assignment2

Polatat

2024-09-08

# Install gtsummary for summary

install.package(“gtsummary”)

# Start analyse the survival factors in first table.

## Define titanic\_data as train.csv for summary identify the abnomral values

install.package(“tidyverse”) # I install this package because it can help to convert the data into much better presentation. library(tidyverse) titanic\_data <- read.csv(“data/titanic/train.csv”)

View(titanic\_data)

# Checking for outliers using exploratory graph for each variables group which is creating a box plot.

Before plotting the graph, I know that there are five important variables, including: Pclass, Sex, Age, SibSp, Parch and Fare. Besides, it can be divided into two types of data.

1. Continuous Variables(Quantitative) : Age and Fare

2. Categorical Variables(Qualitative): SibSp,Parch and Pclass

Note:Sibsp = siblings/spouses aboard the Titanic Parch = parents/ children aboard the Titanic Fare = passenger fare(the revenue earned from carrying a passenger) Pclass = ticket class.

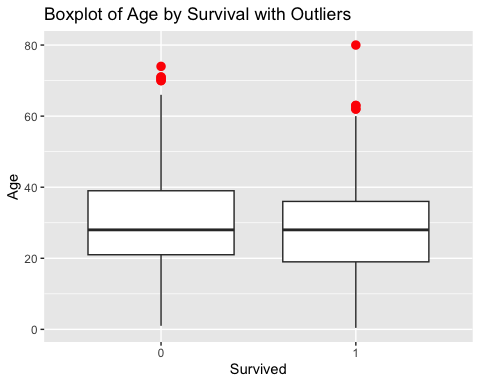
Box plots in fare, age, Parch and SibSp vs survival and non-survival variables.

I do not plot Pclass because the class consists of 1,2,3 ticket class, so there is no outlier.

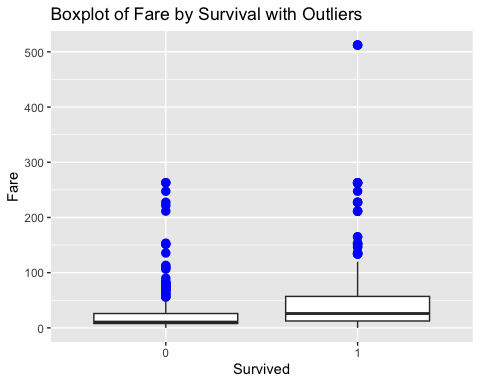
# Set the CRAN mirror  
options(repos = c(CRAN = "https://cloud.r-project.org")) # I need to set the CRAN mirror because R need to detect the installed package that have already had. Without of it, R will throw an error to let me install it everytime, and I cannot knit the file into word document.  
  
# Install the package (only do this if the package isn't already installed)  
if (!requireNamespace("ggplot2", quietly = TRUE)) {  
 install.packages("ggplot2")  
}  
  
# Load ggplot2  
library(ggplot2)

ggplot(titanic\_data, aes(x = as.factor(Survived), y = Age)) +  
 geom\_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 3) +  
 labs(x = "Survived", y = "Age", title = "Boxplot of Age by Survival with Outliers")

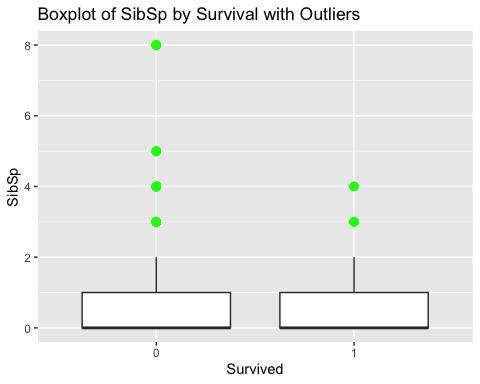
## Warning: Removed 177 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



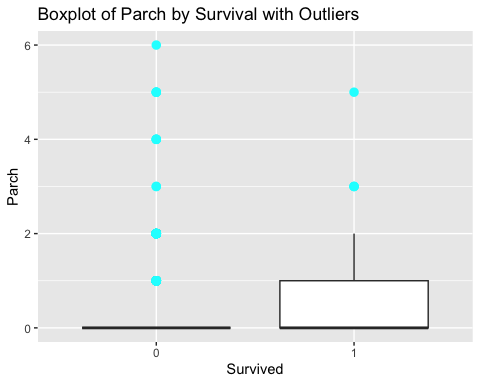
ggplot(titanic\_data, aes(x = as.factor(Survived), y = Fare)) +  
 geom\_boxplot(outlier.colour = "blue", outlier.shape = 16, outlier.size = 3) +  
 labs(x = "Survived", y = "Fare", title = "Boxplot of Fare by Survival with Outliers")



ggplot(titanic\_data, aes(x = as.factor(Survived), y = SibSp)) +  
 geom\_boxplot(outlier.colour = "green", outlier.shape = 16, outlier.size = 3) +  
 labs(x = "Survived", y = "SibSp", title = "Boxplot of SibSp by Survival with Outliers")



ggplot(titanic\_data, aes(x = as.factor(Survived), y = Parch)) +  
 geom\_boxplot(outlier.colour = "cyan", outlier.shape = 16, outlier.size = 3) +  
 labs(x = "Survived", y = "Parch", title = "Boxplot of Parch by Survival with Outliers")



# Creating a gtsummary to analyse the survival factors from train\_data

library(gtsummary)  
  
table1 <- titanic\_data %>%  
 select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare) %>%  
 tbl\_summary(by = Survived)

I do not use ticket ,cabin and embarked information because most of them are not digit data , and also they are not significant factors for survival factor analysis.

## The Interpretation Of Table1

According to the table1, two main headers are composed of survival and non-survival passengers. The non-survival passengers represent 0 total of 549, while the survival passengers are 1 a total of 342 They correlate with the others variables to predict the survival chance, including pclass, sex, age, sibsp, parch, and fare.

Let’s explain each variable, which have two groups that comprise categorical(pcalss,sibsp,parch,and sex) and continuous(age and fare) variaables.

## Categorical variables

These variables are for counting frequency, so percentages is the suitable unit for measure them. Due to the fact that they are arranged by category or groups.

1. **Pclass**( ticket-class) It is clear that pclass is one of the important survival factors. Ticket class 3 has the highest casualty at 372(68%) , while people who have the highest survival rate are first class at 136 (40%) followed by 119(35%), and 87(25%) for class three and class two. Thus, the more people pay for ticket class, the more they may survive the incident.
2. **Sibsp (sibling or spouses) and Parch (parents and children)** Clearly, if passengers travel alone, they will have the highest chance of dying at 398 (72%) and 445(81%) for Sibsp and Parch respectively. They, however, also have the highest survival rate at 210(61%) and 233(68%) respectively, so most passengers coming on this trip alone while some of them come with their families may help each other to survive. In addition, people who come with one member of families will have a chance to survive more than the other number of individuals in the families.
3. **Sex** Sex variable show a significant difference between deceased and alive passengers. Evidently, women have more survival chance than men at 68% in the survival group, while the non-survival group show 85% for males compared to females at 15%. Therefore, it can be interpreted that female has more survival chance than male.

## Continuous variables

These two variables do not measure in number(percentage) like categorical variables because they are continuous values, so we measure median and standard deviation instead.

1. **Age** In non-survival passengers have a mean(SD) of 31(14), whereas survival passengers are 28(15). These values show that they are not vital factors for survival because the mean(SD) from both of them are very similar. It can be seen that the age mean of survival people are younger than casualties although they are very identical. In addition, we do not have deceased passengers aged data higher than survivors aged data by 73.
2. **Fare** This variable is associated with pclass because the more the passengers paid for fare, the higher class they got. In the non-survival group, it has a mean(SD) at 22(31), while survival is 48(67); so it is strongly different in mean and sd. The higher the SD, the more variety of data range. Hence, people who were in the surviving group paid for the first class or second class more than the others. In the non-survival group, on the other hand, most of the people spent in the same range which most of them were at the third class.

**More…** The small SD indicates that the data range from variable is close to the mean. The large SD indicates that the data range differs greatly from the mean.

## Conclusion

Overall, most variables are important to identify the chance of survival. Even though there is missing a number of age data for 177 as 19% of the total. It is still possible to predict that people who paid the highest fare for the first-class ticket will have a much higher chance of surviving. From sex, and age information, women and young children will have a chance to survive more than the others. Parch and Sibsp also show that most people travelled alone on this trip, but passengers who came with one member of families will likely have a higher chance to survive than people who were on board with families of more than or equal to two.

table1

| **Characteristic** | **0** N = 549*1* | **1** N = 342*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*n (%); Median (Q1, Q3) | | |