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

```
train<- read.csv("../data/train.csv")

str(train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 416 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 525 596 662 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

Change type for cathegorical vars

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

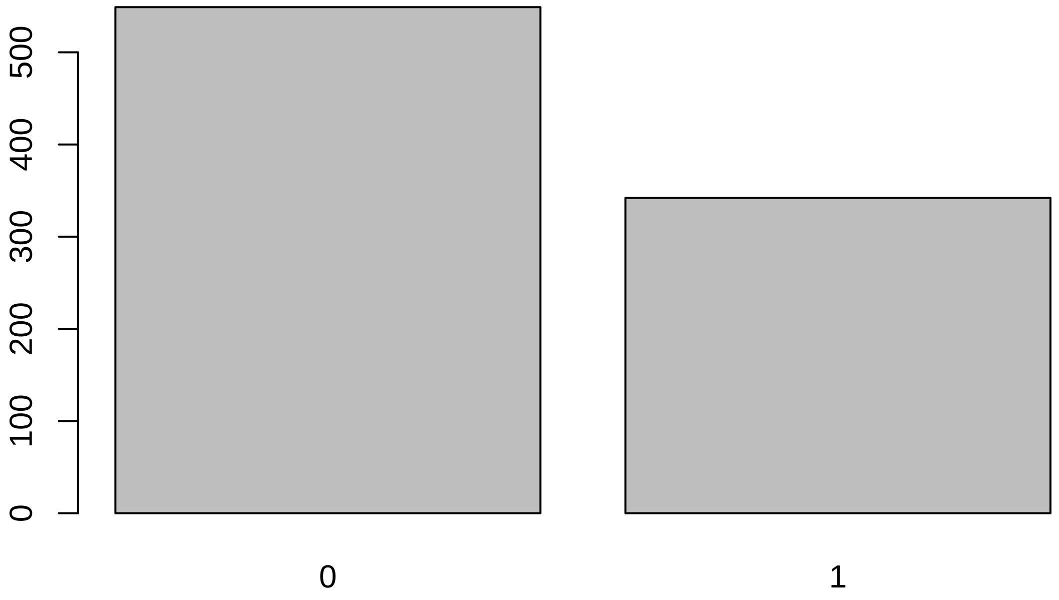
```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 416 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 525 596 662 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
summary(train)
```

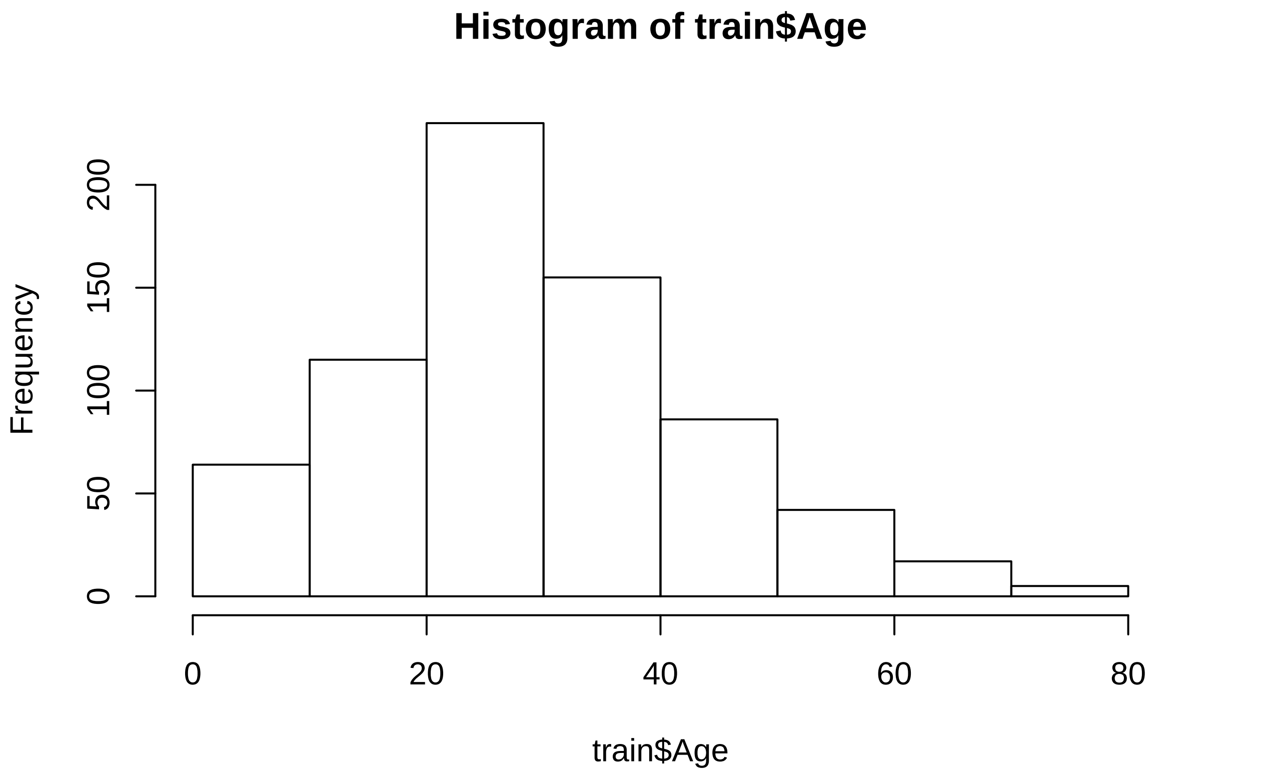
```
##   PassengerId    Survived Pclass
##  Min.   :  1.0   0:549    1:216
##  1st Qu.:223.5   1:342    2:184
##  Median :446.0            3:491
##  Mean   :446.0
##  3rd Qu.:668.5
##  Max.   :891.0
##
##                                     Name         Sex           Age
##  Abbing, Mr. Anthony                  :  1   female:314   Min.   : 0.42
##  Abbott, Mr. Rossmore Edward          :  1   male  :577   1st Qu.:20.12
##  Abbott, Mrs. Stanton (Rosa Hunt)     :  1                Median :28.00
##  Abelson, Mr. Samuel                  :  1                Mean   :29.70
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00
##  Adahl, Mr. Mauritz Nils Martin       :  1                Max.   :80.00
##  (Other)                              :885                NA's   :177
##      SibSp           Parch             Ticket          Fare
##  Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91
##  Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
##  Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
##                                   (Other) :852
##          Cabin     Embarked
##             :687    :  2
##  B96 B98    :  4   C:168
##  C23 C25 C27:  4   Q: 77
##  G6         :  4   S:644
##  C22 C26    :  3
##  D          :  3
##  (Other)    :186
```

Plots

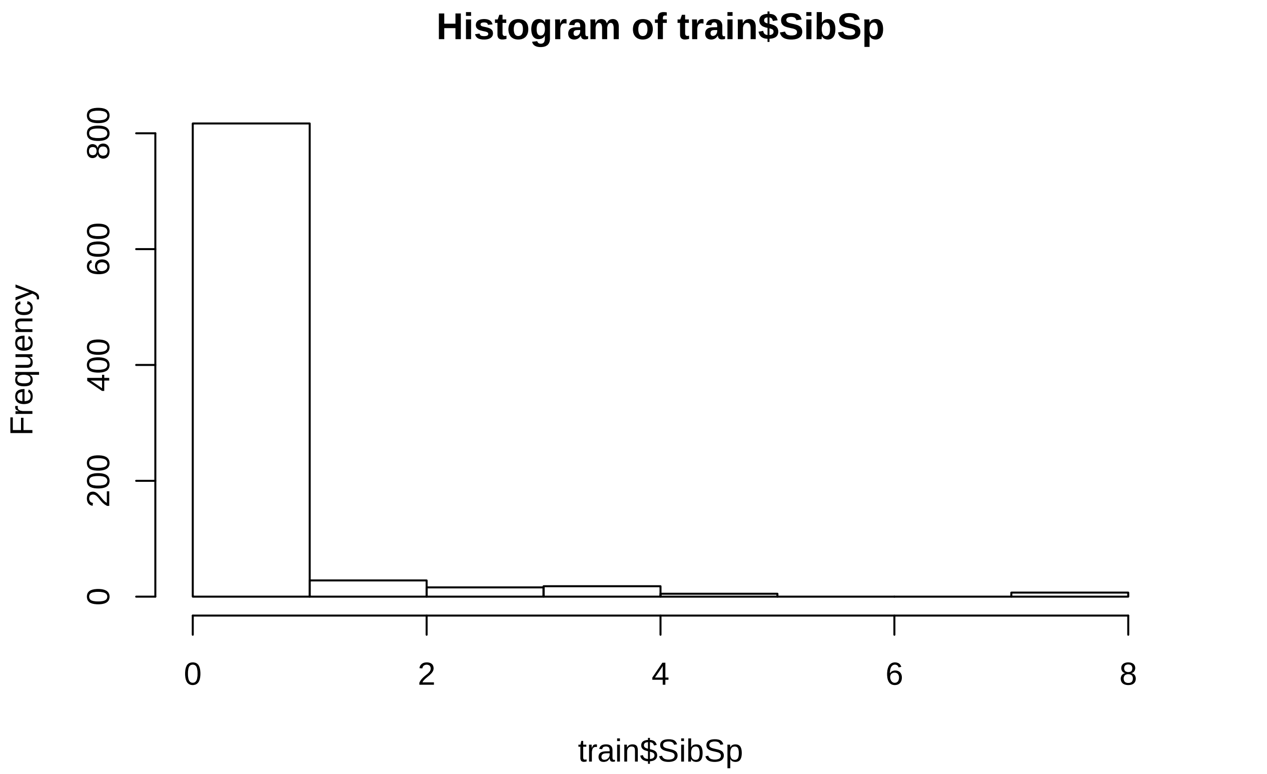
```r
plot(train$Survived)
```



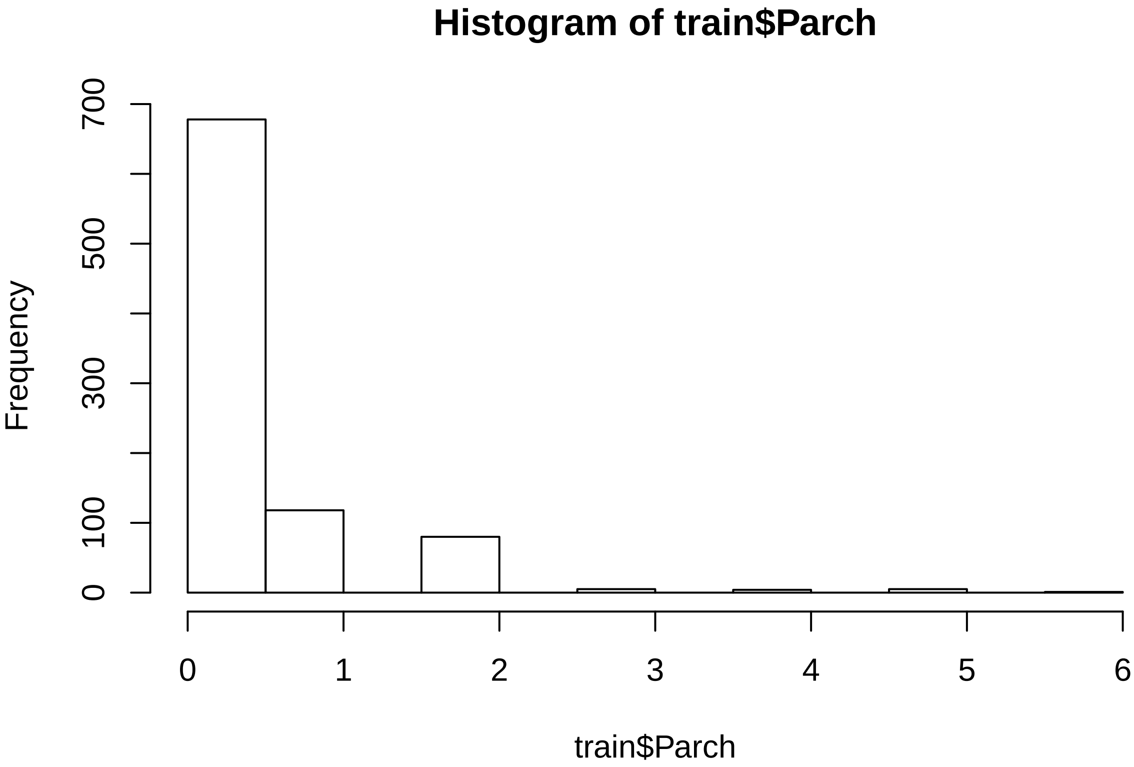
```r
hist(train$Age)
```



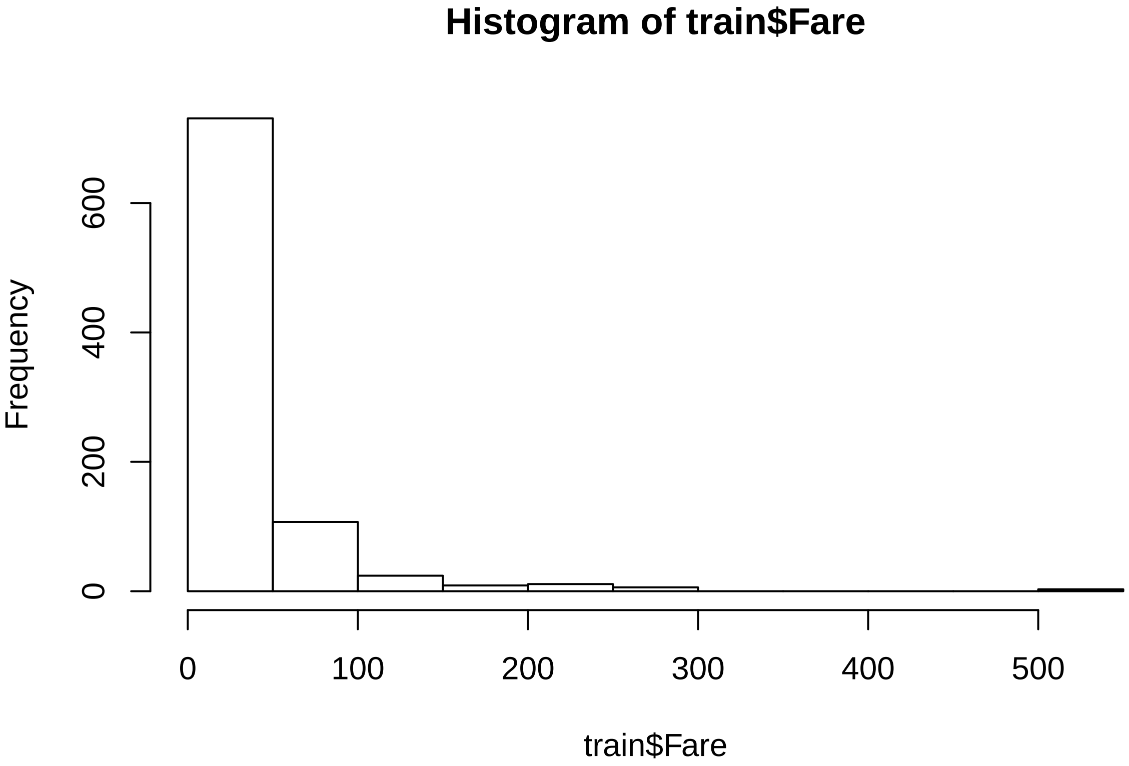
```r
hist(train$SibSp)
```



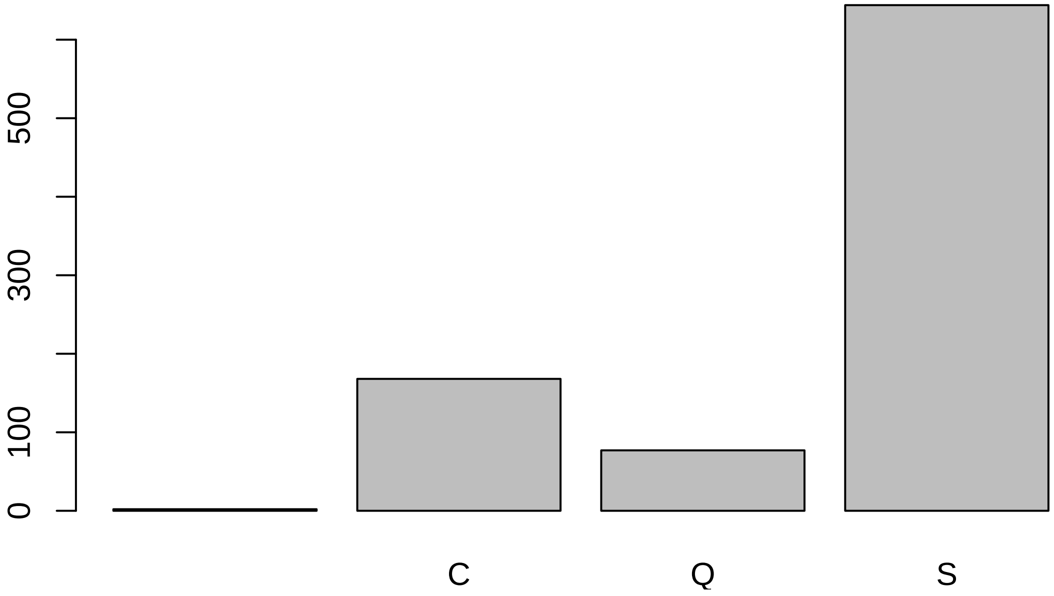
```r
hist(train$Parch)
```



```r
hist(train$Fare)
```



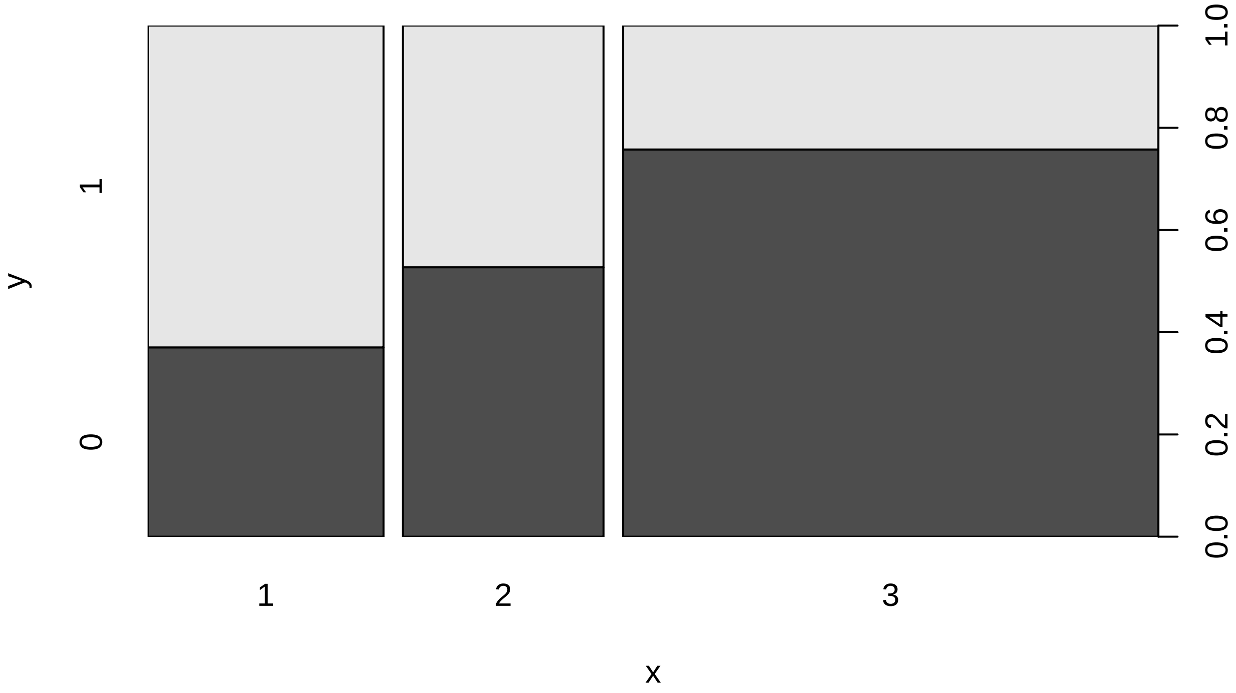
```r
plot(train$Embarked)
```



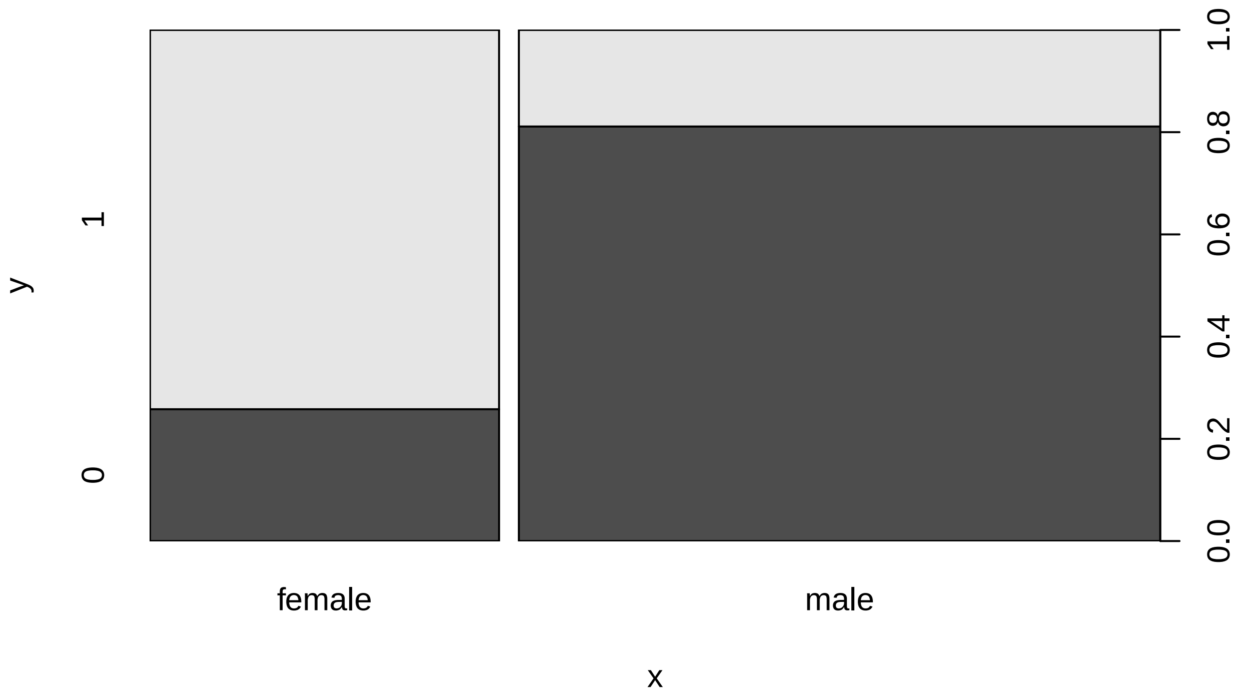
Bidimiensional distributions

Survival vs others

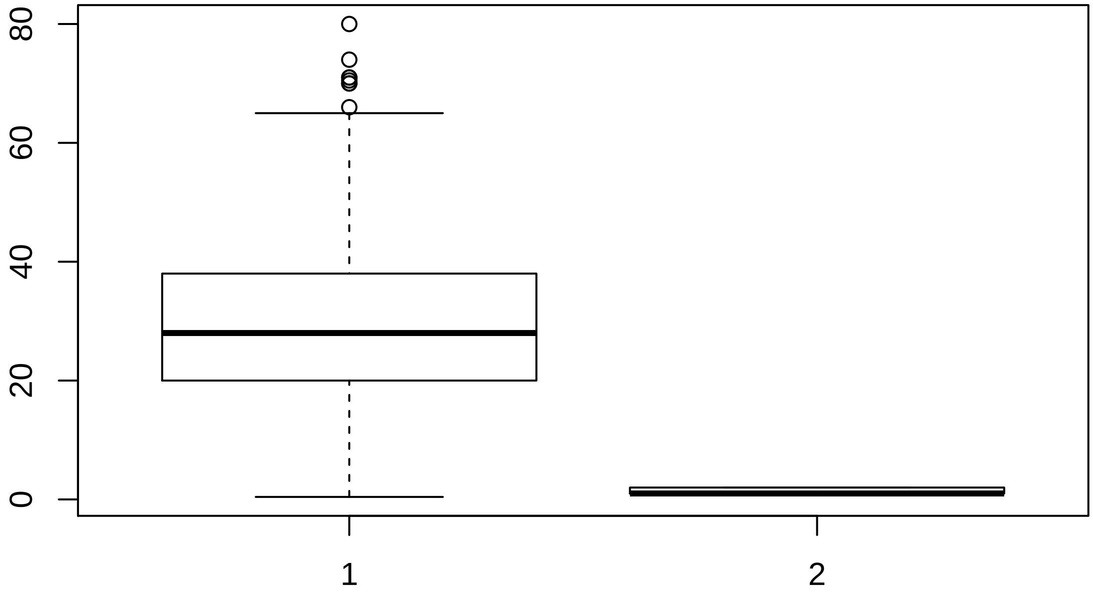
```r
with(train,plot(Pclass,Survived))
```



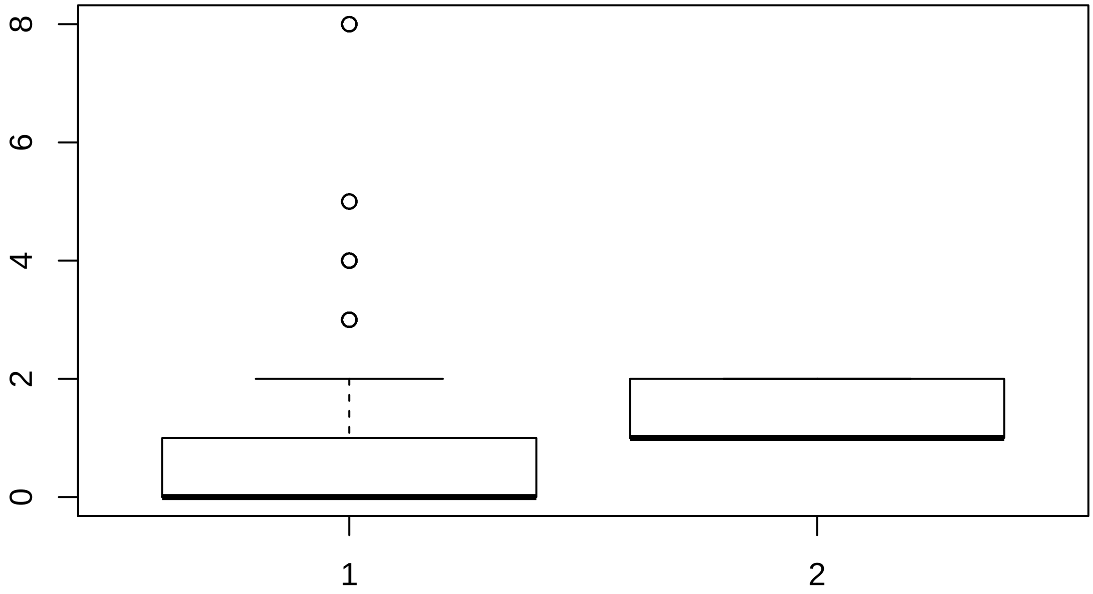
```r
with(train,plot(Sex,Survived))
```



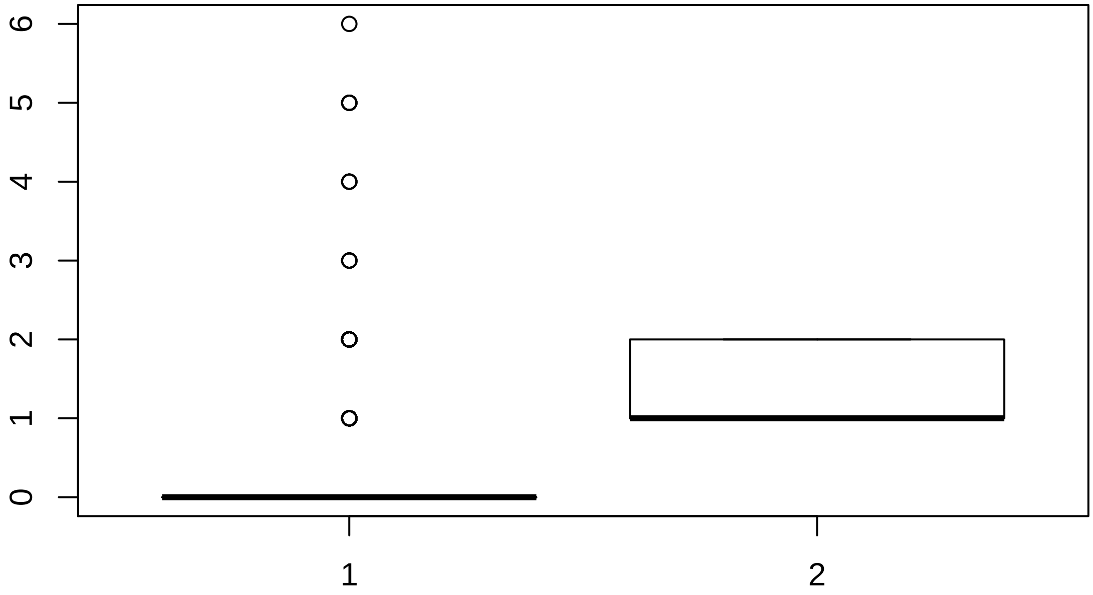
```
with(train,boxplot(Age,Survived))
```



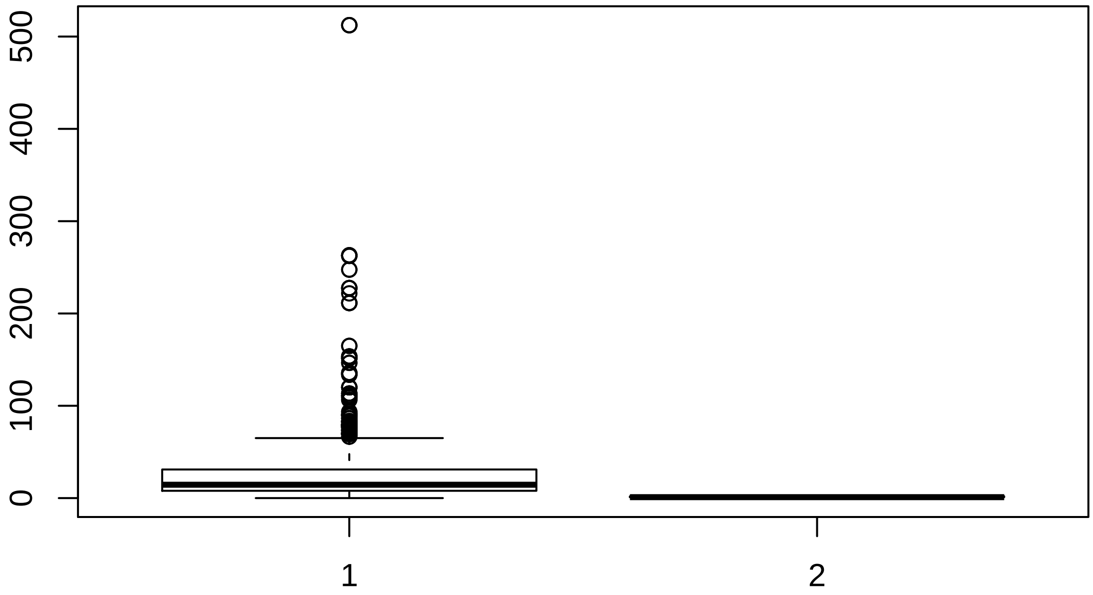
```
with(train,boxplot(SibSp,Survived))
```



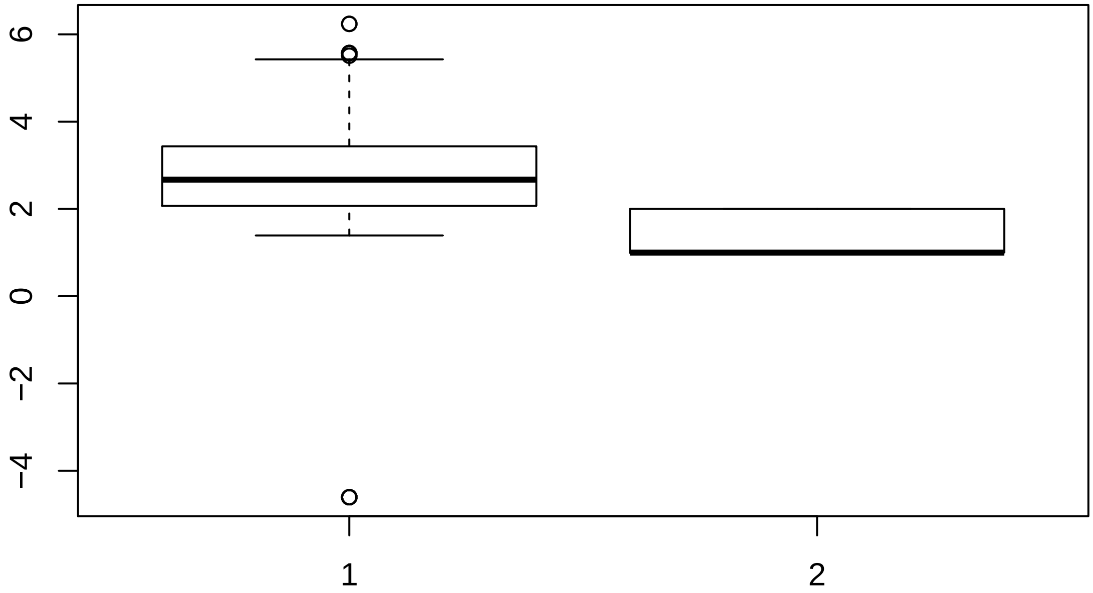
```r
with(train,boxplot(Parch,Survived))
```



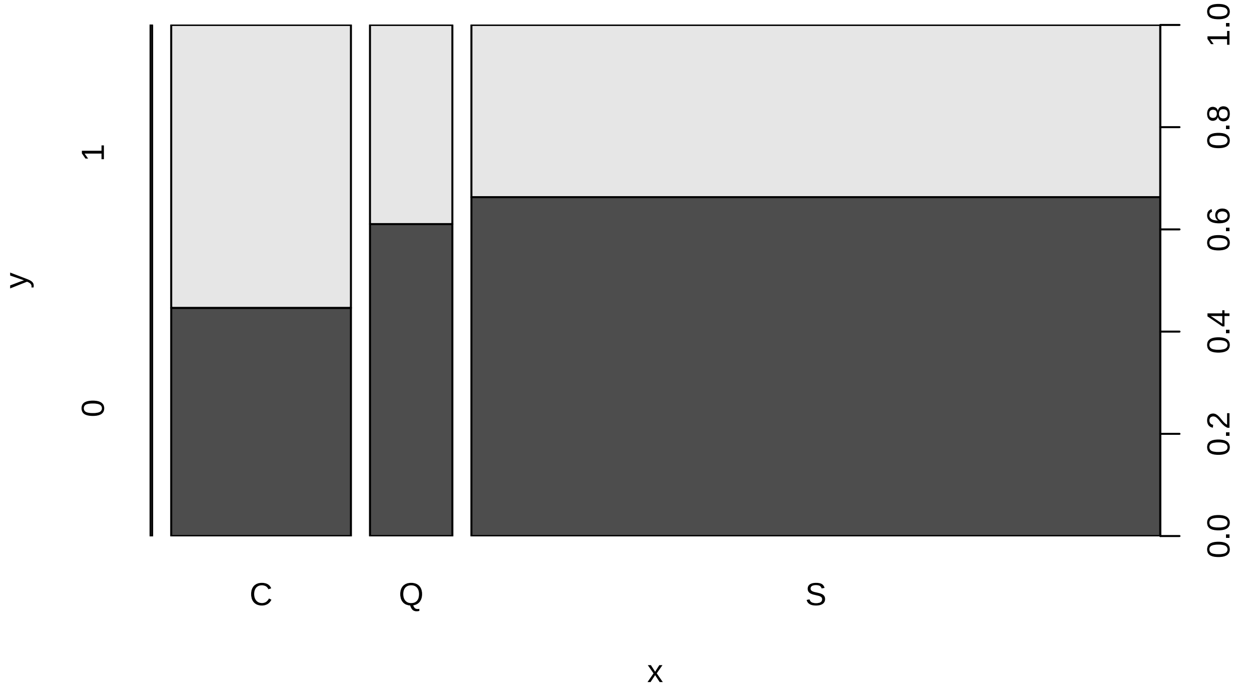
```r
with(train,boxplot(Fare,Survived))
```



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



```r
with(train,plot(Embarked,Survived))
```



Check for data anomalies

No age?

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

```
##   PassengerId    Survived Pclass
##  Min.   :  6.0   0:125    1: 30 
##  1st Qu.:230.0   1: 52    2: 11 
##  Median :452.0            3:136 
##  Mean   :435.6                  
##  3rd Qu.:634.0                  
##  Max.   :889.0                  
##                                                                    
##                                                Name         Sex     
##  Baumann, Mr. John D                             :  1   female: 53  
##  Boulos, Mr. Hanna                               :  1   male  :124  
##  Boulos, Mrs. Joseph (Sultana)                   :  1               
##  Bourke, Miss. Mary                              :  1               
##  Bradley, Mr. George ("George Arthur Brayton"):  1               
##  Brewe, Dr. Arthur Jackson                       :  1               
##  (Other)                                         :171               
##       Age          SibSp           Parch             Ticket   
##  Min.   : NA   Min.   :0.000   Min.   :0.0000   CA. 2343:  7  
##  1st Qu.: NA   1st Qu.:0.000   1st Qu.:0.0000   4133    :  4  
##  Median : NA   Median :0.000   Median :0.0000   1601    :  3  
##  Mean   :NaN   Mean   :0.565   Mean   :0.1808   239853  :  3  
##  3rd Qu.: NA   3rd Qu.:0.000   3rd Qu.:0.0000   371110  :  3  
##  Max.   : NA   Max.   :8.000   Max.   :2.0000   2661    :  2  
##  NA's   :177                                    (Other) :155  
##       Fare            Cabin     Embarked
##  Min.   :  0.00          :158     : 0
##  1st Qu.:  7.75   A14     :  1   C:38
##  Median :  8.05   A19     :  1   Q:49
##  Mean   : 22.16   A32     :  1   S:90
##  3rd Qu.: 24.15   B102    :  1
##  Max.   :227.53   B78     :  1
##                   (Other): 14
```

(no cabin for these people)

impostate the mean for these values

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

```
summary(train$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   22.00   29.70   29.70   35.00   80.00
```

Null in "embarked"?

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

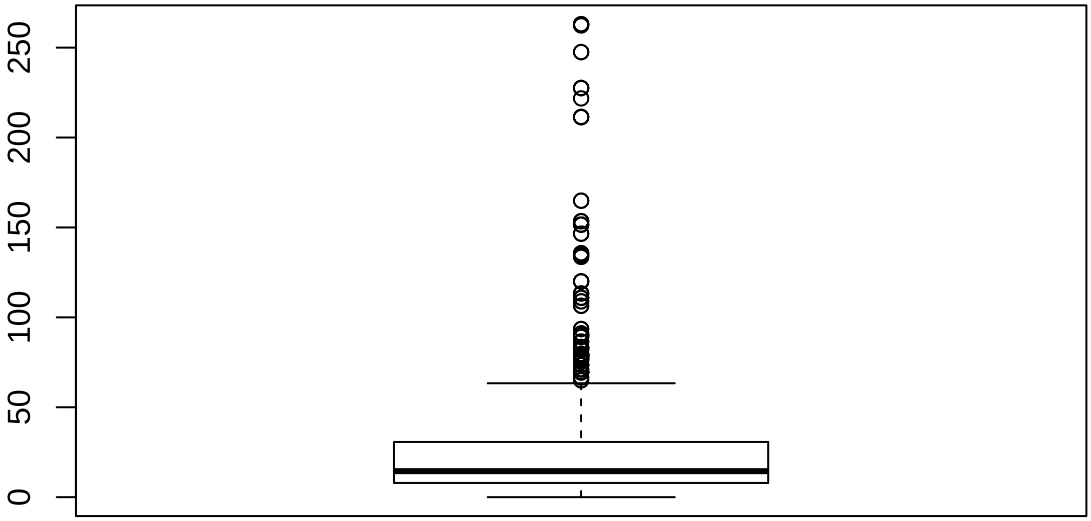
```
##     PassengerId Survived Pclass                                      Name
## 62           62        1      1                       Icard, Miss. Amelie
## 830         830        1      1 Stone, Mrs. George Nelson (Martha Evelyn)
##        Sex Age SibSp Parch Ticket Fare Cabin Embarked
## 62  female  38     0     0 113572   80   B28
## 830 female  62     0     0 113572   80   B28
```

action -> remove these values

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

possible outlier in Fare

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


0
50
100
150
200
250