6737932\_SIRE506\_Assignment02

2024-09-08

library(readr)  
train <- read\_csv("/Users/bu/Documents/SIRE/506/Assignment02/Dataset/train.csv")

## Rows: 891 Columns: 12  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Name, Sex, Ticket, Cabin, Embarked  
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Note: Package “readr” is installed to read CSV file by using read\_csv() function.

# Summary Statistic for Titanic by Survival Status

A comprehensive overview of major characteristic of 891 passengers during titanic diaster has been analyzed and categorized base on survival status. Figure 1 reveal that their are 342 survivors whereas there are 549 non-survivors, the number of female passengers show the significant survivals rate as 68% which is greater than male(32%). The reflection of passengers whose traveling none or at least one Siblings/Spouses aboard (SibSp) or Parents/Children aboard (Patch) had the higher potential of survival rates. Moreover, passenger in first-class got the highest survivals rate with 40% followed by 35% of thired-class and least survival at second-class which is 25% of passengers. The median of age has been calculated and indicates central value at 28 years old, however, the unknown ages can affects the analysis of age in the survival pattern. The fare paid of the survived passenger tended to higher than who not survived. The presence of these finding suggests that fare paid could be one of the factors that determined the survival of passengers which could related to the passengers class because the higher paid can indicate the greater service and location in the cabin, therefore, those factors can influence the chance of servive throughout this diaster.

Note: Essential R packages that has been install are “gt”, “dplyr”, and “gtsummary”.

library(gt)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

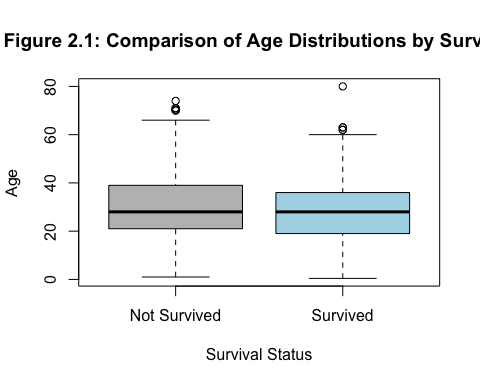
library(gtsummary)  
summary\_table <- train%>% select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare) %>% tbl\_summary(by = Survived)  
summary\_table\_gt <- summary\_table %>%  
 as\_gt() %>%  
 tab\_caption(caption = "Figure 1: Summary Statistic for Titanic by Survival Status") %>%  
 tab\_options(table.width = pct(100))  
summary\_table\_gt <- summary\_table\_gt %>%  
 tab\_footnote(  
 footnote = " Explanatory Notes:  
 Survived: Passenger did not survive (0), Survived passenger (1).   
 Pclass: Passenger class (1, 2, or 3).  
 Sex: Gender of the passenger.   
 Age: Age of the passenger.  
 SibSp: Number of siblings and spouses aboard the Titanic.  
 Parch: Number of children and parents aboard the Titanic.  
 Fare: The price paid (in pounds) for the ticket.",  
 locations = cells\_column\_labels(columns = everything())  
 )  
summary\_table\_gt

| **Characteristic***1* | **0** N = 549*2,1* | **1** N = 342*2,1* |
| --- | --- | --- |
| Pclass |  |  |
| 1 | 80 (15%) | 136 (40%) |
| 2 | 97 (18%) | 87 (25%) |
| 3 | 372 (68%) | 119 (35%) |
| Sex |  |  |
| female | 81 (15%) | 233 (68%) |
| male | 468 (85%) | 109 (32%) |
| Age | 28 (21, 39) | 28 (19, 36) |
| Unknown | 125 | 52 |
| SibSp |  |  |
| 0 | 398 (72%) | 210 (61%) |
| 1 | 97 (18%) | 112 (33%) |
| 2 | 15 (2.7%) | 13 (3.8%) |
| 3 | 12 (2.2%) | 4 (1.2%) |
| 4 | 15 (2.7%) | 3 (0.9%) |
| 5 | 5 (0.9%) | 0 (0%) |
| 8 | 7 (1.3%) | 0 (0%) |
| Parch |  |  |
| 0 | 445 (81%) | 233 (68%) |
| 1 | 53 (9.7%) | 65 (19%) |
| 2 | 40 (7.3%) | 40 (12%) |
| 3 | 2 (0.4%) | 3 (0.9%) |
| 4 | 4 (0.7%) | 0 (0%) |
| 5 | 4 (0.7%) | 1 (0.3%) |
| 6 | 1 (0.2%) | 0 (0%) |
| Fare | 11 (8, 26) | 26 (12, 57) |
| *1* Explanatory Notes: Survived: Passenger did not survive (0), Survived passenger (1). Pclass: Passenger class (1, 2, or 3). Sex: Gender of the passenger. Age: Age of the passenger. SibSp: Number of siblings and spouses aboard the Titanic. Parch: Number of children and parents aboard the Titanic. Fare: The price paid (in pounds) for the ticket. | | |
| *2*n (%); Median (Q1, Q3) | | |

## Identify outliers and the spread of continuoues variables

In order to visuailzing the distribution of continuous variables and identifying outliers, the box plot of the age distribution of passengers (figure 2.1) and the fare paid of passengers (figure 2.2) are use to observed the correlation coefficient of demographic in titanic dataset. Additionally, the box plot analysis are use to support and clarify the data from figure 1.

boxplot(Age ~ Survived,   
 data = train,   
 main = "Figure 2.1: Comparison of Age Distributions by Survival",  
 xlab = "Survival Status",  
 ylab = "Age",  
 names = c("Not Survived", "Survived"),  
 col = c("gray", "lightblue"))



boxplot(Fare ~ Survived,   
 data = train,   
 main = "Figure 2.2: Comparison of Fare Distributions by Survival",  
 xlab = "Survival Status",  
 ylab = "Fare Amount",  
 names = c("Not Survived", "Survived"),  
 col = c("gray", "lightblue"))

