Assignment 1 Fundamentals of Machine Learning: Titanic Data Set

In this assignment I analyze the Titanic Data Set using R.Below shows Importing of the data and cleaning out the NA data , calculate descriptive statistics , transform variables and create visualizations as part of the requirements for Assignment 1

# Introduction

## Load Required Libraries

library(readr)  
library(ggplot2)  
library(tidyr)  
  
  
## Download and Import Dataset  
  
dataset\_url <- "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv"#Download data set  
titanic\_data <- read\_csv(dataset\_url) #Import the data set to R

## `curl` package not installed, falling back to using `url()`  
## 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.

## Handle Missing Values NA rows removed to ensure accurate statistical calculation.   
   
  
## Descriptive Statistics  
titanic\_data<-drop\_na((titanic\_data))  
summary(titanic\_data$Age) #Print out descriptive statistics for a selection of quantitative and categorical variables

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.92 24.00 36.00 35.67 47.50 80.00

summary(titanic\_data$Fare)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 29.70 57.00 78.68 90.00 512.33

summary(titanic\_data$SibSp)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.4645 1.0000 3.0000

#Mean and sd for the quantitative variables  
mean(titanic\_data$Age)

## [1] 35.67443

sd(titanic\_data$Age)

## [1] 15.64387

mean(titanic\_data$Fare)

## [1] 78.68247

sd(titanic\_data$Fare)

## [1] 76.34784

### Categorical Variables  
#frequency counts - descriptive statistics for categorical variables  
table(titanic\_data$Sex)

##   
## female male   
## 88 95

table(titanic\_data$Survived)

##   
## 0 1   
## 60 123

table(titanic\_data$Embarked)

##   
## C Q S   
## 65 2 116

#Provide the propirtions-descriptive statistics for the categorical variables  
prop.table(table(titanic\_data$Sex))

##   
## female male   
## 0.4808743 0.5191257

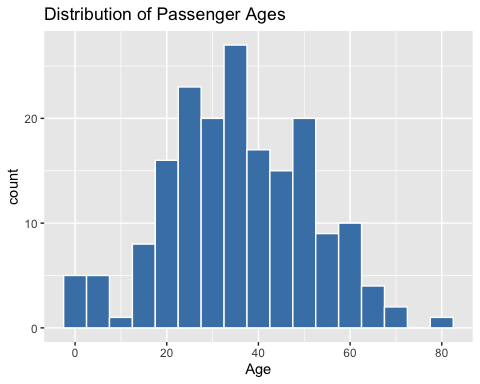
prop.table(table(titanic\_data$Survived))

##   
## 0 1   
## 0.3278689 0.6721311

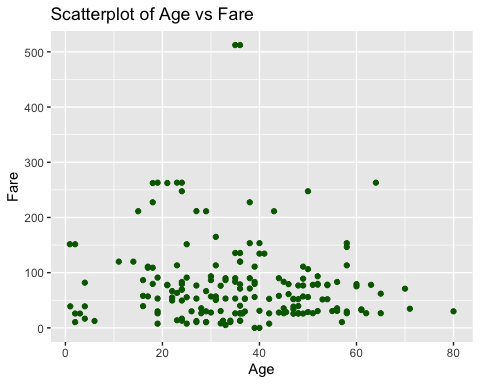
## Transform a Variable  
#Transform the numerical survival data to "yes" or "no"  
titanic\_data$SurvivedLabel <- ifelse(titanic\_data$Survived == 1, "Yes", "No")  
#showing the result  
  
### Histogram of Age  
table(titanic\_data$SurvivedLabel)

##   
## No Yes   
## 60 123

#histogram of age   
ggplot(titanic\_data,aes(x=Age))+geom\_histogram(binwidth=5,fill="steelblue",color="white")+labs(title = "Distribution of Passenger Ages",x="Age",y="count")



### Scatterplot: Age vs. Fare  
ggplot(titanic\_data,aes(x=Age,y=Fare))+geom\_point(color="darkgreen")+labs(title="Scatterplot of Age vs Fare",x="Age", y="Fare")



## Conclusion This R Markdown document demonstrates how to work with a real-world dataset using R. I loaded and cleaned the Titanic dataset, summarized variables, transformed a column, and created visualizations. This assignment helped reinforce key data analysis skills in R and the Markdown documentation format.

Additional references:

"Data Visualization with R in 36 Minutes." YouTube, uploaded by StatQuest with Josh Starmer, 9 Oct. 2023, <https://www.youtube.com/watch?v=McL9MMwmIZY>

"Learn R in 39 Minutes." YouTube, uploaded by R Programming 101, 9 Oct. 2023, <https://youtu.be/yZ0bV2Afkjc>.