assignment2\_report

Chutiwan Pankram

2024-09-11

## Install packages for summary

install.packages(“gt”) install.packages(“gtsummary”) install.packages(“dplyr”)

## Import Dataset  
library(readr)  
train <- read\_csv("~/GitHub/chu\_SIRE506/assign2/data/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.

View(train)  
  
# create table of summary statistic  
library(gt)  
library(gtsummary)  
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

table1 <- train %>%  
 select(Survived, Age, Fare, Pclass, Sex, SibSp, Parch) %>%  
 tbl\_summary(  
 by = Survived,  
 label = list(  
 SibSp ~ "Number of siblings/spouses aboard",   
 Parch ~ "Number of parents/children aboard",  
 Pclass ~ "Ticket class"  
 ),  
 # Customize the summary statistics for numeric variables to show IQR  
 statistic = list(  
 all\_continuous() ~ "{median} ({IQR})"  
 )  
 ) %>%  
 modify\_header(label = "Variable") %>%  
 modify\_spanning\_header(all\_stat\_cols() ~ "Survival Status")   
  
 # Convert to gt table and add caption  
table1\_gt <- as\_gt(table1) %>%  
 tab\_caption(  
 caption = "Figure 1: Summary Statistic for Titanic by Survival Status"  
 )

# Summary Statistic for Titanic by Survival Status

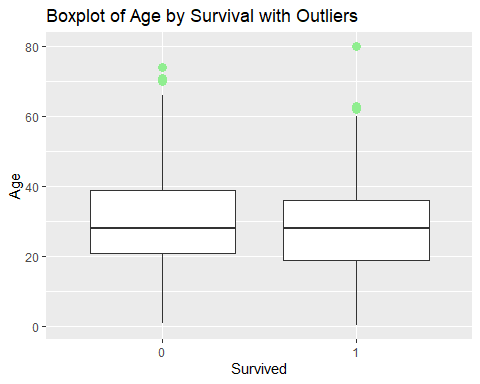
The table 1 shown summarizes the statistics for the Titanic by survival status from passengers total of 891. In the table provided, 0 represents a total of 549 non-survival passengers, and 1 represents a total of 342 survival passengers. The variables include age, fare, ticket class, sex, and number of family members aboard (both siblings/spouses and parents/children). For each variable, the table provides the median values, interquartile ranges (IQR), and percentages for both survivors and non-survivors, allowing a comparison of categorical (pclass, sibsp, parch, and sex) and ticket-related factors between the two groups. Categorical variables measure frequency, so percentages are suitable for analysis. For example, pclass (ticket class) shows that class 3 has the highest casualty rate at 68%, while first-class passengers have a higher survival rate at 40%. Sibsp and parch (siblings or spouses) indicate that those traveling alone have a higher chance of dying (72% for sibsp and 81% for parch) but also a significant survival rate (61% and 68%, respectively). The sex variable shows that women have a higher survival rate at 68%, compared to 15% for men in the non-survival group. Continuous variables, measured by median and standard deviation, include age and fare. The average age of non-survival passengers is 31 (SD 14), while survival passengers are 28 (SD 15), suggesting age is not a significant survival factor. Fare is associated with pclass; non-survival passengers spent an average of 22 (SD 31), while survival passengers spent 48 (SD 67), indicating that higher fares and better classes are linked to higher survival chances. The standard deviation shows variability in fare payments, with higher variability in the survival group. In conclusion, from all the variables used to predict the survival chance, it can be concluded that those who paid the highest fare, which is first class, have a higher chance of survival. The summary statistics show that sex and age indicate that women and children have a higher chance of survival than others. In Parch and Sinsp’s analysis, they show that most passengers traveling alone have a higher chance of survival than those traveling with a family.

table1\_gt

|  | Survival Status | |
| --- | --- | --- |
| Variable | **0** N = 549*1* | **1** N = 342*1* |
| Age | 28 (18) | 28 (17) |
| Unknown | 125 | 52 |
| Fare | 11 (18) | 26 (45) |
| Ticket class |  |  |
| 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%) |
| Number of siblings/spouses aboard |  |  |
| 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%) |
| Number of parents/children aboard |  |  |
| 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%) |
| *1*Median (IQR); n (%) | | |

library(ggplot2)  
# create boxplot of age by survival with outliers  
boxplot\_age <- ggplot (train, aes(x = as.factor(Survived), y = Age)) +  
 geom\_boxplot(outlier.colour = "lightgreen", outlier.shape = 16, outlier.size = 3) +  
 labs(x = "Survived", y = "Age", title = "Boxplot of Age by Survival with Outliers")  
print(boxplot\_age)

## Warning: Removed 177 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



# create boxplot of fare by survival with outliers  
boxplot\_fare <- ggplot (train, aes(x = as.factor(Survived), y = Fare)) +  
 geom\_boxplot(outlier.colour = "lightslateblue", outlier.shape = 16, outlier.size = 3) +  
 labs(x = "Survived", y = "Fare", title = "Boxplot of Fare by Survival with Outliers")  
print(boxplot\_fare)

