Titanic Exploratoin

Chakrit Jiarasatit

2024-09-12

options(repos = c(CRAN = "https://cloud.r-project.org"))

### Install tidyverse and gtsummary packages

install.packages("tidyverse")

##   
## The downloaded binary packages are in  
## /var/folders/4q/n5g5xz054rngj5q6xy1lq8680000gn/T//RtmpZZg3Kv/downloaded\_packages

install.packages("gtsummary")

##   
## The downloaded binary packages are in  
## /var/folders/4q/n5g5xz054rngj5q6xy1lq8680000gn/T//RtmpZZg3Kv/downloaded\_packages

### Dowload essential packages

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(gtsummary)

### Import Titanic data

titanic\_data <- read.csv("../data/titanic/train.csv")

### Summary Titanic data

summary(titanic\_data)

## PassengerId Survived Pclass Name   
## Min. : 1.0 Min. :0.0000 Min. :1.000 Length:891   
## 1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000 Class :character   
## Median :446.0 Median :0.0000 Median :3.000 Mode :character   
## Mean :446.0 Mean :0.3838 Mean :2.309   
## 3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :891.0 Max. :1.0000 Max. :3.000   
##   
## Sex Age SibSp Parch   
## Length:891 Min. : 0.42 Min. :0.000 Min. :0.0000   
## Class :character 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000   
## Mode :character Median :28.00 Median :0.000 Median :0.0000   
## Mean :29.70 Mean :0.523 Mean :0.3816   
## 3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :80.00 Max. :8.000 Max. :6.0000   
## NA's :177   
## Ticket Fare Cabin Embarked   
## Length:891 Min. : 0.00 Length:891 Length:891   
## Class :character 1st Qu.: 7.91 Class :character Class :character   
## Mode :character Median : 14.45 Mode :character Mode :character   
## Mean : 32.20   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

### Boxplot of Titanic passengers data

### 1. The Age distribution of passengers aboard the Titanic

#### Did not survive (“0”)

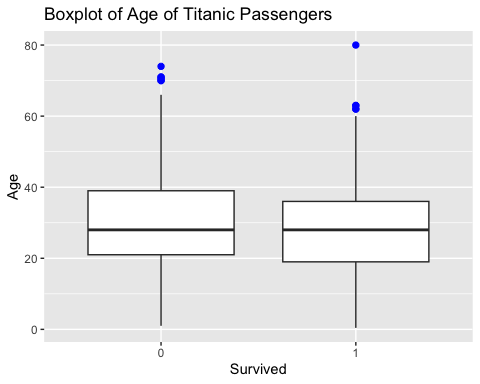
* Median age : Approximately 30 years.
* Interquartile range (IQR) : Spans from about 22 to 37 years.
* Outliers : Several outliers are represented by blue dots above the upper whisker, indicating passengers significantly older than the majority who did not survive.

#### Survived (“1”)

* Median age : Slightly younger, around 28 years.
* Interquartile range (IQR) : Extends from about 19 to 36 years.
* Outliers : There are outliers above the upper whisker, indicating passengers significantly older than the majority who survived.

library(ggplot2)  
ggplot(titanic\_data, aes(x = as.factor(Survived), y = Age)) +  
 geom\_boxplot(outlier.colour = "blue", outlier.shape = 20, outlier.size = 3) +  
 labs(title = "Boxplot of Age of Titanic Passengers", x = "Survived", y = "Age")

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



### 2. The distribution of the Parch (number of parents/children aboard) among passengers on the Titanic

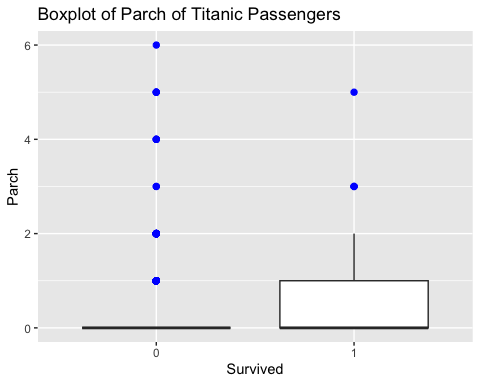
#### Did not survive (“0”)

* Median parch: The median number of parents/children aboard for non-survivors is 0.
* Interquartile Range (IQR) : The IQR spans from 0 to 0, indicating that the middle 50% of non-survivors had no parents or children aboard.
* Outliers : There are several outliers above the upper whisker, indicating passengers with higher numbers of parents/children aboard who did not survive. These outliers range up to 6.

#### Survived (“1”)

* Median parch : The median number of parents/children aboard for survivors is also 0.
* Interquartile range (IQR) : The IQR extends from 0 to 1, showing that the middle 50% of survivors had between 0 and 1 parents or children aboard.
* Outliers : There are outliers above the upper whisker, indicating passengers with higher numbers of parents/children aboard who survived. These outliers range from 3 up to 5.

library(ggplot2)  
ggplot(titanic\_data, aes(x = as.factor(Survived), y = Parch)) +  
 geom\_boxplot(outlier.colour = "blue", outlier.shape = 20, outlier.size = 3) +  
 labs(title = "Boxplot of Parch of Titanic Passengers", x = "Survived", y = "Parch")



### 3. The distribution of the Fare (the fare paid by passengers) among passengers on the Titanic

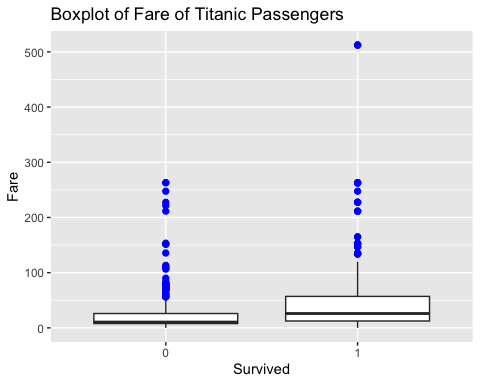
#### Did not survive (“0”)

* Median fare : The median fare for non-survivors is relatively low.
* Interquartile Range (IQR) : The IQR spans from a low fare to a moderate fare, indicating that the middle 50% of non-survivors paid relatively low fares.
* Outliers : There are several outliers above the upper whisker, indicating that some non-survivors paid significantly higher fares.

#### Survived (“1”)

* Median fare : The median fare for survivors is higher compared to non-survivors.
* Interquartile range (TQR) : The IQR extends from a low fare to a higher fare, showing that the middle 50% of survivors paid a wider range of fares.
* Outliers : There are numerous outliers above the upper whisker, indicating that many survivors paid significantly higher fares.

library(ggplot2)  
ggplot(titanic\_data, aes(x = as.factor(Survived), y = Fare)) +  
 geom\_boxplot(outlier.colour = "blue", outlier.shape = 20, outlier.size = 3) +  
 labs(title = "Boxplot of Fare of Titanic Passengers", x = "Survived", y = "Fare")



### 4. The distribution of the SipSp (number of parents/children aboard) among passengers on the Titanic

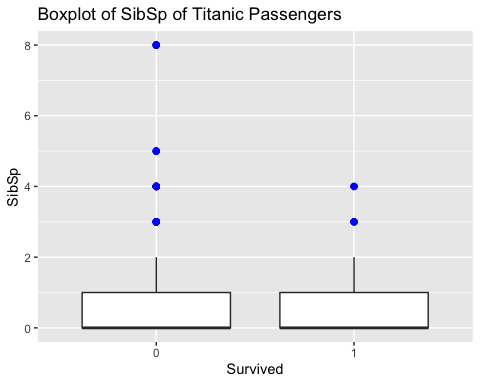
#### Did not survive (“0”)

* Median SipSp : The median number of siblings/spouses aboard for non-survivors is 0.
* Interquartile Range (IQR) : The IQR spans from 0 to 1, indicating that the middle 50% of non-survivors had between 0 and 1 siblings or spouses aboard.
* Outliers : There are several outliers above the upper whisker, indicating passengers with higher numbers of siblings/spouses aboard who did not survive. These outliers range up to 8.

#### Survived (“1”)

* Median SibSp : The median number of siblings/spouses aboard for survivors is also 0.
* Interquartile range (IQR) : The IQR extends from 0 to 1, showing that the middle 50% of survivors had between 0 and 1 siblings or spouses aboard.
* Outliers : There are outliers above the upper whisker, indicating some passengers with higher numbers of siblings/spouses aboard who survived. These outliers range from 3 up to 4.

library(ggplot2)  
ggplot(titanic\_data, aes(x = as.factor(Survived), y = SibSp)) +  
 geom\_boxplot(outlier.colour = "blue", outlier.shape = 20, outlier.size = 3) +  
 labs(title = "Boxplot of SibSp of Titanic Passengers", x = "Survived", y = "SibSp")



### Table 1 Shows the survival rates of Titanic passengers based on Pclass, Sex, SibSp, Parch

### 1. Analysis of passenger class (Pclass) and survival rates on the Titanic

* Highest survival rates in first Class : Passengers in first class had the highest survival rates, likely due to better access to lifeboats and priority during evacuation.
* Third class & Second Class : Interestingly, third-class passengers had a higher survival rate (35%) compared to second-class passengers (25%). This could be due to various factors such as the location of cabins, the demographics of the passengers, or the effectiveness of evacuation procedures for different classes.
* Socioeconomic Factors : The overall trend still suggests that higher-class passengers had better survival chances, but the higher survival rate in third class compared to second class is an important observation that warrants further investigation.

### 2. Analysis of sex and survival rates on the Titanic

* Higher survival rates for females : Female passengers had a significantly higher survival rate (68%) compared to male passengers (32%). This trend suggests that women were given priority during the evacuation process, which aligns with the “women and children first” protocol often followed during maritime disasters.
* Lower survival rates for males : Male passengers had a much lower survival rate, which could be due to the prioritization of women and children for lifeboat spaces.

### 3. Analysis of SibSp and survival rates on the Titanic

* Higher survival rates with 1 Sibling/Spouse : Passengers with one sibling or spouse aboard had the highest survival rate (53.6%). This suggests that having a close family member might have provided better chances of survival. <>/p
* Decreasing survival rates with more Siblings/Spouses : As the number of siblings or spouses increases, the survival rates generally decrease. This trend could be due to the challenges of coordinating and ensuring the safety of larger family groups during the evacuation.

### 4. Analysis of parch and survival rates on the Titanic

* Higher survival rates with 1-3 Parents/Children : Passengers with one to three parents or children aboard had relatively higher survival rates, with the highest being 60% for those with three parents or children. This suggests that having a small family group might have provided better chances of survival.
* Decreasing survival rates with more Parents/Children : As the number of parents or children increases beyond three, the survival rates generally decrease. This trend could be due to the challenges of coordinating and ensuring the safety of larger family groups during the evacuation.

table1 <- titanic\_data %>%  
 select(Survived, Pclass, Sex, SibSp, Parch) %>%  
 tbl\_summary(by = Survived,   
 statistic = list(all\_categorical() ~ "{n} ({p}%)",  
 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%) |
| 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%) |
| *1*n (%) | | |

### Conclusion

The Titanic data gives us some clear insights into who had better chances of survival. First-class passengers had the best survival rates, likely because they had better access to lifeboats and were given priority during evacuation. Interestingly, third-class passengers had a higher survival rate (35%) compared to second-class passengers (25%), which might be due to factors like cabin location and evacuation procedures. Women had a much higher survival rate (68%) compared to men (32%). This follows the “women and children first” rule that was often used during ship evacuations. Family size also mattered. Passengers with one sibling or spouse had the highest survival rate (53.6%), while those with more siblings or spouses had lower chances, likely because it was harder to stay together during the evacuation. Similarly, passengers with one to three parents or children had higher survival rates, with the highest being 60% for those with three parents or children. However, survival rates dropped as the number of parents or children increased beyond three. Overall, the data shows that class, gender, and family size all played a role in who survived the Titanic disaster. These insights help us understand the factors that influenced survival during this tragic event.