D3

2025-05-22

```
setwd("/dades/eric.diez/6Q/ADEI")
dd <- read.csv("adult_def.csv", stringsAsFactors = TRUE);
names(dd)
```

```
##  [1] "age"            "workclass"      "fnlwgt"         "edu_num"
##  [5] "marital"        "occupation"     "relationship"   "race"
##  [9] "sex"            "cap_gain"       "cap_loss"       "hours_week"
## [13] "native_country" "income"         "income_integer"
```

```
dim(dd)
```

```
## [1] 48842    15
```

```
summary(dd)
```

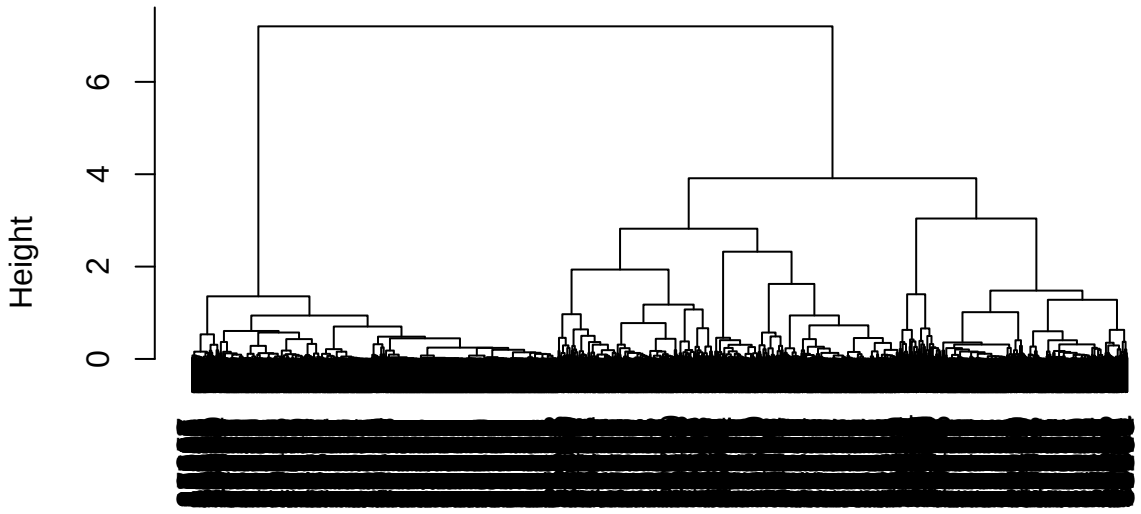
```
##       age        workclass         fnlwgt           edu_num
##  Min.   :17.00   Fed  : 1432   Min.   :  12285   Min.   : 1.00
##  1st Qu.:28.00   Loc  : 3136   1st Qu.: 117550   1st Qu.: 9.00
##  Median :37.00   NoPay:   31   Median : 178144   Median :10.00
##  Mean   :38.64   Priv :36705   Mean   : 189664   Mean   :10.08
##  3rd Qu.:48.00   SelfI: 1695   3rd Qu.: 237642   3rd Qu.:12.00
##  Max.   :90.00   SelfN: 3862   Max.   :1490400   Max.   :16.00
##                  State: 1981
##     marital          occupation            relationship
##  Div    : 6633   Prof     : 8981   Husband       :19716
##  Married:22416   CraftRep : 6112   Not-in-family :12583
##  NevMarr:16117   ExecMan  : 6086   Other-relative: 1506
##  Sep    : 2158   AdminCler: 5611   Own-child     : 7581
##  Widow  : 1518   Sales    : 5504   Unmarried     : 5125
##                  Other    : 4923   Wife          : 2331
##                  (Other)  :11625
##                  race           sex           cap_gain        cap_loss
##  Amer-Indian-Eskimo:  470   Female:16192   Min.   :    0   Min.   :   0.0
##  Asian-Pac-Islander: 1519   Male  :32650   1st Qu.:    0   1st Qu.:   0.0
##  Black             : 4685                  Median :    0   Median :   0.0
##  Other             :  406                  Mean   : 1079   Mean   :  87.5
##  White             :41762                  3rd Qu.:    0   3rd Qu.:   0.0
##                                            Max.   :99999   Max.   :4356.0
##
##    hours_week    native_country   income      income_integer
##  Min.   : 1.00   Other: 5010    <=50K:37155   Min.   :27850
##  1st Qu.:40.00   USA  :43832    >50K :11687   1st Qu.:40820
##  Median :40.00                                Median :44870
##  Mean   :40.42                                Mean   :45684
##  3rd Qu.:45.00                                3rd Qu.:49835
##  Max.   :99.00                                Max.   :80040
##
```

```r
set.seed(123)
samp<-sample(48842,5000)

dd<-dd[samp,]
#set a list of numerical variables
names(dd)
```

```
##  [1] "age"            "workclass"      "fnlwgt"         "edu_num"
##  [5] "marital"        "occupation"     "relationship"   "race"
##  [9] "sex"            "cap_gain"       "cap_loss"       "hours_week"
## [13] "native_country" "income"         "income_integer"
```

```r
attach(dd)
```

```r
#euclidean distance si totes son numeriques
dcon<-data.frame (age,edu_num,cap_gain,cap_loss,hours_week,income_integer)

d  <- dist(dcon[1:6,])

#move to Gower mixed distance to deal
#simoultaneously with numerical and qualitative data

library(cluster)

#dissimilarity matrix
#do not include in actives the identifier variables nor the potential response variable

actives <- c("age", "workclass", "edu_num",
             "marital", "occupation", "relationship", "race", "sex",
             "cap_gain", "cap_loss", "hours_week", "native_country")
actives <- c(1:2,4:13)
dissimMatrix <- daisy(dd[,actives], metric = "gower", stand=TRUE)
distMatrix <- as.dist(dissimMatrix^2)

h1 <- hclust(distMatrix,method="ward.D2")  # NOTICE THE COST
#versions noves "ward.D" i abans de plot: par(mar=rep(2,4)) si se quejara de los margenes del plot

plot(h1)
```



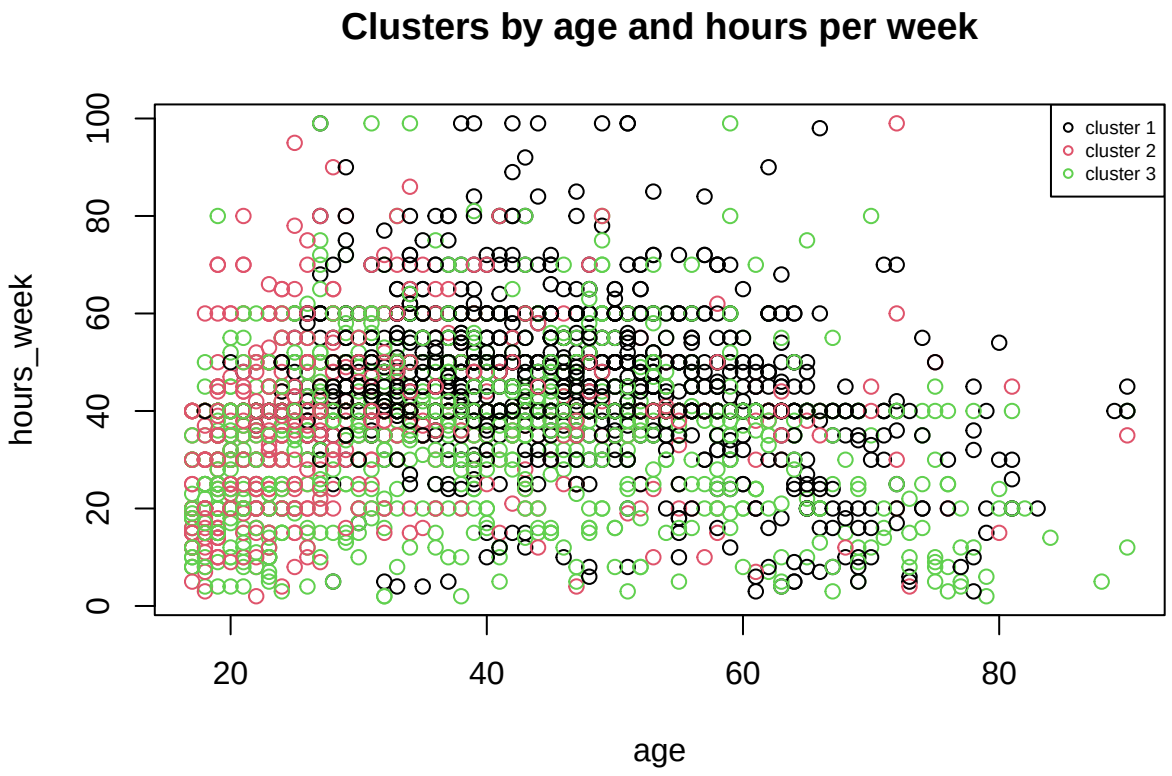

```
k<-3

c2 <- cutree(h1, k=k)
```

Añadir cluster al

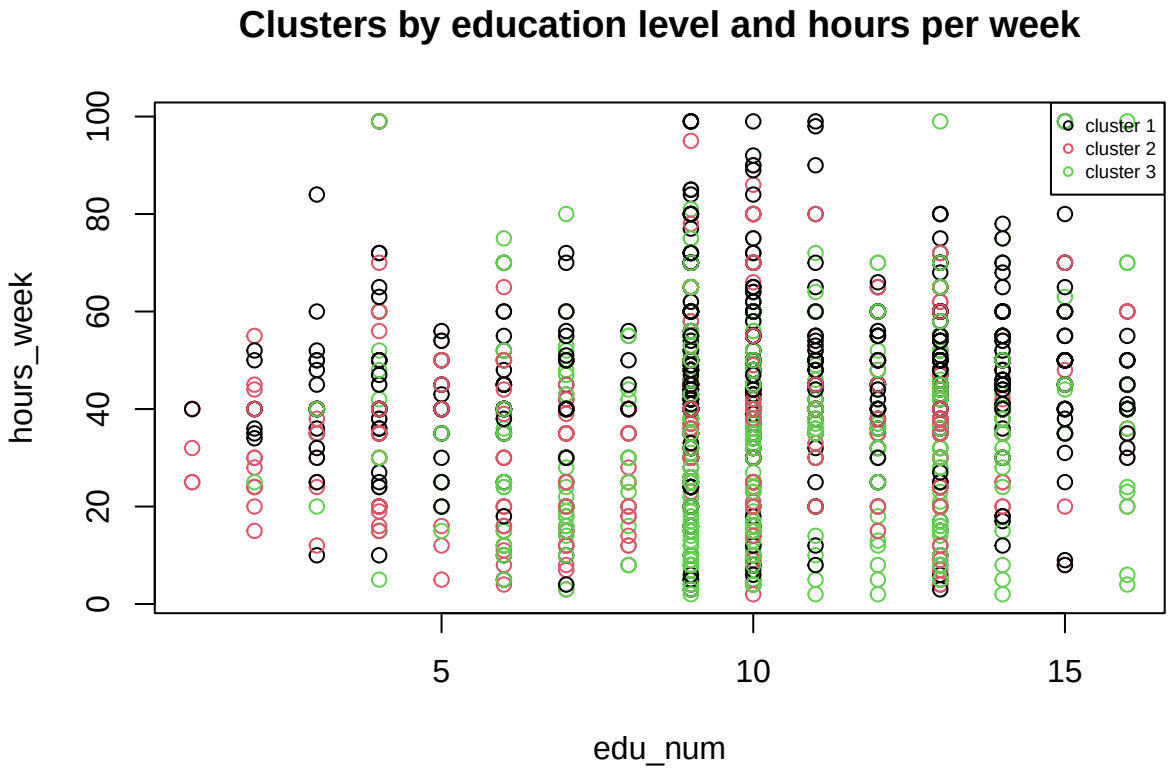
```
dd$cluster <- as.factor(c2)

# Análisis descriptivo de los clusters
table(dd$cluster)
```

```
##
##    1    2    3
## 1957 1221 1822
```

```
# LETS SEE THE PARTITION VISUALLY

c1<-c2
# Basic scatter plots for three key numeric variables (modified version)
plot(age, hours_week, col=c1, main="Clusters by age and hours per week")
legend("topright", paste("cluster", 1:k), pch=1, col=1:k, cex=0.6)
```

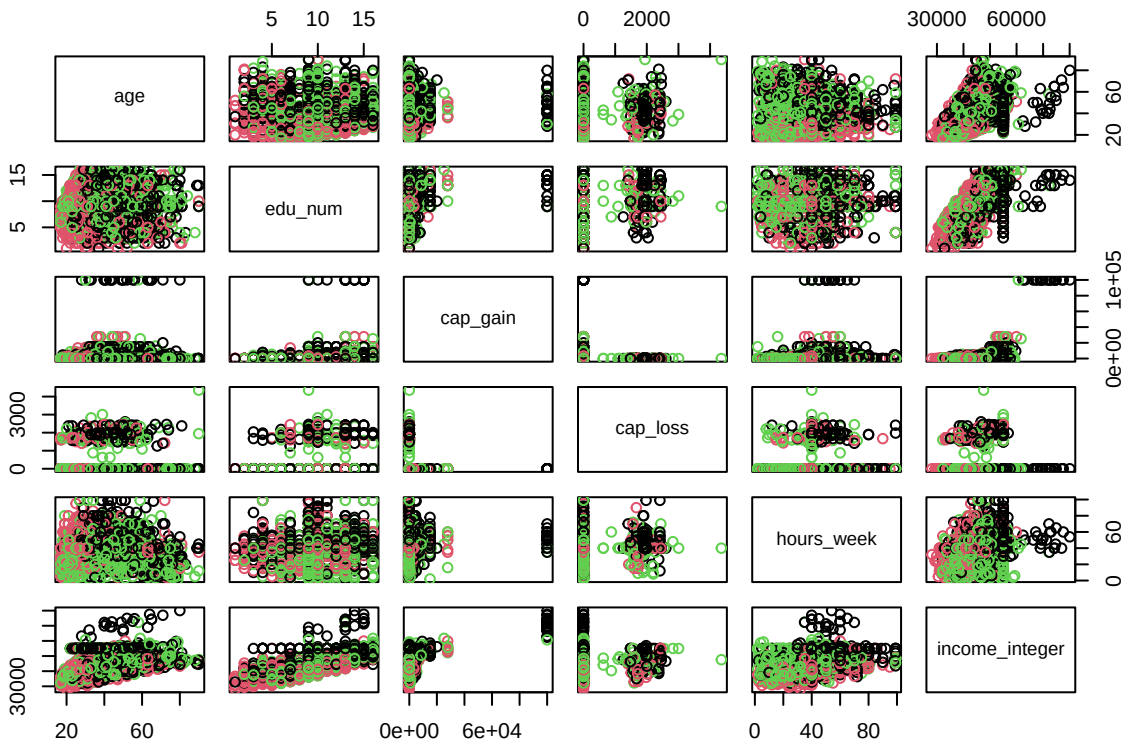



```
plot(edu_num, hours_week, col=c1, main="Clusters by education level and hours per week")
legend("topright", paste("cluster", 1:k), pch=1, col=1:k, cex=0.6)
```




```
plot(age, edu_num, col=c1, main="Clusters by age and education level")
legend("topright", paste("cluster", 1:k), pch=1, col=1:k, cex=0.6)
```


```
pairs(dcon[,1:6], col=c1)
```

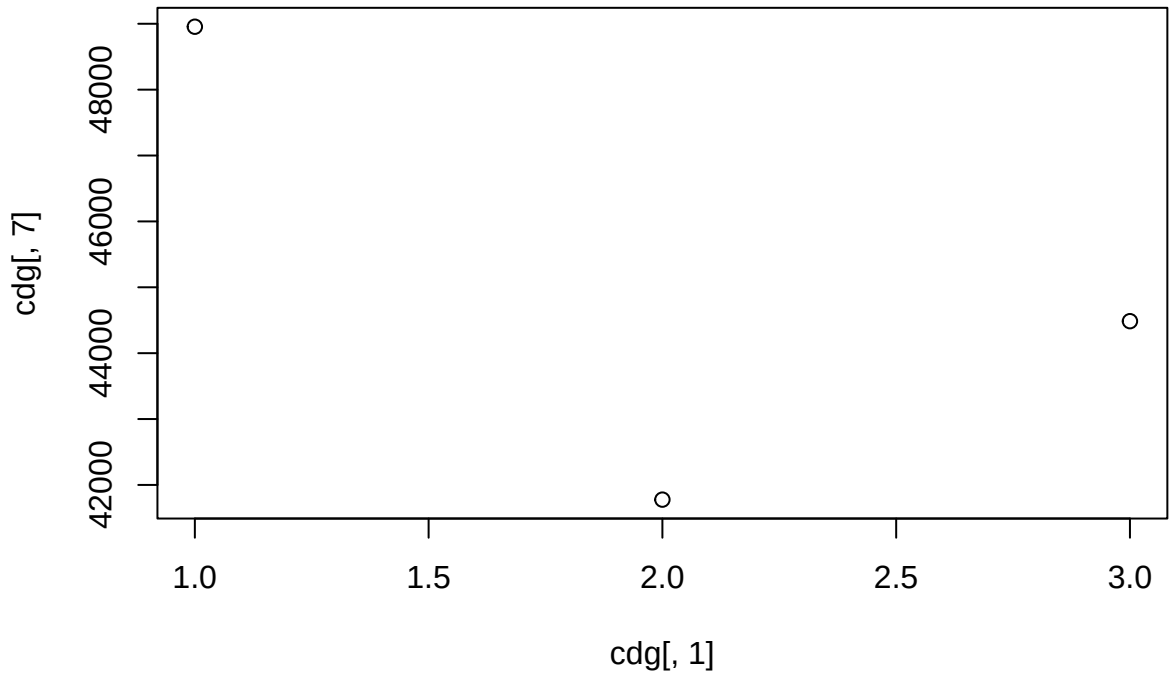


LETS SEE THE QUALITY OF THE HIERARCHICAL PARTITION

```
cdg <- aggregate(as.data.frame(dcon), list(c1), mean)
cdg
```

```
##   Group.1      age   edu_num  cap_gain  cap_loss hours_week income_integer
## 1       1 44.17527 10.273889 1922.9918 138.15330   43.99540       48954.94
## 2       2 30.63718  9.594595  297.0614  54.32514   38.30794       41777.02
## 3       3 37.61690 10.135565  544.7629  65.65642   37.42591       44485.70
```

```
plot(cdg[,1], cdg[,7])
```




```
#Profiling plots

#Calcula els valor test de la variable Xnum per totes les modalitats del factor P
ValorTestXnum <- function(Xnum,P){
  #freq dis of fac
  nk <- as.vector(table(P));
  n <- sum(nk);
  #mitjanes x grups
  xk <- tapply(Xnum,P,mean);
  #valors test
  txk <- (xk-mean(Xnum))/(sd(Xnum)*sqrt((n-nk)/(n*nk)));
  #p-values
  pxk <- pt(txk,n-1,lower.tail=F);
  for(c in 1:length(levels(as.factor(P)))){if (pxk[c]>0.5){pxk[c]<-1-pxk[c]}}
  return (pxk)
}
```

```r
ValorTestXquali <- function(P,Xquali){
  taula <- table(P,Xquali);
  n <- sum(taula);
  pk <- apply(taula,1,sum)/n;
  pj <- apply(taula,2,sum)/n;
  pf <- taula/(n*pk);
  pjm <- matrix(data=pj,nrow=dim(pf)[1],ncol=dim(pf)[2], byrow=TRUE);
  dpf <- pf - pjm;
  dvt <- sqrt(((1-pk)/(n*pk))%*%t(pj*(1-pj)));
  #i hi ha divisions iguals a 0 dona NA i no funciona
  zkj <- dpf
  zkj[dpf!=0]<-dpf[dpf!=0]/dvt[dpf!=0];
  pzkj <- pnorm(zkj,lower.tail=F);
  for(c in 1:length(levels(as.factor(P)))){for (s in 1:length(levels(Xquali))){if (pzkj[c,s]> 0.5){pzkj
  return (list(rowpf=pf,vtest=zkj,pval=pzkj))
}
```

```r
#source("file")
#dades contain the dataset
dades <- dd[, setdiff(names(dd),
                      c("fnlwgt", "income", "cluster"))]
#dades<-dd[filtro,]
#dades<-df
K<-dim(dades)[2]
par(ask=TRUE)


#P must contain the class variable
#P<-dd[,3]
P<-c2
#P<-dd[,18]
nameP<-"classe"
#P<-df[,33]

nc<-length(levels(factor(P)))
nc
```

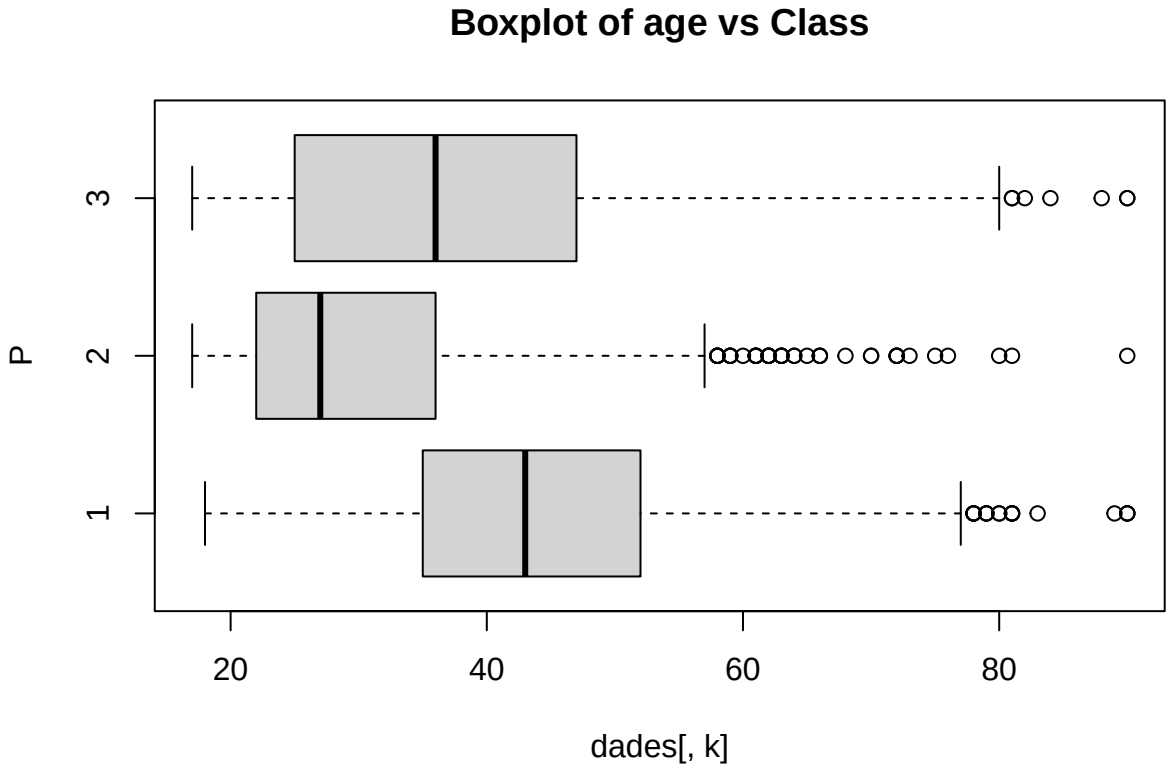
```
## [1] 3
```

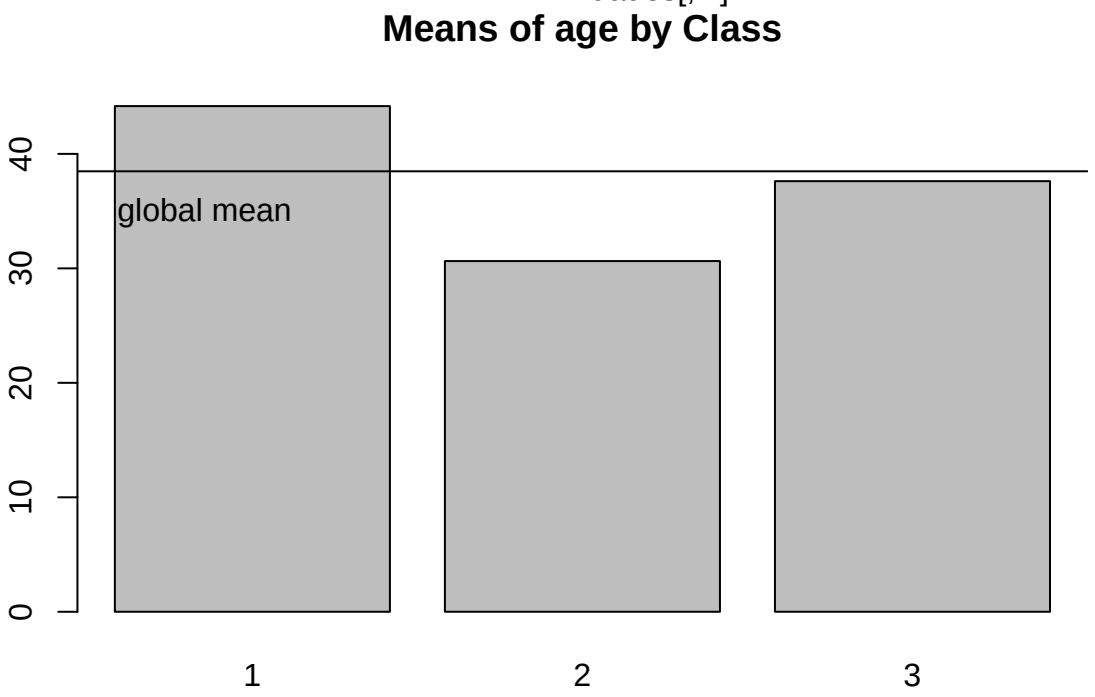
```r
pvalk <- matrix(data=0,nrow=nc,ncol=K, dimnames=list(levels(P),names(dades)))
nameP<-"Class"
n<-dim(dades)[1]

for(k in 1:K){
  if (is.numeric(dades[,k])){
    print(paste("Anàlisi per classes de la Variable:", names(dades)[k]))

    boxplot(dades[,k]~P, main=paste("Boxplot of", names(dades)[k], "vs", nameP ), horizontal=TRUE)

    barplot(tapply(dades[[k]], P, mean),main=paste("Means of", names(dades)[k], "by", nameP ))
    abline(h=mean(dades[[k]]))
    legend(0,mean(dades[[k]]),"global mean",bty="n")
    print("Estadístics per groups:")
    for(s in levels(as.factor(P))) {print(summary(dades[P==s,k]))}
```

```
    o<-oneway.test(dades[,k]~P)
    print(paste("p-valueANOVA:", o$p.value))
    kw<-kruskal.test(dades[,k]~P)
    print(paste("p-value Kruskal-Wallis:", kw$p.value))
    pvalk[,k]<-ValorTestXnum(dades[,k], P)
    print("p-values ValorsTest: ")
    print(pvalk[,k])
  }else{
    if(class(dd[,k])=="Date"){
      print(summary(dd[,k]))
      print(sd(dd[,k]))
      #decide breaks: weeks, months, quarters...
      hist(dd[,k],breaks="weeks")
    }else{
      #qualitatives
      print(paste("Variable", names(dades)[k]))
      table<-table(P,dades[,k])
      #   print("Cross-table")
      #   print(table)
      rowperc<-prop.table(table,1)

      colperc<-prop.table(table,2)
      #  print("Distribucions condicionades a files")
      # print(rowperc)

      #ojo porque si la variable es true o false la identifica amb el tipus Logical i
      #aquest no te levels, por tanto, coertion preventiva

      dades[,k]<-as.factor(dades[,k])


      marg <- table(as.factor(P))/n
      print(append("Categories=",levels(as.factor(dades[,k]))))

      #from next plots, select one of them according to your practical case

      #with legend
      plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of classes by",names(dades)[k]))
      paleta<-rainbow(length(levels(dades[,k])))
      for(c in 1:length(levels(dades[,k]))){lines(colperc[,c],col=paleta[c]) }
      legend("topright", levels(dades[,k]), col=paleta, lty=2, cex=0.6)

      #condicionades a classes
      print(append("Categories=",levels(dades[,k])))

      #with legend
      plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of classes by",names(dades)[k]))
      paleta<-rainbow(length(levels(dades[,k])))
      for(c in 1:length(levels(dades[,k]))){lines(rowperc[,c],col=paleta[c]) }
      legend("topright", levels(dades[,k]), col=paleta, lty=2, cex=0.6)

      #amb variable en eix d'abcisses
      marg <-table(dades[,k])/n
```

```
      print(append("Categories=",levels(dades[,k])))
      paleta<-rainbow(length(levels(as.factor(P))))

      #with legend
      plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of classes by",names(dades)[k]), las=3)
      for(c in 1:length(levels(as.factor(P)))){lines(rowperc[c,],col=paleta[c])}
      legend("topright", levels(as.factor(P)), col=paleta, lty=2, cex=0.6)

      #condicionades a columna

      #with legend
      plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of classes by",names(dades)[k]), las=3)
      for(c in 1:length(levels(as.factor(P)))){lines(colperc[c,],col=paleta[c])}
      legend("topright", levels(as.factor(P)), col=paleta, lty=2, cex=0.6)

      table<-table(dades[,k],P)
      print("Cross Table:")
      print(table)
      print("Distribucions condicionades a columnes:")
      print(colperc)

      #diagrames de barres apilades

      paleta<-rainbow(length(levels(dades[,k])))

      barplot(table(dades[,k], as.factor(P)), beside=FALSE,col=paleta )
      legend("topright",levels(as.factor(dades[,k])),pch=1,cex=0.5, col=paleta)

      #diagrames de barres adosades

      barplot(table(dades[,k], as.factor(P)), beside=TRUE,col=paleta)
      legend("topright",levels(as.factor(dades[,k])),pch=1,cex=0.5, col=paleta)

      print("Test Chi quadrat: ")
      print(chisq.test(dades[,k], as.factor(P)))

      print("valorsTest:")
      print( ValorTestXquali(P,dades[,k]))
      #calcular els pvalues de les quali
    }
  }
}#endfor
```

```
## [1] "Anàlisi per classes de la Variable: age"
```

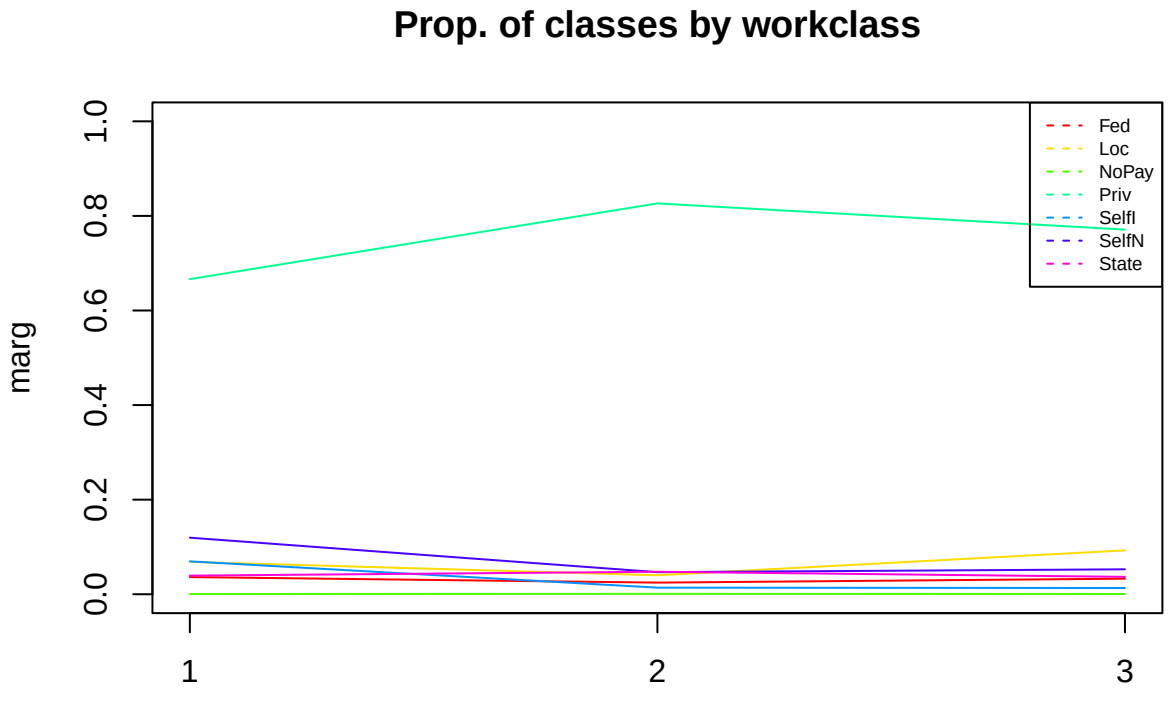





```
## [1] "Estadístics per groups:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   35.00   43.00   44.18   52.00   90.00
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   22.00   27.00   30.64   36.00   90.00
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   25.00   36.00   37.62   47.00   90.00
## [1] "p-valueANOVA: 3.37964367791504e-190"
```

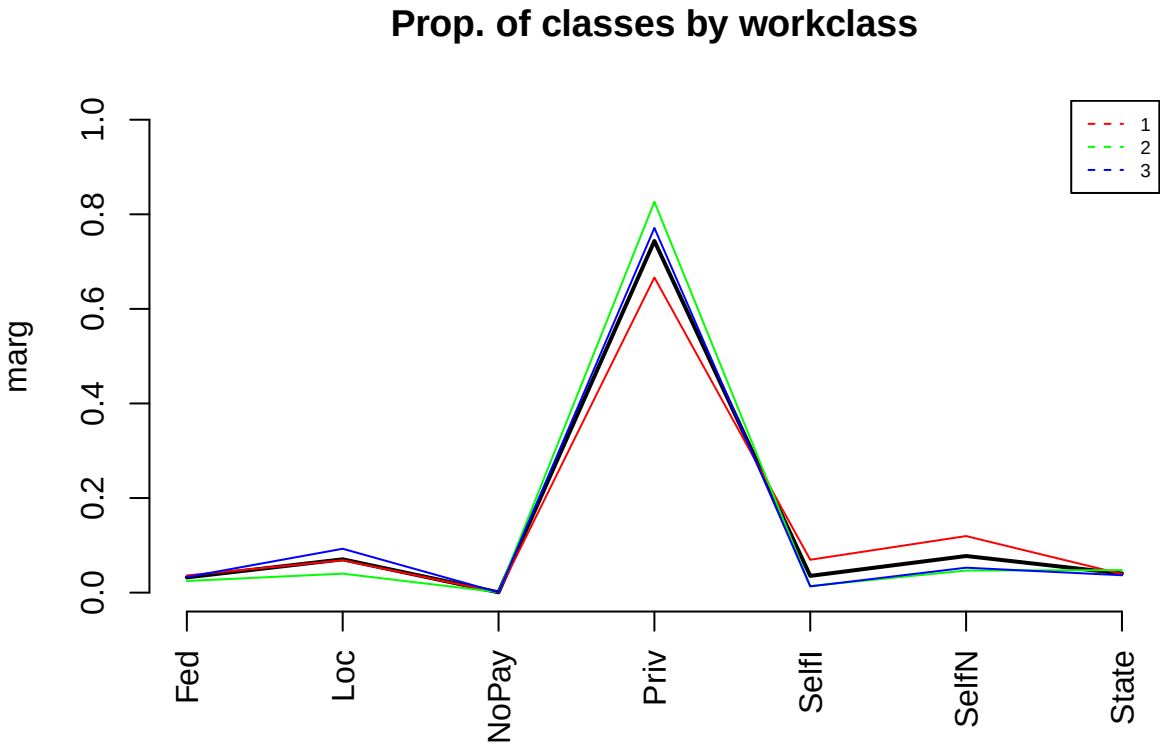
```
## [1] "p-value Kruskal-Wallis: 9.67398091244748e-189"
## [1] "p-values ValorsTest: "
## [1] 9.361532e-115  0.000000e+00  4.265215e-04
## [1] "Variable workclass"
## [1] "Categories=" "Fed"         "Loc"         "NoPay"       "Priv"
## [6] "SelfI"       "SelfN"       "State"
```

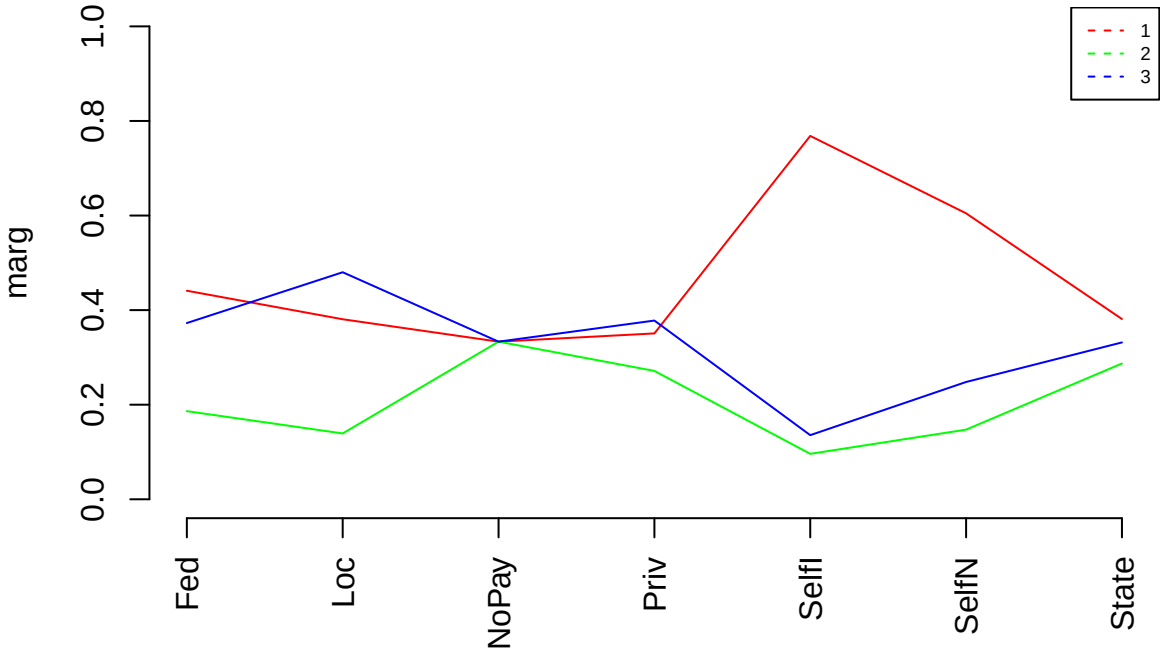

```
## [1] "Categories=" "Fed"         "Loc"         "NoPay"       "Priv"
## [6] "SelfI"       "SelfN"       "State"
```




```
## [1] "Categories=" "Fed"         "Loc"         "NoPay"       "Priv"
```

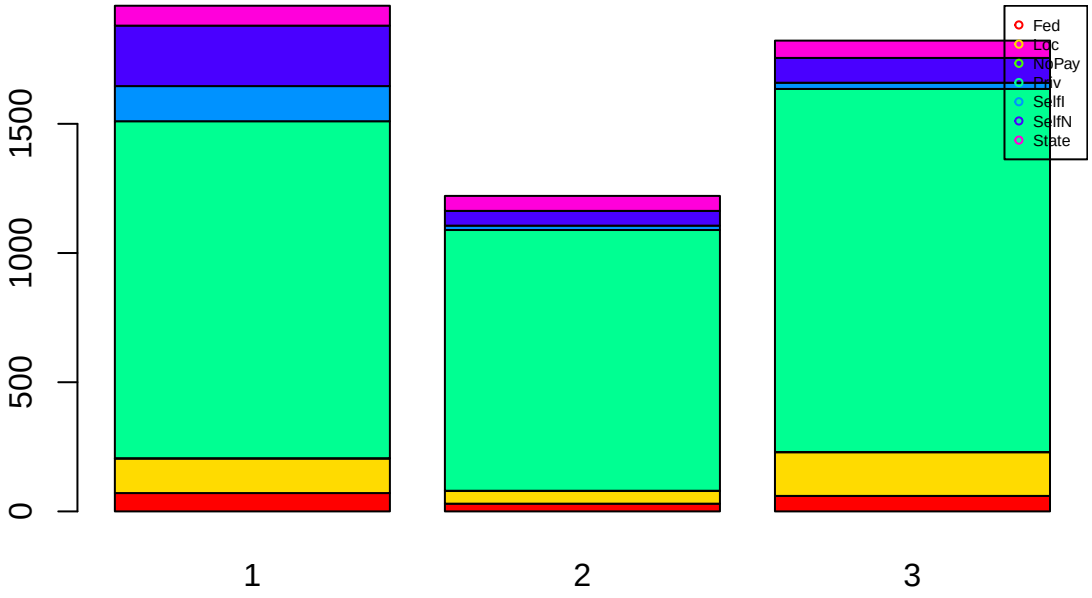
```
## [6] "SelfI"       "SelfN"       "State"
```



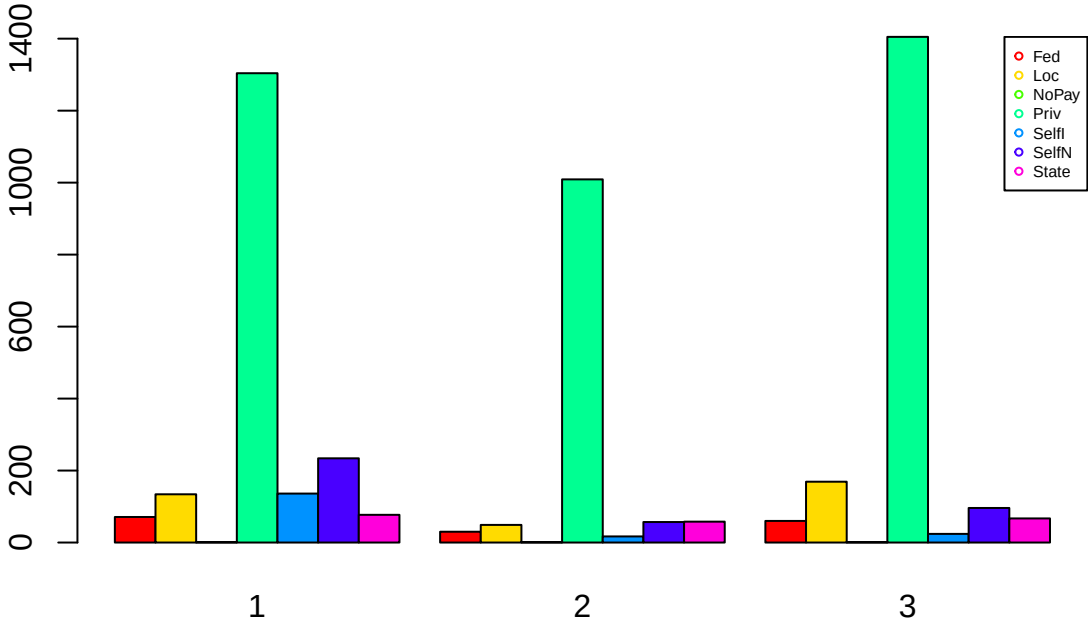



```
## [1] "Cross Table:"
##         P
##            1    2    3
##   Fed     71   30   60
##   Loc    134   49  169
```

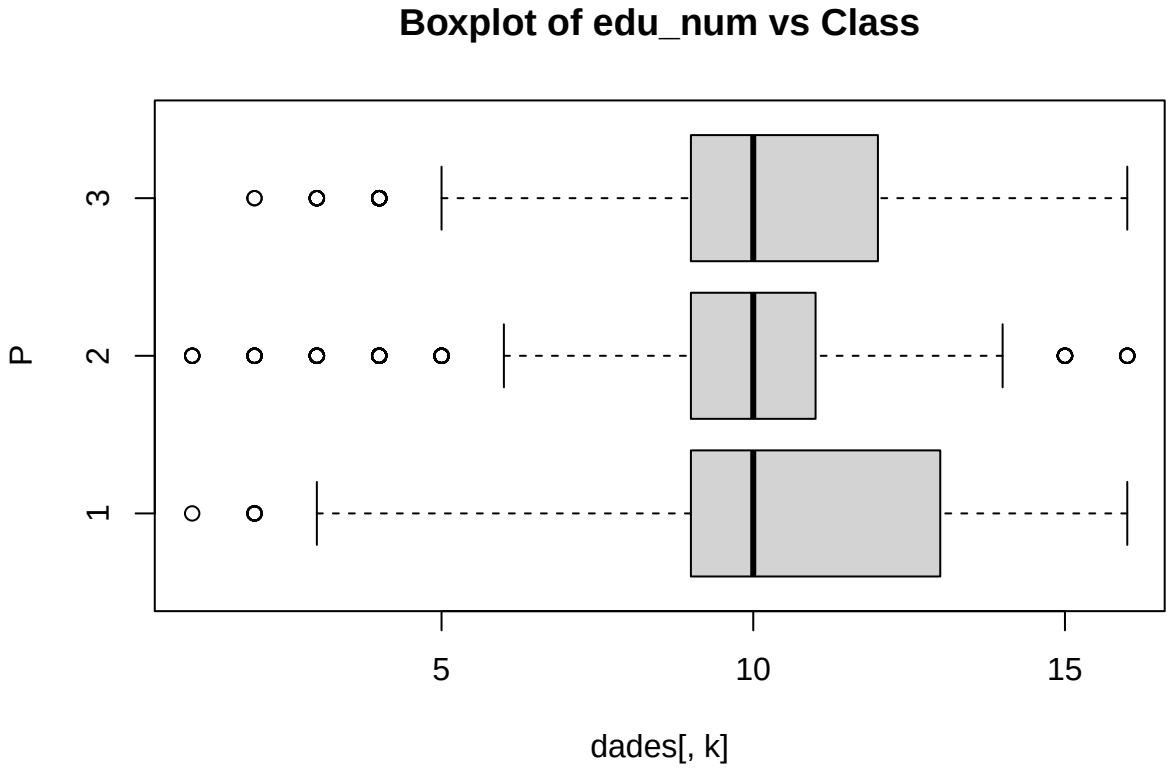
```
##   NoPay    1    1    1
##   Priv  1304 1009 1405
##   SelfI  136   17   24
##   SelfN  234   57   96
##   State   77   58   67
## [1] "Distribucions condicionades a columnes:"
##
## P         Fed       Loc     NoPay      Priv     SelfI     SelfN     State
##   1 0.4409938 0.3806818 0.3333333 0.3507262 0.7683616 0.6046512 0.3811881
##   2 0.1863354 0.1392045 0.3333333 0.2713825 0.0960452 0.1472868 0.2871287
##   3 0.3726708 0.4801136 0.3333333 0.3778913 0.1355932 0.2480620 0.3316832
```

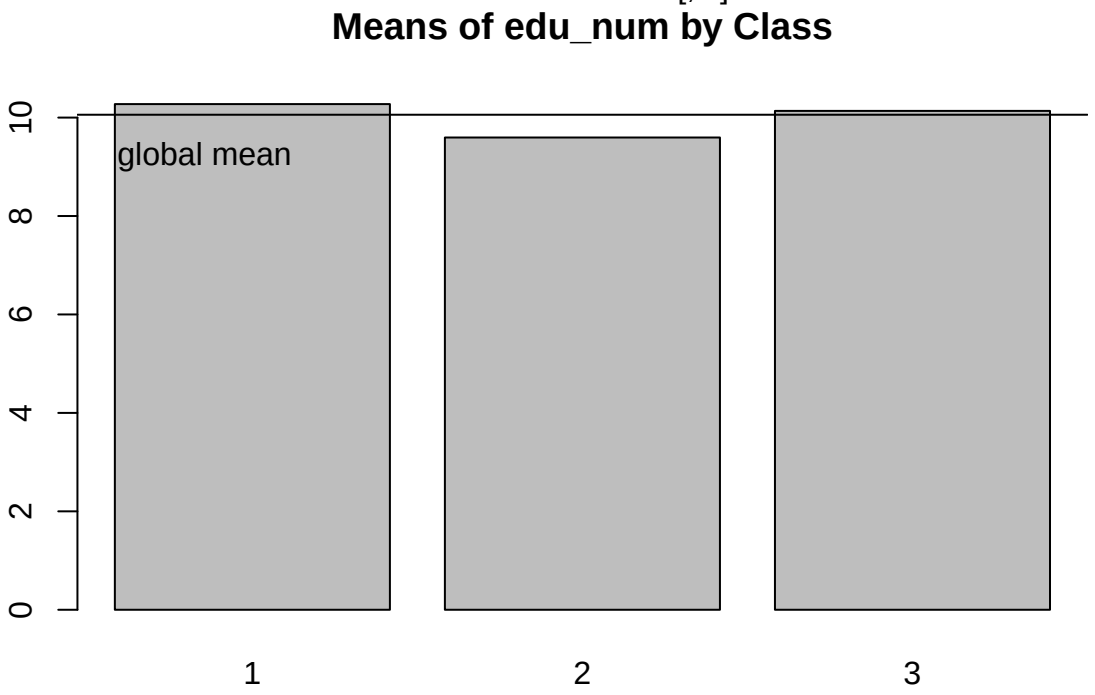



```
## [1] "Test Chi quadrat: "
## Warning in chisq.test(dades[, k], as.factor(P)): L'aproximació Chi-quadrat pot
## ser incorrecta
```



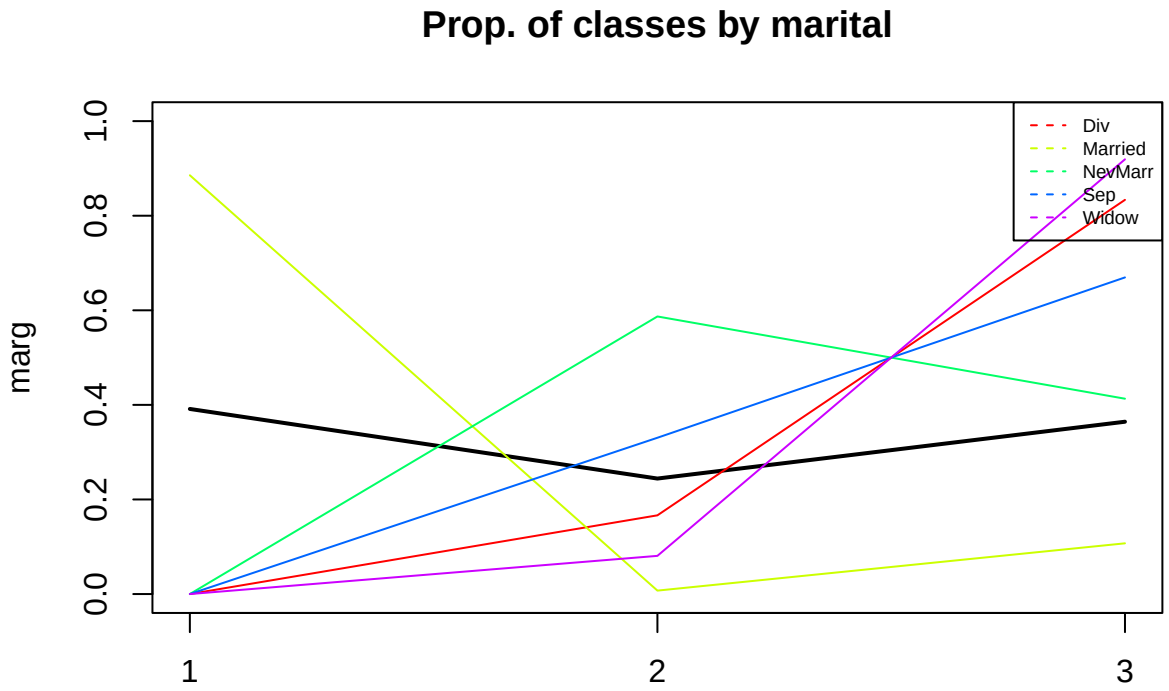
```
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 243.13, df = 12, p-value < 2.2e-16
##
## [1] "valorsTest:"
## $rowpf
##    Xquali
## P            Fed          Loc        NoPay         Priv        SelfI
##   1 0.0362800204 0.0684721513 0.0005109862 0.6663260092 0.0694941237
##   2 0.0245700246 0.0401310401 0.0008190008 0.8263718264 0.0139230139
##   3 0.0329308452 0.0927552141 0.0005488474 0.7711306257 0.0131723381
##    Xquali
## P          SelfN        State
##   1 0.1195707716 0.0393459377
##   2 0.0466830467 0.0475020475
##   3 0.0526893524 0.0367727772
##
## $vtest
##    Xquali
## P           Fed         Loc       NoPay        Priv       SelfI       SelfN
##   1   1.3106026  -0.4273346  -0.2061301 -10.0353860  10.4624576   8.9487790
##   2  -1.7372278  -4.7557334   0.3594635   7.6191799  -4.6714582  -4.6201420
##   3   0.2216598   4.6787160  -0.1118419   3.3757449  -6.4402209  -4.9509571
##    Xquali
## P         State
##   1  -0.3035708
##   2   1.4497802
##   3  -0.9863264
##
## $pval
##    Xquali
## P            Fed          Loc        NoPay         Priv        SelfI
##   1 9.499603e-02 3.345678e-01 4.183446e-01 0.000000e+00 6.423963e-26
##   2 4.117348e-02 9.886364e-07 3.596242e-01 1.276463e-14 1.495345e-06
##   3 4.122894e-01 1.443385e-06 4.554744e-01 3.680807e-04 5.964984e-11
##    Xquali
## P          SelfN        State
##   1 1.797181e-19 3.807275e-01
##   2 1.917388e-06 7.355991e-02
##   3 3.692469e-07 1.619865e-01
##
## [1] "Anàlisi per classes de la Variable: edu_num"
```



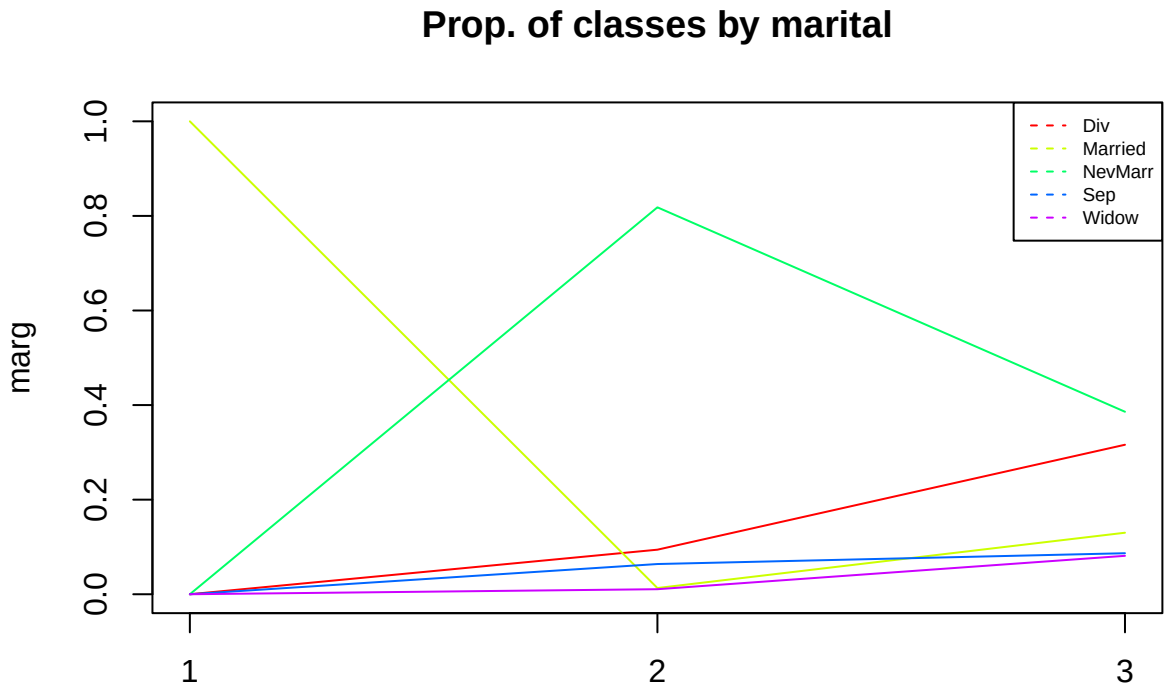



```
## [1] "Estadístics per groups:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.27   13.00   16.00
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   9.000  10.000   9.595  11.000  16.000
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    9.00   10.00   10.14   12.00   16.00
## [1] "p-valueANOVA: 1.231187841333e-11"
```

```
## [1] "p-value Kruskal-Wallis: 1.0090777949887e-08"
## [1] "p-values ValorsTest: "
## [1] 8.904730e-07 2.289280e-13 5.182494e-02
## [1] "Variable marital"
## [1] "Categories=" "Div"         "Married"     "NevMarr"     "Sep"
## [6] "Widow"
```

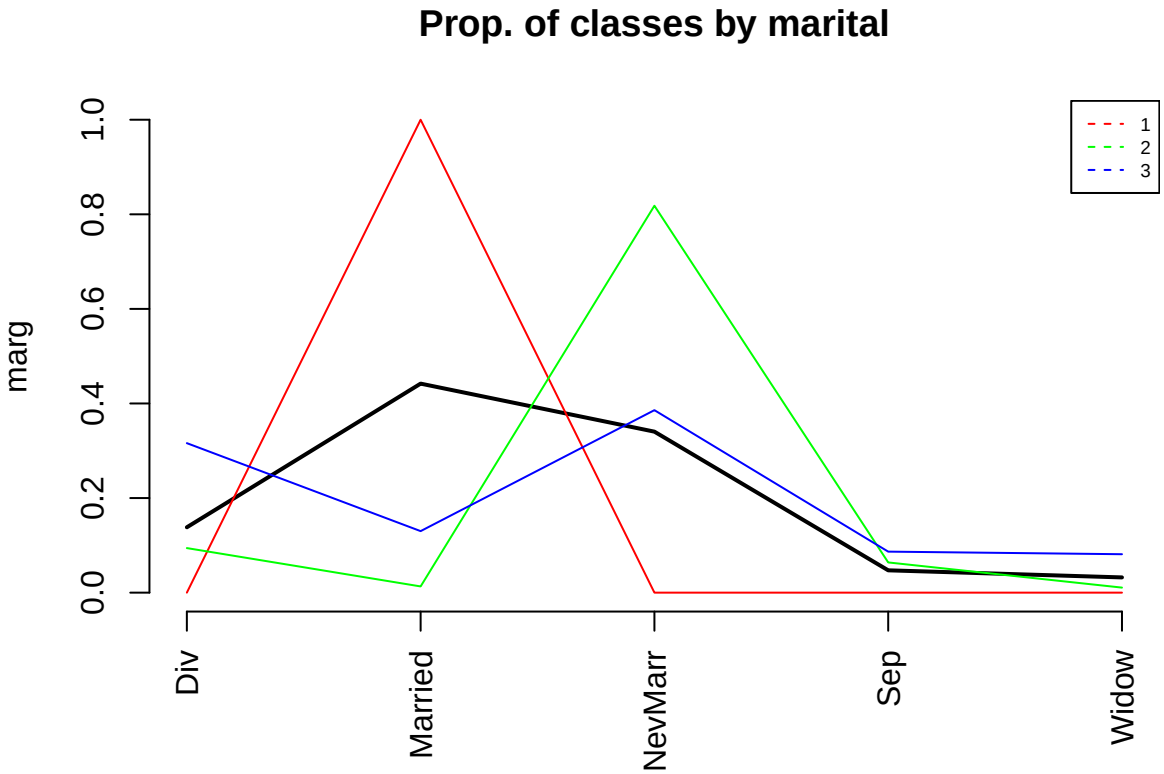


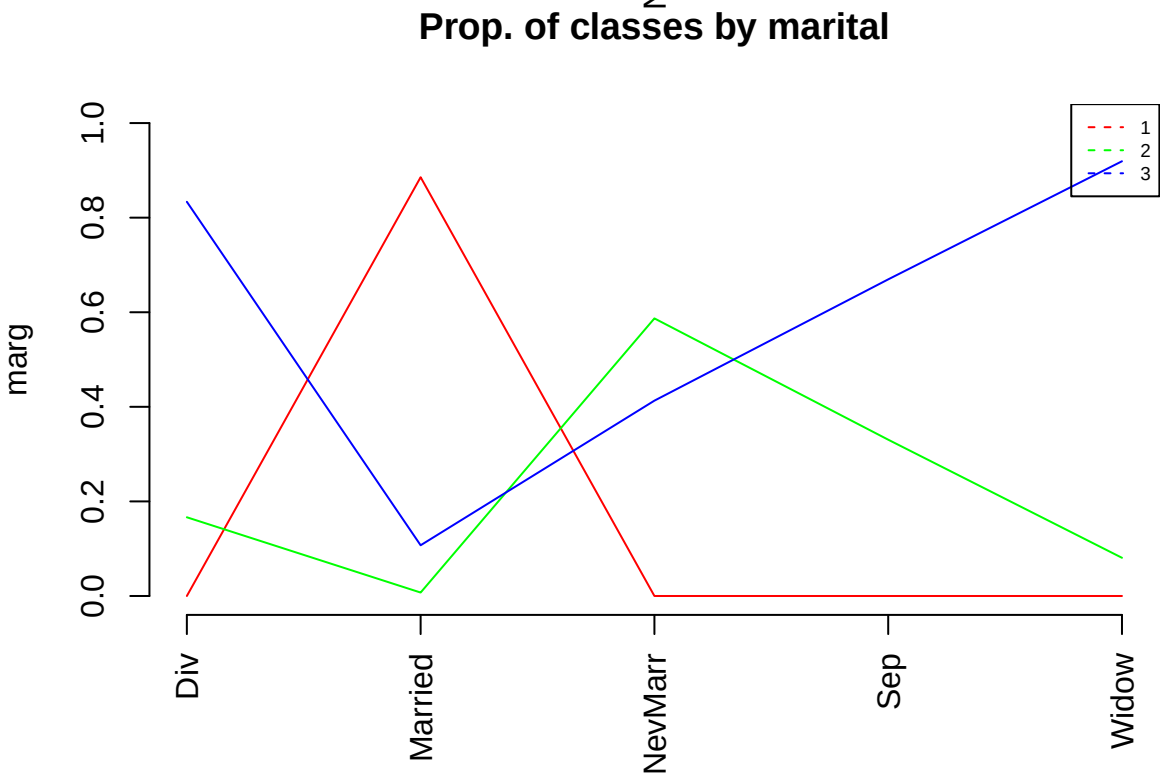

```
## [1] "Categories=" "Div"         "Married"     "NevMarr"     "Sep"
## [6] "Widow"
```




```
## [1] "Categories=" "Div"         "Married"     "NevMarr"     "Sep"
```

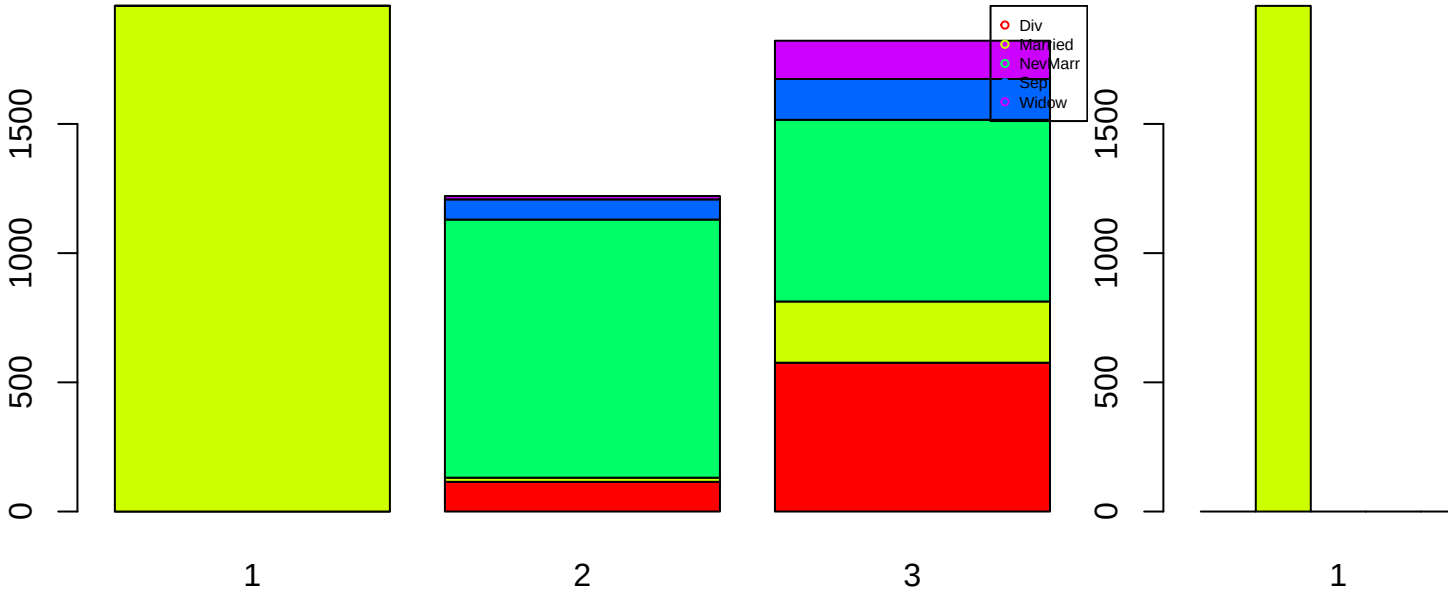
```
## [6] "Widow"
```





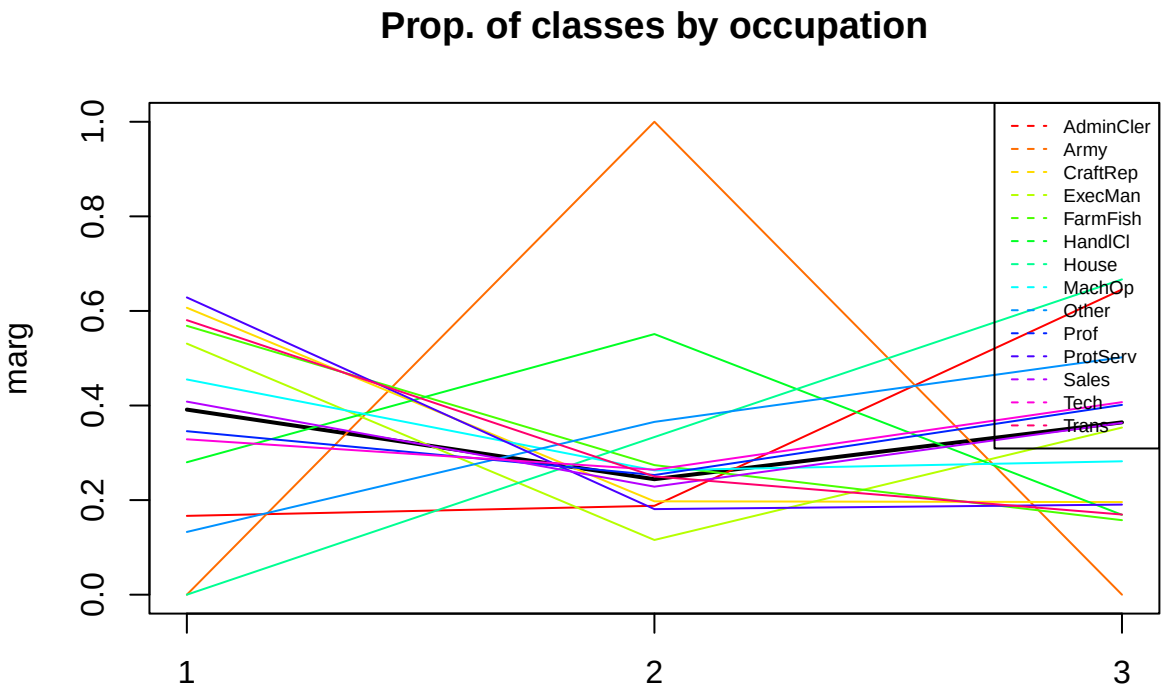

```
## [1] "Cross Table:"
##          P
##              1    2    3
##   Div        0  115  576
```

```
##   Married 1957   16  237
##   NevMarr    0  999  703
##   Sep        0   78  158
##   Widow      0   13  148
## [1] "Distribucions condicionades a columnes:"
##
## P           Div     Married     NevMarr         Sep       Widow
##   1 0.000000000 0.885520362 0.000000000 0.000000000 0.000000000
##   2 0.166425470 0.007239819 0.586956522 0.330508475 0.080745342
##   3 0.833574530 0.107239819 0.413043478 0.669491525 0.919254658
```

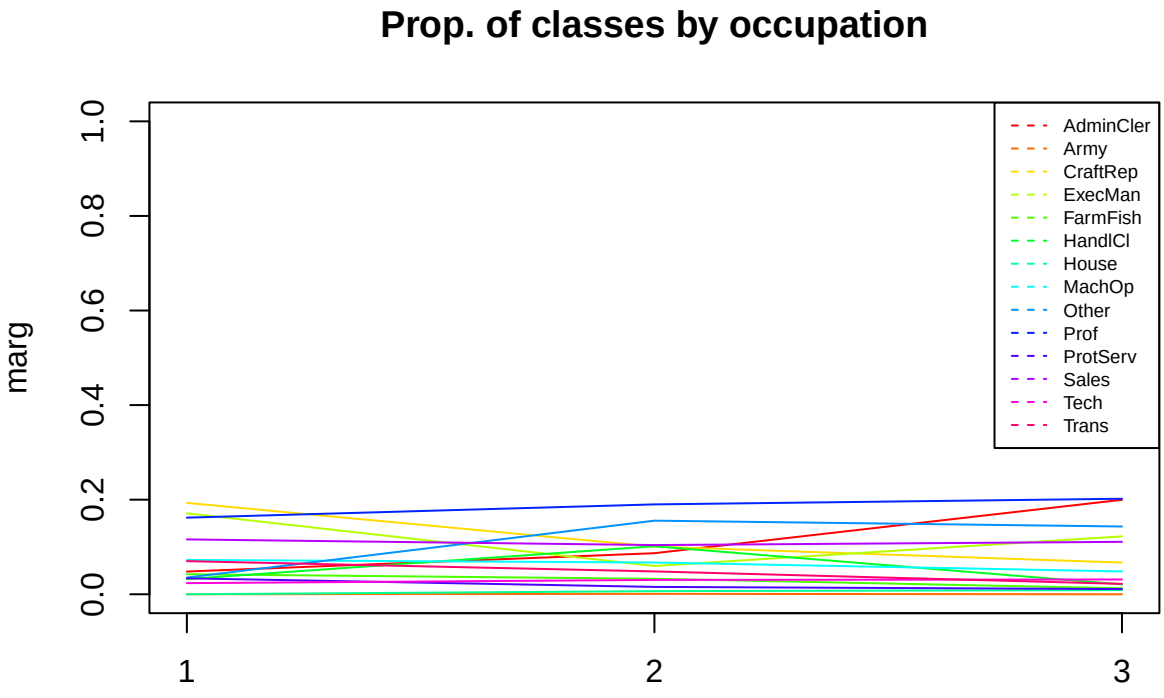



```
## [1] "Test Chi quadrat: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 4865.3, df = 8, p-value < 2.2e-16
##
## [1] "valorsTest:"
## $rowpf
##    Xquali
## P           Div    Married    NevMarr        Sep      Widow
##   1 0.00000000 1.00000000 0.00000000 0.00000000 0.00000000
##   2 0.09418509 0.01310401 0.81818182 0.06388206 0.01064701
##   3 0.31613611 0.13007684 0.38583974 0.08671789 0.08122942
##
## $vtest
##    Xquali
## P           Div    Married    NevMarr        Sep      Widow
##   1 -22.708070  63.714078 -40.736517 -12.621168 -10.343430
##   2  -5.126216 -34.711880  40.527415   3.161785  -4.907284
##   3  27.605084 -33.628046   5.134320   9.977096  14.870249
##
## $pval
##    Xquali
```

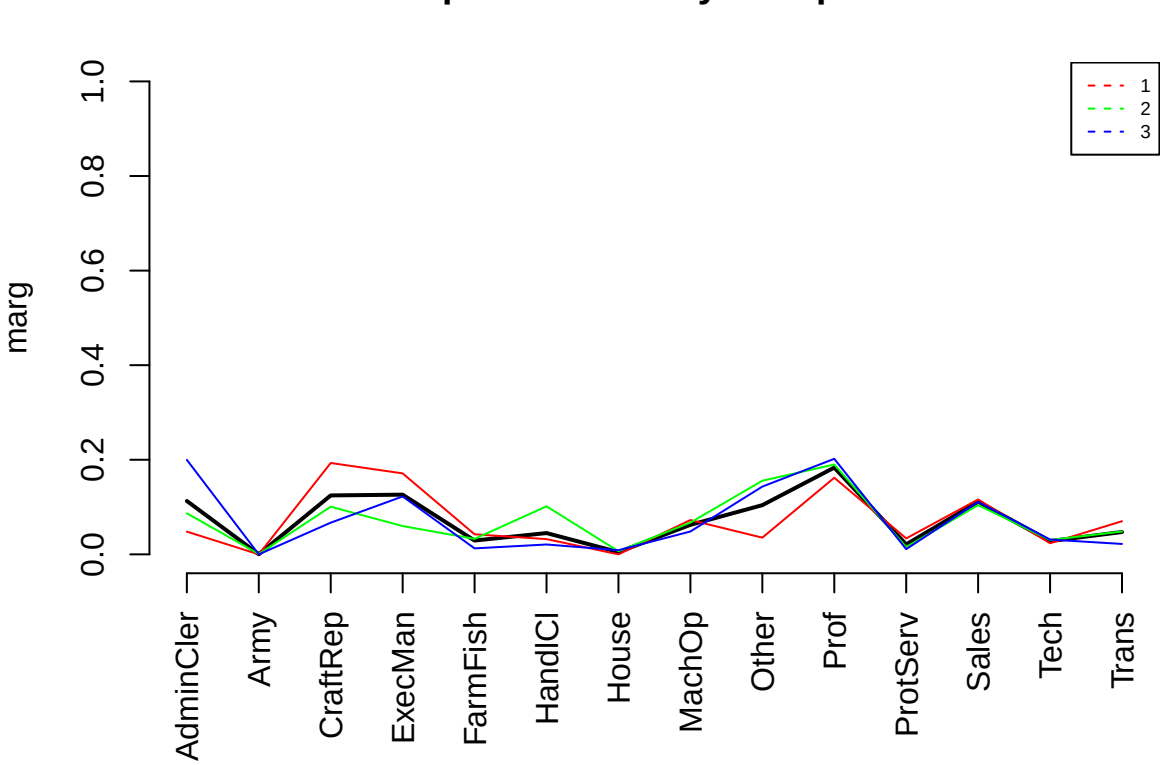
```
## P             Div       Married       NevMarr           Sep         Widow
##   1  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
##   2  1.478117e-07  0.000000e+00  0.000000e+00  7.840272e-04  4.617323e-07
##   3 4.834171e-168  0.000000e+00  1.415828e-07  9.600218e-24  2.571091e-50
##
## [1] "Variable occupation"
##  [1] "Categories=" "AdminCler"   "Army"        "CraftRep"    "ExecMan"
##  [6] "FarmFish"    "HandlCl"     "House"       "MachOp"      "Other"
## [11] "Prof"        "ProtServ"    "Sales"       "Tech"        "Trans"
```

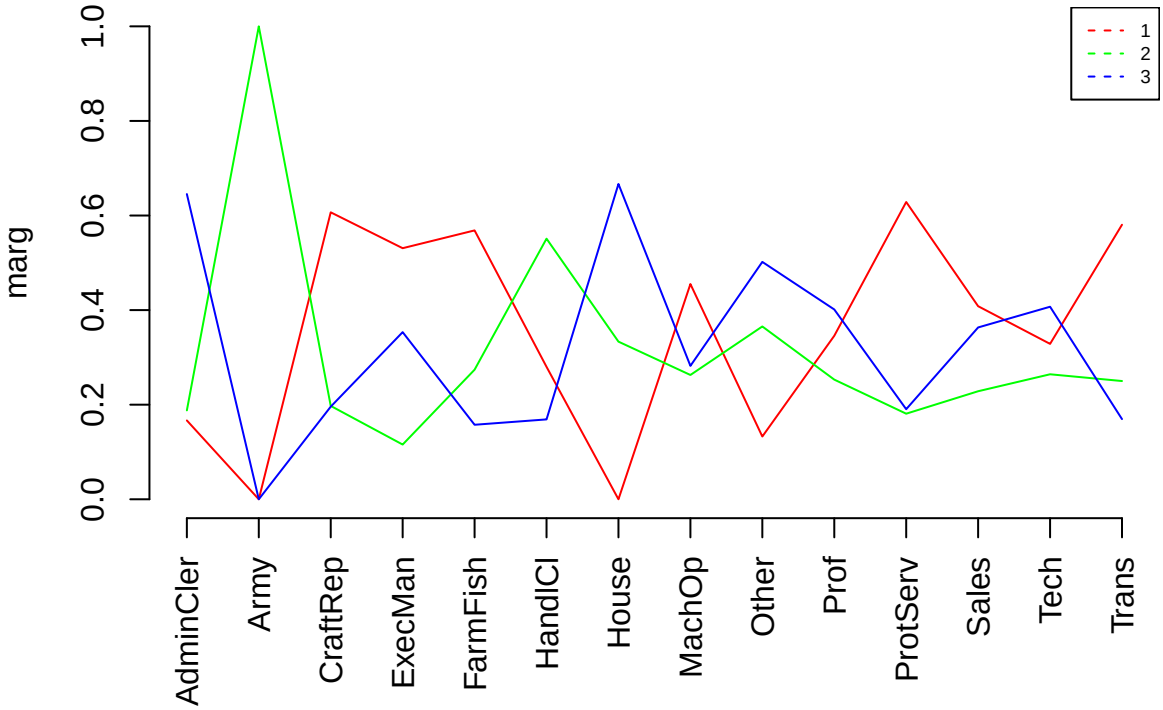



```
##  [1] "Categories=" "AdminCler"   "Army"        "CraftRep"    "ExecMan"
##  [6] "FarmFish"    "HandlCl"     "House"       "MachOp"      "Other"
## [11] "Prof"        "ProtServ"    "Sales"       "Tech"        "Trans"
```

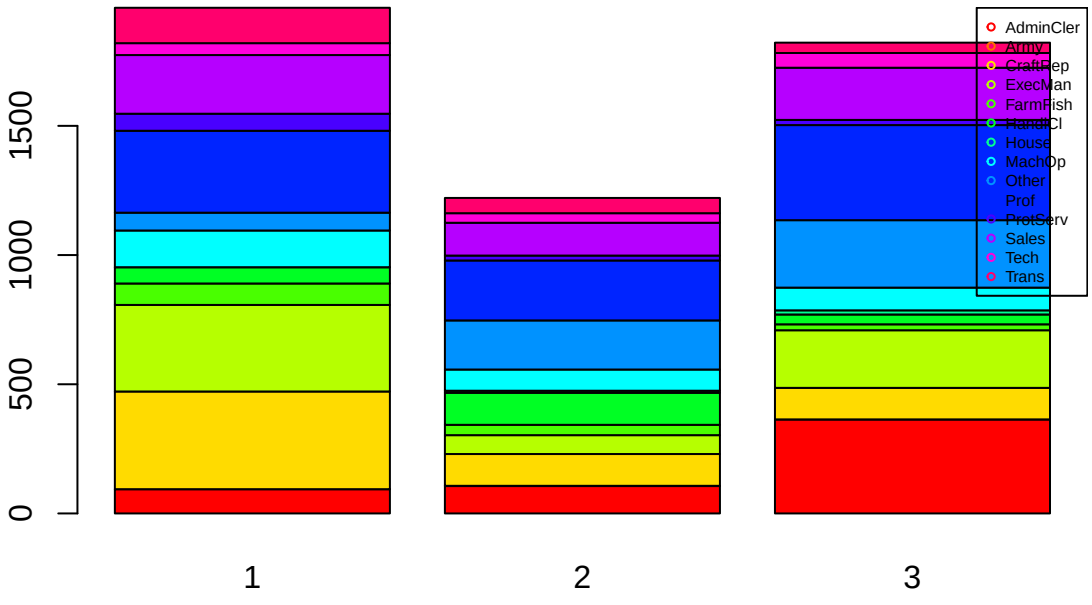



```
##  [1] "Categories=" "AdminCler"   "Army"       "CraftRep"   "ExecMan"
##  [6] "FarmFish"    "HandlCl"     "House"      "MachOp"     "Other"
## [11] "Prof"        "ProtServ"    "Sales"      "Tech"       "Trans"
```



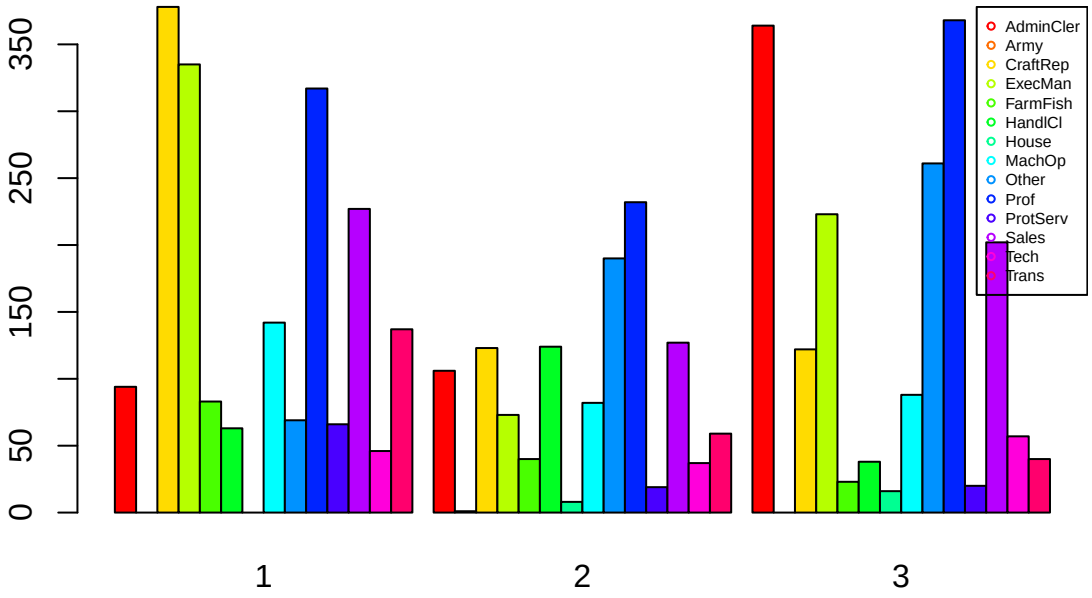



```
## [1] "Cross Table:"
##            P
##               1   2   3
##   AdminCler  94 106 364
##   Army        0   1   0
##   CraftRep  378 123 122
##   ExecMan   335  73 223
##   FarmFish   83  40  23
##   HandlCl    63 124  38
##   House       0   8  16
##   MachOp    142  82  88
##   Other      69 190 261
##   Prof      317 232 368
##   ProtServ   66  19  20
##   Sales     227 127 202
##   Tech       46  37  57
##   Trans     137  59  40
## [1] "Distribucions condicionades a columnes:"
##
## P   AdminCler      Army  CraftRep   ExecMan  FarmFish   HandlCl     House
##   1 0.1666667 0.0000000 0.6067416 0.5309033 0.5684932 0.2800000 0.0000000
##   2 0.1879433 1.0000000 0.1974318 0.1156894 0.2739726 0.5511111 0.3333333
##   3 0.6453901 0.0000000 0.1958266 0.3534073 0.1575342 0.1688889 0.6666667
##
## P      MachOp     Other      Prof  ProtServ     Sales      Tech     Trans
##   1 0.4551282 0.1326923 0.3456925 0.6285714 0.4082734 0.3285714 0.5805085
##   2 0.2628205 0.3653846 0.2529989 0.1809524 0.2284173 0.2642857 0.2500000
##   3 0.2820513 0.5019231 0.4013086 0.1904762 0.3633094 0.4071429 0.1694915
```



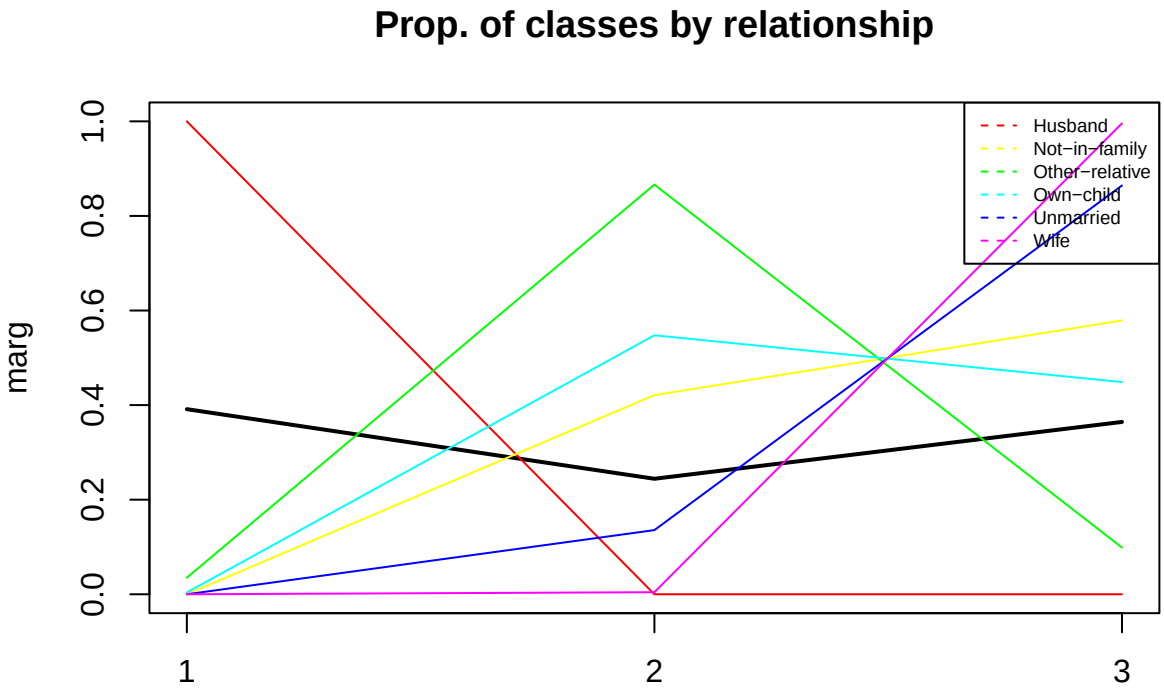

```
## [1] "Test Chi quadrat: "
```

```
## Warning in chisq.test(dades[, k], as.factor(P)): L'aproximació Chi-quadrat pot
## ser incorrecta
```

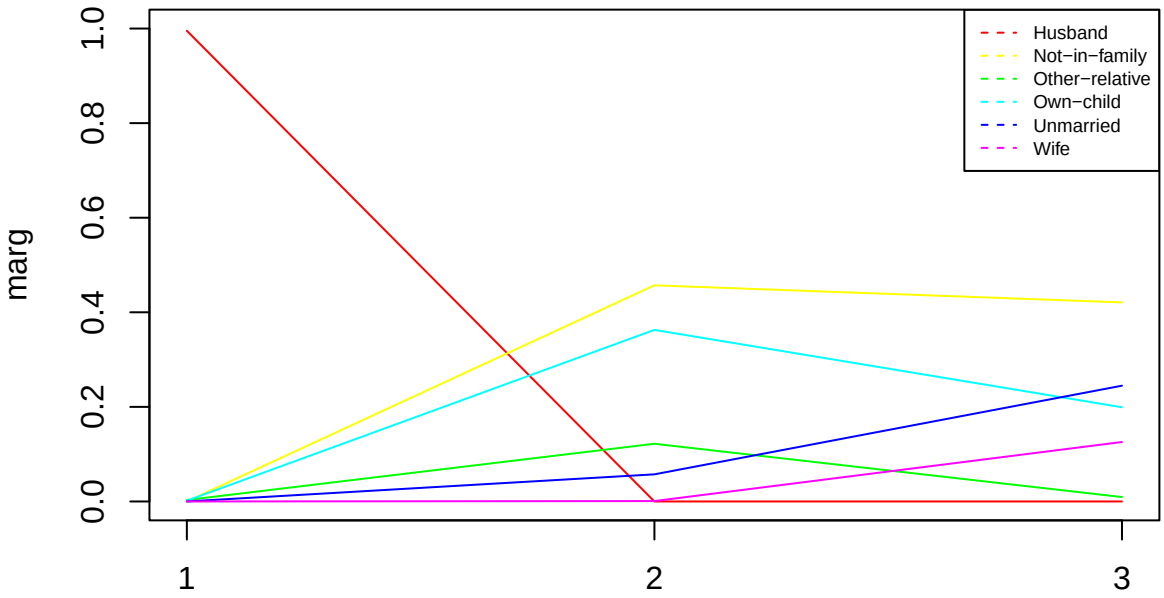



```
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 811.03, df = 26, p-value < 2.2e-16
##
## [1] "valorsTest:"
## $rowpf
##    Xquali
## P       AdminCler         Army     CraftRep      ExecMan     FarmFish
##   1 0.0480327031 0.0000000000 0.1931527849 0.1711803781 0.0424118549
##   2 0.0868140868 0.0008190008 0.1007371007 0.0597870598 0.0327600328
```

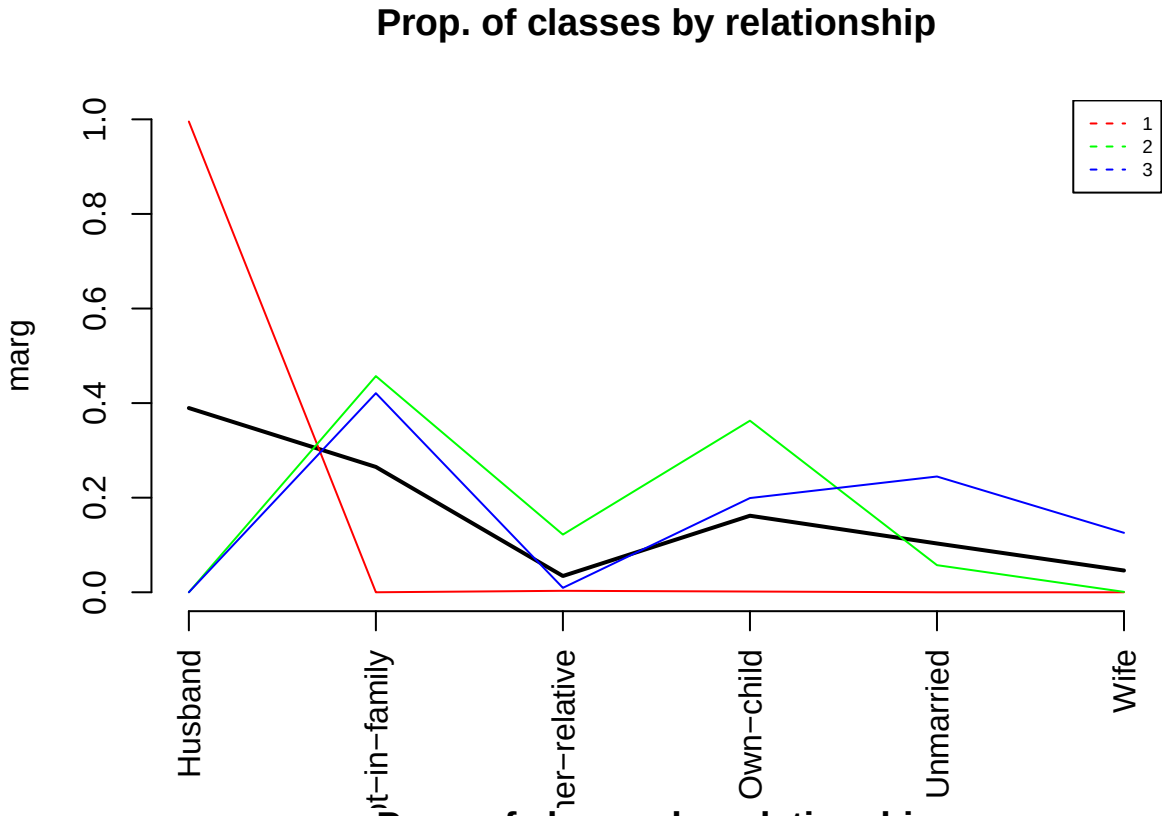
```
##   3 0.1997804610 0.0000000000 0.0669593853 0.1223929748 0.0126234907
##    Xquali
## P        HandlCl        House       MachOp        Other         Prof
##   1 0.0321921308 0.0000000000 0.0725600409 0.0352580480 0.1619826265
##   2 0.1015561016 0.0065520066 0.0671580672 0.1556101556 0.1900081900
##   3 0.0208562020 0.0087815587 0.0482985730 0.1432491767 0.2019758507
##    Xquali
## P       ProtServ        Sales         Tech        Trans
##   1 0.0337250894 0.1159938682 0.0235053654 0.0700051099
##   2 0.0155610156 0.1040131040 0.0303030303 0.0483210483
##   3 0.0109769484 0.1108671789 0.0312843030 0.0219538968
##
## $vtest
##    Xquali
## P      AdminCler         Army     CraftRep      ExecMan     FarmFish
##   1 -11.60967051  -0.80202493  11.77042783   7.68098370   4.44976345
##   2  -3.30161976   1.75943833  -2.90412499  -8.03843303   0.84986917
##   3  14.72104468  -0.75725243  -9.34433831  -0.61380766  -5.27131555
##    Xquali
## P        HandlCl        House       MachOp        Other         Prof
##   1  -3.50346981  -3.93817376   2.38190151 -12.76969794  -3.13827716
##   2  10.96543617   1.01885953   0.79064925   6.79546034   0.68632933
##   3  -6.23562355   3.08432096  -3.12136252   6.88401993   2.56996140
##    Xquali
## P       ProtServ        Sales         Tech        Trans
##   1   5.03256379   0.86469134  -1.54494024   6.09803249
##   2  -1.52465370  -0.91884304   0.56110210   0.21247452
##   3  -3.74266723  -0.05668115   1.06589225  -6.37389231
##
## $pval
##    Xquali
## P      AdminCler         Army     CraftRep      ExecMan     FarmFish
##   1 0.000000e+00 2.112693e-01 2.772131e-32 7.893579e-15 4.298246e-06
##   2 4.806414e-04 3.925154e-02 1.841406e-03 4.440892e-16 1.976989e-01
##   3 2.361536e-49 2.244493e-01 0.000000e+00 2.696712e-01 6.772468e-08
##    Xquali
## P        HandlCl        House       MachOp        Other         Prof
##   1 2.296194e-04 4.105207e-05 8.611752e-03 0.000000e+00 8.497205e-04
##   2 2.801487e-28 1.541348e-01 2.145744e-01 5.398361e-12 2.462527e-01
##   3 2.249910e-10 1.020087e-03 9.000814e-04 2.909336e-12 5.085492e-03
##    Xquali
## P       ProtServ        Sales         Tech        Trans
##   1 2.419817e-07 1.936041e-01 6.118036e-02 5.369096e-10
##   2 6.367275e-02 1.790888e-01 2.873640e-01 4.158684e-01
##   3 9.103864e-05 4.773996e-01 1.432362e-01 9.214496e-11
##
## [1] "Variable relationship"
## [1] "Categories="    "Husband"        "Not-in-family"  "Other-relative"
## [5] "Own-child"      "Unmarried"      "Wife"
```

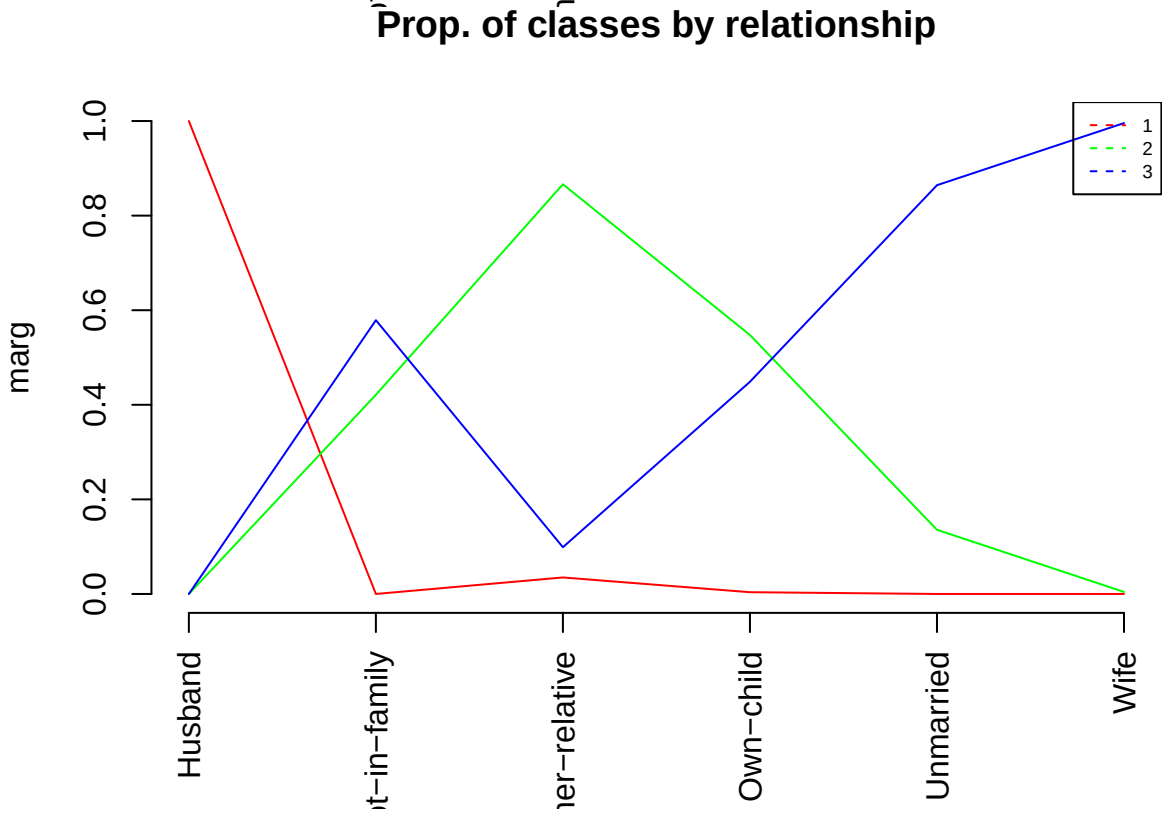



```
## [1] "Categories="    "Husband"        "Not-in-family"  "Other-relative"
## [5] "Own-child"      "Unmarried"      "Wife"
```



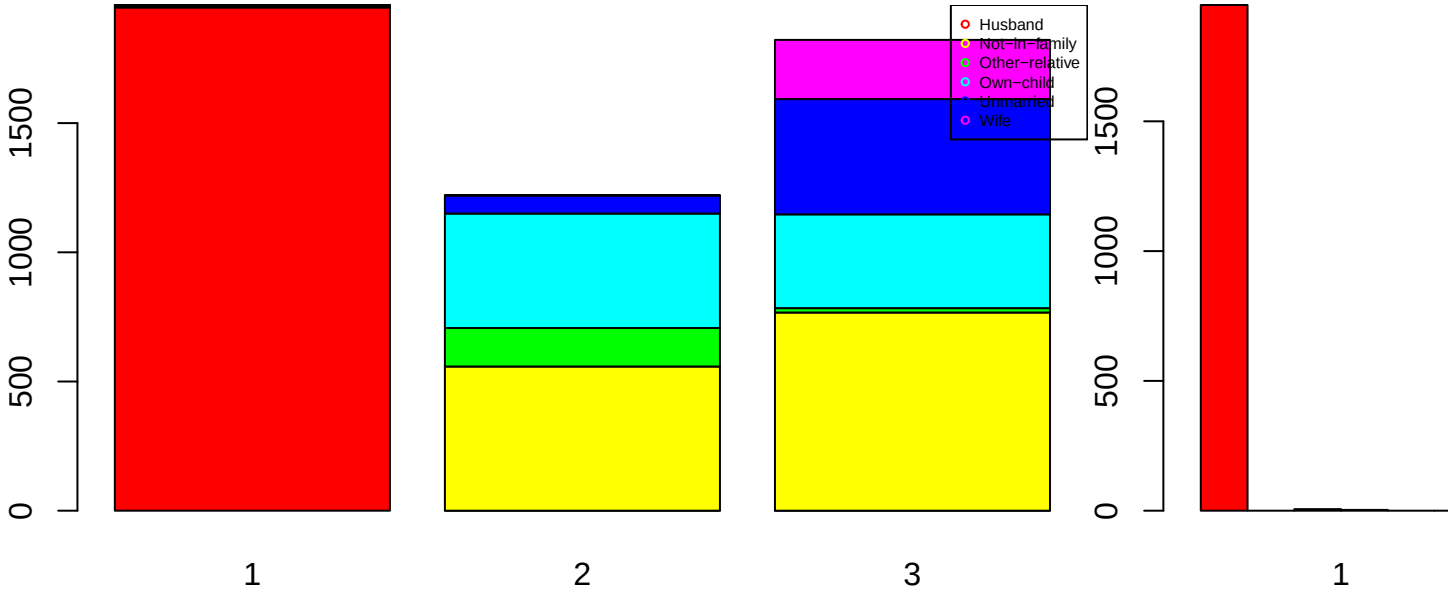

```
## [1] "Categories="    "Husband"        "Not-in-family"  "Other-relative"
## [5] "Own-child"      "Unmarried"      "Wife"
```





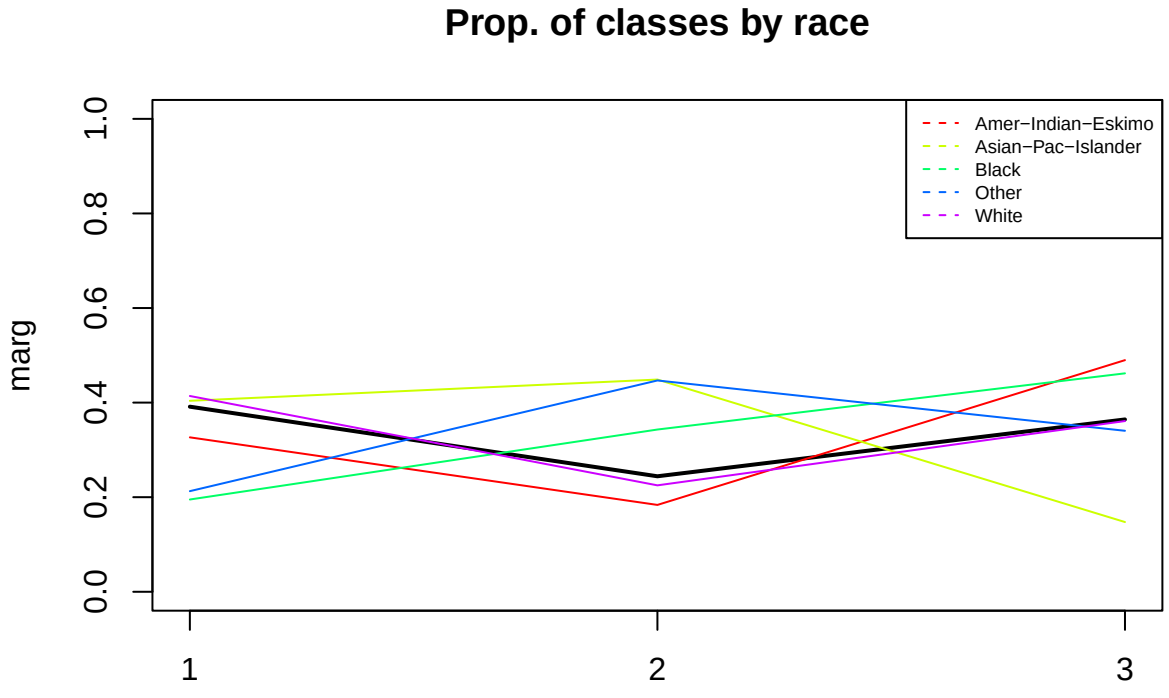

```
## [1] "Cross Table:"
##                   P
##                       1     2     3
```

```
##   Husband        1948    0    0
##   Not-in-family     0  558  767
##   Other-relative    6  149   17
##   Own-child         3  443  363
##   Unmarried         0   70  446
##   Wife              0    1  229
## [1] "Distribucions condicionades a columnes:"
##
## P       Husband Not-in-family Other-relative   Own-child   Unmarried
##   1 1.000000000   0.000000000    0.034883721 0.003708282 0.000000000
##   2 0.000000000   0.421132075    0.866279070 0.547589617 0.135658915
##   3 0.000000000   0.578867925    0.098837209 0.448702101 0.864341085
##
## P          Wife
##   1 0.000000000
##   2 0.004347826
##   3 0.995652174
```

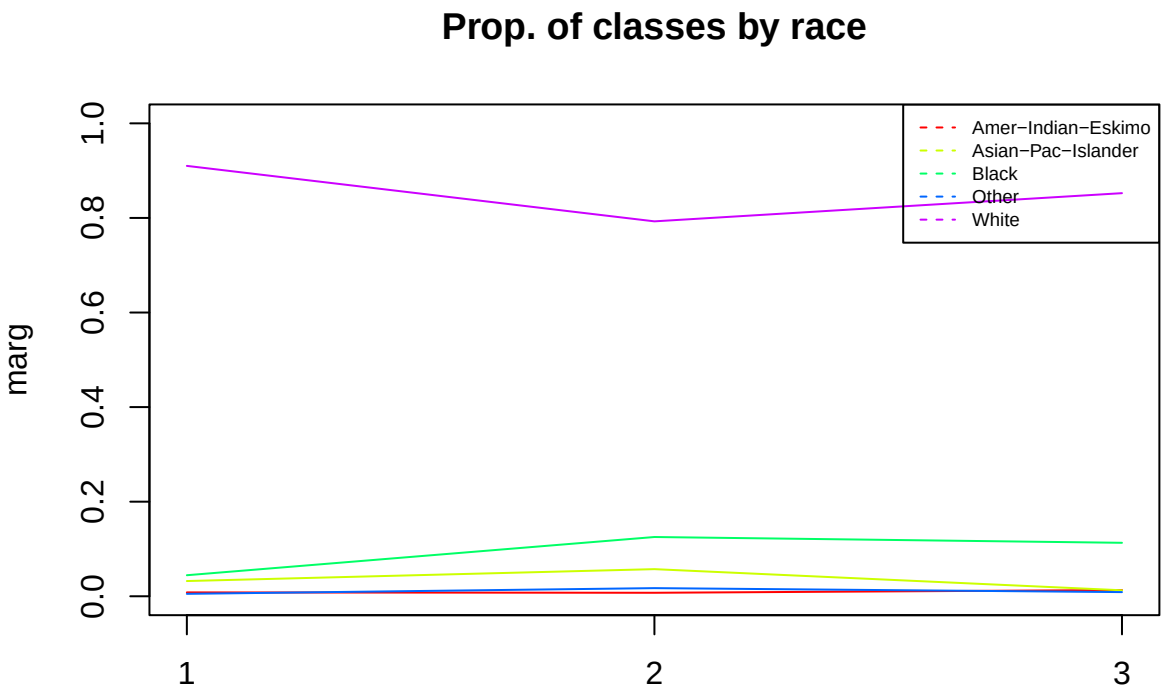



```
## [1] "Test Chi quadrat: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 5854.3, df = 10, p-value < 2.2e-16
##
## [1] "valorsTest:"
## $rowpf
##    Xquali
## P        Husband Not-in-family Other-relative    Own-child    Unmarried
##   1 0.9954011242  0.0000000000   0.0030659172 0.0015329586 0.0000000000
##   2 0.0000000000  0.4570024570   0.1220311220 0.3628173628 0.0573300573
##   3 0.0000000000  0.4209659715   0.0093304061 0.1992316136 0.2447859495
##    Xquali
## P           Wife
##   1 0.0000000000
```

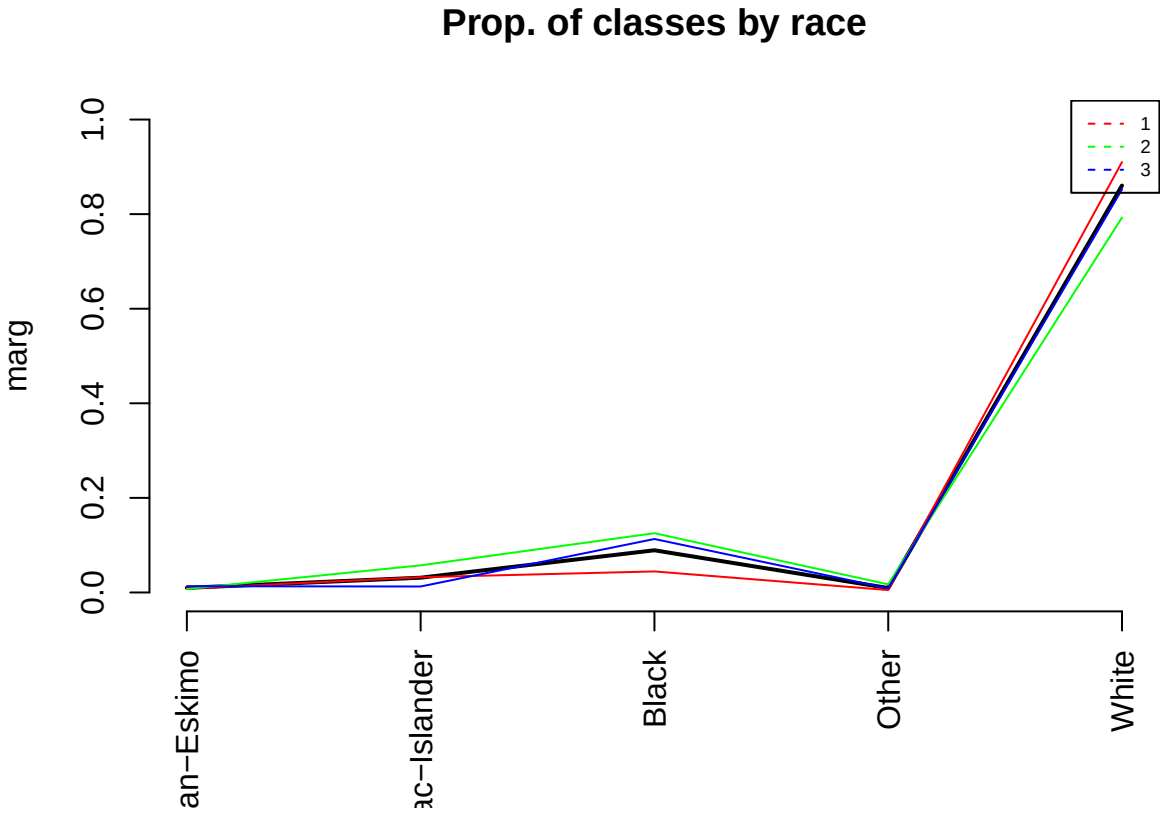
```
##   2 0.0008190008
##   3 0.1256860593
##
## $vtest
##    Xquali
## P      Husband Not-in-family Other-relative  Own-child  Unmarried       Wife
##   1  70.443800    -34.049341      -9.749191 -24.678035 -19.236298 -12.451858
##   2 -32.111205     17.486165      19.325686  21.939406  -6.060311  -8.668763
##   3 -42.774452     18.921061      -7.364643   5.441990  24.918090  20.366270
##
## $pval
##    Xquali
## P         Husband Not-in-family Other-relative     Own-child     Unmarried
##   1  0.000000e+00  0.000000e+00   0.000000e+00  0.000000e+00  0.000000e+00
##   2  0.000000e+00  9.132125e-69   1.633078e-83 5.465957e-107  6.792915e-10
##   3  0.000000e+00  3.824933e-80   8.881784e-14  2.634438e-08 2.368874e-137
##    Xquali
## P            Wife
##   1  0.000000e+00
##   2  0.000000e+00
##   3  1.665531e-92
##
## [1] "Variable race"
## [1] "Categories="        "Amer-Indian-Eskimo" "Asian-Pac-Islander"
## [4] "Black"              "Other"              "White"
```

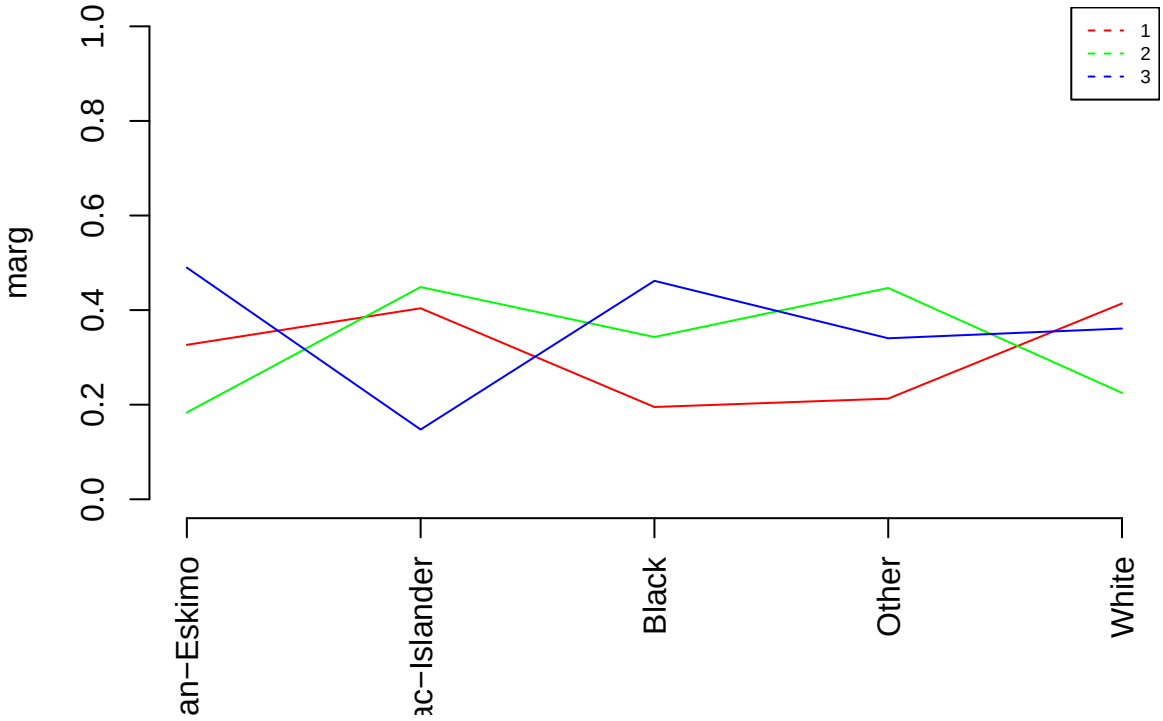



```
## [1] "Categories="        "Amer-Indian-Eskimo" "Asian-Pac-Islander"
## [4] "Black"              "Other"              "White"
```

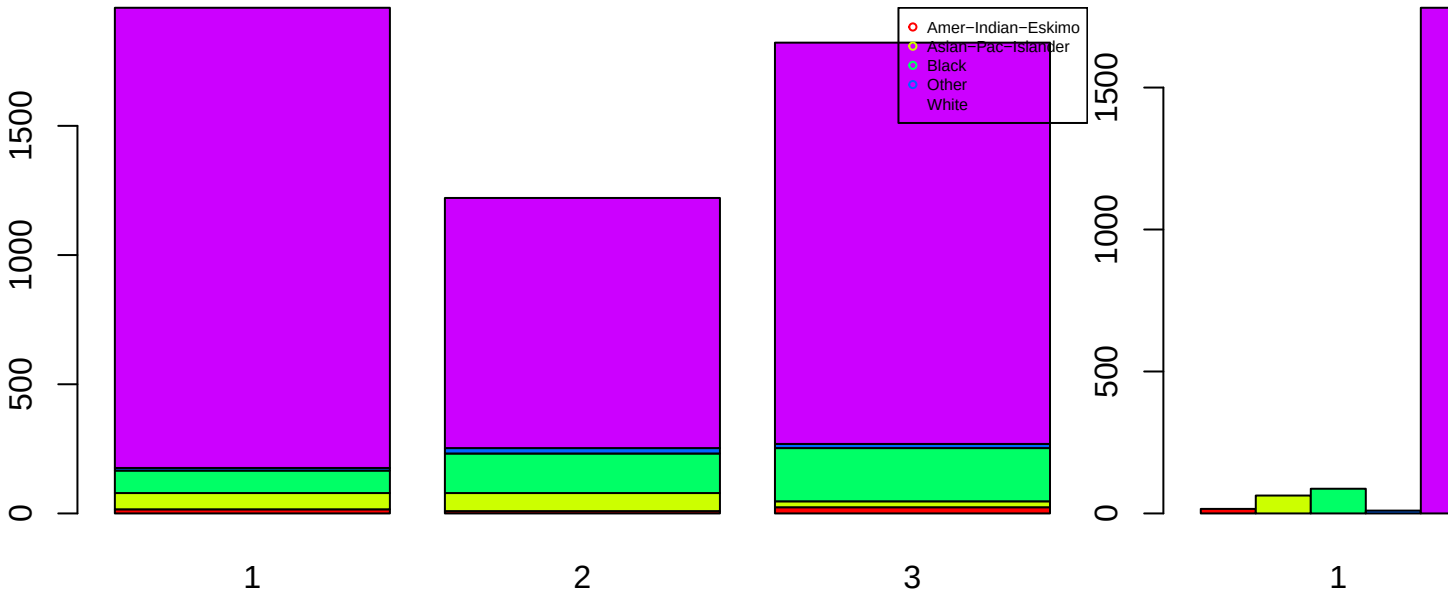



```
## [1] "Categories="        "Amer-Indian-Eskimo" "Asian-Pac-Islander"
## [4] "Black"              "Other"              "White"
```



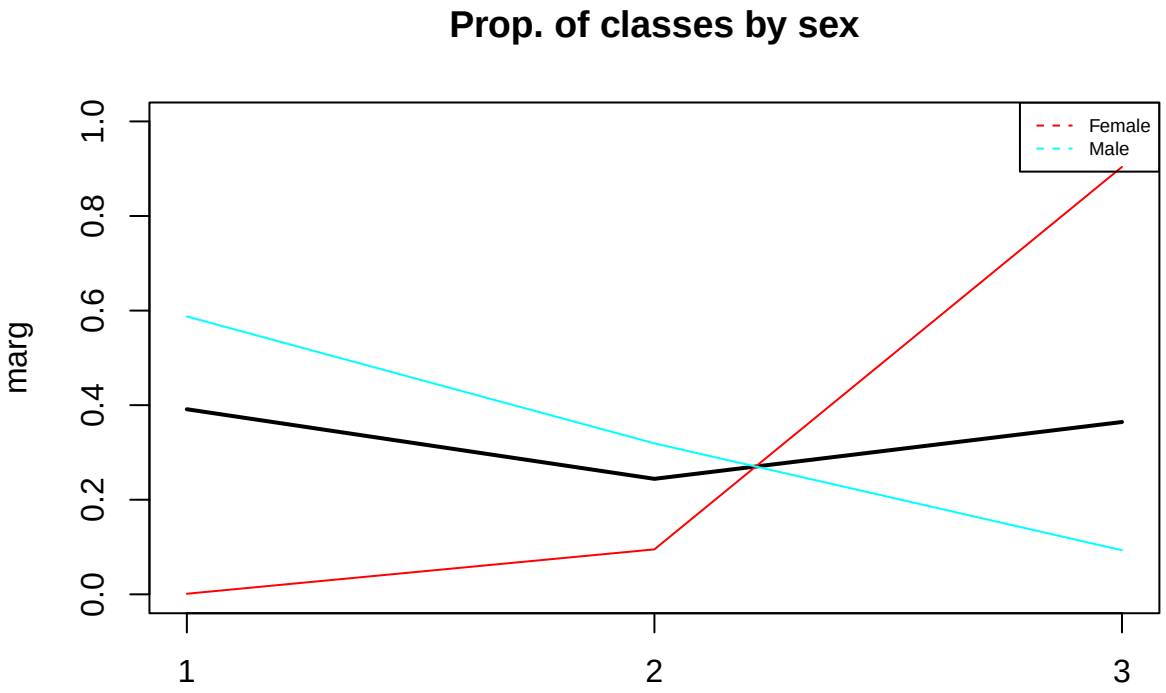



```
## [1] "Cross Table:"
##                       P
##                         1    2    3
##   Amer-Indian-Eskimo   16    9   24
##   Asian-Pac-Islander   63   70   23
##   Black                87  153  206
##   Other                10   21   16
##   White              1781  968 1553
## [1] "Distribucions condicionades a columnes:"
##
## P   Amer-Indian-Eskimo Asian-Pac-Islander     Black     Other     White
##   1          0.3265306          0.4038462 0.1950673 0.2127660 0.4139935
##   2          0.1836735          0.4487179 0.3430493 0.4468085 0.2250116
##   3          0.4897959          0.1474359 0.4618834 0.3404255 0.3609949
```

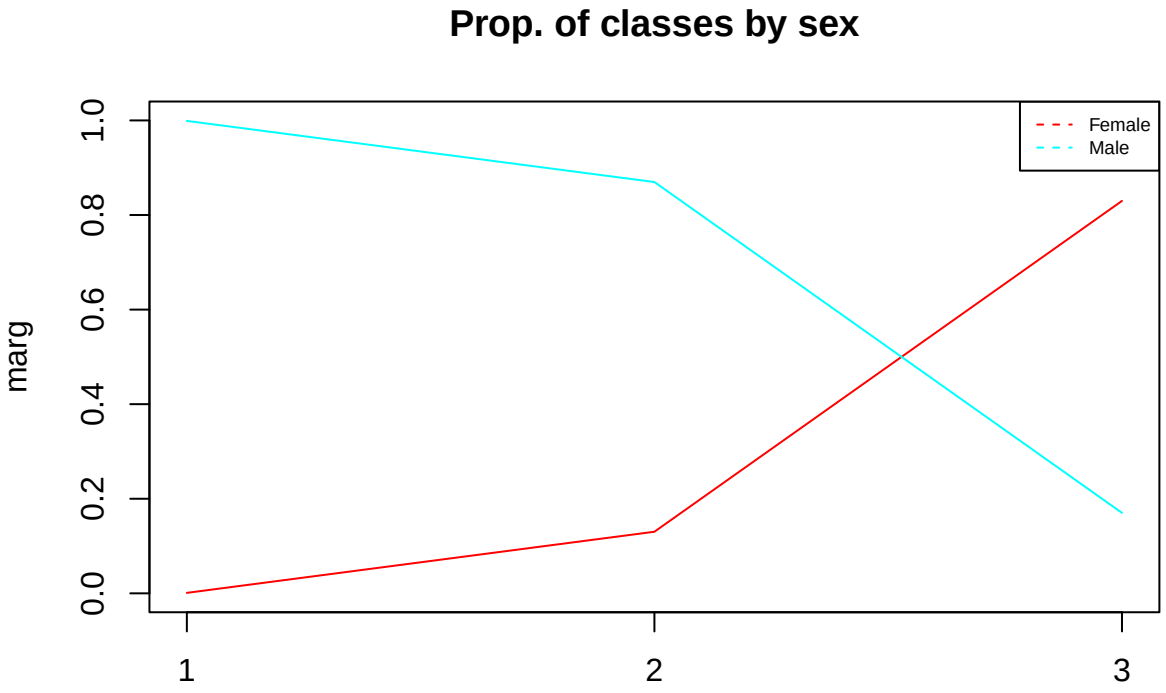



```
## [1] "Test Chi quadrat: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 147.75, df = 8, p-value < 2.2e-16
##
## [1] "valorsTest:"
## $rowpf
##    Xquali
## P   Amer-Indian-Eskimo Asian-Pac-Islander       Black       Other       White
##   1        0.008175779        0.032192131 0.044455800 0.005109862 0.910066428
##   2        0.007371007        0.057330057 0.125307125 0.017199017 0.792792793
##   3        0.013172338        0.012623491 0.113062569 0.008781559 0.852360044
##
## $vtest
##    Xquali
## P   Amer-Indian-Eskimo Asian-Pac-Islander      Black      Other      White
##   1         -0.9349750          0.3235966 -8.9016856 -2.5210857  8.1264236
##   2         -0.9910740          6.0408825  5.0915893  3.2484873 -7.8406934
##   3          1.8329003         -5.7207371  4.4823472 -0.3431371 -1.2420612
##
## $pval
##    Xquali
## P   Amer-Indian-Eskimo Asian-Pac-Islander        Black        Other
##   1       1.749006e-01       3.731217e-01 0.000000e+00 5.849668e-03
##   2       1.608247e-01       7.663678e-10 1.775373e-07 5.801019e-04
##   3       3.340870e-02       5.303145e-09 3.691325e-06 3.657477e-01
##    Xquali
## P          White
##   1 2.210710e-16
##   2 2.220446e-15
##   3 1.071070e-01
##
## [1] "Variable sex"
```

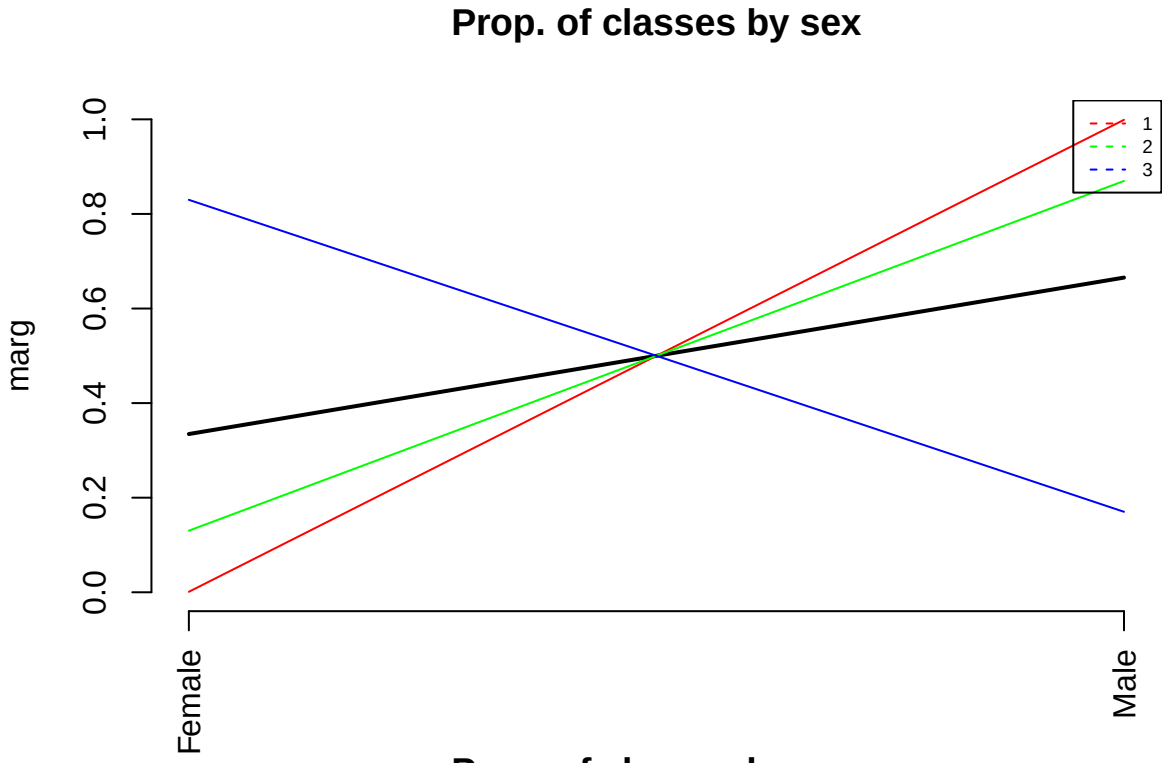
```
## [1] "Categories=" "Female"      "Male"
```

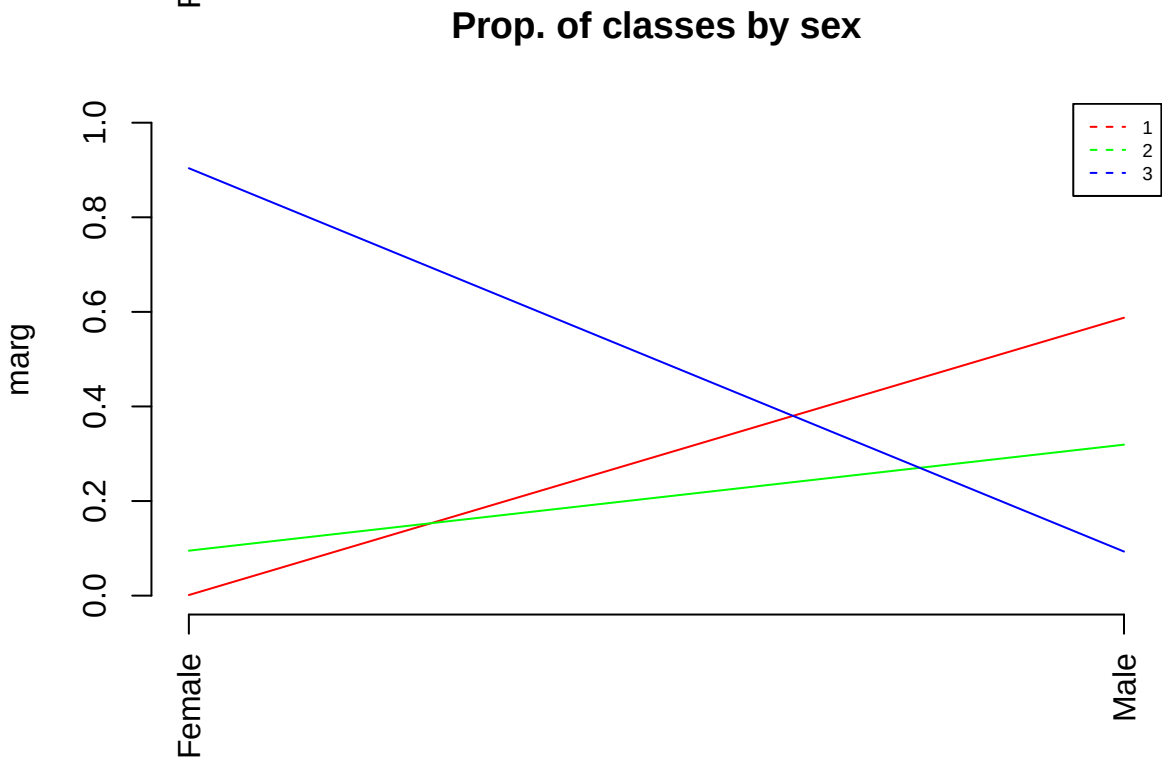



```
## [1] "Categories=" "Female"      "Male"
```



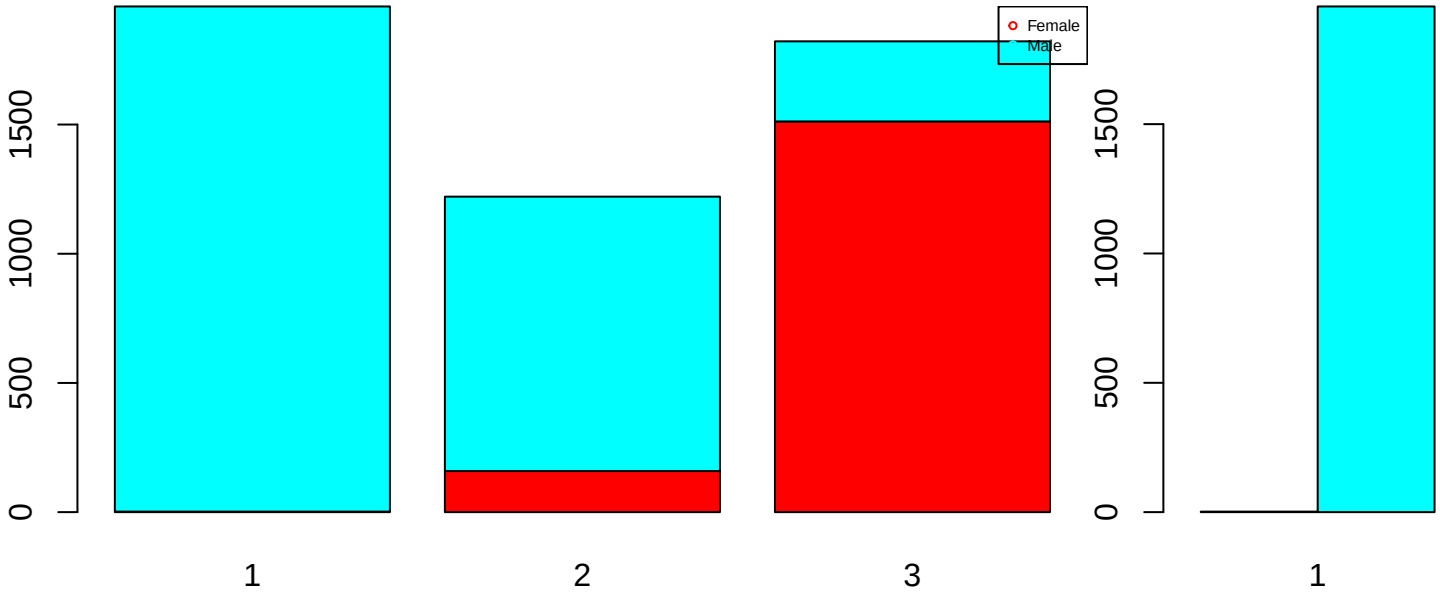

```
## [1] "Categories=" "Female"      "Male"
```





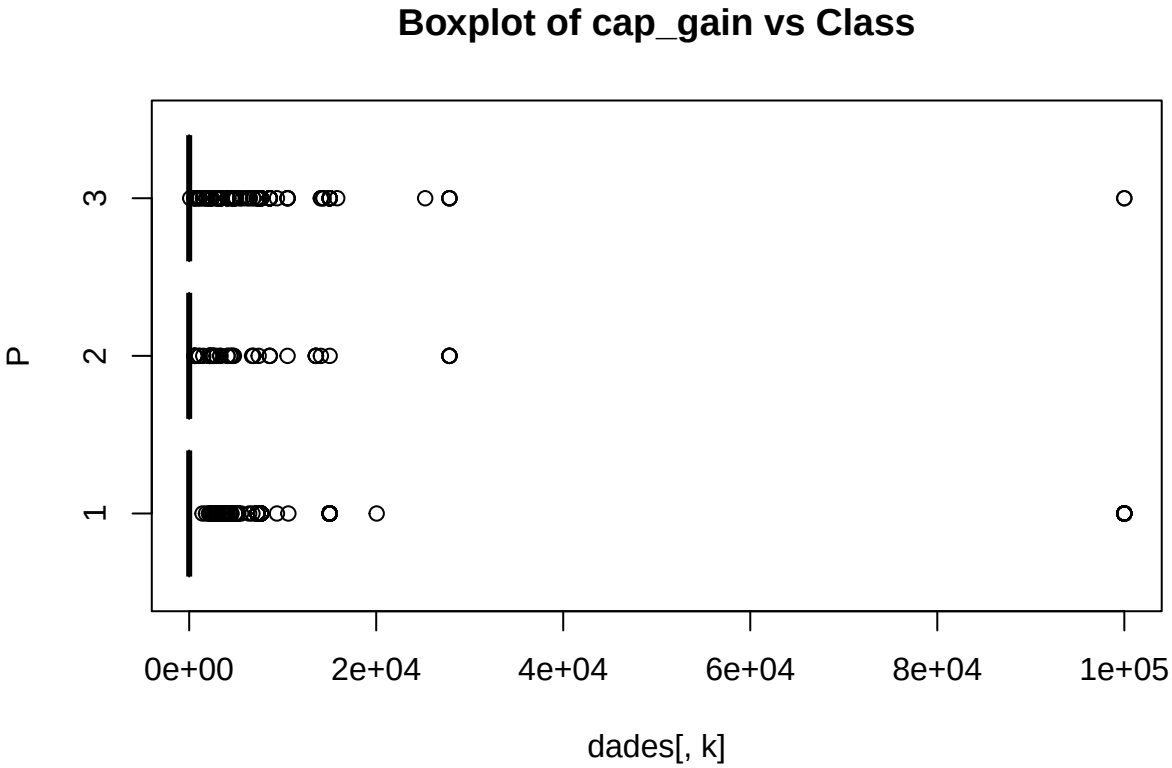

```
## [1] "Cross Table:"
##         P
##             1    2    3
##   Female    2  159 1512
##   Male   1955 1062  310
## [1] "Distribucions condicionades a columnes:"
```

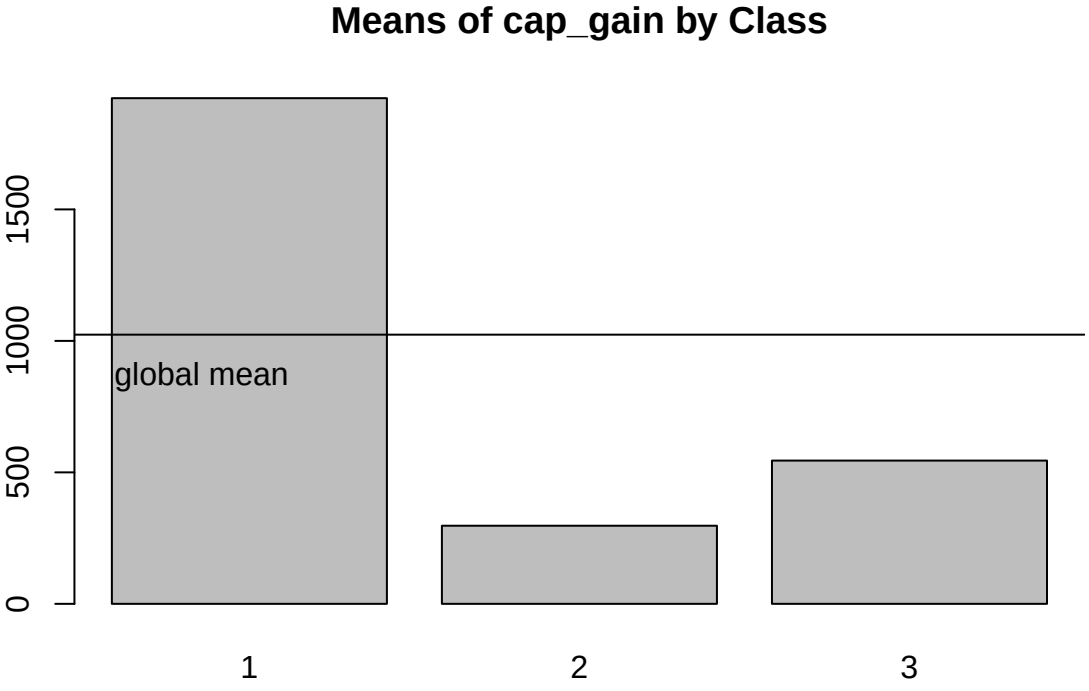
```
##
## P        Female        Male
##   1 0.001195457 0.587616471
##   2 0.095038852 0.319206492
##   3 0.903765690 0.093177036
```




```
## [1] "Test Chi quadrat: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 3214.4, df = 2, p-value < 2.2e-16
##
## [1] "valorsTest:"
## $rowpf
##    Xquali
## P        Female        Male
##   1 0.001021972 0.998978028
##   2 0.130221130 0.869778870
##   3 0.829857300 0.170142700
##
## $vtest
##    Xquali
## P      Female      Male
##   1 -40.08875  40.08875
##   2 -17.40949  17.40949
##   3  56.19643 -56.19643
##
## $pval
##    Xquali
## P         Female         Male
##   1 0.000000e+00 0.000000e+00
##   2 0.000000e+00 3.495277e-68
##   3 0.000000e+00 0.000000e+00
##
```

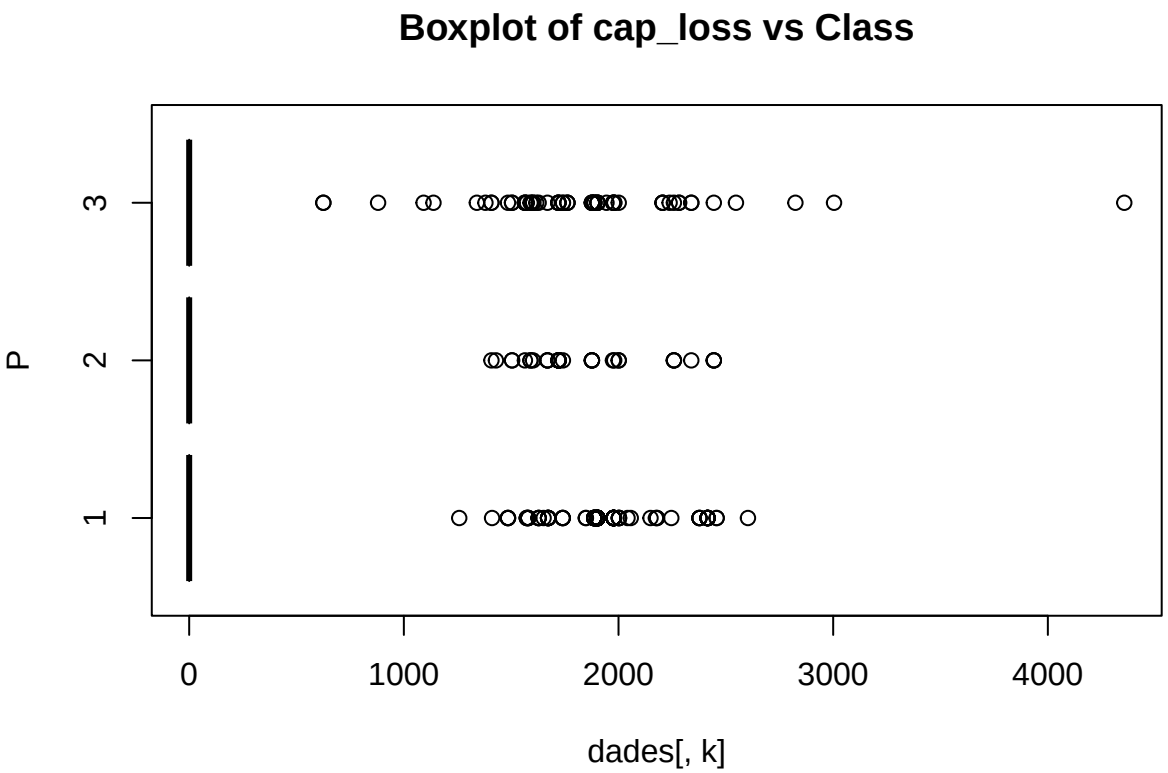
```
## [1] "Anàlisi per classes de la Variable: cap_gain"
```





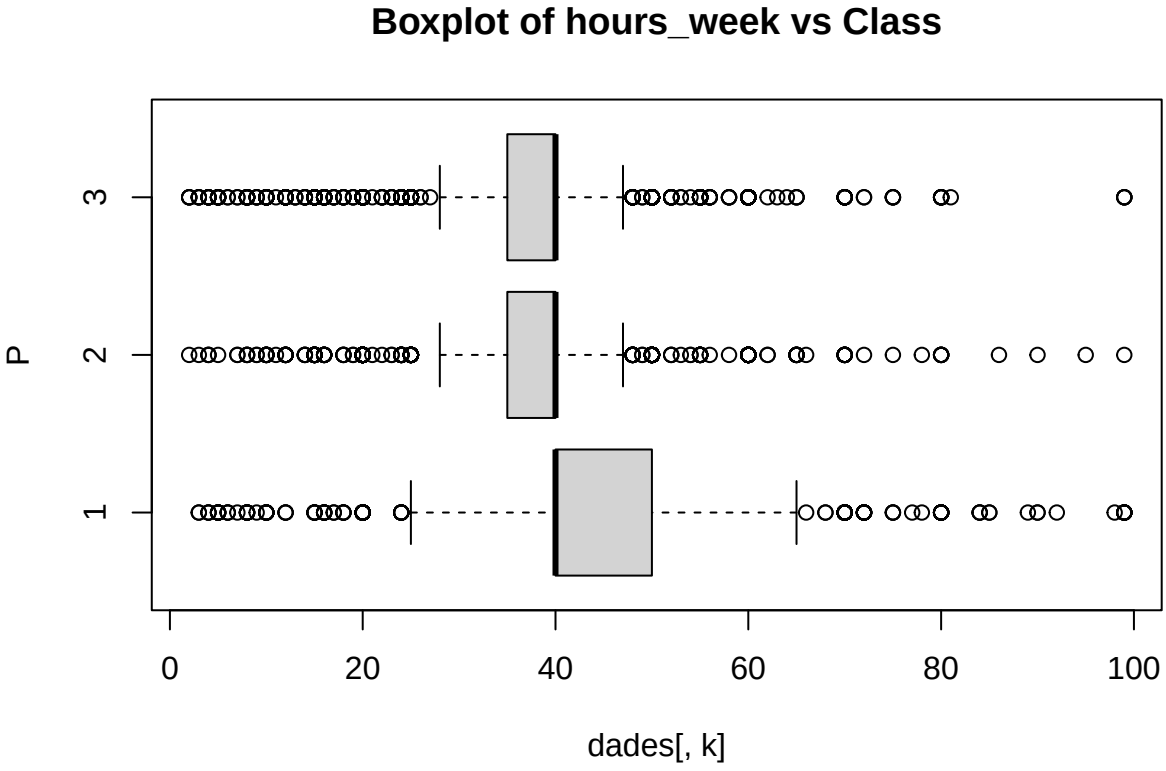

```
## [1] "Estadístics per groups:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1923       0   99999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   297.1     0.0 27828.0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   544.8     0.0 99999.0
```

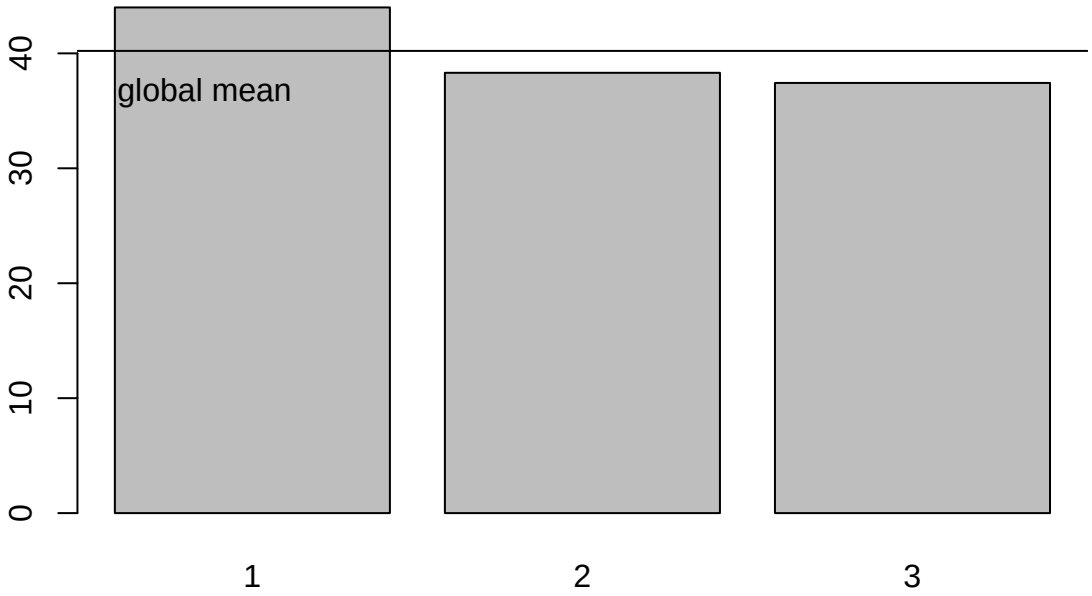
```
## [1] "p-valueANOVA: 2.01817862825246e-10"
## [1] "p-value Kruskal-Wallis: 1.64878128351474e-19"
## [1] "p-values ValorsTest: "
## [1] 6.363026e-13 2.332985e-05 1.746724e-04
## [1] "Anàlisi per classes de la Variable: cap_loss"
```



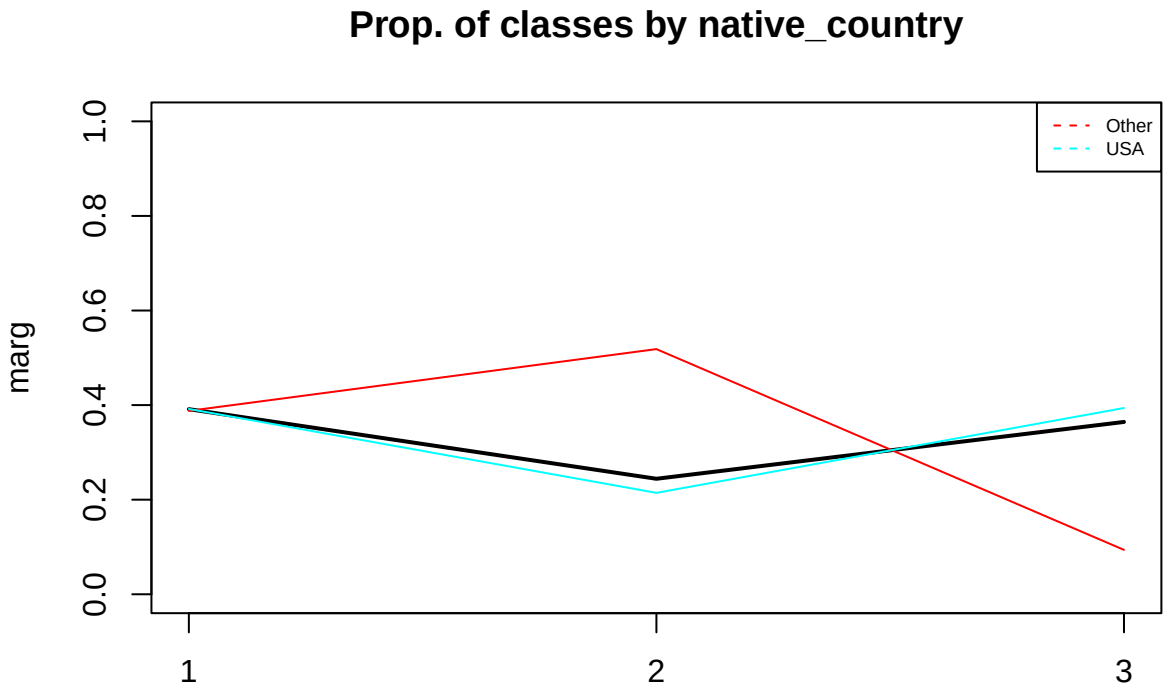

```
## [1] "Estadístics per groups:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   138.2     0.0  2603.0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   54.33    0.00 2444.00
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   65.66    0.00 4356.00
## [1] "p-valueANOVA: 9.61265012137871e-09"
## [1] "p-value Kruskal-Wallis: 2.10298545751866e-09"
## [1] "p-values ValorsTest: "
## [1] 6.588531e-11 1.635098e-04 4.530013e-04
## [1] "Anàlisi per classes de la Variable: hours_week"
```

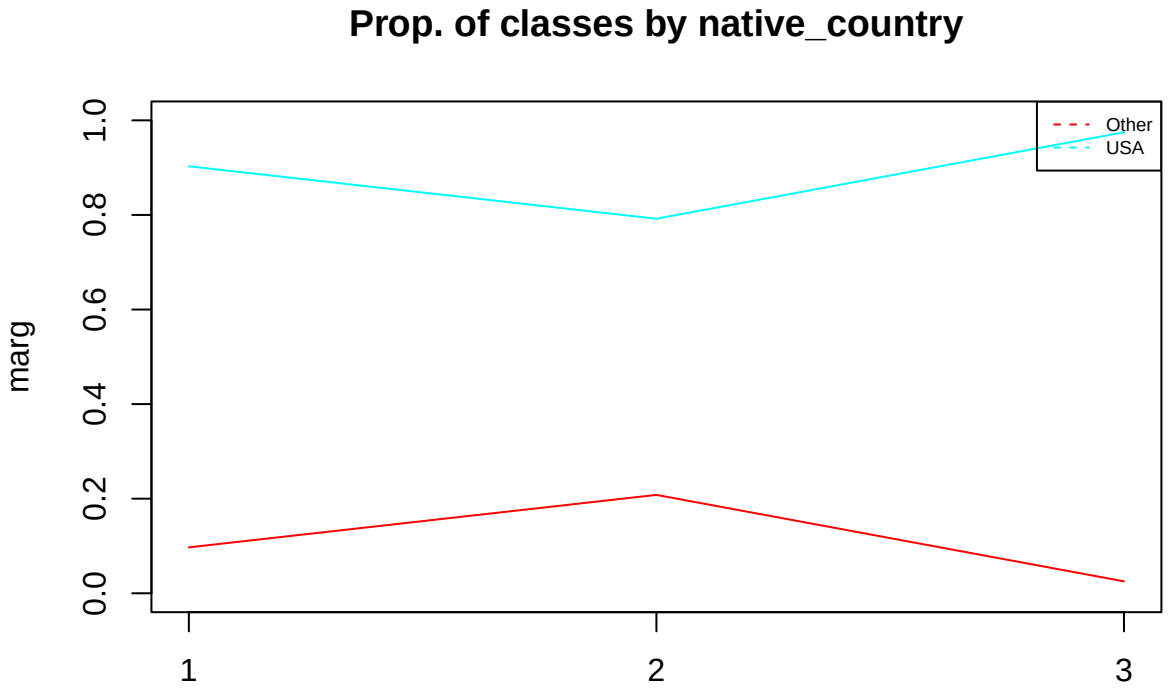




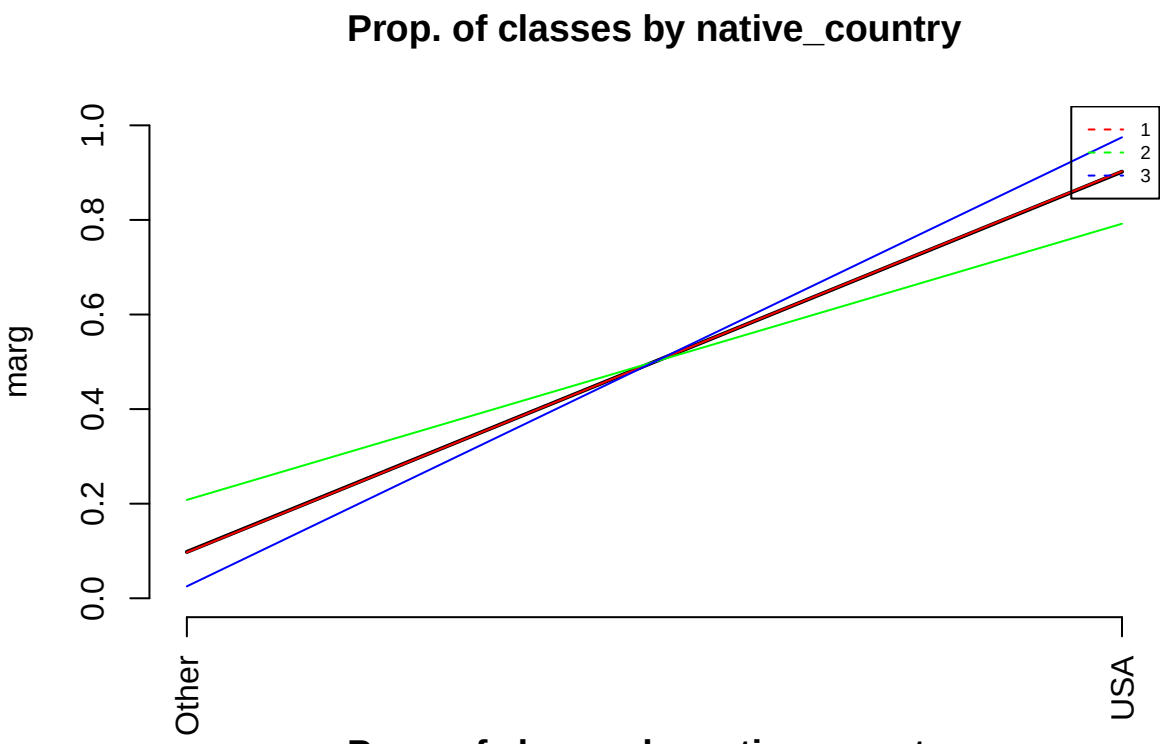

```
## [1] "Estadístics per groups:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3      40      40      44      50      99
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   35.00   40.00   38.31   40.00   99.00
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   35.00   40.00   37.43   40.00   99.00
## [1] "p-valueANOVA: 1.14181725805632e-66"
## [1] "p-value Kruskal-Wallis: 6.19411685035705e-85"
## [1] "p-values ValorsTest: "
## [1] 6.124643e-64 5.816605e-10 0.000000e+00
## [1] "Variable native_country"
## [1] "Categories=" "Other"       "USA"
```

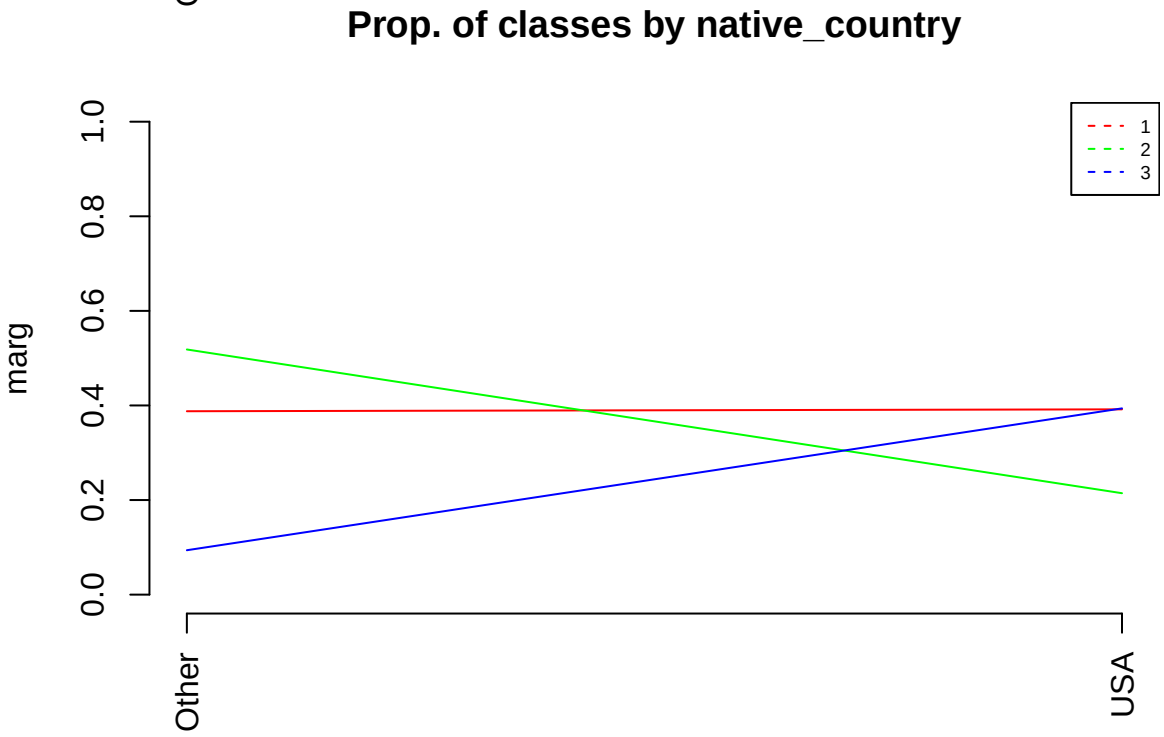



```
## [1] "Categories=" "Other"       "USA"
```



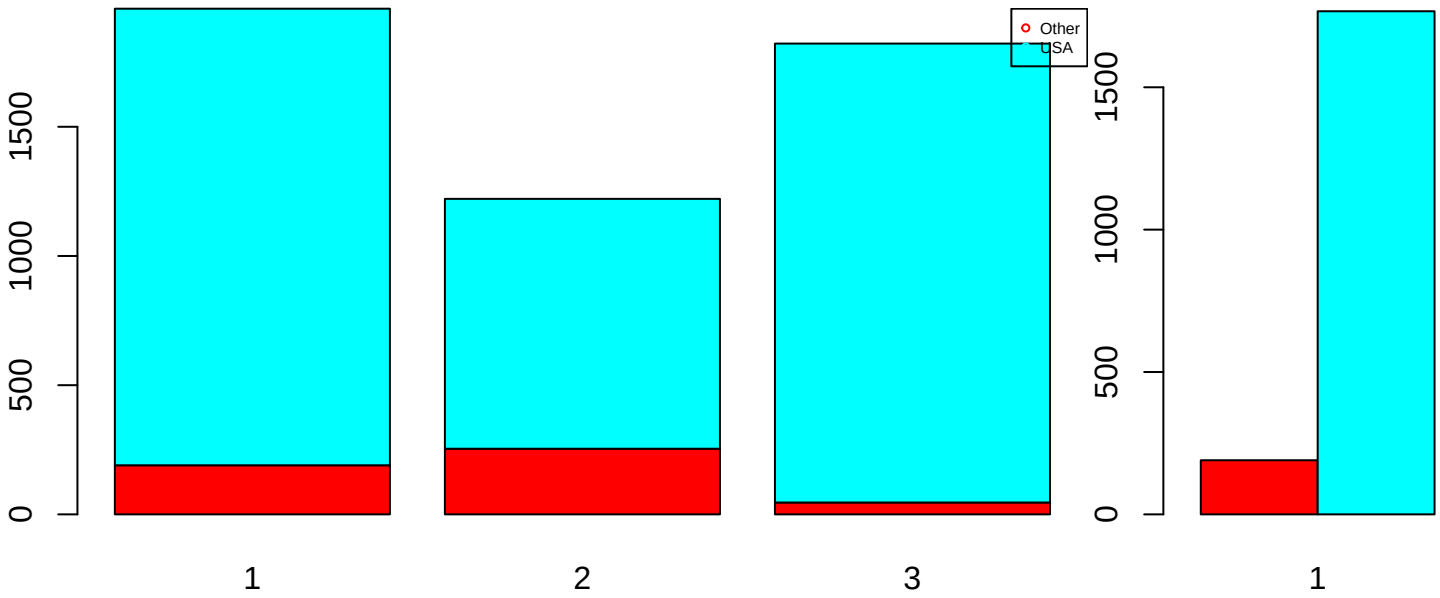

```
## [1] "Categories=" "Other"       "USA"
```



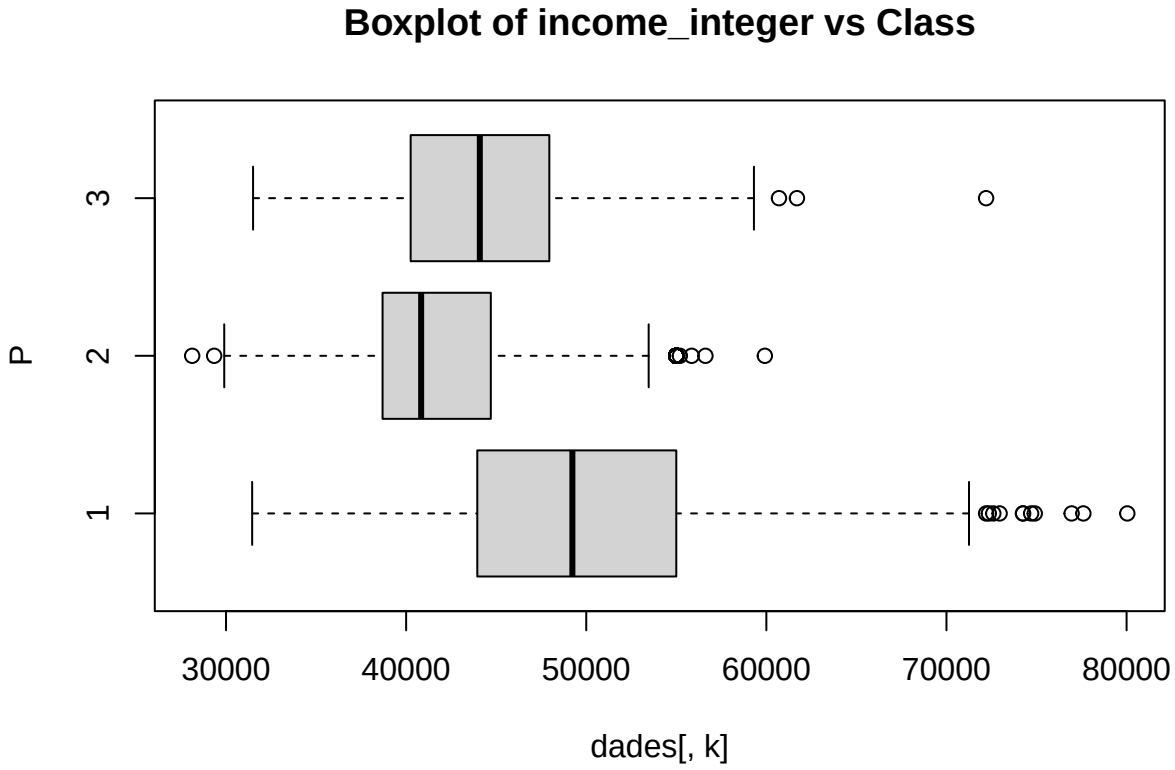



```
## [1] "Cross Table:"
##         P
##             1    2    3
##   Other   190  254   46
##   USA    1767  967 1776
## [1] "Distribucions condicionades a columnes:"
##
## P         Other        USA
```

```
##   1 0.38775510 0.39179601
##   2 0.51836735 0.21441242
##   3 0.09387755 0.39379157
```




```
## [1] "Test Chi quadrat: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 276.33, df = 2, p-value < 2.2e-16
##
## [1] "valorsTest:"
## $rowpf
##    Xquali
## P         Other        USA
##   1 0.09708738 0.90291262
##   2 0.20802621 0.79197379
##   3 0.02524698 0.97475302
##
## $vtest
##    Xquali
## P          Other         USA
##   1  -0.1740619   0.1740619
##   2  14.8742209 -14.8742209
##   3 -13.1013773  13.1013773
##
## $pval
##    Xquali
## P           Other           USA
##   1 4.309084e-01 4.309084e-01
##   2 2.422967e-50 0.000000e+00
##   3 0.000000e+00 1.616682e-39
##
## [1] "Anàlisi per classes de la Variable: income_integer"
```




```
## [1] "Estadístics per groups:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31450   43950   49230   48955   55000   80040
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28120   38690   40833   41777   44700   59916
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31500   40250   44090   44486   47950   72200
## [1] "p-valueANOVA: 2.33607932990687e-236"
```

```
## [1] "p-value Kruskal-Wallis: 8.99199691151354e-225"
## [1] "p-values ValorsTest: "
## [1] 5.076755e-183  0.000000e+00  0.000000e+00
```

```
#descriptors de les classes més significatius. Afegir info qualits
for (c in 1:length(levels(as.factor(P)))) {
  if(!is.na(levels(as.factor(P))[c])){
    print(paste("P.values per class:",levels(as.factor(P))[c]));
    print(sort(pvalk[c,]), digits=3)
  }
}
```

```
## [1] "P.values per class: 1"
##      workclass        marital     occupation   relationship           race
##       0.00e+00       0.00e+00       0.00e+00       0.00e+00       0.00e+00
##            sex native_country income_integer            age     hours_week
##       0.00e+00       0.00e+00      5.08e-183      9.36e-115       6.12e-64
##       cap_gain       cap_loss        edu_num
##       6.36e-13       6.59e-11       8.90e-07
## [1] "P.values per class: 2"
##            age      workclass        marital     occupation   relationship
##       0.00e+00       0.00e+00       0.00e+00       0.00e+00       0.00e+00
##           race            sex native_country income_integer        edu_num
##       0.00e+00       0.00e+00       0.00e+00       0.00e+00       2.29e-13
##     hours_week       cap_gain       cap_loss
##       5.82e-10       2.33e-05       1.64e-04
## [1] "P.values per class: 3"
##      workclass        marital     occupation   relationship           race
##       0.000000       0.000000       0.000000       0.000000       0.000000
##            sex     hours_week native_country income_integer       cap_gain
##       0.000000       0.000000       0.000000       0.000000       0.000175
##            age       cap_loss        edu_num
##       0.000427       0.000453       0.051825
```