Titanic.R

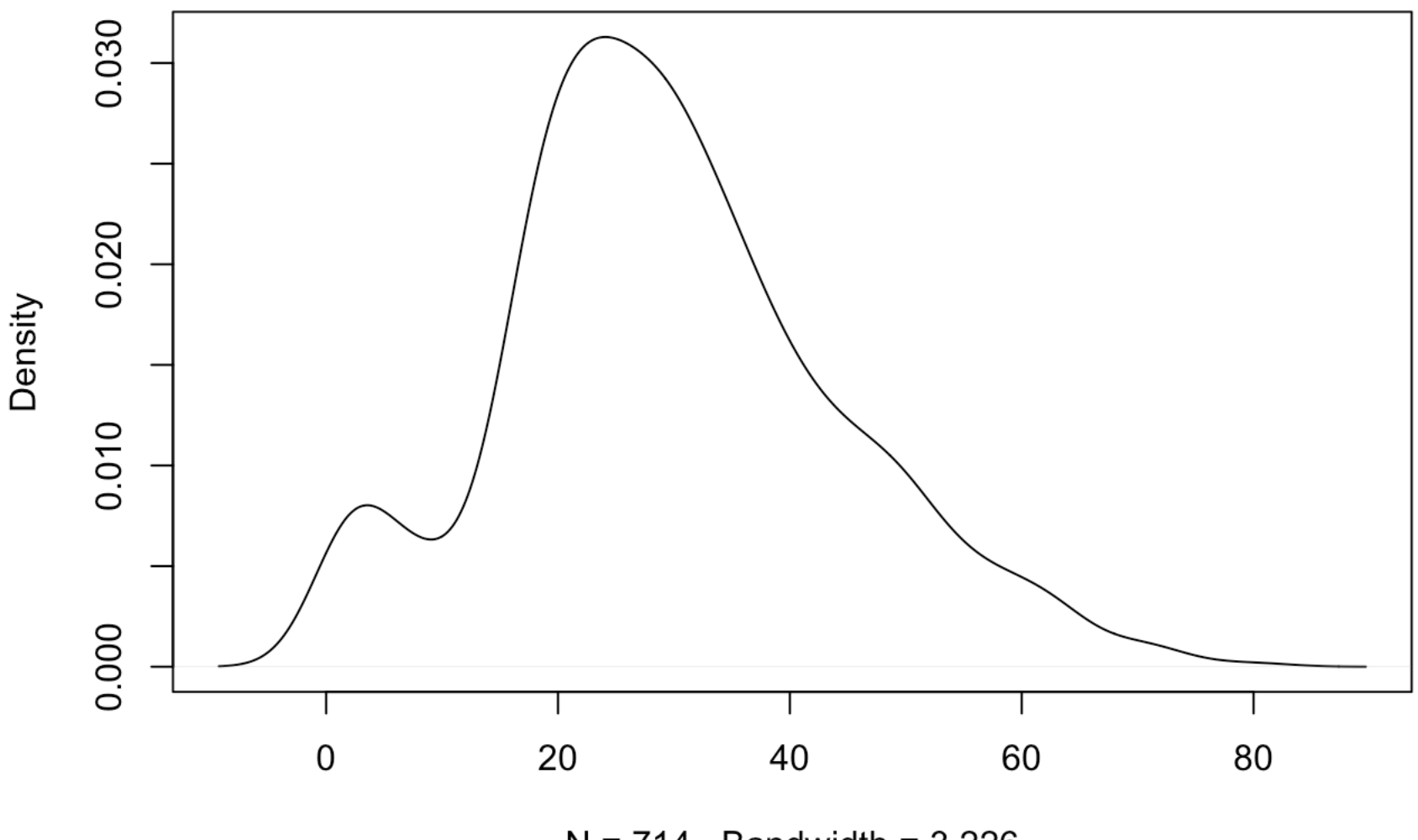
JingbinXu

Wed Dec 21 12:14:57 2016

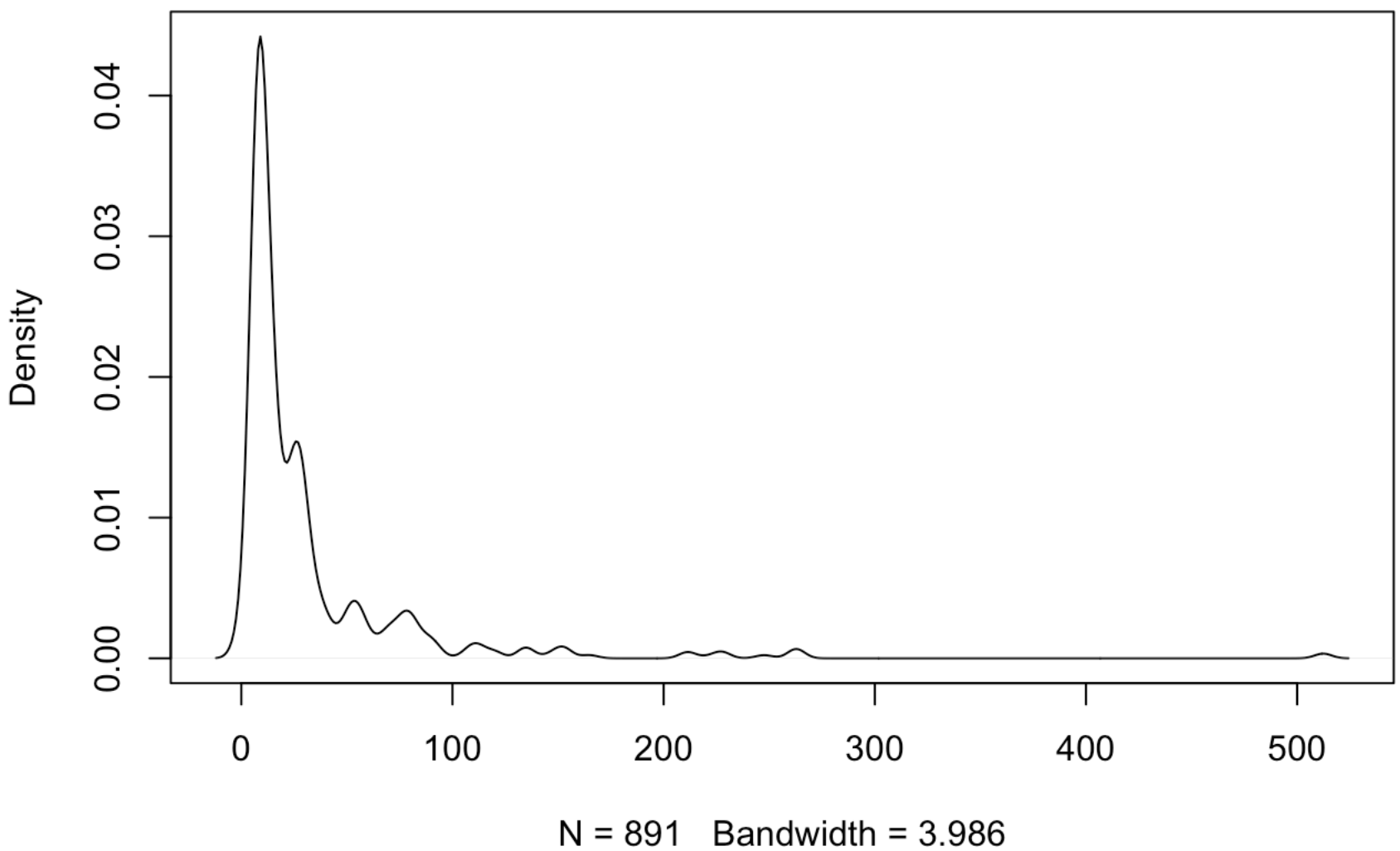
```
##Inputting the Test and Train datasets into Rstudio
traindata<-read.csv("train.csv")
testData<-read.csv("test.csv")
head(traindata)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp
## 1                            Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                             Heikkinen, Miss. Laina female  26     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                           Allen, Mr. William Henry   male  35     0
## 6                                   Moran, Mr. James   male  NA     0
##   Parch            Ticket    Fare Cabin Embarked
## 1     0         A/5 21171  7.2500              S
## 2     0          PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250              S
## 4     0            113803 53.1000  C123        S
## 5     0            373450  8.0500              S
## 6     0            330877  8.4583              Q
```

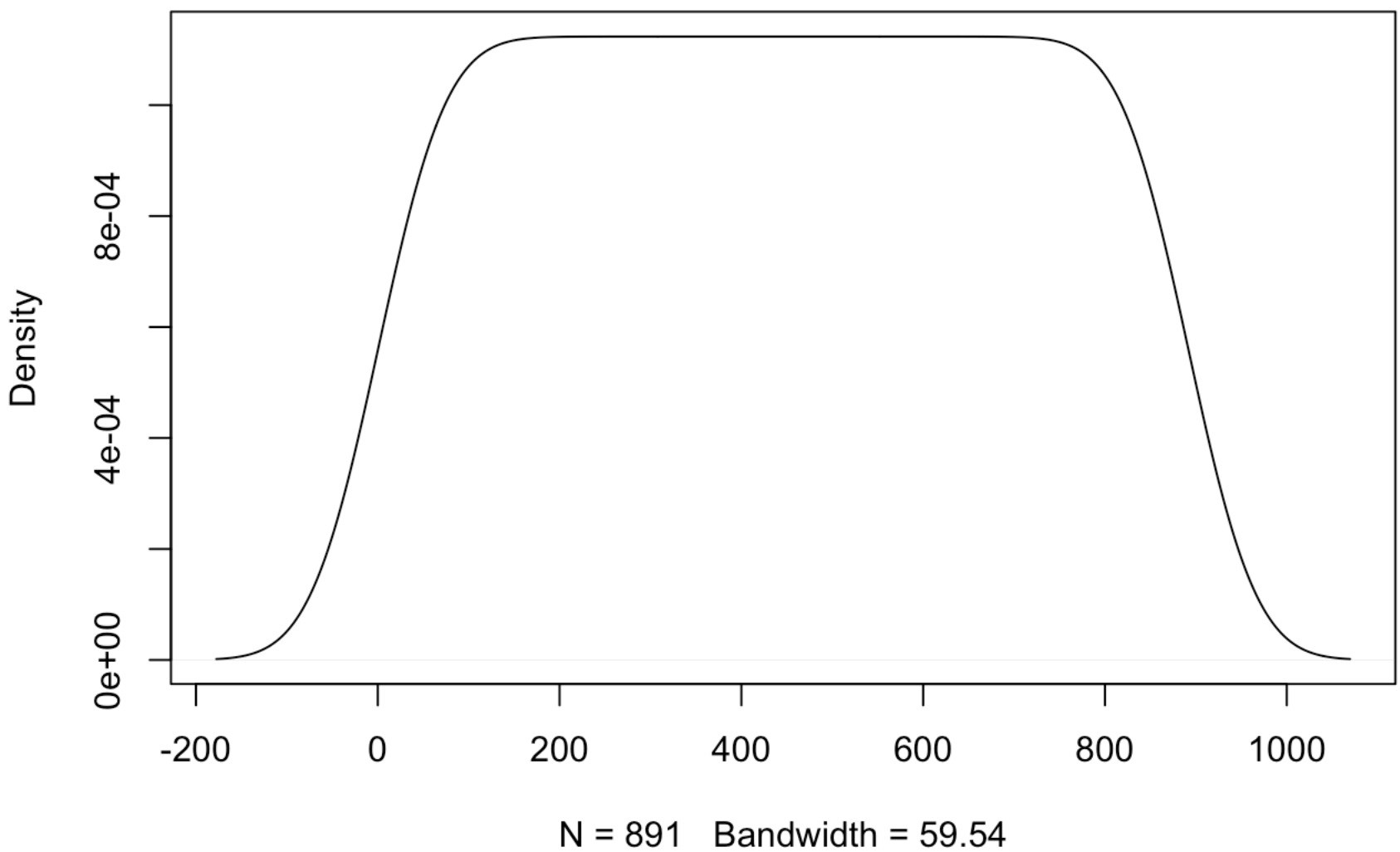
```
##Making basic visualizations in Rstudio
plot(density(traindata$Age,na.rm=TRUE))
```



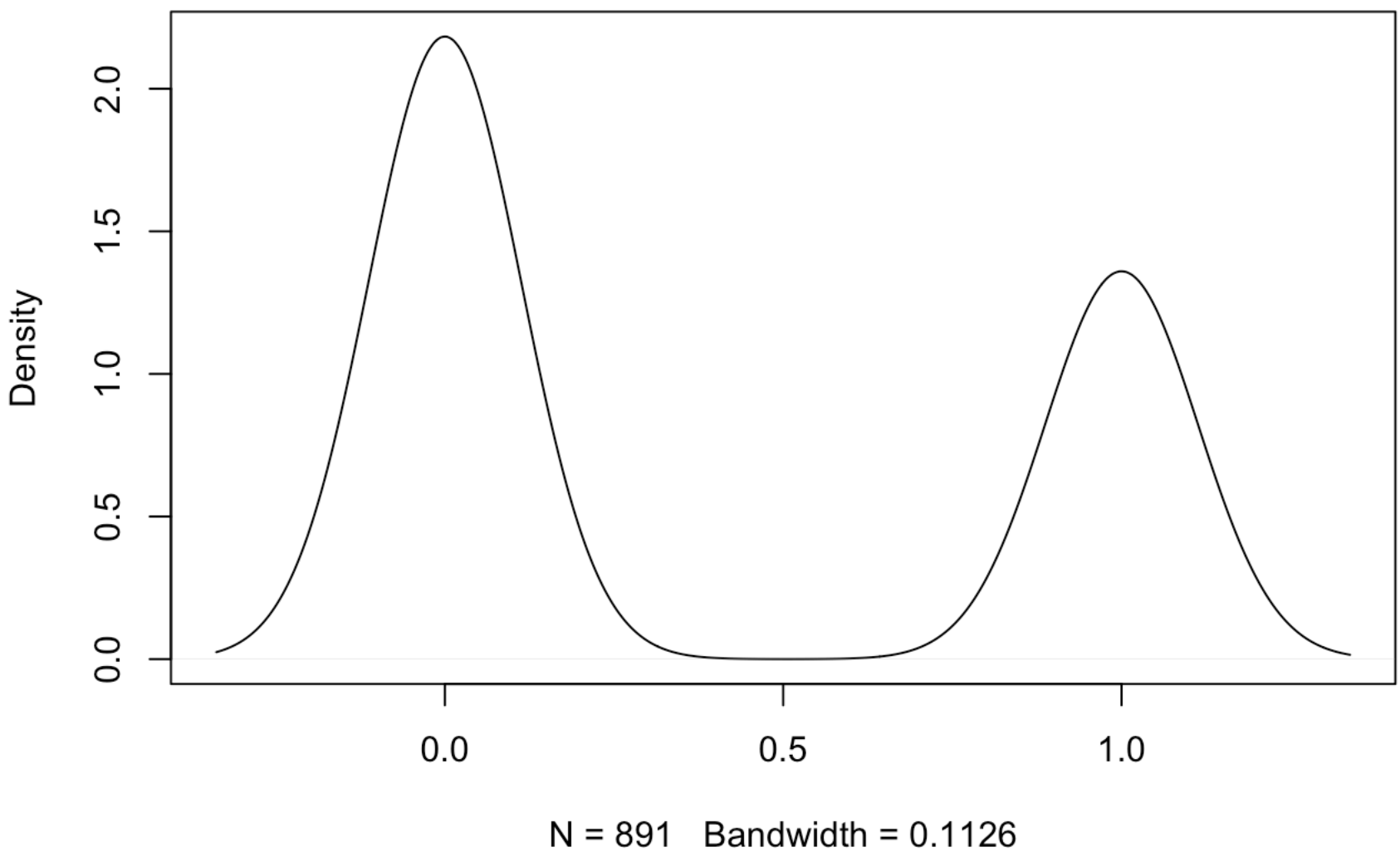

```
plot(density(traindata$Fare,na.rm=TRUE))
```



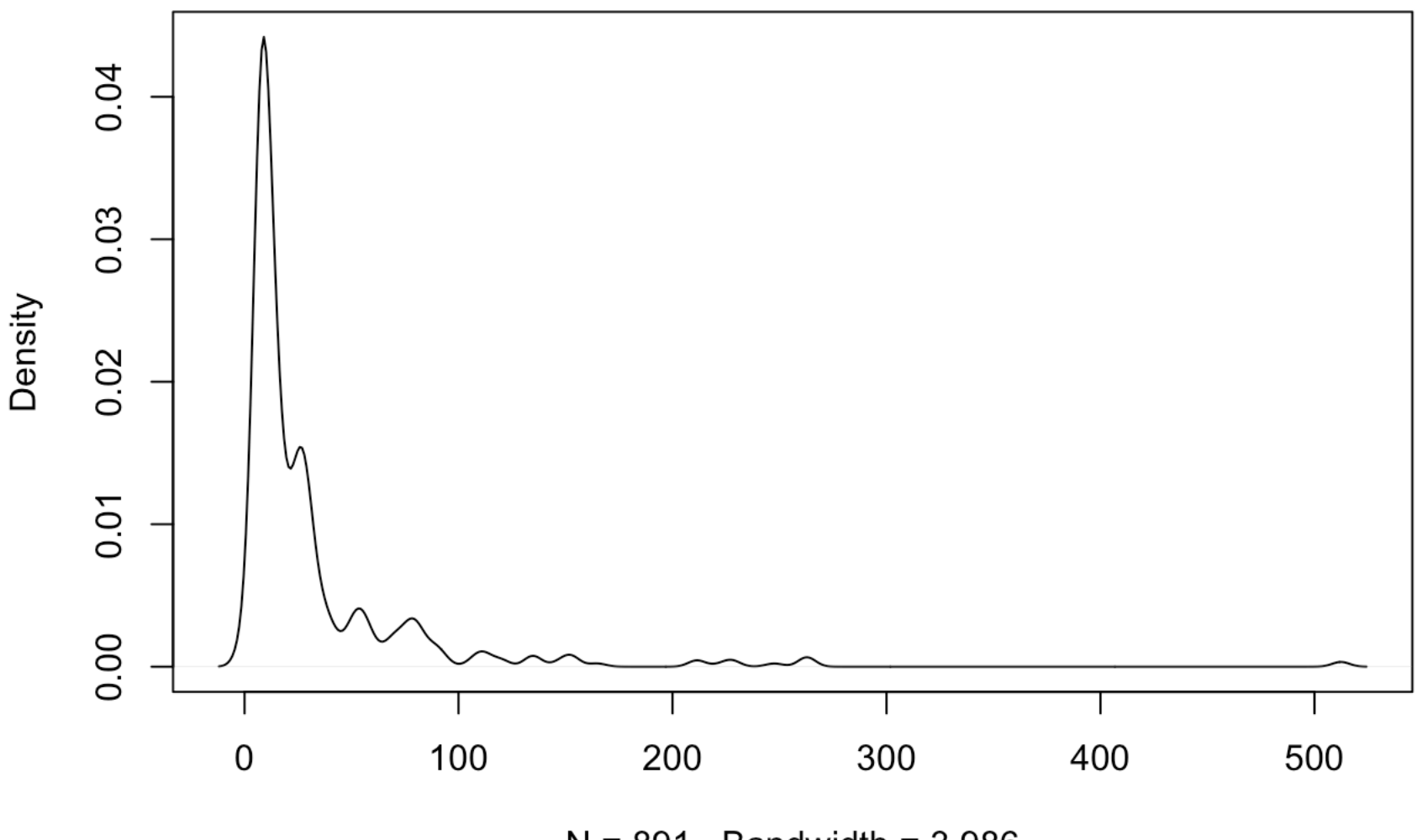

```
plot(density(traindata$PassengerId,na.rm=TRUE))
```




```
plot(density(traindata$Survived,na.rm=TRUE))
```




```
plot(density(traindata$Fare,na.rm=TRUE))
```




```
#By first plotting the density we are able to get a sense of how the overall data fee
l and get a few bague answers:
#1.where is the general center?
#Is there a skew?
#Does is generally take higher values?
#Where are most of the values concentrated?

##Survival Rate by Sex Barplot
counts<-table(traindata$Survived,traindata$Sex)
barplot(counts,xlab = "Gender",ylab = "Number of People",main = "survived")
```


```
counts[2]/(counts[1]+counts[2])
```

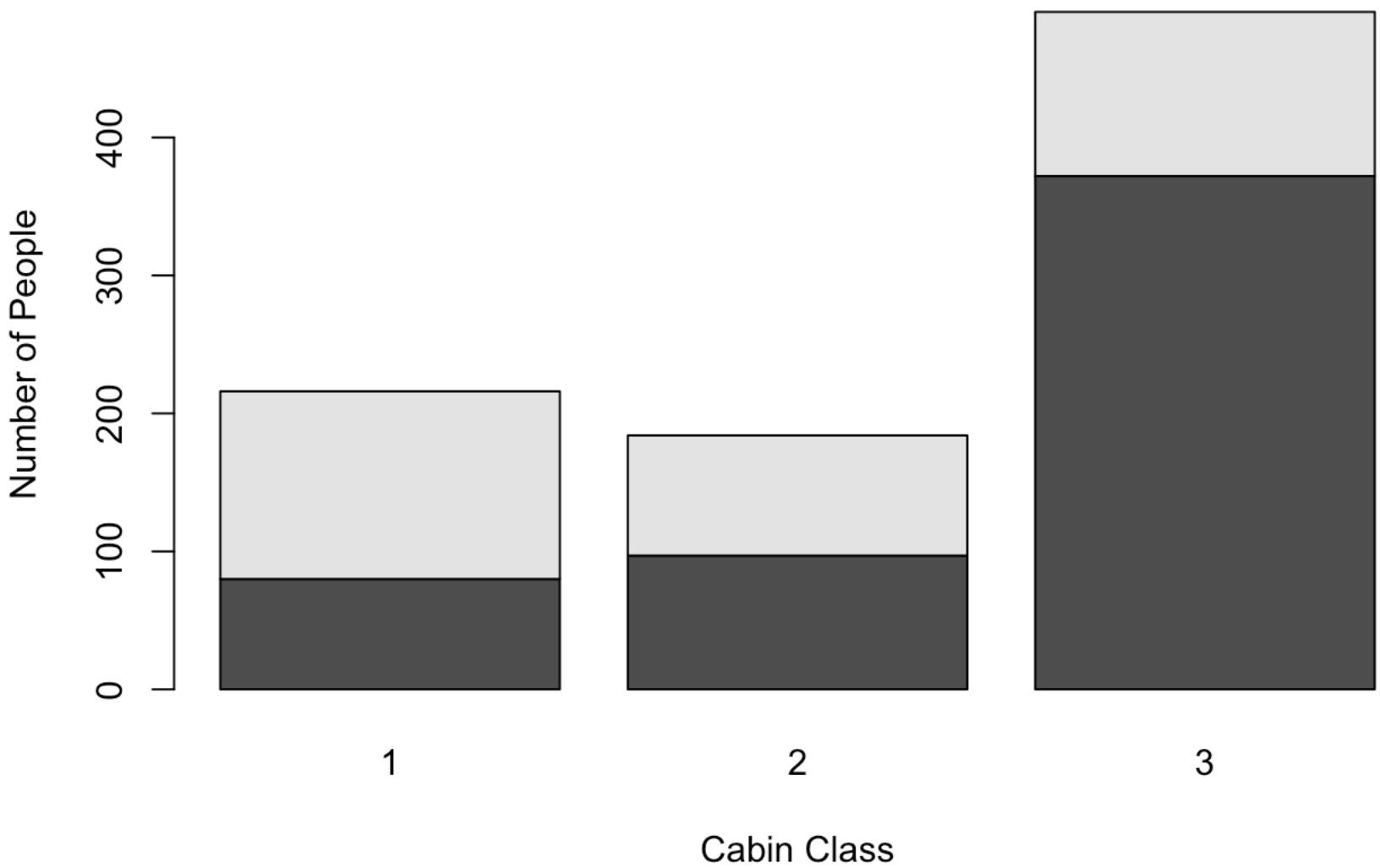
```
## [1] 0.7420382
```

```
counts[4]/(counts[3]+counts[4])
```

```
## [1] 0.1889081
```

```
#74.2% of women survived versus 18.9% of men.

##Suvival Rate by Passaenger Class Barplot
Pclass_survival<-table(traindata$Survived,traindata$Pclass)
barplot(Pclass_survival,xlab="Cabin Class",ylab="Number of People",main="survived and
deceased between male and female")
```




```
Pclass_survival[2]/(Pclass_survival[1]+Pclass_survival[2])
```

```
## [1] 0.6296296
```

```
Pclass_survival[4]/(Pclass_survival[3]+Pclass_survival[4])
```

```
## [1] 0.4728261
```

```
Pclass_survival[6]/(Pclass_survival[5]+Pclass_survival[6])
```

```
## [1] 0.2423625
```

```
#63.0%, 1st class
#47.4%, 2nd class
#24.2%, 3rd class

##Conclusion
#The key idea is that we are trying to determine if any variables are realated to wha
t we are trying to predict: Suvived.
```