Titanic Survaval Analysis Project

In this proejct I analyse a dataset of Titanic passengers and make predictions as to which criteria made people more likely to survive.

I took the Titanic Dataset from Udacity link, which in turn linked to the dataset stored on Kaggle.com.

I will answer the following questions:

1) What was the gender distribution in the group? Who were more likely to survive depending on the gender?

2) What was the age distribution in the group? Passengers of what age were more likely to survive?

3) What was the travel class distribution in the group? Passents of which travel class were more likely to survive?

4) Is there a correlation between the existence of children/parents and the chances of survival? Were passengers with children/parents more likely to survive than those travelling alone?

5) Who had the maximum chances of survival in terms of gender/age/class/existence of children/parents?

In [2]:

```python
# Extract the passenger data from a csv file and store it as a pa

import pandas as pd
import numpy as np

titanic_data = pd.read_csv('titanic_data.csv')

#Let's explore the dataset
titanic_data.head()
```

Out[2]:

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

The description of the fields from Kaggle.com

Data Dictionary

Variable Definition Key

survival Survival 0 = No, 1 = Yes

pclass Ticket class 1 = 1st (Upper), 2 = 2nd (Middle), 3 = 3rd (Lower)

sex Sex

Age Age in years. Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp # of siblings / spouses aboard the Titanic Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)

parch # of parents / children aboard the Titanic Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

ticket Ticket number

fare Passenger fare

cabin Cabin number

embarked Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

Data wrangling

In [3]:

```
# Look at the end of the dataset for reference and additional inf
titanic_data.tail()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Pai
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	

In [4]:

```
# Identify any duplicates
print ('The number of duplicate entries is {}.'.format(titanic_da
```

```
The number of duplicate entries is 0.
```

In [5]:

```
# Let's see some statistics information about the dataset
titanic_data.describe()
```

Out[5]:

	PassengerId	Survived	Pclass	Age	SibSp	
count	891.000000	891.000000	891.000000	714.000000	891.000000	8!
mean	446.000000	0.383838	2.308642	29.699118	0.523008	
std	257.353842	0.486592	0.836071	14.526497	1.102743	
min	1.000000	0.000000	1.000000	0.420000	0.000000	
25%	223.500000	0.000000	2.000000	20.125000	0.000000	
50%	446.000000	0.000000	3.000000	28.000000	0.000000	
75%	668.500000	1.000000	3.000000	38.000000	1.000000	
max	891.000000	1.000000	3.000000	80.000000	8.000000	

```
From the above three tables we can see that some entries lack
information about the age, while some - the number of the cabin.
The information about the age is important for us, while the
number of the cabin is irrelevant.
```

Remove unnecessary colums

The columns 'PassengerId', 'Name', 'Ticket', 'Fare', 'Cabin' and 'Embarked' are not relevant for my analysis and will be removed.

In [6]:

```
titanic_data_short = titanic_data.drop(['PassengerId', 'Name', 'T
```

In [7]:

```
# Display the resulting dataframe
titanic_data_short.head()
```

Out[7]:

	Survived	Pclass	Sex	Age	SibSp	Parch
0	0	3	male	22.0	1	0
1	1	1	female	38.0	1	0
2	1	3	female	26.0	0	0
3	1	1	female	35.0	1	0
4	0	3	male	35.0	0	0

Data analysis

In this section I will try to answer the above questions based on the analysis of the Titanic dataset.

1) What was the gender distribution in the group? Is there a correlation between gender and the chances of survival?

In [491]:

```
gender = titanic_data_short['Sex'].value_counts()

print ('There were {} females and {} males on board.'.format(gend
```

```
There were 314 females and 577 males on board.
```

In [492]:

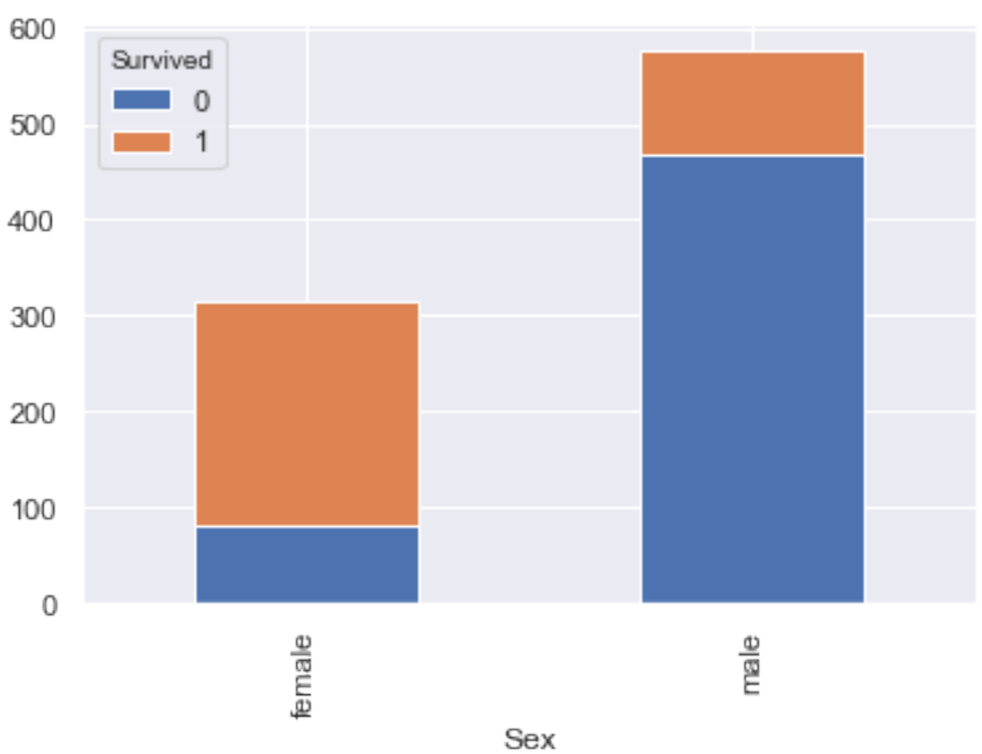
```
group_by_gender_survived = titanic_data_short.groupby(['Sex','Sur
group_by_gender_survived
```

Out[492]:

```
Sex     Survived
female  0            81
        1           233
male    0           468
        1           109
dtype: int64
```

In [493]:

```
# Visualise the distribution of total males and females and those
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
%matplotlib inline
titanic_data_short.groupby(['Sex','Survived']).size().unstack().p
plt.show()
```



In [494]:

```
# Calculate the percentage of survived in a group
# Returns survival rate/percentage of gender
def survival_rate_gender(gender):
    # Take the gender and return the survival rate
    grouped_by_gender = titanic_data_short.groupby(['Sex']).size(
    grouped_by_gender_survived = titanic_data_short.groupby(['Sur
    survival_rate_gender = (grouped_by_gender_survived / grouped_

    return survival_rate_gender
```

In [495]:

```
print('The average survival rate for females was {}%.'.format(sur
print('The average survival rate for males was {}%.'.format(survi
```

```
The average survival rate for females was 74.2%.
The average survival rate for males was 18.89%.
```

Conclusion to Question 1:

Women were much likely to survive than men with the survival rates being 74.2% vs. 18.9%

2) What was the age distribution in the group? Is there a correlation between the age and the chances of survival?

In [16]:

```
print ('We saw before that the information about the age was miss
       format(titanic_data_short['Survived'].count()- titanic_dat
```

```
We saw before that the information about the age was
missing for 177 entries.
```

In [17]:

```
# Identify the missing age entries and remove them from the datas
no_age_entries = pd.isnull(titanic_data_short['Age'])
titanic_data_short[no_age_entries].head()
```

Out[17]:

	Survived	Pclass	Sex	Age	SibSp	Parch
5	0	3	male	NaN	0	0
17	1	2	male	NaN	0	0
19	1	3	female	NaN	0	0
26	0	3	male	NaN	0	0
28	1	3	female	NaN	0	0

In [18]:

```
titanic_data_short_with_age = titanic_data_short.dropna()
titanic_data_short_with_age.head()
```

Out[18]:

	Survived	Pclass	Sex	Age	SibSp	Parch
0	0	3	male	22.0	1	0
1	1	1	female	38.0	1	0
2	1	3	female	26.0	0	0
3	1	1	female	35.0	1	0
4	0	3	male	35.0	0	0

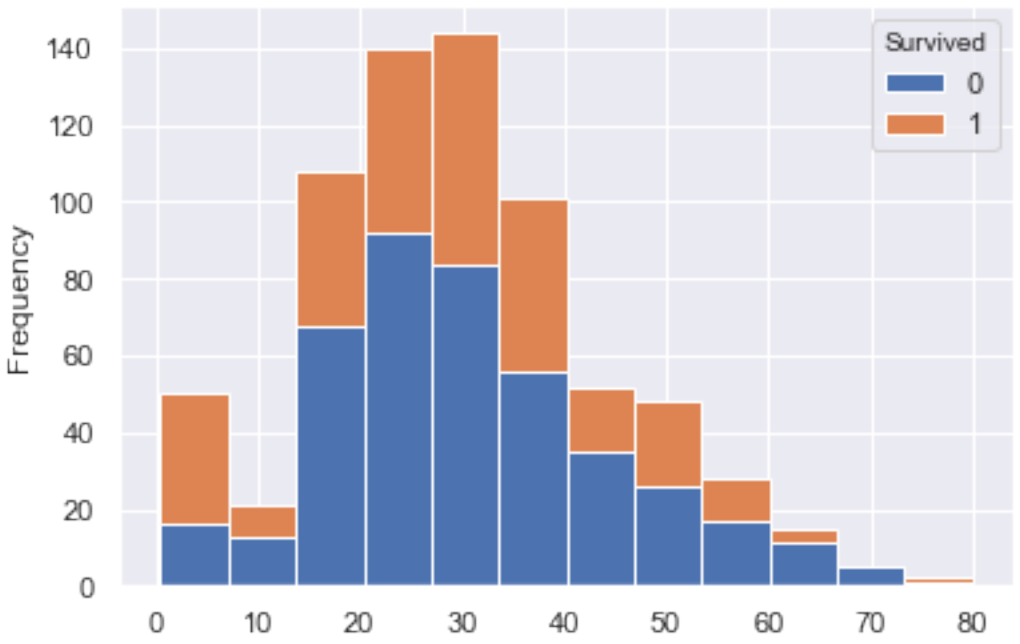
We see that the largest group was from approximately 20 y.o. till approximately 33 y.o.

In [496]:

```
titanic_data_short_with_age.pivot(columns='Survived').Age.plot(ki
```

Out[496]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a25f14b
70>
```

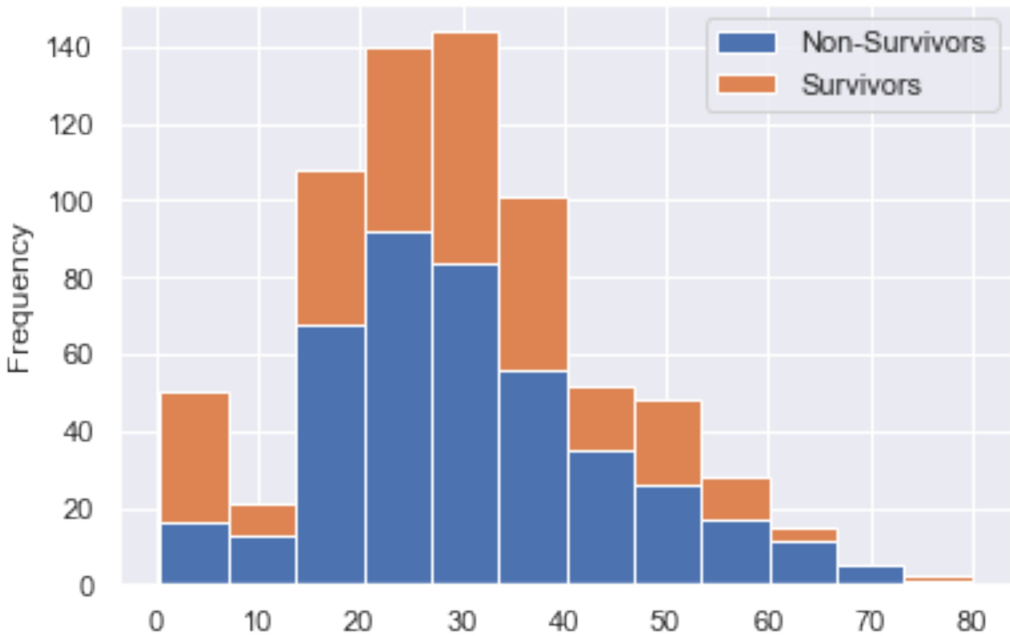


In [27]:

```
pd.DataFrame({'Non-Survivors': titanic_data_short_with_age.groupb
              'Survivors':titanic_data_short_with_age.groupby('Su
```

Out[27]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a24d37c
50>
```

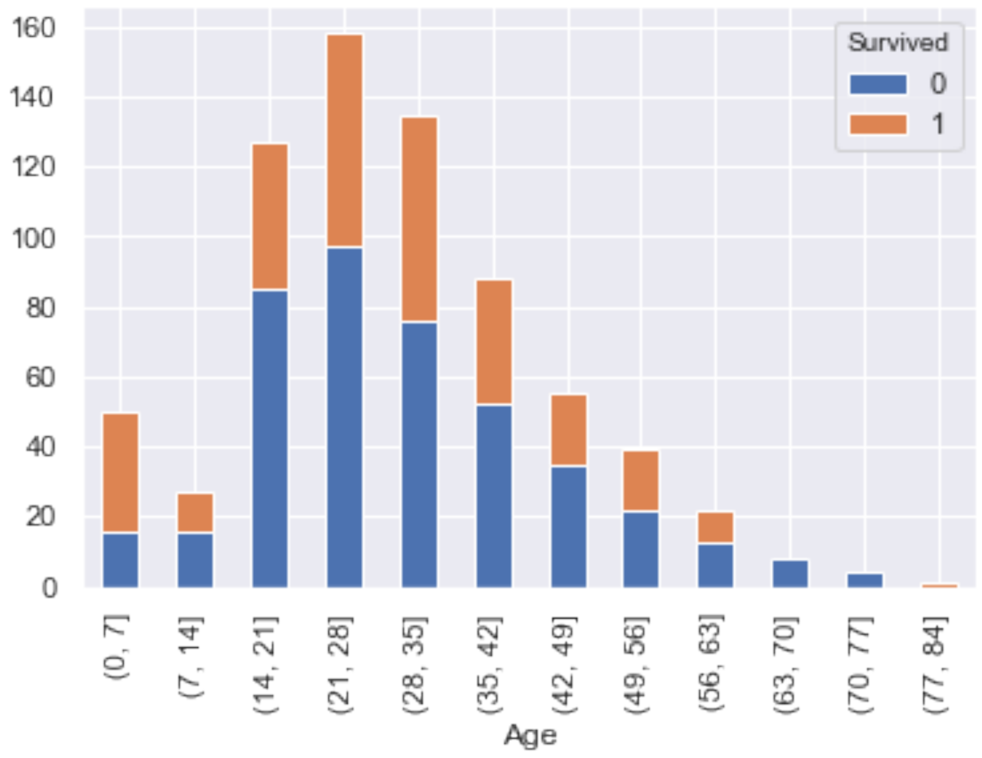


In [28]:

```
titanic_data_short_with_age.groupby(['Survived', pd.cut(titanic_d
        .size().unstack(0).plot.bar(stacked=True)
```

Out[28]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2553e7
80>
```



From the above histograms we see that almost in all age categories the number of survived was lower than the number of drowned, except for young children (0-7 years). In the category from 63 to 77 all passengers died and a very little number above 77 y.o. survived. Let's explore these facts a bit deeper.

First, let's calculated the mean age of those who survived and the mean age of those who did not survive. Then we calculate the percentage of survivors in the age category from 0 to 7 y.o. And try to identify how many persons above 77 years there are who survived.

In [497]:

```
titanic_data_short_with_age.groupby('Survived')['Age'].mean()
```

Out[497]:

```
Survived
0    30.626179
1    28.343690
Name: Age, dtype: float64
```

The mean for those who did not survive (28.34) is very close to those who did not survive (30.62). This prompts me to conclude that, apart for children, there was not much dependence of the survival chances on the age.

In [498]:

```
bins= [0,8,15,64,78,85]
labels = ['Children','Teens','Main','Elder','Oldest']
titanic_data_short_with_age.loc[:,'AgeGroup'] = pd.cut(titanic_da
titanic_data_short_with_age.head()
```

Out[498]:

	Survived	Pclass	Sex	Age	SibSp	Parch	AgeGroup	Category
0	0	3	male	22.0	1	0	Main	Adult
1	1	1	female	38.0	1	0	Main	Adult
2	1	3	female	26.0	0	0	Main	Adult
3	1	1	female	35.0	1	0	Main	Adult
4	0	3	male	35.0	0	0	Main	Adult

In [499]:

```
# Calculate the survaval rate for each age group.
age_groups = titanic_data_short_with_age.groupby(['AgeGroup','Sur
```

In [229]:

```
age_survaval = pd.DataFrame({'Non-Survivors': titanic_data_short_
              'Survivors':titanic_data_short_with_age.groupby('Su
not_survived = titanic_data_short_with_age[titanic_data_short_wit
print('In total {} people did not survive.'.format(not_survived))
```

```
In total 424 people did not survive.
```

In [500]:

```
# The distribution among the age groups is as follows:
agegroup = titanic_data_short_with_age['AgeGroup'].value_counts()
agegroup
```

Out[500]:

```
Main        623
Children     50
Teens        28
Elder        12
Oldest        1
Name: AgeGroup, dtype: int64
```

In [501]:

```
# Return the survival rate for every age group
def survival_rate_age(age_group):
    # Take the age group and return the survival rate
    age_group_total = agegroup[age_group].astype('float')
    try:
        grouped_by_age_survived = titanic_data_short_with_age.grc
        return (grouped_by_age_survived / age_group_total * 100).
    except KeyError as exc:
        return 0
```

In [502]:

```
print('The average survival rates were as follows:')
print('for children - {}%.'.format(survival_rate_age('Children'))
print('for teens - {}%.'.format(survival_rate_age('Teens')))
print('for adults - {}%.'.format(survival_rate_age('Main')))
print('for elderly people - {}%.'.format(survival_rate_age('Elder
print('for seniors - {}%.'.format(survival_rate_age('Oldest')))
```

```
The average survival rates were as follows:
for children - 68.0%.
for teens - 39.29%.
for adults - 39.17%.
for elderly people - 0%.
for seniors - 100.0%.
```

Conclusion to Question 2:

Children under 7 y.o. were the most likely to survive with the survival rate being 68%. The survival rate for seniors of 100.0% is incidental. There was only one senior person in the group and this person survived.

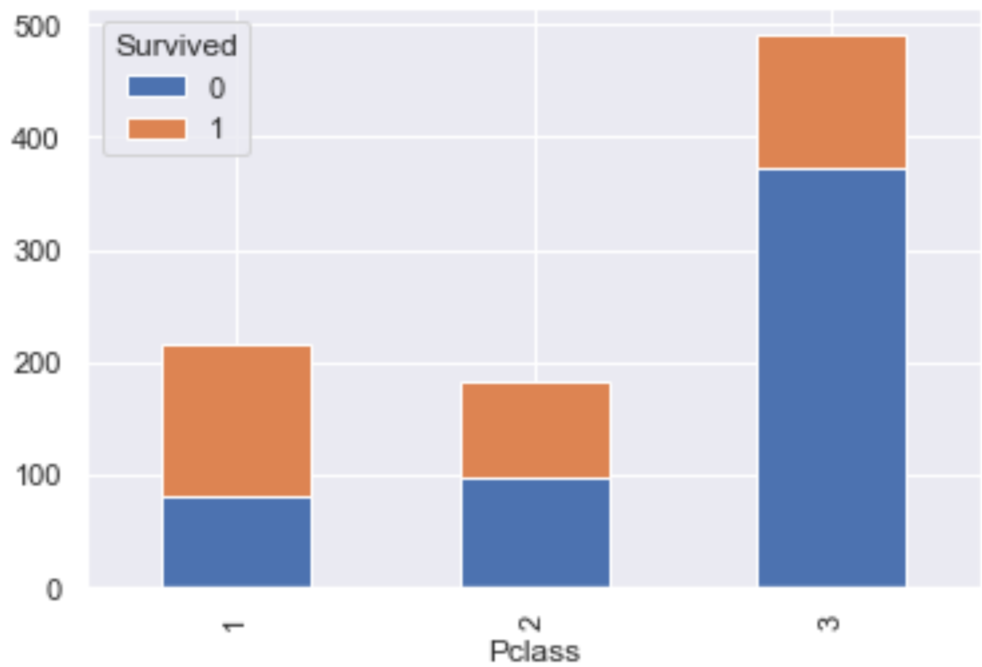
However, in the group of elderly people from 64 to 77 y.o. nobody survived. So I will make a tentative conlcusion that for older people the changes of survival were very low.

Question No. 3

What was the travel class distribution in the group? Is there a correlation between the travel class and the chances of survival?

In [476]:

```
# Visualise the distribution of total passenders by travel class
titanic_data_short.groupby(['Pclass','Survived']).size().unstack(
plt.show()
```



In [477]:

```
print ('There were {} 1st class passengers, {} 2nd class passenge
       .format(distribution_by_class[1], distribution_by_class[2]
```

```
There were 216 1st class passengers, 184 2nd class p
assengers and 491 3rd class passengers on board.
```

In [474]:

```
# Calculate the percentage of survived in a given travel class
# Returns survival rate for the travel class
def survival_rate_tclass(tclass):
    # Take the travel class and return the survival rate
    grouped_by_tclass = titanic_data_short.groupby(['Pclass']).si
    grouped_by_tclass_survived = titanic_data_short.groupby(['Sur
    survival_rate_tclass = (grouped_by_tclass_survived / grouped_

    return survival_rate_tclass
```

In [478]:

```
print('Out of the total of {} passengers, the survaval rates depe
      .format(len(titanic_data_short)))
for i in range(1, 4):
    print('For travel class {} passenders: {}%'.format(i, surviva
```

```
Out of the total of 891 passengers, the survaval rat
es depending on the travel class were:

For travel class 1 passenders: 62.96%
For travel class 2 passenders: 47.28%
For travel class 3 passenders: 24.24%
```

Conclusion to Question 3:

The passengers in the 1st travel class had the highest survival rate of almost 63%, while the passengers from the 3rd travel class had the lowest chances to survive of a little over 24%.

Question No. 4

Is there a correlation between the existence of children/parents and the chances of survival? Is there a correlation between the existence of siblings/spouses and the chances of survival?

```
At first I will divide the population into those who had no
children/parents and those who had children/parents, and
calculate the survival rates for each group.
```

In [342]:

```
bins = [0, 0.5, 10]
labels = ['alone','par_chil']
titanic_data_short.loc[:,'Relations'] = pd.cut(titanic_data_short
titanic_data_short.tail()
```

Out[342]:

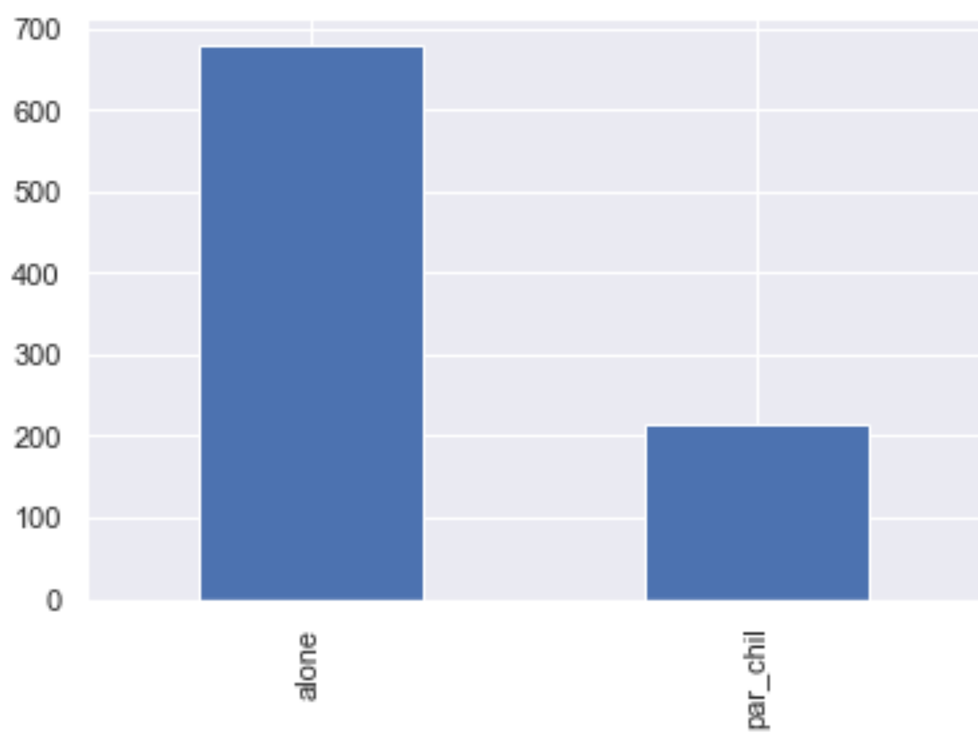
	Survived	Pclass	Sex	Age	SibSp	Parch	Relations
886	0	2	male	27.0	0	0	alone
887	1	1	female	19.0	0	0	alone
888	0	3	female	NaN	1	2	par_chil
889	1	1	male	26.0	0	0	alone
890	0	3	male	32.0	0	0	alone

In [343]:

```
distribution_by_relations = titanic_data_short['Relations'].value
distribution_by_relations.plot(kind = 'bar')
```

Out[343]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a265537
b8>
```



In [344]:

```
print ('The numbers of passengers travelling alone or with parent
       format(distribution_by_relations))
```

```
The numbers of passengers travelling alone or with p
arents/children are:
alone        678
par_chil     213
Name: Relations, dtype: int64
```

In [345]:

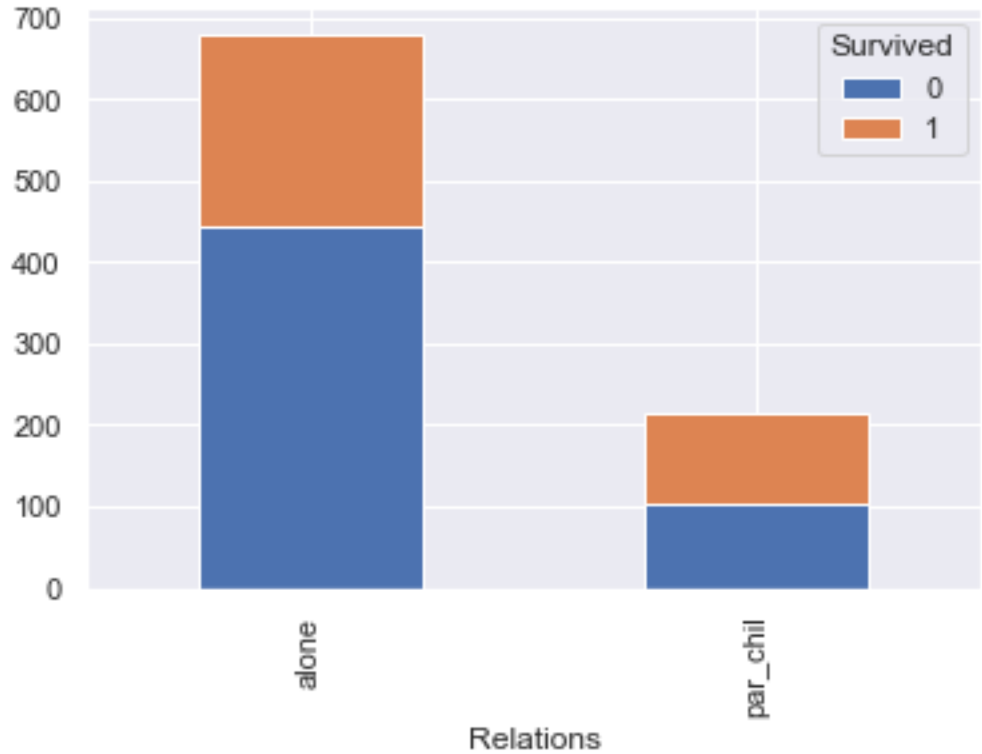
```
# Calculate the survided in each group
group_by_relations = titanic_data_short.groupby('Relations', as_i
group_by_relations['Survived'].sum()
```

Out[345]:

	Relations	Survived
0	alone	233
1	par_chil	109

In [346]:

```
# Visualise the distribution of total passenders by relations and
titanic_data_short.groupby(['Relations','Survived']).size().unsta
plt.show()
```



In [372]:

```
print ('There were {} passengers travelling alone and {} passenge
       .format(alone, par_chil))
```

```
There were 678 passengers travelling alone and 213 p
assengers travelling with parents or children on boa
rd.
```

In [429]:

```
# Calculate the percentage of survived depending on family relati
# Returns survival rate for the given family category
def family_survival_rate(category):
    # Take the family category (alone or with parents/children) a
    grouped_by_relation = titanic_data_short.groupby(['Relations'
    grouped_by_relation_survived = titanic_data_short.groupby(['S
    survival_rate_relation = (grouped_by_relation_survived / grou

    return survival_rate_relation
```

In [430]:

```
print('Out of the total of {} passengers, the survaval rates depe
      .format(len(titanic_data_short)))
print('For passenders travelling alone the survival rate was: {}%
print('For passenders travelling with parents/children the surviv
```

```
Out of the total of 891 passengers, the survaval rat
es depending on the family relations were:

For passenders travelling alone the survival rate wa
s: 34.37%
For passenders travelling with parents/children the
survival rate was: 51.17%
```

Conclusion to Question 4: Passengers travelling with children/parents were more likely to survive (51%) than passengers travelling alone (34%).

Question No. 5: Who had the maximum chances of survival in terms of gender/age/class/existence of children/parents?

From the above analysis and histograms we already saw that women, 1st class passengers and children were more likely to survive than other passengers. Analysing the chances of survivel for elderly people does not make much sense as we don't have a representative group of elderly people and saw that most of them died.

So, for the purposes of this analysis I will compare the survival rate for people with most chances to survive: women (of the main age category) and children travelling in the 1st class to the survival rate for men (from the main age category) and children travelling in the 3rd class.

In [431]:

```
titanic_data_short_with_age.groupby(['AgeGroup', 'Sex', 'Pclass',
```

Out[431]:

```
AgeGroup  Sex     Pclass  Survived
Children  female  1       0             1
                  2       1             7
                  3       0             5
                          1            11
          male    1       1             2
                  2       1             8
                  3       0            10
                          1             6
Teens     female  1       1             1
                  2       1             3
                  3       0             9
                          1             2
          male    1       1             1
                  2       1             1
                  3       0             8
                          1             3
Main      female  1       0             2
                          1            81
                  2       0             6
                          1            58
                  3       0            41
                          1            34
          male    1       0            54
                          1            36
                  2       0            82
                          1             6
                  3       0           194
                          1            29
Elder     male    1       0             7
                  2       0             2
                  3       0             3
Oldest    male    1       1             1
dtype: int64
```

In [445]:

```
# Return survival rate depending on the agegroup, travel class ar
def specific_survival_rate(agegroup, pclass, sex):

    titanic_data_grouped_total = \
    titanic_data_short_with_age.groupby(['AgeGroup','Pclass', 'Se
    try:
        titanic_data_grouped_survived = \
    titanic_data_short_with_age.groupby(['AgeGroup','Pclass','Sur
        return (titanic_data_grouped_survived / titanic_data_grou
    except KeyError as exc:
        return 0

# Return survival rate for children depending on the travel class
def children_survival_rate(pclass, agegroup = 'Children'):
    titanic_data_grouped_total = \
    titanic_data_short_with_age.groupby(['Pclass','AgeGroup']).si
    try:
        titanic_data_grouped_survived = \
    titanic_data_short_with_age.groupby(['Pclass','AgeGroup','Sur
        return (titanic_data_grouped_survived / titanic_data_grou
    except KeyError as exc:
        return 0
```

Now that we have all the functions ready, let's print the survival rates of the categories of interest to us.

In [463]:

```
for i in range(1,4):
    print ('The survival rate for women travelling in class {} wa
for i in range(1,4):
    print ('The survival rate for children travelling in class {}
for i in range(1,4):
    print ('The survival rate for men travelling in class {} was:
```

```
The survival rate for women travelling in class 1 wa
s: 97.59.
The survival rate for women travelling in class 2 wa
s: 90.62.
The survival rate for women travelling in class 3 wa
s: 45.33.
The survival rate for children travelling in class 1
was: 66.67.
The survival rate for children travelling in class 2
was: 100.0.
The survival rate for children travelling in class 3
was: 53.12.
The survival rate for men travelling in class 1 was:
40.0.
The survival rate for men travelling in class 2 was:
6.82.
The survival rate for men travelling in class 3 was:
13.0.
```

Conclusion to Question 5: Women travelling in the first and second classes had the highest survival rate - from 90.6 to over 97%.

Children in the second class were rescued all and children in the 3rd class had a comparable survival rate as children in the 1st class. I make a tentative conclusion from this that children were let to the salvation boats on equal terms from all three travel classes.

Men had the lowest survival rate. Even in the 1st travel class the survival rate was 40%, going down to 7% for the 2nd class and 13% for the 3rd class.

Conclusion

Women travelling in the 1st and 2nd classes and all children were let to the salvation boats on a priority basis. No conclusion can be made as to the differences in the terms of salvation of men in the 2nd and 3rd classes. Same holds true about saving children. The dataset doesn't let us reliably conclude why so many children died in the 1st and 3rd classes and every child travelling in the 2nd class survived.