assignment2

Nattha Bunyamani

2024-09-12

## 1.Coding

### Import a CSV file (train.csv, Titanic dataset)

library(readr)  
train <- read\_csv("../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)

### Install used packages

for creating Table 1 -> install.packages(“gtsummary”), install.packages(“gt”), install.packages(“dplyr”)

for creating Boxplot -> install.packages(“ggplot2”)

### Create table of summary statistics

library(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

#create a gtsummary table  
table1 <- train %>%  
 select(Survived, Age, Fare, SibSp, Parch, Sex, Pclass) %>%  
 tbl\_summary(  
 by = Survived,  
 statistic = list(  
 all\_continuous() ~ c("{median} ({IQR})"),   
 all\_categorical() ~ "{n} ({p}%)"  
 )  
 ) %>%  
 modify\_header(label = "\*\*Characteristic\*\*") %>%   
 modify\_spanning\_header(all\_stat\_cols() ~ "\*\*Survival Status\*\*") %>%  
   
#convert a gtsummary table into a gt table object, for adding header and footnote  
 as\_gt() %>%   
 gt::tab\_header(  
 title = "Table 1: Summary statistics of Titanic Passenger Data by Survival Status"  
 ) %>%  
 gt::tab\_footnote(  
 footnote = "0 = Not survived, 1 = Survived",  
 locations = gt::cells\_title(groups = "title"  
 ))

### Create boxplot for checking Outliers

library(ggplot2)   
  
#create a boxplot for checking Outliers in Age  
boxplot\_age <- ggplot (train, aes(x = as.factor(Survived), y = Age)) +  
 geom\_boxplot(outlier.colour = "seagreen", outlier.shape = 17, outlier.size = 3) +  
 scale\_x\_discrete(labels = c("0" = "Not Survived", "1" = "Survived")) +  
 labs(x = "Survival Status", y = "Age", title = "Boxplot 1: Distribution of Age by Survival Status (With Outliers)")  
  
#create a boxplot for checking Outliers in Fare  
boxplot\_fare <- ggplot (train, aes(x = as.factor(Survived), y = Fare)) +  
 scale\_x\_discrete(labels = c("0" = "Not Survived", "1" = "Survived")) +  
 geom\_boxplot(outlier.colour = "coral3", outlier.shape = 17, outlier.size = 3) +  
 labs(x = "Survival Status", y = "Fare", title = "Boxplot 2: Distribution of Fare by Survival Status (With Outliers)")

## 2. Interpreting the information

The Titanic dataset provides the information of 891 passengers, along with 12 characteristics, including passenger ID, name, age, sex, passenger fare (fare), ticket number, cabin number, number of siblings or spouses aboard (SibSp), number of parents or children aboard (Parch), port of embarkation (Embark, with C = Cherbourg, Q = Queenstown, S = Southampton), and passenger class (Pclass, with 1 = upper class, 2 = middle class, 3 = lower class) along with the survival status of each passenger (0 = not survived, 1 = survived).

According to Table 1, six key characteristics, including age, fare, SibSp, Parch, sex, and Pclass, are selected for comparison with survival status as their potential influence on the survival rate. There are 342 people, or around 38% of the passengers, who survived the incident, whereas 549 people (62%) did not survive. For continuous variables like age and fare, the median and IQR are used to represent the central tendency and variability since the data contain many outliers, as shown in Boxplot 1 and 2. The median age is almost the same for both survivors and non-survivors, which is 28 years, with an IQR of 18 years for non-survivors and 17 years for survivors. However, there are 177 missing records of age (125 for non-survivors and 52 for survivors). From the result, age may not be suitable for predicting the influence on the survival rate, as the amount of missing information is significant. The median fares of the non-survivors and survivors are 11 pounds with an IQR of 18 pounds, and 26 pounds with an IQR of 45 pounds, respectively. It can be inferred that people who paid more for their tickets tend to survive the incident. For discrete variables like SibSp and Parch, as well as categorical variables like sex and Pclass, percentages are used to represent the proportion of each category. For both SibSp and Parch, passengers with 0 SibSp or Parch have the survival rate of 61% or 68% for survivors and 72% and 81% for non-survivors, respectively. However, the survival rates tend to decrease as the number of SibSp or Parch increases. As regards sex, the majority of female passengers who survived is 68%, compared to 15% of those who did not survive. In contrast, 85% of non-survivors are male, while male survivors are around 32%. It can be inferred that females are more likely to survive compared to males. With regard to Pclass, the amount of 40% of upper-class passengers survived, while 68% of lower-class passengers did not survive. The middle class has a similar survival rate between non-survivors and survivors (18% and 25%, respectively), demonstrating that the survival rate is higher among upper-class passengers than in other classes.

From the analysis, it can be concluded that fare, SibSp, Parch, sex, and Pclass can be used as predictors of survival. Higher fares are linked to a higher chance of surviving since passengers who paid more for their tickets tend to survive. The number of siblings/spouses (SibSp) or parents/children (Parch) aboard is inversely correlated with the survival rate. The survival rate increases with decreasing the total number of passengers on board. Moreover, females were more likely to survive compared to males. Finally, passengers in upper classes (Pclass 1) had a higher chance of survival compared to those in other classes (Pclass 2 and 3).

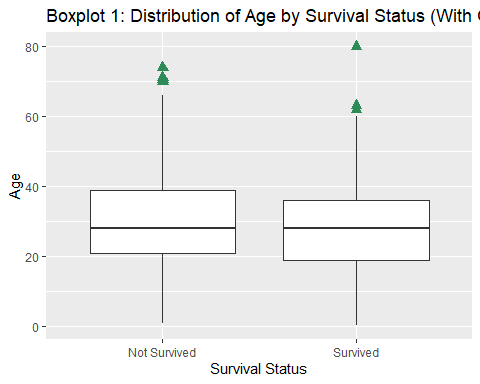
table1

Table 1: Table 1: Summary statistics of Titanic Passenger Data by Survival Status*1*

|  | **Survival Status** | |
| --- | --- | --- |
| **Characteristic** | **0** N = 549*2* | **1** N = 342*2* |
| Age | 28 (18) | 28 (17) |
| Unknown | 125 | 52 |
| Fare | 11 (18) | 26 (45) |
| 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%) |
| Sex |  |  |
| female | 81 (15%) | 233 (68%) |
| male | 468 (85%) | 109 (32%) |
| Pclass |  |  |
| 1 | 80 (15%) | 136 (40%) |
| 2 | 97 (18%) | 87 (25%) |
| 3 | 372 (68%) | 119 (35%) |
| *1*0 = Not survived, 1 = Survived | | |
| *2*Median (IQR); n (%) | | |

boxplot\_age

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



boxplot\_fare

