GR5206 HW1 mz2692

Steven

9/14/2018

Problem 1

Part 1:Importing Data into R

i.

```
#Import data into titanic dataframe
file="//Users//zhongming//Downloads//Titanic.txt"
titanic=read.table(file,header=T,as.is=T)
titanic=as.data.frame(titanic)
```

ii.

```
#get the dimension of titanic
dim(titanic)
```

```
## [1] 891  12
```

Ans: It contains 891 rows and 12 columns.

iii.

```
#Subsetting the titanic dataframe
Survived.Word=titanic["Survived"]

#convert dummy data into characters
Survived.Word[Survived.Word==0]="died"
Survived.Word[Survived.Word==1]="survived"
```

Part 2:Exploring the Data in R

i.

```
#using apply to get mean
mean=sapply(subset(titanic,select=c(Survived,Age,Fare)),mean)
mean
```

```
##   Survived        Age       Fare
##  0.3838384         NA 32.2042080
```

Ans: The mean of Survived tells us the propotion of people survived. The Age mean value is NA because the titanic[“Age”] is a list and it contains NA.

ii.

```
#subset and get the data that are female and survived.
d2=subset(titanic,Survived==1,Sex=="female")

#calculate the proportion of survived female in total dataset
female_survived1=round(nrow(d2)/nrow(titanic),digits = 2)
female_survived1
```

```
## [1] 0.38
```

iii.

```
#subset and get data that are survived.
d3=subset(titanic,Survived==1,select=Sex)

#calculate the propotion of survived feamle in total survived people
female_survived2=round(length(d3[d3=="female"])/length(d3[,]),digits=2)
female_survived2
```

```
## [1] 0.68
```

iv.

```
#calculate the different survival rate of 3 classes
classes <- sort(unique(titanic$Pclass))
Pclass.Survival <- vector("numeric", length = 3)
names(Pclass.Survival) <- classes
for (i in 1:3) {
  Pclass.Survival[i]=round(colMeans(subset(titanic,Pclass==i,select=Survived)),digits = 2)
}
Pclass.Survival
```

```
##    1    2    3
## 0.63 0.47 0.24
```

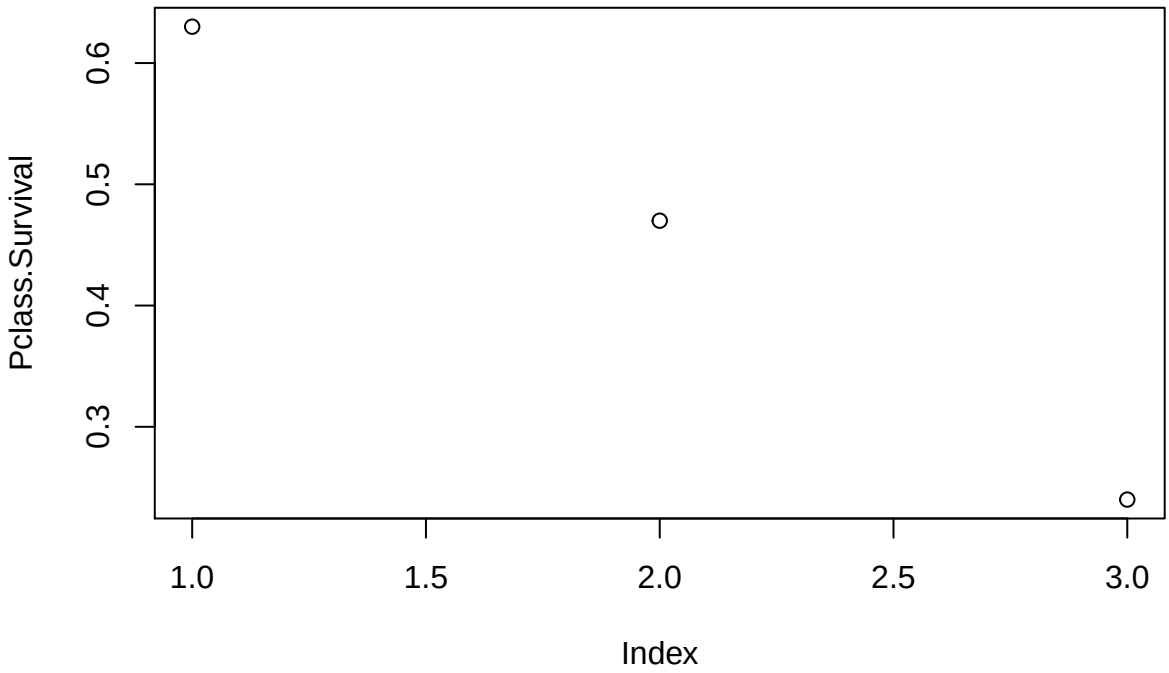
v.

```
#using tapply to get the same result as question iv.
Pclass.Survival2=round(tapply(titanic[["Survived"]],titanic[["Pclass"]],mean),digits = 2)
Pclass.Survival2==Pclass.Survival
```

```
##    1    2    3
## TRUE TRUE TRUE
```

vi.

```
plot(Pclass.Survival)
```




```
cor.test(classes,Pclass.Survival[1:3])
```

```
##
##  Pearson's product-moment correlation
##
## data:  classes and Pclass.Survival[1:3]
## t = -9.65, df = 1, p-value = 0.06574
## alternative hypothesis: true correlation is not equal to 0
## sample estimates:
##        cor
## -0.9946736
```

Ans: According to the correlation test, there seems to be a relationship between survival rate and classes.