange(0,0.7,0.1))
plt.title('Survived Ratio by Different Age Groups')
```

```
Populating the interactive namespace from numpy and matplotlib
```

Out[39]:
```
<matplotlib.text.Text at 0x119074bd0>
```



从不同年龄组人群的对比来看，50岁以下的成年人大约有550人；未成年人和老年人只有100人和大约60人。由于样本过少，这使得分析未成年人和老年人本身就会存在一定的偏差。未成年组的生还率要远远高于成年组和老年组，达到55%。而成年组和老年组的生还率分别为38%和34%，成年组仅仅略高于老年组。这表明在泰坦尼克号发生撞击灾难后，船上船员与乘客将乘坐救生艇逃生的机会更多的给了儿童和少年。某些原因可能恰好使得船上的大部分老年人遇难了，而老年人在灾难中的生还率未必真的很低。年人的生还率低可能与老年人身体条件相对较差，轮船发生撞击后自我逃生能力比较弱有关。

性别与生还率

In [40]:
```
sex_dummy = sex=='female'
sex_dummy.head()
sex_dummy.mean()
```

Out[40]:
```
0.35241301907968575
```

891名乘客中有35%为女性乘客

In [41]:
```
survived.mean()
```

Out[41]: 0.3838383838383838

所有乘客的平均生还率约为38.4%

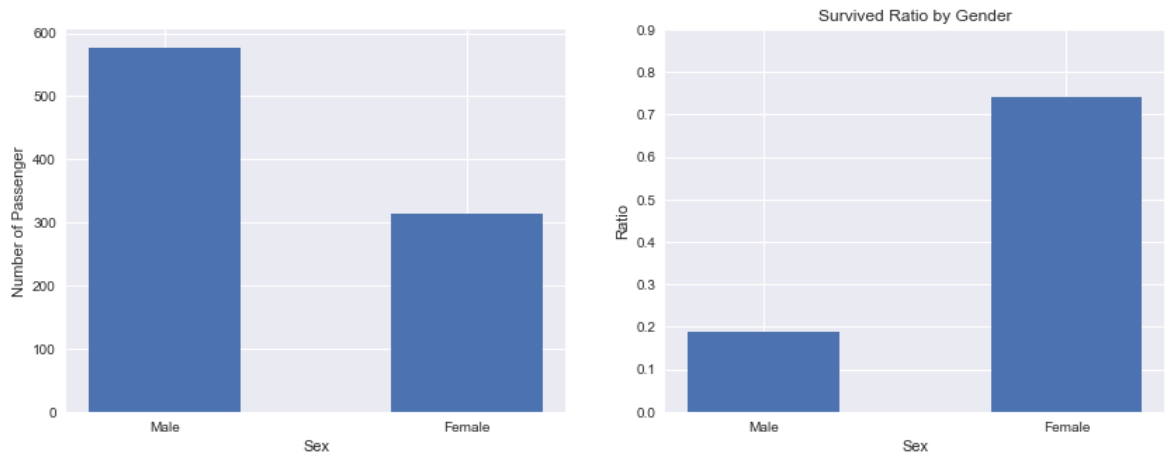
In [42]:
```
survived_sex = survived.groupby(sex_dummy)
survived_sex.mean()
```

Out[42]:
```
Sex
False    0.188908
True     0.742038
Name: Survived, dtype: float64
```

In [43]:
```
%pylab inline
plt.figure(figsize = (14,5))
plt.subplot(1,2,1)
plt.bar(np.arange(2),survived_sex.count(),0.5)
plt.ylabel('Number of Passenger')
plt.xlabel('Sex')
plt.xticks(np.arange(2), ('Male','Female'))
plt.subplot(1,2,2)
plt.bar(np.arange(2),survived_sex.mean(),0.5)
plt.xlabel('Sex')
plt.ylabel('Ratio')
plt.xticks(np.arange(2), ('Male','Female'))
plt.yticks(np.arange(0,1,0.1))
plt.title('Survived Ratio by Gender')
```

```
Populating the interactive namespace from numpy and matplotlib
```

Out[43]: <matplotlib.text.Text at 0x119410610>



从男女人数来看，男性人数比女性多200人以上；而男性的生还率仅为18.9%，女性的生还率为74.2%，女性的生还率远远高于男性。这也印证了在逃生过程中优先让妇女登上救生艇这一说法。正如资料中提到的，”在船的左舷，救生船只载妇女和儿童。在右舷，则是妇女优先逃生之后允许男性登艇。“

旅客客舱等级与生还率

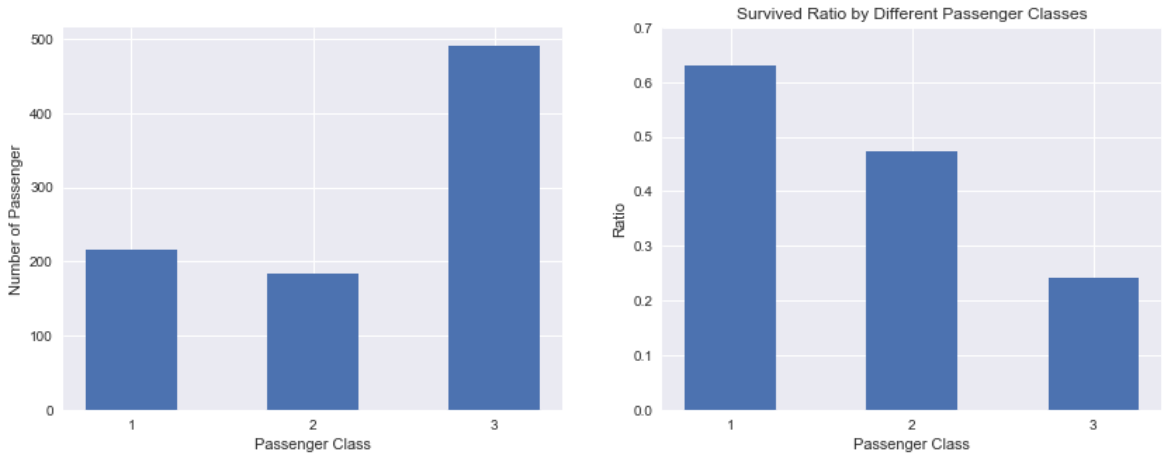
In [44]:
```
survived_pclass = survived.groupby(pclass)
survived_pclass.mean()
```

Out[44]:
```
Pclass
1    0.629630
2    0.472826
3    0.242363
Name: Survived, dtype: float64
```

In [45]:
```
%pylab inline
plt.figure(figsize = (14,5))
plt.subplot(1,2,1)
plt.bar(np.arange(3),survived_pclass.count(),0.5)
plt.ylabel('Number of Passenger')
plt.xlabel('Passenger Class')
plt.xticks(np.arange(3), ('1','2','3'))
plt.subplot(1,2,2)
plt.bar(np.arange(3),survived_pclass.mean(),0.5)
plt.xlabel('Passenger Class')
plt.ylabel('Ratio')
plt.xticks(np.arange(3), ('1','2','3'))
plt.yticks(np.arange(0,0.8,0.1))
plt.title('Survived Ratio by Different Passenger Classes')
```

```
Populating the interactive namespace from numpy and matplotlib
```

Out[45]:
```
<matplotlib.text.Text at 0x11965c850>
```



从资料中查阅得知，泰坦尼克号三等舱位在船身较下层也最便宜；二等舱与一等舱则在船身较上的位置。在事故发生之后，三等舱位的乘客逃生需要更长的时间，到达甲板也较为困难。从总体人数上来看，一等舱和二等舱乘客的样本远少于三等舱，人数少本身也减少了逃生的难度。而从不同等级客舱旅客的生还率来看，一等舱旅客的生还率为63%， 显著高于二等和三等舱，其中三等舱的生还率只有24.2%。从侧面印证了资料中的描述。

总结

在分析了年龄、性别与客舱等级三个因素和泰坦尼克号乘客生还率的关系后，发现未成年人、女性和头等、二等舱乘客的生还率更高。这一定程度上印证了当时船上组织妇女和儿童优先登上救生艇这一说法。而头等舱由于位于轮船较好的位置，也更加利于逃到甲板乘坐救生艇。

此外，由于变量的选择以及数据样本的关系，以上的分析得出的结论仍然存在一些限制。
首先，船上某一类人群的样本偏少，无法代表整个人口，可能使得分析的结论不够准确。有意思的现象的是老年乘客的生还率比较低，尽管人们也会优先让老年登上救生艇，但是一方面老年人的绝对人数很少（从图中可以看出只有大约60多人），过少的样本会存在着一定的样本选择偏差，因此并不能认为老年人在事故中的生存率一定会很低。当然，考虑到老年人的健康较差，轮船发生撞击、天气寒冷都可能是老年人逃生的不利因素。
其次，对数据缺失值的处理方式也会导致偏差。年龄这一变量存在着一定程度的缺失，这也影响了对各个年龄组人群生还率的分析。
最后，一些难以观察到但是对逃生起到关键因素的变量在数据中并没有体现出来，而这可能会影响结论。例如，尽管数据中可以区分仓位等级，但是难以根据船票号码确定乘客在船的哪个位置。在船沉没的过程中，左舷和右舷、船中部和头尾部（船体在中间发生断裂）可能对生还率有着很大的影响。

总的来看，在泰坦尼克号发生撞击到沉没的几个小时里，船上的船员和男性乘客还是发扬了人性中善的一面，将更好的逃生机会留给了妇孺。同时，舱位等级更高的乘客也有着更好的逃生机会。

注：所有文献资料均查阅自百度百科词条-泰坦尼克号 http://baike.baidu.com/item/泰坦尼克号/5677 (http://baike.baidu.com/item/泰坦尼克号/5677)