ANALYSIS OF TITANIC DATASET USING PYTHON

Dataset link

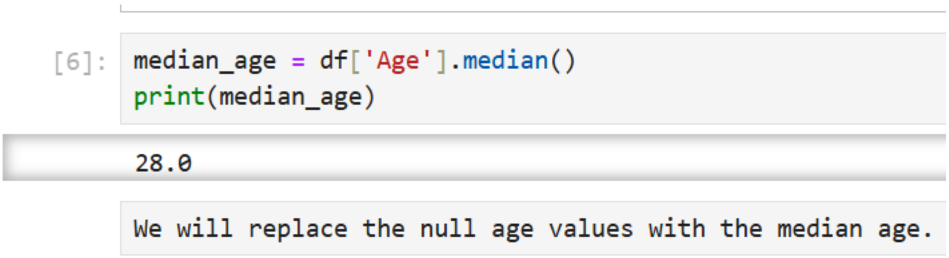
GitHub link

Variable	Definition
survival	Survival 0 = No, 1 = Yes
pclass	Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex
Age	Age in years
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation, C = Cherbourg, Q = Queenstown, S = Southampton

1. Data Cleaning

Nulls were found in the columns ‘Age’, ‘cabin’ and ‘embarked’.

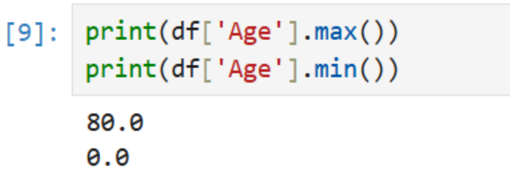
The missing age values were replaced with the median age.



```
[6]: median_age = df['Age'].median()
     print(median_age)

     28.0
```

We will replace the null age values with the median age.



```
[9]: print(df['Age'].max())
     print(df['Age'].min())
```

```
80.0
0.0
```

The reason we chose median is because the age range is widely distributed from 0 to 80. A mean value could have been skewed depending on the count of old and young people.

The NULLs in embarked were replaced with ‘Other’ and the column cabin had lots of missing information so it was removed.

For the analysis, columns ticket, Name and cabin were not considered.

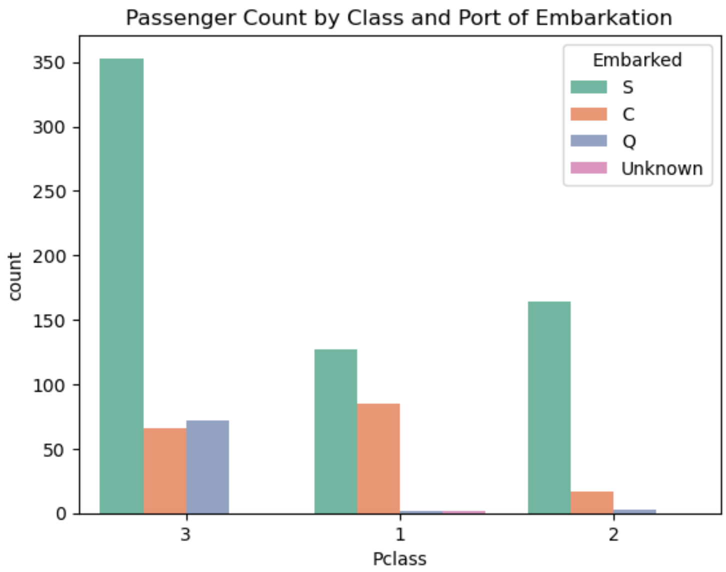
A new column named ‘family size’ was added to the table.

FamilySize = sibsp + parch + 1 (the person itself)

2. Analysis and Visualization

Customer Demographic:

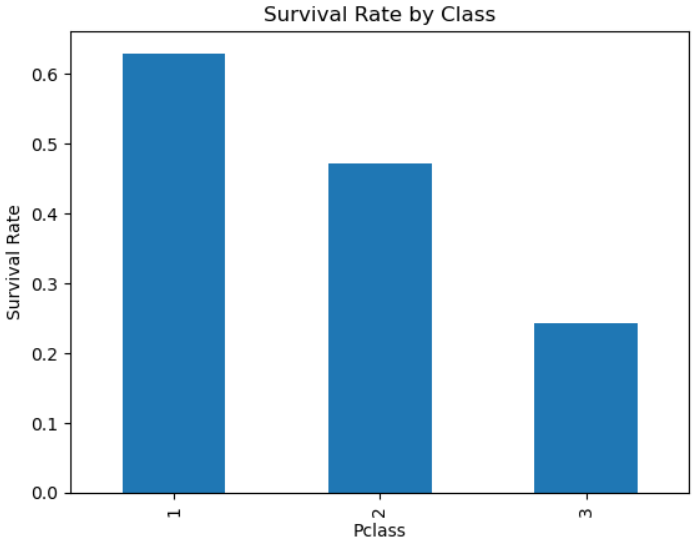
- The mean age on the ship was 29 years old. The mean family size was 2 people, and the mean fare paid was $32.20.
- 314 out of the 891 people were females and the rest 577 were males.
- 216 were sitting in 1st class followed by 184 in 2nd and 491 in 3rd. Most people had 3rd class tickets.
- Most of them boarded from port S: 644, followed by Q: 77 and C: 168. Other: 2
  C = Cherbourg, Q = Queenstown, S = Southampton.



The bar graph above shows that most of the people with a 3rd class or a 2nd class ticket boarded at port S. Port C was used mostly by the 1st class.

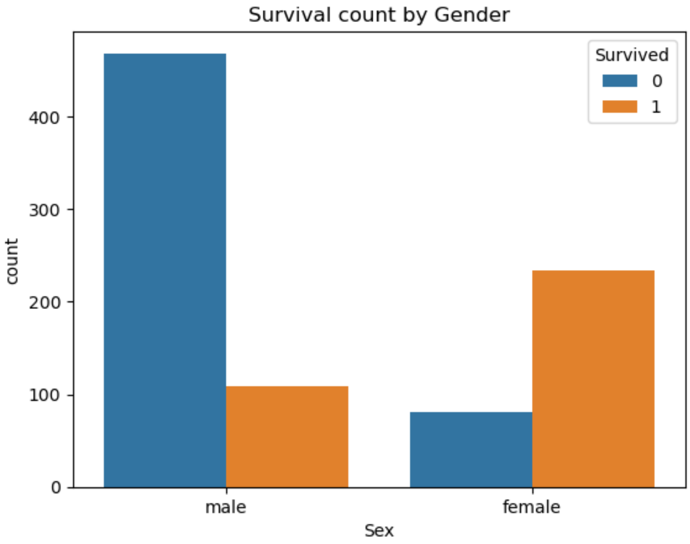
Survival rate by class:

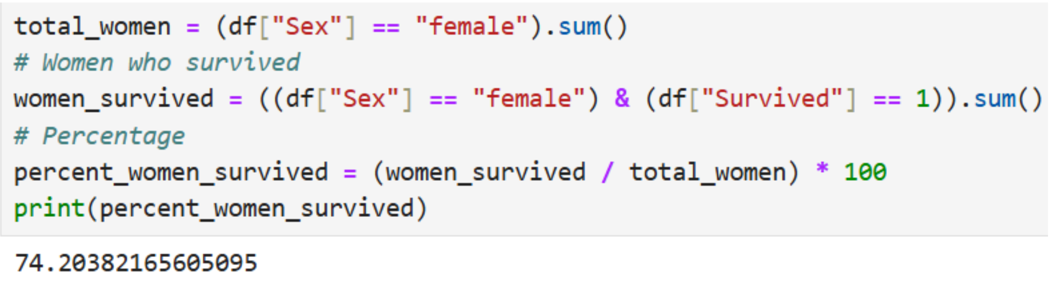
The chart below shows that around 60% passengers from 1st class survived and only 20 to 30 percent from 3rd class lived. Which means 1st class was prioritized. This is not shocking considering the financial bias that we see even in the present day.



Survival by Gender:

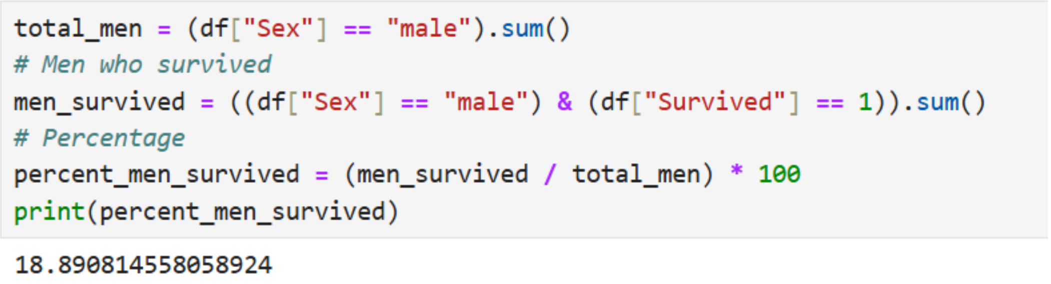
It’s a known fact that mostly female passengers survived the titanic wreck. The graph proves it. 70% of the women on board survived, meanwhile only 19% men survived.





```
total_women = (df["Sex"] == "female").sum()
# Women who survived
women_survived = ((df["Sex"] == "female") & (df["Survived"] == 1)).sum()
# Percentage
percent_women_survived = (women_survived / total_women) * 100
print(percent_women_survived)
```

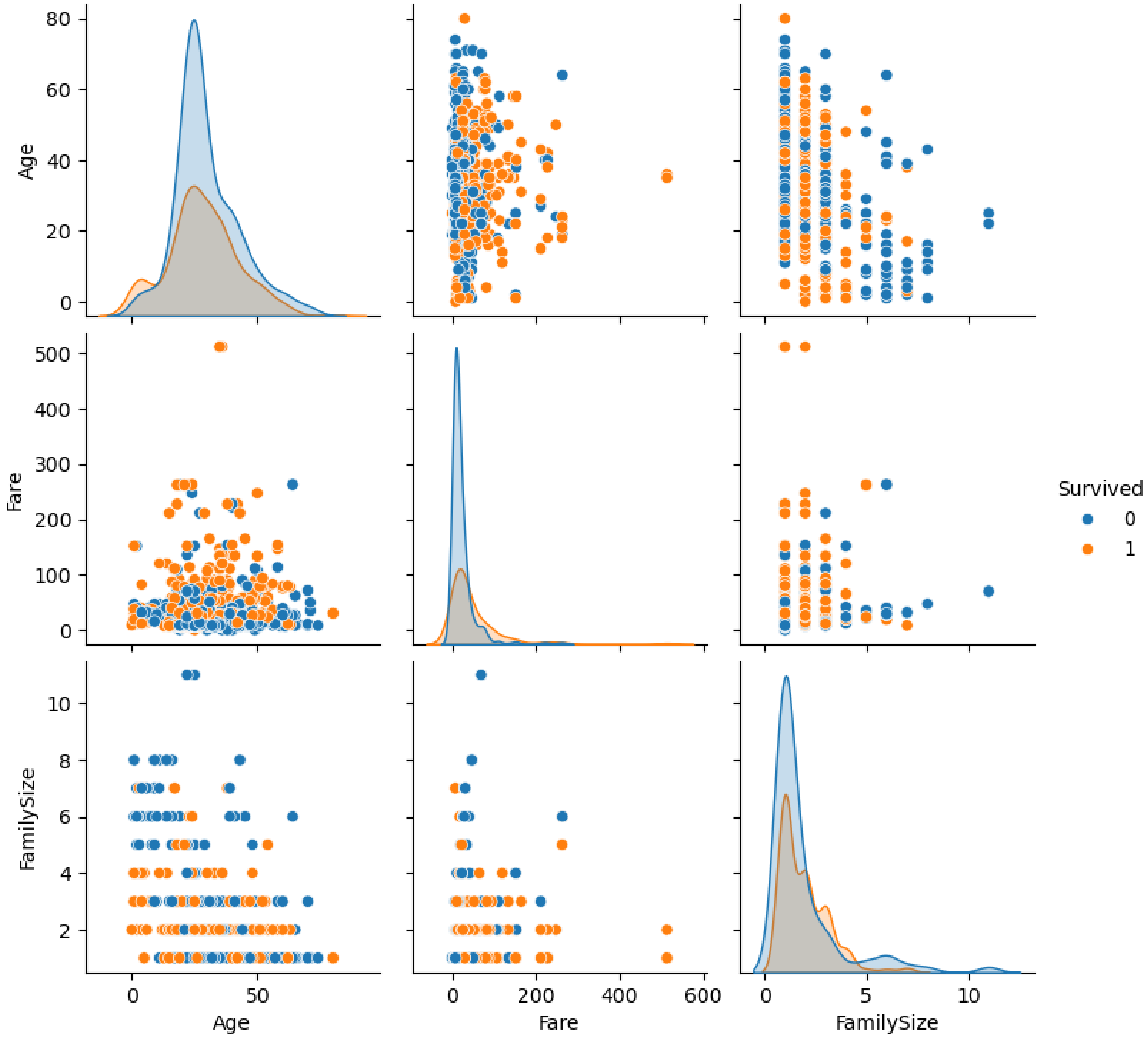
```
74.20382165605095
```



```
total_men = (df["Sex"] == "male").sum()
# Men who survived
men_survived = ((df["Sex"] == "male") & (df["Survived"] == 1)).sum()
# Percentage
percent_men_survived = (men_survived / total_men) * 100
print(percent_men_survived)
```

```
18.890814558058924
```

Pair plot of Age, Fare and Family Size.



- Top Graphs

The first graph shows age distribution by survival. It shows most deaths occurred from the range 20 to 30 and most survivors were aged somewhat below. Survival probability seems higher for very young and slightly better for older adults compared to the peak non-survival age. The next graph also shows that people of all ages who paid low fare failed to live. Third chart shows that people from many different age groups with a smaller family size survived.

- Middle Graphs

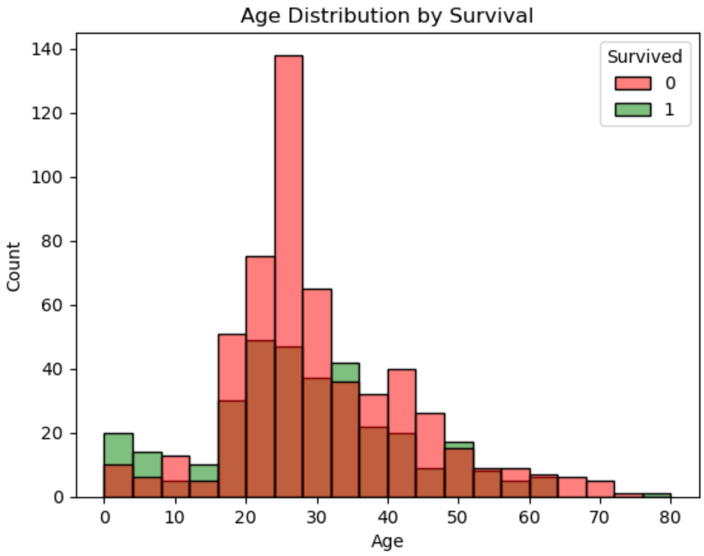
First graph shows survival rate increases with fare for almost all age groups. Children with low fares also seemed to have survived, meaning children were definitely looked after. In the middle graph, distribution for those who survived (orange) is wider and is extending into the higher fare values. This shows high fare significantly increased the chance of survival. Third graph shows higher probability of survival for small families who paid higher fare. The orange dots are moving upwards with fare.

- Bottom Graphs

Out of people who paid low fares, smaller families survived. Families with a size of greater than 5 couldn't be saved. Survival was high for smaller families.

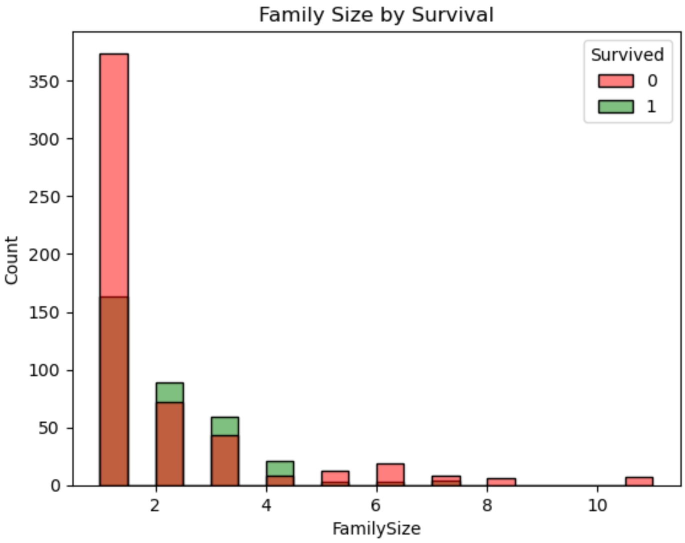
Age distribution by survival:

The histogram below shows that mostly people aged 0 to 15 survived and very few from the age range 30 to 40 years old also did. Range 20-30 has the lowest survival rate.



Family size by survival:

The histogram shows that passengers with a family size of 2 to 5 most likely survived. People from larger families sadly did not, as evacuating large families is very hard. Also, majority of the people were travelling alone but they also have a lower survival rate.


Family Size by Survival
Survived
Count
FamilySize