Assignment 3 Part-2

Name: P.N.Bhavani

Id: 100437867

Introduction:

An outline of the Titanic dataset's exploratory data analysis (EDA) is given in this report. Understanding the structure of the dataset, dealing with missing values, displaying data distributions, and examining correlations between variables are all steps in the EDA process.

Data Structure: The primary columns in the dataset are as follows:
PassengerId: Each passenger's unique identification number.
Survived: Shows if the traveler made it through (1) or not (0).
Pclass: Class of Passengers (1 being Upper, 2 being Middle, and 3 being Lower).
Sex: The passenger's gender.
Age: The passenger's age.
SibSp: The total number of spouses or siblings on board the Titanic.
Parch: The total number of parents and kids on the Titanic.
Fare: The passenger pays the fare.
Port of embarkation: Cherbourg (C), Queenstown (Q), and Southampton (S).

Summary Statistics:

	PassengerId	Survived	Pclass	Age	SibSp	\
count	891.000000	891.000000	891.000000	714.000000	891.000000	
mean	446.000000	0.383838	2.308642	29.699118	0.523008	
std	257.353842	0.486592	0.836071	14.526497	1.102743	
min	1.000000	0.000000	1.000000	0.420000	0.000000	
25%	223.500000	0.000000	2.000000	20.125000	0.000000	
50%	446.000000	0.000000	3.000000	28.000000	0.000000	
75%	668.500000	1.000000	3.000000	38.000000	1.000000	
max	891.000000	1.000000	3.000000	80.000000	8.000000	

	Parch	Fare
count	891.000000	891.000000
mean	0.381594	32.204208
std	0.806057	49.693429
min	0.000000	0.000000
25%	0.000000	7.910400
50%	0.000000	14.454200
75%	0.000000	31.000000
max	6.000000	512.329200

An overview of the dataset's distribution of different parameters, including passenger class, age, number of parents/children on board, number of siblings/spouses on board, and fare, is given by these statistics.

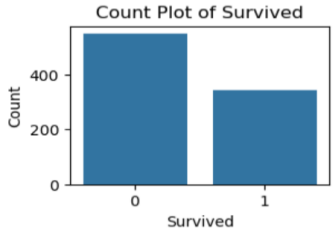
Categorical and Numerical variables distribution:

Based on the inputs/values and their usage we can divide columns to Categorical and Numerical variables

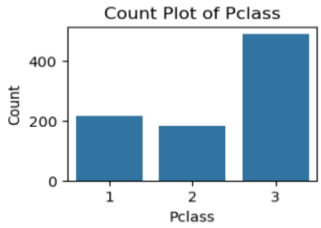
Categorical variables: Survived, Pclass, Sex, Embarked

Numerical variables: PassengerId, Age, SibSp, Parch, Fare

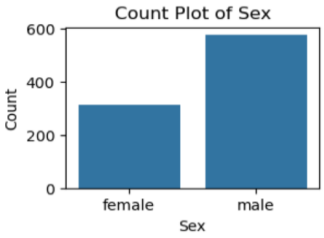
Plotting bar charts for categorical variables:



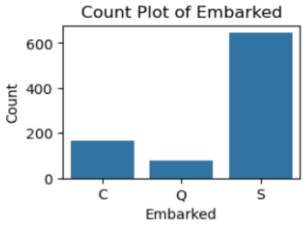
This shows a higher number of individuals who did not survive compared to those who did.



The majority of the passengers are in the third class, followed by the first class and then the second class.

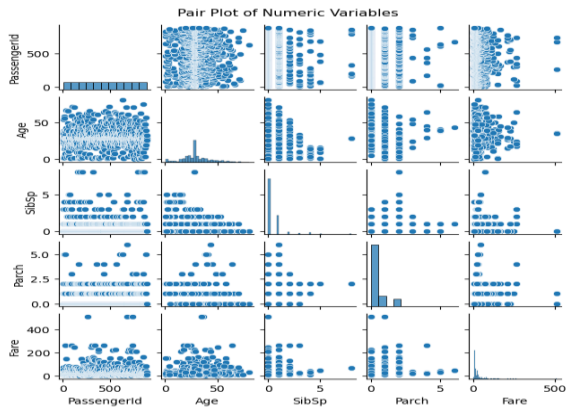


The dataset has a higher number of male entries than female entries.



The majority of individuals in the dataset embarked from location S. This could indicate that location S was a major hub or port during the time of the data collection.

Plotting pair plots for numerical variables:

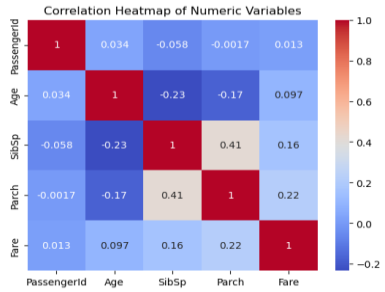


Family Size connection: SibSp and Parch have a positive connection, meaning that travelers who bring their spouses or siblings are also more likely to bring their parents or kids.

Fare Distribution: Some passengers pay extremely high fares, while the distribution of fares is greatly skewed. The median fare may be a more accurate indicator of central trend because this skewness can affect the average fare.

Age Distribution: The distribution of ages is skewed, with a high proportion of young adults, which may be important for more demographic research.

Correlation matrix and heatmap:



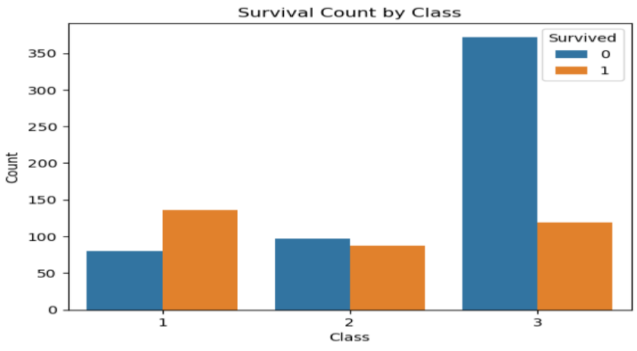
Age: Has a somewhat negative connection with Parch (-0.17) and SibSp (-0.23). This implies that younger travelers typically bring along more parents, kids, and siblings or spouses.
has a weakly positive association (0.097) with fare, suggesting that there is a minor tendency for somewhat older passengers to pay higher fares.

SibSp: Shows a moderately significant connection with Parch (0.41), indicating that passengers are more likely to have parents or kids on board if they have more siblings or spouses.indicates that customers with more siblings or spouses typically pay somewhat higher fares, as evidenced by the weakly positive correlation (0.16) between this variable and fare.

Parch: Indicates that passengers who have more parents or kids on board typically pay higher rates, as evidenced by its weakly positive association (0.22) with fare.

Hypothesis:

a) **Analysis for above Survival Rate vs. Class of Passenger plot below:**



Class 1 (Upper SES):

- There are more survivors (orange bars) than non-survivors (blue bars).
- This indicates that passengers in Class 1 had a higher survival rate.

Class 2 (Middle SES):

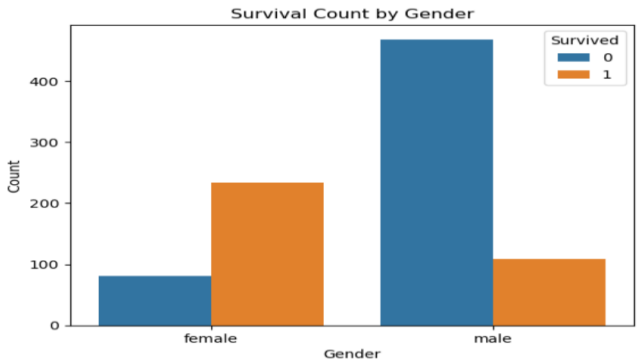
- The number of survivors and non-survivors is fairly balanced.
- This suggests that passengers in Class 2 had a moderate survival rate.

Class 3 (Lower SES):

- There are significantly more non-survivors (blue bars) compared to survivors (orange bars).
- This indicates that passengers in Class 3 had a much lower survival rate.

Conclusion: Based on the bar chart analysis, it can be recommended that the survival rate is linked to the class of the passenger (which can be used as a proxy for SES): Passengers in First Class (Upper SES) had the best survival rate. Passengers in Second Class (Middle SES) had a medium survival rate. Passengers in Third Class (Lower SES) had the worst survival rate. This conclusion sheds light on the effects of social class on the chances of survival during the Titanic disaster, those in higher socio economic classes had better chances of survival.

b) Analysis for above Survival Rate vs. Gender of Passenger plot below:

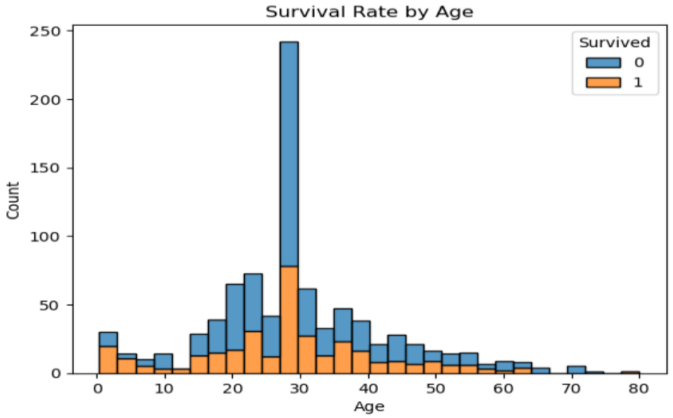


- **Females:** A higher number of females survived compared to those who did not survive. This suggests that females had a higher survival rate.
- **Males:** A significantly higher number of males did not survive compared to those who did survive. This indicates that males had a lower survival rate.

Conclusion:

The analysis concludes that the survival rate was significantly influenced by gender, with females having a greater survival rate than males. This implies that there is a correlation between passenger survival rate and gender.

c) Analysis for above Survival Rate vs. Age of Passenger plot below:



- **Younger Passengers:** There is a noticeable higher survival rate among younger passengers, particularly those around age 0-10. This suggests that younger children had a better chance of survival.
- **Middle-Aged Passengers:** The age group around 20-40 shows a mix of survivors and non-survivors, with a significant portion of non-survivors. The survival rate in this age range is relatively lower.
- **Older Passengers:** Passengers aged 50 and above show fewer survivors, indicating that older passengers had a lower survival rate.

Conclusion:

There was a higher survival percentage among younger passengers (0–10 years old).
With a combination of survivors and non-survivors, middle-aged passengers (20–40 years old) had a reduced survival rate.
The survival rate was lowest for passengers who were 50 years of age or older.
This implies that while older passengers had a reduced chance of surviving, younger passengers—especially children—had a higher chance. The likelihood of surviving the Titanic disaster was significantly influenced by age.