INVESTIGATE A DATA SET

Awad Bin-Jawed

http://www.titanicuniverse.com/wp-content/uploads/2014/03/Titanic-sinking.jpg

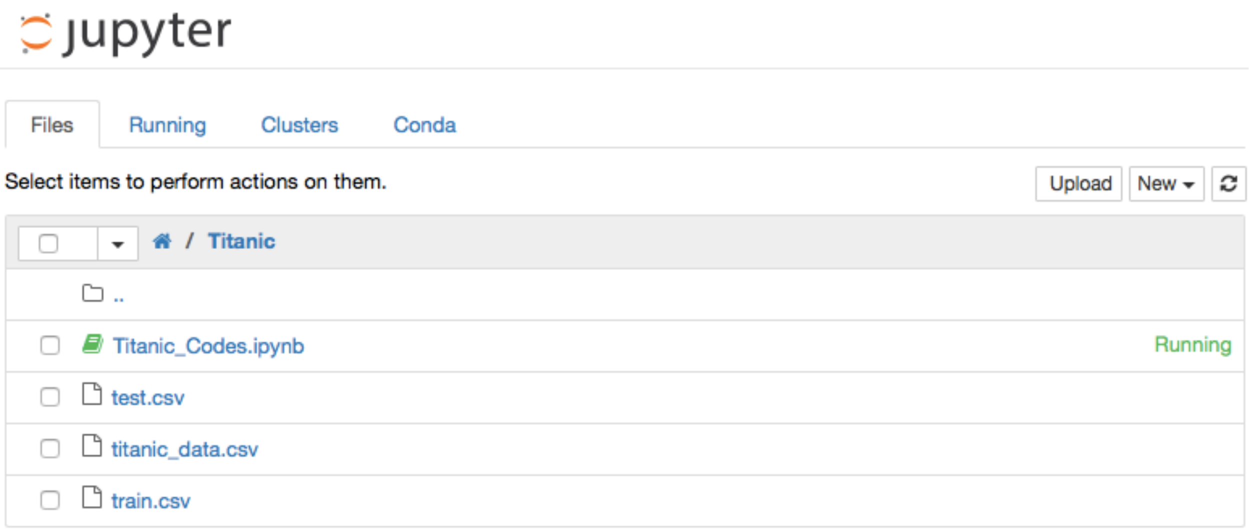
The extensive Titanic database of passenger statistics sparked my curiosity of what affected the fate of those aboard. The factors I wanted to explore were gender, age, and class. Before delving into the provided data, I felt the “factors” that shaped the survival of passengers were nothing more than local customs and ethics of “women and children first” or “noblesse oblige.” A wealthy male passenger who enjoyed affluence and prestige was last priority after a female crew member who washed dishes or cleaned toilets for a living. Then I am reminded of that controversial statement by comedian Bill Burr on why he supports the gender wage gap: “In the unlikely event that we are in a sinking Titanic, for whatever reason, you (women) get to leave with the kids, and I have to stay.” Not that I agree with him, of course. It was, after all, a ridiculous statement, which is why people laugh at it. As for “ethics” and “noblesse oblige”… again, these were my initial perceptions before I looked at the actual data.

<u>QUESTION PHASE</u>: Some interesting questions that occurred to me:

1) “If there were female stewardess and staff members, were they given the same priority as female passengers along with the children?”
2) “Was there an arbitrary age number to determine if a young male is man or child?”
3) “How many emergency boats were there, and how many could each carry?”
4) “Did class make a difference?”

Ideally, the best metrics for would have been the apparent physical attributes, such as gender. This is where I began to divide the cohorts of passengers, but not after first getting the total numbers. I took the raw data of all survivors and compared their classes, and drew conclusions on their likelihoods of survival.

In Anaconda Notebook, I configured the Jupyter Notebook with a folder called “Titanic” in the localhost-8888 option:



Select items to perform actions on them.

/ Titanic

..

Titanic_Codes.ipynb

test.csv

titanic_data.csv

train.csv

Within this notebook, I tested out the scripts of codes under Titanic_Codes.ipynb.

Also, within the Titanic folder, I nested the following CSV files, which I felt were relevant and found from kaggle.com :
-test.csv
-titanic_data.csv
-train.csv

The “test” and “train” files were very helpful in me gaining practice in altering and manipulating the code structure and table alignments. Once I felt confident (without saving changes) I applied them to the “titanic data” file in Excel.

I used Python’s UnicodeCSV library to analyze the CSV files:

I examined titanic_data.csv

First, let’s find the number of passengers aboard total:

```
import unicodecsv


# Read the data from titanic_data.csv and store the results:


def read_csv(filename):
    with open(filename, 'rb') as f:
        reader = unicodecsv.DictReader(f)
        return list(reader)


titanic_data = read_csv(‘titanic_data.csv')
```

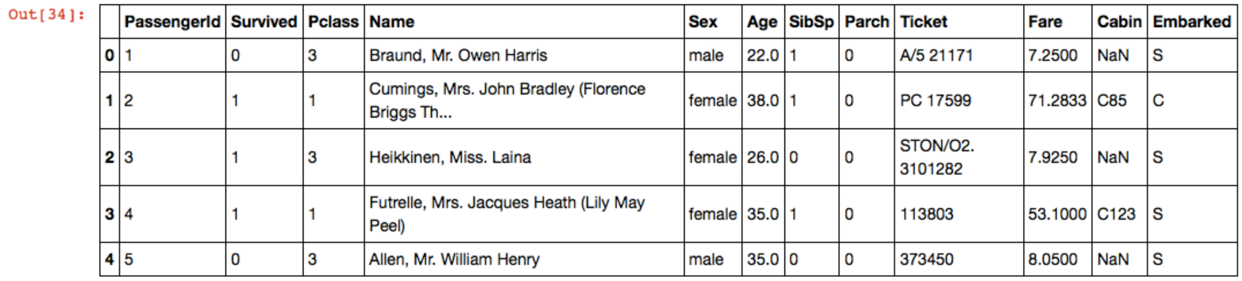
```
len(titanic_data)
Out[1]: 891
```

… so there were a total of 891 passengers aboard.

WRANGLING PHASE:

Now I’m interested in splitting the male passengers from the female passengers:

```
# Import Titanic CSV file and transform it to Pandas Data Frame:
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt

TitanicDF = pd.read_csv('titanic_data.csv')
TitanicDF.head()
```



Out[34]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Per Kaggle, the following legends are variables with specific definitions:

Variable	**Definition**
survival	Survival
pclass	Ticket class
sex	Sex
Age	Age in years
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation

For my own purpose, I don’t need all the columns. The only ones I’m interested in are:

- Survival
- Pclass
- Sex
- Age

In order to minimize the amount of clutter, I won’t be looking at:

- SibSp
- Parch
- Ticket
- Fare
- Embarked
- passengerId
- Name
- Cabin

```
# Dropping Columns:
 columns = [“PassengerID”, “Name”, “SibSp”, “Parch”, “Ticket”, “Fare”, “Embarked”, “Cabin”]
TitanicDF.drop(columns, axis = 1, inplace = True, errors = ‘ignore’)

print(TitanicDF.head( ) )
```

```
  Survived  Pclass     Sex  Age
0        0       3    male   22
1        1       1  female   38
2        1       3  female   26
3        1       1  female   35
4        0       3    male   35
```

```
# Save the Data to CSV for future reference if desired:
TitanicDF.to_csv(‘shortTitanic_data.csv’ , index = False, header = True)
```

EXPLORE PHASE:

Step 1: Define male and female variables:

```
# Set variables for both genders:
malepassengers = TitanicDF[TitanicDF.Sex == 'male']
femalepassengers = TitanicDF[TitanicDF.Sex == 'female']
```

Step 2: print the counts for the male and female passengers:

```
# Get the number of males:
print(‘males’)
print(malepassengers.count()[‘Sex’])

# Get the number of females:
print(‘females’)
print(femalepassengers.count()[‘Sex’])
```

```
males
577
females
314
```

We see that the majority of passengers were males, so could the disproportionately higher death rate for males simply be due to a disproportionately higher male count? First of all, what was the actual survival rate for the male passengers? We look at two factors: their **sex** (**male** & **female**) and if they **survived** (**yes** or **no**).

```
# Get the comparative probability of survival for both genders:
maleprobability = malepassengers.groupby('Sex').Survived.mean()
femaleprobability = femalepassengers.groupby('Sex').Survived.mean()

malepercent = maleprobability[0]*100
femalepercent = femaleprobability[0]*100
```

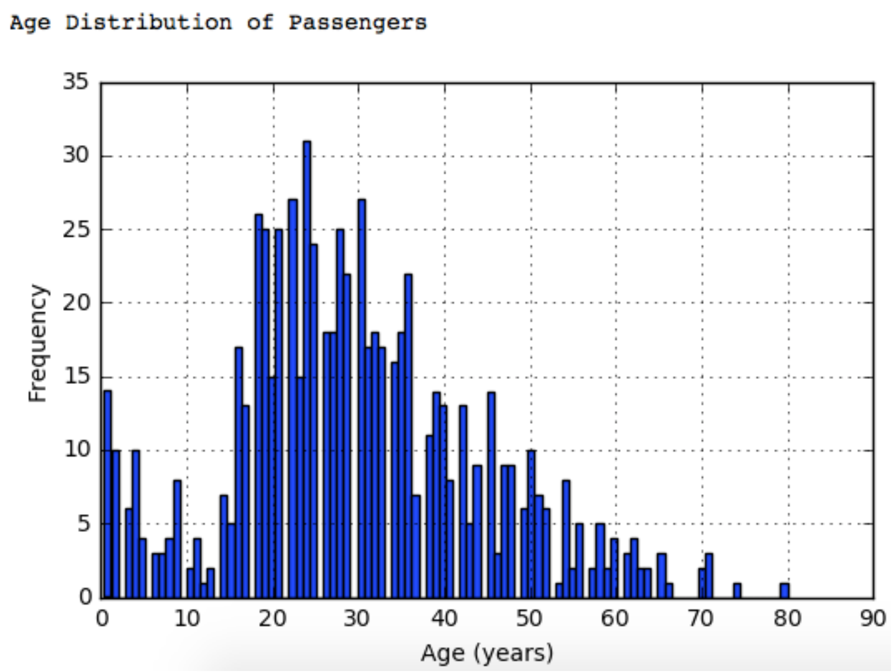
```
print(maleprobability[0])
print(femaleprobability[0])
```

```
0.188908145581
0.742038216561
```

… so we see that although the female passengers were outnumbered by the male passengers (577 to 314), their survival rate was immensely higher (~74% compared to ~19%).

Let’s get a visual aid of the age distribution with an x-axis range between 0 to 90. There weren’t any passengers on board past the age of 100:

```
# Display the frequency for age of passengers:
TitanicDF['Age'].hist(bins = 100)
print "Age Distribution of Passengers"
plt.ylabel("Frequency")
plt.xlabel("Age (years)")
plt.show( )
```



```
Age Distribution of Passengers
```

… based on the plot, we can get a sense of the distribution being heavily concentrated between 20 and 35… let’s calculate the numbers:

```
# For the max, min and mean of passenger ages:
youngestPassenger = TitanicDF['Age'].min()
print(youngestPassenger)


oldestPassenger = TitanicDF['Age'].max()
print(oldestPassenger)


avgPassenger = TitanicDF['Age'].mean()
print(avgPassenger)
```

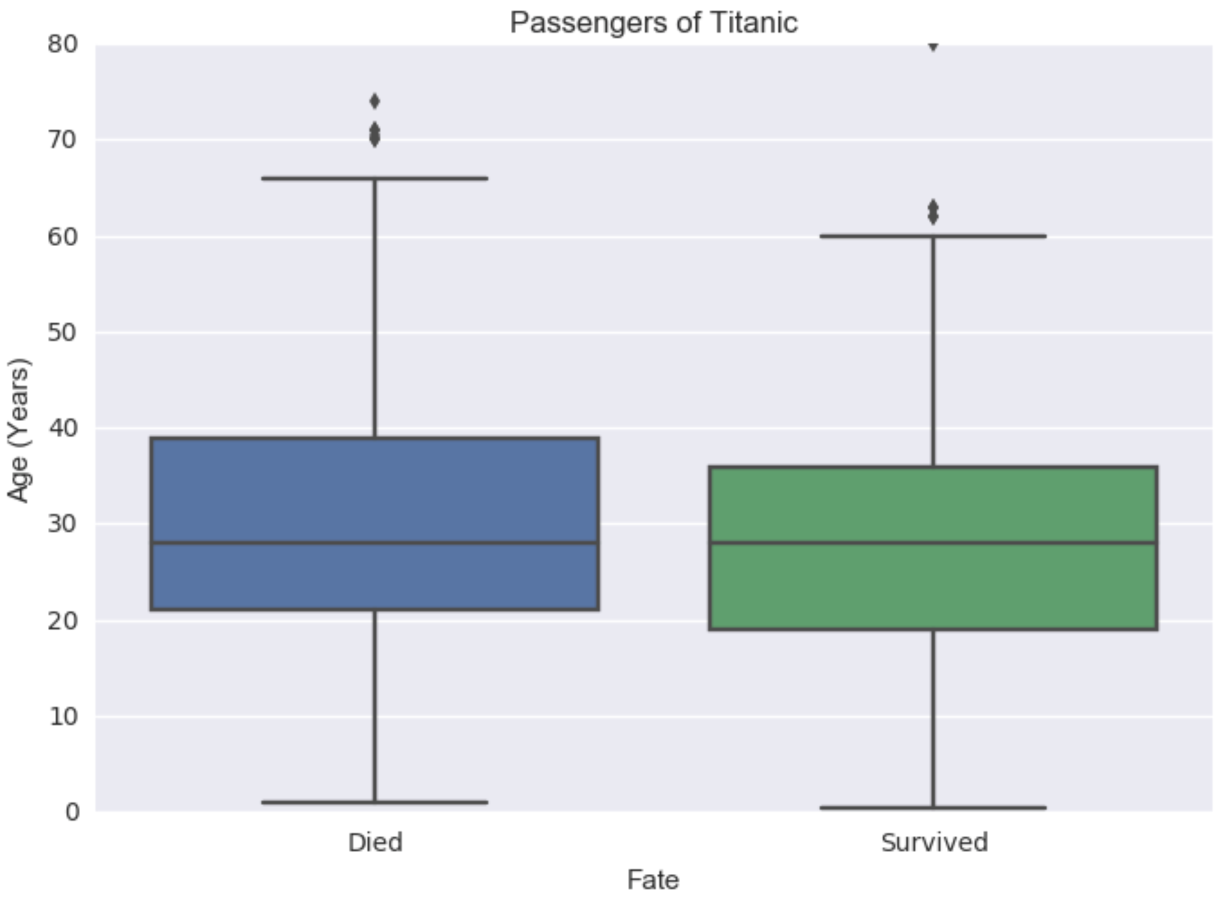
```
0.42
80.0
29.6991176471
```

The youngest passenger was 0.42 (~5 months old); the oldest was 80 years, and the average age was ~29 years of age, closely matching our impression from the plot. It’s a nice visual aid, but does age really affect the outcome of survival?

First, I need to import the seaborn as sns to calculate some important data.

Then cross-examine age and survival together using Age and Survived columns in the Excel file **titanic_data.csv** :

```
import seaborn as sns
SBA = sns.boxplot(data= TitanicDF, x= 'Survived', y= 'Age')


SBA.set(title= 'Passengers of Titanic',
        xlabel = 'Fate',
        ylabel = 'Age (Years)',
        xticklabels = ['Died', 'Survived'])
```



Based on the box-and-whisker plot, there doesn't seem to be much of an age disparity between those who **died** and those who **survived**. Although just one person would be significant enough of a loss to make headlines in this great tragedy, the figure of 891 affords us a set of data with enough validity to make these intriguing investigations of how age didn't really play a factor in likelihood of survival.

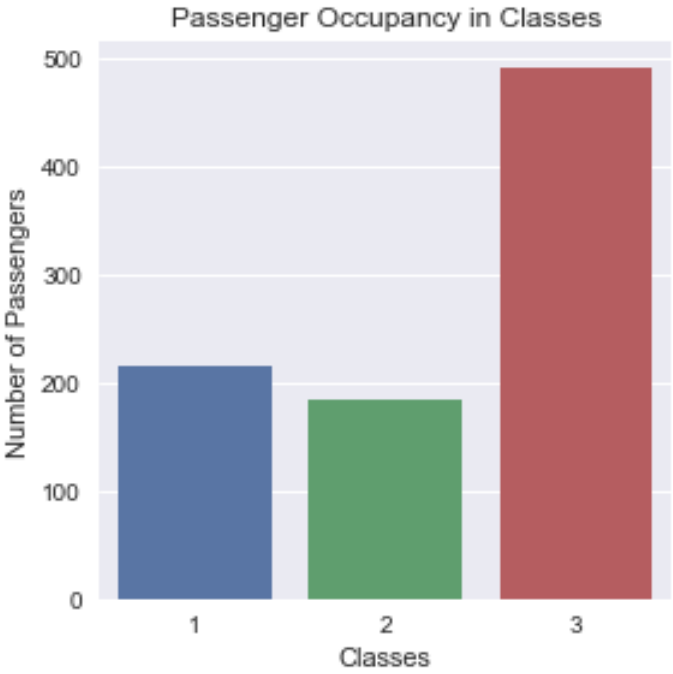
We can’t ignore that rooms were strategically placed around the ship based on class.

Even during design, before construction was completed, there was much engineering invested on locations. This was catered to those willing to spend extra for perks[1] such as space, additional rooms, readily-available amenities, windows for better view.

Was there a correlation between class and survival rates?

Let’s first see how many occupants took up each class:

```
# Explore number of passengers in Classes 1, 2, and 3:
GCSC= sns.factorplot('Pclass', order = [1, 2, 3], data = TitanicDF, kind = 'count')
sns.plt.title("Passenger Occupancy in Classes")
GCSC.despine(left = True)
GCSC.set_xlabels("Classes")
GCSC.set_ylabels("Number of Passengers")

plt.show()
```

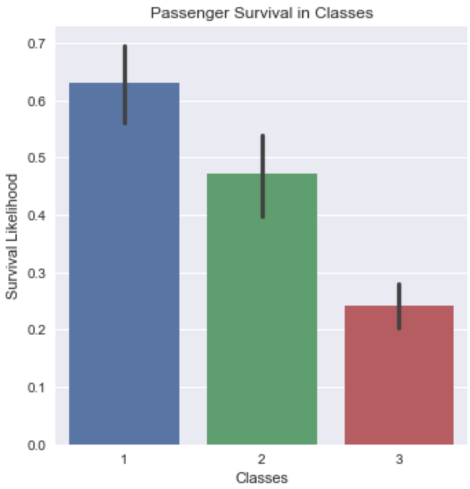


… it appears the vast majority of the passengers purchased tickets for class 3.

Next we will see the survival outcome for each class:

```
# Define the Data Frames for the the Classes:
def define_pClassProb(dataFrameIN, numClass):
    classEntries = dataFrameIN[dataFrameIN.Pclass == numClass]
    sClassEntries = classEntries[classEntries.Survived == 1]
    cClassEntries = (classEntries.count(numeric_only = True)['Pclass']).astype(float)
    cSClassEntries = (sClassEntries.count(numeric_only = True)['Pclass']).astype(float)
    return (cSClassEntries/cClassEntries)

print("Class 1:")
print("Class 2:")
print("Class 3:")

print(define_pClassProb(TitanicDF, 1))
print(define_pClassProb(TitanicDF, 2))
print(define_pClassProb(TitanicDF, 3))

# Print out the Survival outcome for Classes 1, 2, and 3:
CS = sns.factorplot( "Pclass", "Survived", order = [1, 2, 3], data = TitanicDF, kind = "bar", size = 5)
sns.plt.title("Passenger Survival in Classes")
CS.despine(left = True)
CS.set_xlabels("Classes")
CS.set_ylabels("Survival Likelihood")

plt.show()
```

```
Class 1:
0.62962962963
Class 2:
0.472826086957
Class 3:
0.242362525458
```



… based on the probabilities, and visual representations in the chart, those with the best chance for survival were in Class 1, then Class 2, and lastly Class 3. So there is a gradual descend in survival likelihood down Classes 1, 2, and 3. But what happens if we split out the genders of each class?

Let’s look at the survival likelihood for males along the 3 Classes:

```
# Define Data Frames for male passengers in Classes 1, 2, and 3:
def define_pClassProbSex(dataFrameIN, numClass, sex):
    classEntries = dataFrameIN[dataFrameIN.Pclass == numClass][TitanicDF.Sex == sex]
    sClassEntries = classEntries[classEntries.Survived == 1]
    cClassEntries = (classEntries.count(numeric_only = True)['Pclass']).astype(float)
    cSClassEntries = (sClassEntries.count(numeric_only = True)['Pclass']).astype(float)
    return (cSClassEntries/cClassEntries)

# Print out the survival likelihood for the male passengers in Classes 1, 2, and 3:
print("Class 1 Survival(MALE): ")
print(define_pClassProbSex(TitanicDF, 1, 'male'))
print("Class 2 Survival(MALE): ")
print(define_pClassProbSex(TitanicDF, 2, 'male'))
print("Class 3 Survival(MALE): ")
print(define_pClassProbSex(TitanicDF, 3, 'male'))
```

```
Class 1 Survival(MALE):
0.368852459016
Class 2 Survival(MALE):
0.157407407407
Class 3 Survival(MALE):
0.135446685879
```

Now the same for survival likelihood for females along the 3 Classes:

```
# Define Data Frames for female passengers in Classes 1, 2, and 3:
def define_pClassProbSex(dataFrameIN, numClass, sex):
    classEntries = dataFrameIN[dataFrameIN.Pclass == numClass][TitanicDF.Sex == sex]
    sClassEntries = classEntries[classEntries.Survived == 1]
    cClassEntries = (classEntries.count(numeric_only = True)['Pclass']).astype(float)
    cSClassEntries = (sClassEntries.count(numeric_only = True)['Pclass']).astype(float)
    return (cSClassEntries/cClassEntries)

# Print out the survival likelihood for the female passengers in Classes 1, 2, and 3:
print("Class 1 Survival(FEMALE): ")
print(define_pClassProbSex(TitanicDF, 1, 'female'))
print("Class 2 Survival(FEMALE): ")
print(define_pClassProbSex(TitanicDF, 2, 'female'))
print("Class 3 Survival(FEMALE): ")
print(define_pClassProbSex(TitanicDF, 3, 'female'))
```

```
Class 1 Survival(FEMALE):
0.968085106383
Class 2 Survival(FEMALE):
0.921052631579
Class 3 Survival(FEMALE):
0.5
```

After comparing the Class gradient between male and female passengers, it appears Class is a factor in likelihood of survival.

Conclusion:

When looking at the origin of the name Titanic, one might scoff at the paradox given its fate with those aboard it. It stems from the Greek word “Titan,” which according to ancient mythology, was a race of divine entities[2] who competed with the Greek gods for supremacy of their dwellings. It’s even documented that one of the employees remarked, “Not even God can sink this ship” in a famous article. Now, the connotation attached to Titanic is “sinking ship.” What we learn 100 years later is we can’t predict what is destined until after it is too late.

As for the questions initially posed in the investigation:

“If there were female stewardess and staff members, were they given the same priority as female passengers along with the children?”

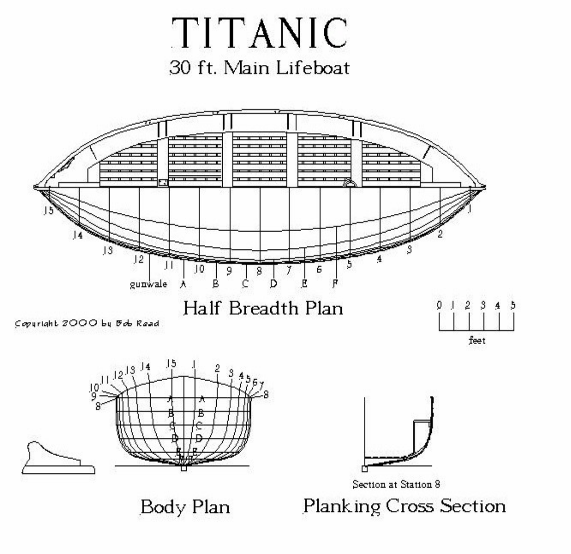
It was on the Udacity module for “Project Details” where we’re informed of the exact figures of passengers and crew members. It states that the CSV Titanic_Data file accounts for 891 passengers of the total 2224 aboard the Titanic. Further investigation on the crew members would certainly arouse intrigue. However, for the purpose of keeping it concise and relevant to the confines within the already provided data, I felt it would be best to not digress and indulge on the crew members. The Titanic_Data is sound and reliable, and contains enough variables to inspire many more investigations.

“Was there an arbitrary age number to determine if a young male is man or child?”

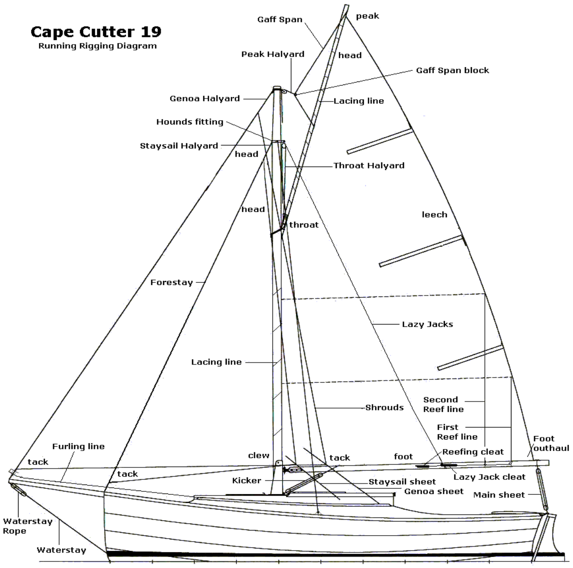
The Titanic was a British passenger ship, and according to the British Judicial Precedent, or simply the “Common Law,” a child is defined as a person under 21.[4] This definition remained in effect until 1970, when it was reduced to under 18. So the Common Law defines adults as 21 and over. However, on the Ticket Certificates, the pricing for children’s tickets is applied to those between 1 and 12.[5] In light of this, calculating the precise number that compare adult males vs women and children might be a bit murky.

“How many emergency boats were there, and how many could each carry?”

The Titanic had 20 boats, 14 of which were lifeboats, 2 cutters, and 4 collapsibles. This was above the minimum limit of 16 boats as dictated by the Board of Trade Regulations. This amount of boats was sufficient for only one-third of the total number (2,224) aboard. As realized on the Titanic the minimum 20 boats was sufficient for 825 passengers.[6]



http://titanic-model.com/db/db-02/br-db-2-lb.html



http://www.capecutter19.com/owners.htm

http://www.berthon.co.uk/about-berthon/berthon-history/

“Did class make a difference?”

Without even wading through published journals of eye witness accounts, we see numeric trends in survival between Classes 1, 2, 3. The passengers with the highest probability of survival were those in Class 1, then Class 2, and then Class 3 in that respective order. Even though women and children were first on the lifeboats, it didn’t skew the numbers. This means that, in the end, gender was not a factor. It was class.

By carefully examining the factors that led up to this disaster, this will hopefully influence future planning on safe travel. In light of the provided data sets, the biggest factor was difference of class. There was a gradual decline in survival based on class, from Class 1, 2, and then 3 in that respective order for both male and female passengers. We hope that class difference simply implies variation in luxuries, not chances of survival.

Sources

1) ***Titanic: A Primary Source History*** by Senan Molony - Publisher: Gareth Stevens Publishing (2005)
2) ***Dictionary of Greek and Roman Biography and Mythology*** by William Smith – Publisher: Spottiswoode and Co (1870)
3) ***Intuition And Beyond: A Step-by-Step Approach to Discovering the Voice of Your Spirit*** by Sharon Klinger - Publisher: Ebury Digital (2011)
4) ***The British Cyclopaedia of Literature, History, Geography, Law and Politics*** by Charles Frederick Partington – Publisher: RareBooksClub (1836)

5) ***The Statistics of the Disaster*** by Lester Mitcham - https://www.encyclopedia-titanica.org/titanic-statistics.html (2001)
6) ***Marine Review***, Volume 42 – Publisher: Penton Publishing Company (1912)

Codes

https://www.udacity.com/api/nodes/5430778793/supplemental_media/l1-starter-codeipynb/download

Data Set

https://d17h27t6h515a5.cloudfront.net/topher/2016/September/57e9a84c_titanic-data/titanic-data.csv