CPSC 4800 - ASSIGNMENT 3.2

EXPLORATORY DATA ANALYSIS ON TITATIC DATA SET

Thi Thu Thuy Tran
Student ID: 100420937

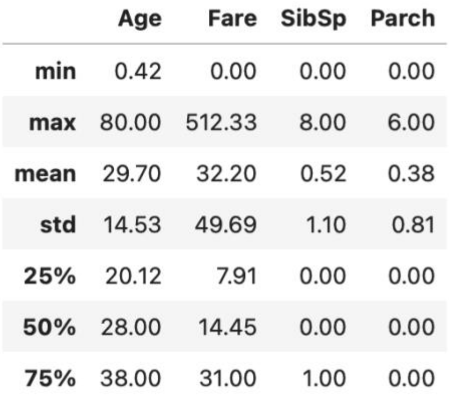
Instructor: Nasim Tabatabaei

Overview

Titanic data set contains information of 891 passengers, including their survivor status, who boarded the ship in 1912. There are 891 observations and 12 varibles:

- **Numerical:** *Float:* Age, Fare. *Integer:* PassesgerId, SibSp, Parch
- **Categorical:** *Ordinal:* Pclass. *Nominal:* Survived, Name, Sex, Ticket, Embarked, Cabin
- The data has 678 missing values for Cabin & 177 missing values for Age

Statistic summary of numerical variables

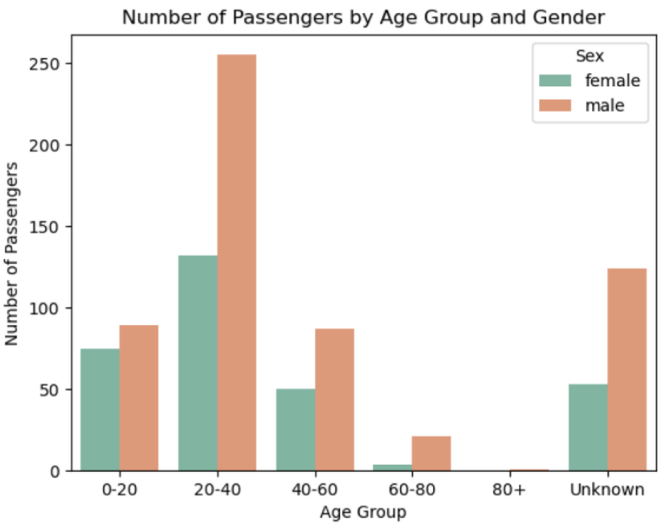


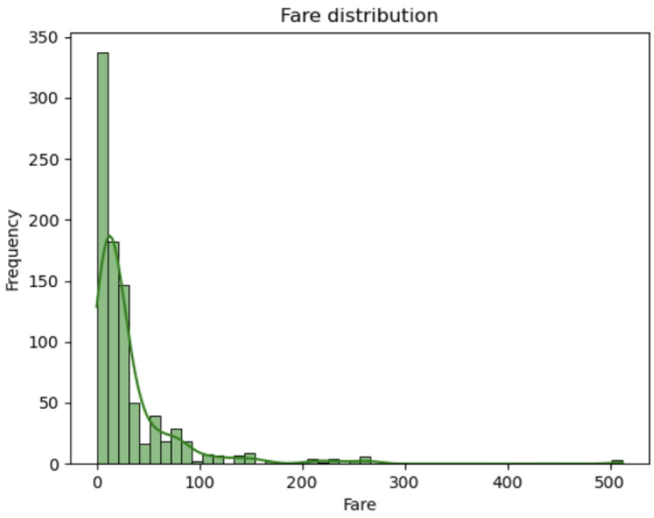
	Age	Fare	SibSp	Parch
min	0.42	0.00	0.00	0.00
max	80.00	512.33	8.00	6.00
mean	29.70	32.20	0.52	0.38
std	14.53	49.69	1.10	0.81
25%	20.12	7.91	0.00	0.00
50%	28.00	14.45	0.00	0.00
75%	38.00	31.00	1.00	0.00

Statistic summary table provides information about the range (min, max), average and variability (IQR) of each numberical variable:

- **Age:** Passenger age range from 5 months old to 80 years old with age average of ~ 30 yeard old
- **Fare:** Ticket is free or upto $512.33 with average of $32.2
- **SubSP:** Passenger travel alone, as a couple or up to 8 siblings travel together
- **Parch:** passenger was travel alone or up to 6 members in a family travel together (1 or 2 parents and the rest are children)

Data distribution





- Majority of passesgers are within 20 to 40 years old

- There are more male passengers than female ones across all age group

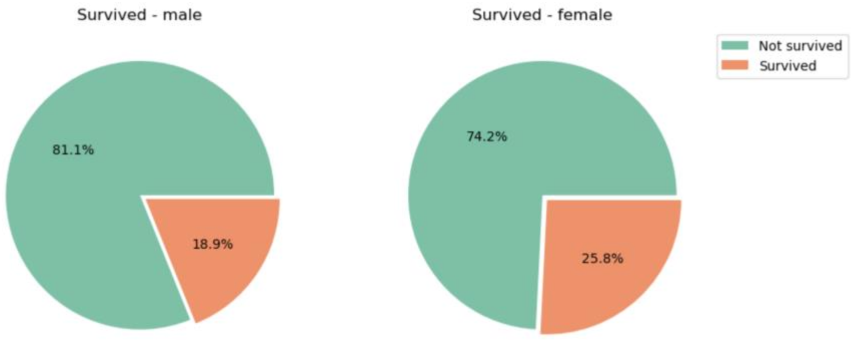
- 20% of passengers did not have their age recorded

- The distribution of ticket is highly skewed: there is a wide range of ticket price but majority of fare is 30 dollars and less.

- There are free tickets but there are also some very expensive tickets and could be more than 500 dollars

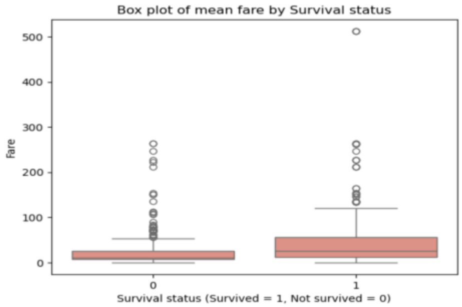
Survival by gender:

The proportion of female passengers who survived is 25.6%, about 6% higher than male counterpart



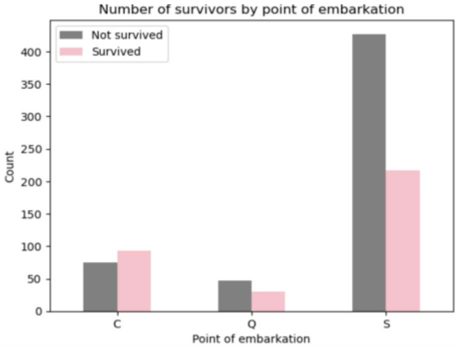
Hypotheses testing

1. **Passengers with higher fares are more likely to survive:** t-test with CI = 95%, plotting box plot



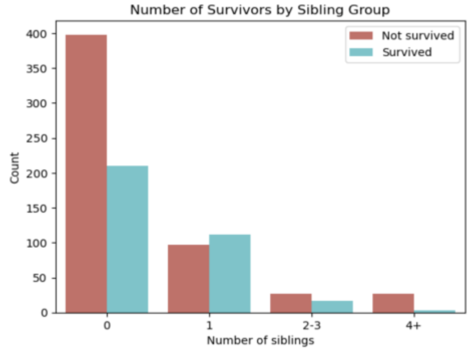
- **t-test = 7.939** is very significant. It suggests the mean fare of survived passenger is significantly different from the mean fare for non-survived ones (7.939 times standard deviation)
- **p_value < 0.05** suggest that statistically, there is significant evidence that there is difference between the mean fare of survived passengers and non-survived passengers
- **Box plot** indicates that on average, survived passengers paid higher fair than non-survived ones.

2. **Survival is associated with point of embark:** chi-squared test with CI = 95%, plotting bar chart



- **chi-squared = 26.489** suggests that there is difference between observed frequencies(or actual counts of survivors and non-survivors for each point of embark) and expected frequencies (expected counts of counts of survivors and non-survivors if there is no relationship between point of embark and survival)
- **p_value < 0.05** suggest that statistically, there is significant evidence that there is relationship between point of embark and survival
- **Bar plot** shows that at embarkation S & Q, more passengers were died than those were survived

3. **Passengers with siblings are more likely to survive:** t-test with CI = 95%, plotting bar chart



- **t-value = -1.054** indicates that the mean number of siblings of those who were survived is slightly less than that of those who were not survived.

- **p_value > 0.05** suggests that statically, there is no significant evidence suggest that passengers with siblings are more likely to survive than those without siblings

- **Bar plot** counts number of survived and died passengers, group by number of siblings