Yamileth Hercules
100385215
CPSC-4800
November 7, 2022

Titanic Exploratory Data Analysis Report

Introduction

Titanic dataset has 891 rows and 12 columns. The objective of this report is to analyze the different Hypotheses of the associations between the variable Survival with Passenger Class, Sex and Age. Therefore, Survival is the dependent variable and the rest of the variables are independent variables.

Survival has 2 categories: 0 (not survived), 1 (survived)

Passenger Class has 3 categories: 1, 2, 3 (Integer values)

Sex has 2 categories: Male and Female (Object type)

Age ranges from 0 to 80 years.(Float type)

Missing Values

The columns that have missing values are:

Column	Frequency	Percentage (%)
Cabin	687	77.10
Age	177	19.86
Embarked	2	0.22

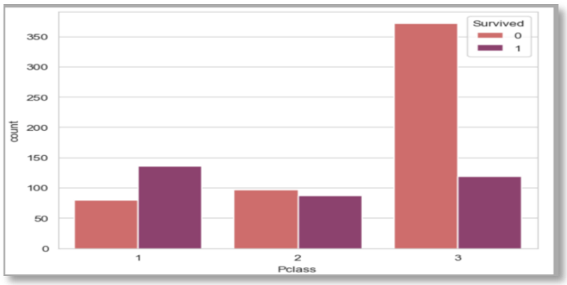
For the Third hypothesis data substitution in Age variable was needed. The substitution was done with the mean.

Summary Statistics

- The percentage of people who died was 61.61% and people who lived were 38.38% .
- In Titanic 577 were female and 314 male. The percentage of female were 64.76% and male were 35.24% .
- People from first class represented 24.24% , second class 20.65% and third class 55.11% of people.
- There are 12 variables in dataframe, 7 of them are categorical variables, the rest numerical.
- The maximun value of the fare was $512.33.
- Age ranges from 0.42 yrs(min) to 80 yrs (max). Age distribution of passenger aboard shows people were between ages 20-30.

Hypothesis 1

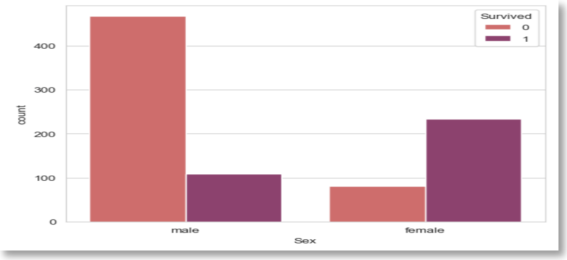
Determine if the survival rate is associated to the class of passenger



- 75% of people from third class died.
- 62.96% percent of people from first class lived.
- The graph clearly show that the 1st class had more posibilities to survived than the rest.
- Indeed, survivability seems to be correlated with the Pclass taking in consideration Chi Square Test.

Hypothesis 2

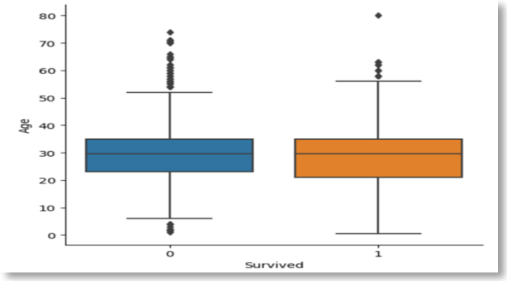
Determine if the survival rate is associated to the gender



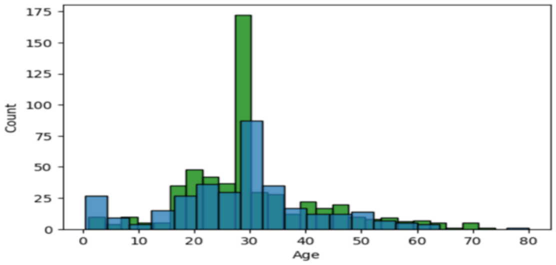
Yamileth Hercules
100385215
CPSC-4800
November 7, 2022

- Just 18.89% of male survived.
- 74.20% of female survived.
- People who were female had more possibilities to live.
- Indeed, survivability seems to be correlated with the Gender, since Chi Squared Test shows that the variables are dependant from each other.

Hypothesis 3
Determine if the survival rate is associated to the Age



- The boxplot shows that there were more extreme values from people who did not survived than people who survived.
- The average age from people who survived were less (28.54) than people who not survived (30.41).



- Almost 60% of children in the range of [0-10] survived.
- The distribution of the people who survived and did not survive are central and almost symmetric with the mean, slightly positive skewed. Showing Children seems had more possibilities to live.
- Indeed, survivability seems to be correlated with the Age, since Chi Squared Test shows that the variables are dependant from each other.

Conclusions

Hypothesis 1: Pclass and Survived are dependent each other. First Class had more possibilities to live than Third Class.
Hypothesis 2: Sex and Survived are dependent each other. Female had more possibilities to live than Male.
Hypothesis 3: Age and Survived are dependent each other. Children in range of (0-10) years old had more possibilities to live than people in other ranges.

To analyze Ages as categorical variable, ranges were generated.
The variables with missing variables (Cabin and embarked were ignore). Age was filled with the mean.