Assignment2

Niramai Arkhom

2024-10-15

# Load required packages, install if necessary  
if (!require("gtsummary")) install.packages("gtsummary")

## Loading required package: gtsummary

if (!require("ggplot2")) install.packages("ggplot2")

## Loading required package: ggplot2

if (!require("dplyr")) install.packages("dplyr")

## Loading required package: 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

if (!require("readr")) install.packages("readr")

## Loading required package: readr

if (!require("gt")) install.packages("gt")

## Loading required package: gt

# Load libraries  
library(gtsummary)  
library(ggplot2)  
library(dplyr)  
library(readr)  
library(gt)  
  
# Read data: Make path relative so it can work on other machines  
titanic\_data <- 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(titanic\_data)  
  
# Create a summary table grouped by 'Survived'  
summary\_table <- titanic\_data %>%  
 select(Survived, Age, Fare, SibSp, Parch, Sex, Pclass) %>%  
 tbl\_summary(  
 by = Survived,  
 statistic = list(  
 all\_continuous() ~ "{median} ({IQR})",   
 all\_categorical() ~ "{n} ({p}%)"  
 )  
 ) %>%  
 modify\_header(label = "\*\*Variable\*\*") %>%  
 modify\_spanning\_header(all\_stat\_cols() ~ "\*\*Survival Status\*\*") %>%  
 as\_gt() %>%  
 tab\_header(  
 title = "Table 1: Summary of Titanic Data by Survival Status"  
 ) %>%  
 tab\_footnote(  
 footnote = "0 = Not Survived, 1 = Survived",  
 locations = cells\_title(groups = "title")  
 )  
  
# Display the summary table in RStudio viewer  
summary\_table

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

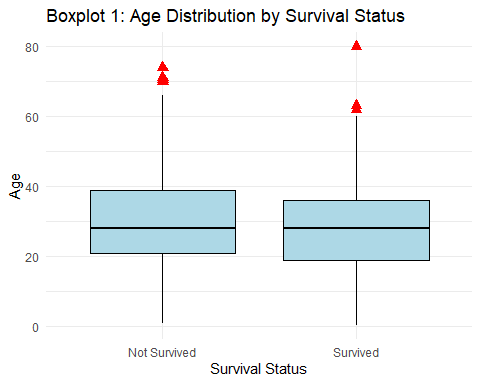
|  | **Survival Status** | |
| --- | --- | --- |
| **Variable** | **0** N = 549*2* | **1** N = 342*2* |
| Age | 28 (18) | 28 (17) |
| Unknown | 125 | 52 |
| Fare | 11 (18) | 26 (45) |
| SibSp |  |  |
| 0 | 398 (72%) | 210 (61%) |
| 1 | 97 (18%) | 112 (33%) |
| 2 | 15 (2.7%) | 13 (3.8%) |
| 3 | 12 (2.2%) | 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 (%) | | |

# Function to generate boxplots  
generate\_boxplot <- function(df, x\_var, y\_var, x\_labels, title, outlier\_col, fill\_col, outlier\_shape, outlier\_size) {  
 ggplot(df, aes\_string(x = x\_var, y = y\_var)) +  
 geom\_boxplot(outlier.colour = outlier\_col, fill = fill\_col, outlier.shape = outlier\_shape, outlier.size = outlier\_size, color = "black") +  
 scale\_x\_discrete(labels = x\_labels) +  
 labs(x = "Survival Status", y = y\_var, title = title) +  
 theme\_minimal()   
}  
  
# Create boxplot for Age vs. Survival Status  
boxplot\_age <- generate\_boxplot(  
 titanic\_data,  
 "factor(Survived)",  
 "Age",  
 c("0" = "Not Survived", "1" = "Survived"),  
 "Boxplot 1: Age Distribution by Survival Status",  
 outlier\_col = "red",   
 fill\_col = "lightblue",   
 outlier\_shape = 17,  
 outlier\_size = 3  
)

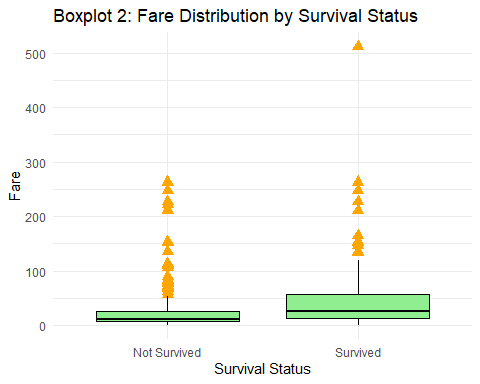
## Warning: `aes\_string()` was deprecated in ggplot2 3.0.0.  
## ℹ Please use tidy evaluation idioms with `aes()`.  
## ℹ See also `vignette("ggplot2-in-packages")` for more information.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

# Create boxplot for Fare vs. Survival Status  
boxplot\_fare <- generate\_boxplot(  
 titanic\_data,  
 "factor(Survived)",  
 "Fare",  
 c("0" = "Not Survived", "1" = "Survived"),  
 "Boxplot 2: Fare Distribution by Survival Status",  
 outlier\_col = "orange",   
 fill\_col = "lightgreen",   
 outlier\_shape = 17,  
 outlier\_size = 3  
)  
  
# Display boxplots in RStudio viewer  
boxplot\_age

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



boxplot\_fare



#end