IS 362 Final Project

Simple Analysis of the Titanic Passenger List

By Matthew Heino

CUNY – SPS

IS 362 Fall 2019

December 13, 2019

Introduction

This presentation seeks to use the Titanic data set that was downloaded from Vanderbilt University's Department of Biostatistics. The link to the data set is the following:

Titanic Data Set

This data set has information about the passenegers of the Titanic. For example, the dataset has information about the passengers including the name, age and sex. Other information about passengers is also available. To view this information please visit the link above. You will need to look at the data description file that is available on the website.

Goals of the Project

- The overall goals of this project is to use as many concepts and tools that were developed during the course. I will make use of the following components from the course:

1. Dataframes to store the data.
2. Reading of data from a CSV.
3. Using a Web API to try to retrieve some articles about some of the passengers of the Titanic.
4. The machine learning aspect of skearn.
5. Transformations of the data to better enable processing.

While this is not an exhaustive list these are the key elements that will be use in the project.

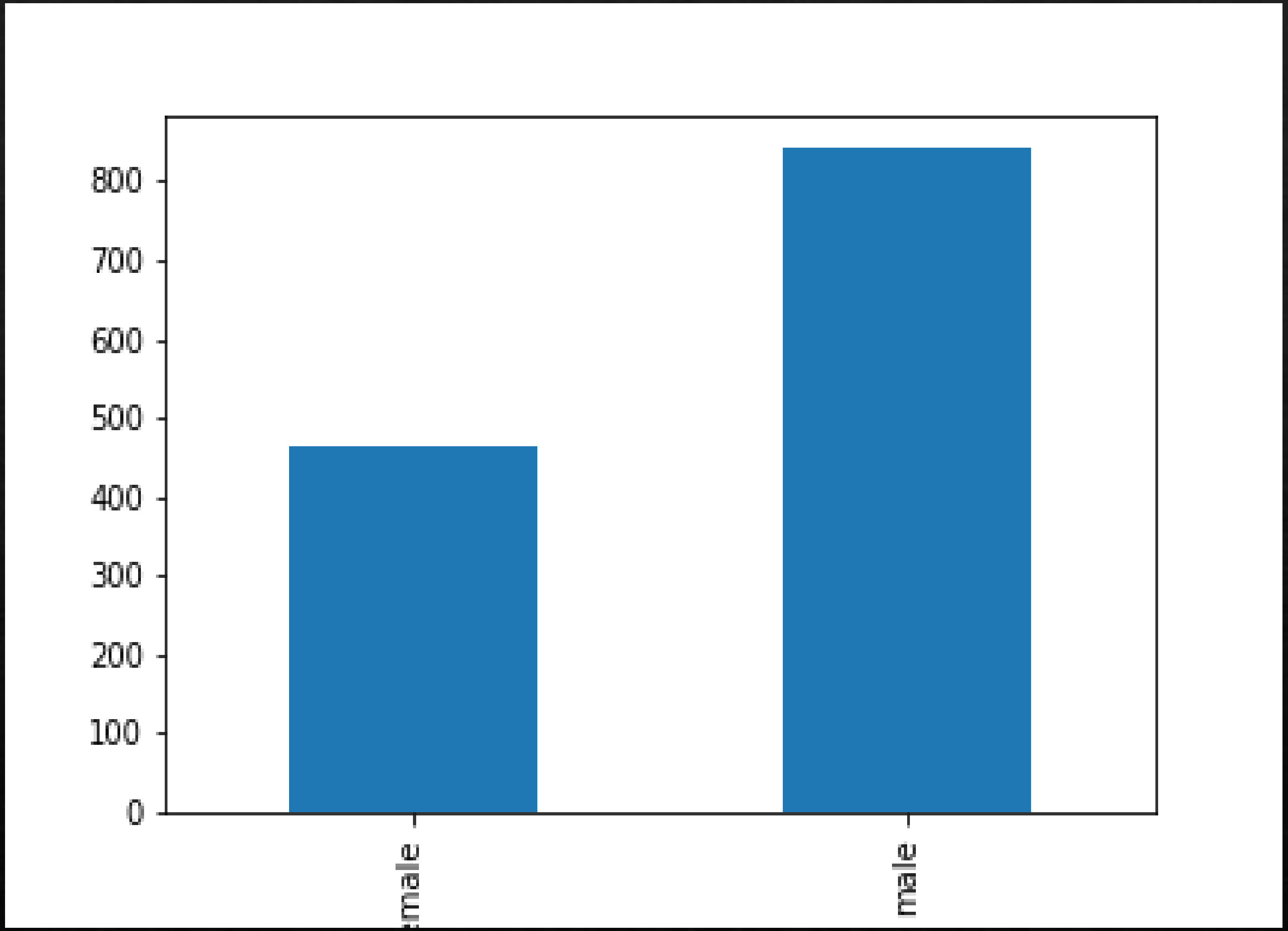
Some Information on the Titanic Passenger List

- This list contains the names and other information about the passengers. It also contains vital statistics like sex and age. It also has information that we will need to help do a prediction. We will make use of 'pclass' throughout the course of this presentation. I will use it to develop some statistics about the passengers. I will group them by this variable exclusively. It seems the most logical way to group to get the desired result for the calculations that will made during the course of the project.

Some Simple Statistics

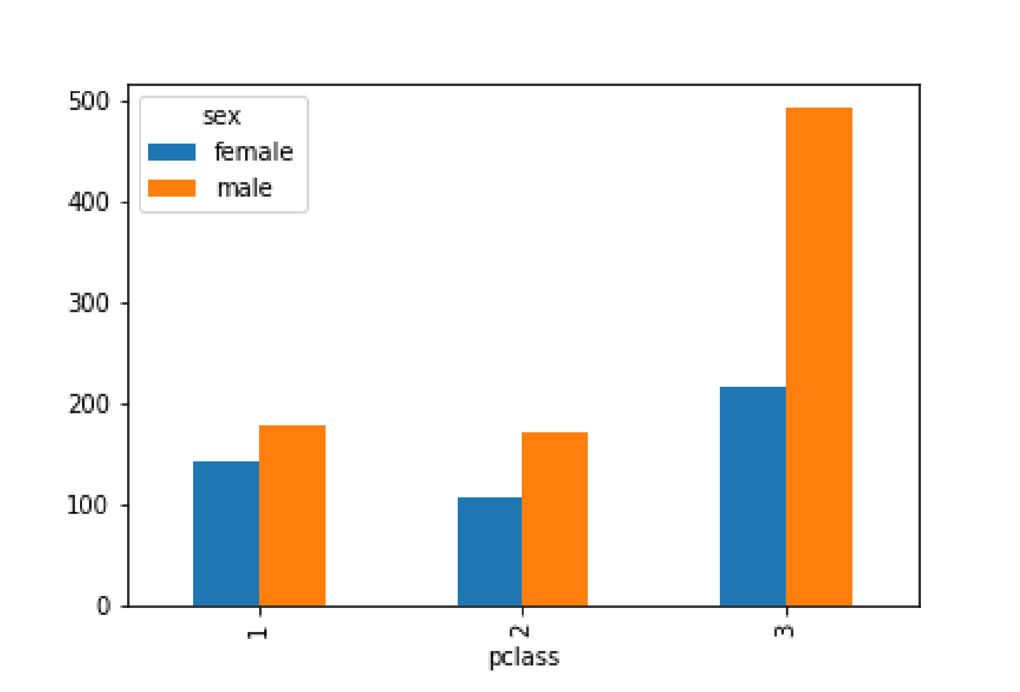
Using a simple code we can determine the total number of entries in the list was 1309. Which is the total number of passengers on the Titanic list. We can see the breakup of the passengers in the graph.

All code for this and future coding segments can be found in the Jupyter Notebook. It is found here: [Titanic Jupyter Notebook]()



How Many men and women were in each of the passenger classes?

Plcass (Passenger Class)	Sex	Count
1	female	144
	male	179
2	female	106
	male	171
3	female	216
	male	493



What were the average ages of the passengers?

The first group we will look at is the average age of males and females on the Titanic. The chart below is given below on the left gives the average age as grouped by sex. The chart on the right gives the average age based on class. There is additional information about how these calcuations were accomplished in the Notebook.

Sex	Age
Female	28.687088
Male	30.585228

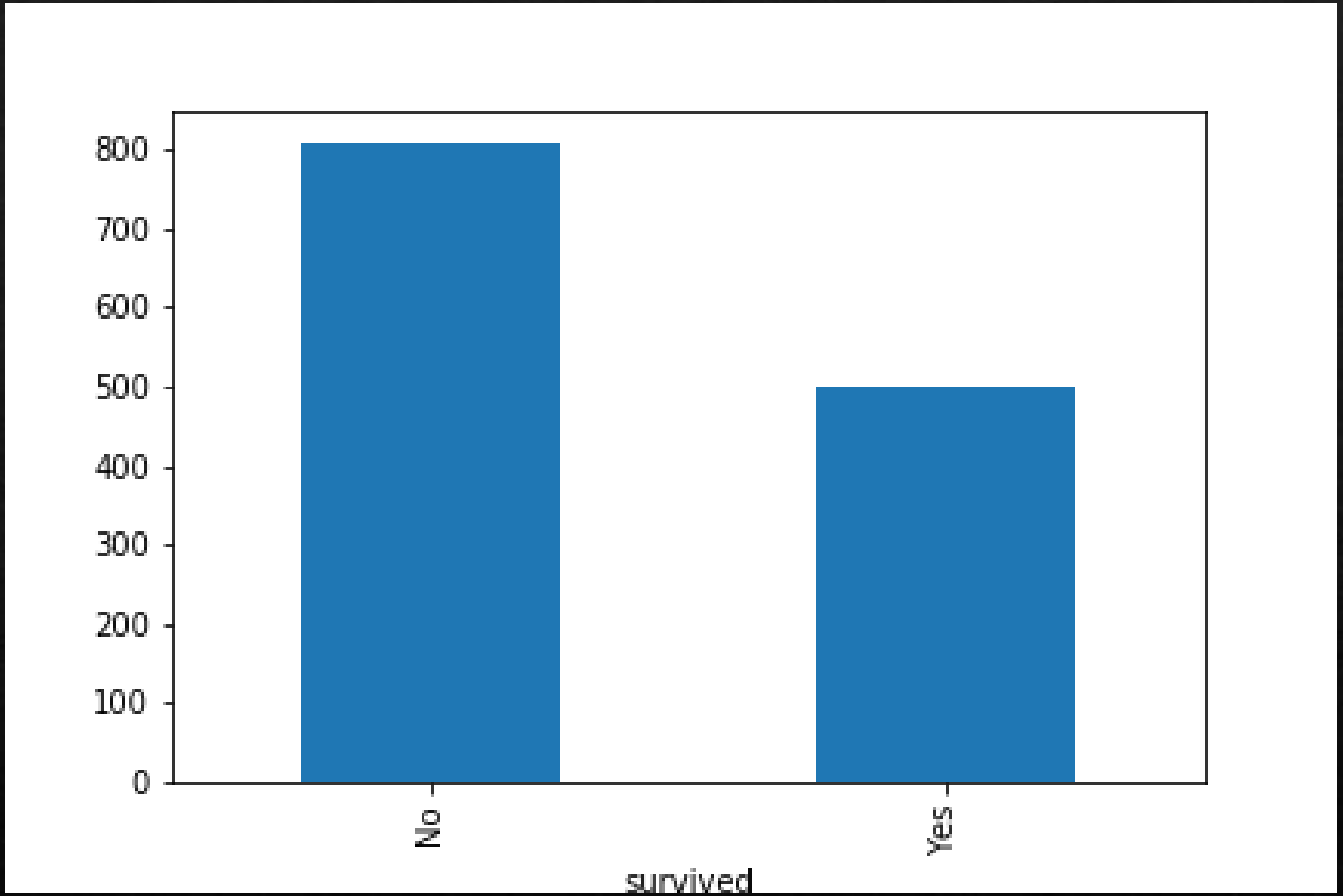
Passenger Class	Age
1	39.159930
2	29.506705
3	24.816367

How were the youngest and the oldest passengers on the Titanic?

- This part of the project looked to find the personal information of the oldest and the youngest passenger on the Titanic. The code to accomplish this is the Notebook, which is linked at the beginning of this presentation. The results were the following:
- Youngest passenger: Miss. Elizabeth Gladys "Millvina" Dean (.17 Years)
- Oldest passenger: Mr. Algernon Henry Wilson Barkworth (age 80 years)
- I also tried to find some articles using the Web API for the New York Times but was unable to find anything related to these passengers to get a little more information about them. I did run a test query using John Jacob Astor to see if I could retrieve information. I was able to get a few articles about him. So , the query string does work as constructed.

How many passengers actually survived?

- We might like to know how many of the passengers actually survived the disaster. When we do the calculations we get the following numbers:
- Survived: 509
- Died: 800
- A graph of the this result is shown below.

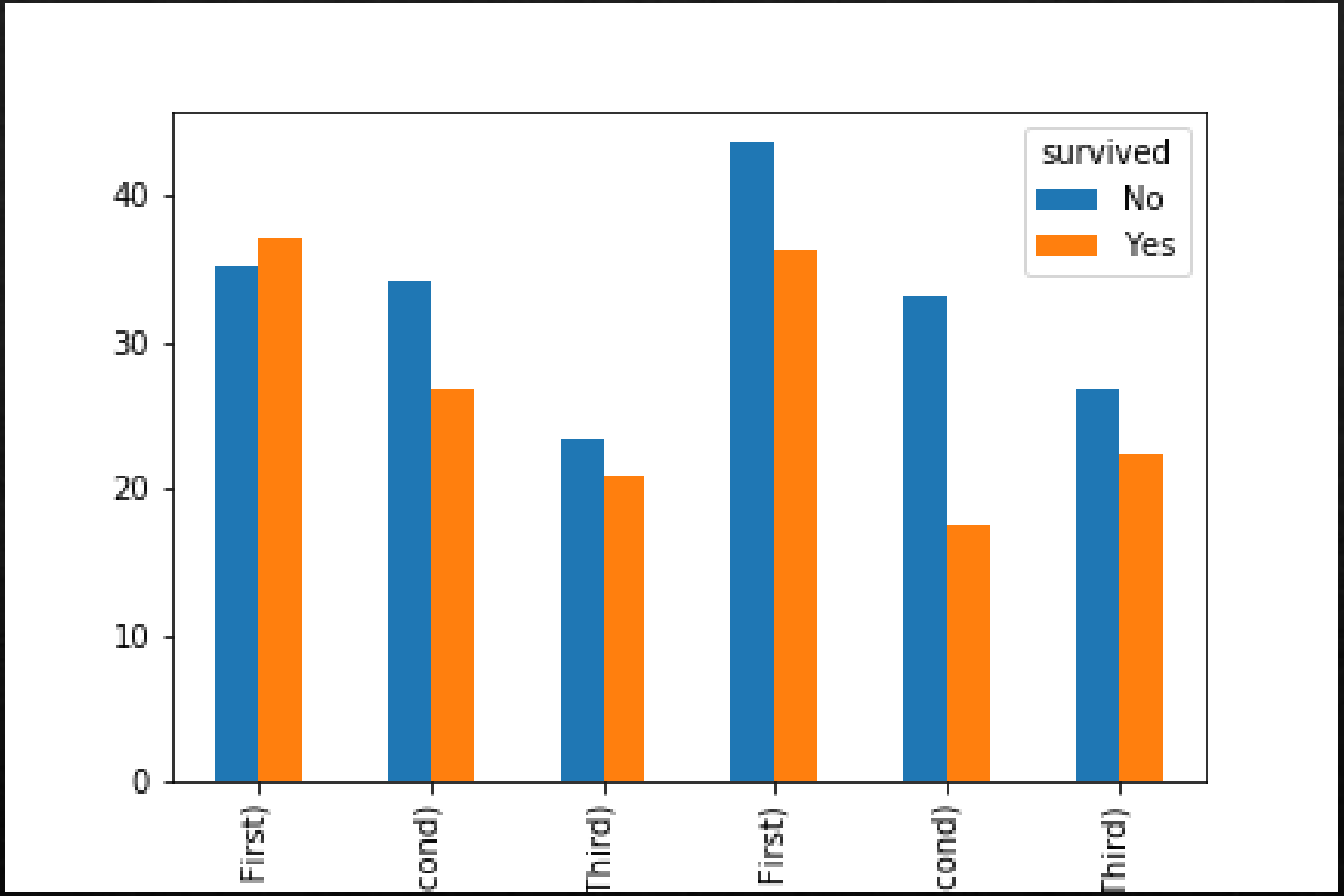


How many passengers actually survived? (cont.)

- We can look at the average age of the survivors of the Titanic based on class. When we do that we get the following information.

Sex	Class	Survived	Age
Femal	First	No	35.200000
		Yes	37.109375
	Second	No	34.090909
		Yes	26.711087
	Third	No	23.418750
		Yes	20.814861
Male	First	No	43.658163
		Yes	36.168302
	Second	No	33.092593
		Yes	17.449130
	Third	No	26.679586
		Yes	22.436441

- This is the visual represnations of the mean ages as broken down by passenger class. (The labels were cutoff, but the full graph can be found in the associated notebook.)



Making Some Predictions

- I was able to make some predictions as to what feature of the data set will make a good indicator of who would survive. The first was using class as a predictor of the whether one would survive. The results were:
  - Using Kneighbors:0.6967150496562261
  - Using Kneighbors:0.6600458365164248
  - Using Logistic regression: 0.5637891520244461
- Using sex as a predictor the results were:
  - Using Kneighbors:0.8074866310160428
  - Using Kneighbors:0.7708174178762414
  - Using Logistic regression: 0.7799847211611918

Conclusion

- Using sex as a predictor was the better course when trying to determine whether a passenger was to survive. With a little insight and knowledge of the how lifeboats were loaded this outcome was not surprising. Using class did only slighty better than 50/50 show to predict whether a passenger was to survive the sinking.
- Summary: The project showed some interresting stats. It is amazing that there was such a a breadth of ages on the Titanic. The youngest not even a year old and the oldest was 80. Overall the ages of the passengers were relatively young. The mean seems to be centered around 30 or so. Most of the passengers should have had most of their lives to look forward to.