Titanic Survival Rates

Author : Kweku E. Acquaye.

Preamble

In this report, a dataset[1] of RMS Titanic passengers available on GitHub was analyzed for insights on variables that impacted survival rates, as part of a course class task[2].

Method

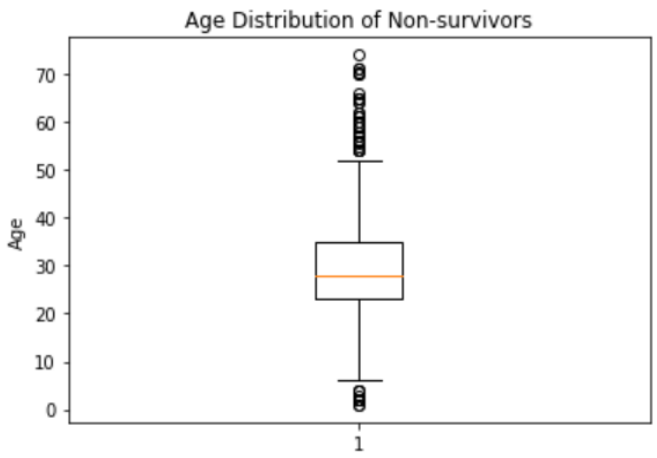
The dataset was loaded directly from GitHub. The chosen medium of analysis is the cloud platform Google Colab, a python Jupyter Notebook utility. Standard Pandas and Numpy libraries were used throughout the analysis.

Questions

Following exploratory data analysis (EDA), the following questions were posed :

1. "What is the age distribution of non-survivors?"
2. "To what extent did class and sex affect survival?"
3. "Can money buy life - to what extent did fare price affect survival?"

Results

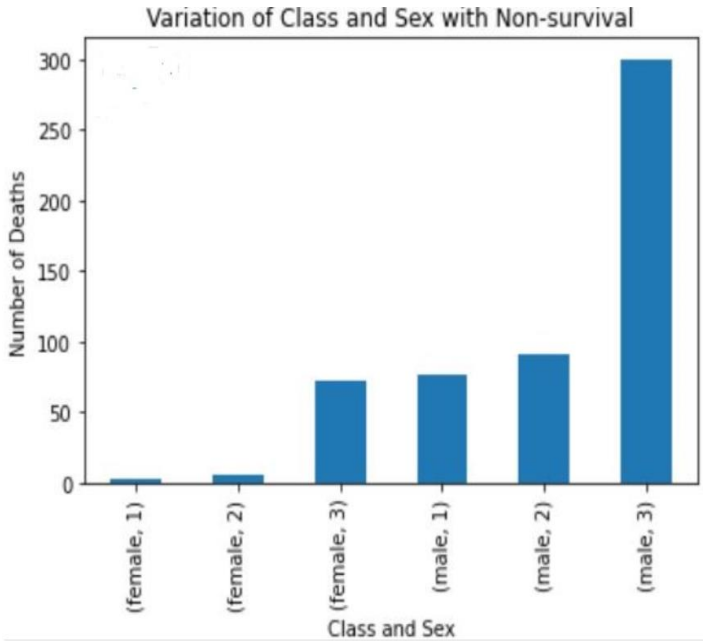



```
dns['age'].describe()

count    549.000000
mean      30.028233
std       12.499986
min        1.000000
25%       23.000000
50%       28.000000
75%       35.000000
max       74.000000
Name: age, dtype: float64
```

```
dns['age'].median()

28.0
```



	survived	0
sex	pclass	
female	1	3
	2	6
	3	72
male	1	77
	2	91
	3	300

Observations and Discussion

From a sample size of 891, EDA revealed missing data in the following features :

Feature	*No. Missing*	*Percentage Missing*
age	: 177	19.9
embarked	: 2	000.2
deck	: 688	77.2
embark_town	: 2	000.2

With 77.2% and 19.9% missing data in some columns, cleaning this dataset by removing all records with missing data would impact the information within significantly. Also in this instance, the author's choice of columns needed for analysis - survival, class, sex, and fare - have no missing values. However the "age" column has missing data, hence 'age' column was cleaned by filling it's empty cells with the median value.

From the analysis (Figure 1 and Table 1), the age distribution of non-survivors was observed to form a Gaussian curve with both mean and median at 28, and interquatile range of 23 to 35 = 12. This distribution is however likely to be co-dependent on second class male and female passengers, more so on third class male and female, but especially so on third class male passengers. The higher the integer value of the class, the lower the chance of survival. This was largely due to the ship's evacuation policy of "children and women first", which in the author's opinion, is a noble policy even today.

In tandem to this was also an ignoble policy - that of class. From the analysis (Figure 2 and Table 2) it was observed that non-survival correlated highly with class integer number, and starkly so with regard to males.

Analysis also determined that variation of fare price with survival rates was significant - the higher the fare price paid, the better one's chances of survival on the Titanic. This was also likely to be tied to class. By far the major contributory factor to non-survival was class in both sexes, but particularly pronounced in males.

Conclusions

Survival rates on the Titanic were impacted significantly by age, class, and sex. Fare price was observed to be directly related to survival rate generally. Hints of bivariate and multivariate relationships with class were not investigated in this task.

Appendix

1. Link to Colab notebook : Kweku's Google Colab Analysis for Titanic Mini-report Task

References

1. Link to dataset : Titanic Reduced Dataset on GitHub

2. Link to task : FutureLearn Class Task