Lab 4A

Nehal Ur Rahman

2023-02-16

#NAME: NEHAL UR RAHMNAN #STUDENT ID: 991691259

##Introduction #In lab 4A we are going to analyze a Titanic dataset which consists of 4 columns.There are 891 rows in the given dataset. This dataset gives us an idea of the demograph of the people in the titanic ship. —

titanic <- read.csv("Titanic1.csv")  
head(titanic)

## Survived Pclass Sex Age  
## 1 Dead third male 22  
## 2 Alive first female 38  
## 3 Alive third female 26  
## 4 Alive first female 35  
## 5 Dead third male 35  
## 6 Dead third male NA

#We are loading the below library to perform various functions.

library(sets)  
library(ggplot2)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ tibble 3.1.8 ✔ dplyr 1.1.0  
## ✔ tidyr 1.3.0 ✔ stringr 1.5.0  
## ✔ readr 2.1.3 ✔ forcats 1.0.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ forcats::%>%() masks stringr::%>%(), dplyr::%>%(), purrr::%>%(), tidyr::%>%(), tibble::%>%(), sets::%>%()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ISLR)  
library(moments)  
library(dplyr)

#We are using the cat function to print the column names of the titanic dataset with the help of colnames function.

cat("Column Names : ", colnames(titanic))

## Column Names : Survived Pclass Sex Age

#As we can see in the output, there are 4 columna in the dataset.The columns are Survived Pclass Sex Age.

cat("\nSummary of data: \n")

##   
## Summary of data:

summary(titanic)

## Survived Pclass Sex Age   
## Length:891 Length:891 Length:891 Min. : 0.42   
## Class :character Class :character Class :character 1st Qu.:20.12   
## Mode :character Mode :character Mode :character Median :28.00   
## Mean :29.70   
## 3rd Qu.:38.00   
## Max. :80.00   
## NA's :177

#Now we use the summary function to print the summary of all the variables but we get the statistical information of the Age column only because it is the only numerical variable. #We also see that there 177 Na’s in the Age variable which will be cleaned later.

cat("\nStructure of data: \n")

##   
## Structure of data:

str(titanic)

## 'data.frame': 891 obs. of 4 variables:  
## $ Survived: chr "Dead" "Alive" "Alive" "Alive" ...  
## $ Pclass : chr "third" "first" "third" "first" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

#We have also used the function colSums to find the sum of na values in the dataset.

colSums(is.na(titanic))

## Survived Pclass Sex Age   
## 0 0 0 177

#We have fetched the structure of the data to get an idea about the datatypes of the variables. So we can see that there are 4 variables out of which 3 are character variables and 1 is a numerical variable.

#We create a variable mage to store the value of mean value of Age variable from the titanic dataset   
mage = mean(titanic$Age, na.rm = TRUE)  
mage

## [1] 29.69912

#As we can see above the mean of the age variable is 29.69912. We get the mean of the Age variable only after removing the NA values which is why we have the code “na.rm=TRUE”.

#Now we create a new dataset called new\_titanic from titanic dataset. We do this so we can keep the originality of the titanic dataset. In this new daatset we put the mean value of age in the Na rows of the Age column. #We have done this to provide a more refined dataset.

new\_titanic <- titanic  
mage -> new\_titanic$Age[is.na(new\_titanic$Age)]   
head(new\_titanic)

## Survived Pclass Sex Age  
## 1 Dead third male 22.00000  
## 2 Alive first female 38.00000  
## 3 Alive third female 26.00000  
## 4 Alive first female 35.00000  
## 5 Dead third male 35.00000  
## 6 Dead third male 29.69912

#Here, we can see that now the Na values in the Age column are replaced by the mean(29.69912).

colSums(new\_titanic[is.na(new\_titanic$Age)])

## numeric(0)

cat("\n")

summary(new\_titanic)

## Survived Pclass Sex Age   
## Length:891 Length:891 Length:891 Min. : 0.42   
## Class :character Class :character Class :character 1st Qu.:22.00   
## Mode :character Mode :character Mode :character Median :29.70   
## Mean :29.70   
## 3rd Qu.:35.00   
## Max. :80.00

#We again check the summary of the the new dataset called new\_titanic to make sure that the age variable does not contaion Na variables.

cat("Index:", which(is.na(titanic$Age)))

## Index: 6 18 20 27 29 30 32 33 37 43 46 47 48 49 56 65 66 77 78 83 88 96 102 108 110 122 127 129 141 155 159 160 167 169 177 181 182 186 187 197 199 202 215 224 230 236 241 242 251 257 261 265 271 275 278 285 296 299 301 302 304 305 307 325 331 335 336 348 352 355 359 360 365 368 369 376 385 389 410 411 412 414 416 421 426 429 432 445 452 455 458 460 465 467 469 471 476 482 486 491 496 498 503 508 512 518 523 525 528 532 534 539 548 553 558 561 564 565 569 574 579 585 590 594 597 599 602 603 612 613 614 630 634 640 644 649 651 654 657 668 670 675 681 693 698 710 712 719 728 733 739 740 741 761 767 769 774 777 779 784 791 793 794 816 826 827 829 833 838 840 847 850 860 864 869 879 889

#We used the which function to find the index of all the missing values in the age column.

colSums(is.na(new\_titanic))

## Survived Pclass Sex Age   
## 0 0 0 0

#Now we see if any column consists of missing values in order to remove any columns missing more than 2 values.

str(new\_titanic)

## 'data.frame': 891 obs. of 4 variables:  
## $ Survived: chr "Dead" "Alive" "Alive" "Alive" ...  
## $ Pclass : chr "third" "first" "third" "first" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ Age : num 22 38 26 35 35 ...

#We are using the str function to check the data type and row values of the new dataset.

table(new\_titanic$Survived == "0")

##   
## FALSE TRUE   
## 496 395

#We use the table function to see how many times 0 has occured in the Survived variable.

table(new\_titanic$Survived)

##   
## 0 Alive Dead   
## 395 342 154

#As we can see above, there were 342 alive and 154 dead in the titanic ship which is not matching with the facts that happened in 1912. According to the facts, majority of the people who were present in the ship were dead. Moreover, we notice that there 395 0 values in the Survival column.

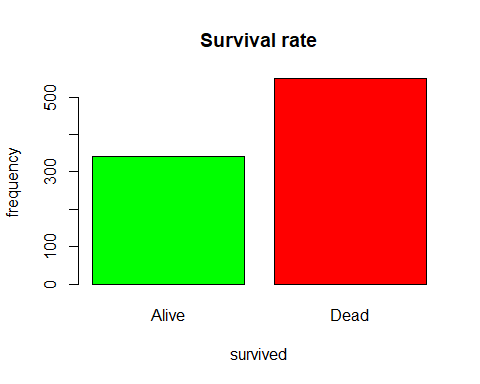
#Therefore, to solve this problem we are going to replace the 0’s by DEAD because we are going to assume that the 0’s in the survival column means that the passenger is dead.

titanic\_sr <- new\_titanic  
titanic\_sr$Survived[titanic\_sr$Survived == "0"] <- "Dead"  
unique(titanic\_sr$Survived) #We use the unique function to display the unique groups in the Survived column

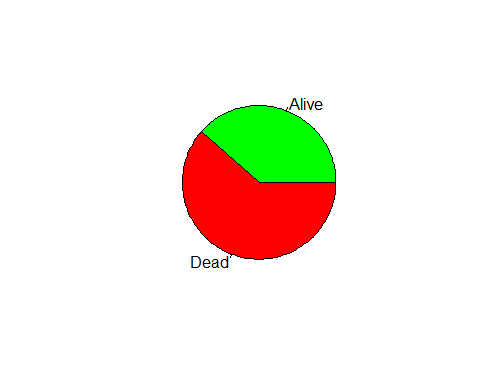
## [1] "Dead" "Alive"

#We now confirm the Survive column contains only alive and dead .

barplot(table(titanic\_sr$Survived), main="Survival rate", xlab="survived", ylab="frequency", col=c("green","red"))

 #We have created a plot for seeing the distribution of Alive and Dead values in the survivbal dataset. #Therefore, according to the dataset that we see above we can infer that majority of the people were dead.

pie(table(titanic\_sr$Survived), col=c("green","red"))

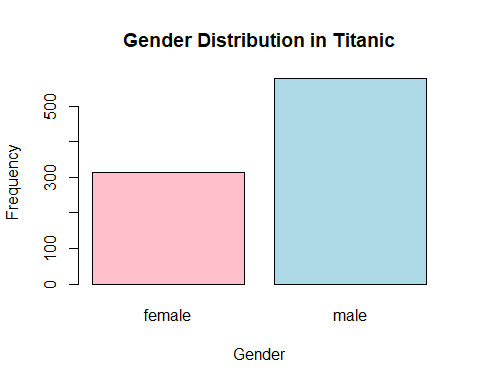
 #We also have created a pie chart to see the distribution of Alive and Dead people. From the pie chart we can see that more than 50% of the people on the Titanic were declared dead.

prop.table(table(titanic\_sr$Survived))

##   
## Alive Dead   
## 0.3838384 0.6161616

#To see the exact percentage of the dead and alive people in the titanic ship, we use the prop.table function. #So here we can see that 61.61% were Dead and 38.38% were Alive

barplot(table(titanic\_sr$Sex), main="Gender Distribution in Titanic", xlab="Gender", ylab="Frequency", col=c("pink","lightblue"))

 #As we can see above, the gender distributuion of the titanic is shown above. We can see that male population was much more than the female population in the titanic ship.

table(titanic\_sr$Survived, titanic\_sr$Sex)

##   
## female male  
## Alive 233 109  
## Dead 81 468

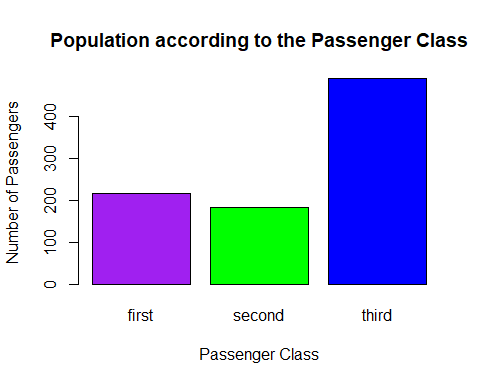
#We also have created a table to show how many male and female were alive or dead in the titanic ship. #We see that more females were alive and more males were dead.

table(titanic\_sr$Survived, titanic\_sr$Pclass)

##   
## first second third  
## Alive 136 87 119  
## Dead 80 97 372

#Here we have created a table to show how many people are alive or dead from the 3 different passenger classes.

barplot(table(titanic\_sr$Pclass), main="Population according to the Passenger Class", xlab="Passenger Class", ylab="Number of Passengers", col=c("purple","green","blue"))

 #Here, we have created a plot for the passenger class in the titanic ship. We can see that most passengers were from the Third class followed by first and second class.

#Conclusion: We have performed a lot of analysis on the given Titanic dataset to refine the data and hence perform various functions as shown above.

#It was not possible to get the the correlation between age and fare vraiable as the fare variable is not present in the dataset.