1. Introduction

The Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

An analysis will be completed to determine what sorts of people were likely to survive. In particular, tools of machine learning will be used to predict which passengers survived the tragedy.

In [1]:
```
import pandas as pd
import numpy as np
import scipy.stats as sp
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC, LinearSVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
```

In [2]:
```
### I printed the data and changed the Header Names ###
titantic_df = pd.read_csv('titantic.csv')
titantic_df.columns = ['Passenger ID', 'Survivor', 'Passenger Class', 'Name',
'Gender', 'Age', '# of Sibilings/ Spouses Aboard', '# of Parents/ Children Abo
ard', 'Ticket Number', 'Ticket Cost', 'Cabin', 'Port of Embarkation']
titantic_df.head()
```

Out[2]:

	Passenger ID	Survivor	Passenger Class	Name	Gender	Age	# of Sibilings/ Spouses Aboard	# of Parents/ Children Aboard	Ticket Numb
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 175
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/ 310128
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450

In [3]:
```
### Next, I reformated the output results ###
## Removed the ticket number because it's not beneficial to my hypothesis
### S = Southampton, C = Cherbourg, Q = Queenstown
### In Cabin, change NaN to not speficied
### Change Passanger Class from 1 = First Class, 2 = Second Class, 3 = Third C
lass
### Ticket Cost, add dollar sign to output
```

In [5]:
```
titantic_df = titantic_df[["Survivor", "Passenger Class", "Gender", "Age",  "#
 of Sibilings/ Spouses Aboard", "# of Parents/ Children Aboard"]]
titantic_df.describe(include="all")
```

Out[5]:

	Survivor	Passenger Class	Gender	Age	# of Sibilings/ Spouses Aboard	# of Parents/ Children Aboard
count	891.000000	891.000000	891	714.000000	891.000000	891.000000
unique	NaN	NaN	2	NaN	NaN	NaN
top	NaN	NaN	male	NaN	NaN	NaN
freq	NaN	NaN	577	NaN	NaN	NaN
mean	0.383838	2.308642	NaN	29.699118	0.523008	0.381594
std	0.486592	0.836071	NaN	14.526497	1.102743	0.806057
min	0.000000	1.000000	NaN	0.420000	0.000000	0.000000
25%	0.000000	2.000000	NaN	NaN	0.000000	0.000000
50%	0.000000	3.000000	NaN	NaN	0.000000	0.000000
75%	1.000000	3.000000	NaN	NaN	1.000000	0.000000
max	1.000000	3.000000	NaN	80.000000	8.000000	6.000000

In [6]:
```
### Since the Passenger count is 891 but there is only a count of 714 in age,
 that tells me there were 177 without an age listed ###
print titantic_df["Gender"].value_counts()
```

```
male      577
female    314
Name: Gender, dtype: int64
```

In [7]:
```
### I removed the blank ages from the data, which takes the count from 891 to
 714 ###
titantic_df = titantic_df[titantic_df["Age"].notnull()]
len(titantic_df)
```

Out[7]: 714

*** 2. Basic Observations

I'm curious about the single variable exploration to use as a base comparison. I chose to do an overview of the data set before looking at more complex analysis. We will explore:

How many males & females were on the Titantic?

What was the youngest, oldest and average age on the Titantic?

*** How many males and females were on board the Titantic?

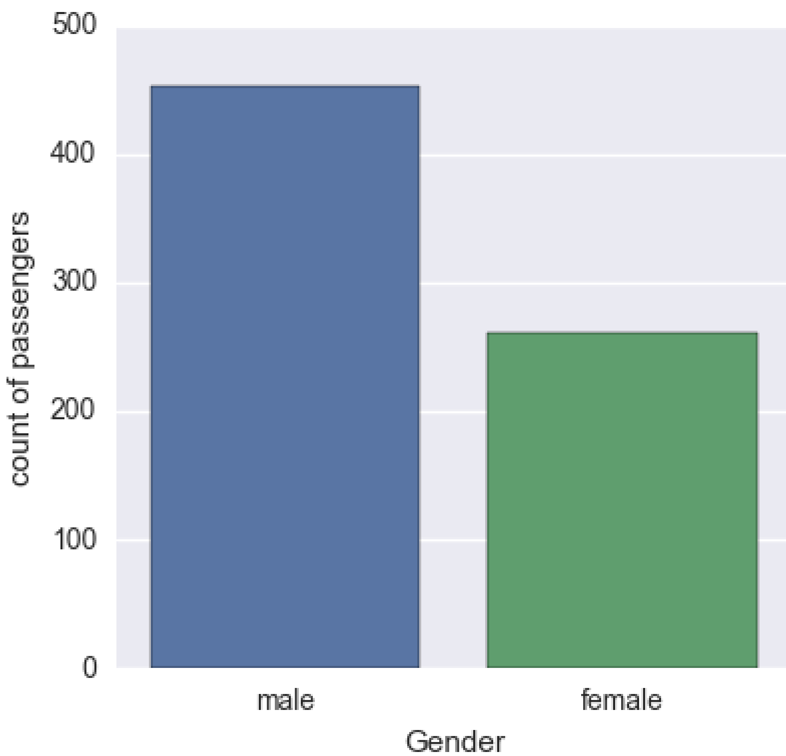
In [8]:
```
menData = titantic_df[titantic_df.Gender == 'male']
womenData = titantic_df[titantic_df.Gender == 'female']
```

In [9]:
```
print("Males: ")
print(menData.count()['Gender'])
print("")
print("Females: ")
print(womenData.count()['Gender'])
```

```
Males: 
453

Females: 
261
```

In [10]:
```
gSSC = sns.factorplot('Gender', data=titantic_df, kind='count')
gSSC.despine(left=True)
gSSC.set_ylabels("count of passengers")
```

Out[10]: <seaborn.axisgrid.FacetGrid at 0xc8abc18>



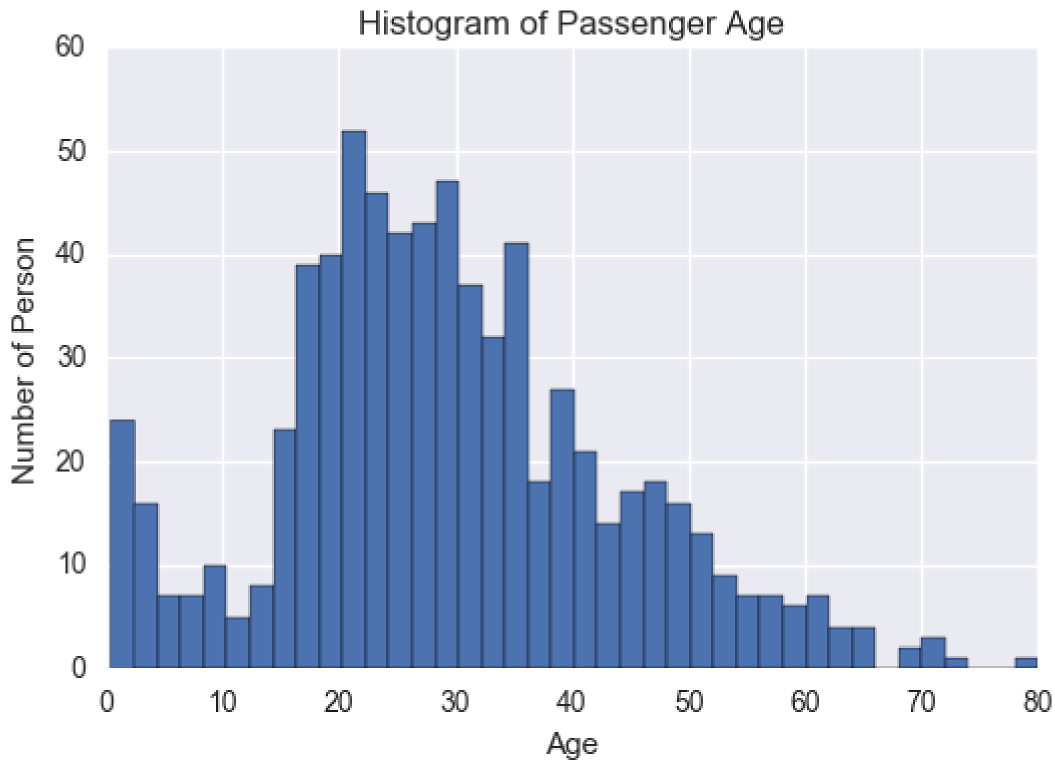
*** Age of Passengers. What is the youngest, oldest and average age of passengers on board the Titantic? What is the distribution of passengers?

In [11]:

```
# Youngest Passenger
print("Youngest Passenger: ")
youngestPassenger = titantic_df['Age'].min()
print(youngestPassenger)
print("")

# Oldest Passenger
print("Oldest Passenger: ")
oldestPassenger = titantic_df['Age'].max()
print(oldestPassenger)
print("")

# Average age of Passenger:
print("Average Age of Passengers: ")
avgPassenger = titantic_df['Age'].mean()
print(avgPassenger)
print("")
```

```
Youngest Passenger: 
0.42

Oldest Passenger: 
80.0

Average Age of Passengers: 
29.6991176471
```

In [12]:
```
titantic_df.Age.hist(bins=40)
plt.xlabel("Age")
plt.ylabel("Number of Person")
plt.title("Histogram of Passenger Age");
```



*** 3A - Women & Children First - Were more women & children survivors?

Women and Children were suppose to get on the lifeboats first. I'm interested in investigating how many women versus men survived and how many adults versus children survived. To do this, I will plot a graph.

I want to see the overview of age based on survivorship. When Survivor = 0, that's the grouping of those who did not survive. This table shows the average age of those who didn't survived was 30.6 years old. Whereas, the average age for the survivors (survivors = 1) was 28.3 years old. Because these stats were similar, I decided to group children as those under 12 and ran another analysis. I will need to define who is a child as those under 12 years old.

In [13]:
```
titantic_df.groupby('Survivor').Age.describe().unstack(level=0)
```

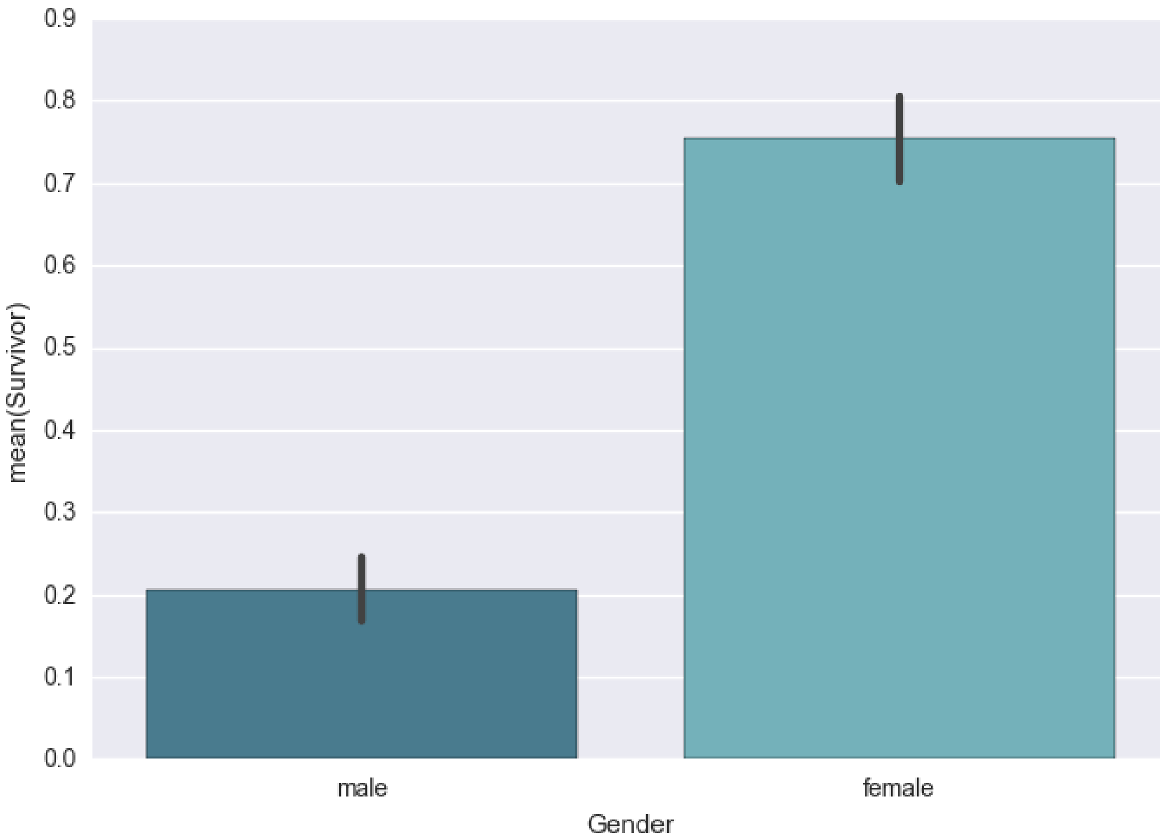
Out[13]:

Survivor	0	1
count	424.000000	290.000000
mean	30.626179	28.343690
std	14.172110	14.950952
min	1.000000	0.420000
25%	21.000000	19.000000
50%	28.000000	28.000000
75%	39.000000	36.000000
max	74.000000	80.000000

In [14]:
```
### I need to define who is a child first ###

def isChild(x):
    if x > 12:
        return "Adult"
    else:
        return "Child, under 12"
titantic_df["IsChild"] = pd.Series(titantic_df["Age"].apply(isChild), index=ti
tantic_df.index)
```

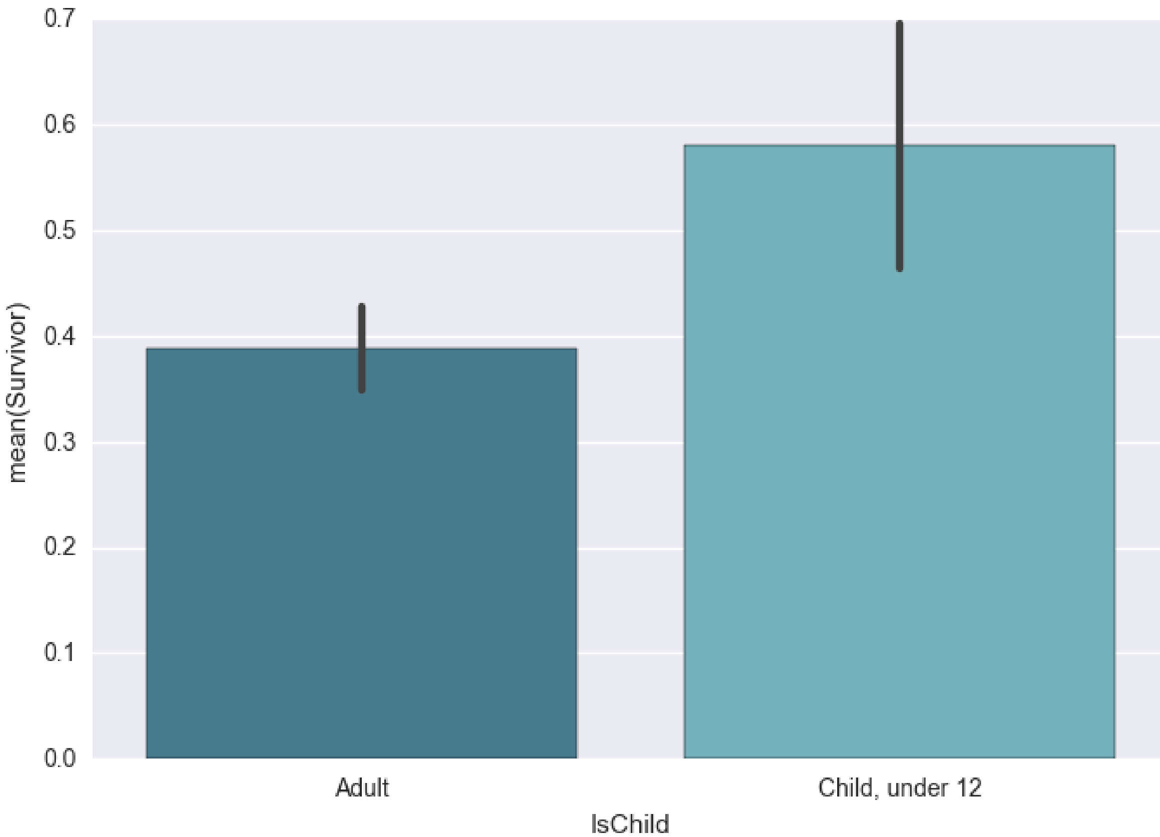
This bar graph shows that yes, more females survived than male.

In [15]:
```
%matplotlib inline
sns.set(style="darkgrid")
sns.barplot(data=titantic_df,x="Gender",y="Survivor", palette="GnBu_d")
sns.plt.show()
```



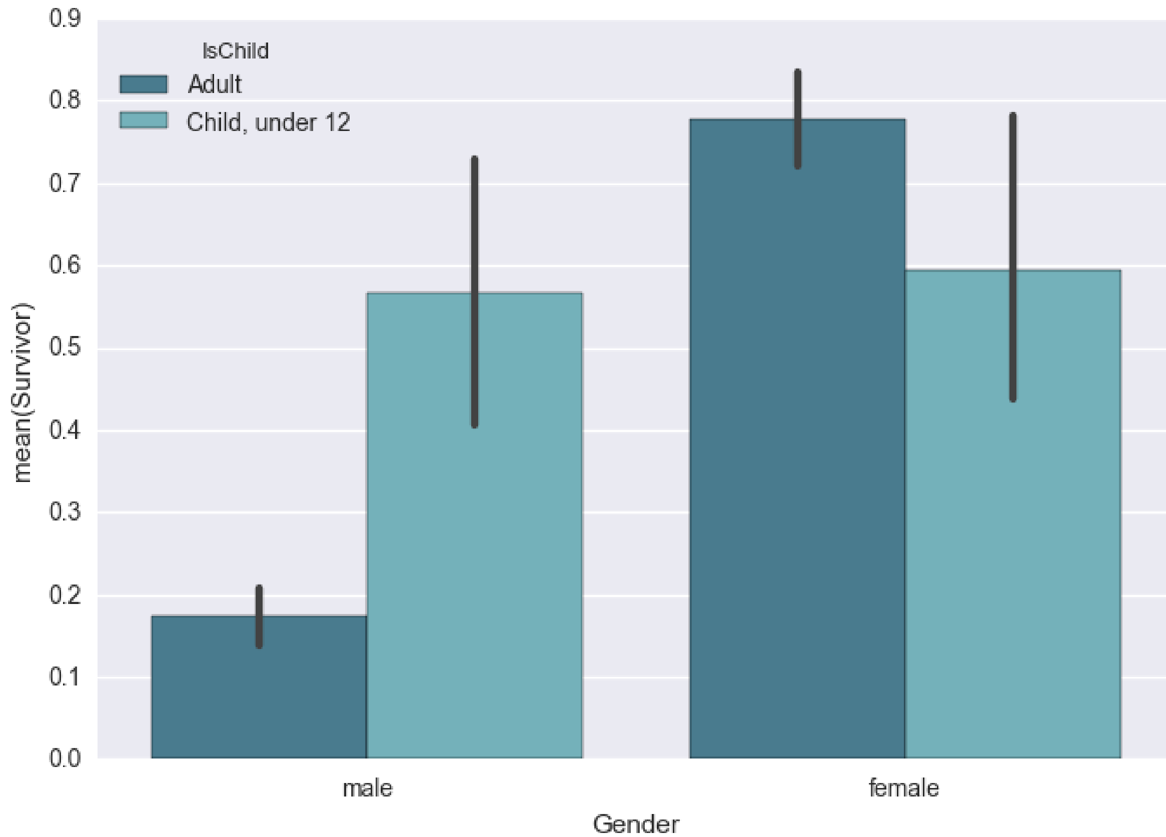
This bar chart shows that yes, if you were a child under 12 years old, you were more likely to survive.

In [16]:
```
%matplotlib inline
sns.set(style="darkgrid")
sns.barplot(data=titantic_df,x="IsChild",y="Survivor", palette="GnBu_d")
sns.plt.show()
```



This doubled bar graph reinforces that both women and children were more likely to survive the Titanic.

In [17]:
```
sns.barplot(data=titantic_df, x="Gender", y="Survivor", hue="IsChild", palette
nBu_d")
sns.plt.show()
```



***3B - Chi Square Test - Goodness of Fit to test women & children data

I decided to test the goodness of fit. We know that there were 2,228 passengers on the Titanic. http://www.titanic-facts.com/passengers-on-the-titanic.html (http://www.titanic-facts.com/passengers-on-the-titanic.html)

Our data file includes information on 891 passengers. We had ages for 714 passengers. Previously, I analyzed the data to give us detail about the raw data. It's possible that the missing values might skew the results. So, I will test if the observed values fit the expected values.

In [18]:
```
titantic_df['WomenChildren'] = np.where((titantic_df.Age <= 12) | (titantic_df
nder == 'female'),1,0)
```

In [19]:
```
def compute_freq_chi2(x,y):

    freqtab = pd.crosstab(x,y)
    print("Frequency table")
    print("============================")
    print(freqtab)
    print("============================")

    chi2,pval,dof,expected = sp.chi2_contingency(freqtab)
    print("ChiSquare test statistic: ",chi2)
    print("p-value: ",pval)
    return
```

In [20]:
```
compute_freq_chi2(titantic_df.Survivor,titantic_df.WomenChildren)
```

```
Frequency table
============================
WomenChildren    0    1
Survivor
0              344   80
1               72  218
============================
('ChiSquare test statistic: ', 222.20201160022424)
('p-value: ', 2.9928112626771852e-50)
```

In the frequency table, it shows the magnitude difference of women and children that survived compared to those who didn't. Since both independent and dependent variable are categorical, I choose the Chi-Square Independece test. For this test to be true, we have to validate the conditions. It's true that all conditions have been met.

Each cell has at least 5 expected cases. Each case only contributes to one cell in the table. If it was a sample, the random sample is less than 10% population; however, this dataset is already a population.

Since we have checked all the condition, we can proceed to the test. And as expected, the chi-square statistic provides a very high number (222.202011), and p value which practically zero. Thus the data provides convincing evidence that whether the passenger woman or children and whether they survived are related.

To take this test farther and to valid the accuracy as predictive model, we observe the accuracy, which tells us that it is 78.7% accurate.

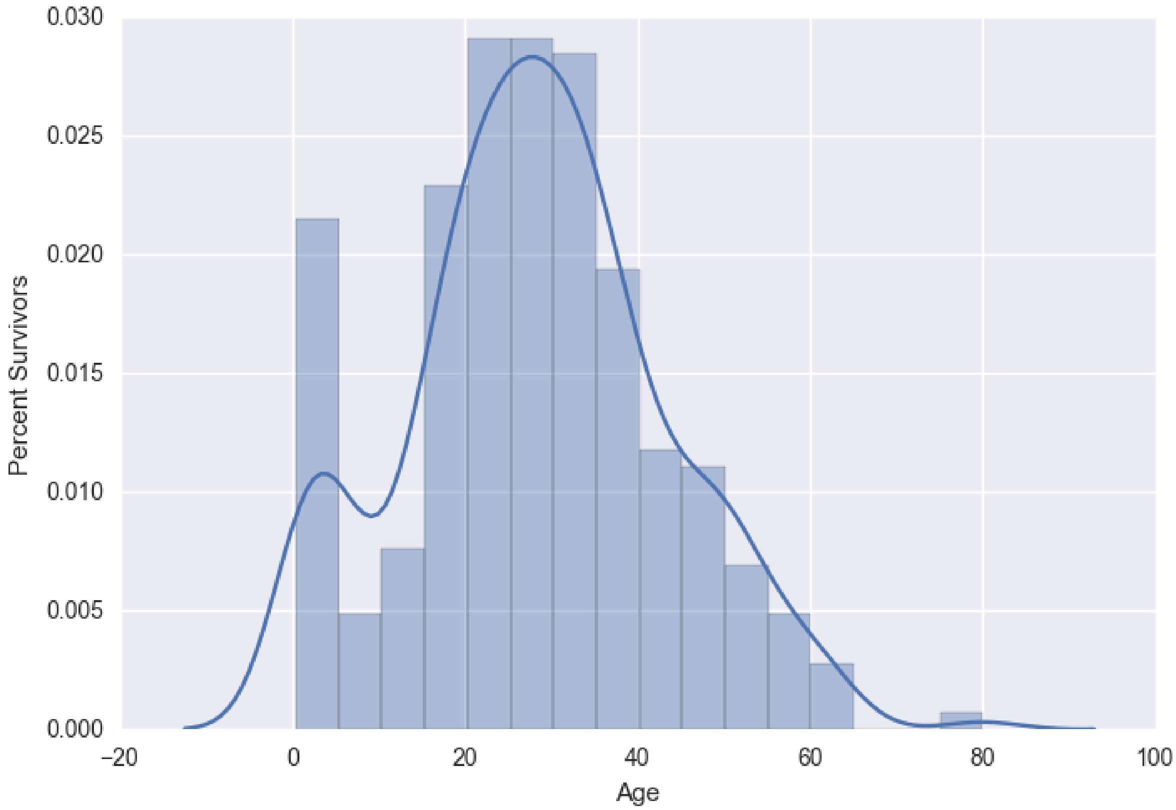
In [21]:
```
(titantic_df['WomenChildren']  == titantic_df.Survivor).mean()
```

Out[21]:
```
0.78711484593837533
```

4. What age was the survivoral rate the highest?

This graph shows the most common survivor age was between 20 and 40 years old. This is most likely because the 20-40 is a common age from the passengers. Above, we learned that the average age was 29 years old.

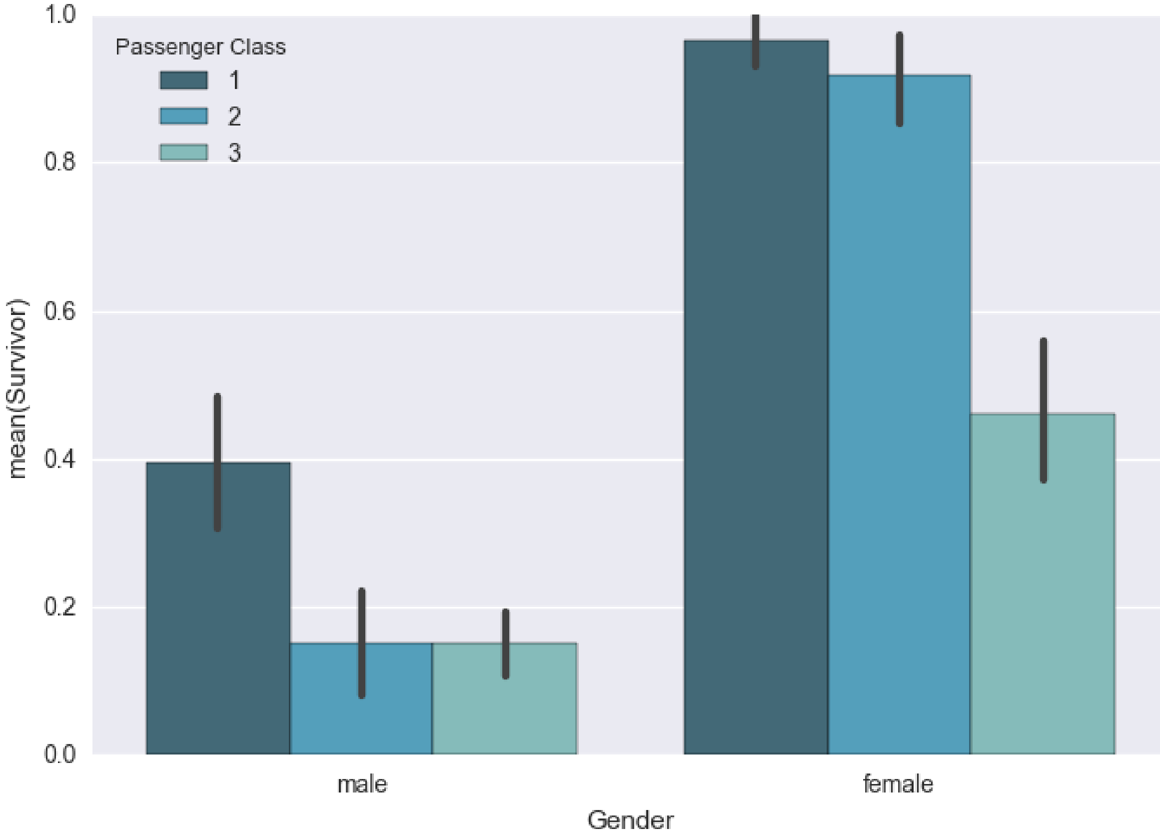
In [22]:
```
survivors_ages = titantic_df[titantic_df["Survivor"] == 1]["Age"]
ax = sns.distplot(survivors_ages)
ax.set(xlabel='Age', ylabel='Percent Survivors')
sns.plt.show()
```



5. If you were in first or second class, did you have a greater likelihood of surviving?

I looked at age a lot; however, I also want to review the classes. This chart shows that you had the great chance of survival if you were a first class passenger. The second class passengers were the 2nd more likely class of passengers to survivor. Therefore, the third class passengers had the lowest amount of survivors.

In [23]:
```
sns.barplot(data=titantic_df, x="Gender", y="Survivor", hue="Passenger Class",
 palette="GnBu_d")
sns.plt.show()
```



Conclusion

The raw data is confusing. The name field doesn't match the gender. For example, The third entry shows "Cumings, Mrs. John Bradley (Florence Briggs Th". It also says she has spouse/sibiling. I would want to ask for a further explaination of the data. For the individuals who have a count in spouse/sibling or a count in children/parents, I'm curious if it is a multiple entry list per cell. If, for example, Mr and Mrs Smith were in a single cell, it would effect the gender and age information.

Initially, I was concerned that the missing age data would effect the Women & Children analysis. The chi test in 3B assured us that our data has a goodness of fit for the women & children analysis. It gave us a p-value of "2.9928112626771852e-50" which is incredible close to zero and tells us that our data is a good fit.

Generally, the results came out as expected. Both women and children were the largest survivor groups in terms of age and gender. This analysis also told us that First Class Passenger were a group of individuals who had the great change of survival.