Udacity: Data Analyst Nanodegree
Student Name: Dave Gonsalves
Student ID: dpgonsalves@att.net
Project P2: Investigate a Dataset: *The Titanic Disaster*
Due date: 04/06/2017

Background Information

The sinking of the RMS Titanic occurred on the night of 14 April through to the morning of 15 April 1912 in the North Atlantic Ocean, four days into the ship's maiden voyage from Southampton to New York City. The largest passenger liner in service at the time, Titanic had an estimated 2,224 people on board when she struck an iceberg at around 23:40 (ship's time) on Sunday, 14 April 1912. Her sinking two hours and forty minutes later at 02:20 (05:18 GMT) on Monday, 15 April resulted in the deaths of more than 1,500 people, which made it one of the deadliest peacetime maritime disasters in history.

Source: https://en.wikipedia.org/wiki/Sinking_of_the_RMS_Titanic

Questions

One reason that this disaster led to mass loss of life was that there were not enough lifeboats for the passengers and crew. Some groups of people, such as women, children, and the upper-class, constituted the larger proportions of survivors. In this regard, I investigated the following five questions:

Q1. How representative is this passenger list compared with the entire set of passengers on the RMS Titanic?

Q2. What was the likelihood for female passengers to survive as compared with male passengers?

Q3. What was the likelihood for children (those <13 yrs. old) to survive as compared to adults?

Q4. What was the likelihood for upper-class (1st class) passengers to survive as compared to other passengers?

Q5. Can graphical representations of data for the various categories of survivors vs. non-survivors provide additional insights?

Data Wrangling
I analyzed a **partial passenger list** for the RMS Titanic labeled *titanic_data.csv* obtained from the Kaggle web site. This list was imported into a Pandas DataFrame for analysis which I've named *passengers_df*.

Data Exploration
The *passengers_df* DataFrame contained all the base data necessary to answer my questions about the Titanic disaster. As part of my inquiry I did append two Boolean columns named **Child** and **Uclass**, either to facilitate the categorization of data, or to simplify visualization. I also noted that there are missing data values for 'Age' and for 'Cabin' in the DataFrame. Only the 177 missing 'Age' values

are relevant to my questions. I discuss how I accommodate for these missing values under my response to Q3
I wrote Pandas code within a Jupyter notebook named *P2 Project – The Titanic Disaster.ipynb*, which I've included with this report. The notebook is comprehensively annotated both with in-line comments and within markdown cells that explain the code and that document answers obtained from the data exploration. The results of the exploration are described as follows:

Q1. How representative is this passenger list compared with the entire set of passengers on the RMS Titanic?

This question was intended to validate that *the partial passenger list* dataset was indeed representative of the full complement of passengers aboard the Titanic.

Result1

Per historic record the percentage of survivors among the total passengers onboard the Titanic was approx. 724/2224 = **33%**
The data in our titanic_data.csv file yielded the percentage of **survivors** to be approx.
342/(342 + 549) = **38%**

Note: This difference, missing data values for 'Age' coupled with the use of a partial passenger list for this analysis limits the confidence in the conclusions drawn from the analysis.

Q2. What was the likelihood for female passengers to survive as compared with male passengers?

Accounts of the Titanic disaster claim that women passengers were given preference over men with respect to occupancy of lifeboats. This question was intended to validate that claim.

Result2

Our sample file of Titanic passengers yielded the following results with regard to likelihood of survival for females vs. males:
The likelihood of survival for **female** passengers was 233/(233 + 81) = **74%**, whereas the likelihood of survival for **male** passengers was 109/(109 + 468) = **19%**

Q3. What was the likelihood for children (those <13 yrs. old) to survive as compared to adults?

Accounts of this tragedy also claim that child passengers were given preference over adult passengers with respect to occupancy of lifeboats. My question was intended to validate this claim. The category child is somewhat imprecise. I chose age 12 as the dividing line between child and adulthood. My choice is based on my understanding that in the 1912 era persons of age 13 and older were often expected to take on many adult responsibilities.

Result3.

Our sample file of Titanic passengers yields the following results with respect to the likelihood of survival for children vs adults:
The likelihood of survival for **child** passengers was 40/(40 + 29) = **58%**, whereas the likelihood of survival for **adult** passengers was 302/(302 + 520) = **37%**

As noted above 177 passengers are missing "Age' values. There is insufficient related data to infer the missing age values. So, the most equitable accommodation is to assume that the missing ages are spread uniformly across all the passenger categories. Thus, I calculate that among the 177

passengers with missing ages 177(69/(891 - 177)) = **17** (the **+/- err**) should be in the category 'Child'. Then, in turn, adjusting for the 177 passengers with missing ages across the relevant categories the adjusted likelihoods of survival become:
For **child** passengers (40 + **10**)/((40 + **10**) + (29 + **7**)) = **58%**
For **adult** passengers (302 – **7**)/((302 – **7**) + (520 – **10**)) = **37%**
Clearly, in this case, this adjustment did not alter the aggregate results

Q4. What was the likelihood for upper-class (1st class) passengers to survive as compared to other passengers?

Much fun has been poked at the wealthy onboard the Titanic for receiving undue preference with respect to occupying the limited lifeboat space. My question was intended to determine if this was, indeed, the case.
Result4.

Our sample file of Titanic passengers yields the following results with respect to likelihood of survival for upper-class (Pclass = 1) passengers vs passengers travelling in other classes:
The likelihood of survival for **upper-class** passengers was 136/(136 + 80) = **63%**, whereas the likelihood of survival for **other classes** of passengers was 206/(206 + 469) = **31%**

While performing this analysis, it occurred to me that the ratio of females among the upper-class may distort the above finding, so I also determine the following:

The percentage of **females in upper-class** was (91 + 3)/(136 + 80) = **44%**, whereas the percentage of **females for partial passenger list** was (233 + 81)/891 = **35%**

Q5. Can graphical representations of data for the various categories of survivors vs. non-survivors provide additional insights?

Visual representations often allow readers to grasp the findings more quickly than the textual information. In this regard, I generated Matplot pie charts and a Seaborn bar chart that segment the various categories of passengers. The pie charts augment the results but does not substitute for them because the textual results reflect likelihood of survival within category, whereas the graphic segmentation shows proportions of each subcategory within the partial passenger list.

Result5.

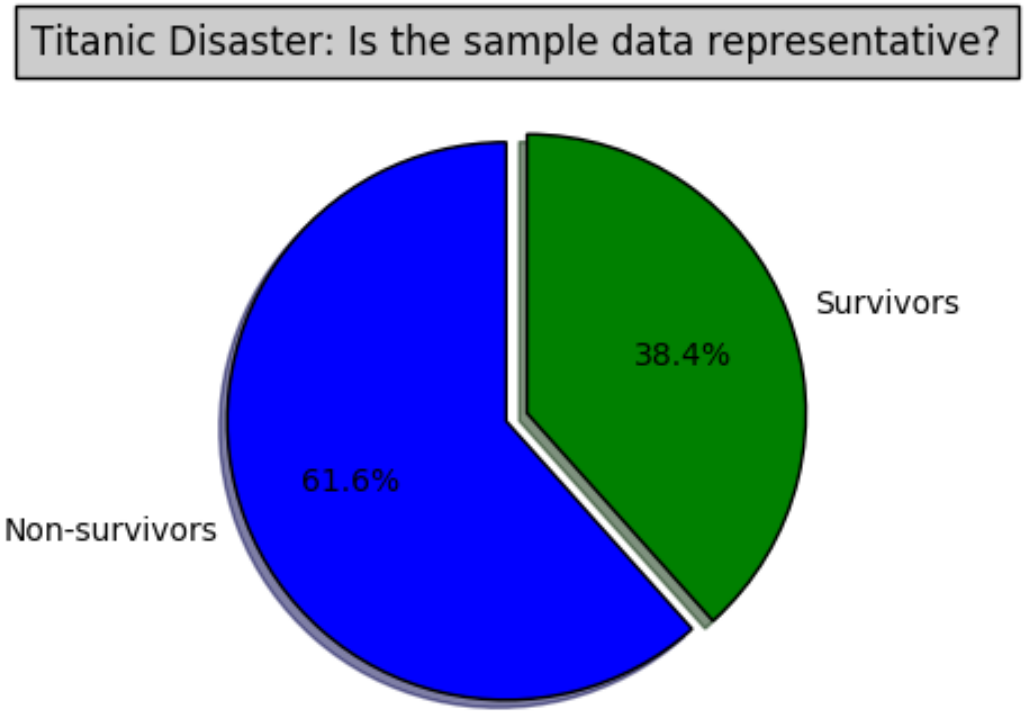
The aforementioned charts appear below.

Conclusions & Visualizations

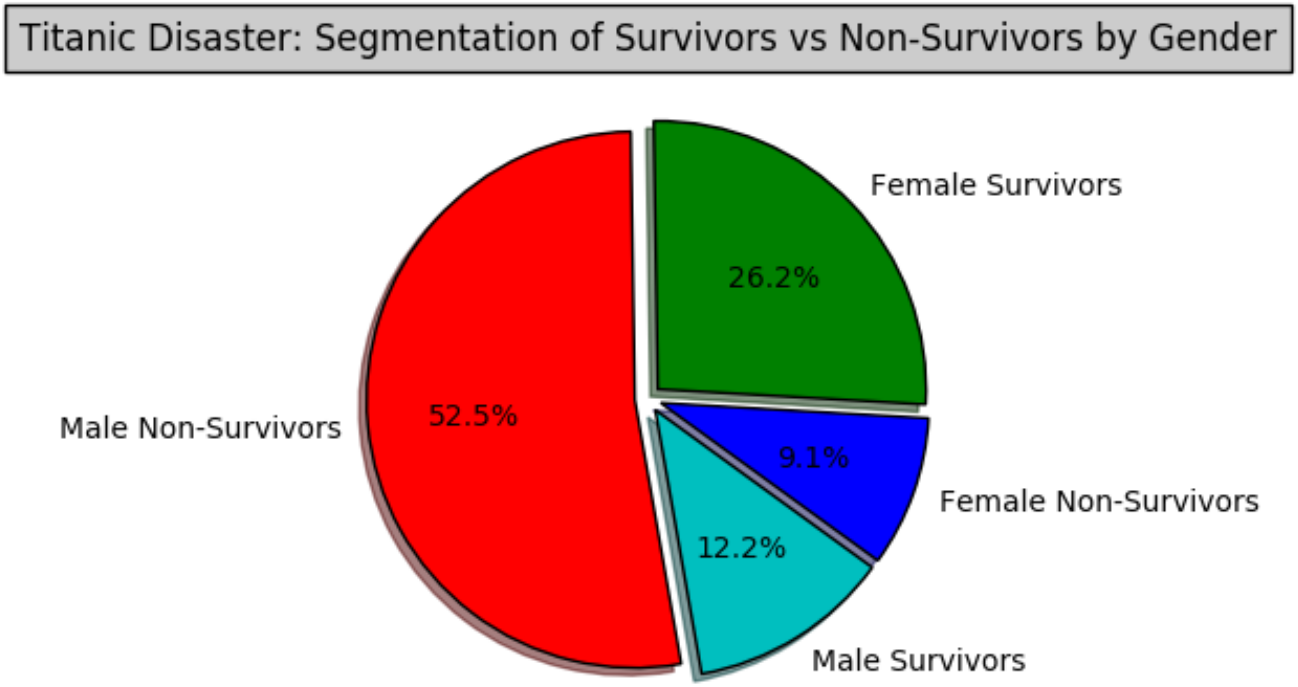
As expected, from historical accounts, I've confirmed that various categories of passengers aboard the Titanic seemed to have had survival advantages. My analysis confirmed the following:

Limitations: as noted under Result1 there is a small difference in the Titanic survival rate based on historical records (33%) vs the survival rate (38%) drawn from the partial passenger list, analyzed here. This difference plus missing data values for 'Age' (handled as described under Result3) limits the confidence in the conclusions drawn from this analysis.

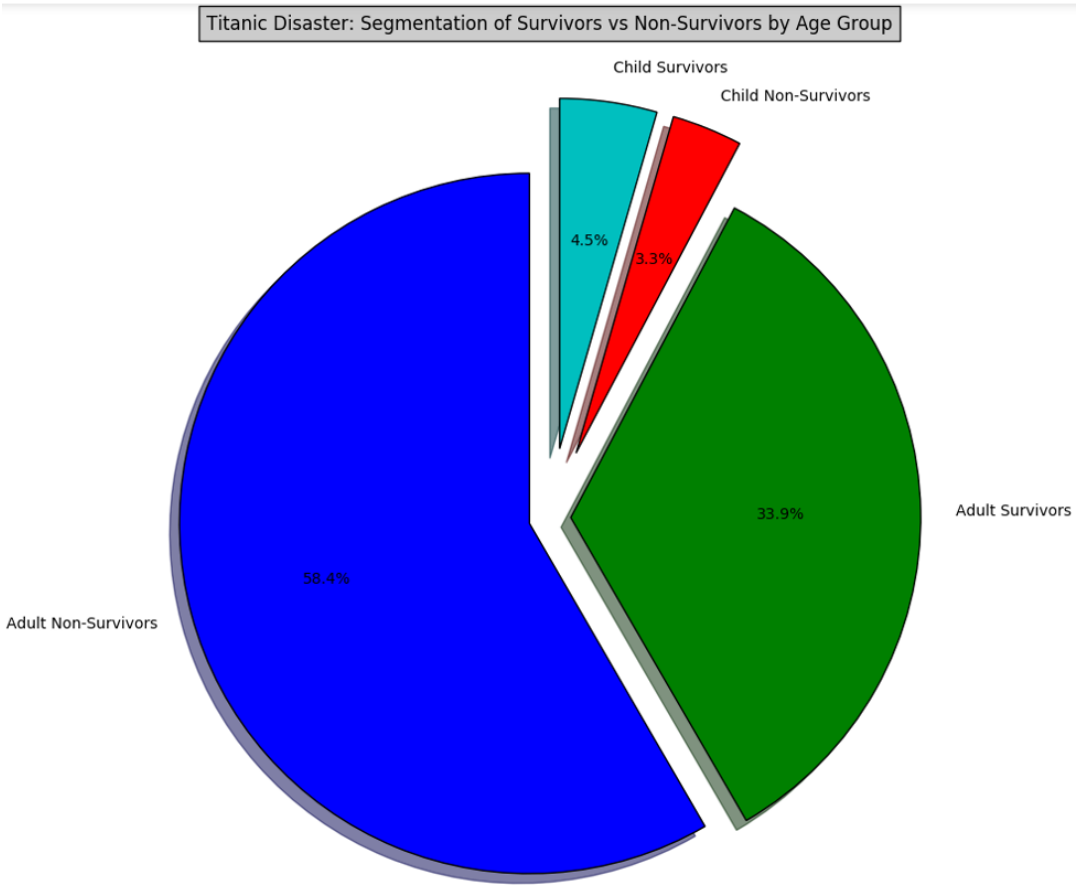
Answer to Q1: Yes, it appears that our sample of Titanic passengers is confirmed to be representative of the total compliment of passengers aboard because the survival rate for the partial list is 38%, which is very similar to the historically reported rate of 33%.



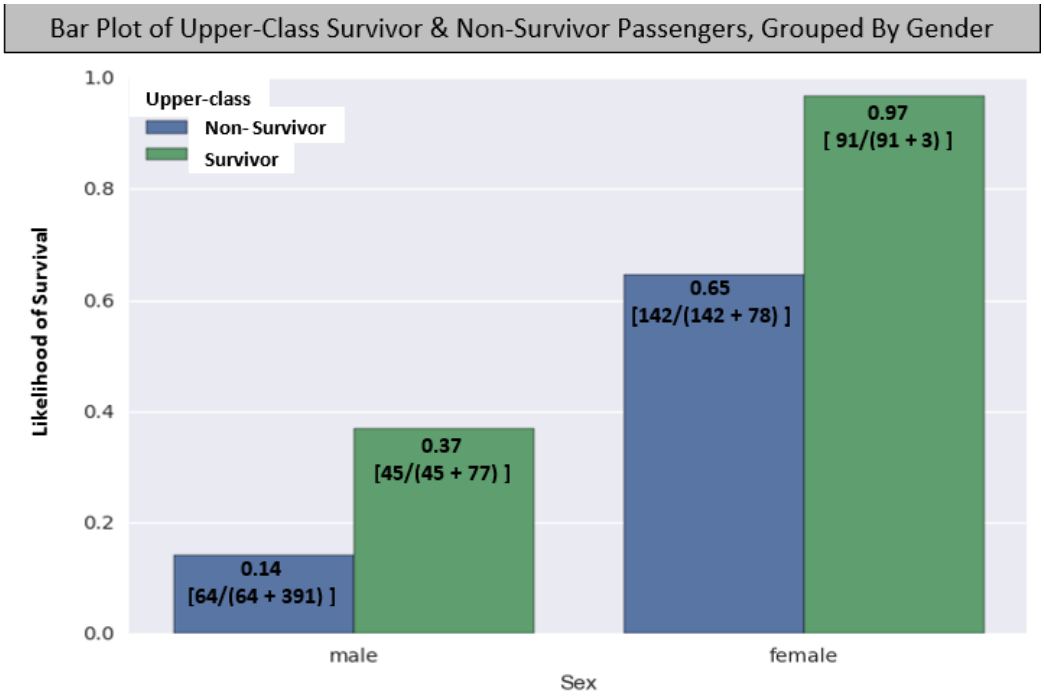
Answer to Q2: The likelihood for a female passenger to survive was 74%, whereas the likelihood for a male was 19%. This represents a significant advantage for females!



Answer to Q3: The likelihood for a child passenger to survive was 58%, whereas the likelihood for an adult passenger to survive was 37%. Children did have a survival advantage, but not as significant as that for females.



Answer to Q4: The likelihood for an upper-class passenger to survive was 63%, whereas the likelihood for an other-class passenger to survive was 31%. Thus, it appears that upper-class passengers had a better likelihood of survival than children.
However, a complicating factor is that females constitute 44% of the upper-class passengers as compared to 35% for the whole dataset.
To better illustrate this factor, I've used a Seaborn plot to show the male vs female mix among the upper-class.



Note: The precise values for the bar heights were calculated from the results of:
upper_class_vs_other_classes_sex = passengers_df.groupby(['Uclass', 'Survived', 'Sex'])['Uclass'].count()

It seems that the higher proportion of women that constituted the upper-class probably makes the advantage upper-class survival relative to children insignificant.
Nevertheless, the upper-class passengers maintained a survival advantage over the other-class adult passengers. The category of passengers that had the greatest survival advantage was upper-class women

Summary
1st Hypothesis (H_0): The adage “women and children first” holds true.
Results:
Based upon analyzing Titanic passenger survivor/non-survivor counts the likelihood of a female passenger on the Titanic to survive was 74%;
Based upon analyzing Titanic passenger survivor/non-survivor counts the likelihood of a child passenger on the Titanic to survive was 58%.

2nd Hypothesis (H_0): The upper-class are the privileged class, even in a crisis.
Results:
Based upon analyzing Titanic passenger survivor/non-survivor counts the likelihood of an upper-class passenger on the Titanic to survive was 63%.
Moreover, the likelihood of an upper-class, female passenger to survive was 97%

Limitations: as noted under Result1 there is a small difference in the Titanic survival rate based on historical records (33%) vs the survival rate (38%) drawn from the partial passenger list, analyzed here. This difference plus missing data values for ‘Age’ (handled as described under Result3) limits the confidence in the conclusions drawn from this analysis.