Descriptive Stats ML-1

Jenny Prashanth Buraga

09/09/2023

This Assignment will focus on getting familiarized with R and Github.

I have taken the dataset about the Real Titanic’s Passengers Attributes. It has data about 887 rows (Passengers) and 8 Columns (Attributes)

#Getting data in R

#Load the library package  
library()  
  
#Load the dataset as a csv file  
ttnc<-read.csv("C:/Users/ssim/Downloads/titanic.csv")  
  
#Select the first 10 rows and 5 columns of the dataframe  
small.ttnc=ttnc[c(1:10),c(1:5)]  
  
#Summary of the small dataframe  
print(small.ttnc)

## Survived Pclass Name Sex  
## 1 0 3 Mr. Owen Harris Braund male  
## 2 1 1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  
## 3 1 3 Miss. Laina Heikkinen female  
## 4 1 1 Mrs. Jacques Heath (Lily May Peel) Futrelle female  
## 5 0 3 Mr. William Henry Allen male  
## 6 0 3 Mr. James Moran male  
## 7 0 1 Mr. Timothy J McCarthy male  
## 8 0 3 Master. Gosta Leonard Palsson male  
## 9 1 3 Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson female  
## 10 1 2 Mrs. Nicholas (Adele Achem) Nasser female  
## Age  
## 1 22  
## 2 38  
## 3 26  
## 4 35  
## 5 35  
## 6 27  
## 7 54  
## 8 2  
## 9 27  
## 10 14

#Structure of the dataframe  
str(small.ttnc)

## 'data.frame': 10 obs. of 5 variables:  
## $ Survived: int 0 1 1 1 0 0 0 0 1 1  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2  
## $ Name : chr "Mr. Owen Harris Braund" "Mrs. John Bradley (Florence Briggs Thayer) Cumings" "Miss. Laina Heikkinen" "Mrs. Jacques Heath (Lily May Peel) Futrelle" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ Age : num 22 38 26 35 35 27 54 2 27 14

## Descriptive Statistics

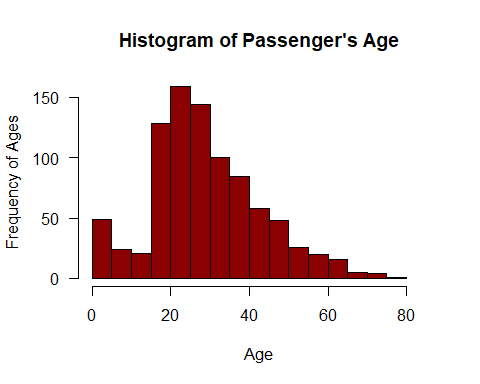
#Quantitative Variables

# 1.Passenger’s Age

#Statistics Summary of Passenger's Age  
summary(ttnc$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.42 20.25 28.00 29.47 38.00 80.00

#Make a Histogram to analyze Passenger's Age Pattern  
hist(ttnc$Age, main="Histogram of Passenger's Age", xlab="Age", border="black", col="darkred", xlim=c(0,85), ylab="Frequency of Ages", las=1, breaks=20)



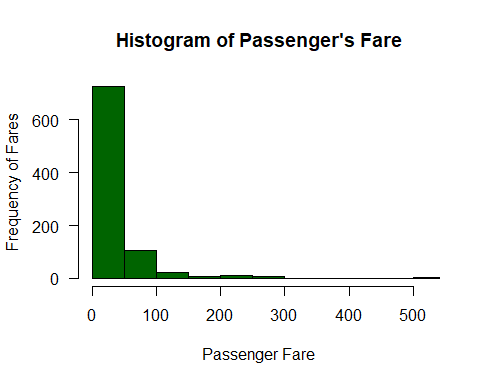
We can see in the Histogram that majority of the passengers in Titanic were aged between 20-40 Years. Also, we can notice that they were very small number people older than 65 Years of age.

# 2.Passenger’s Fare

#Statistics Summary of Passenger's Fare paid to travel on Titanic.  
summary(ttnc$Fare)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 7.925 14.454 32.305 31.137 512.329

#Make a Histogram to analyze Passenger's Fare Patterns  
hist(ttnc$Fare, main="Histogram of Passenger's Fare", xlab="Passenger Fare", border="black", col="darkgreen", xlim=c(0,520), ylab="Frequency of Fares", las=1, breaks=10)



Here we can see that a vast majority of people on Titanic paid Fares lower than 50 bucks. Also, very small number of passengers paid Fares around 500 bucks which is indiciated by the tail of the histogram on the right meaning they were Rich People.

#Categorical Variables

#Passenger’s Sex

#Summary of Passenger's Sex  
table(ttnc$Sex)

##   
## female male   
## 314 573

#Here we will now create a pie chart and also transform the variable to make a Percentage Pie Chart  
str(ttnc$Sex)

## chr [1:887] "male" "female" "female" "female" "male" "male" "male" "male" ...

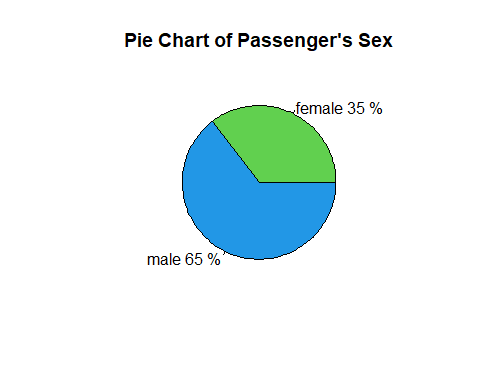
#We will now transform this variable which is in Character form to Factor form  
Gender=as.factor(ttnc$Sex)  
  
table(Gender)

## Gender  
## female male   
## 314 573

Freqgender=table(Gender)  
  
#Calculate Percentage of Males & Females  
percent=round(Freqgender/887\*100)  
percent

## Gender  
## female male   
## 35 65

#Creating PieChart showing Passenger's Gender  
lbl=paste(names(Freqgender),percent,"%",sep=" ")  
  
pie(Freqgender, main="Pie Chart of Passenger's Sex",col=c(3,4),labels=lbl)

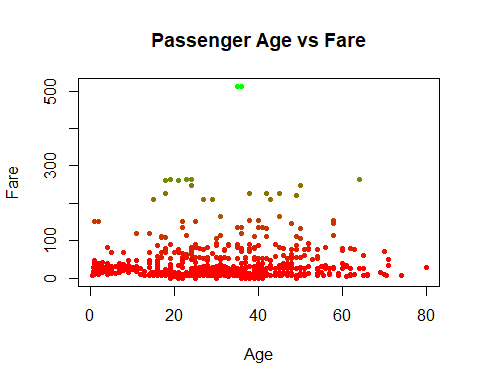


So, here I have taken a Categorical Variable which was in Character form and transformed it to factor type variable to calculate the percentage with which I have created a Pie Chart.

This Pie Chart shows that majority of the passengers aboard the Titanic were Males.

## Scatterplot

#Function to generate a continuous color palette  
mycolor<-colorRampPalette(c("red","green"))  
  
#Add color palette based on Passenger Fare  
mycolor2<-mycolor(10)[as.numeric(cut(ttnc$Fare, breaks=10))]  
  
#Plot Age vs Fare  
plot(ttnc$Age,ttnc$Fare, main="Passenger Age vs Fare", xlab="Age", ylab="Fare", col=mycolor2, pch=20)



This Scatterplot shows the distribution of Passenger Age and Fare they paid. Interesting point in this scatterplot is the highest Fare was paid by Passengers that were around 35 Years of age.

Dataset Source: © Stanford 2016 | Created by Chris. CS109 has been developed by many talented teachers