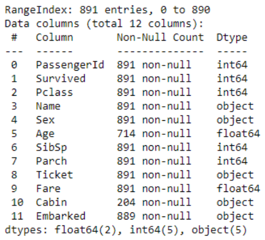
Jorge Ramirez
ID 10376776
CPSC 4800 W01
3/11/22

1. TITANIC DATASET

The Titanic was a British ship that sank on April 15, 1912, after striking an iceberg off the coast of Newfoundland. Of the 2,240 passengers and crew on board, more than 67% lost their lives.

We received a dataset containing observations of this well-known incident to evaluate a series of hypotheses for this report.



```
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
```

This first glimpse indicates the types of variables we are working with; forward this point, we will focus only on the following ones:

- Survived: an integer that indicates if the passenger **survived (1) or not (0)**
- Pclass: an integer that indicates the passenger's class
- Sex: a string variable that indicates the gender of each passenger
- Age: a float variable that indicates the age of each passenger

The mean replaced all Missing values in our Age variable.

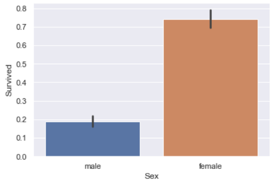
2. OUR HYPOTHESIS

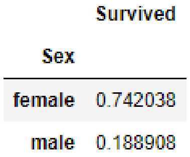
The hypothesis we need to test using statistical techniques are the following:

a. Determine the survival rate is associated with age
b. Determine if the survival rate is associated with gender
c. Determine if the survival rate is associated with the class of passenger

2.1 Determine the survival rate is associated with gender

We related these variables through a bar plot to evaluate this hypothesis, showing us which gender had the best chance of surviving the incident.



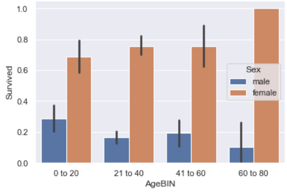


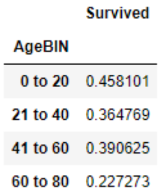
	Survived
Sex	
female	0.742038
male	0.188908

We test that both variables are associated through this plot, being females with better chances to survive (74%).

2.2 Determine if the survival rate is associated with age

As stated before, we managed all missing values related to the "Age" variable; but to come up with a simple conclusion not relating hundred of data points, we binned our variable "Age" into four groups.



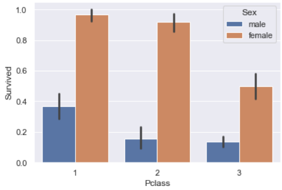


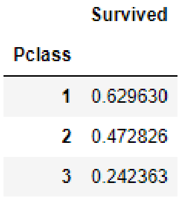
	Survived
AgeBIN	
0 to 20	0.458101
21 to 40	0.364769
41 to 60	0.390625
60 to 80	0.227273

As with our first hypothesis, age is associated with the survival rate. The passengers between 0 to 40 years have the best chance to survive, particularly women between 0 to 20 years old.

2.3 Determine if the survival rate is associated with the class of passenger

Finally, the passenger class had an essential role in the survival rate, being first-class passengers with the best survival chances. And again, following the pattern of or last hypothesis, first-class women had the best survival chances overall.





	Survived
Pclass	
1	0.629630
2	0.472826
3	0.242363

3. Correlations.

Among our numerical data, we calculated the following correlation indexes:

Survival – Passenger class (Correlation index = -0.034)

- Considering that 3 in our dataset indicates third class, we can say, despite the sign of the index, that the lower the class, the lower the survival probability.

Survival – Age (Correlation index = -0.072)

- Indicating the older the passenger was, the less the survival probability.