Analysis of Titanic Data

The Titanic disaster claimed several lives and besides the fact that there were errors made by the captain and crew questions remain as to whether other factors also affected survival. The analysis that follows examines how the independent variables age, gender, and cabin location may have affected the dependent variable survival. Specifically three questions are considered.

- Does the distribution of ages of people who survived differ from the distribution of ages of people who did not?
- What gender was most likely to survive?
- Is there a relationship between cabin location and survival status?

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


## Load Data from CSV
filename = 'titanic_data.csv'
titanic_df = pd.read_csv(filename)
```

In [2]:
```
#Some commands to begin investigating the data

titanic_df.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	N
				Futrelle, Mrs. Jacques							

3	4	1	1	Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N

In [3]:
```
titanic_df.tail()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN

Fixing Data Types and Adding a column

By examining the first and last 5 records of the dataframe we can see that in some cases age is NaN so in analysis this is noted. Also "Sex" is a string field and it may be useful to represent gender numerically. "Cabin" is a combination of deck level and a number which will be a challenge to analyze in this form. The code below will "clean" this data in the following manner to address these issues.

- The column "gender" will be added to the dataframe and if "Sex" is female, gender = 1 and if male then gender = 0
- Any rows that have a cabin specified will have the cabin field split into two components, the first being the letter in position 0 and positions 1 to just before the first space a number. In some of the rows multiple cabins are listed separated by a space. Only the first cabin listed will be used. Two new columns Cabin Deck and Cabin Number will be created in the

resulting dataframe. A few of the rows had a deck in the "Cabin" field but no number as we can't be certain of the exact location.

In [4]:
```
#Create a series with an integer(0 for males, 1 for females)
#rather than True and False, add it to the
#dataframe based on what is in the Sex column
titanic_df["Gender"] = (titanic_df["Sex"] == "female").astype(int)
titanic_df.head(10)
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	N
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	N
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth	female	27.0	0	2	347742	11.1333	N

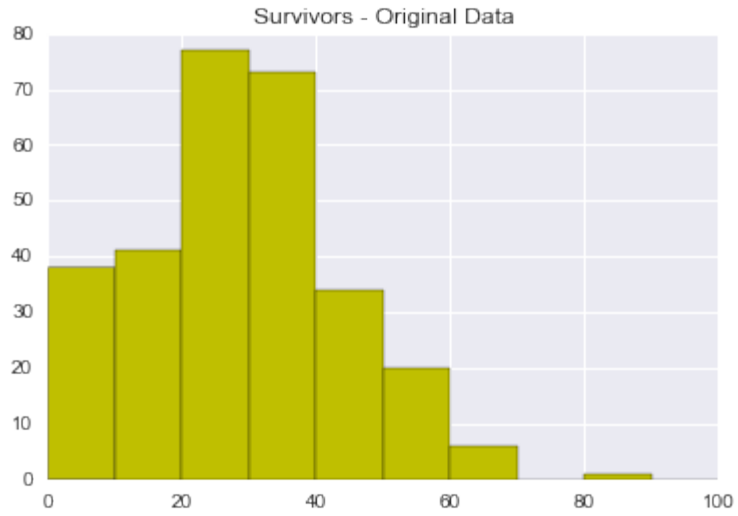
				Vilhelmina Berg)							
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	N

Initial Analysis

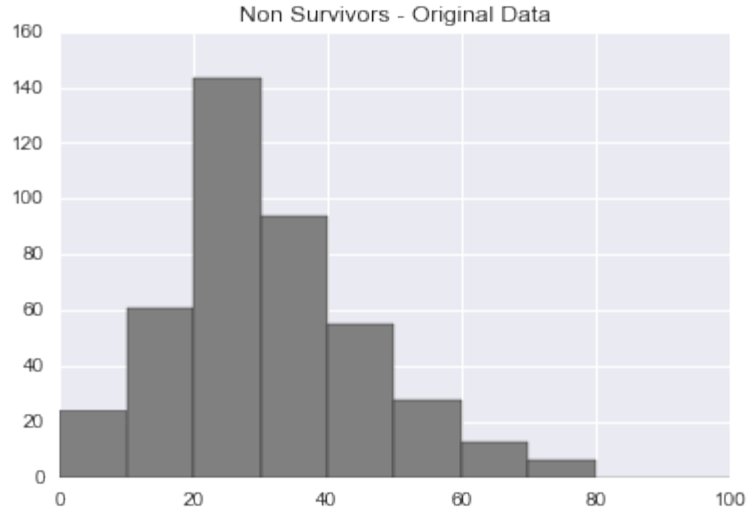
Histograms of Initial Analysis of Age distribution by survival status.

In [5]:
```
#create two data frames: one for survivors and another for non-survivors
survived_df1 = titanic_df.loc[(titanic_df["Survived"] == 1), ["Age"]]
died_df1 = titanic_df.loc[(titanic_df["Survived"] == 0), ["Age"]]
```

In [6]:
```
plt.hist(survived_df1["Age"], range= [0, 100], color = 'y')
plt.title('Survivors - Original Data')
plt.show()
```



In [7]:
```
plt.hist(died_df1["Age"], range= [0, 100], color= 'grey')
plt.title('Non Survivors - Original Data')
plt.show()
```



Question 1:

Does the distribution of ages of people who survived differ from the distribution of ages of people who did not?

This question can be answered at this time before we continue with cleaning the data as we are concerned only with the distribution of ages by survival status. Only individuals whose age is known are included in the analysis.

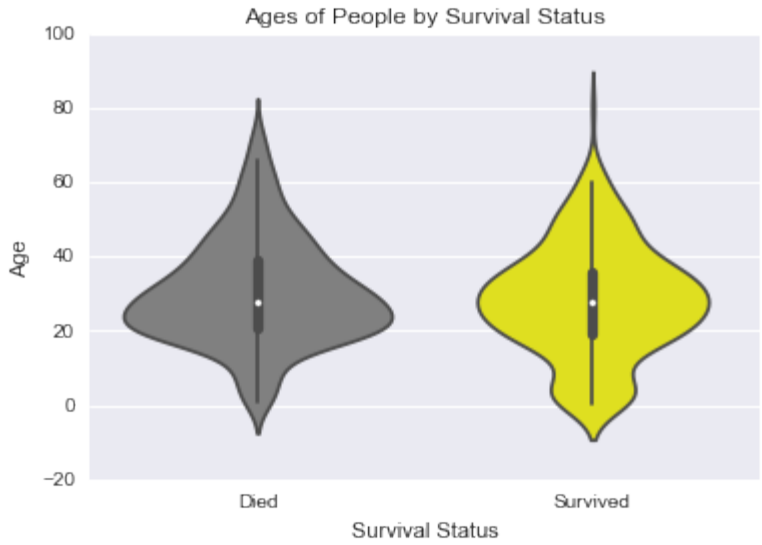
The median is selected over the mean as the measure of center as the histograms indicate some skew on the positive side indicating the median is a better indicator of center.

A graph of violin boxplots was selected to compare the data.

In [8]:
```
#Determine the counts of the individuals included in this
#analysis by survival status
survivor_count = survived_df1[(survived_df1["Age"].notnull())].count()
print "Number of survivors included", survivor_count
died_count = died_df1[(died_df1["Age"].notnull())].count()
print "Number of non survivors included", died_count
excluded_people = len(titanic_df) - survivor_count - died_count
print "Individuals missing", excluded_people
#Compute the median for each data frame
survived_median = survived_df1.median()
died_median = died_df1.median()

print "Survivors Median", survived_median
print "Non Survivors Median", died_median

#Create the violin Plot 'yellow'])

sns.set_palette(["grey", "yellow"])
ax = sns.violinplot(x="Survived", y="Age", data=titanic_df)

ax.set_xlabel('Survival Status')
plt.xticks([0, 1], ("Died", "Survived"))

plt.title('Ages of People by Survival Status')
```

```
plt.show()
```

```
Number of survivors included Age    290
dtype: int64
Number of non survivors included Age    424
dtype: int64
Individuals missing Age    177
dtype: int64
Survivors Median Age    28.0
dtype: float64
Non Survivors Median Age    28.0
dtype: float64
```



891 individuals were included in the data set. Age was missing for 177 of them. The violin plots indicate the overall shape of the distributions by survival status. The medians are the same for each group but the shape of the plots indicate a bit more spread in ages of survivors than non_survivors. It seems of the majority of individuals who died were in their late teens to late 30's. The survivor plot shows a slight increase in the number of survivors in the younger years compared to those who died.

Fixing Data Types Continued

- The following procedure continues with the cleaning described above.

In [9]:
```
#Split up the cabin field if it contains a value

def fix_row(in_row):

    #Deck A is the highest level on the ship so set it to 8
    #because we want the graph to be a rough resemblance of the ship
    deck_map = {'A': 8, 'B': 7, 'C': 6, 'D': 5, 'E':4, 'F':3, 'G':2, 'T':
1}

    #Now split up the cabin field if it is not null
    if not (pd.isnull(in_row["Cabin"])):
```

```
        #first split the cabin up by spaces - only using 1st cabin listed
        cabin = in_row["Cabin"].split()
        deck = cabin[0][0]
        cabin_num = cabin[0][1:]
        #if the cabin number is missing on deck F, assign a room
        #appropriate to general location
        if cabin_num == '' and deck == 'F':
            cabin_num = random.randint(1, 100)

        #if the cabin_number is blank then don't include this row
        #in the data to be graphed
        if cabin_num == '':
            cabin_num = np.NAN
        else:
            in_row["Deck"] = deck_map[deck]
            in_row["CabinNum"] = int(cabin_num)

    return in_row

clean_df = titanic_df.apply(fix_row, axis = 1)
clean_df.head(6)
```

Out[9]:

	Age	Cabin	CabinNum	Deck	Embarked	Fare	Gender	Name	Parch	PassengerId
0	22.0	NaN	NaN	NaN	S	7.2500	0	Braund, Mr. Owen Harris	0	1
1	38.0	C85	85.0	6.0	C	71.2833	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	2
2	26.0	NaN	NaN	NaN	S	7.9250	1	Heikkinen, Miss. Laina	0	3
3	35.0	C123	123.0	6.0	S	53.1000	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	4
4	35.0	NaN	NaN	NaN	S	8.0500	0	Allen, Mr. William Henry	0	5
5	NaN	NaN	NaN	NaN	Q	8.4583	0	Moran, Mr. James	0	6

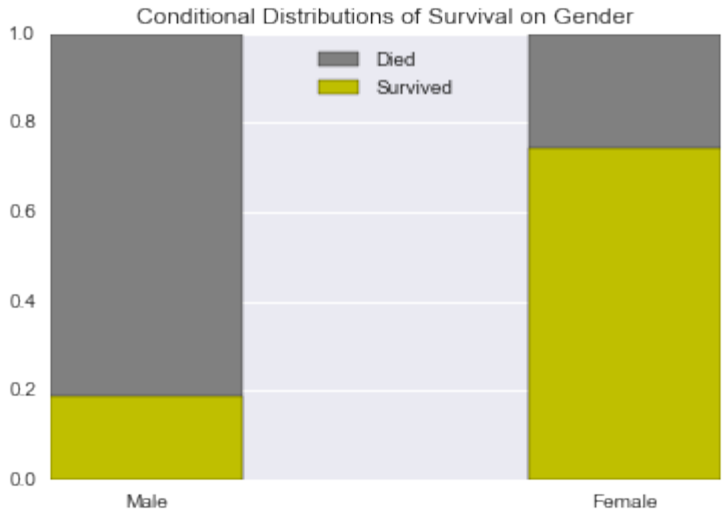
In [10]:
```
#This routine takes as input the column to select on, desired value,
#and column in list form.  It is used to help address the questions.  It re
turns the requested data set.
def get_df(col, col_value, cols):
    return clean_df.loc[(clean_df[col] == col_value), cols]
```

Question 2

What gender was most likely to survive? That is given only a persons gender determine the likelihood that they survived. In order to answer this question two new data frames for each gender were created. For each of these groups calculate the percents who survived. Conditional distributions total 100% and this fact was used to create the bar graphs depicting survival percents by gender.

In [11]:
```
#Split the data into two sets by gender
#select just the Survived column
male_df = get_df("Gender", 0, ["Survived"])
#male_df = clean_df.loc[(clean_df["Gender"] == 0), ["Survived"]]

female_df = get_df("Gender", 1, ["Survived"])
#female_df = clean_df.loc[(clean_df["Gender"] == 1), ["Survived"]]

#Determine the percent who survived
def get_percent(in_df):

    s_count = in_df.loc[(in_df["Survived"] == 1), "Survived"].count()
    s_percent = 1. * s_count/len(in_df["Survived"])
    return s_percent

s_percents = []
#From each of the gender buckets get the percent who survived
s_percents.append(get_percent(male_df))

s_percents.append(get_percent(female_df))

genders = [0 , .25]

d_percents = [1, 1]     #Relative frequencies always total 1 or 100%
#print len(male_df), len(female_df)
print "Male, Female Survival", s_percents

plt.bar(genders, d_percents, width = .1, color = 'grey', tick_label = ["Ma
le", "Female"], align='center')
plt.bar(genders, s_percents, width = .1, color = 'y', tick_label = ["Male"
, "Female"], align ='center')
plt.title("Conditional Distributions of Survival on Gender")
plt.legend(["Died", "Survived"], loc = 9)   #center the legend
plt.show()
```

```
Male, Female Survival [0.18890814558058924, 0.7420382165605095]
```



As indicated by the percents and bar graph it appears a relationship may exist between gender and surival status. Less than 20% of 577 males survived while almost 75% of 314 females survived. To be certain significance test such as a chi-square test of association could be performed. This would show whether or not the relationship between gender and survival is statistically significant but we could not say an individual's gender caused them to survive or die as this is not a controlled experiment but merely a study of existing data.

Question 3

Is there a relationship between cabin location and survival status? A scene in the Hollywood movie, "Titanic", shows a chilling moment when crew of the Titanic lock people in the lower decks. This scene dramatizes the speculation by several that those in lower socio-economic brackets were less likely to survive the disaster. While checking if a relationship between class and survival status would be very useful it also seems worthwile to see if a relationship between cabin location and survival class exists.

Notes

As stated above "Cabin" was separated into two components. The first letter of the cabin corrresponds to the deck and the second the number. This work was done in the procedure fix_row.

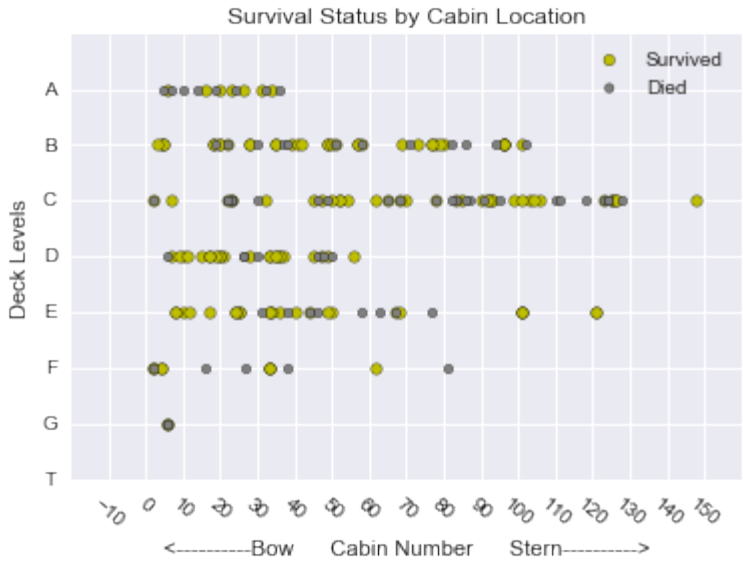
Reviewing images from https://www.encyclopedia-titanica.org/titanic-deckplans/ indicate the following:

- The lower room numbers are located at the bow of the ship.
- "A" deck is at the top of the ship and is exclusively first class. "T" is at the bottom.
- Third class cabins are located closer to the tails of the ship and first class at mid-ship with second class cabins between them.

In [12]:
```
#Call the procedure to create the desired dataframe
survived_df = get_df("Survived", 1, ["Cabin", "CabinNum", "Deck"])
died_df = get_df("Survived", 0, ["Cabin", "CabinNum", "Deck"])

#Select only the rows where cabin is not null
```

```
s_df = survived_df[survived_df["Cabin"].notnull()]
d_df = died_df[died_df["Cabin"].notnull()]

print "Survivors", s_df["Cabin"].count()
print "Non Survivors", d_df["Cabin"].count()

#Create the graph
deck_labels = ["T", "G", "F", "E", "D", "C", "B", "A"]
fig, ax = plt.subplots()
plt.scatter        (s_df["CabinNum"], s_df["Deck"], s=30, c='y')
plt.scatter(d_df["CabinNum"], d_df["Deck"], s=20, c = 'grey')
plt.xticks(range(-10, 160, 10), rotation=-40)
plt.yticks([1,2,3,4,5,6,7,8], deck_labels)
plt.title("Survival Status by Cabin Location")
ax.set_xlabel("<----------Bow      Cabin Number      Stern---------->")
ax.set_ylabel("Deck Levels")
plt.legend(["Survived", "Died"])
plt.show()
```

```
Survivors 136
Non Survivors 68
```



This graph indicates that a relationship between cabin location and survival status doesn't seem to exist but before that claim can be made with certainty more research is required. Part of the issue with determing if in fact a relationship exists is that so few cabin assignments were actually listed in the given data. As indicated with the "Survivors" and "Non Survivors" counts and the graph, cabin numbers are known by more survivors than those who died. This seems to makes sense as survivors could have self reported their cabin number while the deceased of course could not. As stated on (2004) Cabin Allocations Encyclopedia Titanica (ref: #3216, accessed 20th June 2016 06:59:34 PM) URL : https://www.encyclopedia-titanica.org/cabins.html, *"The allocation of cabins on the Titanic is a source of continuing interest and endless speculation. Apart from the recollections of survivors and a few tickets and boarding cards, the only authoritative source of cabin data is the incomplete first class passenger list recovered with the body of steward Herbert Cave."*

Summary

Of the three questions asked the data only seems to support that a relationship between gender and survival status may exist. Age ranges and distributions were very similar to both "Survivors" and "Non Survivors". The lack of information available on "Cabin" location really hindered the ability to say if a relationship does or does not exist with survival status.

In [ ]:
```
##Reference List

(2004) Cabin Allocations Encyclopedia Titanica (ref: #3216, accessed 20th June 2016 06:59:34 PM) URL : https://www.encyclopedia-titanica.org/cabins.html,

https://docs.python.org/2/library/

http://matplotlib.org/

https://www.encyclopedia-titanica.org/

https://www.kaggle.com/c/titanic/data

http://stackoverflow.com

https://stanford.edu/~mwaskom/software/seaborn/generated/seaborn.violinplot.html

http://www.titanicstory.com/shipspec.htm
```

In [ ]:
```

```