Can the logistic regression predict which passenger will survive on the Titanic?

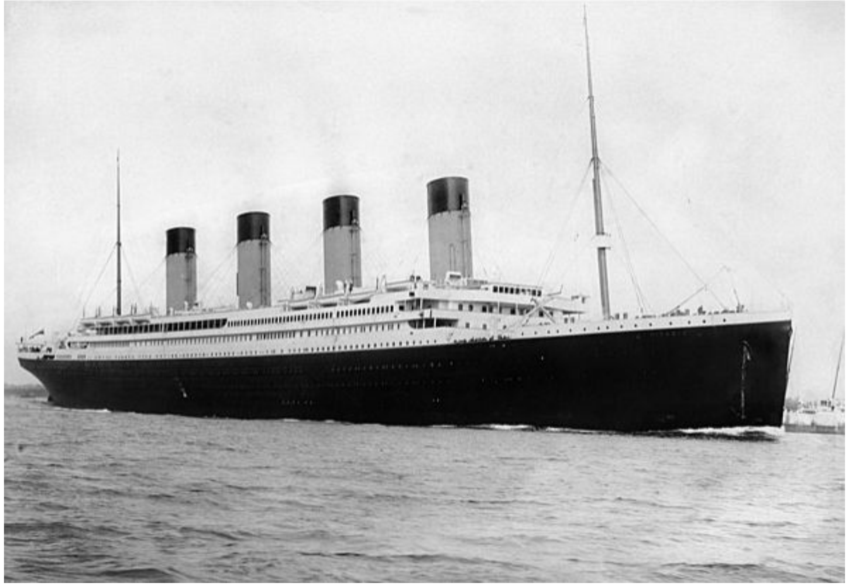
Content

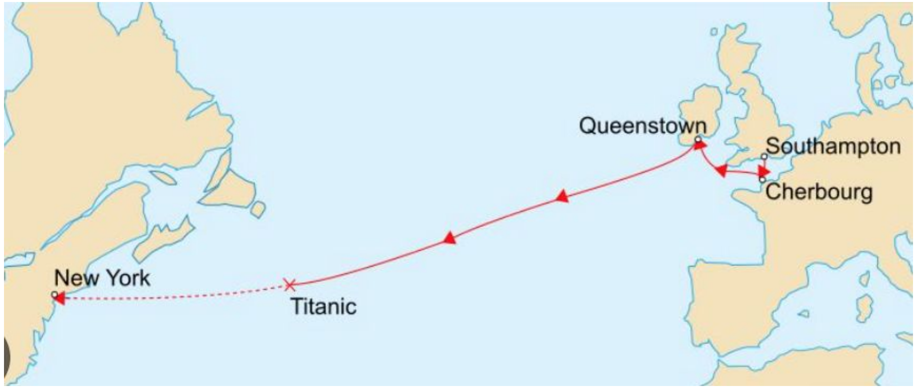
1. Backstory of Titanic
2. The dataset
3. Correlation Matrix
4. Exploratory Analysis
5. Building Logistic Regression Model
6. Evaluation of model
7. Conclusion
8. Appendix

Backstory of titanic

RMS Titanic: Luxurious, most advanced ship passenger at that time

- Left Southampton on 10th April 1912 to NY
- Hit North Atlantic Ocean iceberg at 11:40 PM
- 3 hours to sink on 14 April 1912





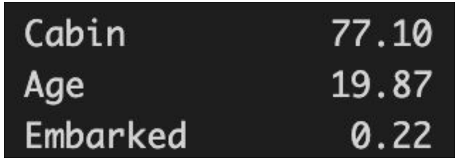
- 2,240 passengers on board incl 885 crew members
- Lifeboats for ⅓ passengers only
- 1500 died

The dataset

Train data for exploratory analysis:

- 891 entries
- 11 features
- 1 target variable: Survived
- 5 categorical + 7 numerical

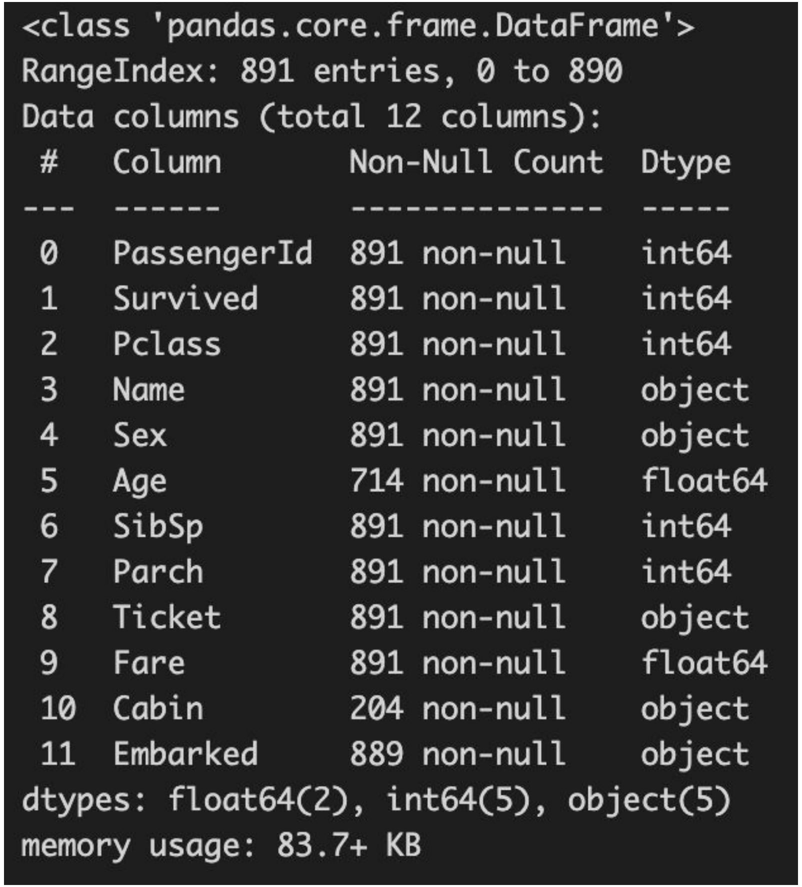
Cleaning:



```
Cabin         77.10
Age           19.87
Embarked       0.22
```

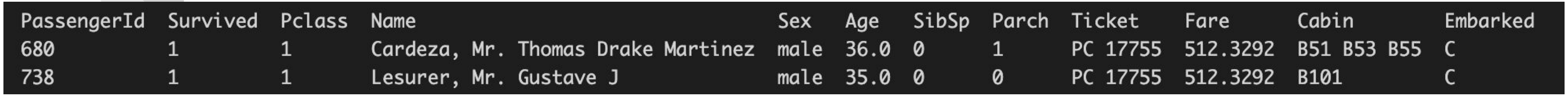
- Missing values (%)
- Deleted ‘Cabin’
- Replaced NaN age with mean
- Age range:0-80
- Embarked deleted 2 rows

=> only 889 entries + 11 columns



```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 83.7+ KB
```

Example

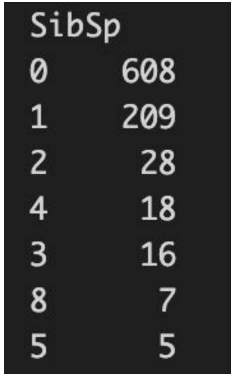


```
PassengerId  Survived  Pclass  Name                                Sex   Age   SibSp  Parch  Ticket    Fare      Cabin        Embarked
680          1         1       Cardeza, Mr. Thomas Drake Martinez  male  36.0  0      1      PC 17755  512.3292  B51 B53 B55  C
738          1         1       Lesurer, Mr. Gustave J              male  35.0  0      0      PC 17755  512.3292  B101         C
```

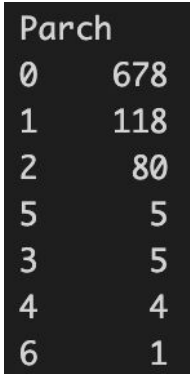
Sibsp = Number of sibling/spouse aboard (0-5)

Parch = Number of parents/children aboard (0-6)

-> Majority were on the boat on their own



```
SibSp
0    608
1    209
2     28
4     18
3     16
8      7
5      5
```



```
Parch
0    678
1    118
2     80
5      5
3      5
4      4
6      1
```

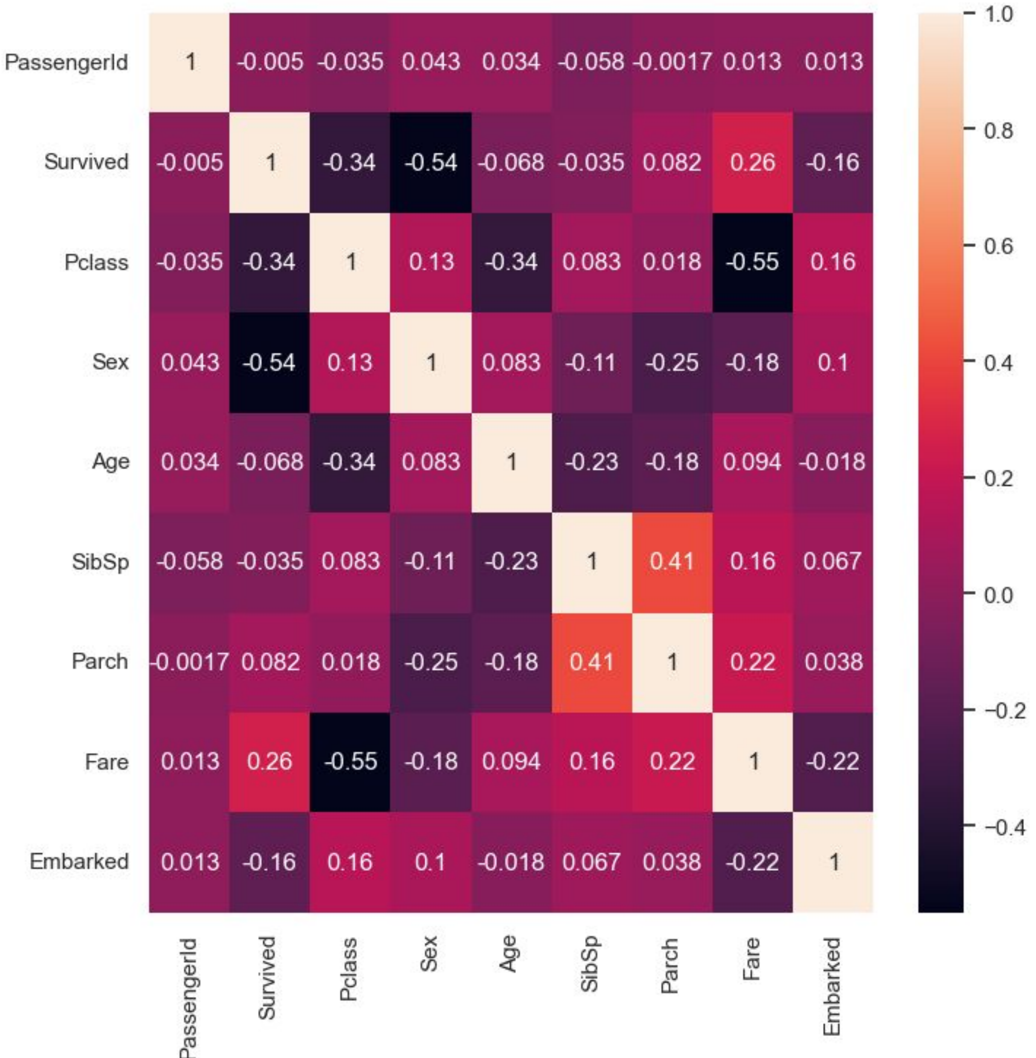
2nd cleaning: Pre-processing

- Encode: ‘Age’ and ‘Embark
- Standardise all

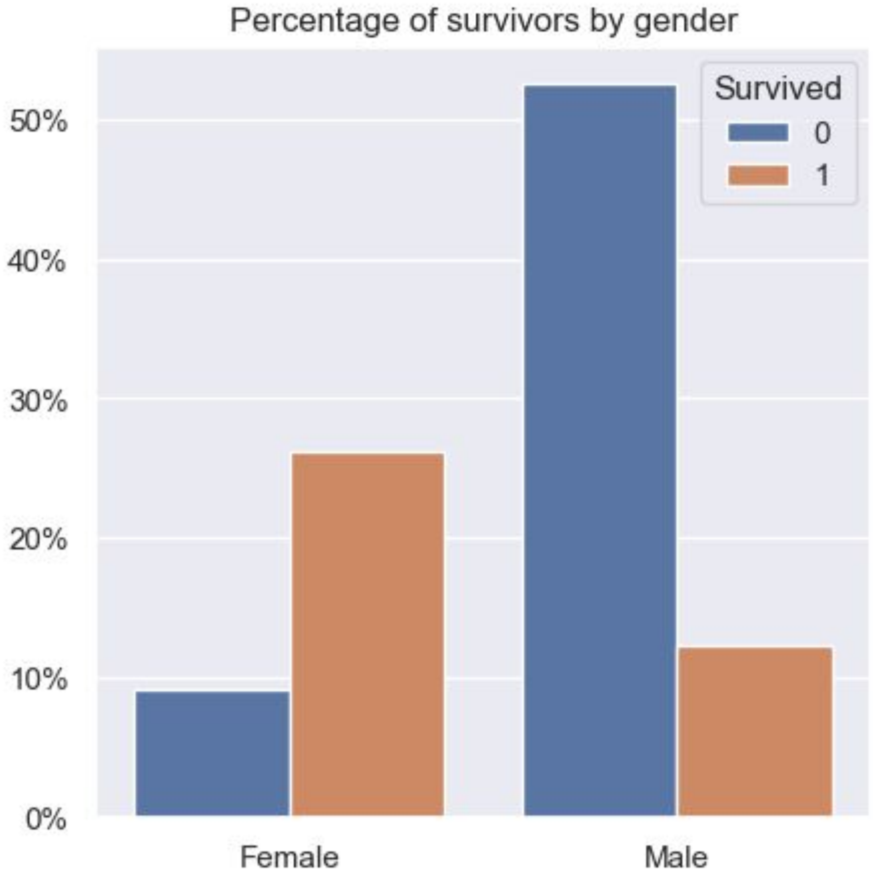
Correlation Matrix

Top 4 correlated variables:

- Sex: -0.54
- Pclass: 0.26
- Fare: 0.26
- Embarked: -0.16
- Remaining: <0.1



	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
PassengerId	1	-0.005	-0.035	0.043	0.034	-0.058	-0.0017	0.013	0.013
Survived	-0.005	1	-0.34	-0.54	-0.068	-0.035	0.082	0.26	-0.16
Pclass	-0.035	-0.34	1	0.13	-0.34	0.083	0.018	-0.55	0.16
Sex	0.043	-0.54	0.13	1	0.083	-0.11	-0.25	-0.18	0.1
Age	0.034	-0.068	-0.34	0.083	1	-0.23	-0.18	0.094	-0.018
SibSp	-0.058	-0.035	0.083	-0.11	-0.23	1	0.41	0.16	0.067
Parch	-0.0017	0.082	0.018	-0.25	-0.18	0.41	1	0.22	0.038
Fare	0.013	0.26	-0.55	-0.18	0.094	0.16	0.22	1	-0.22
Embarked	0.013	-0.16	0.16	0.1	-0.018	0.067	0.038	-0.22	1

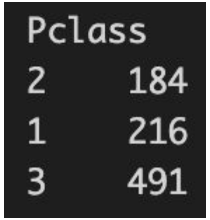


1)**<u>Sex:</u>** gender male/female

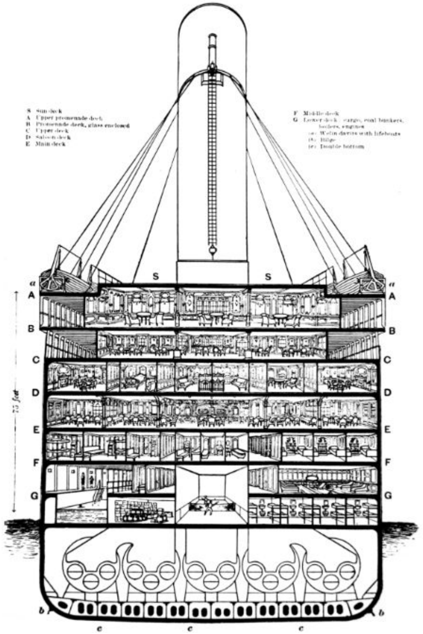
Correlation: -0.54

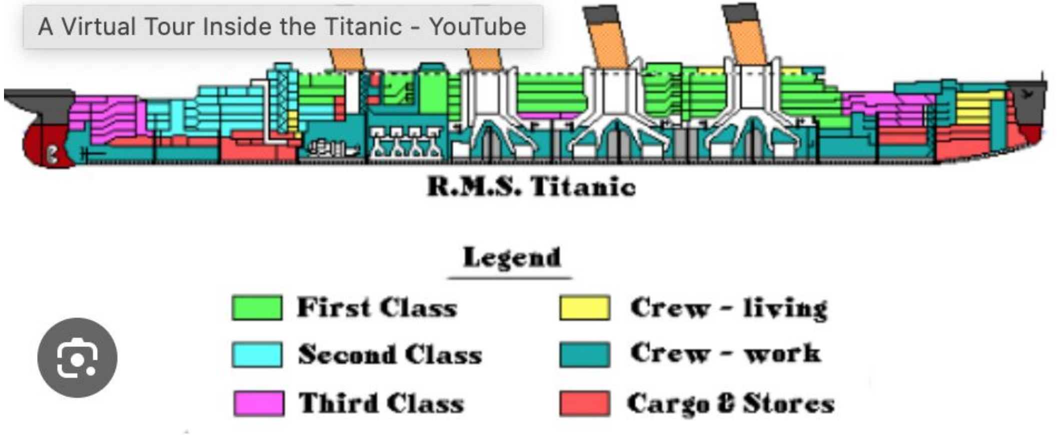
- Female survival rate: 26.15%
- Male survival: 12.23%

2) **PClass** = Proxy for socioeconomic classes , correlation at 0.26



```
Pclass
2      184
1      216
3      491
```



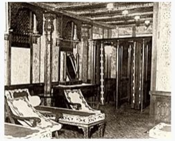


he Forward First Class ;rand Staircase of *Titanic*'s sister ship RMS *Olympic*. *Titanic*'s taircase would have ooked nearly identical. lo known photos of *Titanic*'s staircase exist.

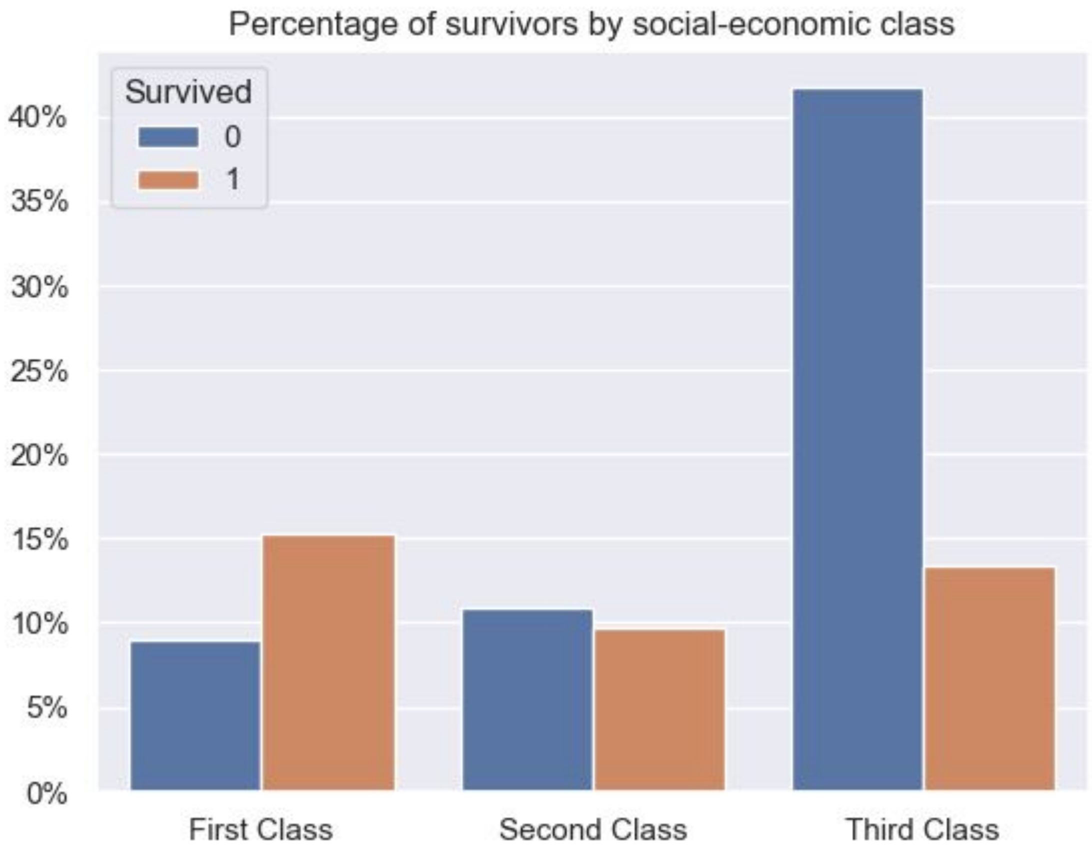
The gymnasium on the boat deck, which was equipped with the latest exercise machines

The à la carte restaurant on B Deck (pictured here on sister ship RMS *Olympic*), run as a concession by Italian-born chef Gaspare Gatti

The First Class lounge of RMS *Olympic*, *Titanic*'s sister ship

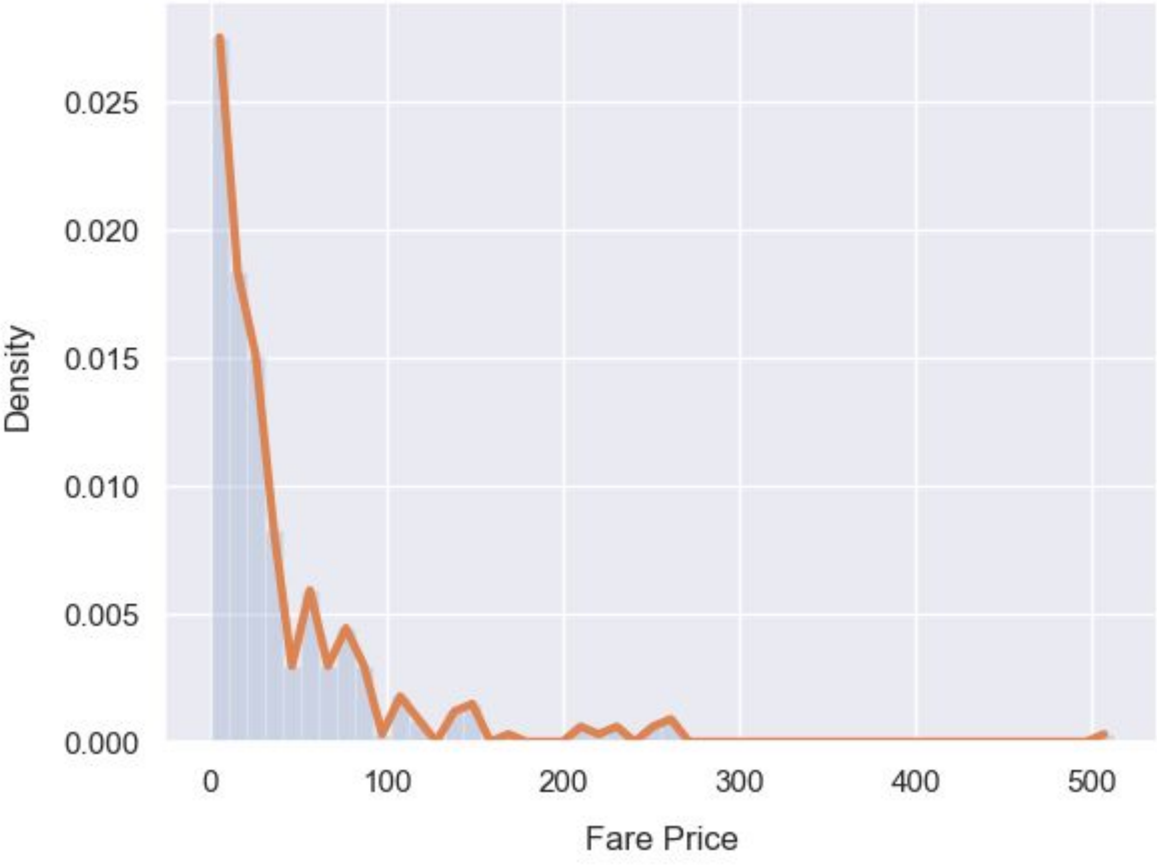


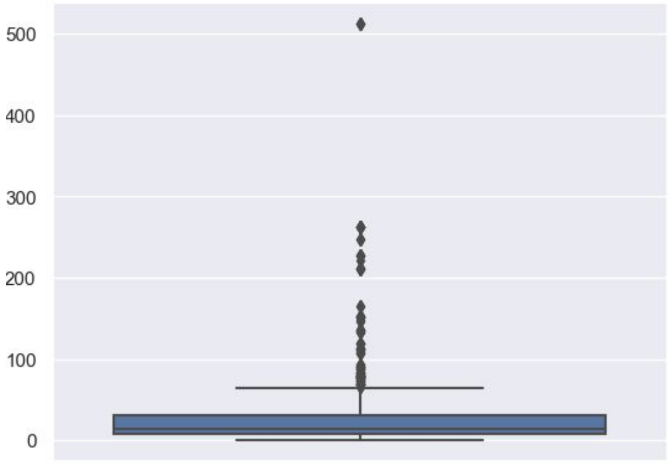
The First Class turkish baths, located along the Starboard side of F-Decl

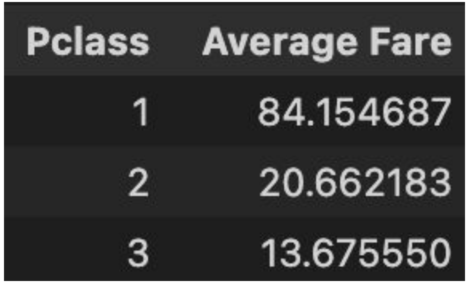


- Highest chance of survival in First Class
- Lowest: Third Class

3) <u>Fare:</u> price tickets ranged from 0.0 to 513 shillings.







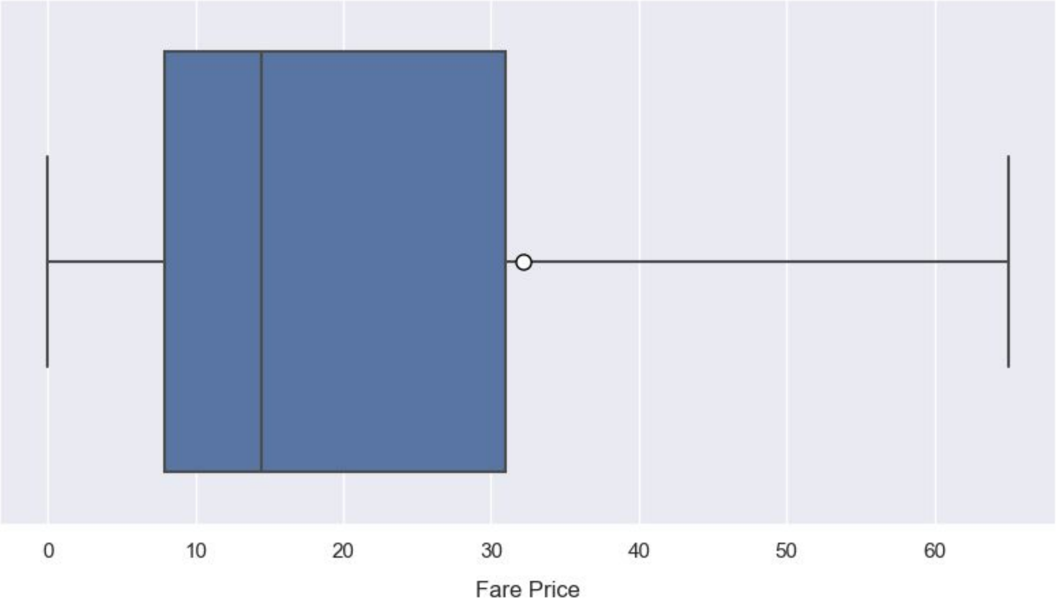
Pclass	Average Fare
1	84.154687
2	20.662183
3	13.675550

Mean: 32 (affected by outliers)

Median: 14

Range: 0 to 513

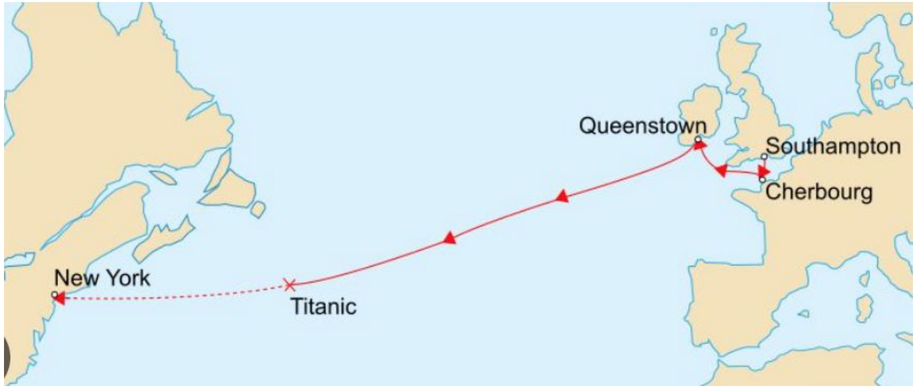
->separating into bins for histogram

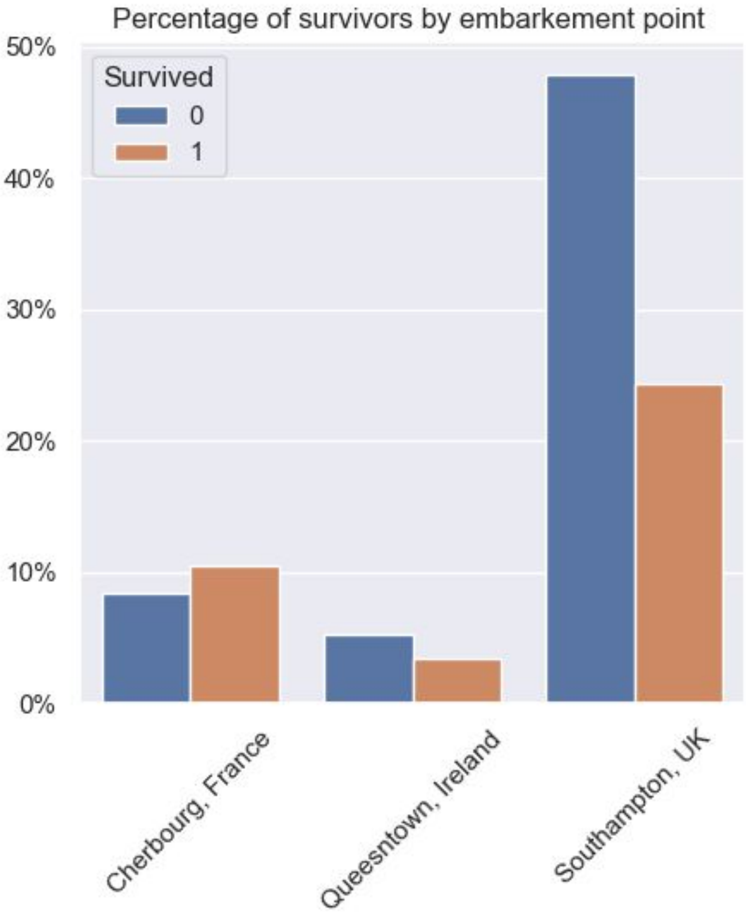


4) **Embarked:** 3 ports to embark from

- Highest survival:

  Boarding from France

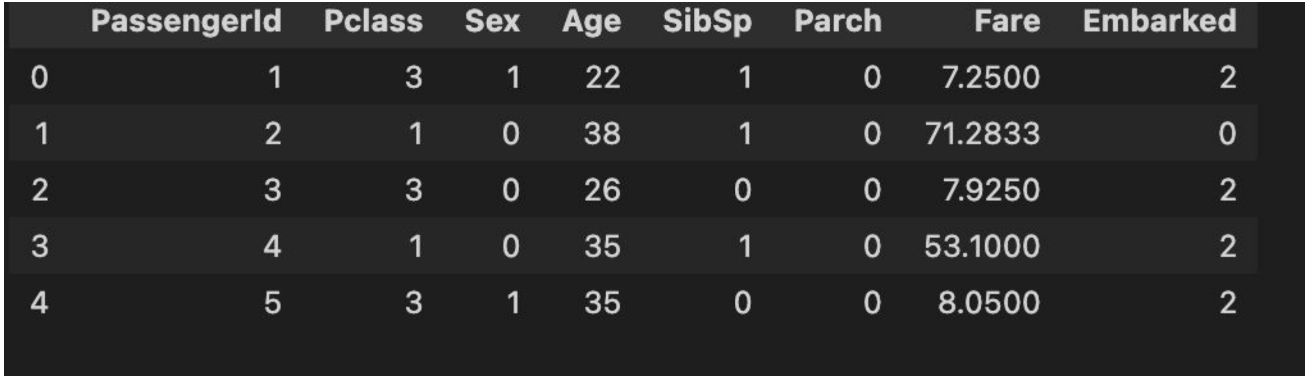




Building Logistic Regression Model -> cleaning

- Dropped ‘Name’ and ‘Ticket’ : words
- Dropped ‘Cabin’ as 77% Nan
- Replaced all missing values of ‘Sex’,’Fare’ with its median and ‘Embarked’ with its mode
- One-hot encoding for ‘Sex’ and ‘Embarked’

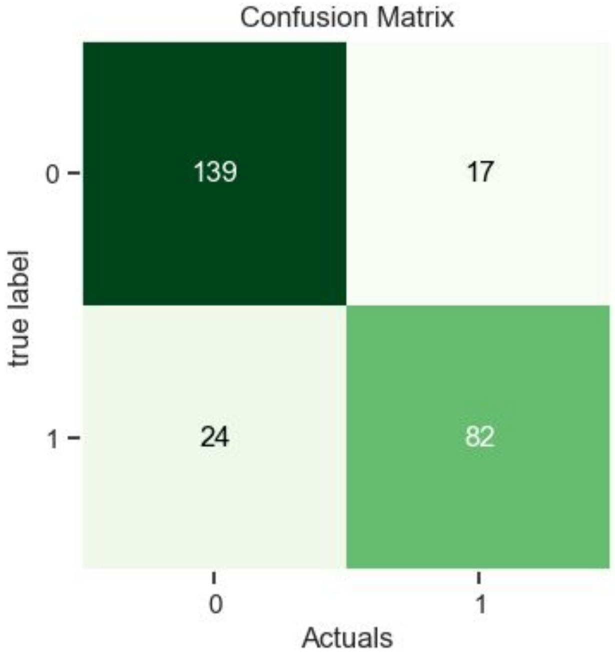
Clean data:

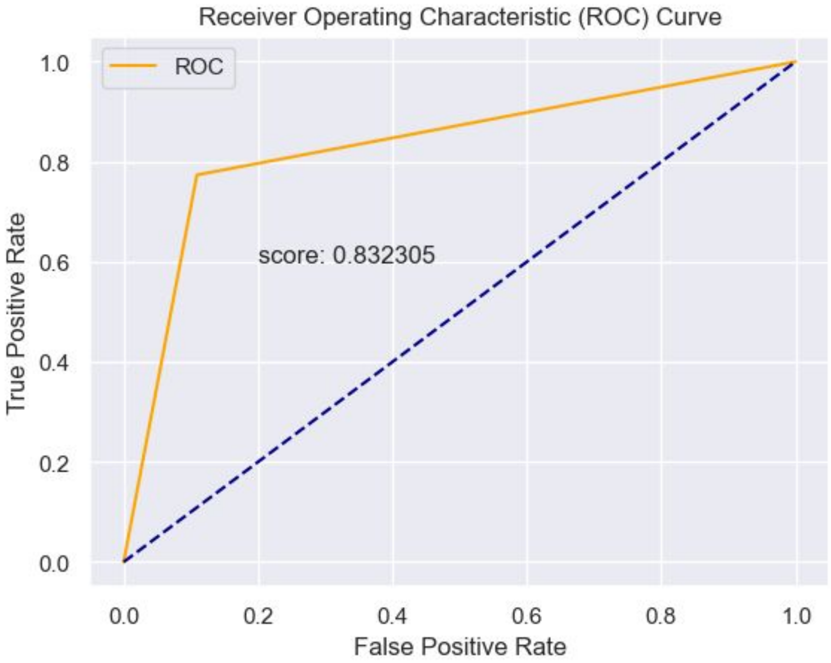


	PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	3	1	22	1	0	7.2500	2
1	2	1	0	38	1	0	71.2833	0
2	3	3	0	26	0	0	7.9250	2
3	4	1	0	35	1	0	53.1000	2
4	5	3	1	35	0	0	8.0500	2

Evaluation of model

Accuracy score: 0.8435





Conclusion

- Our logistic regression model's accuracy is at 84%.

- There are room for improvements therefore there could be further cleaning such as converting age into classes.

Appendix

Github

Week 6

Date: 09/06/2023