Analysis Project in R Studio for Titanic Data set

Marwah Alisaee

2023-07-19

## Introduction

### downlode the CSV from kaggle imports a 891 observation data set which contains 12 variables.The variables are PassengerId, Survived, Pclass, Name, Sex, Age, Sibsp (number of siblings and spouses), Parch (number of parents and children), Ticket, Fare, Cabin, and Embarked (which had port of embarkation).

#### lode the data set:

library(readr)  
tested <- read\_csv("tested.csv")

## Rows: 418 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(tested)

#### first 5 rows:

head(tested)

## # A tibble: 6 × 12  
## PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin  
## <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>  
## 1 892 0 3 Kelly,… male 34.5 0 0 330911 7.83 <NA>   
## 2 893 1 3 Wilkes… fema… 47 1 0 363272 7 <NA>   
## 3 894 0 2 Myles,… male 62 0 0 240276 9.69 <NA>   
## 4 895 0 3 Wirz, … male 27 0 0 315154 8.66 <NA>   
## 5 896 1 3 Hirvon… fema… 22 1 1 31012… 12.3 <NA>   
## 6 897 0 3 Svenss… male 14 0 0 7538 9.22 <NA>   
## # ℹ 1 more variable: Embarked <chr>

#### 

#### Display the structure of each variable in PassengerId

str(tested$PassengerId)

## num [1:418] 892 893 894 895 896 897 898 899 900 901 ...

str(tested$Age)

## num [1:418] 34.5 47 62 27 22 14 30 26 18 21 ...

#### 

#### describe() functions

#### Determine variable median

median\_age <- median(tested$Age, na.rm=T)

#### Store original data in a variable in case we make a mistake or want it for

#### comparison

original\_age\_data <- tested$Age

#### 

#### replace NAs with median age

tested$Age[is.na(tested$Age)] <- median\_age  
modified\_age\_data <- tested$Age

#### Comparison between original data and adjusted data

summary(original\_age\_data)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.17 21.00 27.00 30.27 39.00 76.00 86

#### 

#### show the summary of data set

summary(tested)

## PassengerId Survived Pclass Name   
## Min. : 892.0 Min. :0.0000 Min. :1.000 Length:418   
## 1st Qu.: 996.2 1st Qu.:0.0000 1st Qu.:1.000 Class :character   
## Median :1100.5 Median :0.0000 Median :3.000 Mode :character   
## Mean :1100.5 Mean :0.3636 Mean :2.266   
## 3rd Qu.:1204.8 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :1309.0 Max. :1.0000 Max. :3.000   
##   
## Sex Age SibSp Parch   
## Length:418 Min. : 0.17 Min. :0.0000 Min. :0.0000   
## Class :character 1st Qu.:23.00 1st Qu.:0.0000 1st Qu.:0.0000   
## Mode :character Median :27.00 Median :0.0000 Median :0.0000   
## Mean :29.60 Mean :0.4474 Mean :0.3923   
## 3rd Qu.:35.75 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :76.00 Max. :8.0000 Max. :9.0000   
##   
## Ticket Fare Cabin Embarked   
## Length:418 Min. : 0.000 Length:418 Length:418   
## Class :character 1st Qu.: 7.896 Class :character Class :character   
## Mode :character Median : 14.454 Mode :character Mode :character   
## Mean : 35.627   
## 3rd Qu.: 31.500   
## Max. :512.329   
## NA's :1

#### 

#### the number of null value in each columns

colSums(is.na(tested))

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 327 0

#### 

#### change data set to data frame

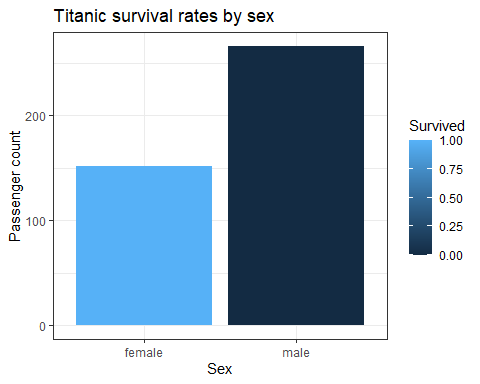
tested = data.frame(tested)

## Titanic servival retes by sex

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.1

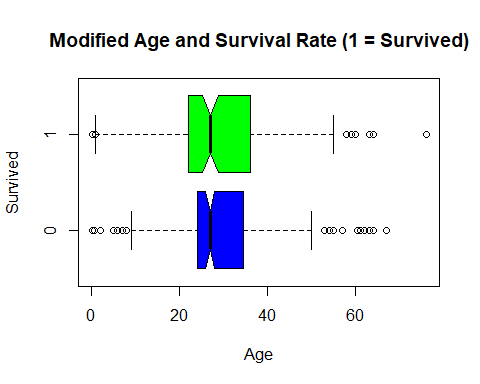
ggplot(tested, aes(x = Sex, fill = Survived)) +  
 theme\_bw()+  
 geom\_bar()+  
 labs(y = "Passenger count",  
 title = "Titanic survival rates by sex")



#### in this par plot the veribels chosen to analyze is the number of pasengers’s survival rate best on the gender; the survival rate of femals is sgnificantly highre than survival rate of mals at 150 female surviving obsite 250 meal dying .

## Modified Age and Survival Rate

# First with the Modified Age Data  
boxplot(tested$Age~tested$Survived, notch=T, horizontal=T, ylab = "Survived",  
 xlab = "Age", main = "Modified Age and Survival Rate (1 = Survived)", col=c("blue", "green"))

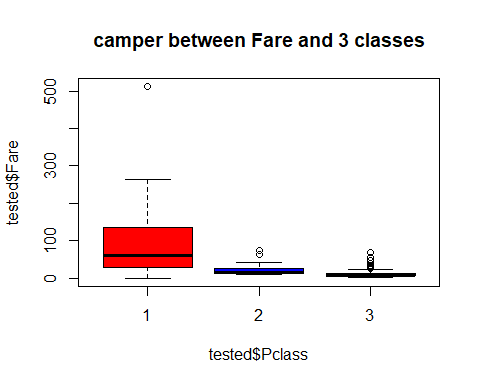


*boxplot that compares the ages of two groups: those who survived (represented by the green boxplot) and those who did not survive (represented by the blue boxplot). The y-axis represents the “Survived” category, and the x-axis represents the “Age” category.*

#### By comparing the boxplots, you can observe any differences or similarities in the distribution of ages between the two groups. For example, you can see if the median age and the spread of ages differ between the survivors and non-survivors.

## camper between Fare and 3 classes

boxplot(tested$Fare~tested$Pclass, col=c("red","blue","green"), main='camper between Fare and 3 classes')



*in this boxplot distrib of the “Fare” variable based on the “Pclass” variable. The “Pclass” variable represents the passenger class, which can take the values 1, 2, or 3, while the “Fare” variable represents the fare paid by each passenger.*

#### The boxplot is divided into three sections, corresponding to each passenger class (1, 2, and 3). The y-axis represents the fare values, and the x-axis represents the passenger class. The red, blue, and green colors represent the different passenger classes.

#### The boxplot provides a visual summary of the fare distribution within each passenger class.

#### By examining the boxplot, you can observe the differences in fare distribution between the passenger classes. It can help you understand the spread, central tendency, and potential outliers in the fare data for each class.