Peter Xinrui Lu
January 10th, 2017

Disclaimer: This was not a controlled experiment, and thus no conclusive causations can be drawn from the data. All results below are observations and casual inferences discovered from the dataset, and should not be treated as definite conclusions.

Introduction

This report was created to record the investigation into the "Titanic" dataset, comprised of various information attributes for each of the passengers in the population sample provided. Overall, the process of data wrangling and analysis were applied to extract meaningful and insightful aspects of the dataset, primarily done using Python's `numpy` and `pandas` libraries. This report contains the questions about the data were posed prior to the investigation, the methods and process used to analyze the data, and the conclusions reached from such analysis.

Structure of Data

The dataset being investigated was contained in a comma-separated values (`.csv`) file, named '`titanic_data`'. It was comprised of various personal attributes of passengers onboard the Titanic in regards to the the ship and their journey. This file had data from a sample of 891 passengers onboard, and information was omitted where not found. More specifically, the dataset had the following attributes:

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)
```

The primary attribute that was most relevant in this investigation is the survival variable, which indicated whether or not the given passenger had survived the wreckage of the

Titanic. Other variables were investigated to extrapolate connections and correlations between them and survival.

Questions Explored From the Data

From the variables in the dataset, it was determined that they could roughly be split into three categories:

- Direct Impact on Survival — **Age, Passenger Class, and Sex:** these variables were thought to have the most direct influence on an individual's survival, and had the strongest correlation to the survival variable.
- Indirect Impact on Survival — **SibSp (Siblings & Spouses) and Parch (Parents & Children):** these variables were thought to not direct contribute to an individual's survival, but were closely tied to direct impact variables, and thus were still relevant in the investigation.
- Low Impact on Survival — **Ticket #, Fare, Port of Destination, and Name:** these variables had either low or negligible impact on survival, and may be loosely tied to direct impact variables.

The investigation was based upon examining the correlation of direct impact variables in relation to survival, in conjunction with how indirect and low impact survival variables could influence these direct impact variables.

Some of the following questions were used as lead-ins to the start of the investigation:

- What impact did each of the direct variables have on survival?
- For each direct variable, what is the percentage of people who survived from categorization on those variables?
- Factoring in the indirect variables, what influences do they have in relation to the direct impact variables on survival?
- What relations, if any, does the low impact variables have on the direct or indirect impact variables?

Methods Used

This data was primarily processed and analyzed through the use of Python's `numpy` and `pandas` libraries, used to parse extract information such as the average survival rate among groups, the correlation between survival and other variables, and others.

Results & Conclusions

From analysis on the dataset, the following were found:

Correlation between various variables and survival:

	Survived
PassengerID	-0.005007
Survived	1
Pclass	**-0.338481**
Sex	**0.543351**
Age	0.010539
SibSp	-0.035322
Parch	0.081629
Fare	**0.257307**

As is shown from the above table, survival had the greatest correlation to the **Sex** variable, with a computed 0.54 correlation coefficient. This indicates that for any singular variable, the **Sex** of an individual had the greatest impact on survival.

Furthermore, it can be seen that the **Passenger Class** variable had a negative correlation coefficient with regards to survival, indicating that individuals who belonged to lower classes (classes start at 1, with 3 being the lowest class) had a lower chance of survival.

The only other statistically relevant variable in relation to survival was **Fare**, which indicated a positive correlation between fare paid and survival rates. This connection can be explored further through the correlation between **Fare** and **Passenger Class**, which had a **-0.5495** correlation coefficient, indicating that lower classes tended to pay less than higher classes, which makes sense considering the definition of **Passenger Class** defined previously.

Average survival rates for various passenger groups, split by direct impact variables:

Passenger Group	Average Survival Rate	Passenger Count
Male Passengers Only	0.188908145581	577
Female Passengers Only	**0.742038216561**	**314**
Children Passengers Only	0.382608695652	316
Adult Passengers Only	0.386075949367	575
Class 1 Passengers Only	**0.62962962963**	**216**
Class 2 Passengers Only	0.472826086957	184
Class 3 Passengers Only	0.242362525458	491

From these values, one can see that the correlation found above between **Sex** and survival is clearly demonstrated by the average survival rates of male passengers compared to female passengers. On average, females out-survived males by roughly **4:1**, a **74%** survival rate for all females compared to just **18%** of all males. However, there is to note that due to the lower number of females, this result may be due to sampling error from the population.

Another interesting statistic to note is the survival rates between the various classes, starting at approximately **63%** of all **Class 1** passengers, to **47%** for **Class 2** passengers, and finally dropping to just **24%** for **Class 3** passengers. This supports to the correlation found above for passenger class and survival, which indicated a negative correlation between lower classes and survival.

Finally, even though there were more adult passengers onboard than children, it can be seen that survival rates for both groups are relatively similar, which seem to indicate that no preference was given to child passengers during rescue.

Average survival rates for various passenger groups, split by indirect impact variables:

Passenger Group	Average Survival Rate	Passenger Count
>0 Siblings and Spouses	0.466431095406	283
No Siblings or Spouses	0.345394736842	608
>0 Parents and Children	0.343657817109	678
No Parents or Children	0.511737089202	213
Dest: Queenstown	0.38961038961	77
Dest: Cherbourg	0.553571428571	168

Passenger Group	Average Survival Rate	Passenger Count
Dest: UNKNOWN	1	2
Dest: Southhampton	0.336956521739	644
Greater than Avg. Fare	**0.597156398104**	**211**
Lesser or Equal to Avg. Fare	**0.317647058824**	**680**

Here, it can be noted that passengers who paid greater than average fare had a far greater proportion of survivors, 59% compared to 32% to those who paid lesser or equal to the average fare. However, due to the fact that the vast majority of people paid lesser or average fare, the finding may also be due to sampling error.

Otherwise, as there was no significant correlation found between the above variables and survival, the values here are presented purely for information, and no concrete conclusions are drawn from the data.

Average survival rates for various passenger groups, split by a combination of variables:

* Note: The table for this data is included at the end, at Appendix I.

Highlights:

- For all groups except female passengers, having siblings/spouses or parents/children onboard tended to have higher average rates of survival than those who did not.
- For both **Class 1** and **Class 2** female passengers, their survival rate was much higher than the average survival rate of all females, **97%** and **92%** respectively compared to the **74%** average. This stands in particular contrast with **Class 3** female passengers, who only had a survival rate of **50%.**
- For **Class 1** passengers who had no parents or children onboard, their survival rate is **36%,** compared to the remarkably higher **78%** of those who did. However, since both sample sizes are relatively low, with those who did have parents or children especially, this may be a result of sample bias.

Conclusion:

Overall, the data shows that survival rates of passengers had a significant correlation in regards to both **Sex** and **Passenger Class**, with being **Female** and **Class 1** particular standouts in regards to high survival rates. Both children and adults had similar average survival rates, which may indicate a lack of preference during rescue for children. Also, it seemed that on average, people who had spouses, siblings, parents or children onboard had survival rates, may be of particular interest, as one would usually expect otherwise.

Appendix I.

Passenger Group	Average Survival Rate	Passenger Count
Male & Passenger Class 1	0.368852459016	122
Male & Passenger Class 2	0.157407407407	108
Male & Passenger Class 3	0.135446685879	347
Female & Passenger Class 1	**0.968085106383**	**94**
Female & Passenger Class 2	**0.921052631579**	**76**
Female & Passenger Class 3	**0.5**	**144**
Adult & Passenger Class 1	0.635294117647	170
Adult & Passenger Class 2	0.416666666667	144
Adult & Passenger Class 3	0.199233716475	261
Child & Passenger Class 1	0.608695652174	46
Child & Passenger Class 2	**0.675**	**40**
Child & Passenger Class 3	0.291304347826	230
Male & Child	0.205128205128	195
Male & Adult	0.180628272251	382
Female & Child	0.677685950413	121
Female & Adult	0.782383419689	193
No Siblings and Spouses & Male	0.168	434
>0 Siblings and Spouses & Male	0.252	143
No Siblings and Spouses & Female	**0.787**	**174**
>0 Siblings and Spouses & Female	**0.686**	**140**
No Siblings and Spouses & Adult	0.342	410
>0 Siblings and Spouses & Adult	0.485	165
No Siblings and Spouses & Child	0.354	198
>0 Siblings and Spouses & Child	0.441	118

Passenger Group	Average Survival Rate	Passenger Count
No Siblings and Spouses & Passenger Class 1	0.562	137
>0 Siblings and Spouses & Passenger Class 1	0.747	79
No Siblings and Spouses & Passenger Class 2	0.417	120
>0 Siblings and Spouses & Passenger Class 2	0.578	64
No Siblings and Spouses & Passenger Class 3	0.236	351
>0 Siblings and Spouses & Passenger Class 3	0.257	140
No Parents or Children & Male	0.165	484
>0 Parents or Children & Male	0.312	93
No Parents or Children & Female	0.788	194
>0 Parents or Children & Female	0.666	120
No Parents or Children & Adult	0.352	471
>0 Parents or Children & Adult	0.519	104
No Parents or Children & Child	0.323	207
>0 Parents or Children & Child	0.504	109
No Parents or Children & Passenger Class 1	**0.358**	**134**
>0 Parents or Children & Passenger Class 1	**0.78**	**50**
No Parents or Children & Passenger Class 2	0.225	381
>0 Parents or Children & Passenger Class2	0.3	110
No Parents or Children & Passenger Class 3	0.607	163
>0 Parents or Children & Passenger Class 3	0.698	53