EDA ON TITANIC DATASET

sharik pathan

Introduction

"Good morning/afternoon/evening everyone, and welcome to my presentation on the Titanic dataset. The sinking of the Titanic remains one of the most tragic and memorable events in history. In 1912, the luxurious ocean liner, the Titanic, sank on its maiden voyage after colliding with an iceberg, resulting in the loss of more than 1,500 lives.

This dataset provides valuable insights into the passengers and crew aboard the Titanic, including demographic information, socio-economic status, and survival rates. Through analysis and visualization of this data, we can gain a deeper understanding of the events that unfolded during this disaster.

In this presentation, we will explore the Titanic dataset and uncover interesting patterns and trends that emerge from the data. We will examine factors such as age, gender, class, and family status to determine their impact on survival rates.

Without further ado, let's dive into the Titanic dataset and see what it can teach us about this tragic event in history."

The aforementioned project was executed under the expert guidance of Professor Anna Androvitsanea and was presented by Sharik Pathan, who is currently enrolled in GTM 05 (Data Science) for the subject of Information and Learning Procedures. The project entailed performing Exploratory Data Analysis using Python in Jupyter Notebook. The resultant analysis provides valuable insights that address the 8 questions set forth by the professor. The source code for this project will be made available on GitHub at the following link. Additionally, the titanic.xls file contains data pertaining to 887 actual passengers aboard the Titanic. This dataset furnishes information regarding the passengers and includes a column that indicates whether each passenger survived (denoted as "1") or did not survive (denoted as "0").

The columns in the dataset are:

passengers: Passenger Identity

survived: Whether passenger survived or not

pclass: Class of ticket

sex: Sex of passenger (Male or Female)

age: Age of passenger

sibSp: Number of sibling and/or spouse travelling with passenger

parch: Number of parent and/or children travelling with passenger

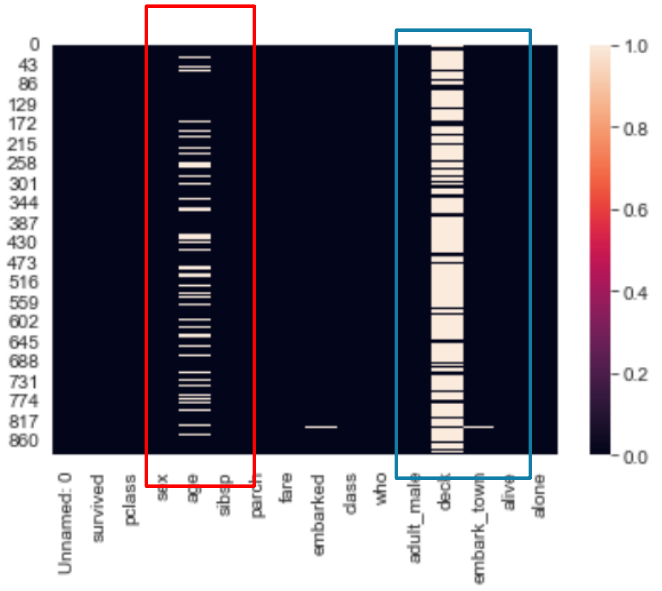
ticket: Ticket number

fare: Price of ticket

cabin: Cabin number

Finding out missing values

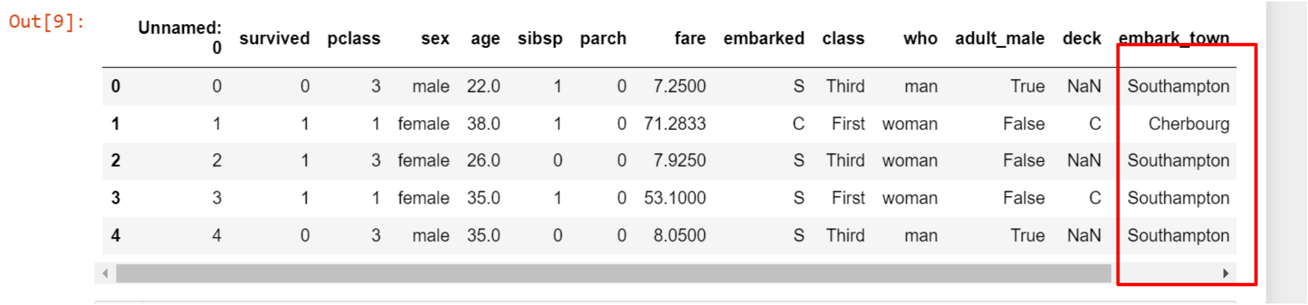
To gain insight into missing values, I utilized a heatmap and employed the seaborn library.



Based on my analysis, it appears that approximately 20 percent of the Age data is missing. This proportion is likely small enough to be replaced reasonably with some form of imputation. However, when examining the Deck column, it became apparent that too much data is missing to make any meaningful conclusions at a basic level. Hence, it was found that both the Age and Deck columns have a significant number of missing values.

Question 1: How many and what where the names of the ports?

There were two ports namely Southampton and Cherbourg.

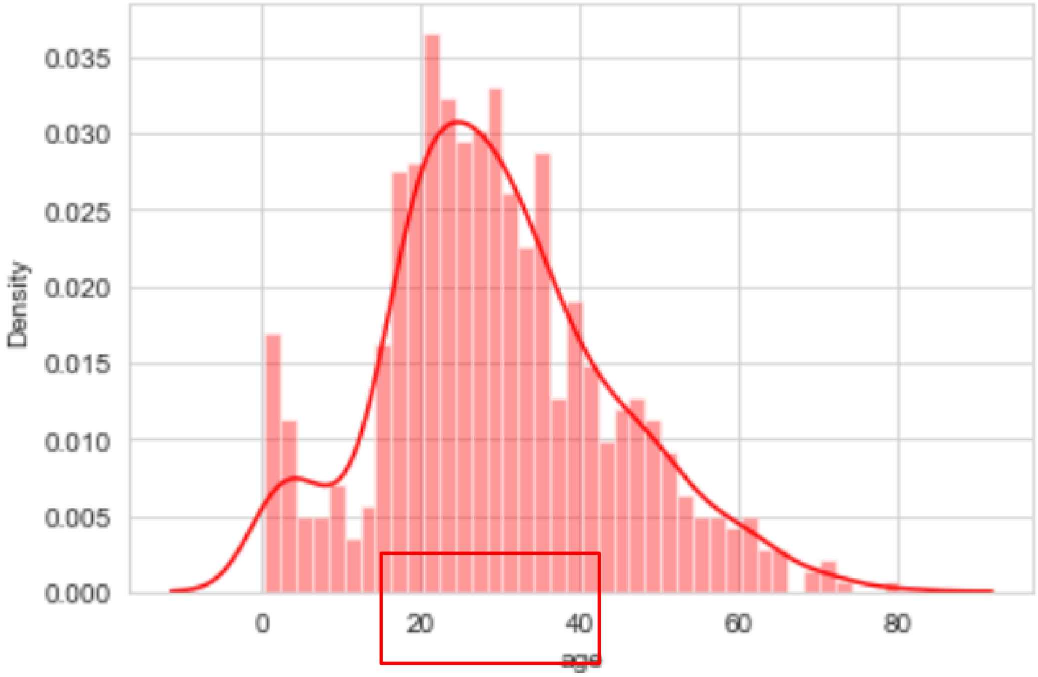


```
Out[9]:
```

	Unnamed: 0	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town
0	0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton
1	1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg
2	2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton
3	3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton
4	4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton

Question 2: Which was the average age of the passengers?

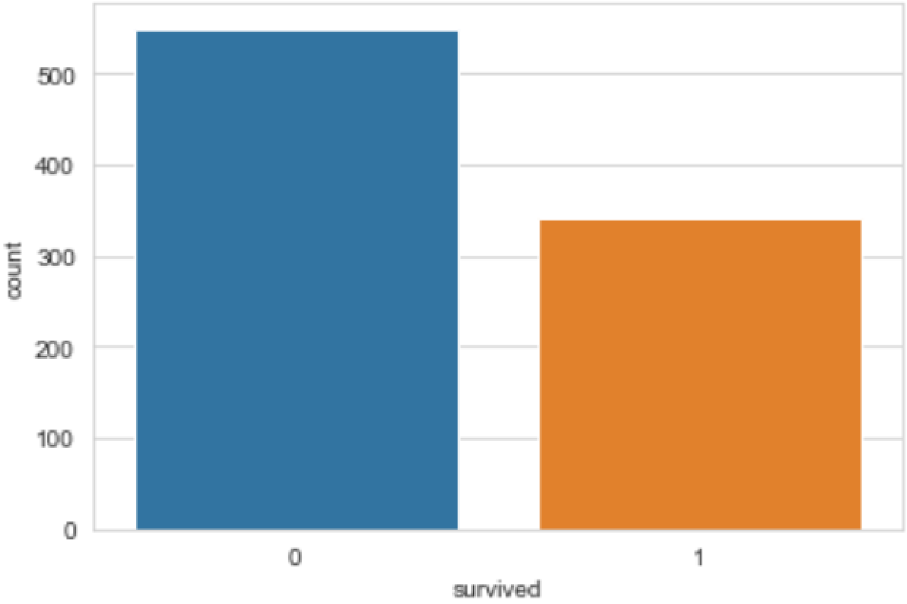
The diagram below depicts that the average age of Titanic passengers fell within the range of 17 to 40 years old. More specifically, the average age of passengers was calculated to be 28.5 years.



Question 3: How many died and how many survived?

I used Seaborn to visualize this Column and Style white grid

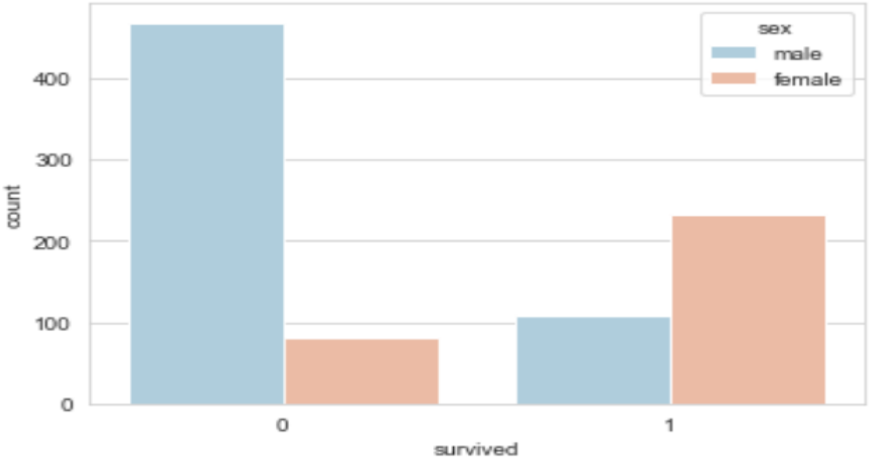
In the diagram below, we can observe that over 550 individuals did not survive, while more than 300 individuals survived. The values of 0 and 1 denote did survive and survived, respectively**.**



Question 4: Is survival connected with any other feature? i.e., age, class, sex etc.? Are i.e., those that are younger more likely to survive?

1. Based on the diagram plotted against sex, it is apparent that over 450 males perished in the disaster, while more females survived. This observation leads us to conclude that, during the Titanic disaster, a higher priority was given to rescuing females.

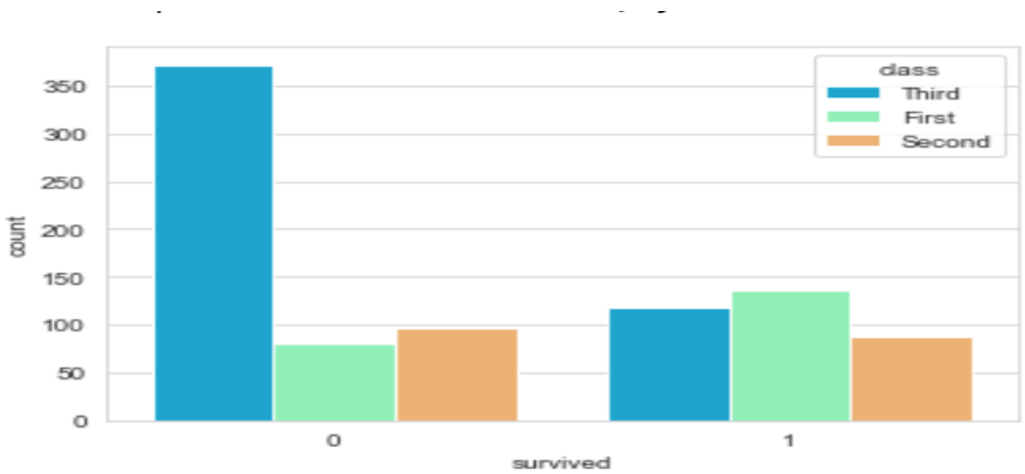
1.Survived vs sex



2. Upon comparing survival rates with passenger class, it is evident that a greater proportion of first-class passengers survived the Titanic disaster, while third-class passengers had the lowest survival rates. However, it is important to exercise caution when interpreting these results as there were more passengers in the second and third-class cabins. Notwithstanding, it is

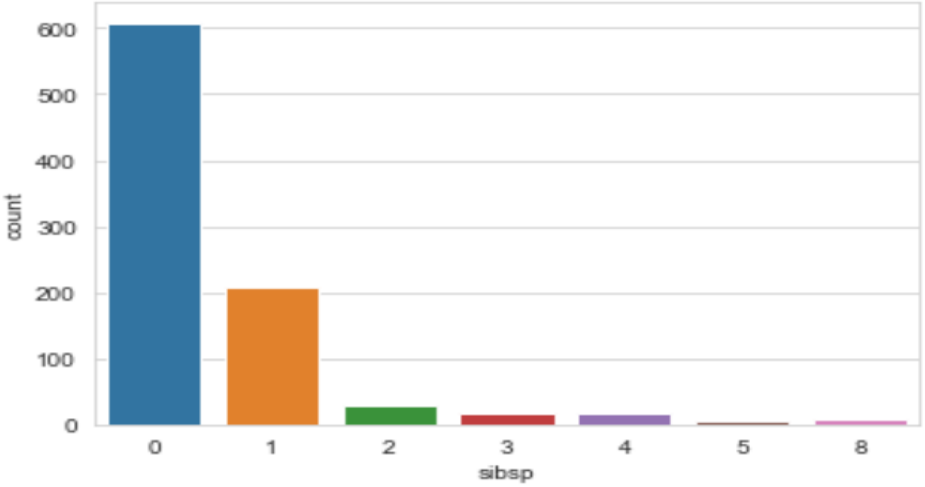
clear that individuals with higher socio-economic status were given priority when it came to rescue operations.

2.Survived vs class



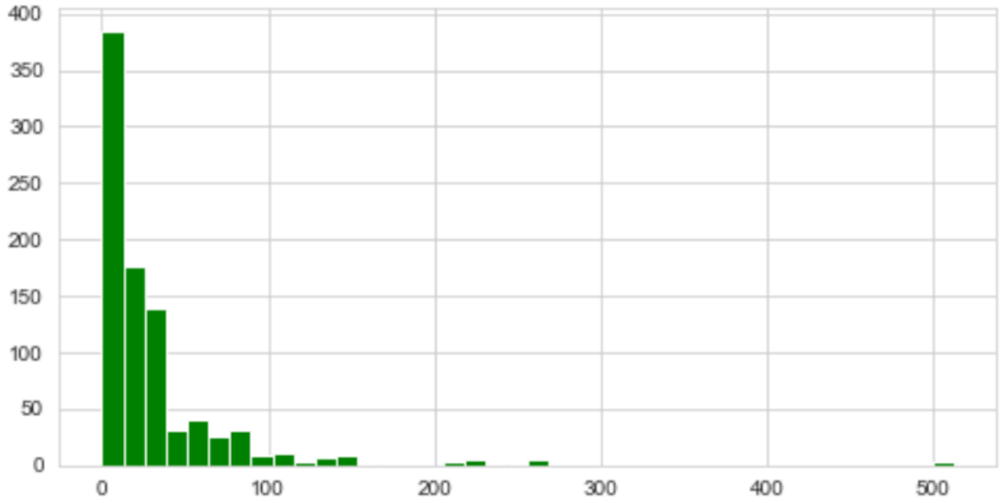
Question 5: How many families where on board?

From the analysis, it appears that there were relatively few families on board the Titanic. The diagram below illustrates that the highest number of individuals who boarded the ship did not have families, followed by those who had a spouse (denoted as 1), then individuals who had one child (denoted as 2), and so forth. Therefore, if we combine the numbers of people who had families, we can estimate that the total would exceed 250 individuals.



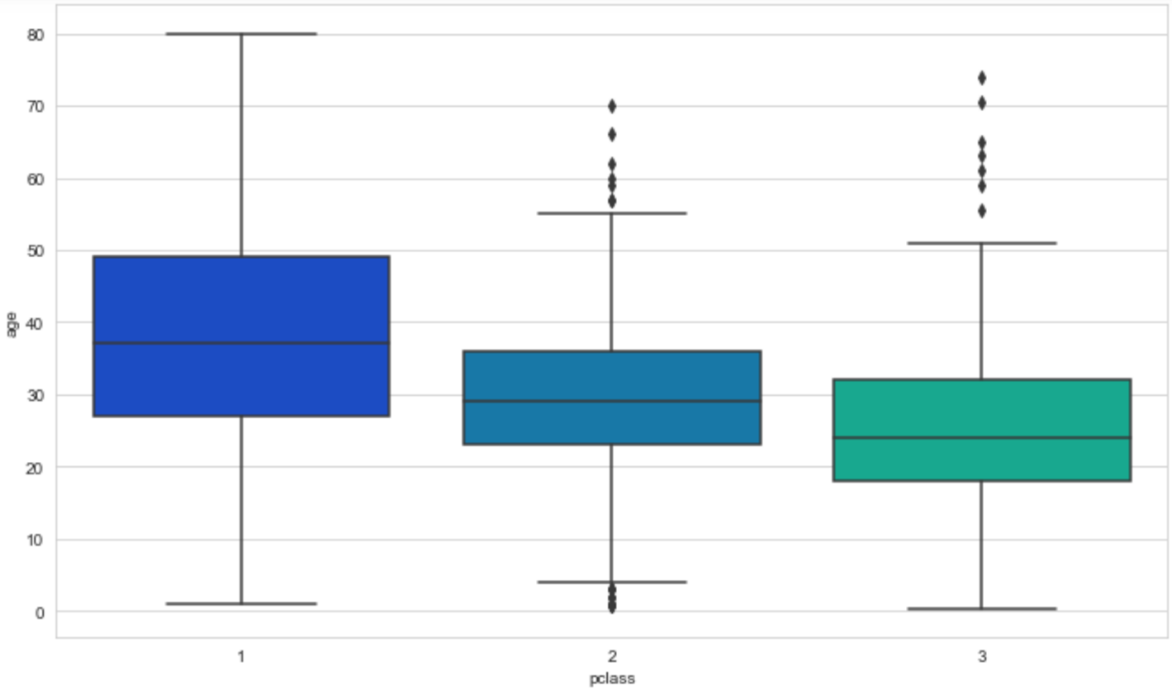
Question 6: What was the average price of the tickets?

We can see that the average fare is 470



Question 7: Which was the average age of those traveling in the first class?

We can see from the box plot the passengers in the first class tend to be older. The range of the passengers between the age group 28 to 49 were in first class excluding the outliners. The average age of passengers in 1st class is 38.



Question 8: Which was the average age of those traveling in the second and third class?

Similarly, we can see that passengers in the 2nd class were in the age group from 25 to 35 and the average age is 29, also excluding the outliners.

Conclusion:

Based on the observations and analyses conducted on the Titanic dataset, several conclusions can be drawn. Firstly, a higher priority was given to rescuing females during the disaster. Secondly, individuals with higher socio-economic status, particularly those in the first-class cabins, were given priority in rescue operations. Thirdly, there were relatively few families on board the Titanic, with the majority of passengers travelling alone. Additionally, the average fare paid by passengers was approximately $470. Finally, there were noticeable differences in the age distribution of passengers across different cabin classes. Passengers in first-class tended to be older than those in second-class cabins.

I think personally the rescue should have been done based on the age.