Questions about Dataset

The following questions can be asked regarding the dataset: 1) Does age group determines the chances of survival? 2) Does gender effects the chances of survival? 3) Are salutation of passenger & chances of survival correlated 4) Are passenger class & chances of survival correlated

Data Wrangling

To examine the effect of Age on the chances of survival, let us first load the data and identify if there are any data entries without age being mentioned.

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%pylab inline
```

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

In [3]:
```
data_titanic=pd.read_csv("C:\Users\Prasanna\Desktop\Titanic\data.csv")
```

In [4]:
```
print(data_titanic)
```

In [5]:
```
data_titanic.isnull().any(axis=1)
```

We see that from above results, there are some entries in the dataset where age is not given any number. We should remove such entries and analyse for dependency for age & the chances of survival as shown below

In [6]:
```
data_titanic_age=data_titanic.filter(['Survived','Age'],axis=1).dropna(axis=0)
```

In [7]:
```
print data_titanic_age
```

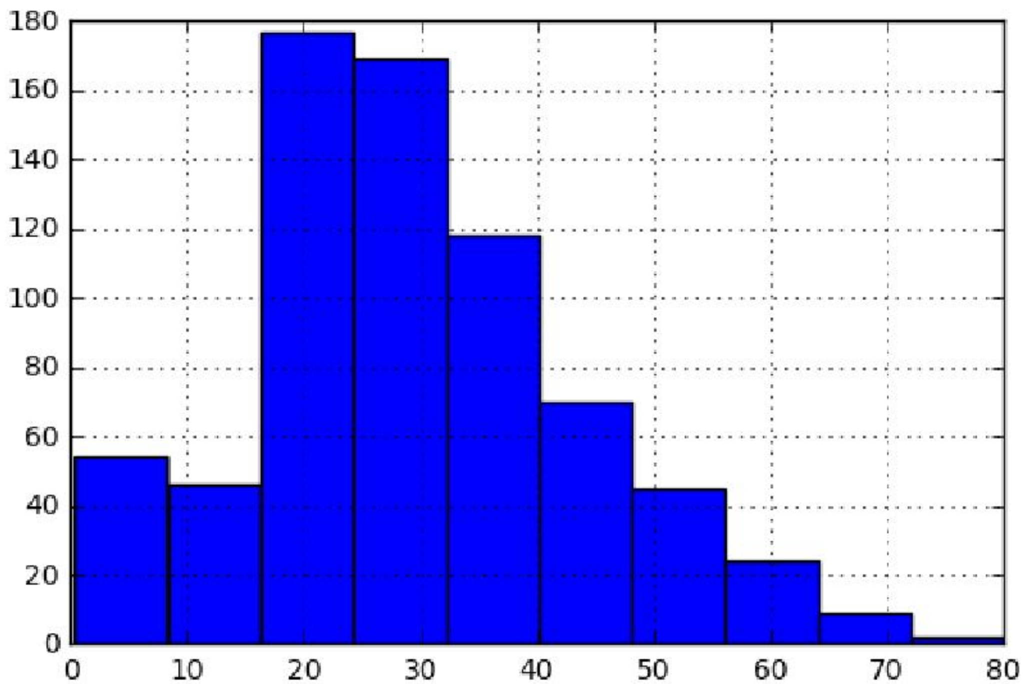
```
     Survived   Age
0           0  22.0
1           1  38.0
2           1  26.0
3           1  35.0
4           0  35.0
6           0  54.0
7           0   2.0
8           1  27.0
9           1  14.0
10          1   4.0
11          1  58.0
12          0  20.0
13          0  39.0
14          0  14.0
15          1  55.0
16          0   2.0
18          0  31.0
20          0  35.0
21          1  34.0
22          1  15.0
23          1  28.0
24          0   8.0
25          1  38.0
27          0  19.0
30          0  40.0
33          0  66.0
34          0  28.0
35          0  42.0
37          0  21.0
38          0  18.0
..        ...   ...
856         1  45.0
857         1  51.0
858         1  24.0
860         0  41.0
861         0  21.0
862         1  48.0
864         0  24.0
865         1  42.0
866         1  27.0
867         0  31.0
869         1   4.0
870         0  26.0
871         1  47.0
872         0  33.0
873         0  47.0
874         1  28.0
875         1  15.0
876         0  20.0
877         0  19.0
879         1  56.0
880         1  25.0
881         0  33.0
882         0  22.0
883         0  28.0
884         0  25.0
```

```
885          0  39.0
886          0  27.0
887          1  19.0
889          1  26.0
890          0  32.0

[714 rows x 2 columns]
```

Let us divide age into 3 groups- Kids, Youngsters & MiddleAge & Old as below: Kids - 0-10 Youngsters- 10-30 MiddleAge - 30-50 Old - 50-_

In [8]:
```
data_titanic_age['Age'].hist()
```

Out[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x889cac8>
```



In [ ]:

In [9]:
```
print data_titanic_age
```

```
     Survived   Age
0           0  22.0
1           1  38.0
2           1  26.0
3           1  35.0
4           0  35.0
6           0  54.0
7           0   2.0
8           1  27.0
9           1  14.0
10          1   4.0
11          1  58.0
12          0  20.0
13          0  39.0
14          0  14.0
15          1  55.0
16          0   2.0
18          0  31.0
20          0  35.0
21          1  34.0
22          1  15.0
23          1  28.0
24          0   8.0
25          1  38.0
27          0  19.0
30          0  40.0
33          0  66.0
34          0  28.0
35          0  42.0
37          0  21.0
38          0  18.0
..        ...   ...
856         1  45.0
857         1  51.0
858         1  24.0
860         0  41.0
861         0  21.0
862         1  48.0
864         0  24.0
865         1  42.0
866         1  27.0
867         0  31.0
869         1   4.0
870         0  26.0
871         1  47.0
872         0  33.0
873         0  47.0
874         1  28.0
875         1  15.0
876         0  20.0
877         0  19.0
879         1  56.0
880         1  25.0
881         0  33.0
882         0  22.0
883         0  28.0
884         0  25.0
```

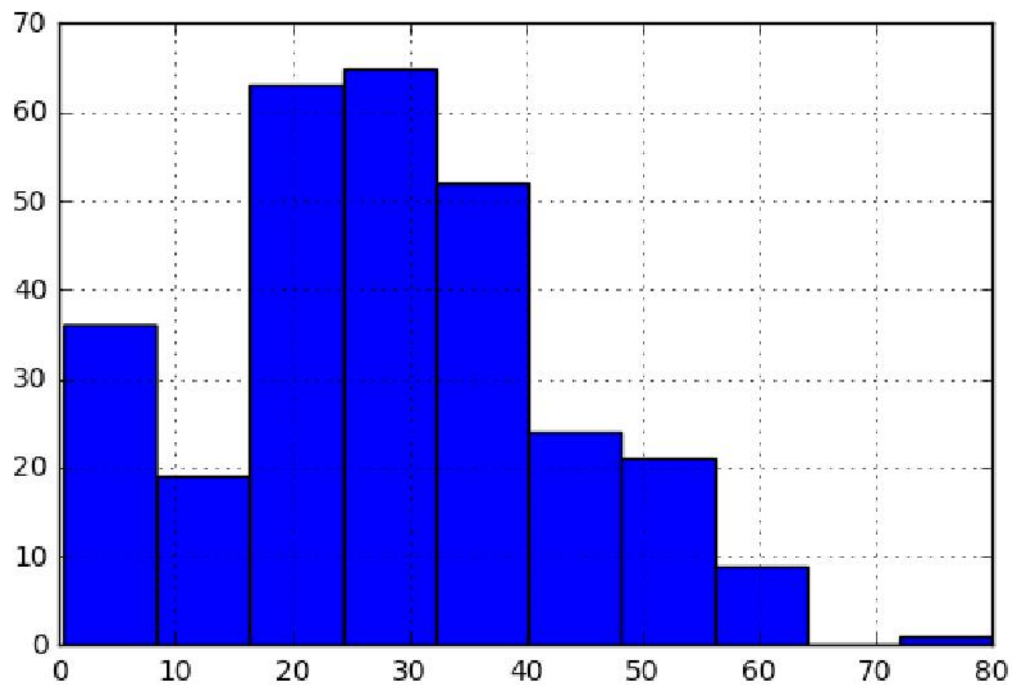
```
885         0  39.0
886         0  27.0
887         1  19.0
889         1  26.0
890         0  32.0

[714 rows x 2 columns]
```

In [10]:
```
data_titanic_age=data_titanic.filter(['Survived','Age'],axis=1).dropna(axis=0)
```

In [11]:
```
data_titanic_age_Survived=data_titanic_age[data_titanic_age['Survived']>0]
```

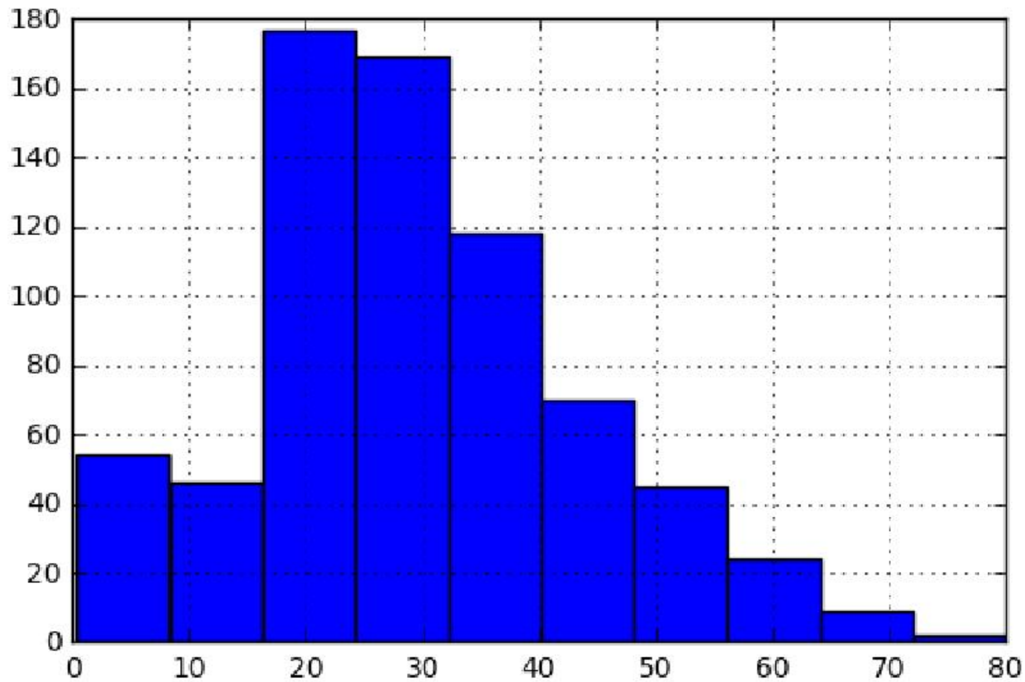
In [12]:
```
data_titanic_age_Survived['Age'].hist()
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x8cae5c0>
```



In [13]:
```
data_titanic_age['Age'].hist()
```

Out[13]: `<matplotlib.axes._subplots.AxesSubplot at 0x8b2ac18>`



In [45]:
```
corrcoef(data_titanic_age['Survived'],data_titanic_age['Age'])
```

Out[45]:
```
array([[ 1.        , -0.07722109],
       [-0.07722109,  1.        ]])
```

Effect of Gender on Chances of Survival

Let us first find total number of males & females

In [14]:
```
Female_total=(sum(data_titanic['Sex']=='female'))
Male_total=(sum(data_titanic['Sex']=='male'))
```

In [15]:
```
Female_Survived=(sum((data_titanic['Sex']=='female') & (data_titanic['Survived']=
Male_Survived=(sum((data_titanic['Sex']=='male') & (data_titanic['Survived']==1))
```

In [16]:
```
Female_Survived_percent=Female_Survived*100/Female_total
Male_Survived_percent=Male_Survived*100/Male_total
```

In [17]:
```
print(Female_Survived_percent)
print(Male_Survived_percent)
```

```
74
18
```

From above calculations, percentage of female survivors is very much larger than that of male survivors. This suggests that the chances of survival for female passengers is relatively higher.

Effect of Title on Survival

Let us first seperate the title from the full name

In [18]:
```
data_titanic['Title']=data_titanic['Name'].str.split(',').apply(pd.Series)[1].str
```

In [19]:
```
data_titanic['TotalCount']=1
```

In [ ]:

In [20]:
```
data_titanic_groupByTitle=data_titanic.groupby(['Title'],as_index=False)['Survive
```

Let us refine more the above data by grouping rare title passengers

In [ ]:

In [21]:
```
data_titanic_groupByTitle[data_titanic_groupByTitle['Survived']<5][['Survived','T
```

Out[21]:
```
Survived      12
TotalCount    27
dtype: int64
```

In [ ]:

In [22]:
```
data_titanic_groupByTitle=data_titanic_groupByTitle[data_titanic_groupByTitle['Su
```

In [23]:
```
data_titanic_groupByTitle.loc[7]=['OtherTitles',12,27]
```

In [24]:
```
print data_titanic_groupByTitle
```

```
          Title  Survived  TotalCount
7   OtherTitles        12          27
8          Miss       127         182
11           Mr        81         517
12          Mrs        99         125
```

In [25]:
```
data_titanic_groupByTitle['PercentSurvived']=(data_titanic_groupByTitle['Survived
```

In [26]:
```
print data_titanic_groupByTitle
```

```
          Title  Survived  TotalCount  PercentSurvived
7   OtherTitles        12          27        44.444444
8          Miss       127         182        69.780220
11           Mr        81         517        15.667311
12          Mrs        99         125        79.200000
```

From the above analysis, there is a high chance that married women passengers would survive than any other passenger

Analysis on Passenger Class versus chance of Survival

Let us group the data by Passenger class and check the total number of passengers survived for each class

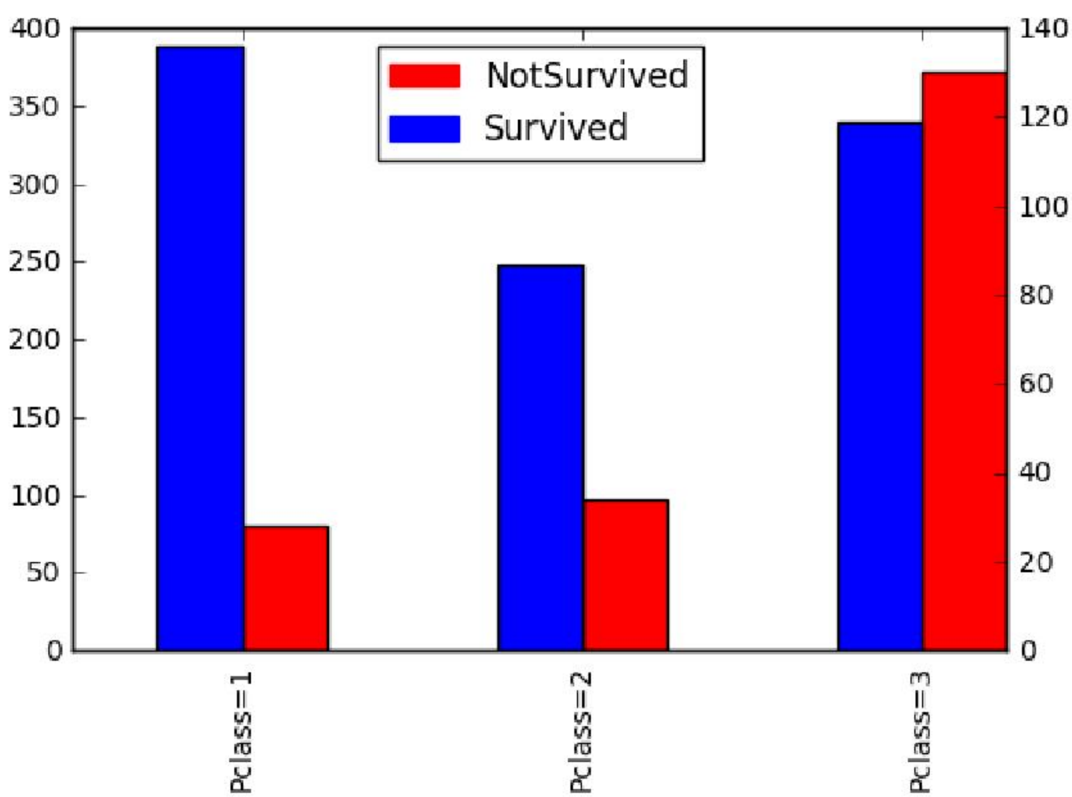
In [27]:
```
import matplotlib.pyplot as plt
data_titanic_Class=data_titanic[['Pclass','Survived','TotalCount']]
print data_titanic_Class
```

	Pclass	Survived	TotalCount
0	3	0	1
1	1	1	1
2	3	1	1
3	1	1	1
4	3	0	1
5	3	0	1
6	1	0	1
7	3	0	1
8	3	1	1
9	2	1	1
10	3	1	1
11	1	1	1
12	3	0	1
13	3	0	1
14	3	0	1
15	2	1	1
16	3	0	1
17	2	1	1
18	3	0	1
19	3	1	1
20	2	0	1
21	2	1	1
22	3	1	1
23	1	1	1
24	3	0	1
25	3	1	1
26	3	0	1
27	1	0	1
28	3	1	1
29	3	0	1
..	...	...	...
861	2	0	1
862	1	1	1
863	3	0	1
864	2	0	1
865	2	1	1
866	2	1	1
867	1	0	1
868	3	0	1
869	3	1	1
870	3	0	1
871	1	1	1
872	1	0	1
873	3	0	1
874	2	1	1
875	3	1	1
876	3	0	1
877	3	0	1
878	3	0	1
879	1	1	1
880	2	1	1
881	3	0	1
882	3	0	1
883	2	0	1
884	3	0	1
885	3	0	1

```
886        2          0            1
887        1          1            1
888        3          0            1
889        1          1            1
890        3          0            1

[891 rows x 3 columns]
```

In [28]:
```
data_titanic_groupByClass=data_titanic_Class.groupby(['Pclass'],as_index=False)['
data_titanic_groupByClass['NotSurvived']=data_titanic_groupByClass['TotalCount']-
```

In [39]:
```
fig = plt.figure()
ax = fig.add_subplot(111)
ax2 = ax.twinx()

data_titanic_groupByClass['NotSurvived'].plot(kind='bar', color='red', ax=ax, pos
data_titanic_groupByClass['Survived'].plot(kind='bar', color='blue', ax=ax2, posi
ax.set_xticklabels(['Pclass=1','Pclass=2','Pclass=3'], minor=False)

import matplotlib.patches as mpatches

NS = mpatches.Patch(color='red', label='NotSurvived')
S = mpatches.Patch(color='blue', label='Survived')
plt.legend(handles=[NS,S], loc=9)
plt.show()
```



As we see from the above plot, the passenger class and the chances of survival are correlated. This might be because the first class passengers are expected to be in the elite group as the ticket fare is high and which in turn might

In [41]:
```
corrcoef(data_titanic_groupByClass['Survived'],data_titanic_groupByClass['Pclass'
```

Out[41]:
```
array([[ 1.        , -0.34164385],
       [-0.34164385,  1.        ]])
```

In [42]:
```
corrcoef(data_titanic_groupByTitle['Survived'],data_titanic_groupByClass['Title']
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-42-b1324c6acc77> in <module>()
----> 1 corrcoef(data_titanic_groupByTitle['Survived'],data_titanic_groupByClas
s['Title'])

C:\Users\Prasanna\Anaconda3\envs\DAND\lib\site-packages\pandas\core\frame.pyc i
n __getitem__(self, key)
   2057             return self._getitem_multilevel(key)
   2058         else:
-> 2059             return self._getitem_column(key)
   2060
   2061     def _getitem_column(self, key):

C:\Users\Prasanna\Anaconda3\envs\DAND\lib\site-packages\pandas\core\frame.pyc i
n _getitem_column(self, key)
   2064         # get column
   2065         if self.columns.is_unique:
-> 2066             return self._get_item_cache(key)
   2067
   2068         # duplicate columns & possible reduce dimensionality

C:\Users\Prasanna\Anaconda3\envs\DAND\lib\site-packages\pandas\core\generic.pyc
in _get_item_cache(self, item)
   1384         res = cache.get(item)
   1385         if res is None:
-> 1386             values = self._data.get(item)
   1387             res = self._box_item_values(item, values)
   1388             cache[item] = res

C:\Users\Prasanna\Anaconda3\envs\DAND\lib\site-packages\pandas\core\internals.p
yc in get(self, item, fastpath)
   3539
   3540             if not isnull(item):
-> 3541                 loc = self.items.get_loc(item)
   3542             else:
   3543                 indexer = np.arange(len(self.items))
[isnull(self.items)]

C:\Users\Prasanna\Anaconda3\envs\DAND\lib\site-packages\pandas\indexes\base.pyc
in get_loc(self, key, method, tolerance)
   2134                 return self._engine.get_loc(key)
   2135             except KeyError:
-> 2136                 return self._engine.get_loc(self._maybe_cast_indexer(ke
y))
   2137
   2138         indexer = self.get_indexer([key], method=method, tolerance=tole
rance)

pandas\index.pyx in pandas.index.IndexEngine.get_loc (pandas\index.c:4443)()

pandas\index.pyx in pandas.index.IndexEngine.get_loc (pandas\index.c:4289)()

pandas\src\hashtable_class_helper.pxi in pandas.hashtable.PyObjectHashTable.get
_item (pandas\hashtable.c:13733)()

pandas\src\hashtable_class_helper.pxi in pandas.hashtable.PyObjectHashTable.get
```

```
_item (pandas\hashtable.c:13687)()

KeyError: 'Title'
```

In [ ]:

In [ ]:

In [ ]: