Data preparation and Exploratory 1

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

Load de data

```
alldata<- read.csv("../data/train.csv")
alldata$dataset<-runif(nrow(alldata))


training <- alldata[alldata$dataset<=.6,]
validation <- alldata[alldata$dataset>.6 & alldata$dataset<=.8,]
testing <- alldata[alldata$dataset>.8,]


training$dataset <-NULL;
validation$dataset <-NULL;
testing$dataset  <-NULL;
alldata<-NULL;


str(training)
```

```
## 'data.frame':    553 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 6 7 8 9 10 12 ...
##  $ Survived   : int  0 1 1 1 0 0 0 1 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 1 3 3 2 1 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 559 520 629 416 581 9
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 1 1 1 ...
##  $ Age        : num  22 38 26 35 NA 54 2 27 14 58 ...
##  $ SibSp      : int  1 1 0 1 0 0 3 0 1 0 ...
##  $ Parch      : int  0 0 0 0 0 0 1 2 0 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 525 596 662 50 276 86 396 345 133 39 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.46 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 131 1 1 1 51 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 3 4 4 4 2 4 ...
```

Change type for cathegorical vars

```
training$Survived<-factor(training$Survived)
training$Pclass<-factor(training$Pclass)
str(training)
```

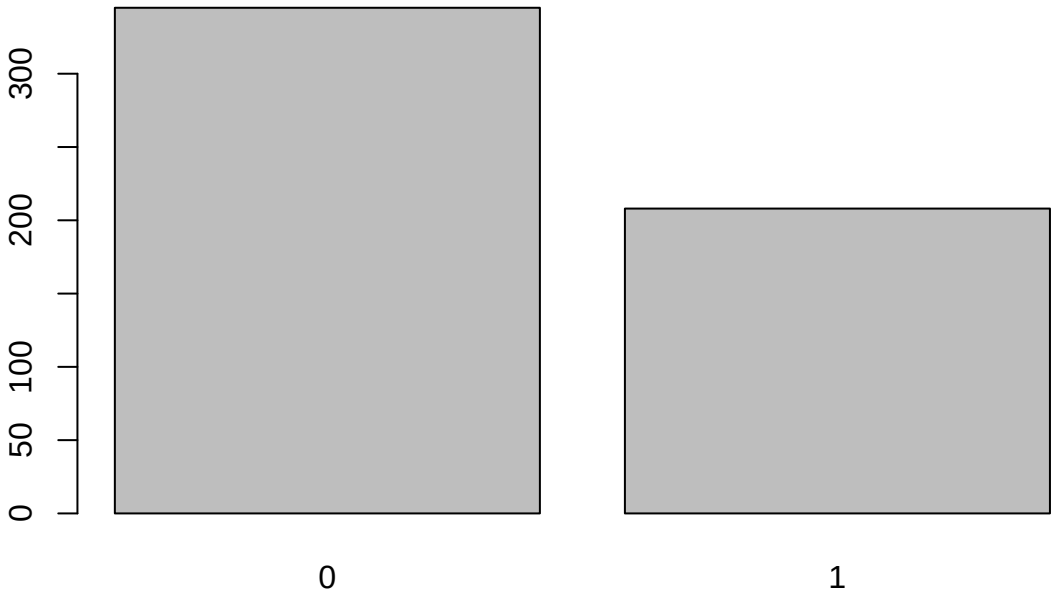
```
## 'data.frame':    553 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 6 7 8 9 10 12 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 1 3 3 2 1 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 559 520 629 416 581 9
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 1 1 1 ...
##  $ Age        : num  22 38 26 35 NA 54 2 27 14 58 ...
##  $ SibSp      : int  1 1 0 1 0 0 3 0 1 0 ...
##  $ Parch      : int  0 0 0 0 0 0 1 2 0 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 525 596 662 50 276 86 396 345 133 39 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.46 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 131 1 1 1 51 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 3 4 4 4 2 4 ...
```

```
summary(training)
```

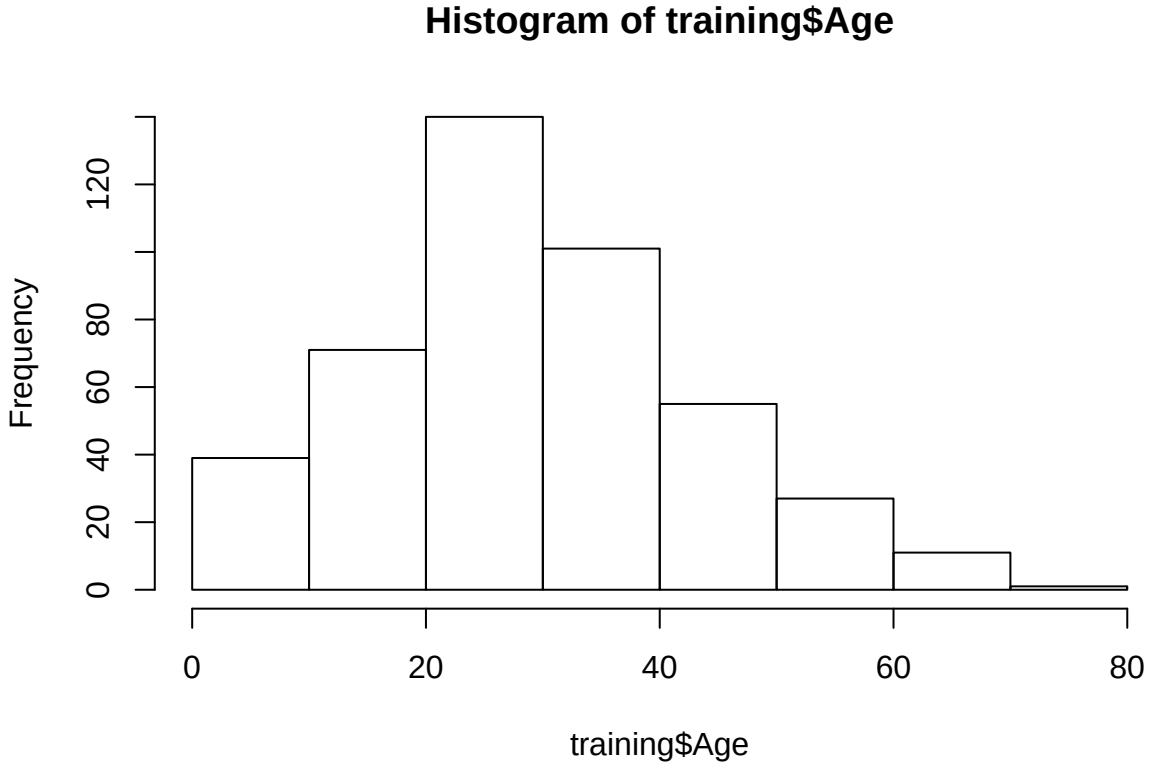
```
##   PassengerId    Survived Pclass
##  Min.   :  1.0   0:345    1:144
##  1st Qu.:214.0   1:208    2:108
##  Median :450.0            3:301
##  Mean   :443.9
##  3rd Qu.:668.0
##  Max.   :890.0
##
##                                                 Name         Sex
##  Abbing, Mr. Anthony                              :  1   female:189
##  Abbott, Mrs. Stanton (Rosa Hunt)                 :  1   male  :364
##  Abelson, Mrs. Samuel (Hannah Wizosky)            :  1
##  Adams, Mr. John                                  :  1
##  Ahlin, Mrs. Johan (Johanna Persdotter Larsson):  1
##  Aks, Mrs. Sam (Leah Rosen)                       :  1
##  (Other)                                          :547
##       Age            SibSp            Parch             Ticket
##  Min.   : 0.42   Min.   :0.0000   Min.   :0.0000   1601    :  6
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   CA. 2343:  5
##  Median :28.00   Median :0.0000   Median :0.0000   113781  :  4
##  Mean   :29.69   Mean   :0.4955   Mean   :0.3237   3101295 :  4
##  3rd Qu.:38.00   3rd Qu.:1.0000   3rd Qu.:0.0000   LINE    :  4
##  Max.   :71.00   Max.   :8.0000   Max.   :5.0000   110152  :  3
##  NA's   :108                                       (Other) :527
##       Fare                Cabin     Embarked
##  Min.   :  0.000             :426    :  0
##  1st Qu.:  7.896   C22 C26   :  3   C:115
##  Median : 14.400   C23 C25 C27:  3   Q: 42
##  Mean   : 33.044   F33       :  3   S:396
##  3rd Qu.: 31.387   B18       :  2
##  Max.   :512.329   B20       :  2
##                    (Other)   :114
```

Plots

```
plot(training$Survived)
```



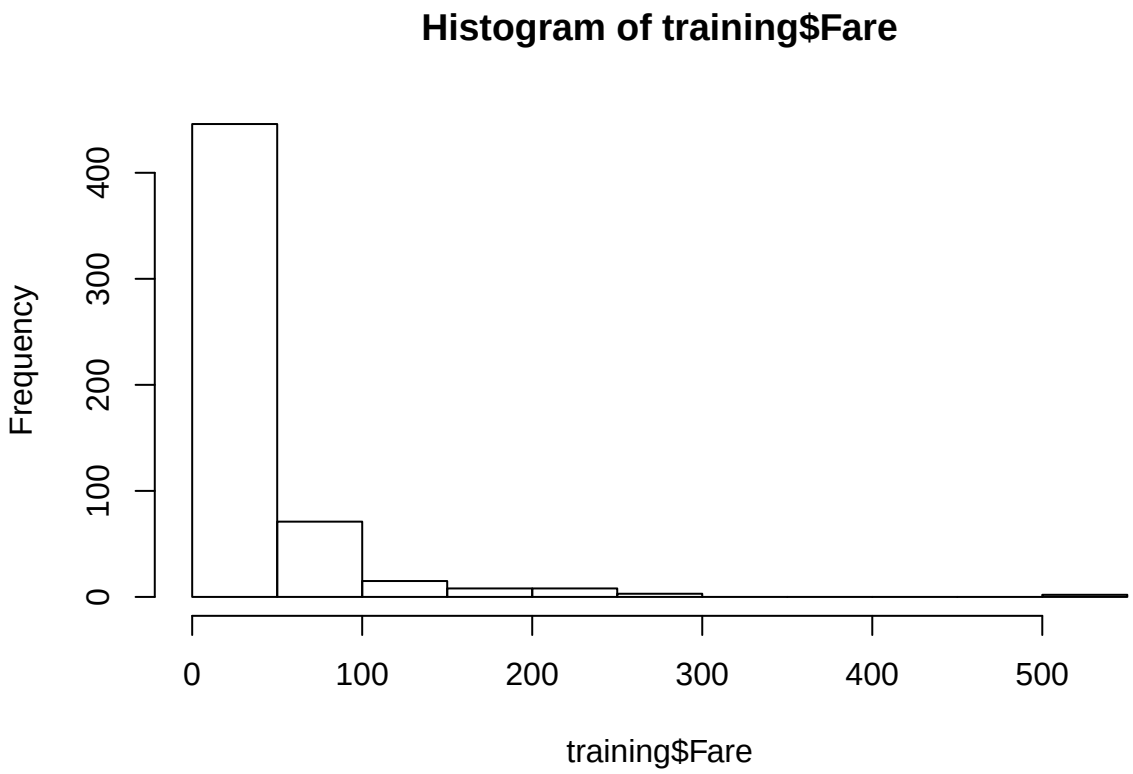

```
hist(training$Age)
```



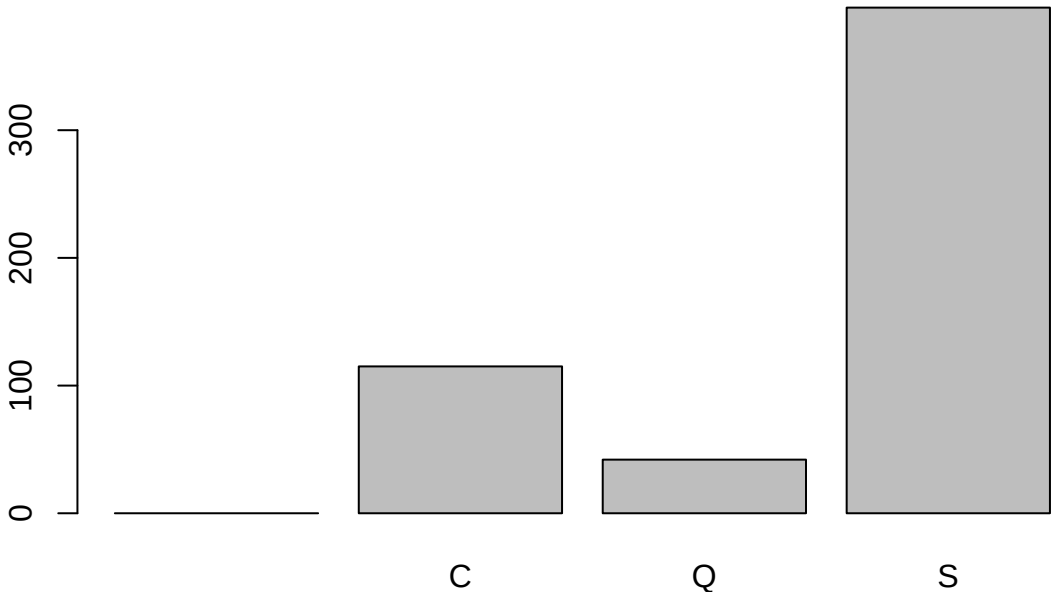

```
hist(training$SibSp)
```


```
hist(training$Parch)
```


```
hist(training$Fare)
```



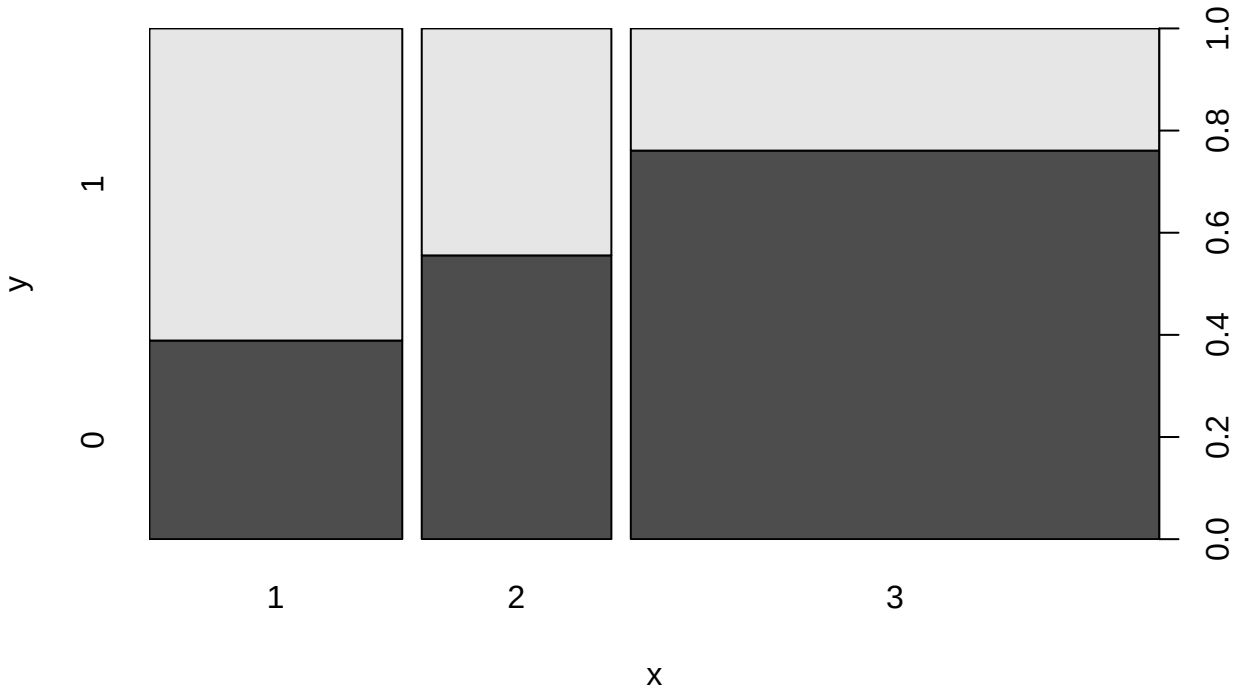

```
plot(training$Embarked)
```



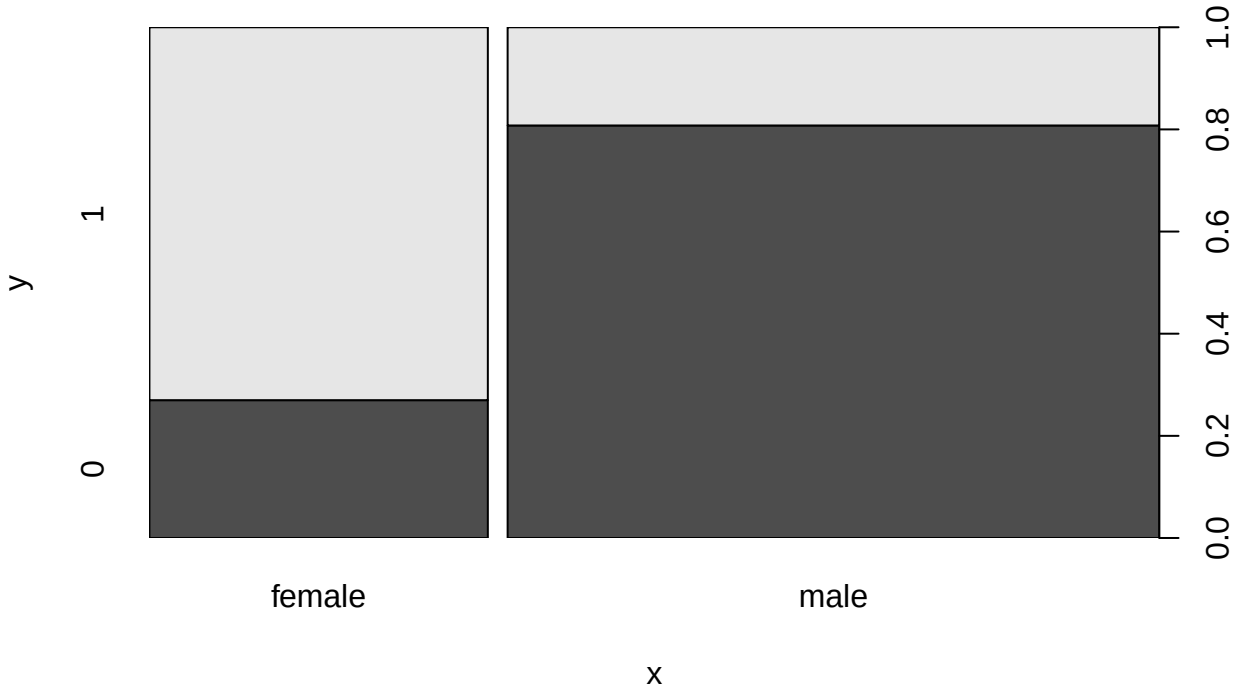
Bidimiensional distributions

Survival vs others

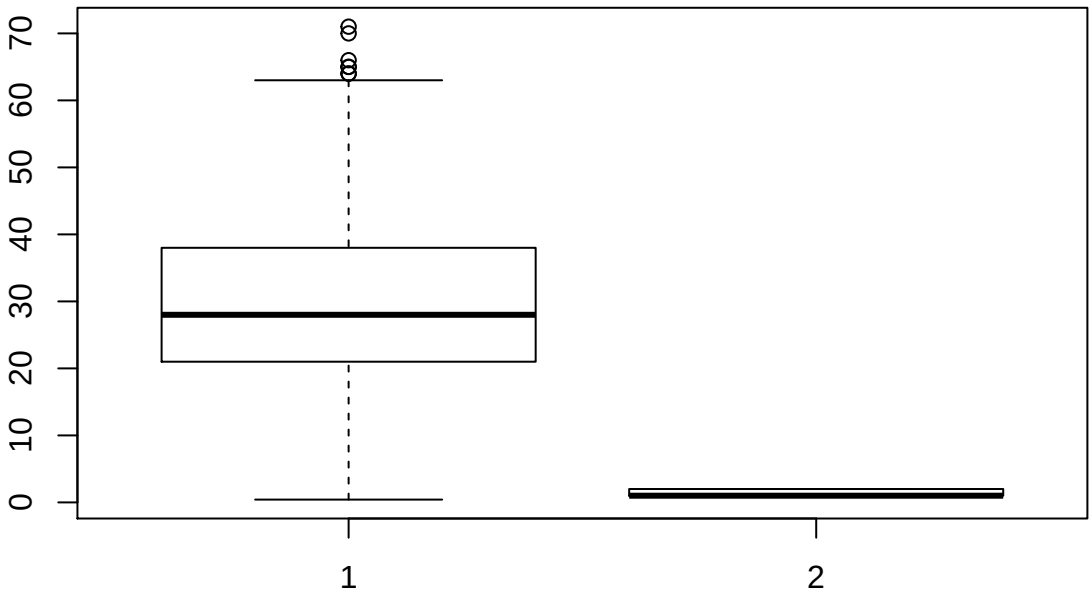
```
with(training,plot(Pclass,Survived))
```



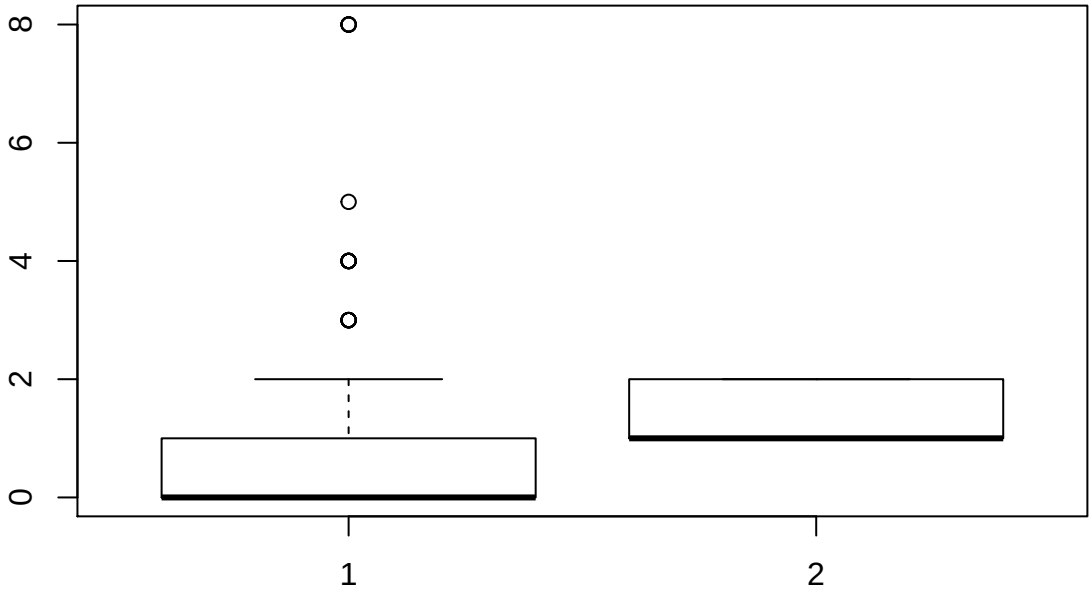

```
with(training,plot(Sex,Survived))
```



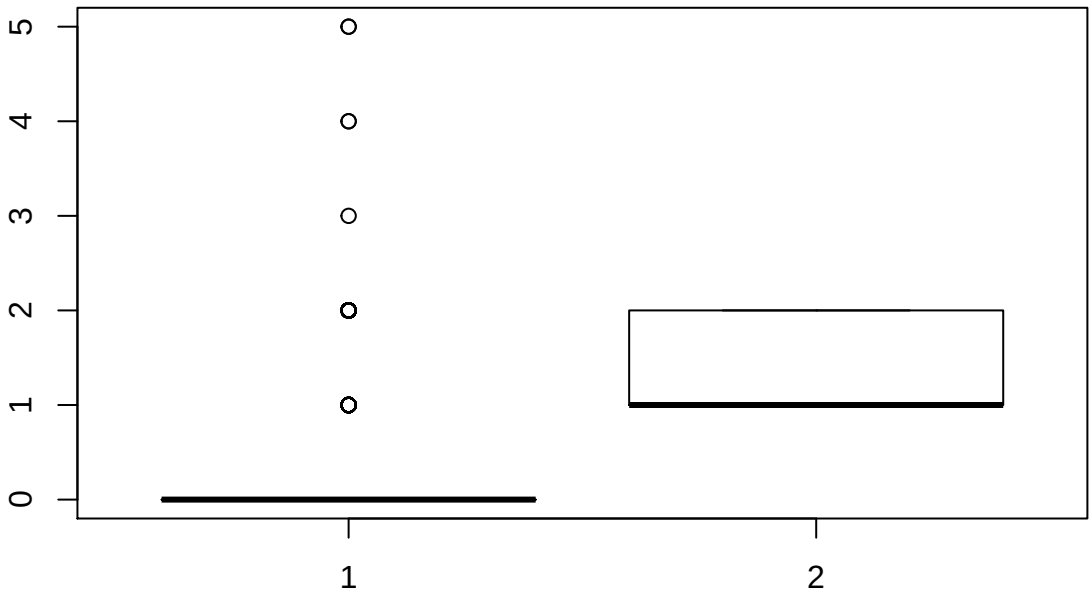

```
with(training,boxplot(Age,Survived))
```



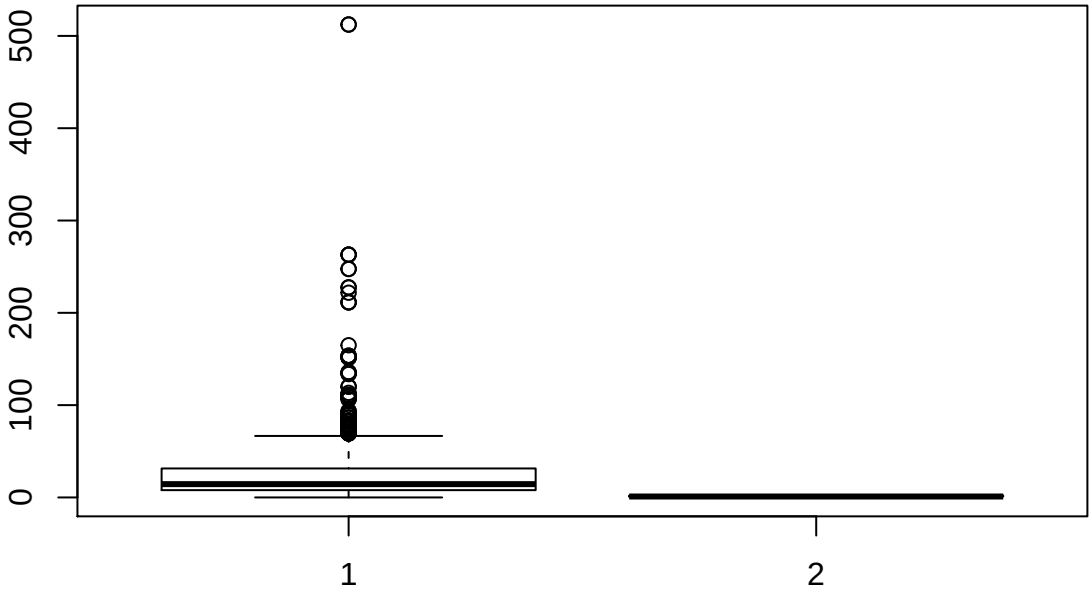

```
with(training,boxplot(SibSp,Survived))
```



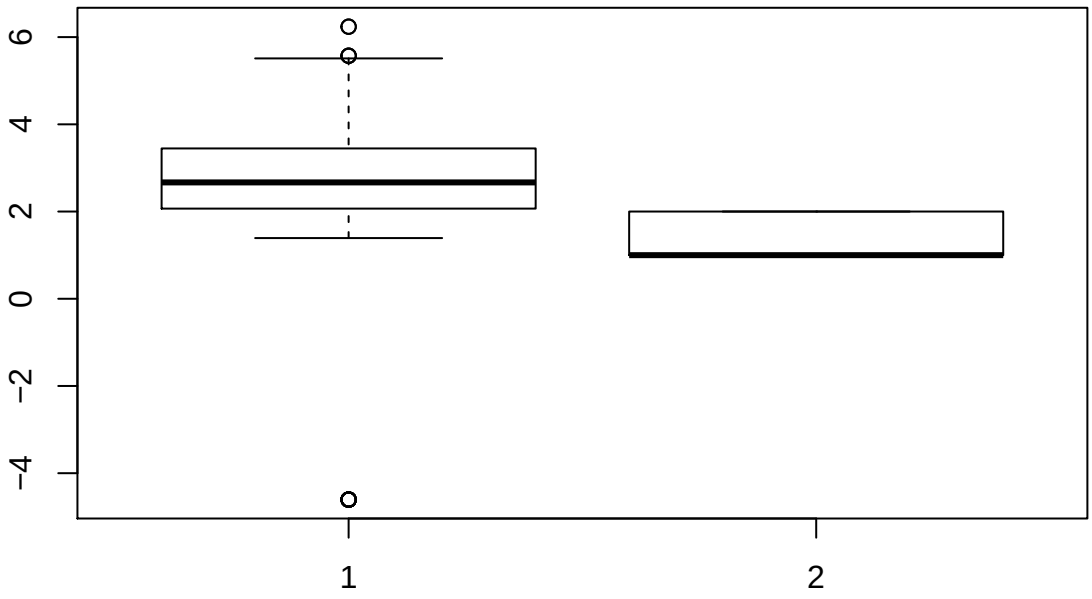

```
with(training,boxplot(Parch,Survived))
```



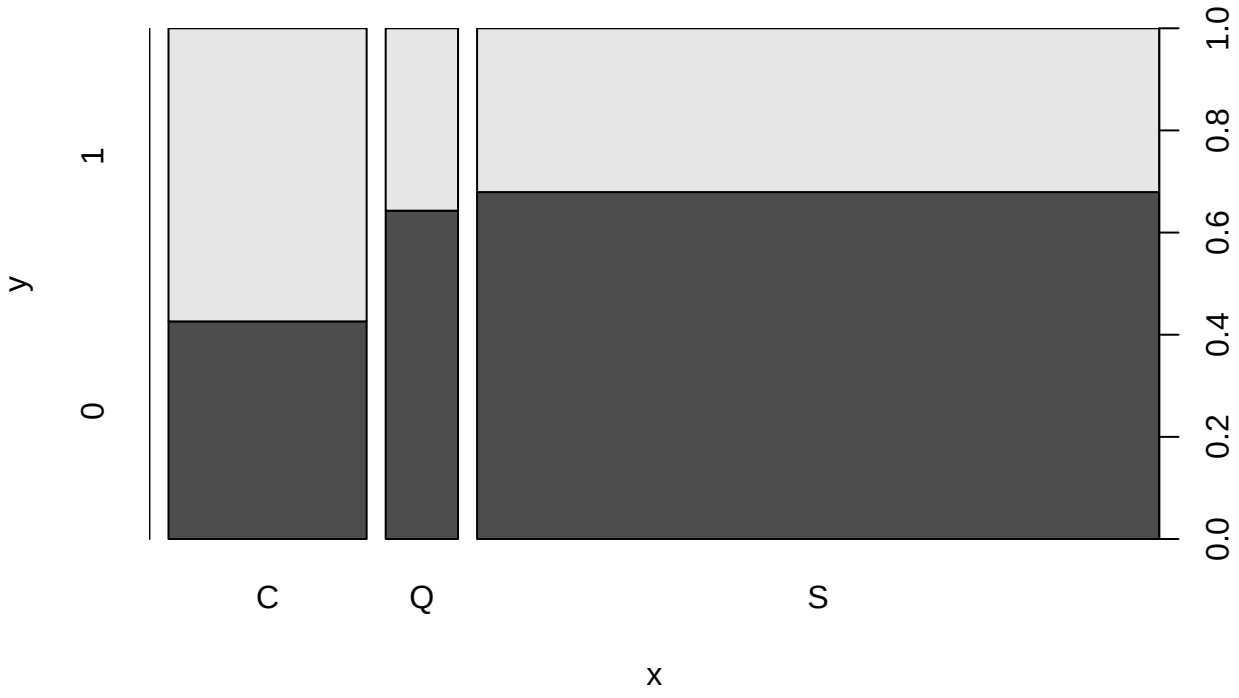

```
with(training,boxplot(Fare,Survived))
```




```
with(training,boxplot(log(Fare+.01),Survived))
```




```
with(training,plot(Embarked,Survived))
```



Check for data anomalies

No age?

```
summary (training[is.na(training$Age),])
```

```
##   PassengerId    Survived Pclass
##  Min.   :  6.0   0:79     1:18
##  1st Qu.:193.2   1:29     2: 7
##  Median :418.5            3:83
```

```
##  Mean   :412.9
##  3rd Qu.:605.5
##  Max.   :879.0
##
##                                             Name         Sex
##  Baumann, Mr. John D                           :  1   female:30
##  Boulos, Mrs. Joseph (Sultana)                 :  1   male  :78
##  Bourke, Miss. Mary                            :  1
##  Bradley, Mr. George ("George Arthur Brayton"):  1
##  Brewe, Dr. Arthur Jackson                     :  1
##  Cairns, Mr. Alexander                         :  1
##  (Other)                                       :102
##       Age          SibSp            Parch             Ticket
##  Min.   : NA   Min.   :0.0000   Min.   :0.0000   CA. 2343: 5
##  1st Qu.: NA   1st Qu.:0.0000   1st Qu.:0.0000   1601    : 3
##  Median : NA   Median :0.0000   Median :0.0000   371110  : 3
##  Mean   :NaN   Mean   :0.5833   Mean   :0.1574   239853  : 2
##  3rd Qu.: NA   3rd Qu.:0.0000   3rd Qu.:0.0000   367230  : 2
##  Max.   : NA   Max.   :8.0000   Max.   :2.0000   4133    : 2
##  NA's   :108                                     (Other) :91
##       Fare             Cabin     Embarked
##  Min.   :  0.000          :103    : 0
##  1st Qu.:  7.750   A14    :  1   C:24
##  Median :  8.081   C124   :  1   Q:30
##  Mean   : 22.334   C52    :  1   S:54
##  3rd Qu.: 25.467   C95    :  1
##  Max.   :221.779   D45    :  1
##                    (Other):  0
```

(no cabin for these people)

impostate the mean for these values

```
training[is.na(training$Age),]$Age<-mean(training$Age, na.rm = TRUE)
```

```
summary(training$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   22.00   29.69   29.69   35.00   71.00
```

Null in "embarked"?

```
training[training$Embarked=="",]
```

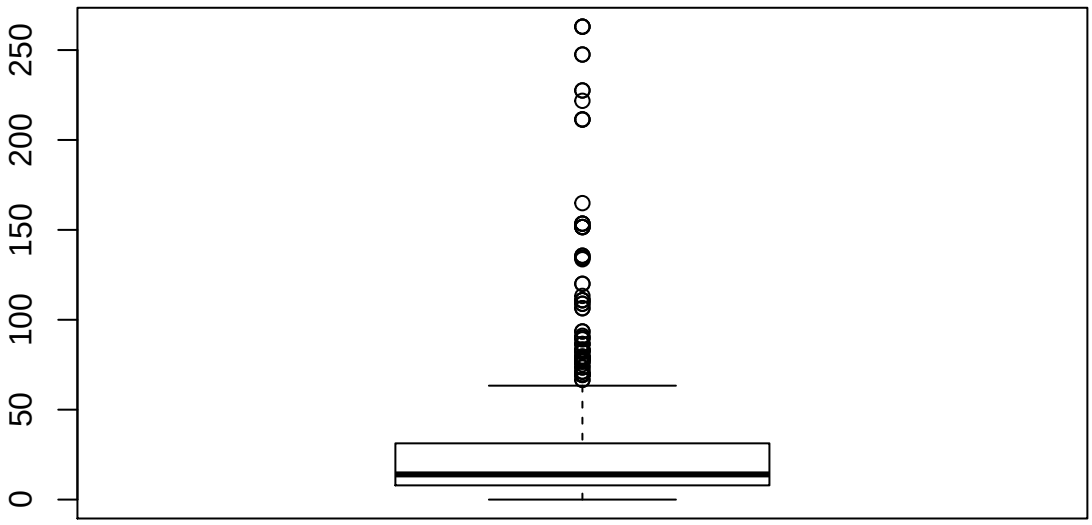
```
##  [1] PassengerId Survived    Pclass      Name        Sex
##  [6] Age         SibSp       Parch       Ticket      Fare
## [11] Cabin       Embarked
## <0 rows> (or 0-length row.names)
```

action -> remove these values

```
train<-training[training$Embarked!="",]
```

possible outlier in Fare

```
boxplot(training[training$Fare<500,]$Fare)
```



Eliminamos ese punto aunque la distrubicion sigue siendo muy sesgada

```
train<-training[training$Fare<500,]
```

```
summary(train)
```

```
##   PassengerId    Survived Pclass
##  Min.   :  1.0   0:345    1:142
##  1st Qu.:213.5   1:206    2:108
##  Median :450.0            3:301
##  Mean   :443.7
##  3rd Qu.:667.5
##  Max.   :890.0
##
##                                                 Name         Sex
##  Abbing, Mr. Anthony                              :  1   female:188
##  Abbott, Mrs. Stanton (Rosa Hunt)                 :  1   male  :363
##  Abelson, Mrs. Samuel (Hannah Wizosky)            :  1
##  Adams, Mr. John                                  :  1
##  Ahlin, Mrs. Johan (Johanna Persdotter Larsson):  1
##  Aks, Mrs. Sam (Leah Rosen)                       :  1
##  (Other)                                          :545
##       Age            SibSp           Parch             Ticket
##  Min.   : 0.42   Min.   :0.0000   Min.   :0.0000   1601    :  6
##  1st Qu.:22.00   1st Qu.:0.0000   1st Qu.:0.0000   CA. 2343:  5
##  Median :29.69   Median :0.0000   Median :0.0000   113781  :  4
##  Mean   :29.67   Mean   :0.4973   Mean   :0.3249   3101295 :  4
##  3rd Qu.:35.00   3rd Qu.:1.0000   3rd Qu.:0.0000   LINE    :  4
##  Max.   :71.00   Max.   :8.0000   Max.   :5.0000   110152  :  3
##                                                    (Other) :525
##       Fare                Cabin     Embarked
##  Min.   :  0.000             :425    :  0
##  1st Qu.:  7.896   C22 C26   :  3   C:113
##  Median : 14.000   C23 C25 C27:  3   Q: 42
##  Mean   : 31.304   F33       :  3   S:396
##  3rd Qu.: 31.275   B18       :  2
##  Max.   :263.000   B20       :  2
##                    (Other)   :113
```

Prepared data:

We discard artifial features and normalize the numerical features

```
normalizationData <- list(
  ageMean=mean(training$Age),
  ageSD=sd(training$Age),
  SibSpMean=mean(training$SibSp),
  SibSpSD=sd(training$SibSp),
  ParchMean=mean(training$Parch),
  ParchSD=sd(training$Parch),
  FareMean=mean(training$Fare),
  FareSD=sd(training$Fare)
)

normalizationData
```

```
## $ageMean
## [1] 29.69328
##
## $ageSD
## [1] 12.81782
##
## $SibSpMean
## [1] 0.4954792
##
## $SibSpSD
## [1] 1.071906
##
## $ParchMean
## [1] 0.323689
##
## $ParchSD
## [1] 0.7059012
##
## $FareMean
## [1] 33.04377
##
## $FareSD
## [1] 51.05901
```

```
training<-data.frame(
Pclass=training$Pclass,
Sex =training$Sex,
AgeNorm=(training$Age-normalizationData$ageMean)/normalizationData$ageSD,
SibSpNorm=(training$SibSp -normalizationData$SibSpMean)/normalizationData$SibSpSD,
ParchNorm=(training$Parch -normalizationData$ParchMean)/normalizationData$ParchSD,
FareNorm=(training$Fare -normalizationData$FareMean)/normalizationData$FareSD,
Embarked=training$Embarked,
Survived=training$Survived
  )
```

Check data cleaned

```
summary(training)
```

```
##  Pclass      Sex         AgeNorm          SibSpNorm
##  1:144   female:189   Min.   :-2.2838   Min.   :-0.4622
##  2:108   male  :364   1st Qu.:-0.6002   1st Qu.:-0.4622
##  3:301                Median : 0.0000   Median :-0.4622
##                       Mean   : 0.0000   Mean   : 0.0000
##                       3rd Qu.: 0.4140   3rd Qu.: 0.4707
##                       Max.   : 3.2226   Max.   : 7.0011
##    ParchNorm          FareNorm        Embarked Survived
##  Min.   :-0.4585   Min.   :-0.64717    :  0    0:345
##  1st Qu.:-0.4585   1st Qu.:-0.49253   C:115    1:208
##  Median :-0.4585   Median :-0.36514   Q: 42
##  Mean   : 0.0000   Mean   : 0.00000   S:396
##  3rd Qu.:-0.4585   3rd Qu.:-0.03244
##  Max.   : 6.6246   Max.   : 9.38689
```

save the clean data

```
save(file = "../processed/training.dat", training)
save(file = "../processed/testing.dat", testing)
save(file = "../processed/validation.dat", validation)
```