Assignment2

Jittin Pomprakay

2024-09-13

# Import titanic\_data

library(readr)  
train <- read\_csv("../Data/train.csv",  
col\_types = cols(Survived = col\_integer(),  
Pclass = col\_integer(), SibSp = col\_integer(),  
Parch = col\_integer()))

# Checking for any outliers or abnormal values in any variables group using box plot

Plotting the graph, I use only 6 important variables, including; Pclass, Sex, Age, SipSp, Parch and Fare. To these variables, they can be categorized into two group of data.

1. Continuous Variables(Quantitative): Only Age and Fare

2. Categorical Variables(Qualitative): SibSp(siblings/spouses aboard the Titanic), Parch(parents/children aboard the Titanic) and Pcalss(Ticket class)

Note: I don’t plot Pclass on the graph because the Pclass only consists of integers 1,2 and 3, so there are no otliers.

## Installing the ggplot2 package (contains a Box Plot maker)

options(repos = c(CRAN = "https://cran.r-project.org"))  
  
# Install the ggplot2 package  
install.packages("ggplot2")

## Installing package into 'C:/Users/jitti/AppData/Local/R/win-library/4.4'  
## (as 'lib' is unspecified)

## package 'ggplot2' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\jitti\AppData\Local\Temp\RtmpgXhGM9\downloaded\_packages

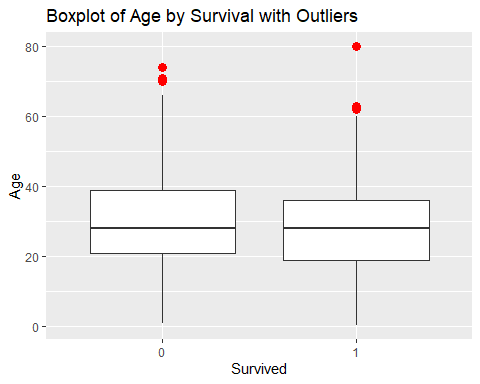
# Load ggplot2  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.2

## Age

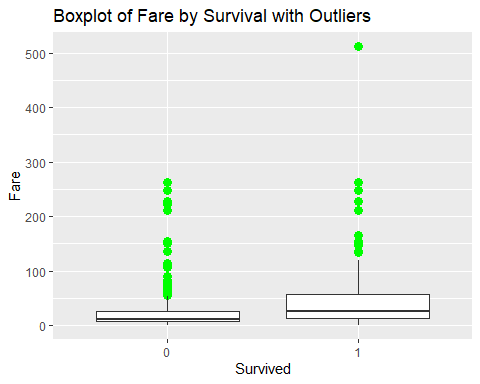
ggplot (train, aes(x = as.factor(Survived), y = Age)) +  
 geom\_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 3) +  
 labs(x = "Survived", y = "Age", title = "Boxplot of Age by Survival with Outliers")

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



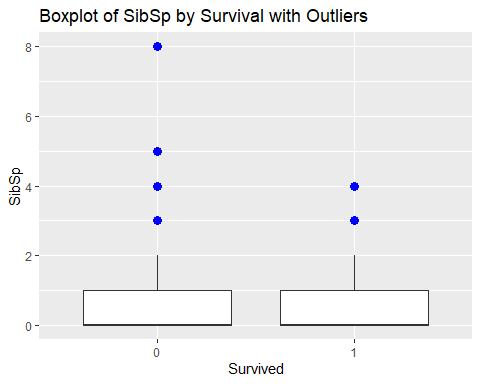
## Fare

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



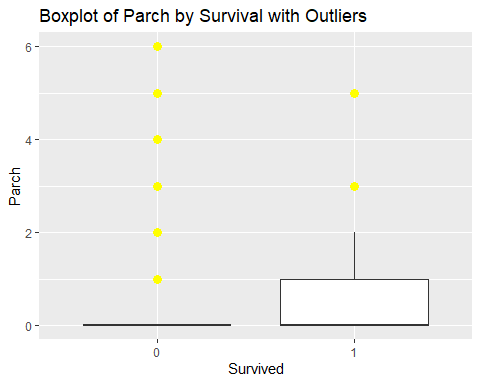
## SibSp

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



## Parch

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



# Creating the table 1 summarizing available data on the titanic\_data

Using gtsummary package to create table 1

## Installing gtsummary package

install.packages("gtsummary")

## Installing package into 'C:/Users/jitti/AppData/Local/R/win-library/4.4'  
## (as 'lib' is unspecified)

## package 'gtsummary' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\jitti\AppData\Local\Temp\RtmpgXhGM9\downloaded\_packages

## Creating the table 1

library(gtsummary)

## Warning: package 'gtsummary' was built under R version 4.4.2

table1 <- train %>%  
select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare) %>%  
tbl\_summary(  
by=Survived,  
statistic = list(all\_continuous() ~ "{mean} ({sd})")  
)  
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 | 31 (14) | 28 (15) |
| 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 | 22 (31) | 48 (67) |
| *1*n (%); Mean (SD) | | |

Note: Ticket, Cabin, and Embarked information aren’t used because they consist of digit data and are not significant variable for survival analysis.

## The interpretation of table 1

According to Table 1, 0 and 1 were used to represent surviving and non-surviving passengers, respectively.The total number of surviving passengers is 549, while the non-surviving is 342. The other variables, including **Pclass**, **Sex**, **Age**, **SibSp**, **Parch**, and **Fare**, used the number of surviving and non-surviving passengers to analyze which variables were more likely to influence survival.

The significant variables were divided into two groups of data, like a box plot, consisting of continuous and categorical variables.

### Categorical Variables

These variables are take a wide range of values, then using them to calculate the mean and SD for statistical analysis.

**Age**

The mean age of non-surviving passengers is 31 years, which is slightly higher than the mean age of surviving passengers, which is 28 years. This may indicate that older passengers had a lower probability of surviving compared to younger passengers. The SD values for both groups also show that age alone may not be the main variable in predicting survival, as they are very similar. Therefore, the other variables also should be considered.

**Fare**

This variable can be linked to the Pclass variable. If they want to get a higher class, they must pay a higher fare. This means that a higher mean fare indicates that most passengers were in a higher class. The mean and SD fare of surviving passengers are significantly different from those of non-surviving passengers. This suggests that passengers who paid more for a higher class had a higher probability of surviving.

### Continuous variables

These variables are for counting frequency, then converting them to percentages for more comprehensible visualization.

**Pclass (Ticket class)**

Pclass variables indicate to the passenger’s ticket class. The data show that The first class has the highest survival rate at 136(40%), followed by class 2 at 87(25%) and class 3 at 119(35%). Especially, ticket class 3 has the highest number of casualties at 372(68%). This means that passengers who paid for higher classes had a better chance of survival.

**Sex**

This variable shows the highest survival rate in females at 233(68%) compared to 109(32%) in males, while males have the highest mortality rate at 468(85%) compared to 81(15%) in females. This means that females have a higher chance of survival compared to males.

**SipSp (Siblings/Spouses aboard the Titanic) and Parch (Parents/Children aboard the Titanic)**

These two variables have similar patterns of information and detail, so I simultaneously interpreted both of them. The passenger who comes alone has the highest mortality rate at 398 (72%) and 445 (81%) for SipSp and Parch, respectively. They also have the highest survival rate at 210 (61%) and 233 (68%). Moreover, both show a similar trend in survival and mortality rates. If we bring one or two parents or siblings, this can increase the probability of surviving, while having more than two parents or siblings can increase the probability of mortality as well.

# Conclusion

All in all, six variables have the potential to be important factors in predicting the survival chances of passengers. Even though the age variable contains some missing information, it is still sufficient for analysis. Based on the available data, younger passengers tend to have a higher chance of survival than older passengers. Surprisingly, females have a significantly higher chance of surviving than males. It cannot be denied that wealth plays a role in increasing the probability of survival, as purchasing power enables one to afford a higher ticket class. According to statistics, most passengers who traveled alone had a lower chance of survival, while traveling with one or two family members increased the chances of surviving.