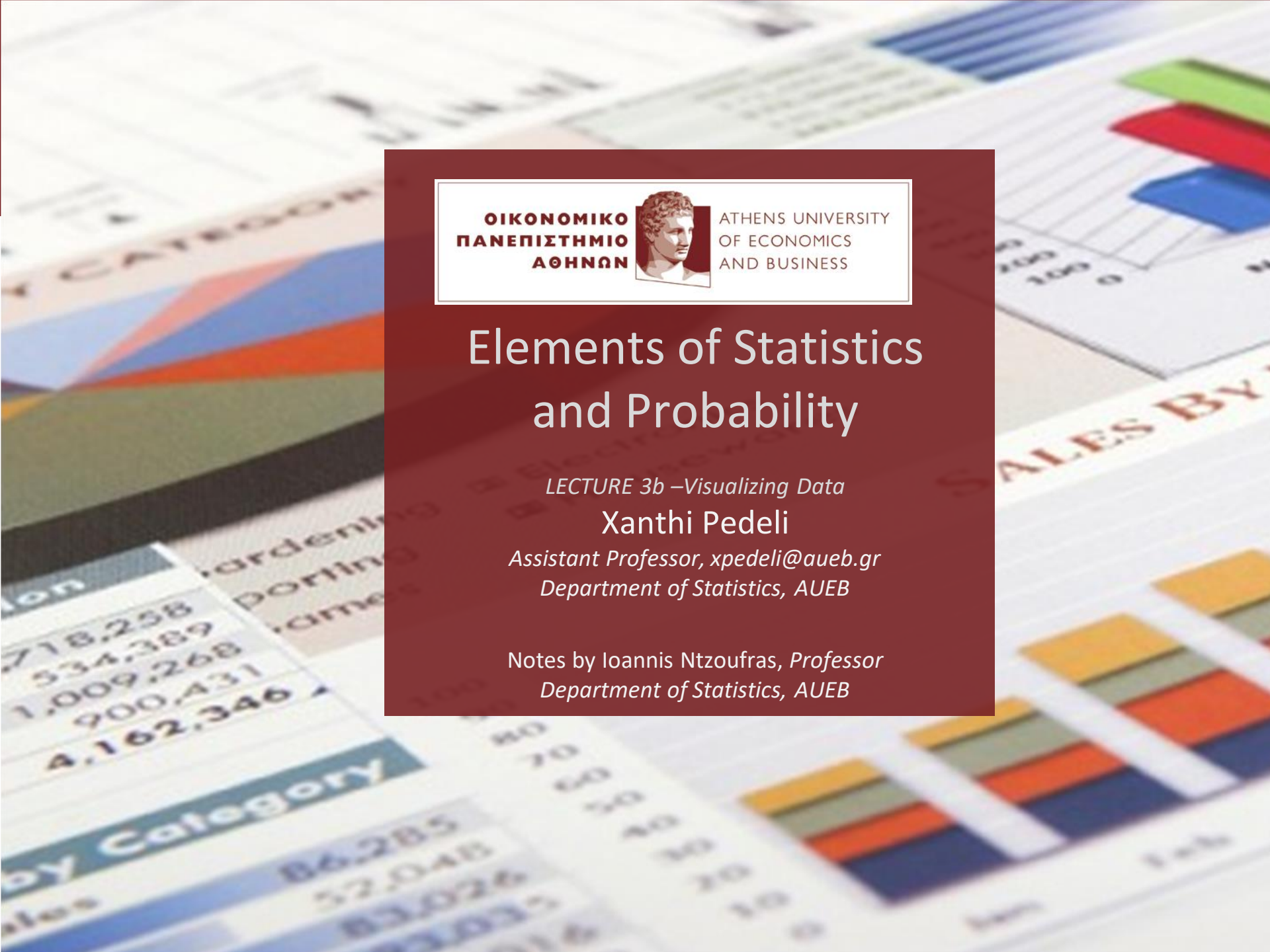


ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Elements of Statistics and Probability

LECTURE 3b –Visualizing Data

Xanthi Pedeli

Assistant Professor, xpedeli@aueb.gr

Department of Statistics, AUEB

Notes by Ioannis Ntzoufras, *Professor*

Department of Statistics, AUEB

3. Data Visualization

- One variable
  - Categorical variable
  - Quantitative variable– Distribution fitting and shape
- Association between two variables
  - Two quantitative
  - One quantitative and one categorical
  - Two categorical

3. Data Visualization

- Many covariates
  - Three quantitative variables (3D scatterplot)
  - Two quantitative variables over a categorical
  - Visualization of multiple quantitative covariates using matrix scatterplots
  - Finding patterns and clusters using faces and stars
- Exploratory data analysis (EDA)

3. Data Visualization

Type of Graph	Type of variable	Graphical representation
One dimension (variable)	Categorical	Bar chart (barplot) Pie chart (pie)
	Quantitative	Histogram (hist) Box-whiskers plot/box-plot (boxplot) qqplot/ pplot
Two dimensions (variables)	Two categorical	Clustered Barcharts Stacked barcharts
	One quantitative & one categorical	Box plots Error bars
	Two quantitative	Scatter plot (plot)
Multidimensional or Multivariate	Quantitative	Matrix scatter-plot Star plots Chernoff faces

3. Data Visualization

3.1. One Categorical variable

- Bar chart (barplot)
- Pie chart (pie)

3.1. One Categorical variable
Bar chart

<u>Salary dataset from Norusis</u>

This is a data file containing information on 474 employees hired by a Midwestern bank between 1969 and 1971.

It was created for an Equal Employment Opportunity (EEO) court case involving wage discrimination.

3.1. One Categorical variable
Bar chart

ΟΠΑ AUEB

Salary dataset from Norusis

ID	Employee code
SALBEG	Beginning salary
SEX	Gender of employee (0=Female; 1=Male)
TIME	Job Seniority (months)
AGE	Age of Employee (years and fraction)
SALNOW	Current salary
EDLEVEL	Educational Level (years)
WORK	Work Experience (years x 100)
JOBCAT	Employment category (1 Clerical, 2 Office trainee, 3 Security officer, 4 College trainee, 5 Exempt employee, 6 MBA trainee 7 Technical)
MINORITY	Minority Classification (0=White; 1=Nonwhite)

3.1. One Categorical variable
Bar chart

ΟΠΑ AUEB

Dataset salary from Norusis

```
library(foreign)
salary<-read.spss('salary.sav', to.data.frame=T)
names(salary)
```

```
> names(salary)
 [1] "id"       "salbeg"   "sex"      "time"     "age"      "salnow"
 [7] "edlevel"  "work"     "jobcat"   "minority" "sexrace"
>
```

```
head(salary)
```

3.1. One Categorical variable
Bar chart

ΟΠΑ AUEB

Salary dataset from Norusis

RGui (64-bit) - [R Console]

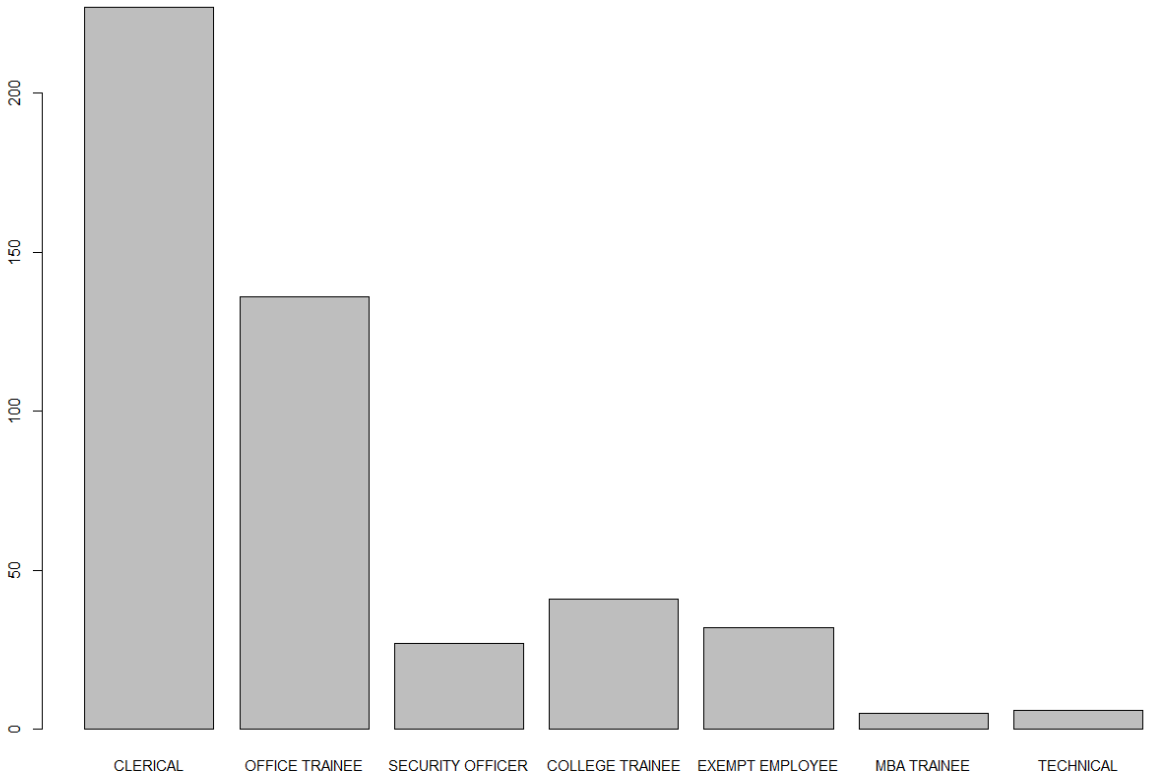
File Edit View Misc Packages Windows Help

```
> head(salary)
  id salbeg   sex time   age salnow edlevel  work          jobcat minority     sexrace
1  1   8400 MALES   81 28.50  16080      16  0.25 COLLEGE TRAINEE    WHITE WHITE MALES
2  2  24000 MALES   73 40.33  41400      16 12.50 EXEMPT EMPLOYEE    WHITE WHITE MALES
3  3  10200 MALES   83 31.08  21960      15  4.08 EXEMPT EMPLOYEE    WHITE WHITE MALES
4  4   8700 MALES   93 31.17  19200      16  1.83 COLLEGE TRAINEE    WHITE WHITE MALES
5  5  17400 MALES   83 41.92  28350      19 13.00 EXEMPT EMPLOYEE    WHITE WHITE MALES
6  6  12996 MALES   80 29.50  27250      18  2.42 COLLEGE TRAINEE    WHITE WHITE MALES
> 
```

3.1. One Categorical variable
Bar chart

Variable – Jobcat

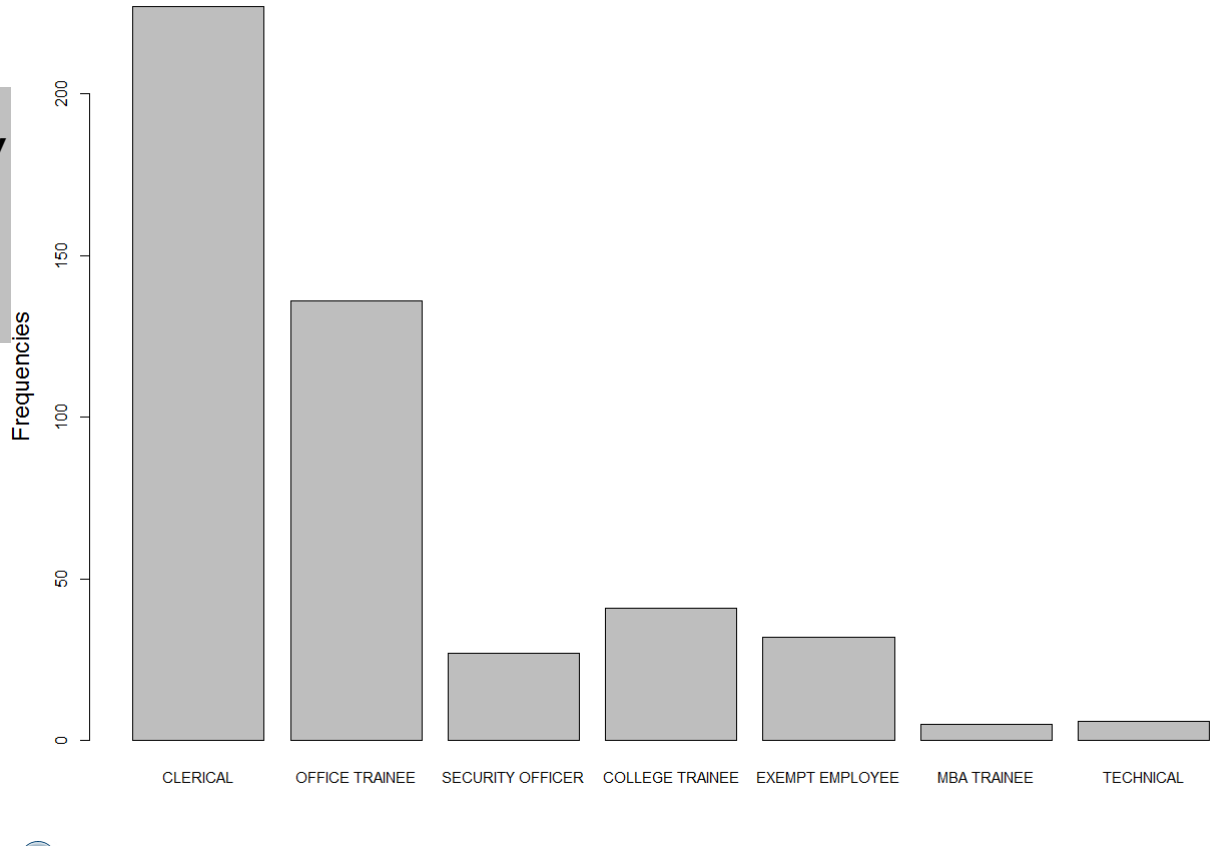
```
barplot(table(salary$jobcat), cex.names=0.9)
```



Adding a label on y axis

Variable – Jobcat

```
barplot(table(salary$jobcat),
cex.names=0.9,
ylab='Frequencies',
cex.lab=1.5 )
```

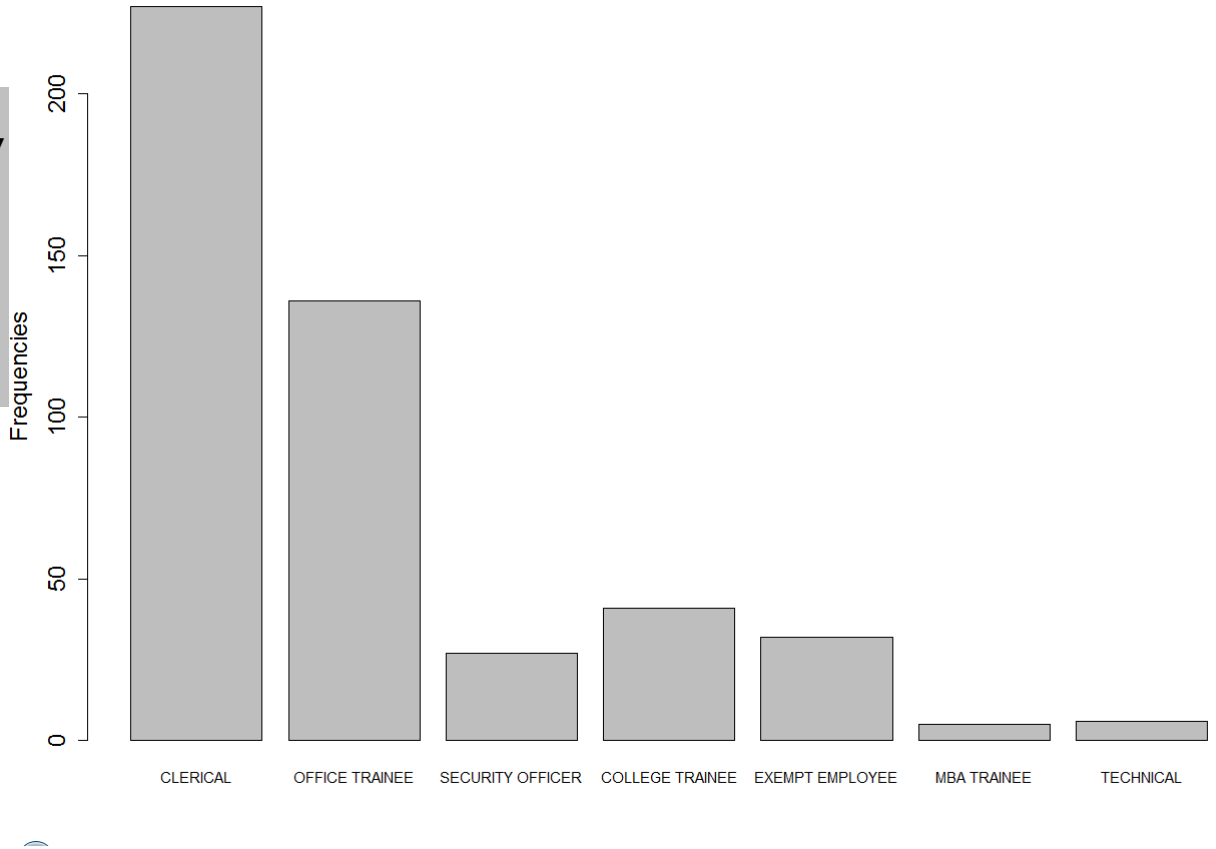


3.1. One Categorical variable
Bar chart

Increasing the size of the y axis numbers

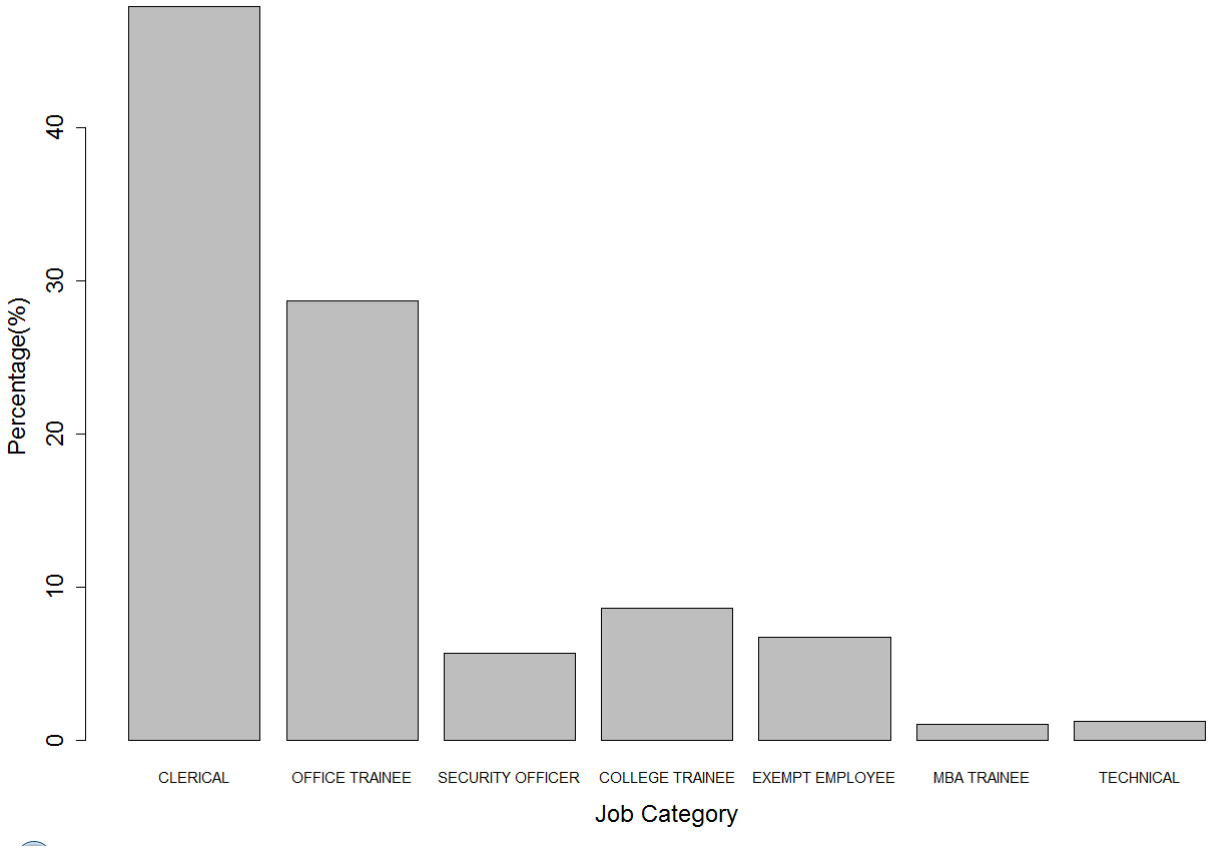
Variable – Jobcat

```
barplot(table(salary$jobcat),
cex.names=0.9,
ylab='Frequencies',
cex.lab=1.5,
cex.axis=1.5)
```



Ploting percentages and an X label

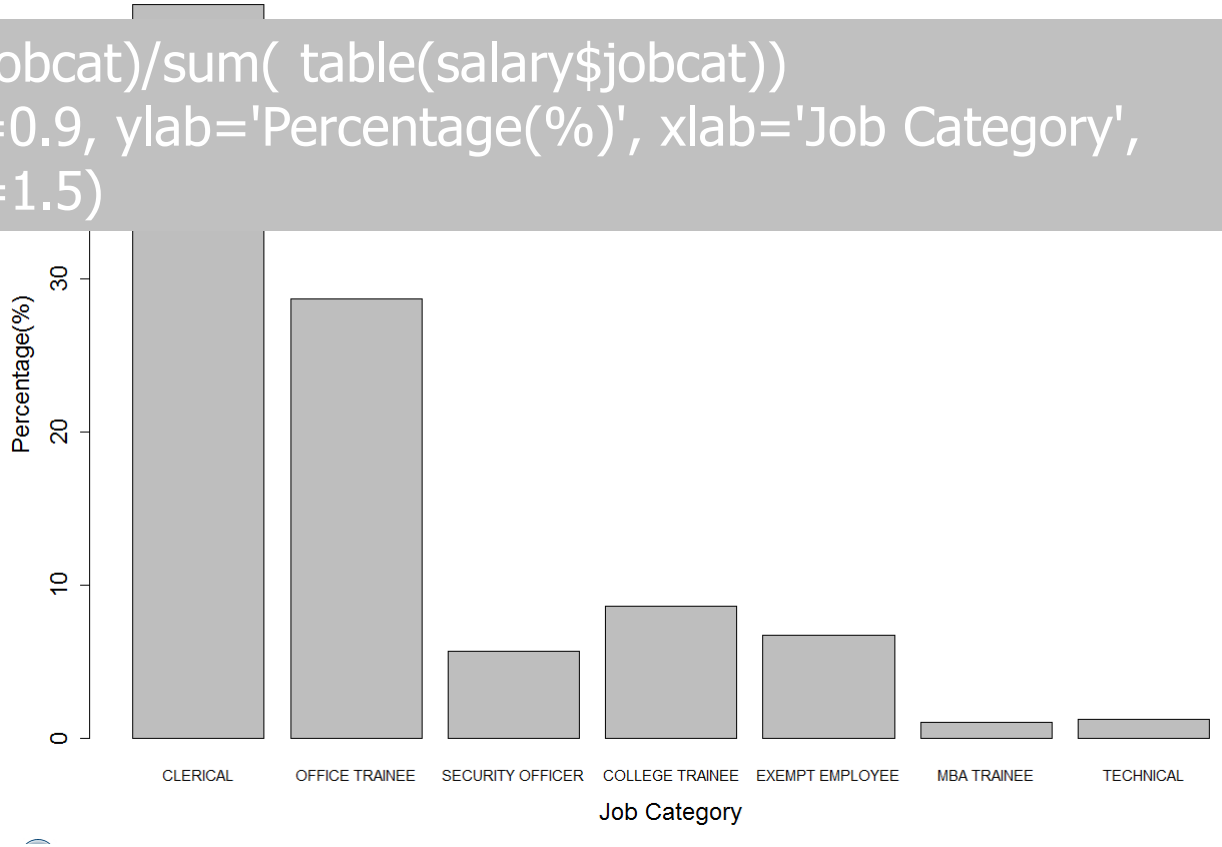
Variable – Jobcat



Ploting percentages and an X label

Variable – Jobcat

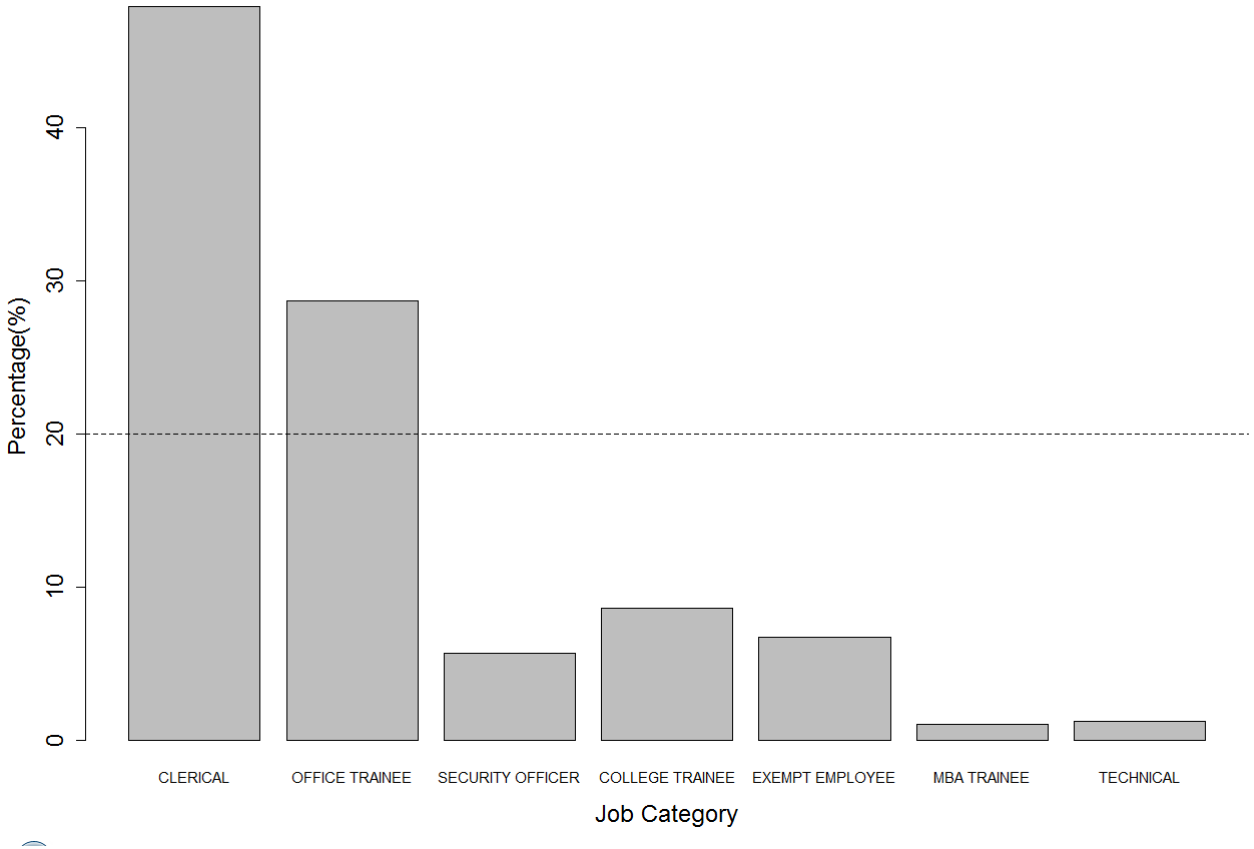
```
x<-100*table(salary$jobcat)/sum( table(salary$jobcat))
barplot(x, cex.names=0.9, ylab='Percentage(%)', xlab='Job Category',
cex.lab=1.5, cex.axis=1.5)
```



Adding a reference line

Variable – Jobcat

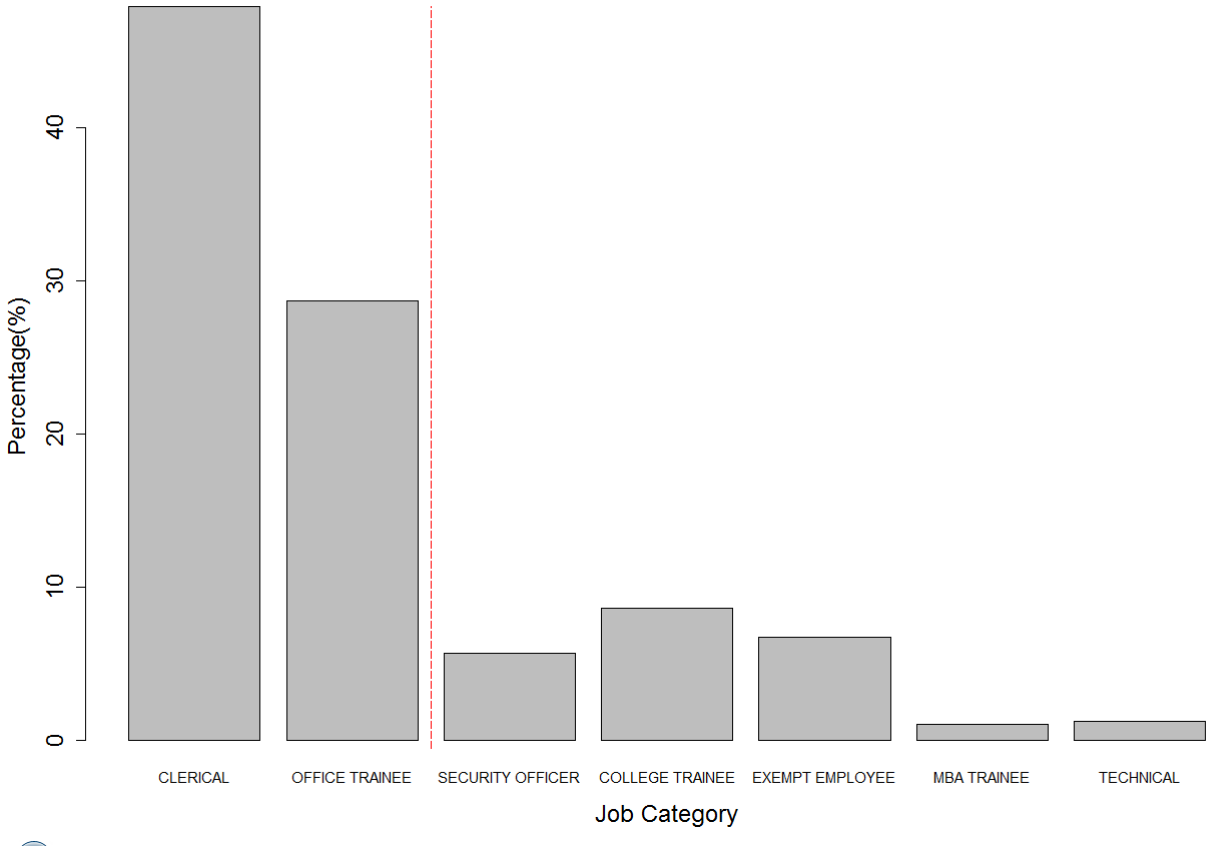
```
abline(h=20, lty=2)
```



Adding a reference line

Variable – Jobcat

```
barplot(…)
abline(v=2.5, lty=5,
col=2)
```

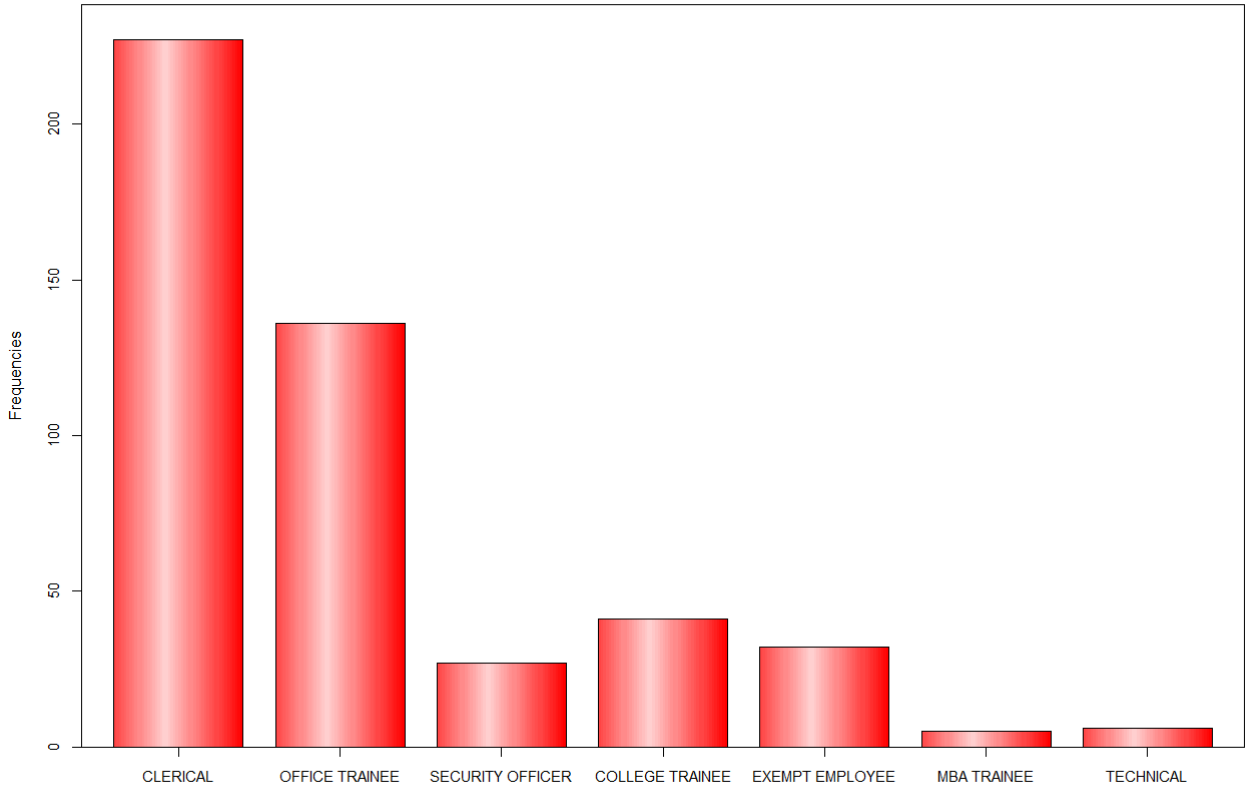


3.1. One Categorical variable
Bar chart

Function Barp in Plotrix library

Variable – Jobcat

```
library(plotrix)
x<-table(salary$jobcat)
barp(x,col=2, cylindrical=T, shadow=F,
names.arg=names(x), cex.axis=0.9)
```



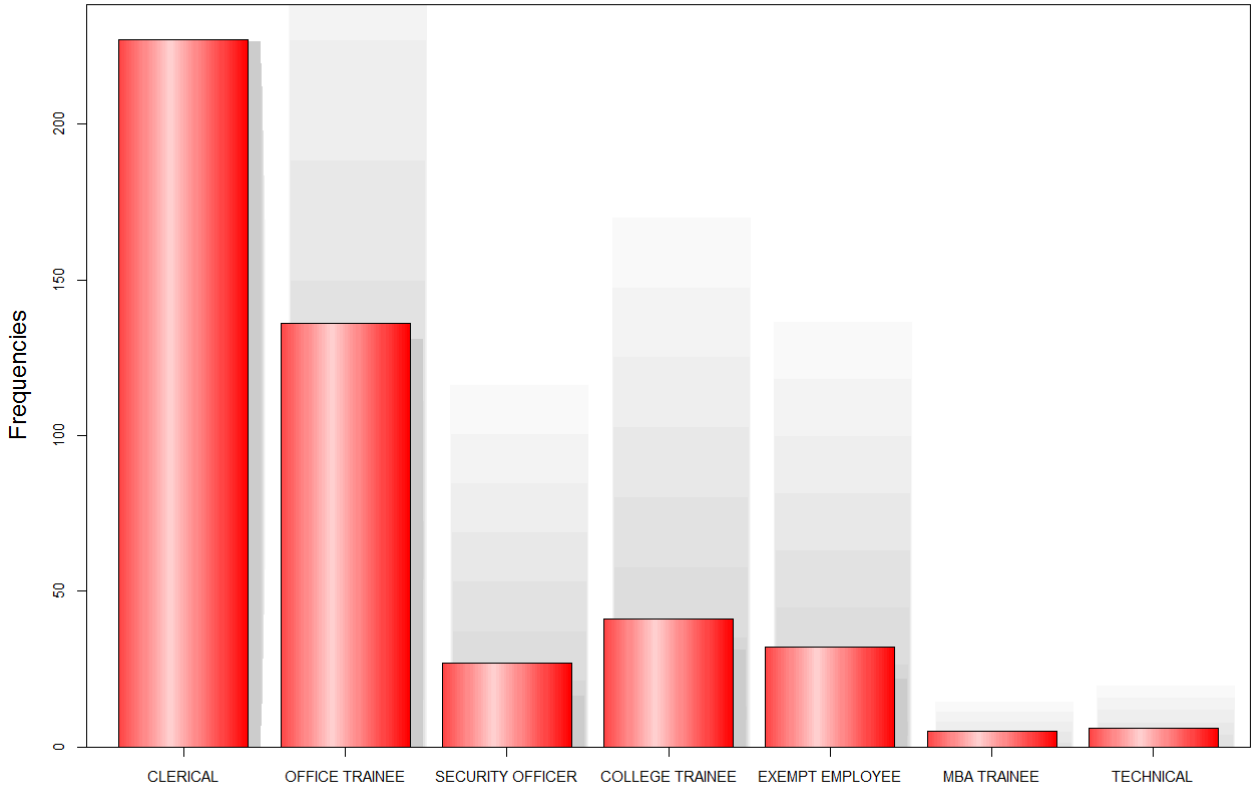
3.1. One Categorical variable
Bar chart

Function Barp in Plotrix library

Variable – Jobcat

Adding a shade effect

```
library(plotrix)
x<-table(salary$jobcat); par(cex.lab=1.5)
barp(x,col=2, cylindrical=T, shadow=T,
names.arg=names(x), cex.axis=0.9)
```

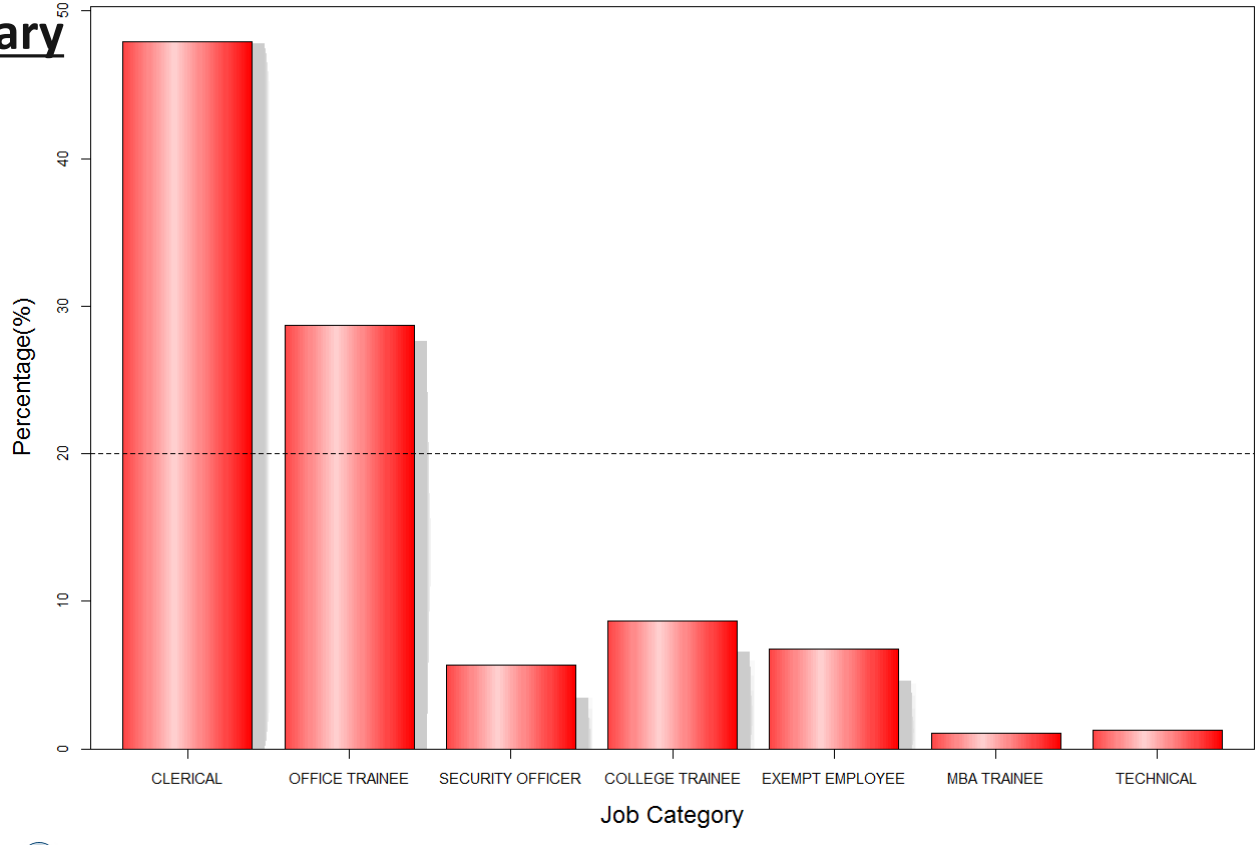


```
x<-100*table(salary$jobcat)/sum( table(salary$jobcat))
par(cex.lab=1.5)
barp(x, col=2, cylindrical=T, shadow=T, names.arg=names(x),
          cex.axis=0.9, ylab='Percentage(%)', xlab='Job Category')
abline(h=20,lty=2)
```

Function Barp in Plotrix library

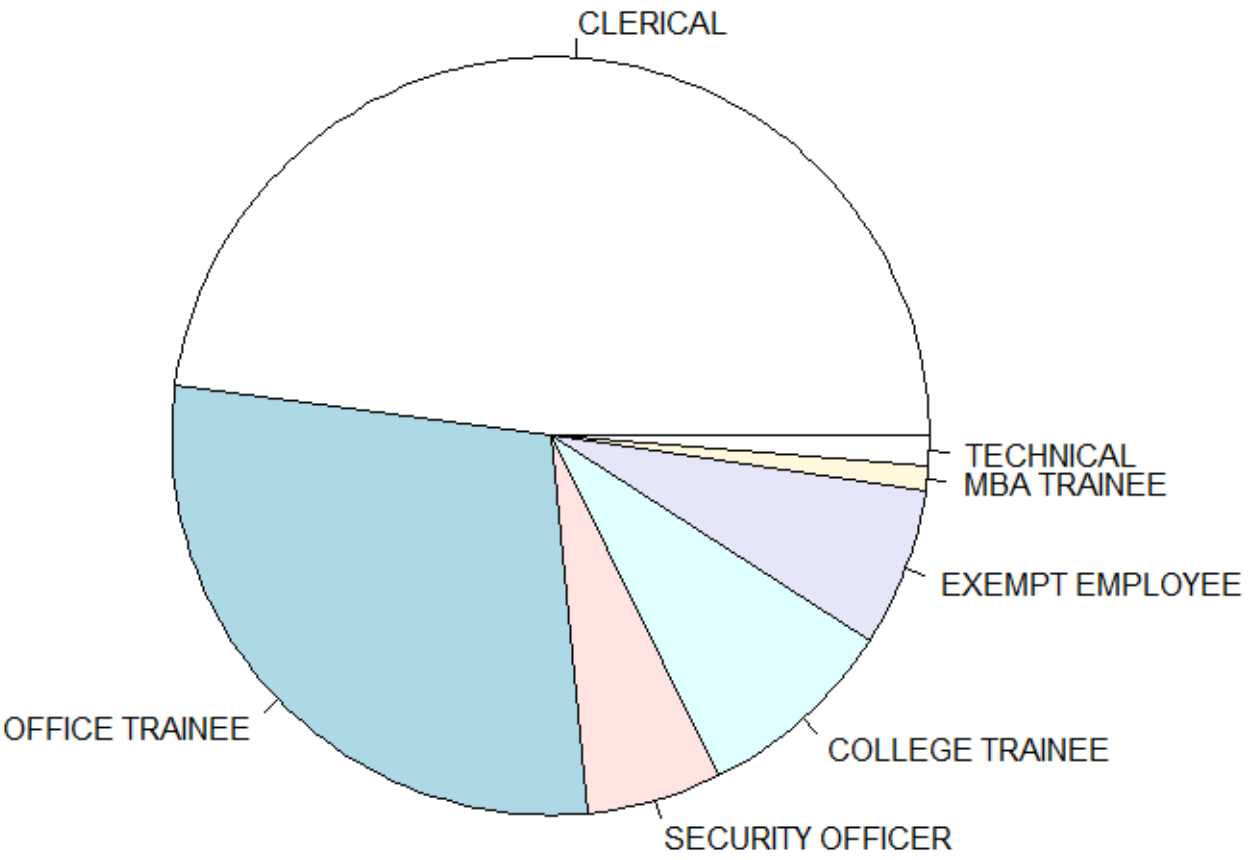
Variable – Jobcat

Adding a labels and a reference line



3.1. One Categorical variable
Pie chart

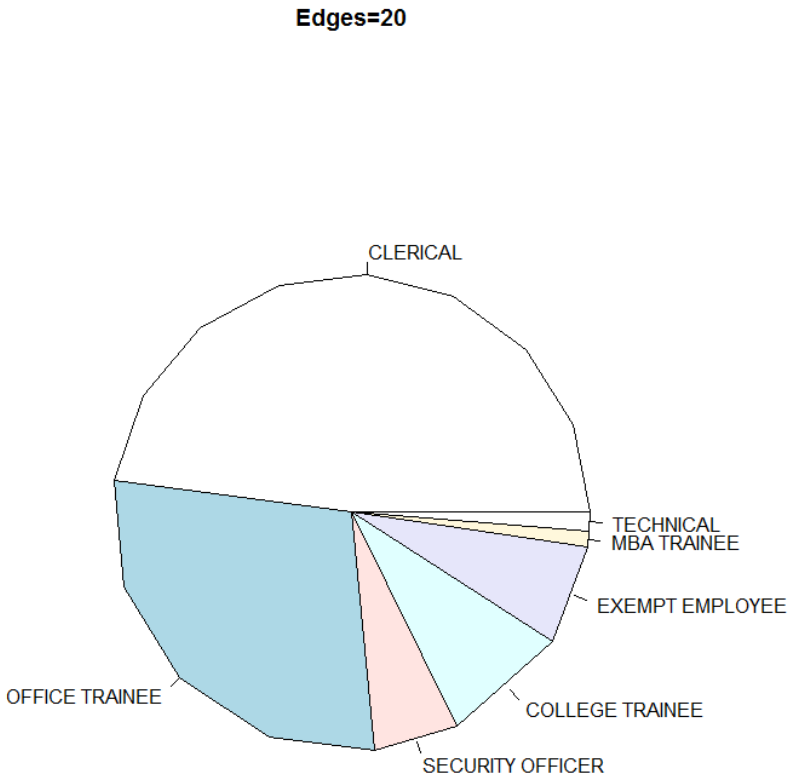
- Salary dataset
- Variable jobcat



Reducing edges

- Salary dataset
- Variable jobcat

```
pie(table(salary$jobcat),
          edges=20,
          main='Edges=20')
```

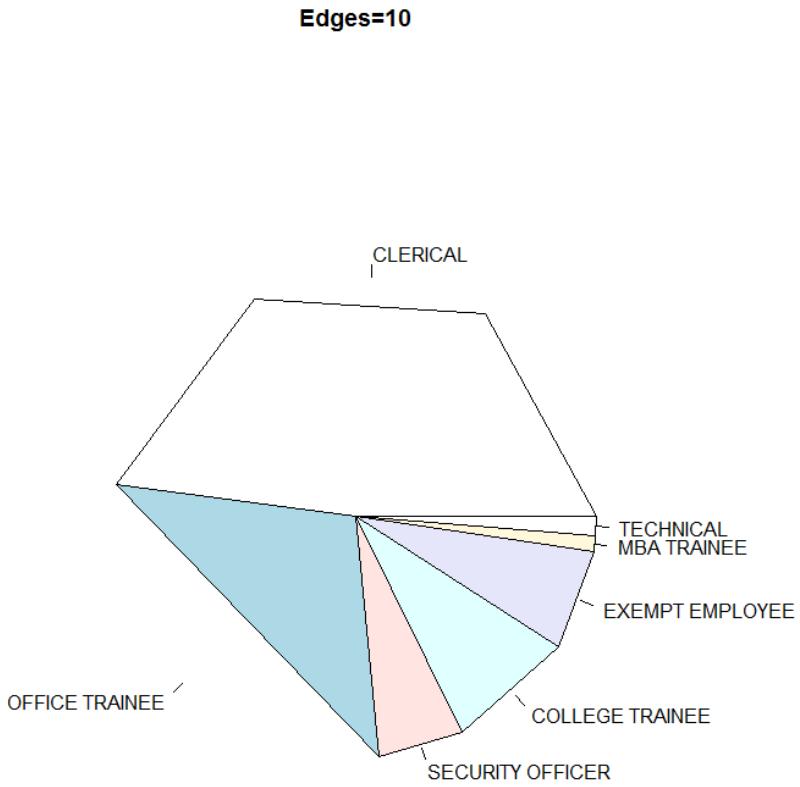


3.1. One Categorical variable
Pie chart

Reducing edges

- Salary dataset
- Variable jobcat

```
pie(table(salary$jobcat),
        edges=10,
        main='Edges=10')
```

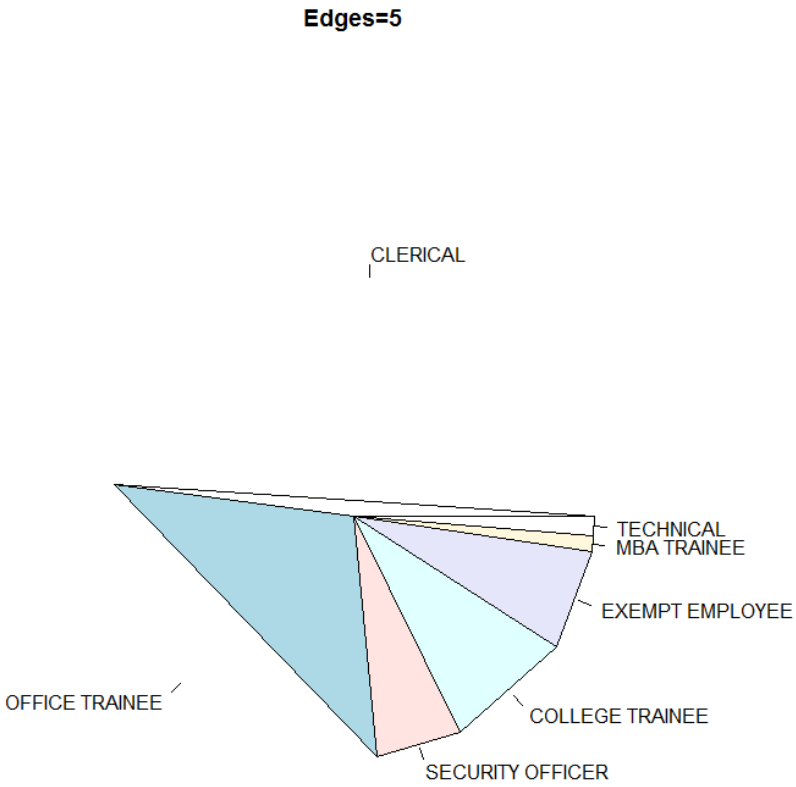


3.1. One Categorical variable
Pie chart

Reducing edges

- Salary dataset
- Variable jobcat

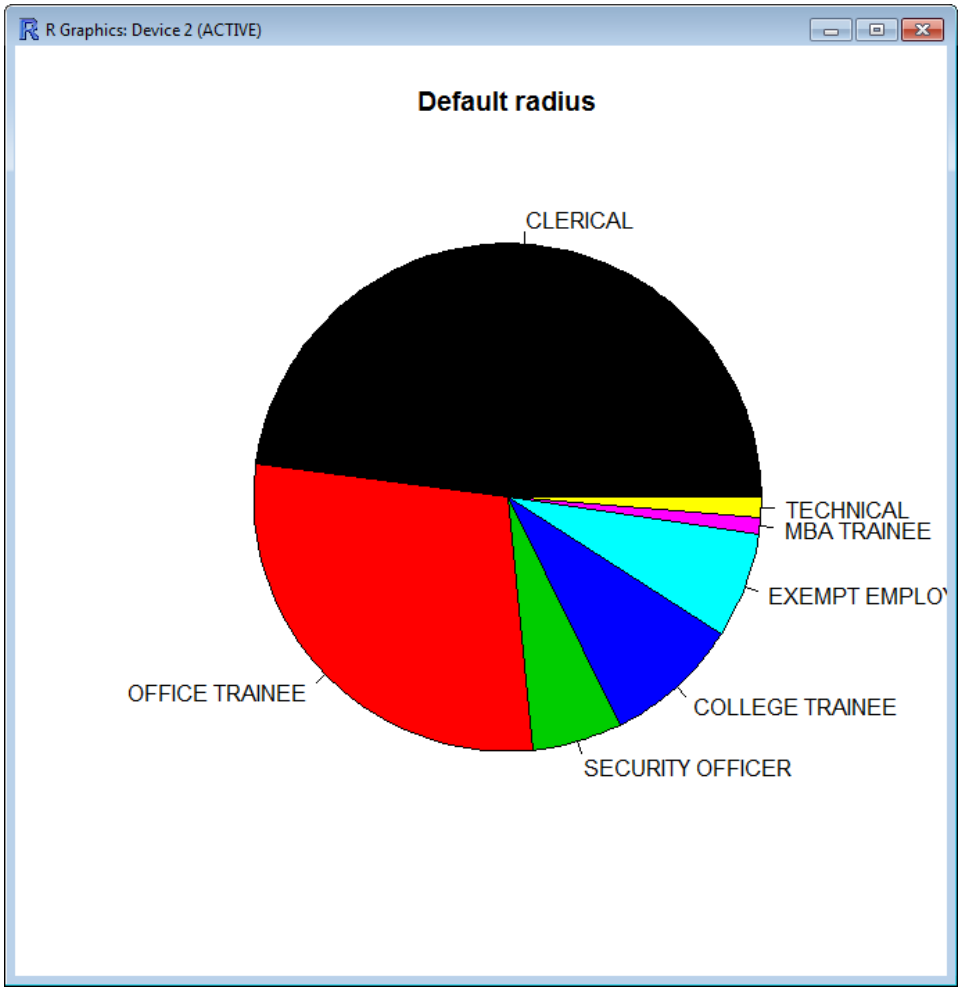
```
pie(table(salary$jobcat),
          edges=5,
          main='Edges=5')
```



Changing colors

- Salary dataset
- Variable jobcat

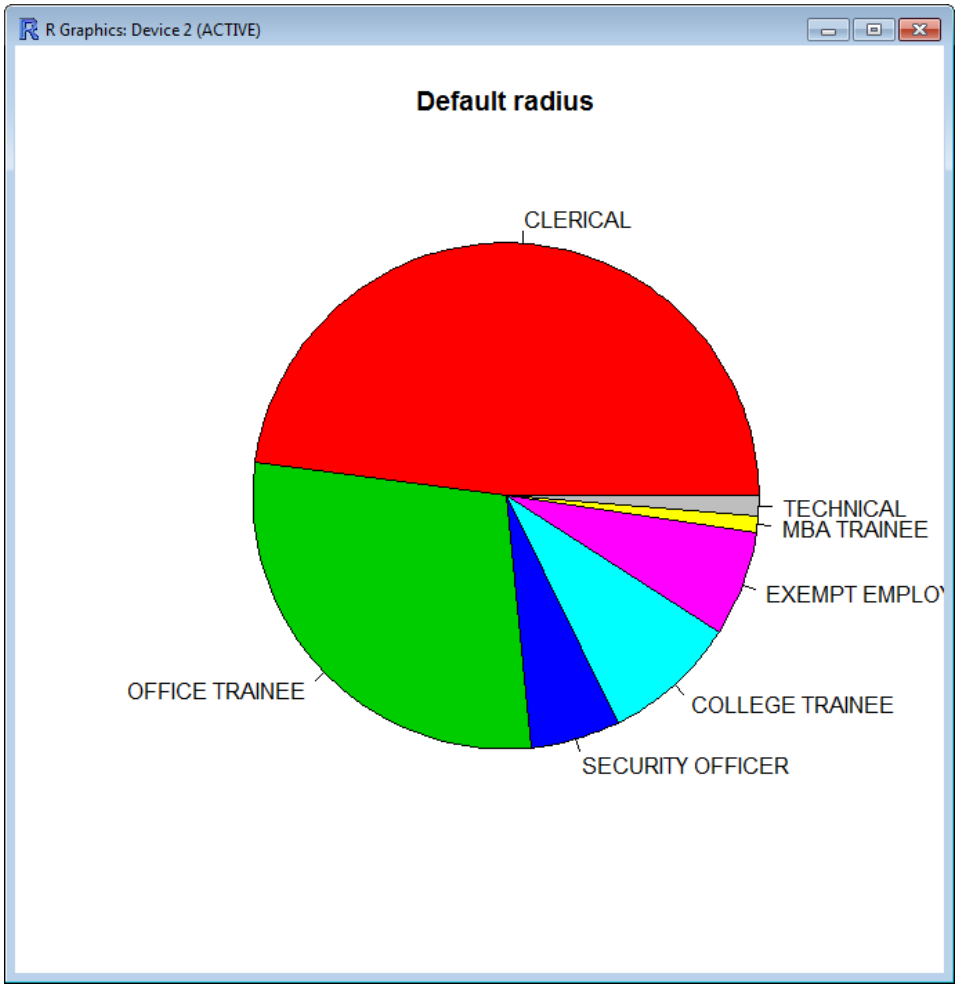
```
pie(table(salary$jobcat),
    col=1:10)
```



Changing colors

- Salary dataset
- Variable jobcat

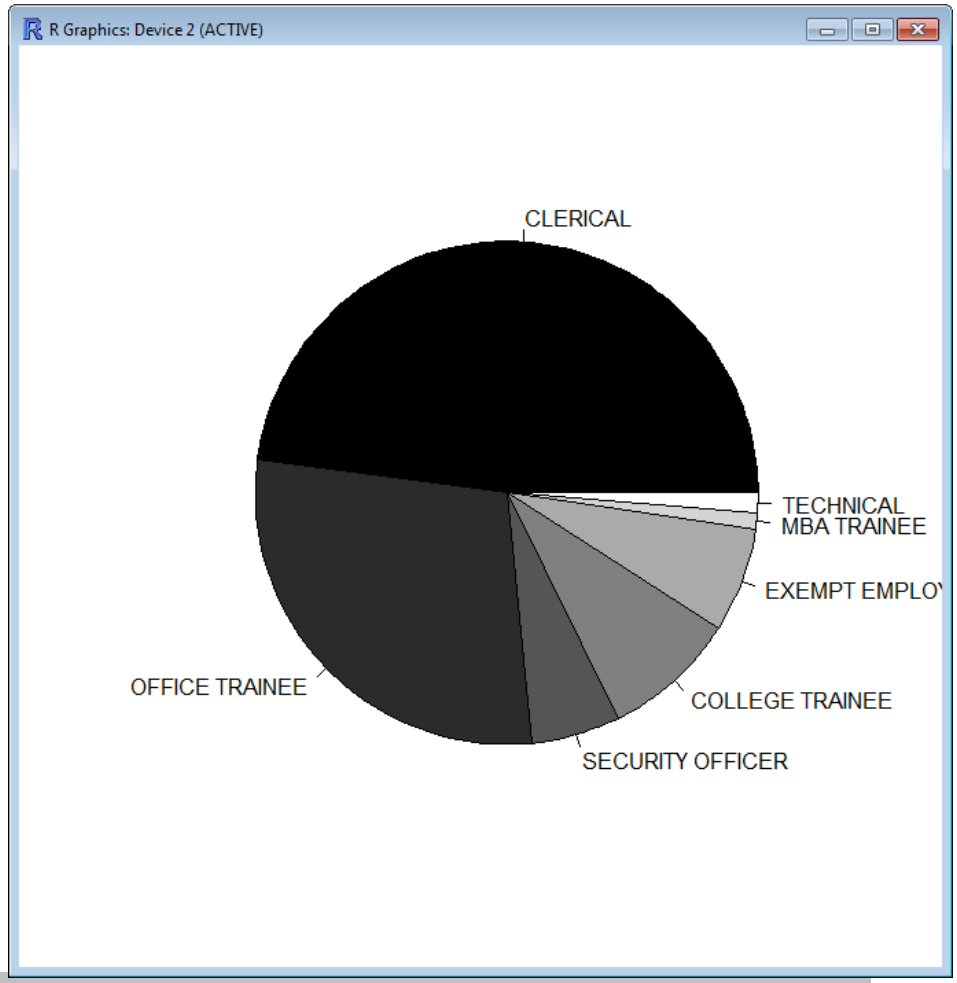
```
pie(table(salary$jobcat),
col=2:10)
```



Using shades of grey

The Grey pallete

- Salary dataset
- Variable jobcat

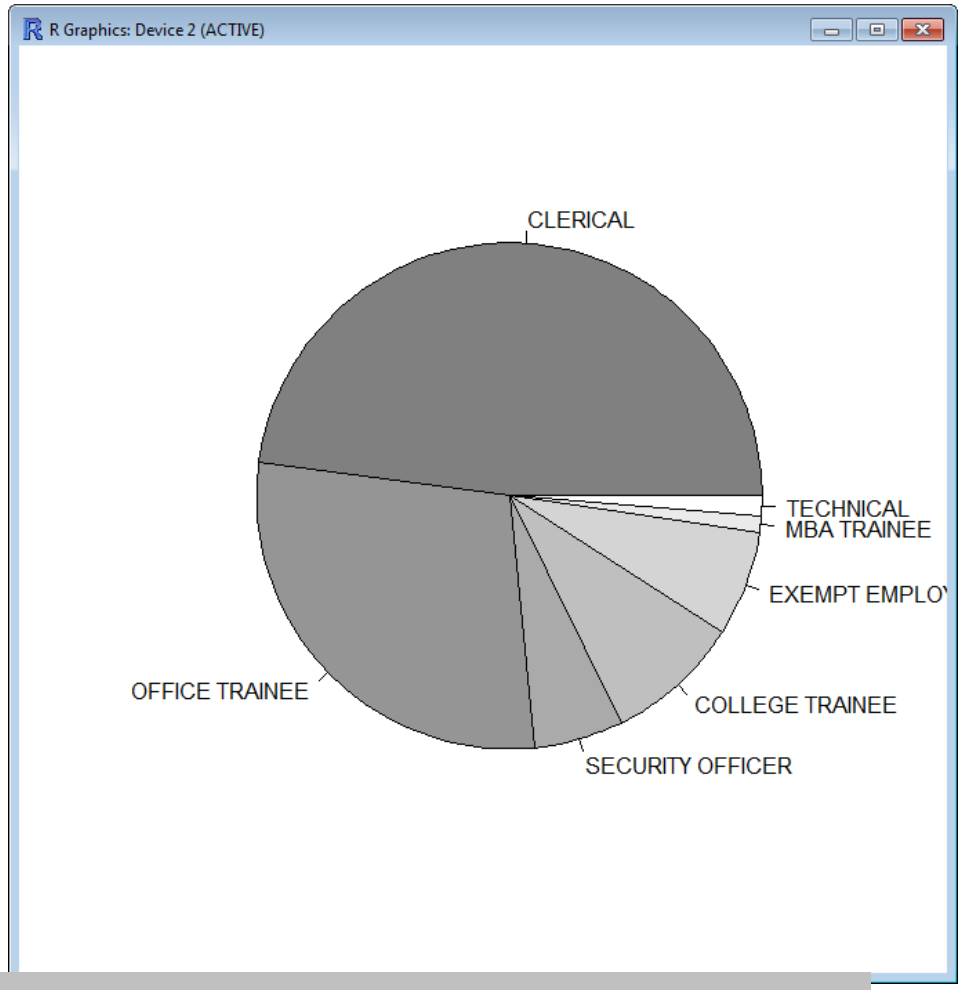



```
pie(table(salary$jobcat), col=grey( seq( 0, 1, length.out=7) ) )
```

Using shades of grey

The Grey palette

- Salary dataset
- Variable jobcat




```
pie(table(salary$jobcat), col=grey( seq( 0.5, 1, length.out=7) ) )
```

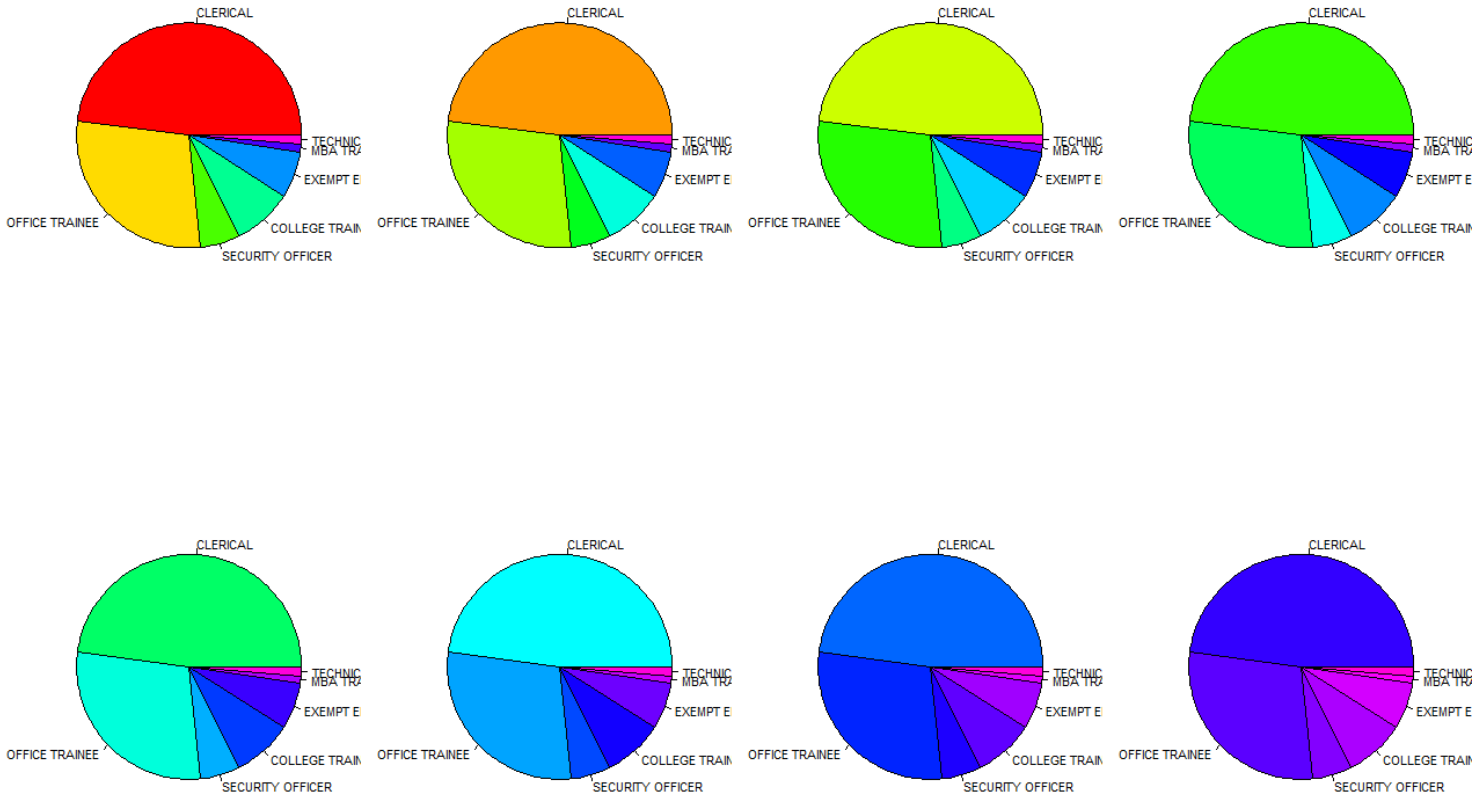
Using different palettes of colors

The rainbow palette

- Salary dataset
- Variable jobcat

```
index<-seq(0,0.7, 0.1)
par(mfrow=c(2,4))
for (i in 1:length(index)){
        pie(table(salary$jobcat), col=rainbow(7, start=index[i]),
                                        radius=1.5)
}
```

3.1. One Categorical variable
Pie chart



3.1. One Categorical variable
Pie chart

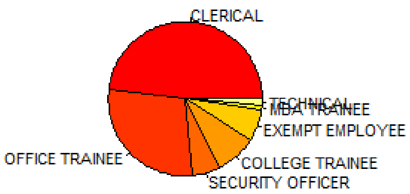
Using different palettes of colors

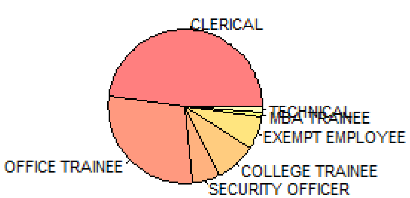
Other color palettes – the parameter alpha determines the transparency

```
pie(table(salary$jobcat), col=heat.colors(7,alpha=1) )
 pie(table(salary$jobcat), col=heat.colors(7,alpha=0.5) )

  pie(table(salary$jobcat), col=terrain.colors(7,alpha=1.) )
  pie(table(salary$jobcat), col=terrain.colors(7,alpha=0.5) )

  pie(table(salary$jobcat), col=topo.colors(7,alpha=0.5) )
  pie(table(salary$jobcat), col=topo.colors(7,alpha=1) )

  pie(table(salary$jobcat), col=cm.colors(7,alpha=1) )
  pie(table(salary$jobcat), col=cm.colors(7,alpha=0.5) )
```

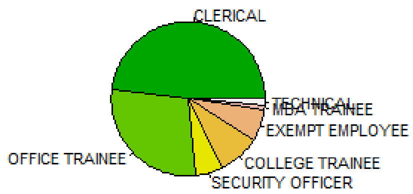
3.1. One Categorical variable
Pie chart

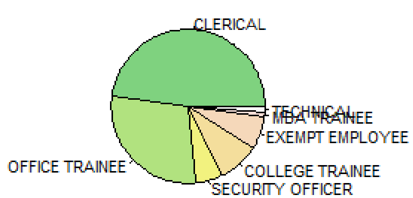
heat.colors



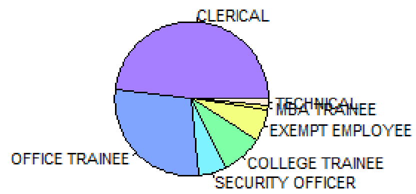


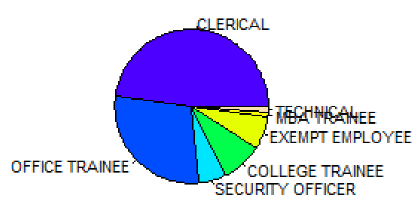
terrain.colors



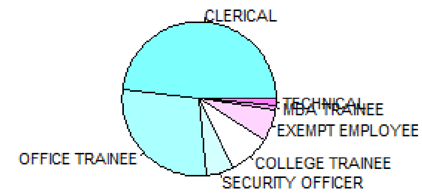


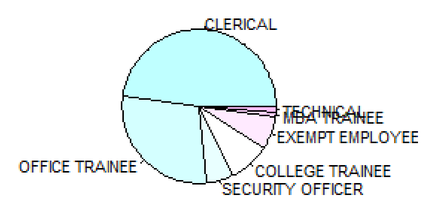
topo.colors





cm.colors

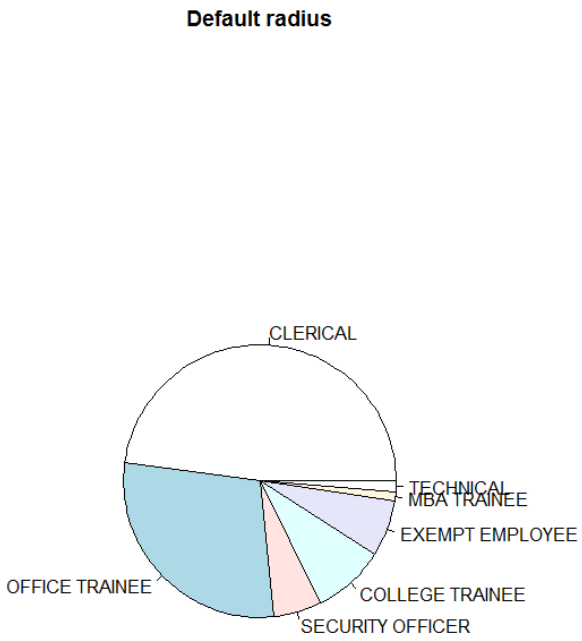


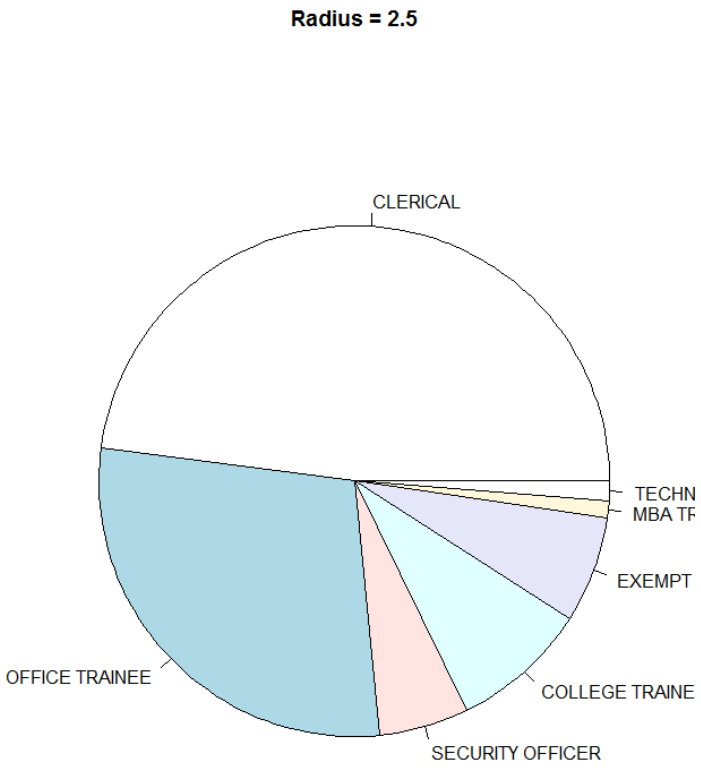


3.1. One Categorical variable
Pie chart

Changing the size of the pie using radius

- Salary dataset
- Variable jobcat

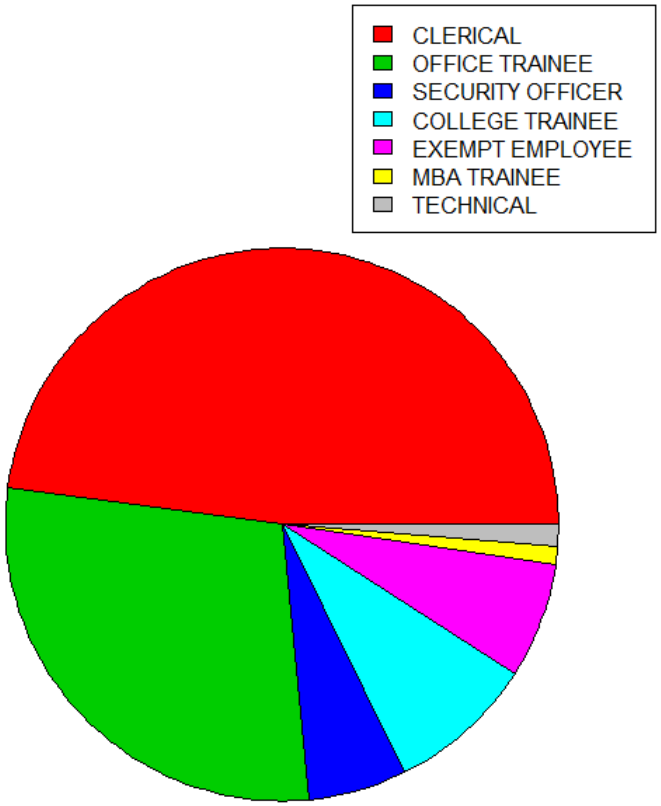





```
pie(table(salary$jobcat), radius=2.5)
```

Adding a legend

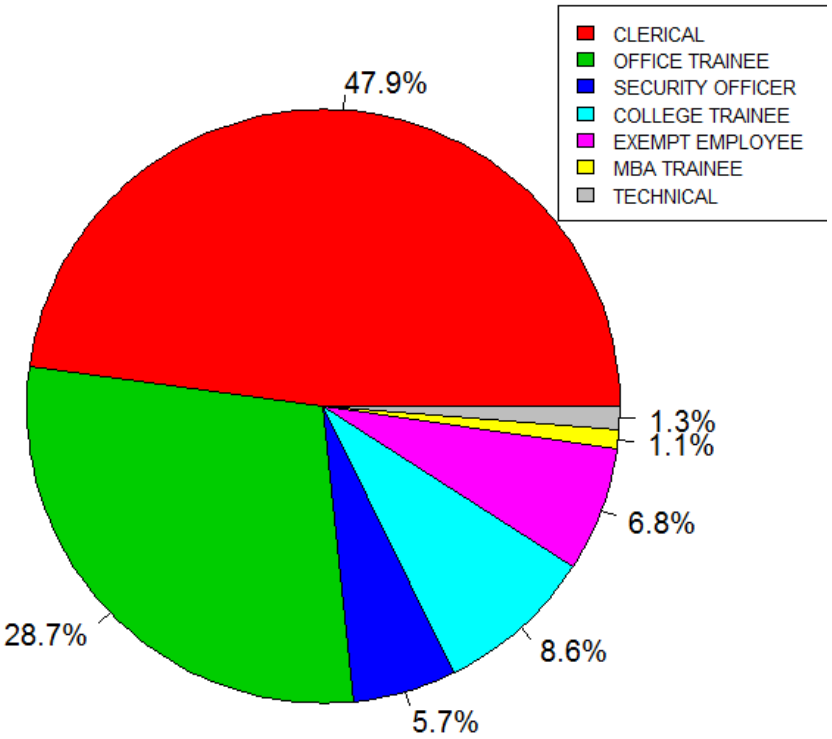
- Salary dataset
- Variable jobcat




```
pie(table(salary$jobcat), label=NA, col=2:8)
legend('topright', fill=2:8, legend=levels(salary$jobcat),cex=1.0)
```

Adding percentages

- Salary dataset
- Variable jobcat



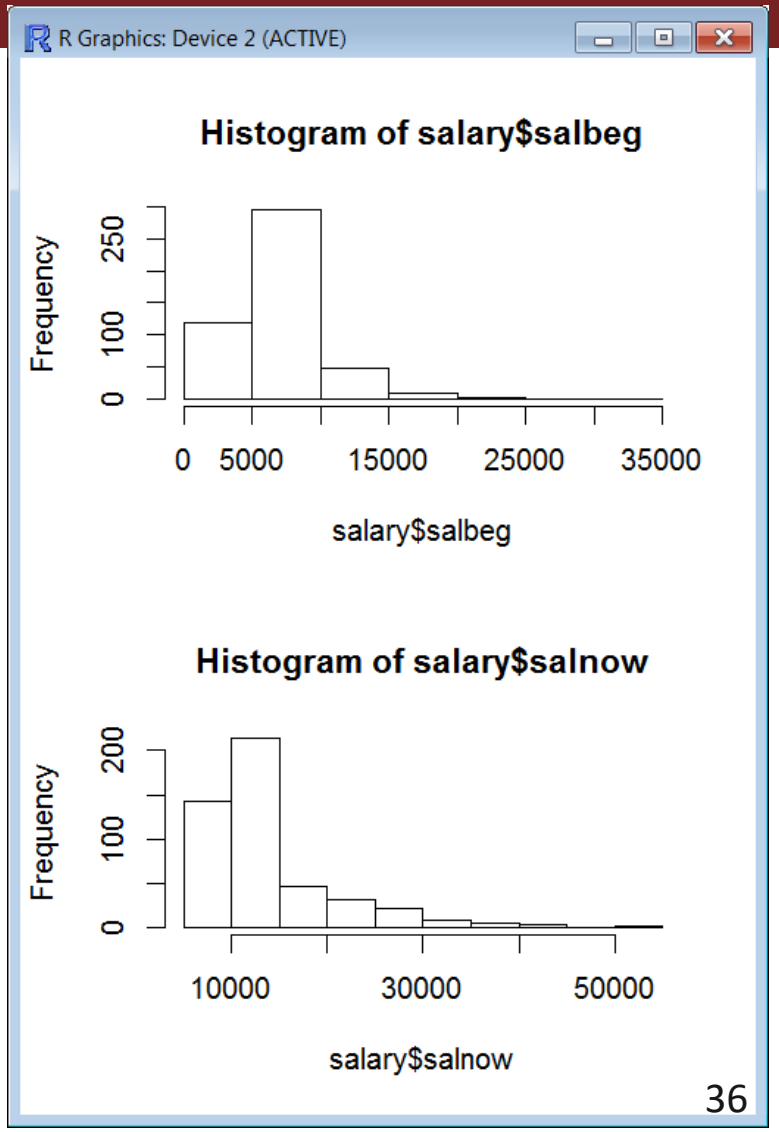

```
x <- 100*table(salary$jobcat)/length(salary$jobcat)
pie(x, label=paste( round(x,1), '%', sep='') , col=2:8)
legend('topright', fill=2:8, legend=levels(salary$jobcat),cex=0.7)
```

- Histogram
- Density plot
- Box-whiskers plot/Box plot
- Qqplot or pplot

3.2. One quantitative variable
Histogram

- Example: salary dataset
- Variables: salbeg & salnow

```
par(mfrow=c(2,1))
hist(salary$salbeg)
hist(salary$salnow)
```

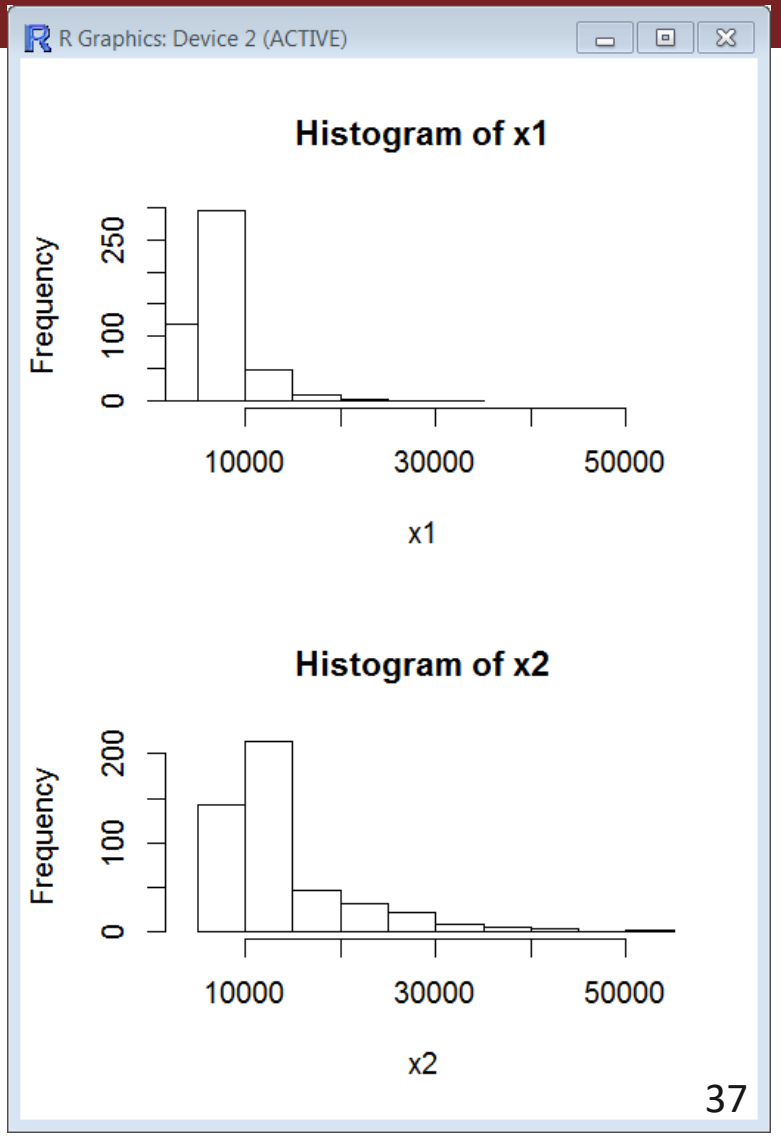


3.2. One quantitative variable
Histogram

Putting common limits in x-axis

- Example: salary dataset
- Variables: salbeg & salnow

```
par(mfrow=c(2,1))
x1<-salary$salbeg
x2<-salary$salnow
x3<-range(c(x1,x2))
hist(x1, xlim=x3)
hist(x2, xlim=x3)
```

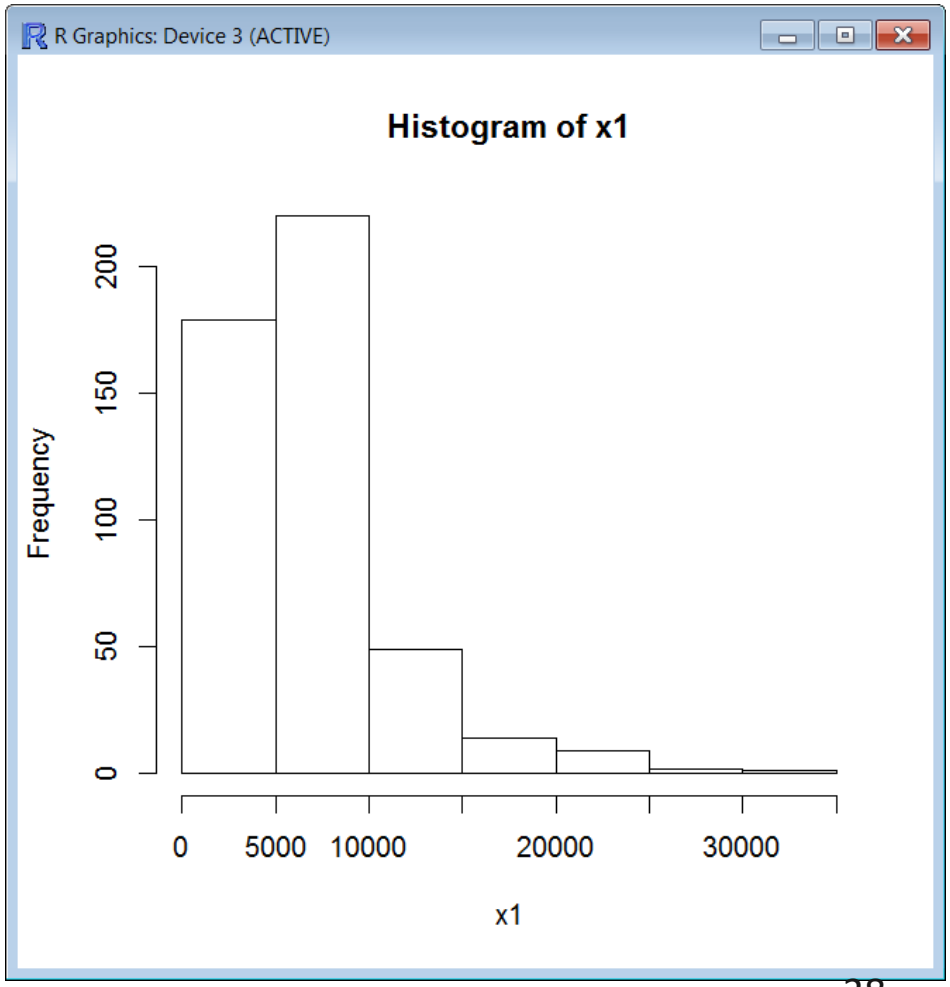


3.2. One quantitative variable
Histogram

Monitoring differences

- Example: salary dataset
- Variable: Salary difference

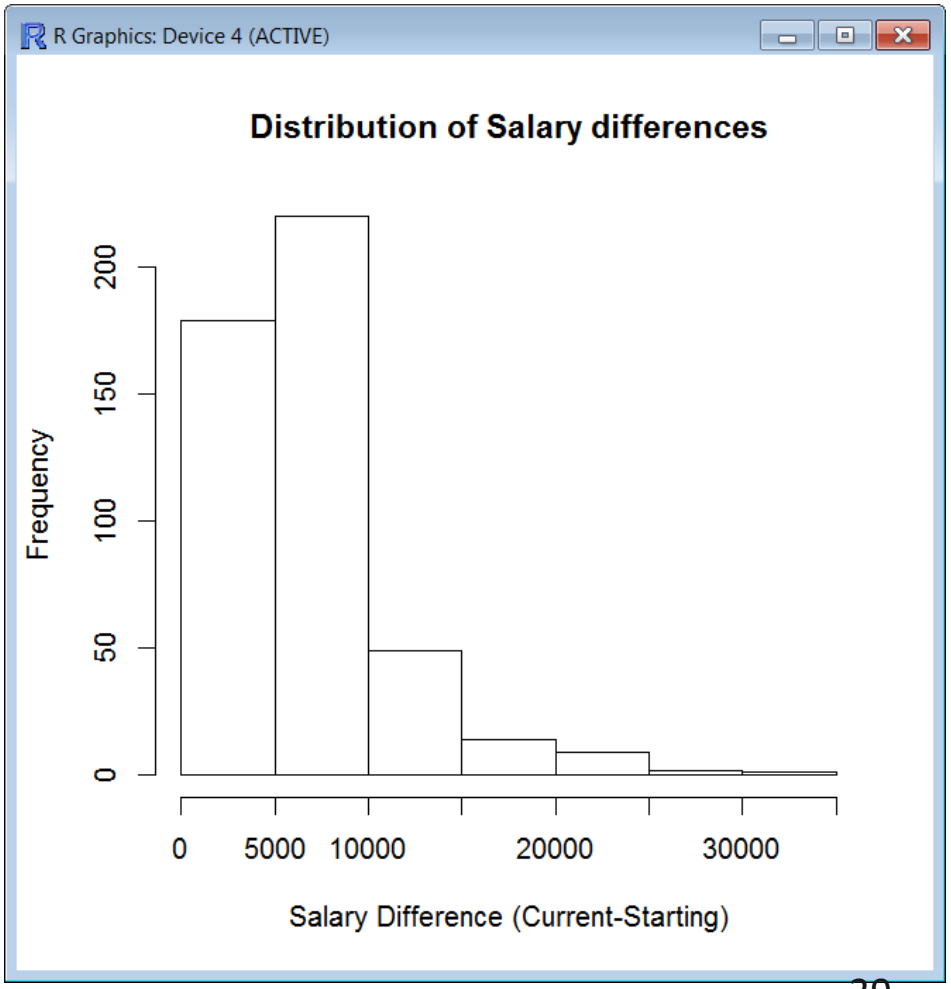
```
win.graph()
x1 <- salary$salnow-salary$salbeg
hist(x1)
```



```
win.graph()
x1 <- salary$salnow-salary$salbeg
hist(x1,  main='Distribution of Salary differences',
          xlab='Salary Difference (Current-Starting)' )
```

Adding Labels

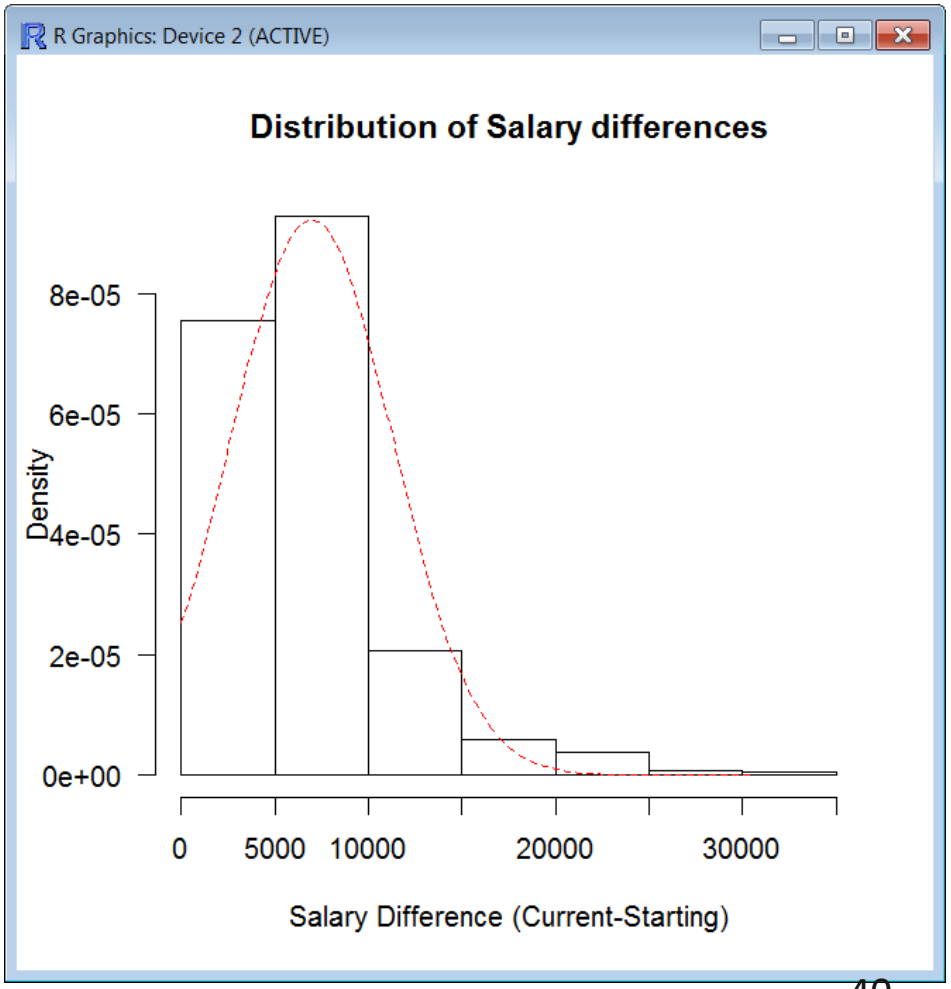
- Example: salary dataset
- Variable: Salary difference



```
x1 <- salary$salnow-salary$salbeg
hist(x1, main='Distribution of Salary differences',
          xlab='Salary Difference (Current-Starting)', probability=TRUE, las=1)
x2<- seq( min(c(0,x1)), max(x1), length.out=100)
lines( x2, dnorm(x2,mean(x1),sd(x1)), lty=2, col=2 )
```

Comparing with the normal density

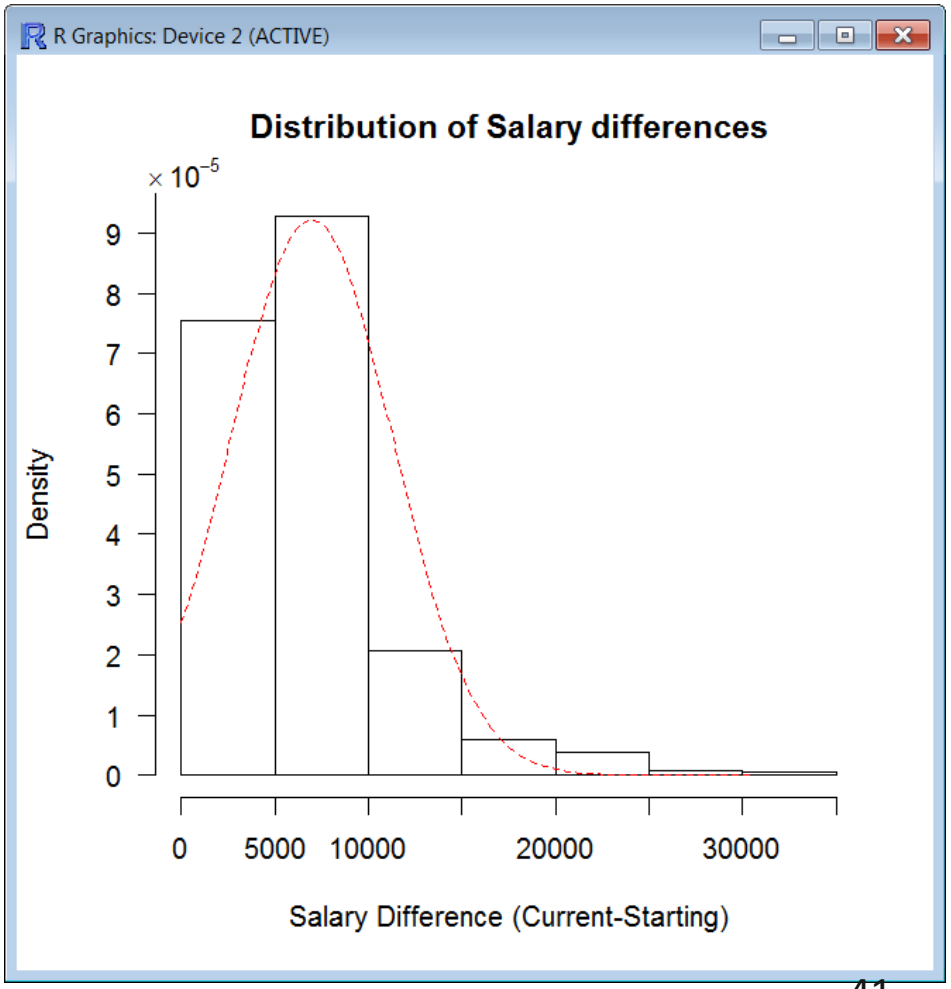
- Example: salary dataset
- Variable: Salary difference



3.2. One quantitative variable
Histogram

Putting nice labels

- Example: salary dataset
- Variable: Salary difference



3.2. One quantitative variable
Histogram

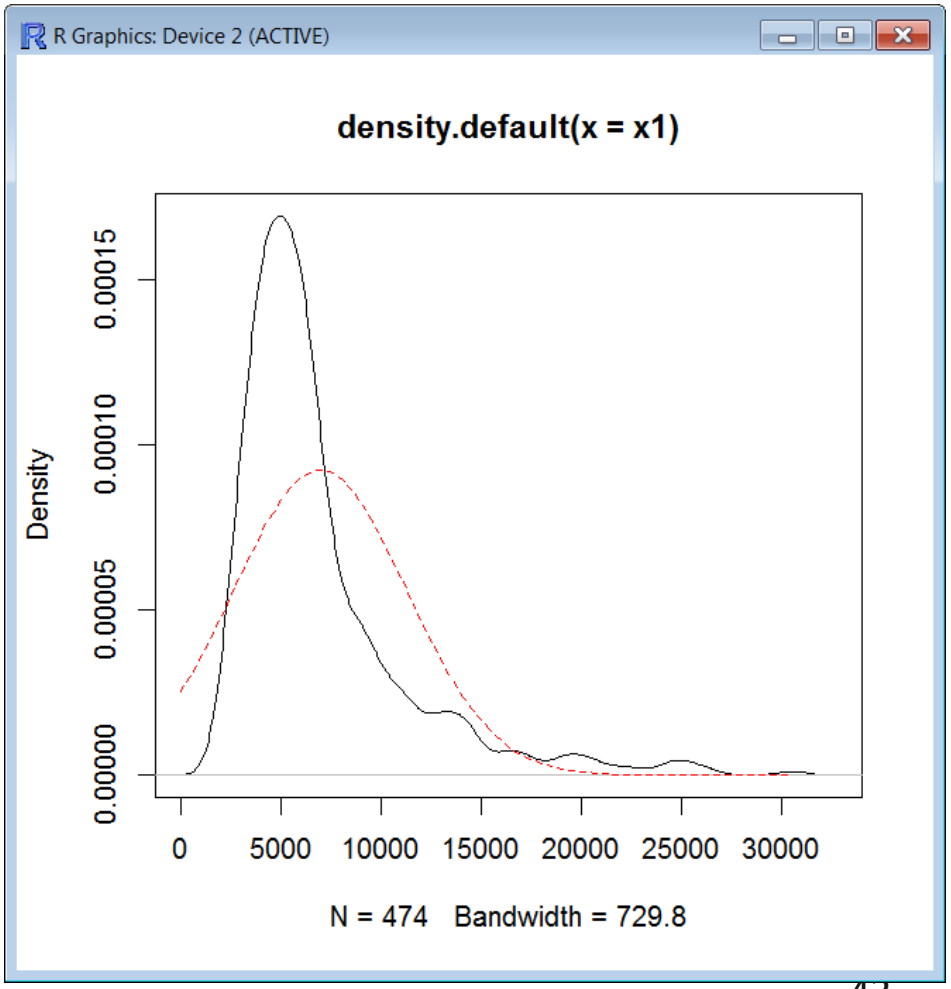
Putting nice labels

```
x1 <- salary$salnow-salary$salbeg
hist(x1, main='Distribution of Salary differences', xlab='Salary Difference (Current-
Starting)', probability=TRUE, axes=FALSE)
axis(1)
temp1<-hist(x1, main='Distribution of Salary differences', xlab='Salary Difference
(Current-Starting)', probability=TRUE, axes=FALSE)
temp2<-pretty(temp1$density,7)
axis(1)
axis(2, at = temp2, labels = 0:10, las=1)
mtext(bquote(" %*% 10^{-5}), at=11)

x2<- seq( min(c(0,x1)), max(x1), length.out=100)
lines( x2, dnorm(x2,mean(x1),sd(x1)), lty=2, col=2 )
```

```
x1 <- salary$salnow-salary$salbeg
plot(density(x1))
x2<- seq( min(c(0,x1)), max(x1), length.out=100)
lines( x2, dnorm(x2,mean(x1),sd(x1)), lty=2, col=2 )
```

Density plots

- Example: salary dataset
- Variable: Salary difference

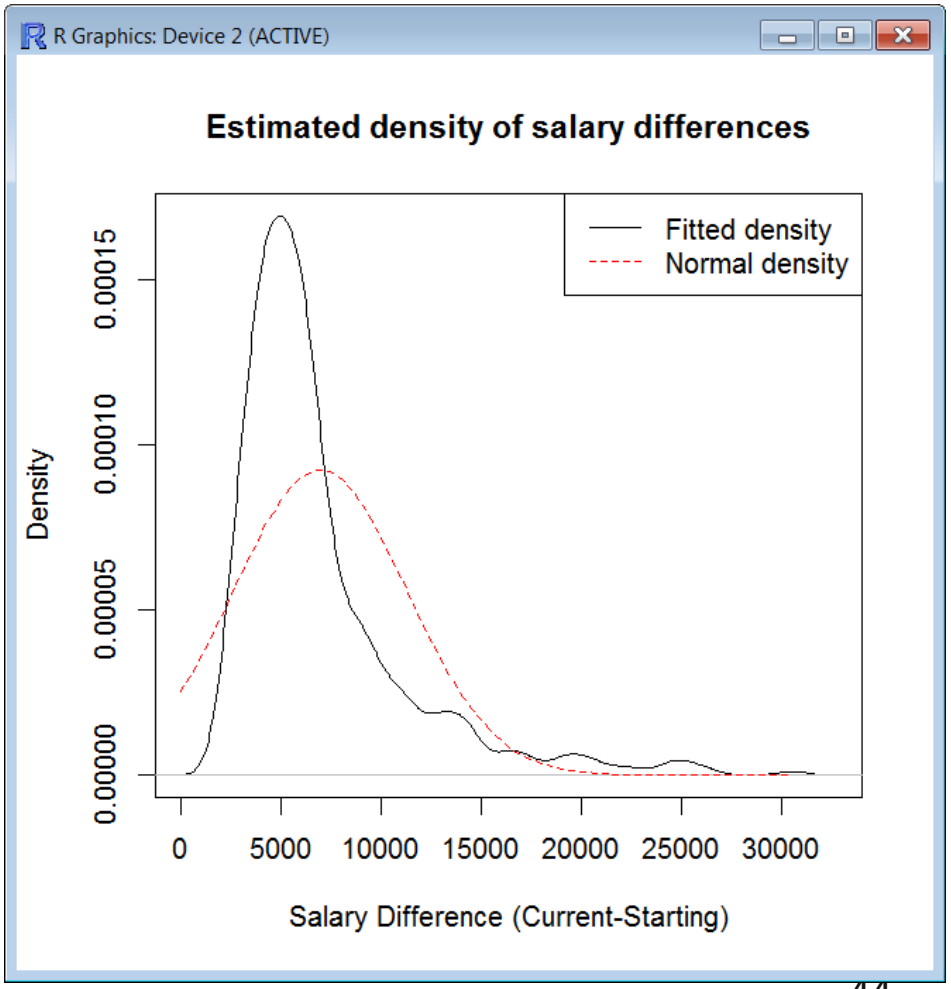


```
x1 <- salary$salnow-salary$salbeg
plot(density(x1), main='Estimated density of salary differences', xlab='Salary Difference
(Current-Starting)')
x2<- seq( min(c(0,x1)), max(x1), length.out=100)
lines( x2, dnorm(x2,mean(x1),sd(x1)), lty=2, col=2 )
legend( "topright", col=1:2, lty=1:2, legend=c("Fitted density", "Normal density") )
```

Density plots

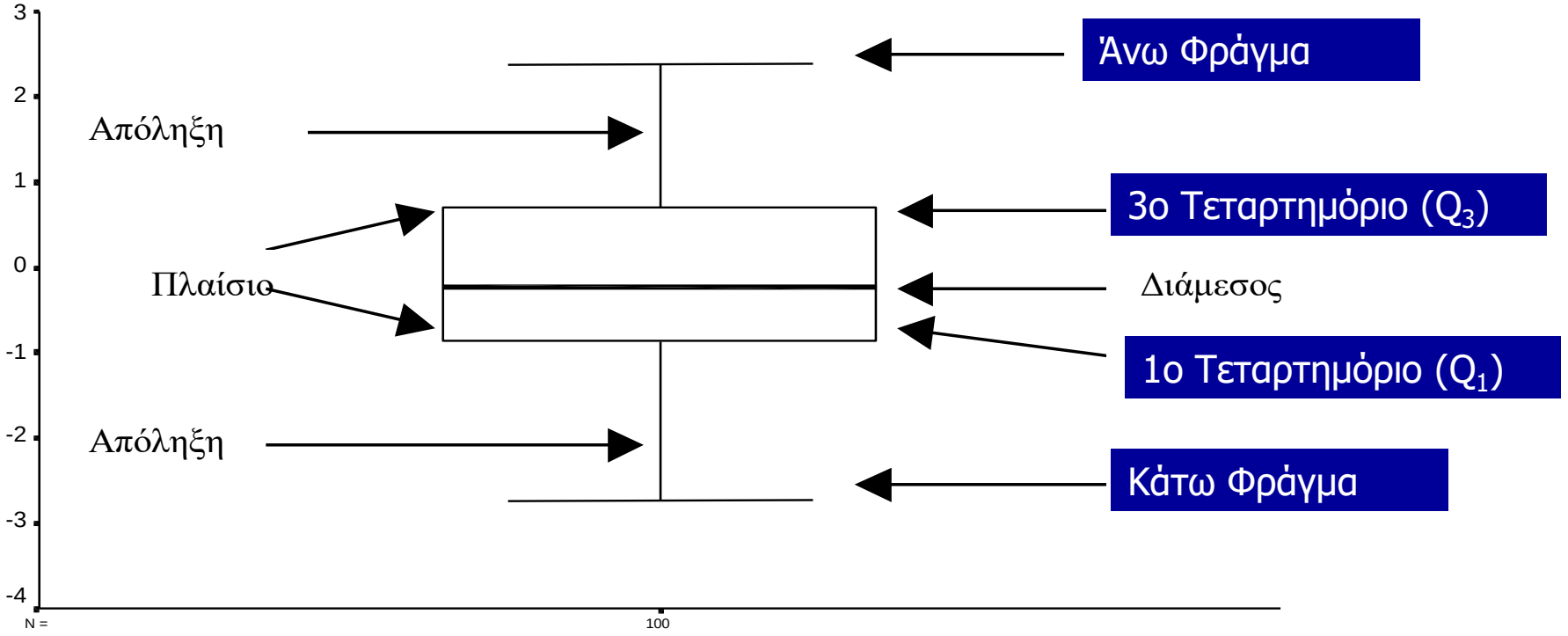
Putting nice labels

- Example: salary dataset
- Variable: Salary difference



Box whiskers plots – Box plots

- o Outliers
- * Extremes (not in R default boxplot function)

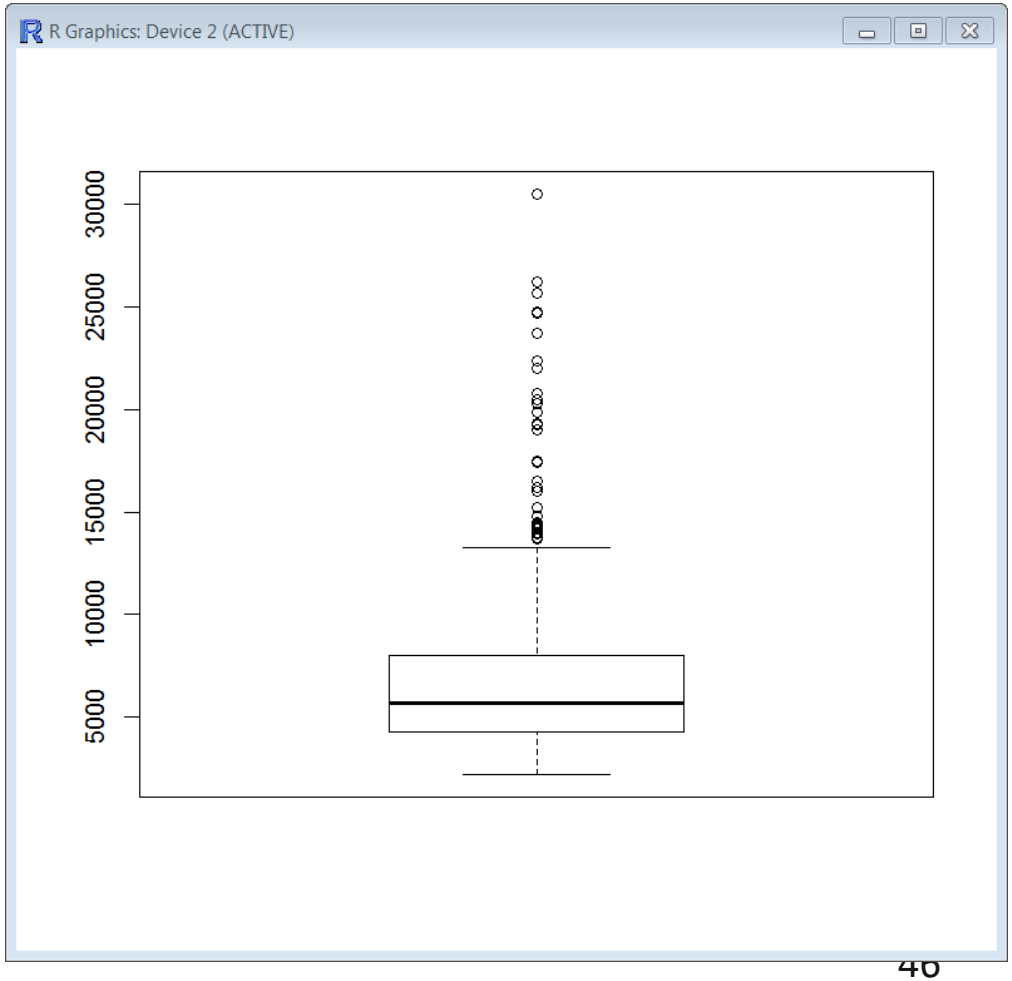


3.2. One quantitative variable
Box plots

The Box plot function

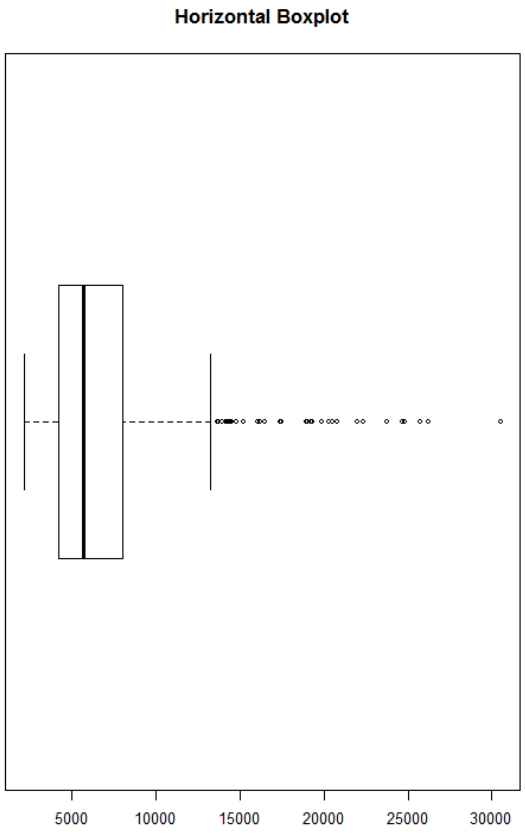
- Example salary
- Variable salary difference

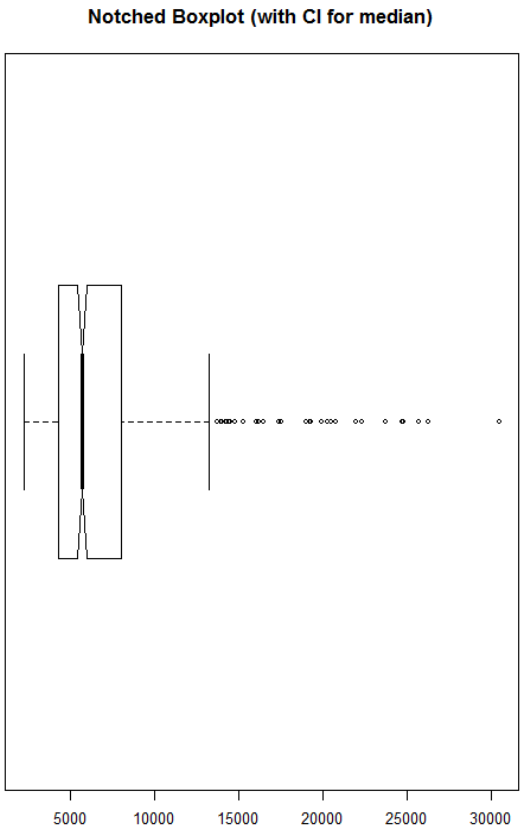
```
x1 <- salary$salnow-salary$salbeg
boxplot(x1)
```

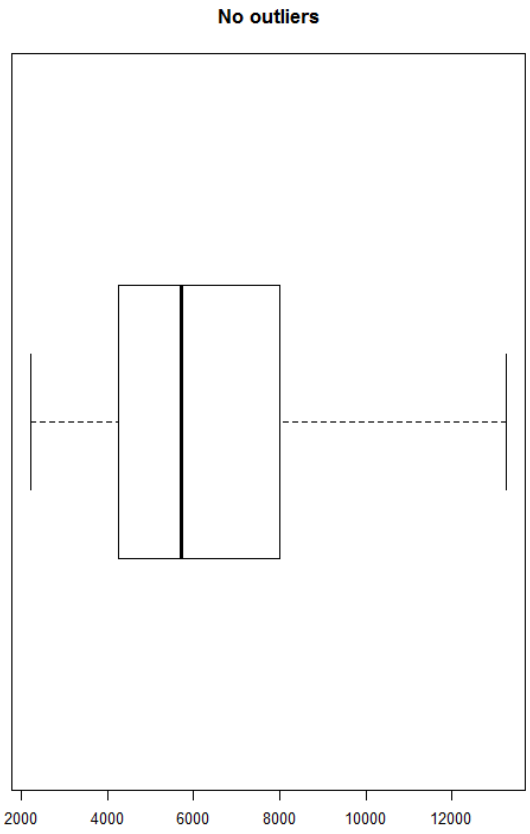


```
par(mfrow=c(1,3))
boxplot(x1, horizontal=T, main='Horizontal Boxplot')
boxplot(x1, horizontal=T, notch=T, main='Notched Boxplot (with CI for median)')
boxplot(x1, horizontal=T, outline=F,  main='No outliers')
```

Variations in boxplots







```
par(mfrow=c(1,3))
boxplot(x1, horizontal=T, main='Horizontal Boxplot')
boxplot(x1, horizontal=T, notch=T, main='Notched Boxplot (with CI for median)')
boxplot(x1, horizontal=T, outline=F,  main='No outliers')
```

RGui (64-bit) - [R Console]

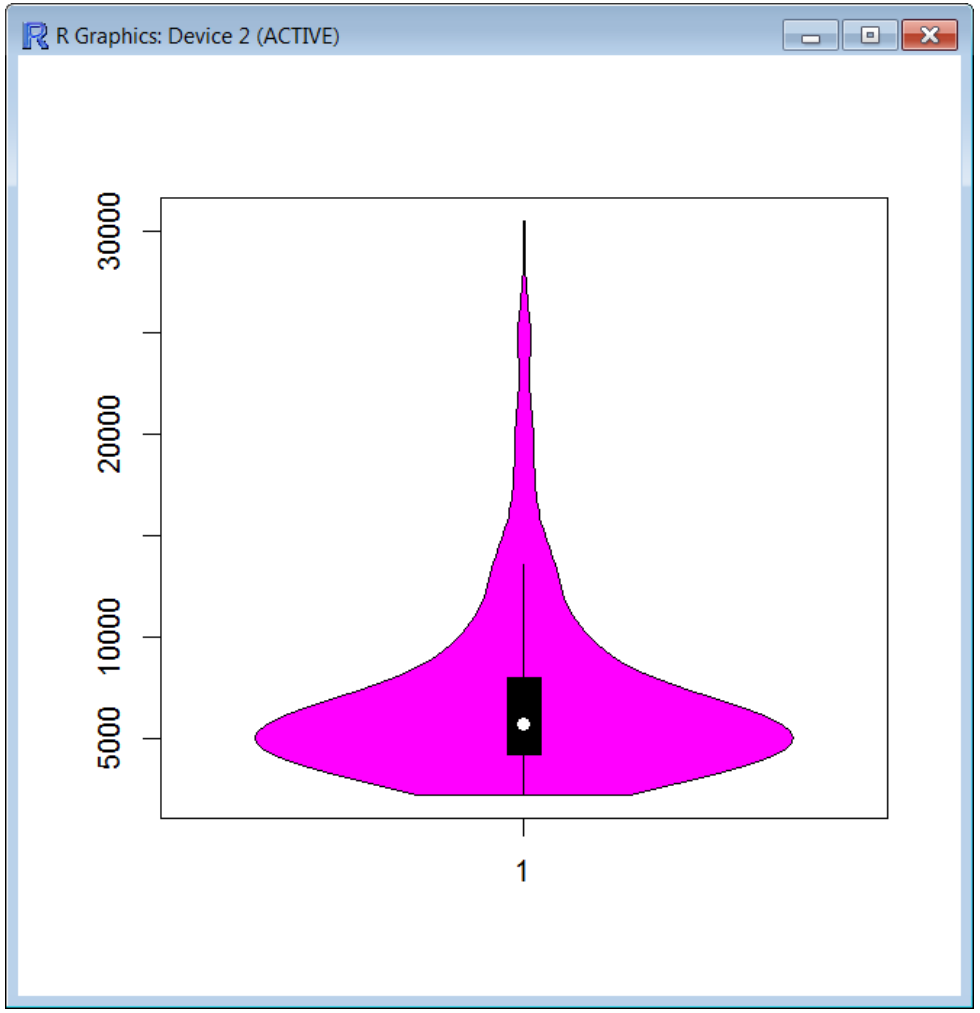
File Edit View Misc Packages Windows Help

```
> temp <- boxplot(x1)
> temp
$stats
      [,1]
[1,]  2220
[2,]  4260
[3,]  5700
[4,]  8008
[5,] 13260

$n
[1] 474

$conf
         [,1]
[1,] 5428.001
[2,] 5971.999

$out
 [1] 17400 14254 16496 19300 14446 20308 16200 14508 16500 22008 25700 19250 14800 13700 15250 19000 24796
[18] 13958 17504 23750 24800 26250 20804 30496 14200 13984 16508 19008 24704 20500 17400 13754 14400 14250
[35] 16050 14308 19904 22360

$group
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

$names
[1] "1"

> 
```

3.2. One quantitative variable
Violin plots

- Example salary
- Variable salary difference

```
# install.packages(vioplot)
library(vioplot)
vioplot(x1)
```



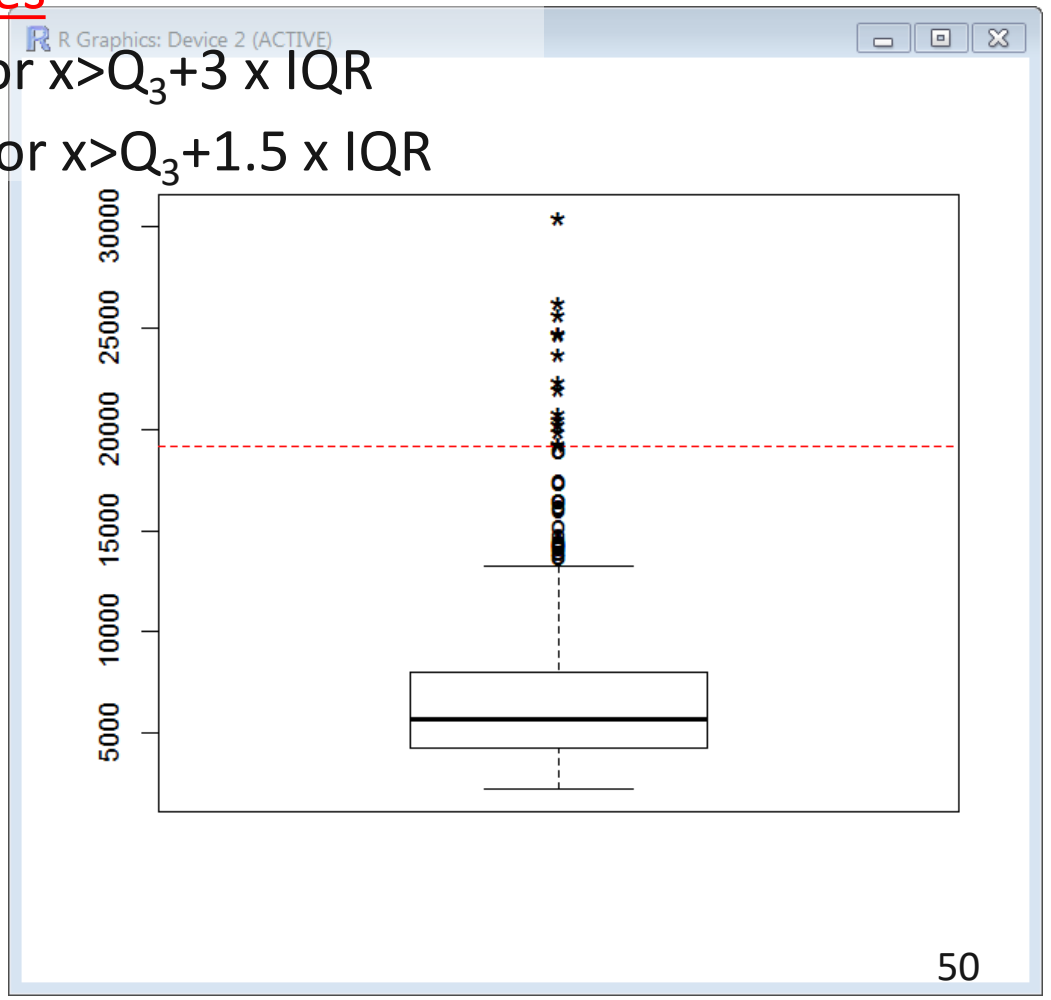
3.2. One quantitative variable
Box plots with extreme values

Making a boxplot with extremes

* => Extremes if $x<Q_1-3 \times IQR$ or $x>Q_3+3 \times IQR$

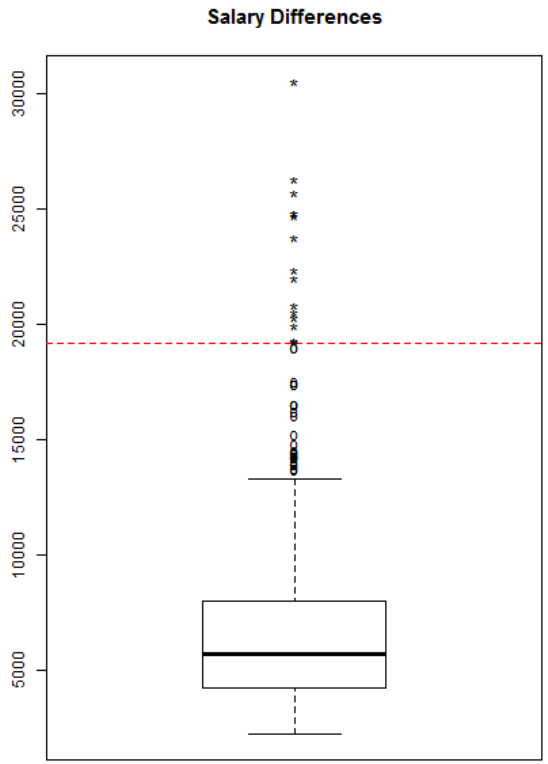
o => Outliers if $x<Q_1-1.5 \times IQR$ or $x>Q_3+1.5 \times IQR$

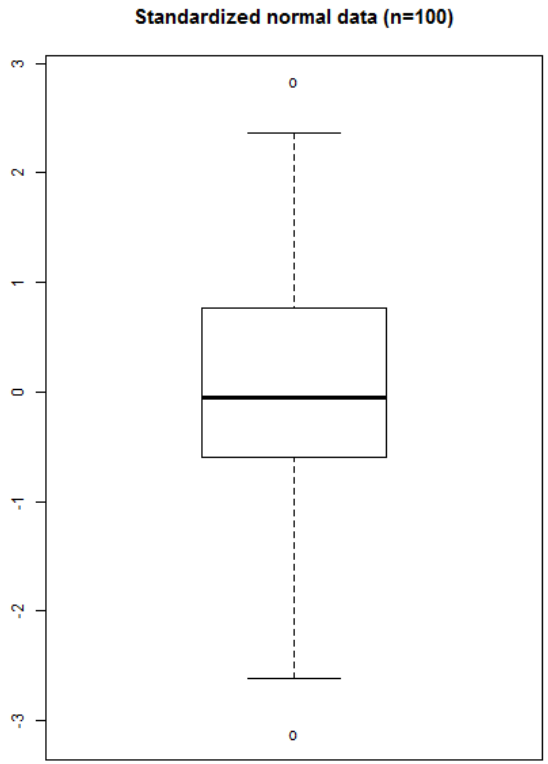
HOME MADE CODE

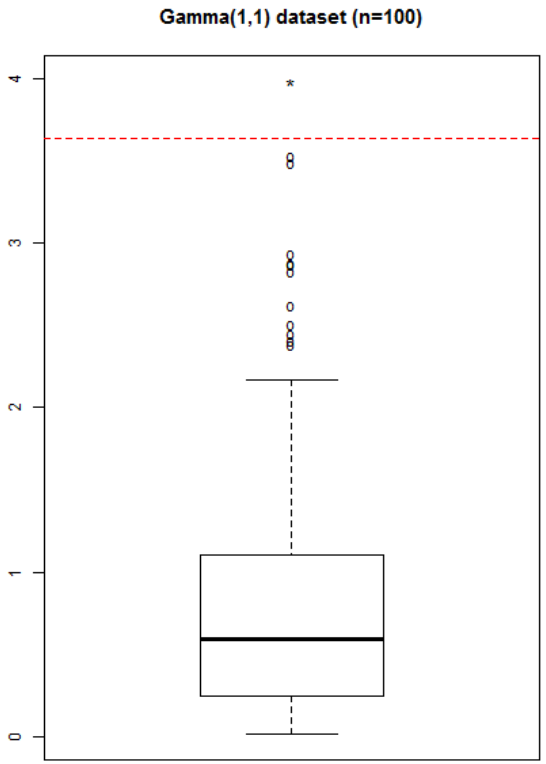


Box plots with extreme values

Home made boxplot function for extremes







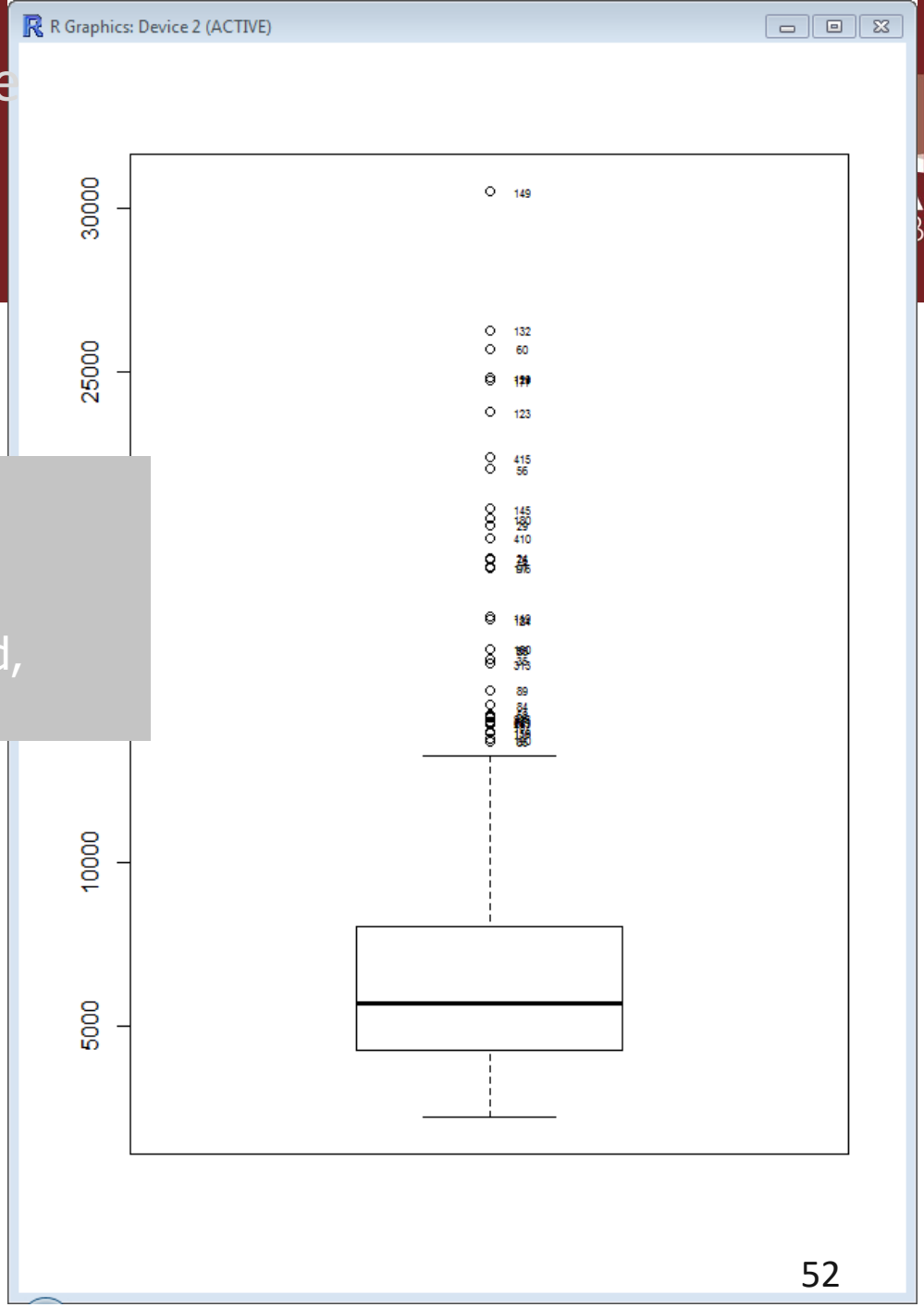

```
par(mfrow=c(1,3))
myboxplot(x1)
title(main='Salary Differences')
```

```
myboxplot(rnorm(100))
title(main='Standardized normal data (n=100)')

myboxplot(rgamma(100,1))
title(main='Gamma(1,1) dataset (n=100)')
```

3.2. One quantitative variable
Box plots with labels

Adding labels to outliers

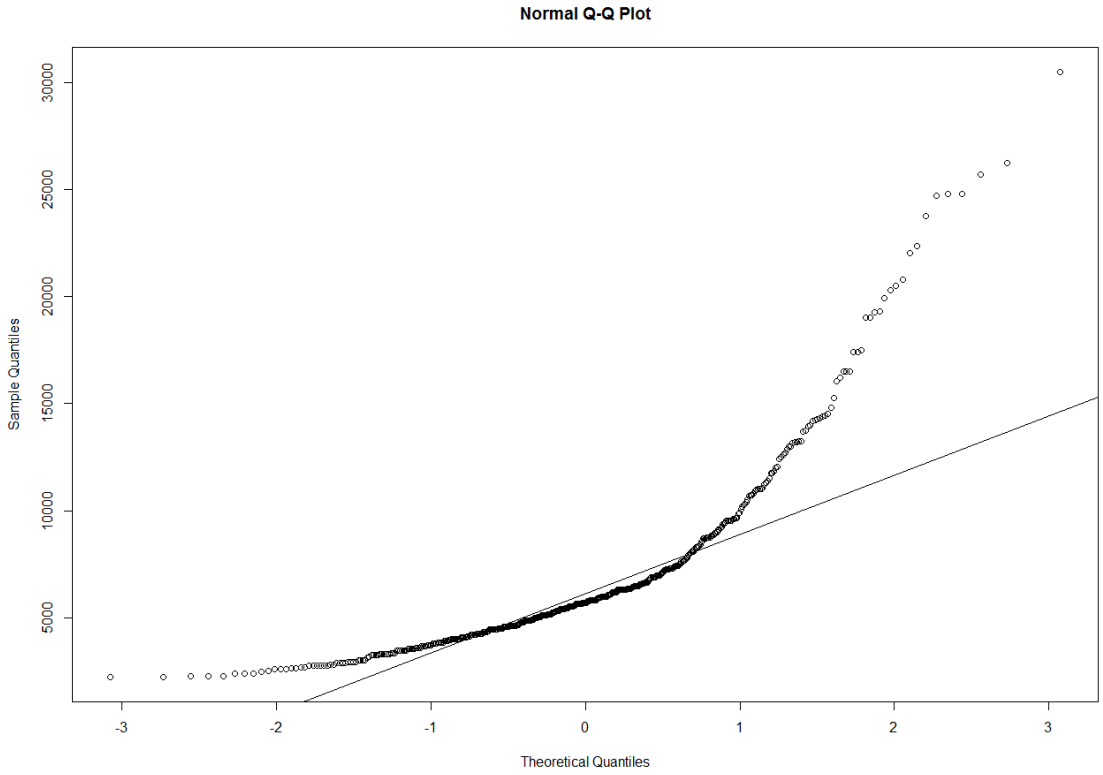
```
boxplot(x1)
temp <- boxplot(x1)
id<-which(x1 %in% temp$out)
text( rep(1.05,length(id)), temp$out, id,
            cex=0.5 )
```



3.2. One quantitative variable
Q-Qplots

- Q-Qplots are used for checking the fit in various distributions
- X axis => expected quantiles based on the distribution we want to compare
- Y axis => observed values – quantiles
- Example: salary dataset; variable: Salary difference

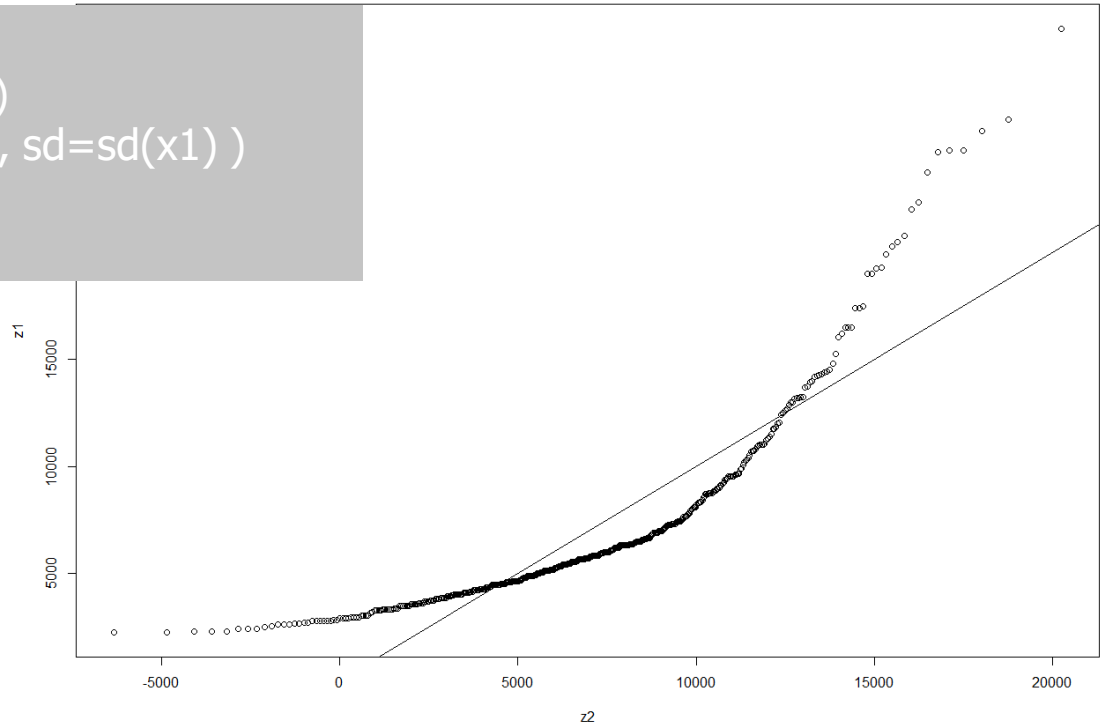
```
qqnorm(x1)
qqline(x1)
```



3.2. One quantitative variable
Q-Qplots

- Q-Qplots are used for checking the fit in various distributions
- X axis => expected quantiles based on the distribution we want to compare
- Y axis => observed values – quantiles
- Example: salary dataset; variable: Salary difference

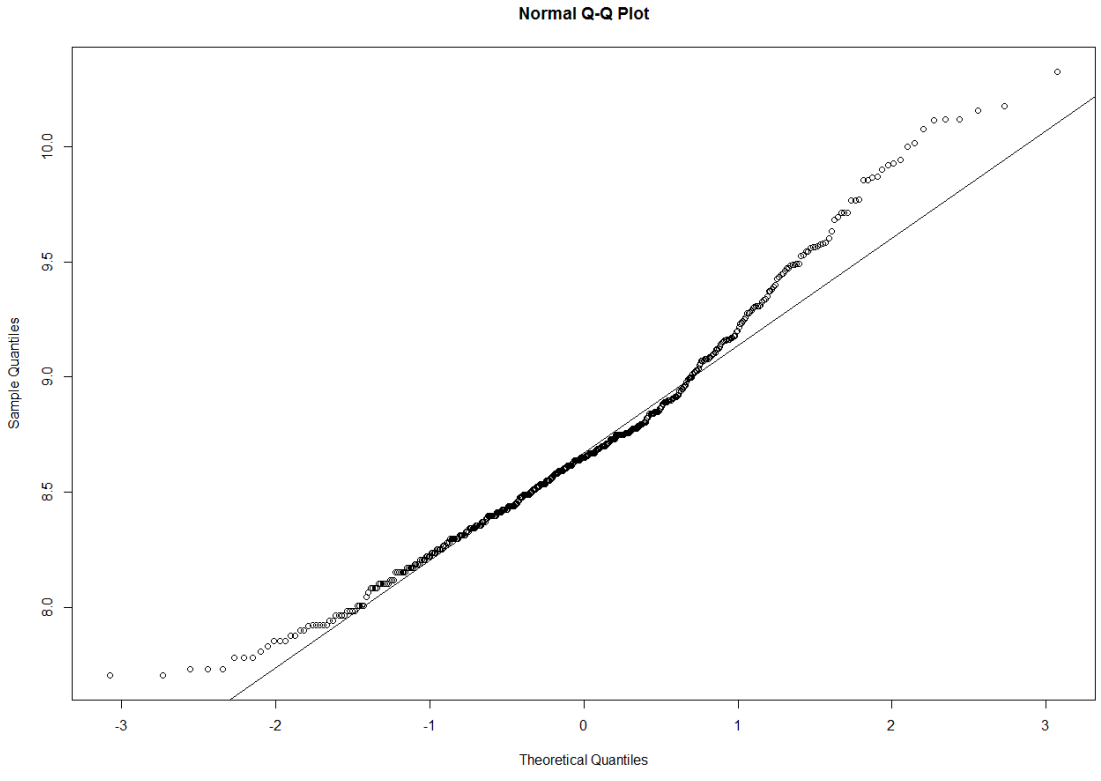
```
z1 <- sort(x1)
p2 <- (1:length(z1)-0.5)/length(z1)
z2 <- qnorm( p2, mean=mean(x1), sd=sd(x1) )
plot(z2,z1)
abline(0,1)
```



3.2. One quantitative variable
Q-Qplots

- Q-Qplots are used for checking the fit in various distributions
- X axis => observed values – quantiles
- Y axis => expected quantiles based on the distribution we want to compare
- Example: salary dataset; variable: Salary difference

```
qqnorm(log(x1))
qqline(log(x1))
```

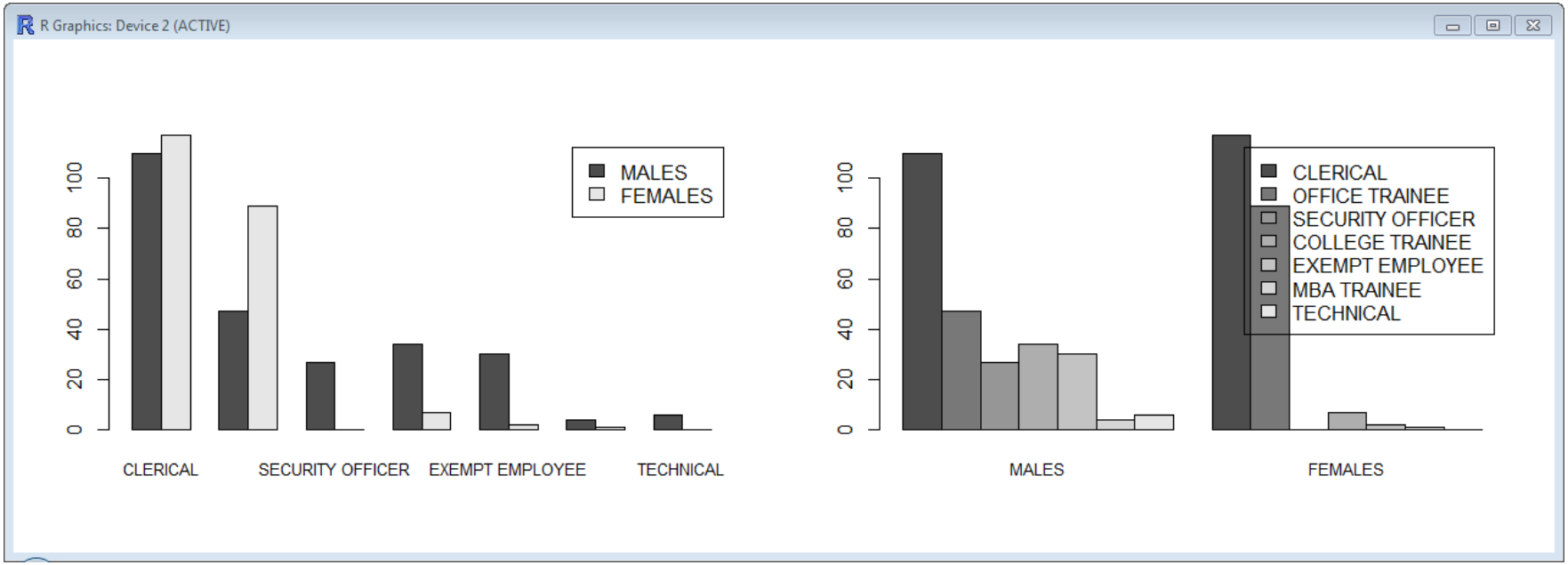


3.3. Visualizing associations between two categorical variables

- Clustered barchart
- Stacked barchart

3.3. Visualizing associations between two categorical variables: Clustered bar chart

- Salary dataset – variables: Sex by Job category

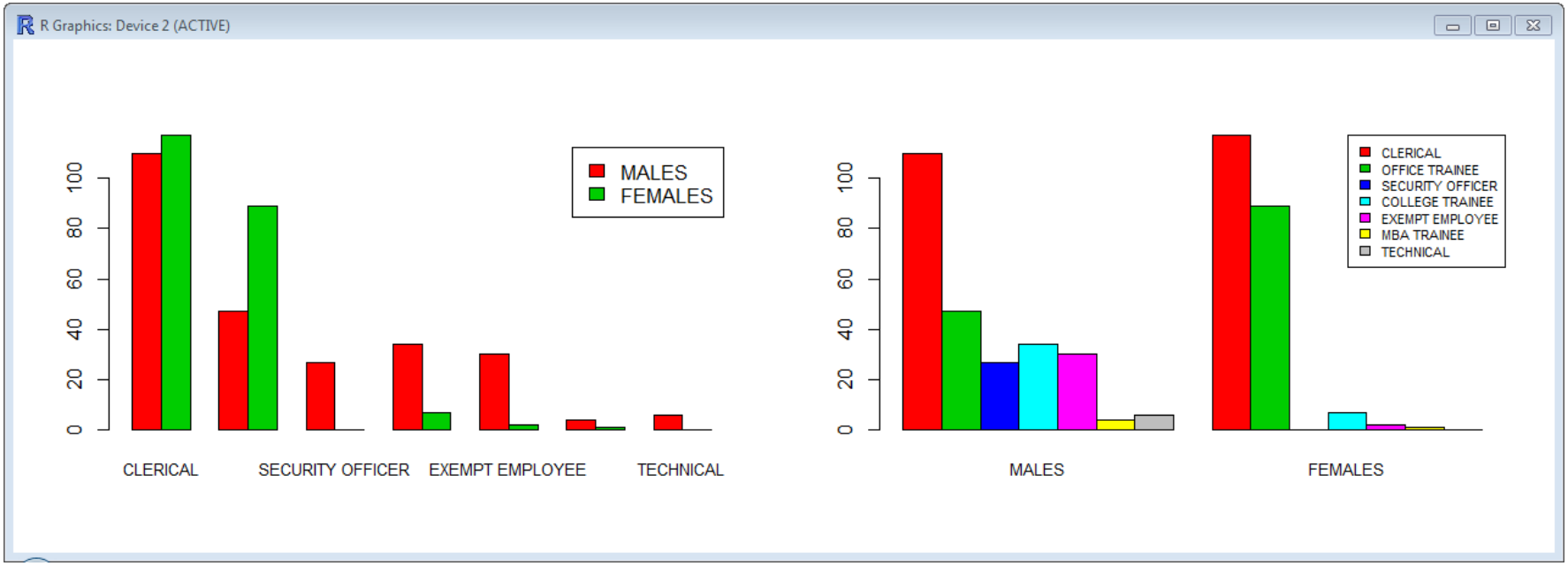



```
par(mfrow=c(1,2))
barplot( table(salary$sex,salary$jobcat), beside=T, cex.names=0.8, legend=TRUE)
barplot( t(table(salary$sex,salary$jobcat)), beside=T, cex.names=0.8, legend=TRUE)
```

3.3. Visualizing associations between two categorical variables: Clustered bar chart

Selecting colors and adjusting the legend

- Salary dataset – variables: Sex by Job category

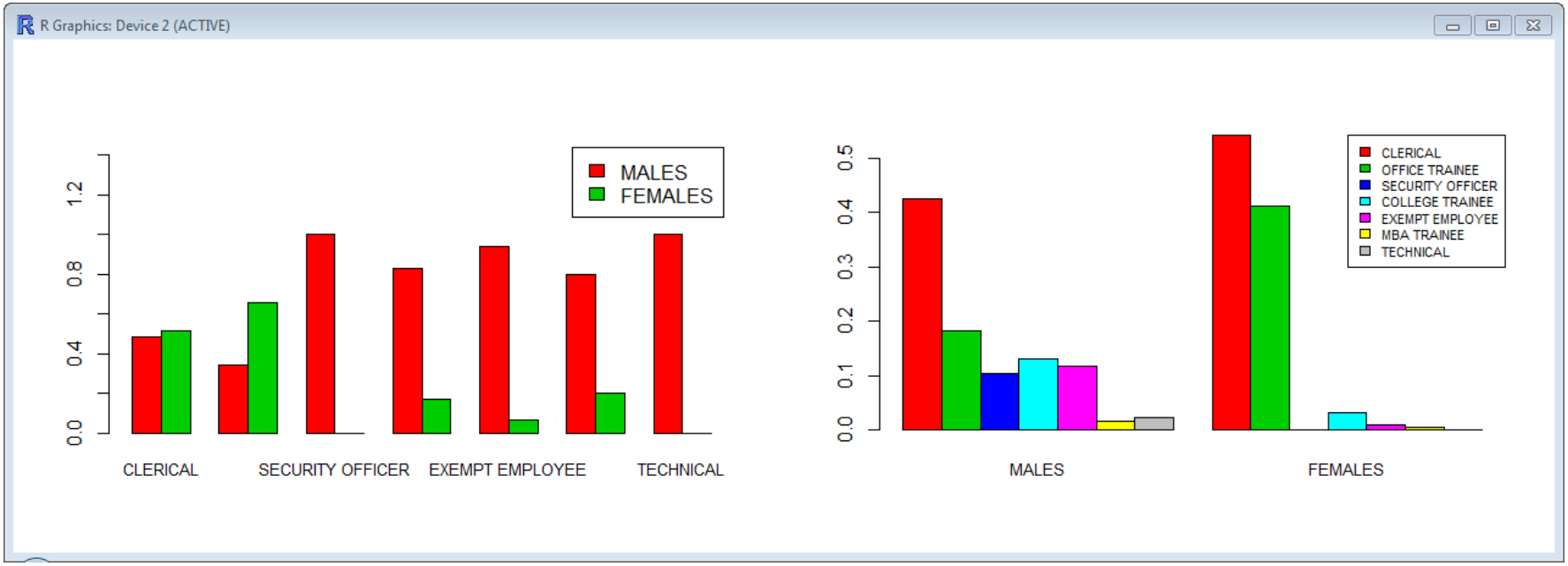



```
par(mfrow=c(1,2))
barplot( table(salary$sex,salary$jobcat), beside=T, cex.names=0.8, legend=TRUE, col=2:3)
barplot( t(table(salary$sex,salary$jobcat)), beside=T, cex.names=0.8, col=2:8)
legend( "topright", legend=levels(salary$jobcat), fill=2:8, cex=0.7 )
```

3.3. Visualizing associations between two categorical variables: Clustered bar chart

Comparing proportions – bars for each group on the x-axis add to one

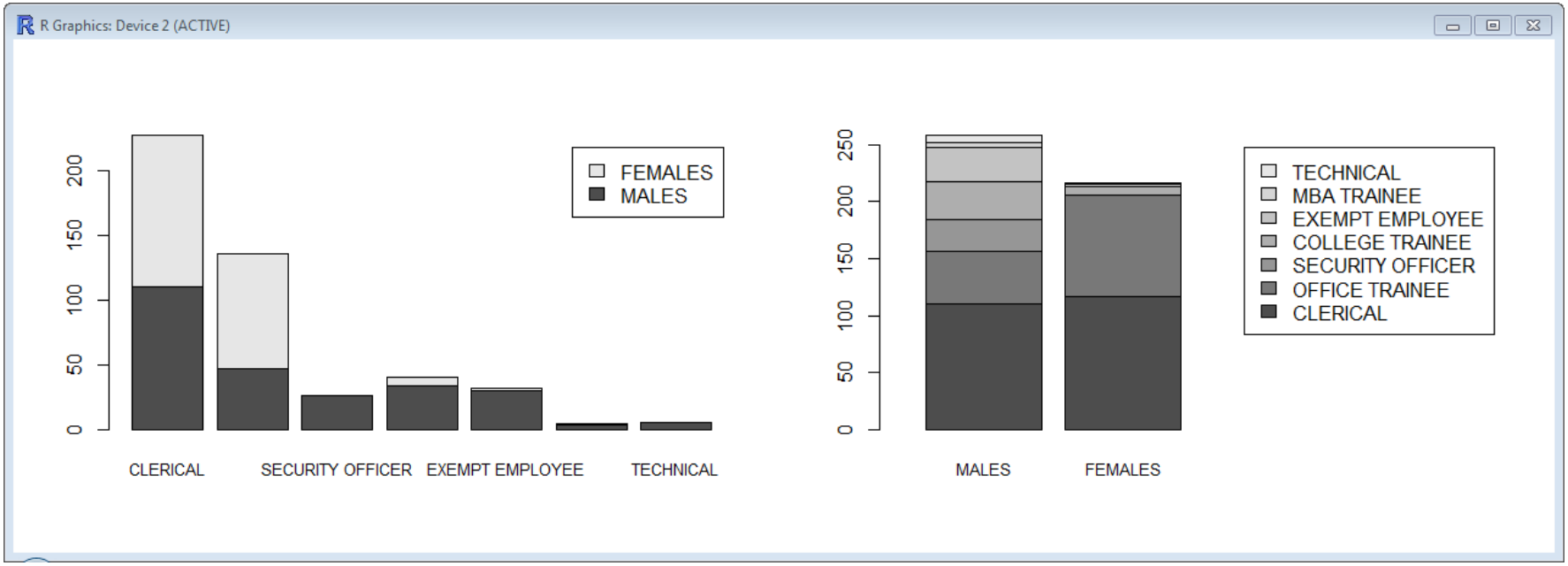
- Salary dataset – variables: Sex by Job category




```
par(mfrow=c(1,2))
x<-table(salary$sex,salary$jobcat)
barplot( apply(x,2,prop.table), beside=T, cex.names=0.8, col=2:3, legend=TRUE, ylim=c(0,1.5))
barplot( apply(x,1,prop.table), beside=T, cex.names=0.8, col=2:8)
legend( "topright", legend=levels(salary$jobcat), fill=2:8, cex=0.7 )
```

3.3. Visualizing associations between two categorical variables: Stacked bar charts

- Salary dataset – variables: Sex by Job category

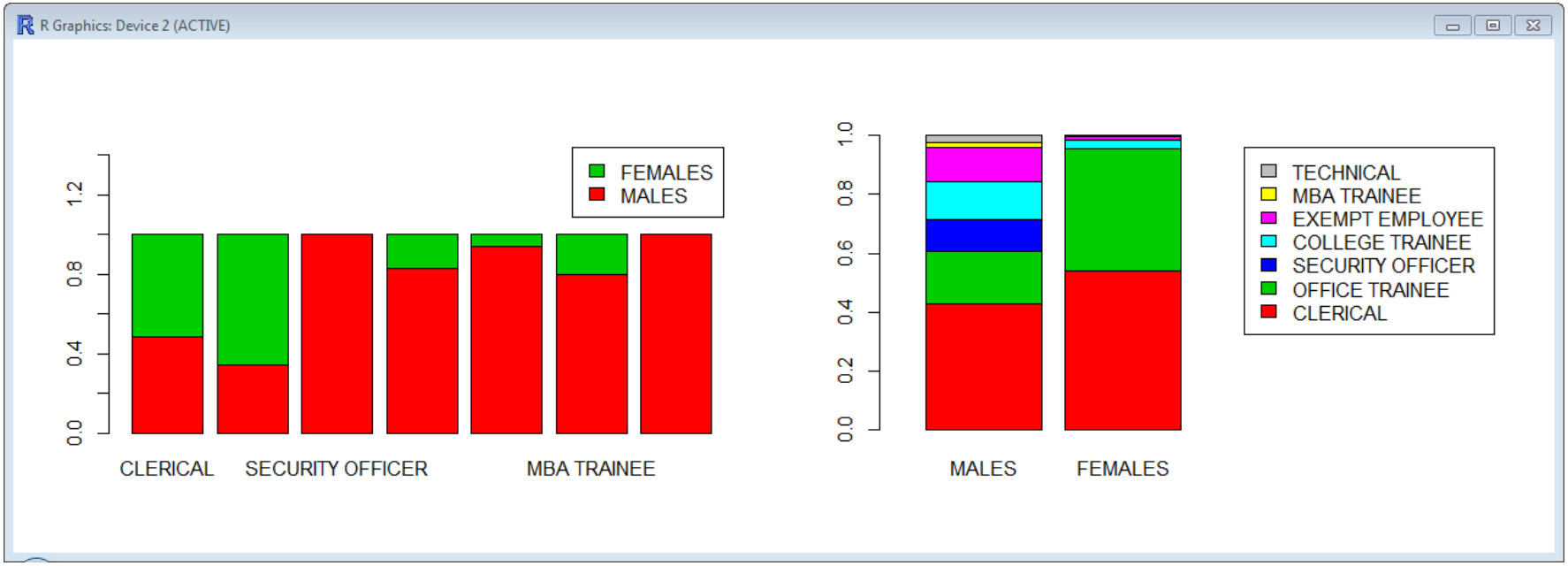



```
par(mfrow=c(1,2))
barplot( table(salary$sex,salary$jobcat), cex.names=0.8, legend=TRUE)
barplot( t(table(salary$sex,salary$jobcat)),  cex.names=0.8, legend=TRUE, xlim=c(0,5))
```

3.3. Visualizing associations between two categorical variables: Clustered bar chart

Comparing proportions

- Salary dataset – variables: Sex by Job category

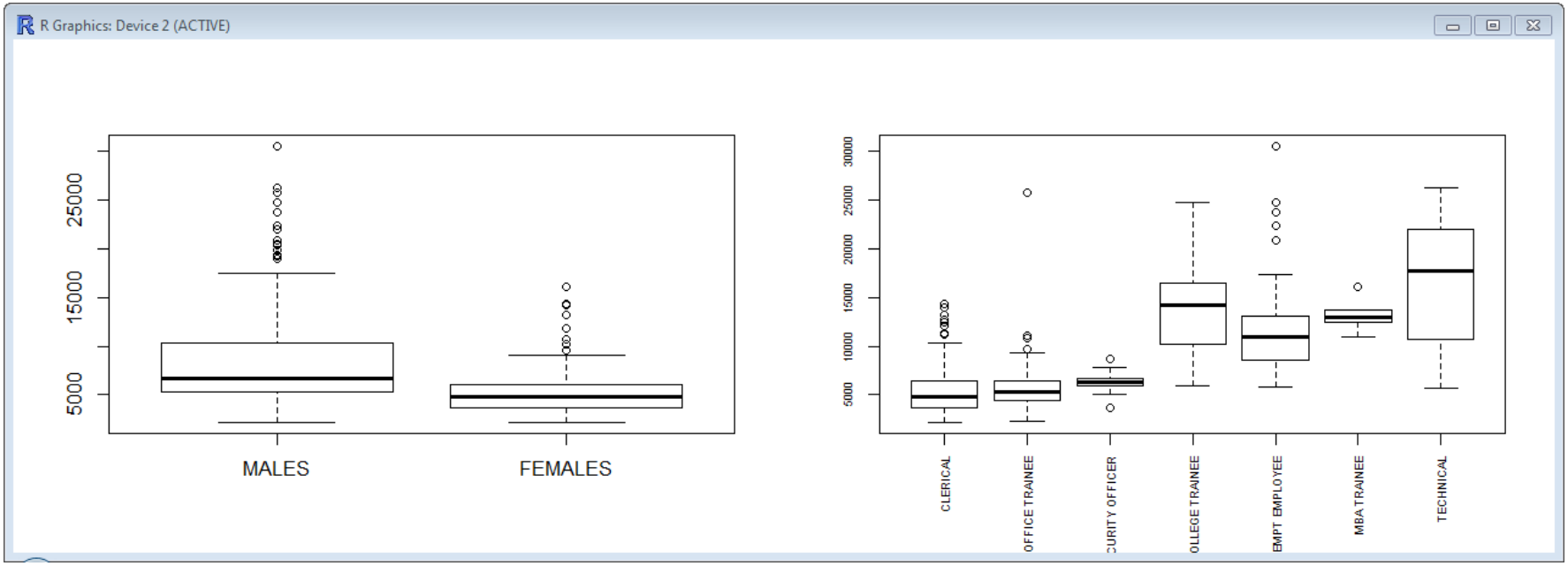



```
par(mfrow=c(1,2))
x<-table(salary$sex,salary$jobcat)
barplot(apply(x,2,prop.table), legend=TRUE, col=2:3, ylim=c(0,1.5))
barplot(apply(x,1,prop.table), legend=TRUE, xlim=c(0,5), col=2:8)
```

3.4. Graphical comparisons of quantitative by levels of a categorical variable

Salary differences by gender and job category

A boxplot for each category

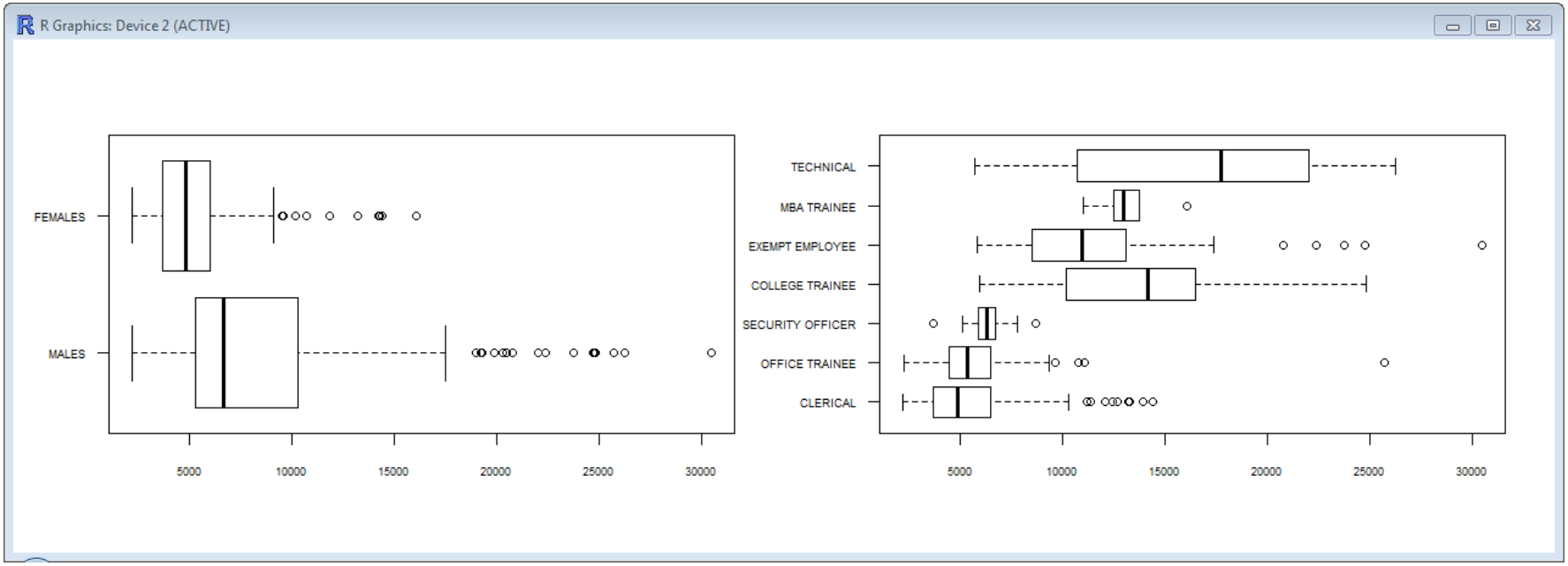



```
x1 <- salary$salnow-salary$salbeg
par(mfrow=c(1,2))
boxplot( split(x1,salary$sex))
boxplot( split(x1,salary$jobcat), cex.axis=0.6, las=3 )
```

3.4. Graphical comparisons of quantitative by levels of a categorical variable

Salary differences by gender and job category

Horizontal box plots

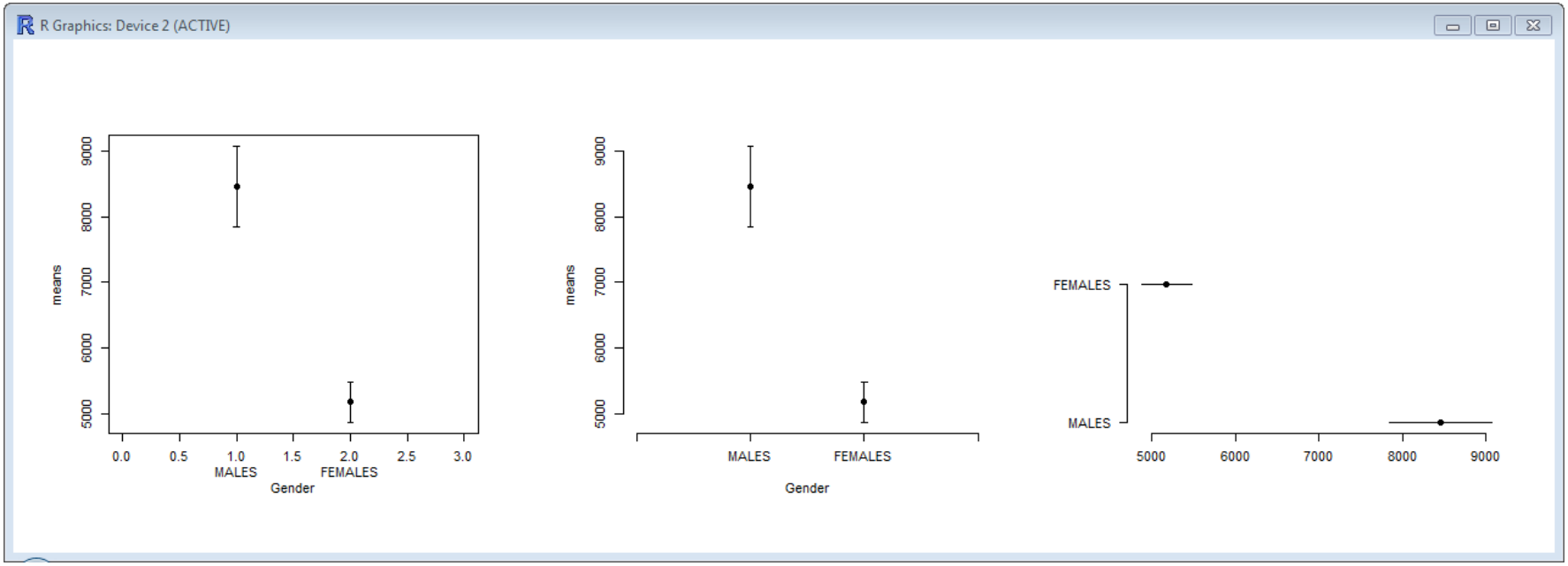



```
x1 <- salary$salnow-salary$salbeg
par(mfrow=c(1,2))
boxplot( split(x1,salary$sex), horizontal=T, las=1, cex.axis=0.6 )
boxplot( split(x1,salary$jobcat), horizontal=T, las=1, cex.axis=0.6 )
```

3.4. Graphical comparisons of quantitative by levels of a categorical variable

Salary differences by gender and job category

Error bars using the command errbar in library Hmisc



3.4. Graphical comparisons of quantitative by levels of a categorical variable

Salary differences by gender and job category

Error bars using the command errbar in library Hmisc

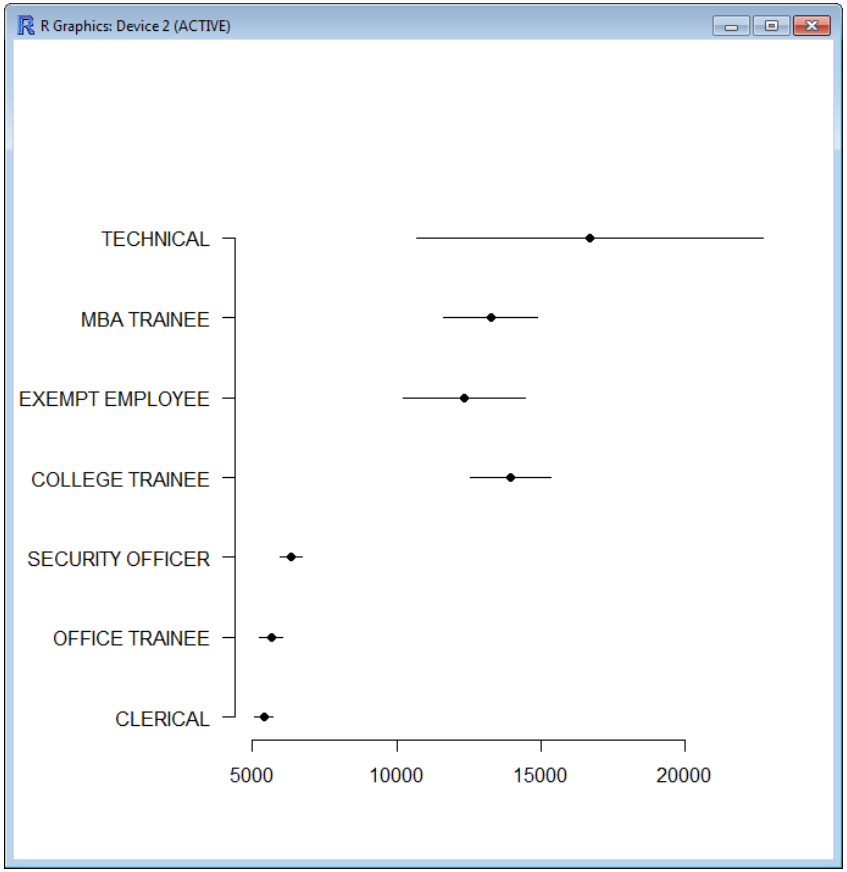
```
library(Hmisc)
a<-0.05
x1 <- salary$salnow-salary$salbeg
means <- by(x1,salary$sex, mean)[1:2]
sds <- by(x1,salary$sex, sd)[1:2]
upper.limit <- means + qnorm(1-a/2)*sds/sqrt(table(salary$sex))
lower.limit <- means + qnorm(a/2)*sds/sqrt(table(salary$sex))
par(mfrow=c(1,3))
# graph A
errbar( 1:2, means, upper.limit, lower.limit, xlim=c(0,3), xlab='Gender')
mtext(levels(salary$sex), side=1, at=c(1,2), line=2.0, cex=0.7)
# graph B
errbar( 1:2, means, upper.limit, lower.limit, xlim=c(0,3), axes=F, xlab='Gender')
axis(2)
axis(1, at=0:3, labels=c('',levels(salary$sex),''))
# graph C
errbar(levels(salary$sex), means, upper.limit, lower.limit)
```

3.4. Graphical comparisons of quantitative by levels of a categorical variable

Salary differences by gender and job category

Error bars using the command errbar in library Hmisc

```
a<-0.05
x1 <- salary$salnow-salary$salbeg
means <- by(x1,salary$jobcat, mean)[1:7]

sds <- by(x1,salary$jobcat, sd)[1:7]
upper.limit <- means + qnorm(1-
    a/2)*sds/sqrt(table(salary$jobcat))
lower.limit <- means +
    qnorm(a/2)*sds/sqrt(table(salary$jobcat))
errbar(levels(salary$jobcat), means,
    upper.limit, lower.limit)
```

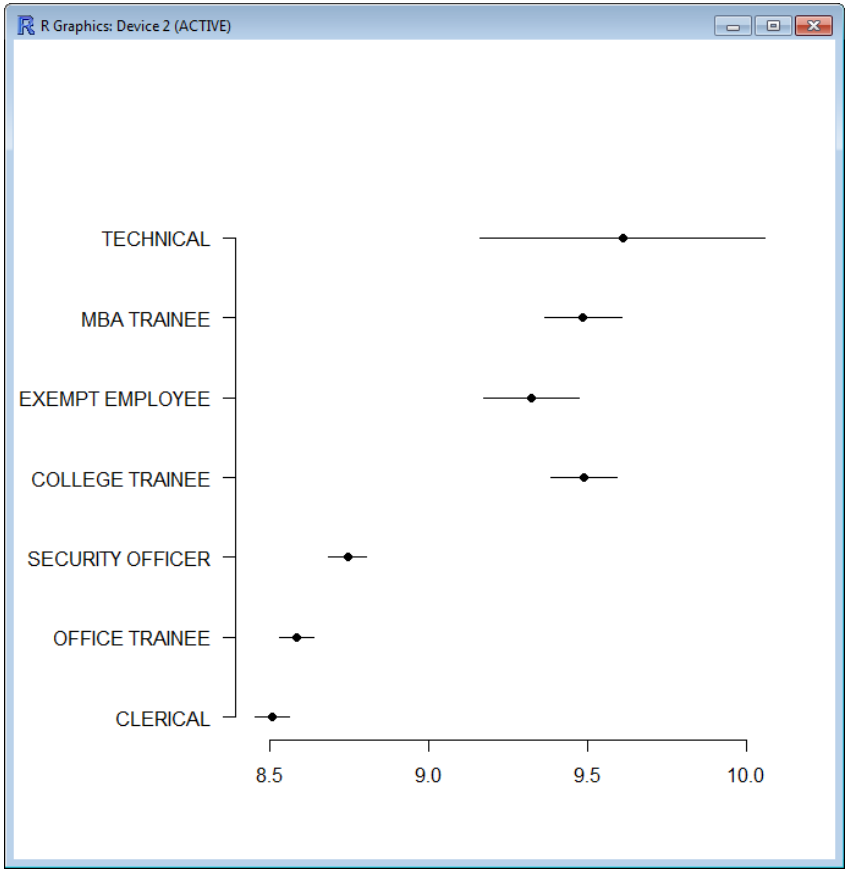


3.4. Graphical comparisons of quantitative by levels of a categorical variable

Log(Salary differences) by job category

Error bars using the command errbar in library Hmisc

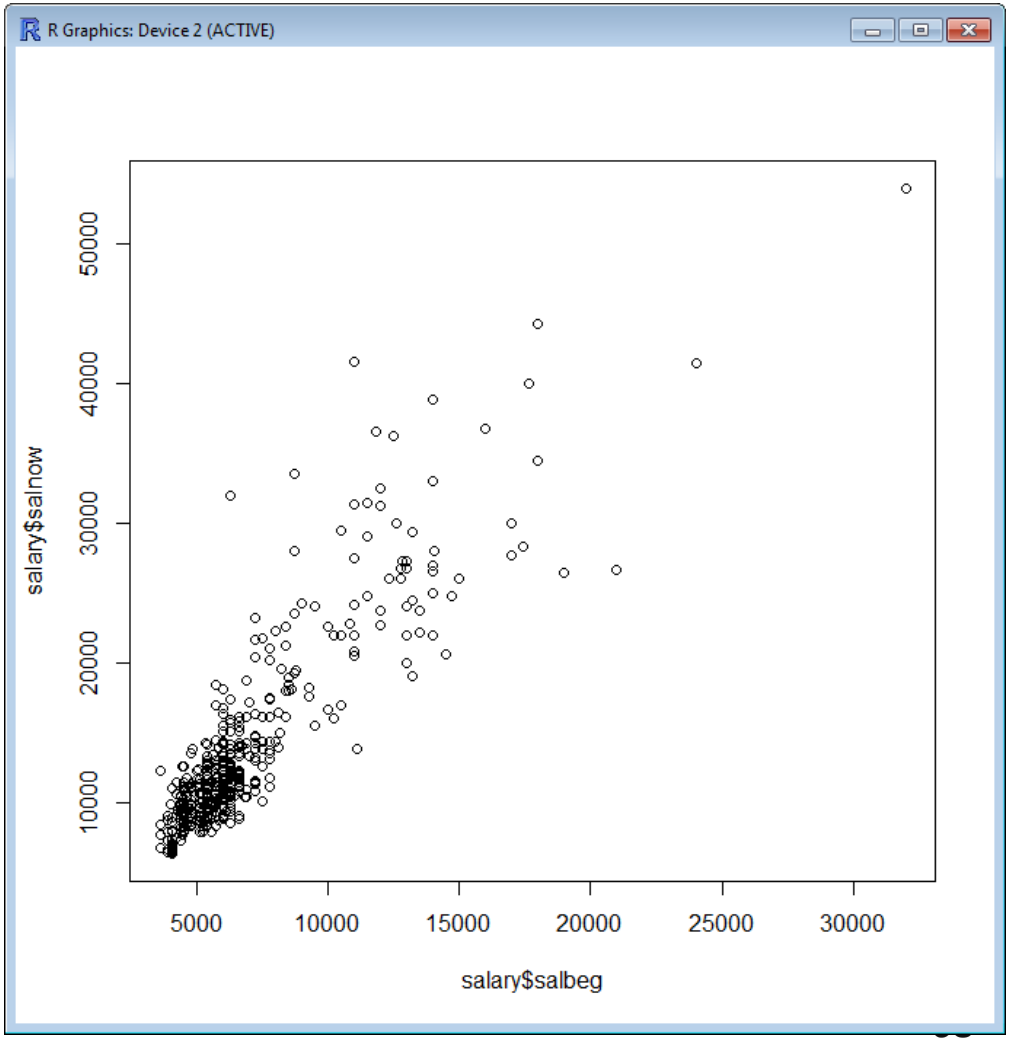
```
a<-0.05
x1 <- log(salary$salnow-salary$salbeg)
means <- by(x1,salary$jobcat, mean)[1:7]

sds <- by(x1,salary$jobcat, sd)[1:7]
upper.limit <- means + qnorm(1-
    a/2)*sds/sqrt(table(salary$jobcat))
lower.limit <- means +
    qnorm(a/2)*sds/sqrt(table(salary$jobcat))
errbar(levels(salary$jobcat), means,
    upper.limit, lower.limit)
```



3.5. Visualization of associations between two quantitative variables: Scatterplots

Salary dataset – salbeg vs salnow

```
plot(salary$salbeg, salary$salnow)
```

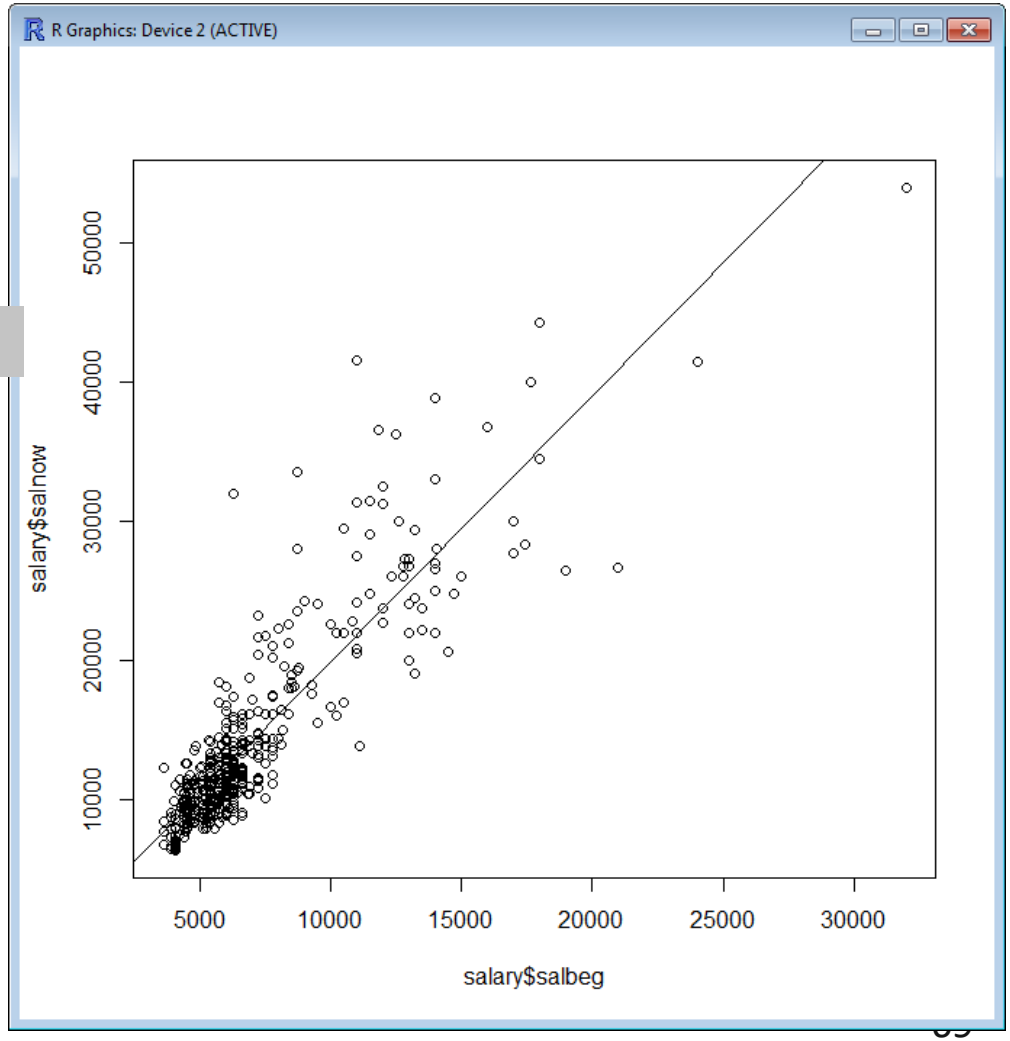


3.5. Visualization of associations between two quantitative variables: Scatterplots

Salary dataset – salbeg vs salnow

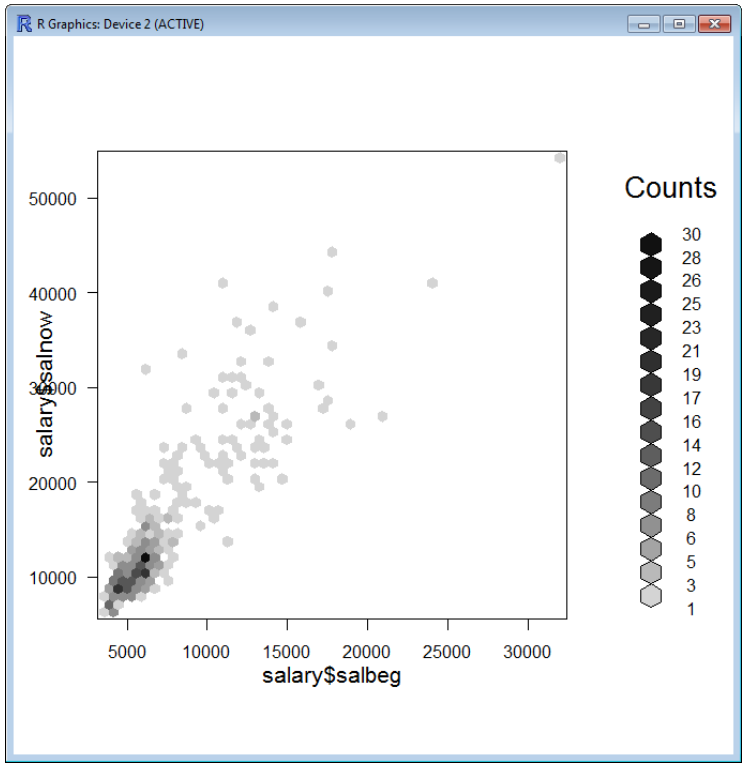
```
plot(salary$salbeg, salary$salnow)
```

```
abline( lm(salnow~salbeg,data=salary) )
```

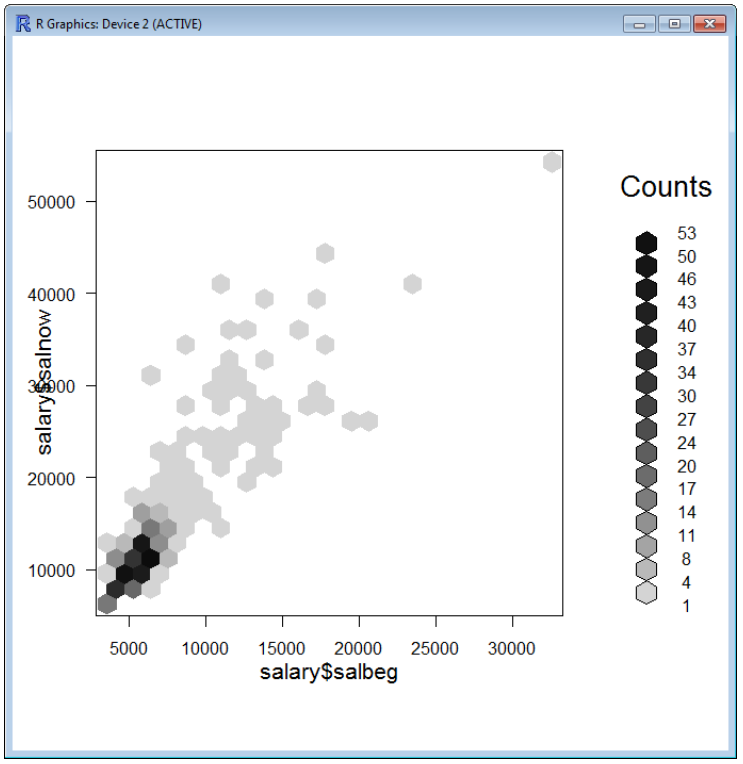


3.5. Visualization of associations between two quantitative variables: Scatterplots

Salary dataset – salbeg vs salnow




```
library(hexbin)
bin<-hexbin(salary$salbeg, salary$salnow,
        xbins=50)
plot(bin)
```

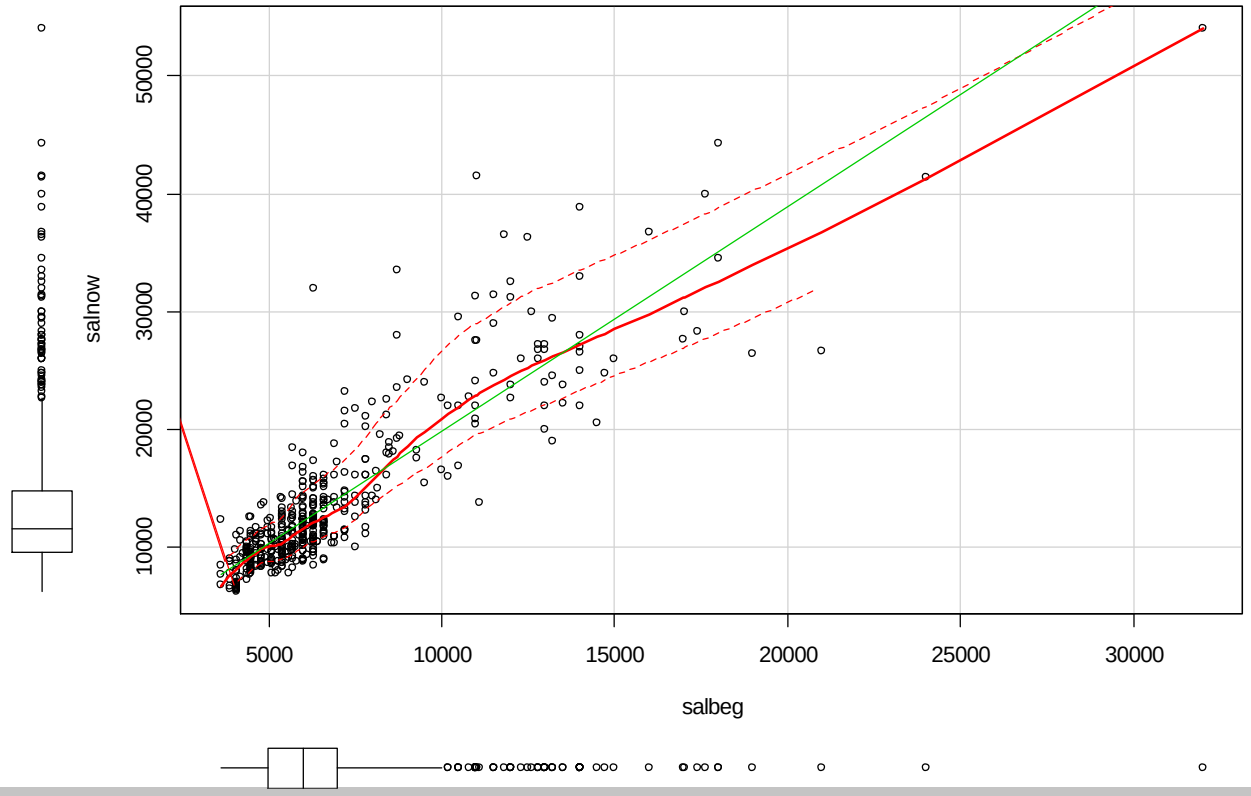



```
bin<-hexbin(salary$salbeg, salary$salnow,
        xbins=25)
plot(bin)
```

3.5. Two quantitative variables & one categorical: Scatterplots

Salary dataset – salbeg vs salnow

Enhanced scatterplot

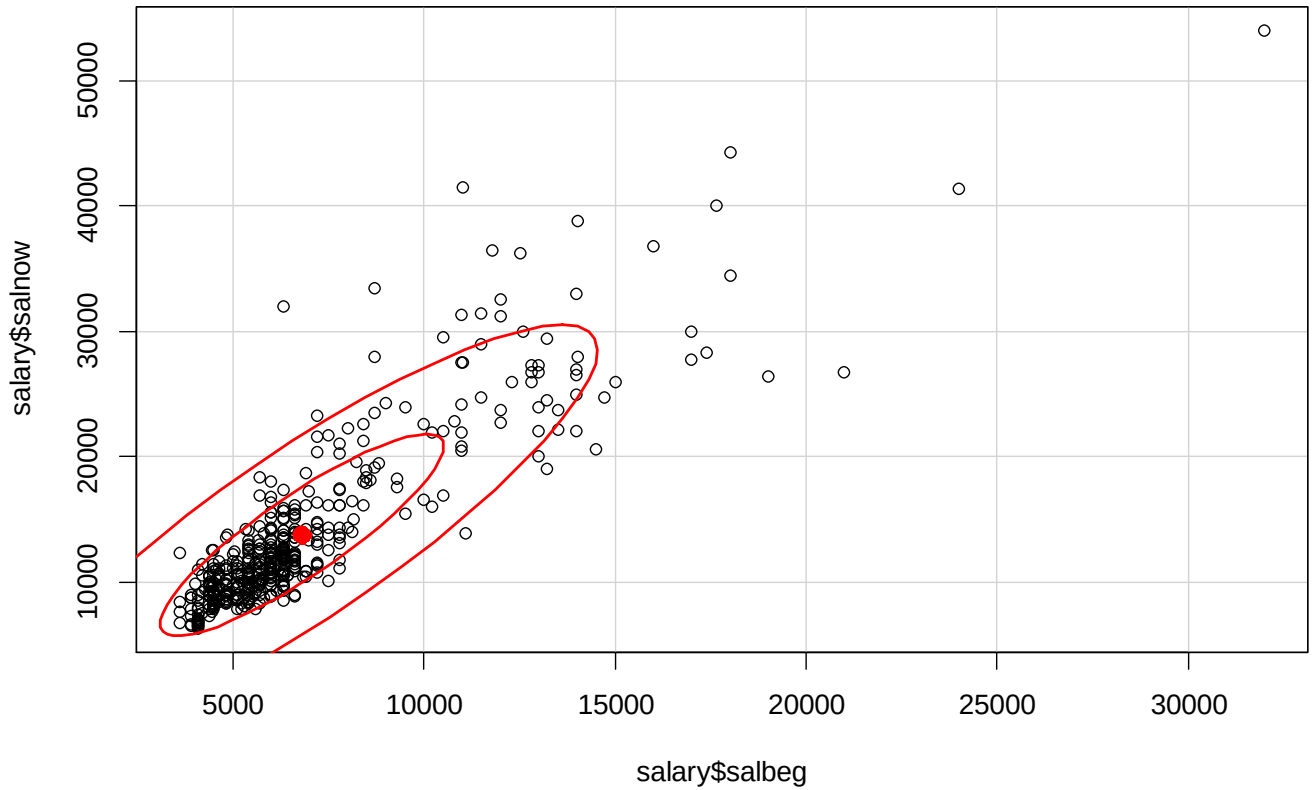



```
library(car)
scatterplot( salnow~salbeg, data=salary)
```

3.5. Two quantitative variables & one categorical: Scatterplots

Salary dataset – salbeg vs salnow

Enhanced scatterplot with data ellipse

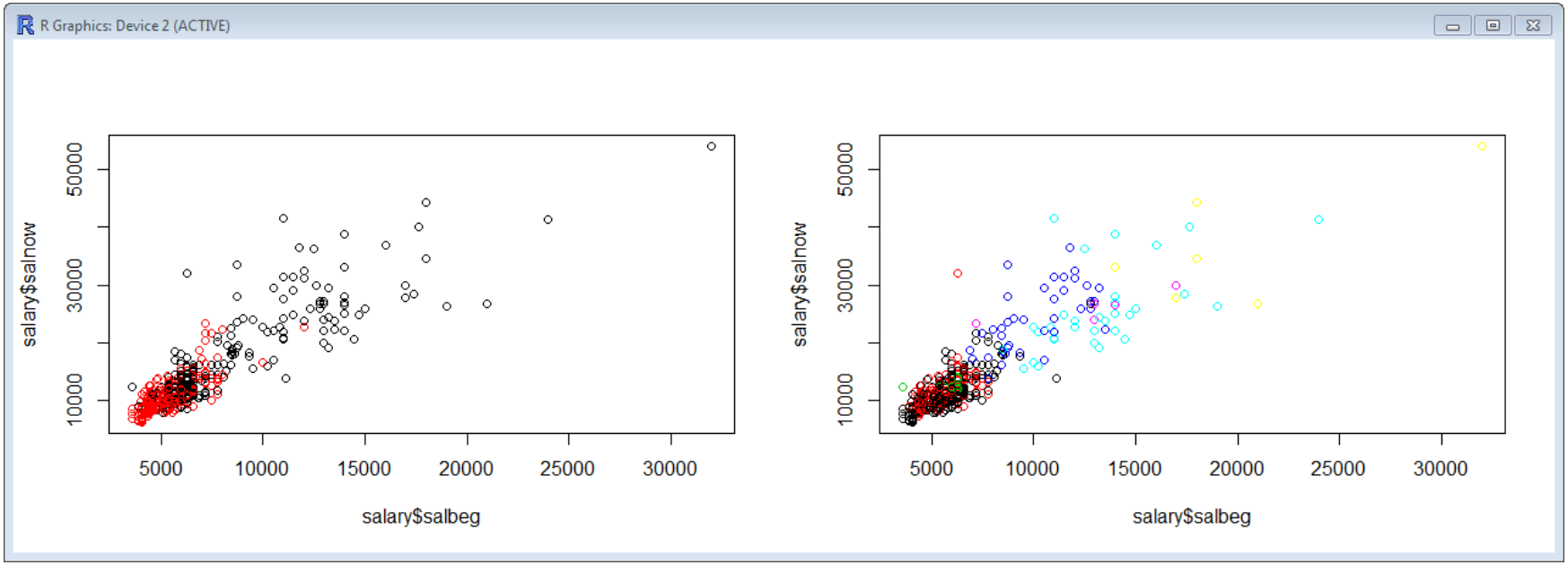



```
library(car)
dataEllipse(salary$salbeg, salary$salnow)
```

3.5. Two quantitative variables & one categorical: Scatterplots

Salary dataset – salbeg vs salnow by gender or jobcat

Categorical is denoted with different color

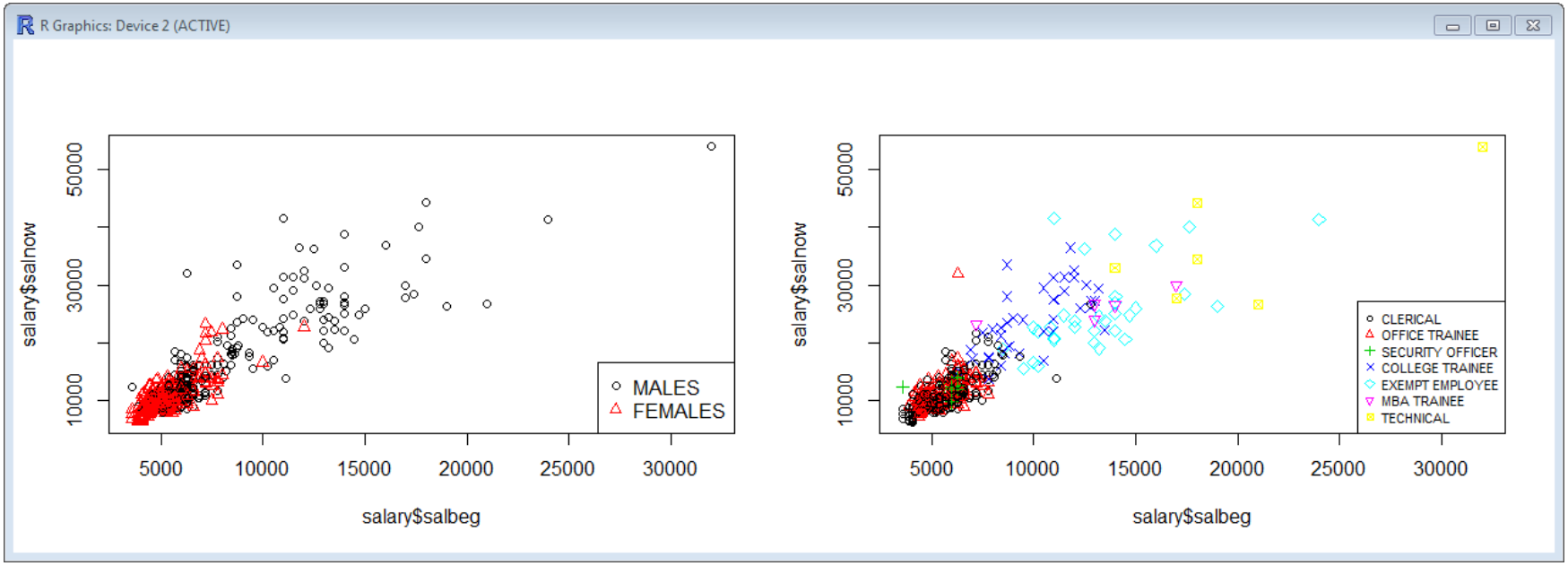



```
par(mfrow=c(1,2))
plot(salary$salbeg, salary$salnow, col=salary$sex)
plot(salary$salbeg, salary$salnow, col=salary$jobcat)
```

3.5. Two quantitative variables & one categorical: Scatterplots

Salary dataset – salbeg vs salnow by gender or jobcat

Categorical is denoted with different color and point type

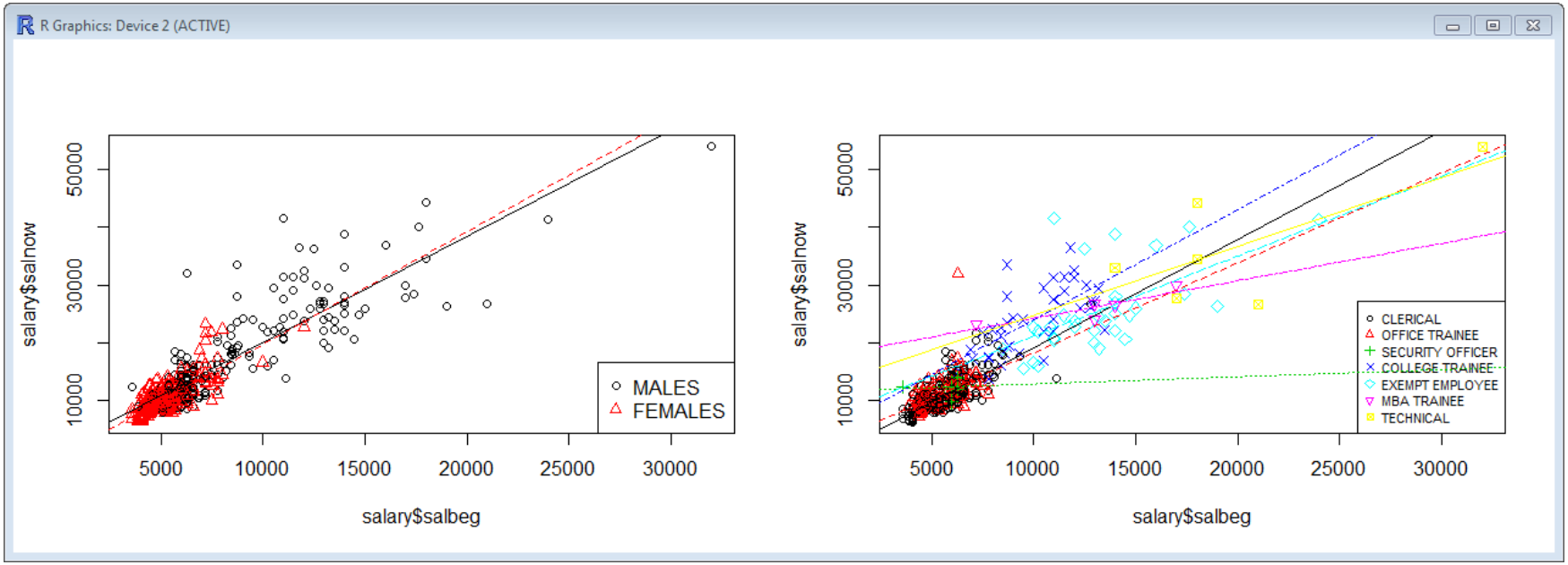



```
par(mfrow=c(1,2))
plot(salary$salbeg, salary$salnow, col=salary$sex, pch=as.numeric(salary$sex))
legend('bottomright', pch=1:2, col=1:2, legend=levels(salary$sex))
plot(salary$salbeg, salary$salnow, col=salary$jobcat, pch=as.numeric(salary$jobcat))
legend('bottomright', pch=1:7, col=1:7, legend=levels(salary$jobcat), cex=0.7)
```

3.5. Two quantitative variables & one categorical: Scatterplots

Salary dataset – salbeg vs salnow by gender or jobcat

Adding regression lines by group



3.5. Two quantitative variables & one categorical: Scatterplots

Salary dataset – salbeg vs salnow by gender or jobcat

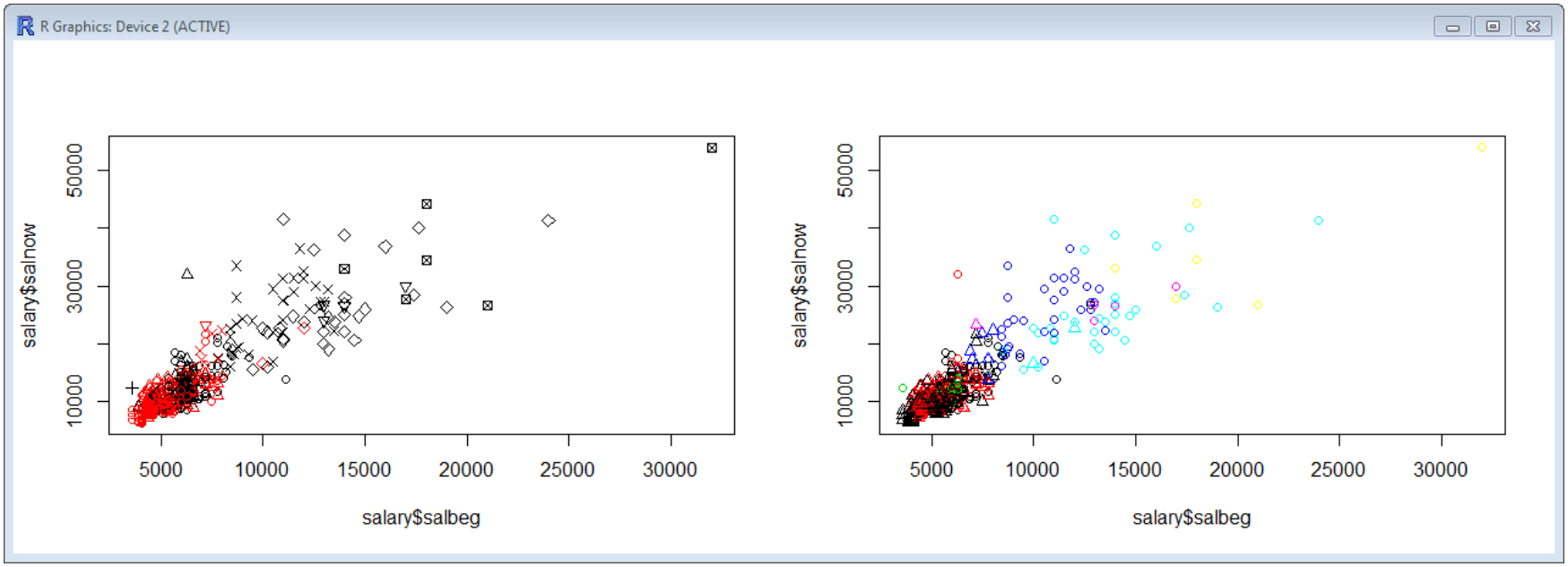
Adding regression lines by group

```
par(mfrow=c(1,2))
plot(salary$salbeg, salary$salnow, col=salary$sex, pch=as.numeric(salary$sex))
legend('bottomright', pch=1:2, col=1:2, legend=levels(salary$sex))
temp <- split(salary, salary$sex)
abline( lm(salnow~salbeg,data=temp[[1]]) )
abline( lm(salnow~salbeg,data=temp[[2]]), col=2,lty=2)
plot(salary$salbeg, salary$salnow, col=salary$jobcat,
     pch=as.numeric(salary$jobcat))
legend('bottomright', pch=1:7, col=1:7, legend=levels(salary$jobcat), cex=0.7)
temp <- split(salary, salary$jobcat)
for (i in 1:7){
     abline( lm(salnow~salbeg,data=temp[[i]]), col=i, lty=i )
}
```

3.5. Two quantitative variables & one categorical: Scatterplots

Salary dataset – salbeg vs salnow by gender and jobcat

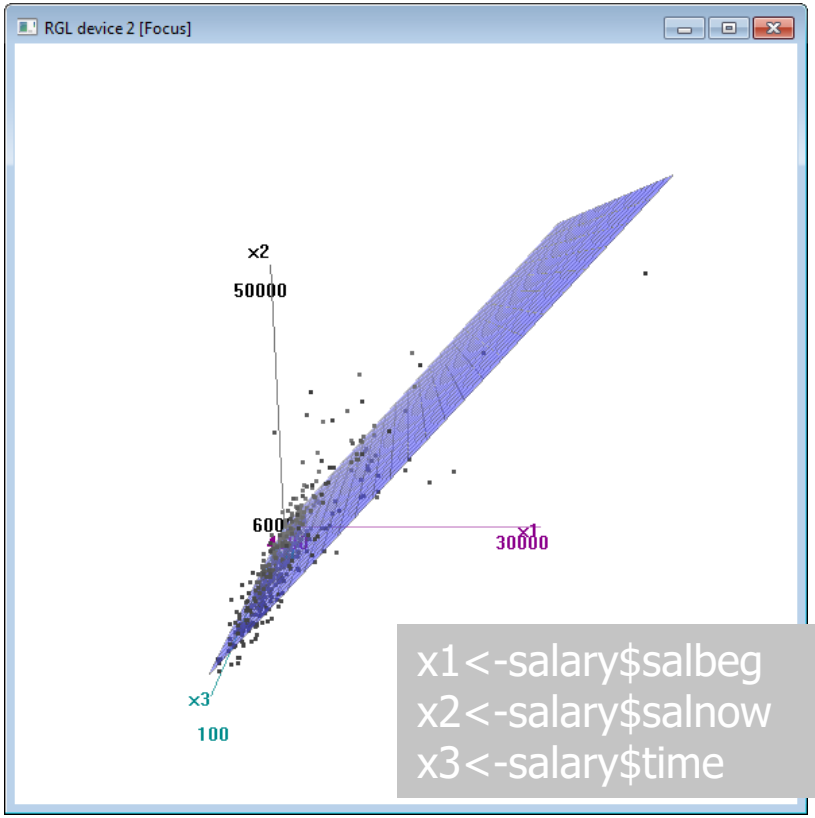
Adding regression lines by group



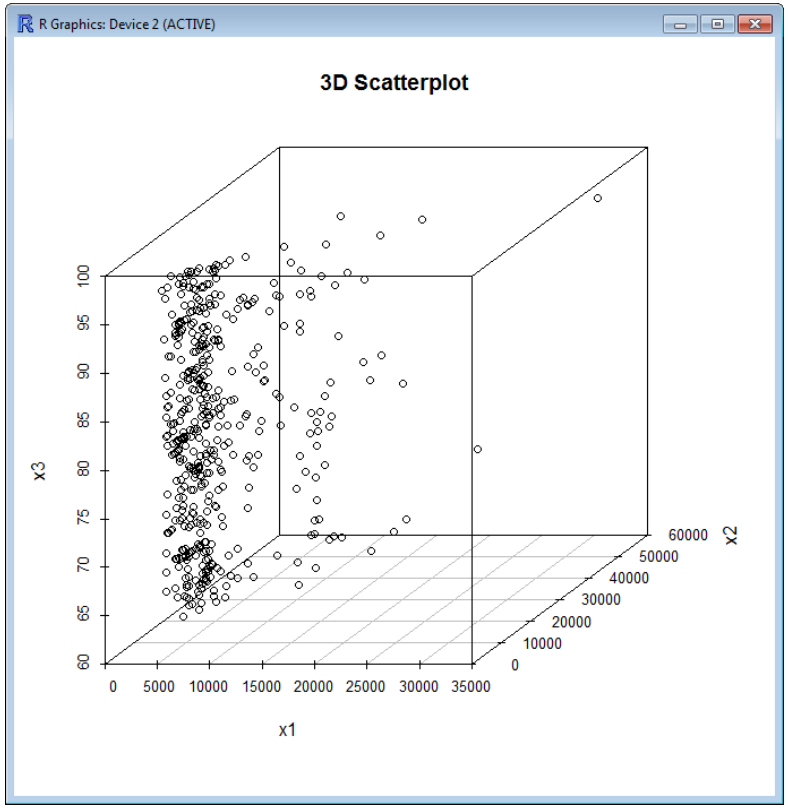

```
par(mfrow=c(1,2))
plot(salary$salbeg, salary$salnow, col=salary$sex, pch=as.numeric(salary$jobcat))
plot(salary$salbeg, salary$salnow, col=salary$jobcat, pch=as.numeric(salary$sex))
```

3.5. Three quantitative variables – 3D scatterplot

Salary dataset – salbeg vs salnow vs time




```
x1<-salary$salbeg
x2<-salary$salnow
x3<-salary$time
```

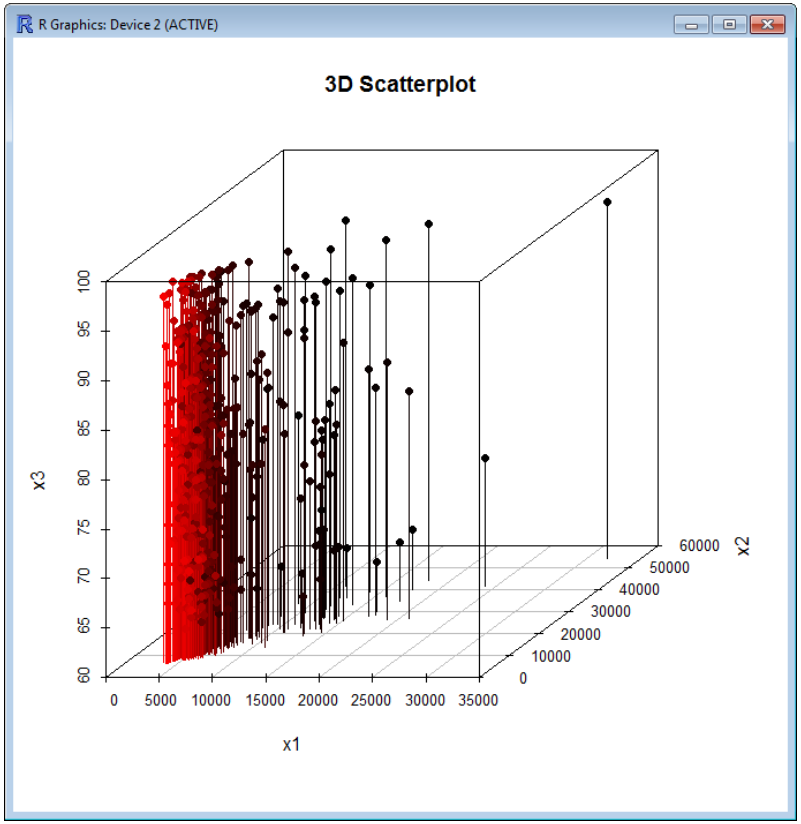


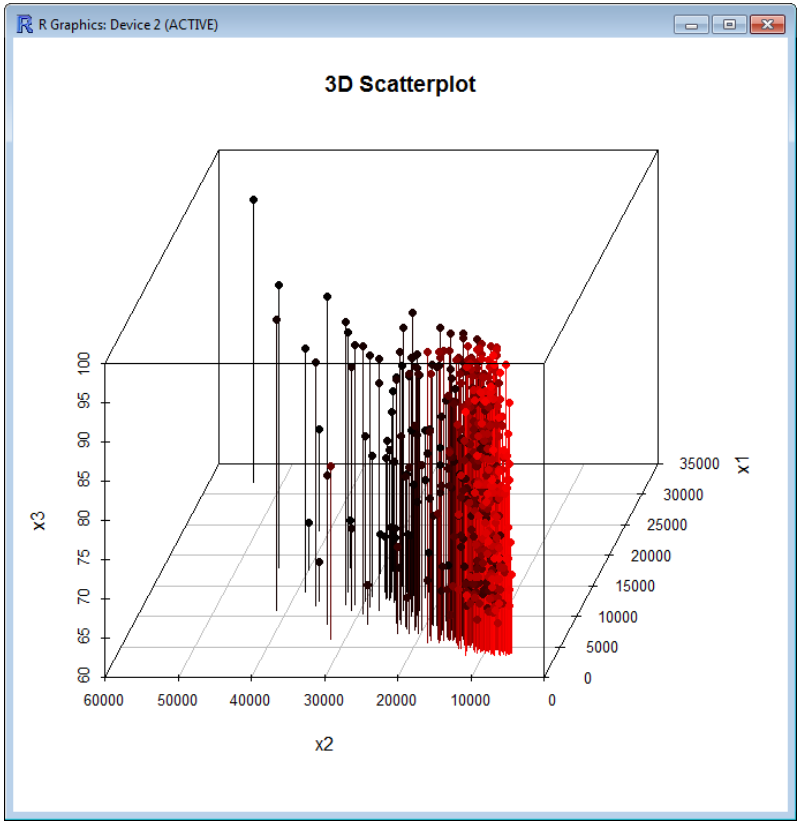

```
library(car)
scatter3d(x1,x2,x3, residuals=F, point.col=1)
```

```
library(scatterplot3d)
scatterplot3d(x1,x2,x3)
```

3.5. Three quantitative variables – 3D scatterplot

Salary dataset – salbeg vs salnow vs time

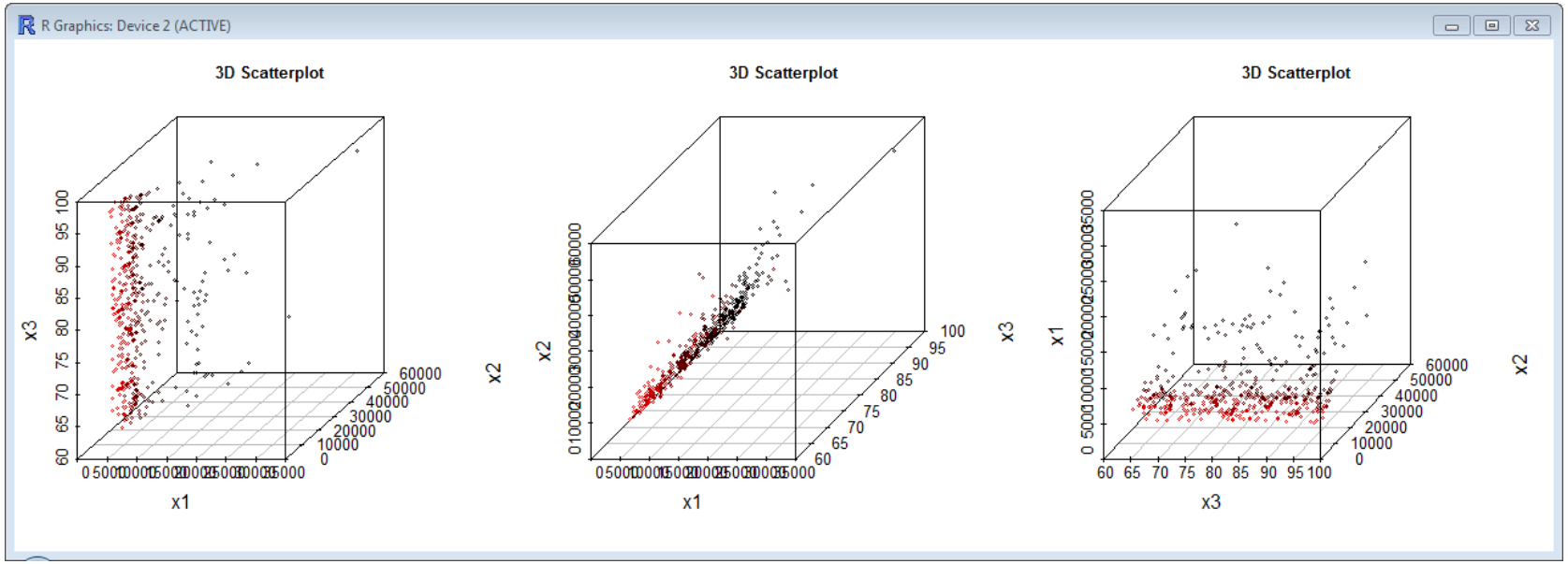





```
scatterplot3d(x1,x2,x3, pch=16, highlight.3d=TRUE,  type="h")
scatterplot3d(x1,x2,x3, pch=16, highlight.3d=TRUE,  type="h", angle=250)
```

3.5. Three quantitative variables – 3D scatterplot

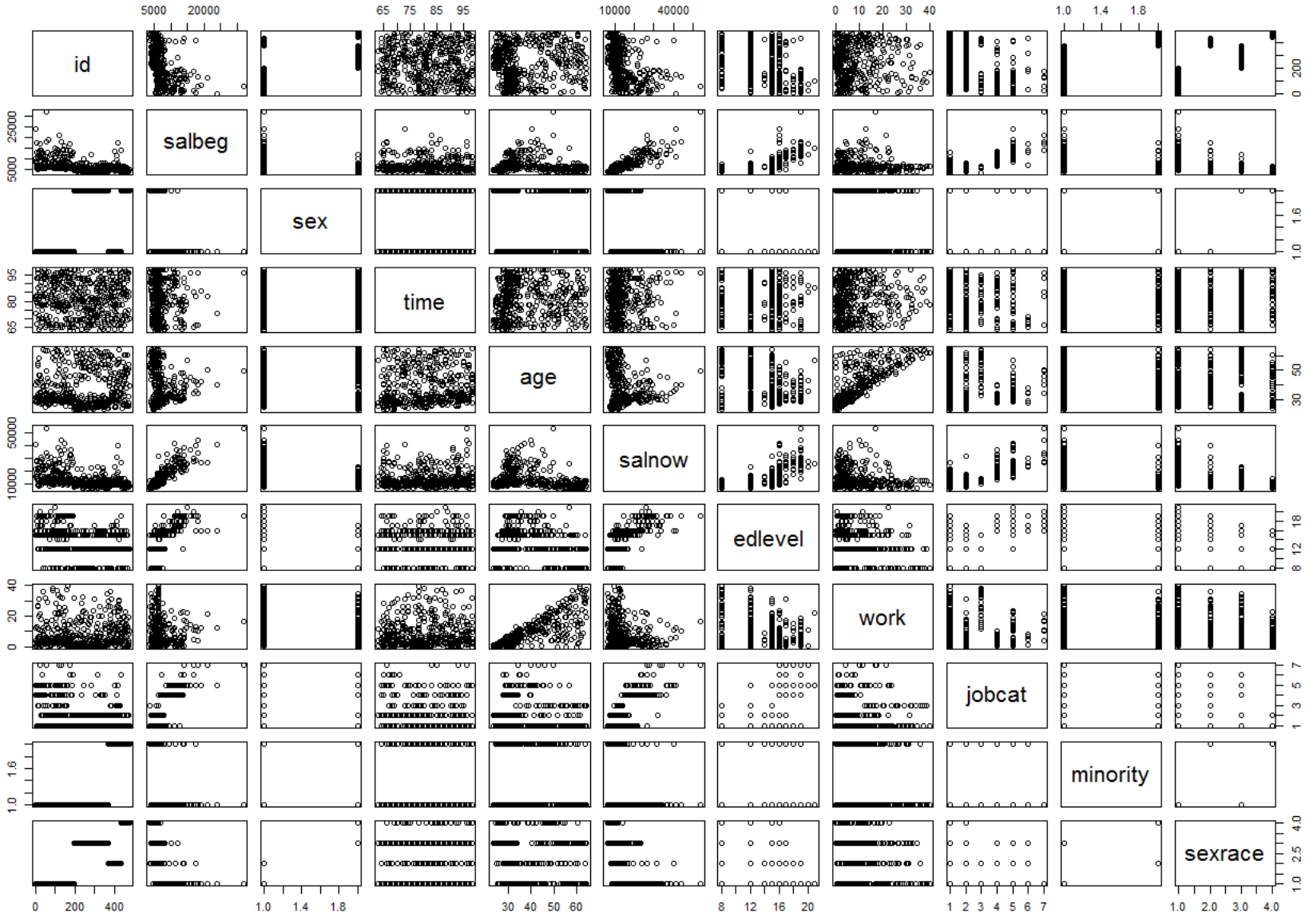
Salary dataset – salbeg vs salnow vs time



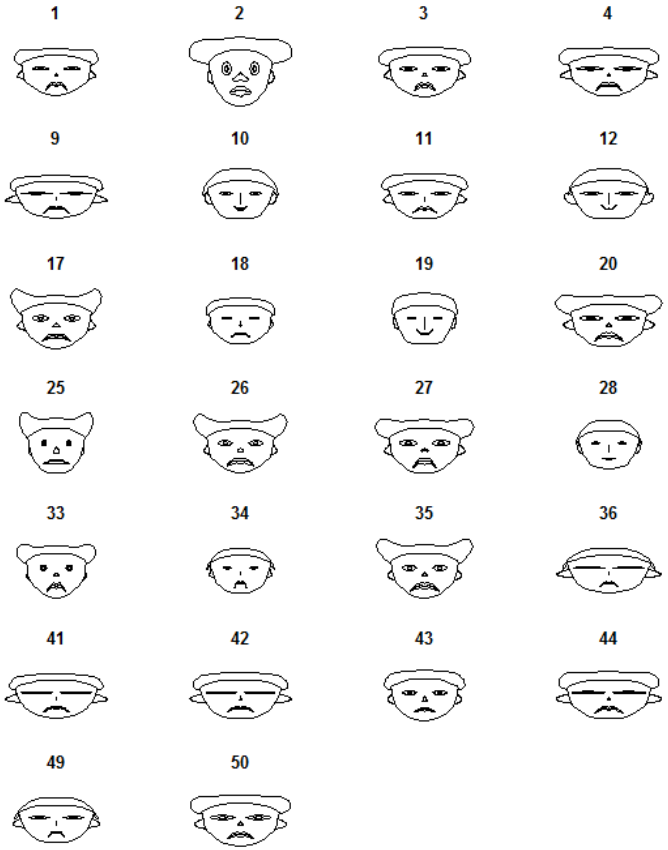

```
library(scatterplot3d)
par(mfrow=c(1,3))
scatterplot3d(x1,x2,x3, main="3D Scatterplot",highlight.3d=TRUE)
scatterplot3d(x1,x3,x2, main="3D Scatterplot",highlight.3d=TRUE)
scatterplot3d(x3,x2,x1, main="3D Scatterplot",highlight.3d=TRUE)
```

3.6. Multivariate graphs
Matrix scatterplot

pairs(salary, cex.labels=2)



3.6. Multivariate graphs
Chernoff faces



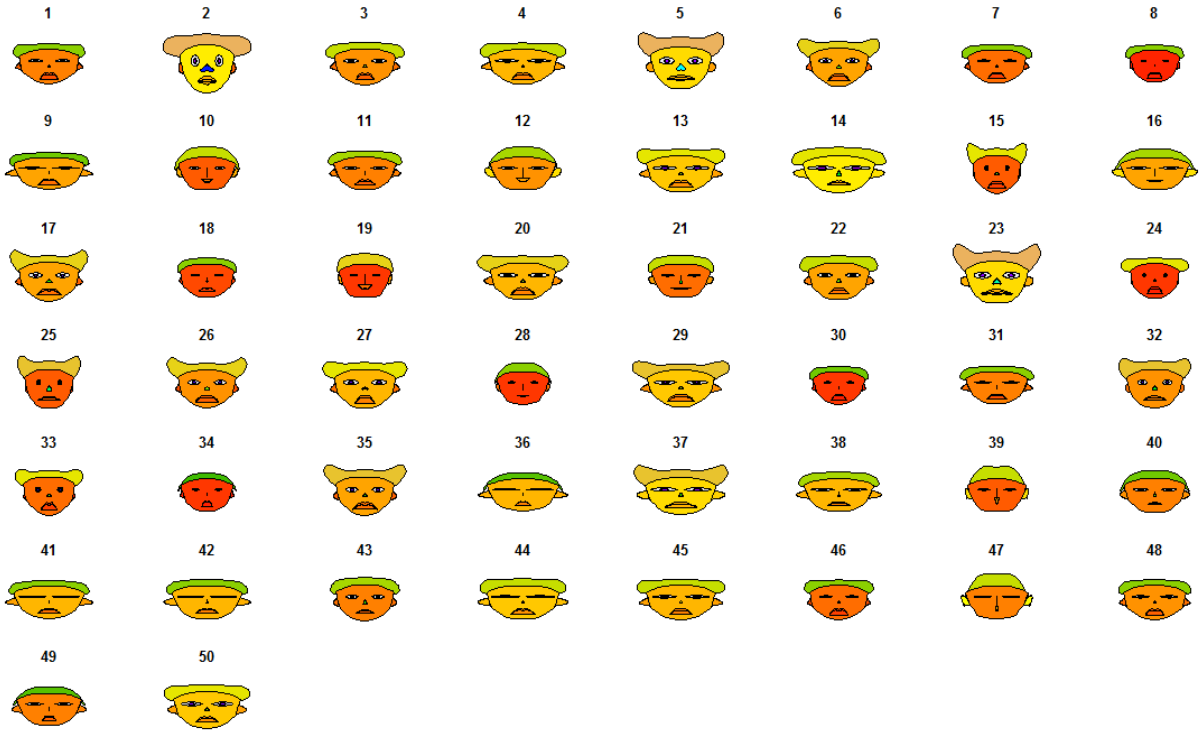

```
> index<-sapply(salary,class)=='numeric'
> sal2<-salary[,index]
> sal2<-sal2[,-1]
> library(aplpack)
> faces(sal2[1:50,], face.type=0)
effect of variables:
 modified item       Var
 "height of face   " "salbeg"
 "width of face    " "time"
 "structure of face" "age"
 "height of mouth  " "salnow"
 "width of mouth   " "edlevel"
 "smiling          " "work"
 "height of eyes   " "salbeg"
 "width of eyes    " "time"
 "height of hair   " "age"
 "width of hair    " "salnow"
 "style of hair    " "edlevel"
 "height of nose   " "work"
 "width of nose    " "salbeg"
 "width of ear     " "time"
 "height of ear    " "age"
>
```

3.6. Multivariate graphs
Chernoff faces


```
effect of variables:
 modified item       Var
 "height of face   " "salbeg"
 "width of face    " "time"
 "structure of face" "age"
 "height of mouth  " "salnow"
 "width of mouth   " "edlevel"
 "smiling          " "work"
 "height of eyes   " "salbeg"
 "width of eyes    " "time"
 "height of hair   " "age"
 "width of hair    "  "salnow"
 "style of hair    "  "edlevel"
 "height of nose   "  "work"
 "width of nose    "  "salbeg"
 "width of ear     "  "time"
 "height of ear    "  "age"
```

face.type=0

3.6. Multivariate graphs
Chernoff faces




```
effect of variables:
 modified item       Var
 "height of face   " "salbeg"
 "width of face    " "time"
 "structure of face" "age"
 "height of mouth  " "salnow"
 "width of mouth   " "edlevel"
 "smiling          " "work"
 "height of eyes   " "salbeg"
 "width of eyes    " "time"
 "height of hair   " "age"
 "width of hair    "  "salnow"
 "style of hair    "  "edlevel"
 "height of nose   "  "work"
 "width of nose    "  "salbeg"
 "width of ear     "  "time"
 "height of ear    "  "age"
```

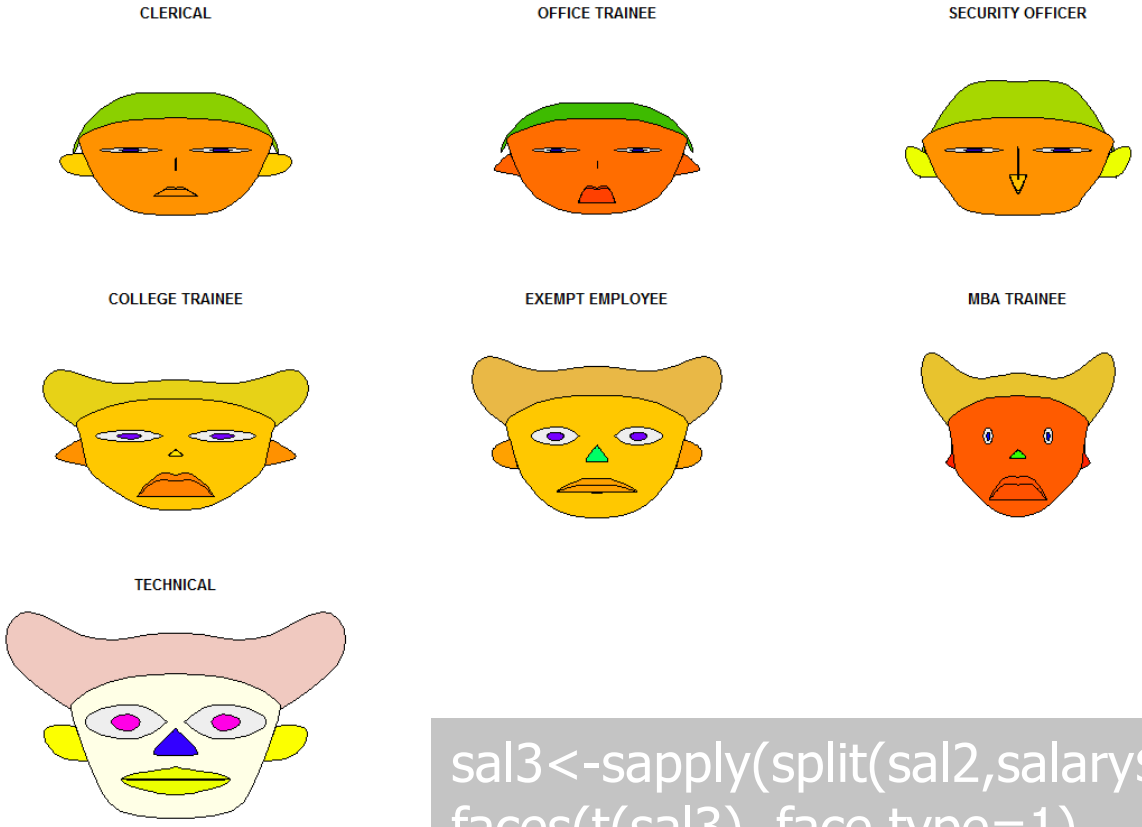
face.type=1

3.6. Multivariate graphs
Chernoff faces


```
effect of variables:
 modified item        Var
 "height of face   " "salbeg"
 "width of face    " "time"
 "structure of face" "age"
 "height of mouth  " "salnow"
 "width of mouth   " "edlevel"
 "smiling          " "work"
 "height of eyes   " "salbeg"
 "width of eyes    " "time"
 "height of hair   " "age"
 "width of hair    " "salnow"
 "style of hair    " "edlevel"
 "height of nose   " "work"
 "width of nose    " "salbeg"
 "width of ear     " "time"
 "height of ear    " "age"
```

face.type=2

Mean characteristics by job category




```
effect of variables:
 modified item       Var
 "height of face   " "salbeg"
 "width of face    " "time"
 "structure of face" "age"
 "height of mouth  " "salnow"
 "width of mouth   " "edlevel"
 "smiling          " "work"
 "height of eyes   " "salbeg"
 "width of eyes    " "time"
 "height of hair   " "age"
 "width of hair    " "salnow"
 "style of hair    " "edlevel"
 "height of nose   " "work"
 "width of nose    " "salbeg"
 "width of ear     " "time"
 "height of ear    " "age"
```

```
sal3<-sapply(split(sal2,salary$jobcat),colMeans)
faces(t(sal3), face.type=1)
```

Mean characteristics by job category


```
effect of variables:
 modified item       Var
 "height of face   " "salbeg"
 "width of face    " "time"
 "structure of face" "age"
 "height of mouth  " "salnow"
 "width of mouth   " "edlevel"
 "smiling          " "work"
 "height of eyes   " "salbeg"
 "width of eyes    " "time"
 "height of hair   " "age"
 "width of hair    " "salnow"
 "style of hair    " "edlevel"
 "height of nose   " "work"
 "width of nose    " "salbeg"
 "width of ear     " "time"
 "height of ear    " "age"
```

```
sal3<-sapply(split(sal2,salary$jobcat),colMeans)
faces(t(sal3), face.type=2)
```

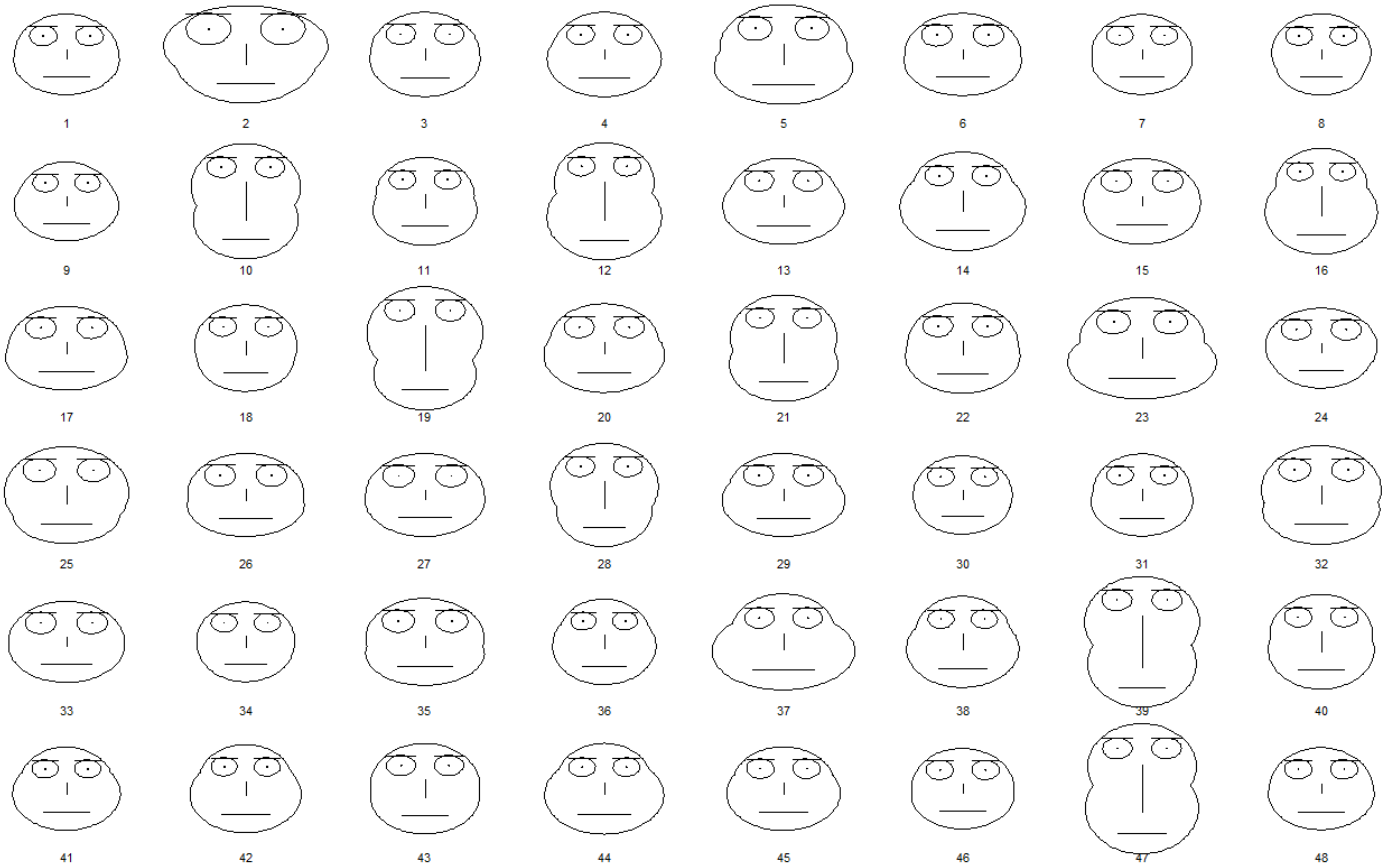
3.6. Multivariate graphs
Chernoff faces


```
library(TeachingDemos)
faces(sal2[1:50,])
```

Details

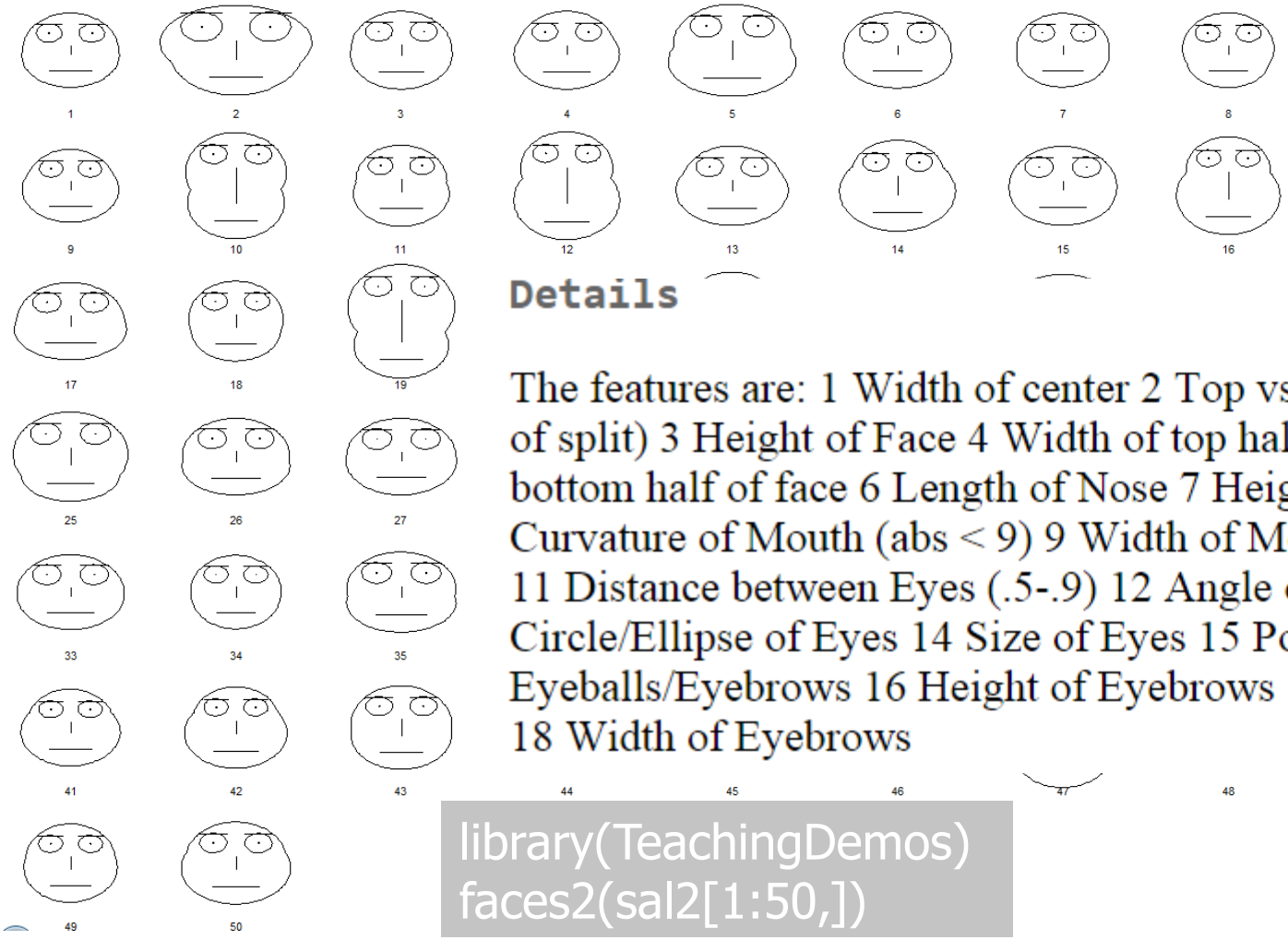
The features paramters of this implementation are: 1-height of face, 2-width of face, 3-shape of face, 4-height of mouth, 5-width of mouth, 6-curve of smile, 7-height of eyes, 8-width of eyes, 9-height of hair, 10-width of hair, 11-styling of hair, 12-height of nose, 13-width of nose, 14-width of ears, 15-height of ears. For details look at the literate program of `faces`

3.6. Multivariate graphs
Chernoff faces




```
library(TeachingDemos)
faces2(sal2[1:50,])
```

3.6. Multivariate graphs
Chernoff faces



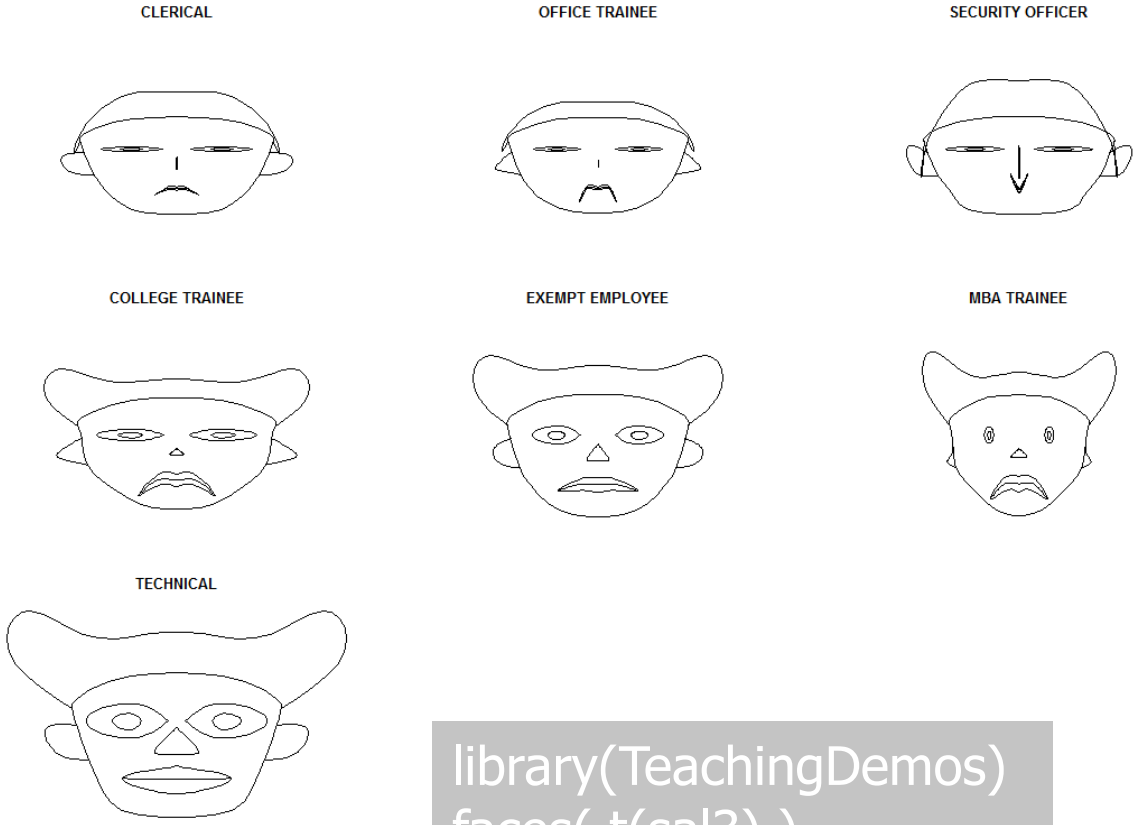
Details

The features are: 1 Width of center 2 Top vs. Bottom width (height of split) 3 Height of Face 4 Width of top half of face 5 Width of bottom half of face 6 Length of Nose 7 Height of Mouth 8 Curvature of Mouth (abs < 9) 9 Width of Mouth 10 Height of Eyes 11 Distance between Eyes (.5-.9) 12 Angle of Eyes/Eyebrows 13 Circle/Ellipse of Eyes 14 Size of Eyes 15 Position Left/Right of Eyeballs/Eyebrows 16 Height of Eyebrows 17 Angle of Eyebrows 18 Width of Eyebrows

```
library(TeachingDemos)
faces2(sal2[1:50,])
```

3.6. Multivariate graphs
Chernoff faces

Mean characteristics by job category

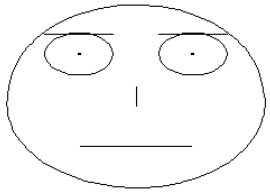



```
library(TeachingDemos)
faces( t(sal3) )
```

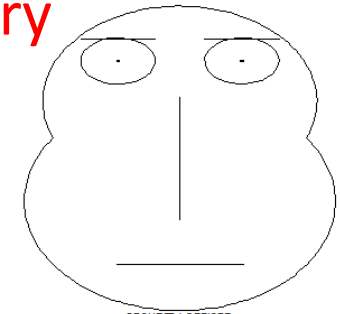
3.6. Multivariate graphs
Chernoff faces

Mean characteristics by job category

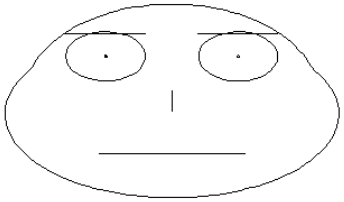
CLERICAL



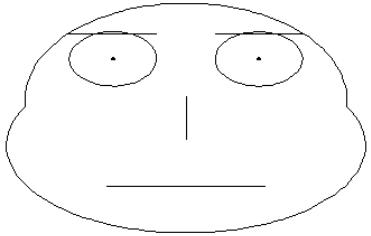
OFFICE TRAINEE



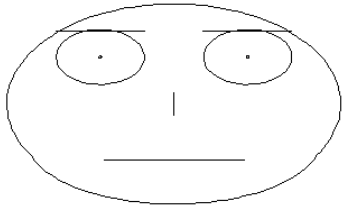
SECURITY OFFICER



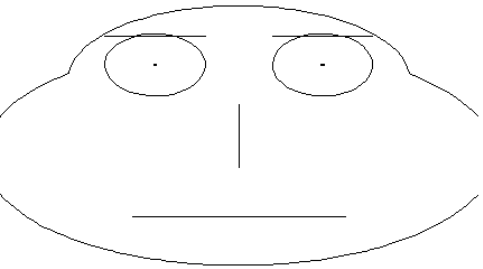
COLLEGE TRAINEE



EXEMPT EMPLOYEE



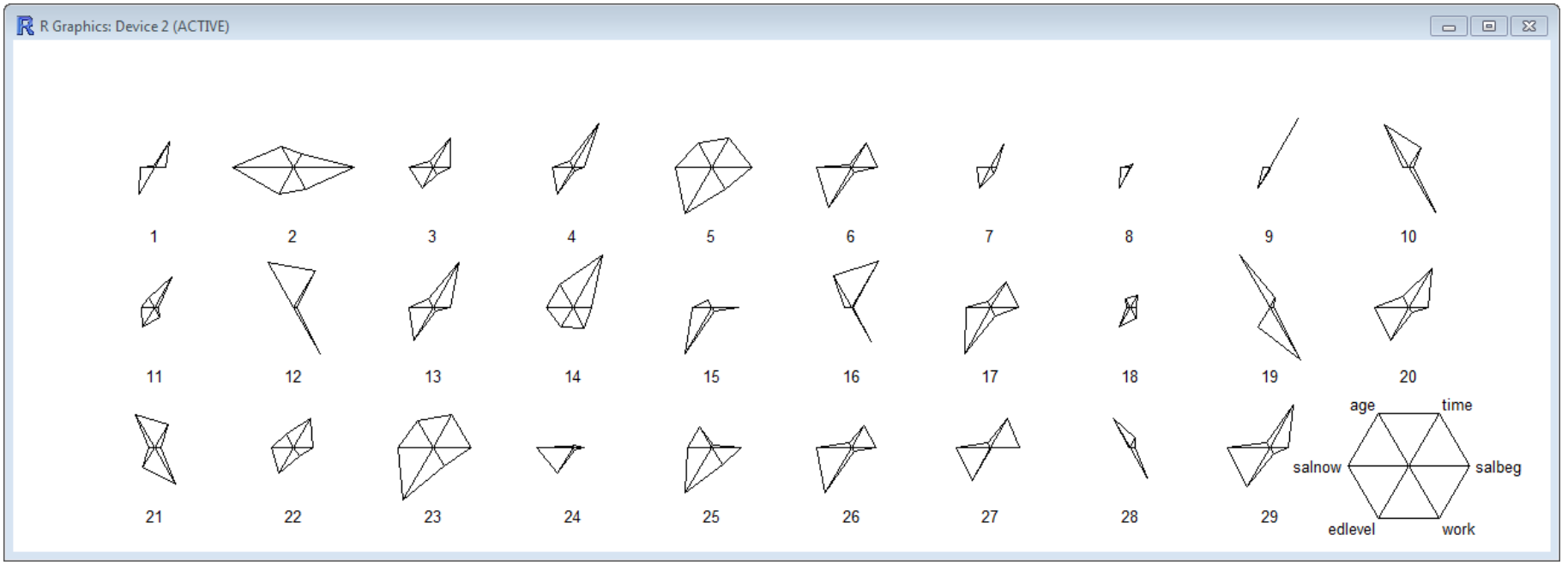
MBA TRAINEE



TECHNICAL

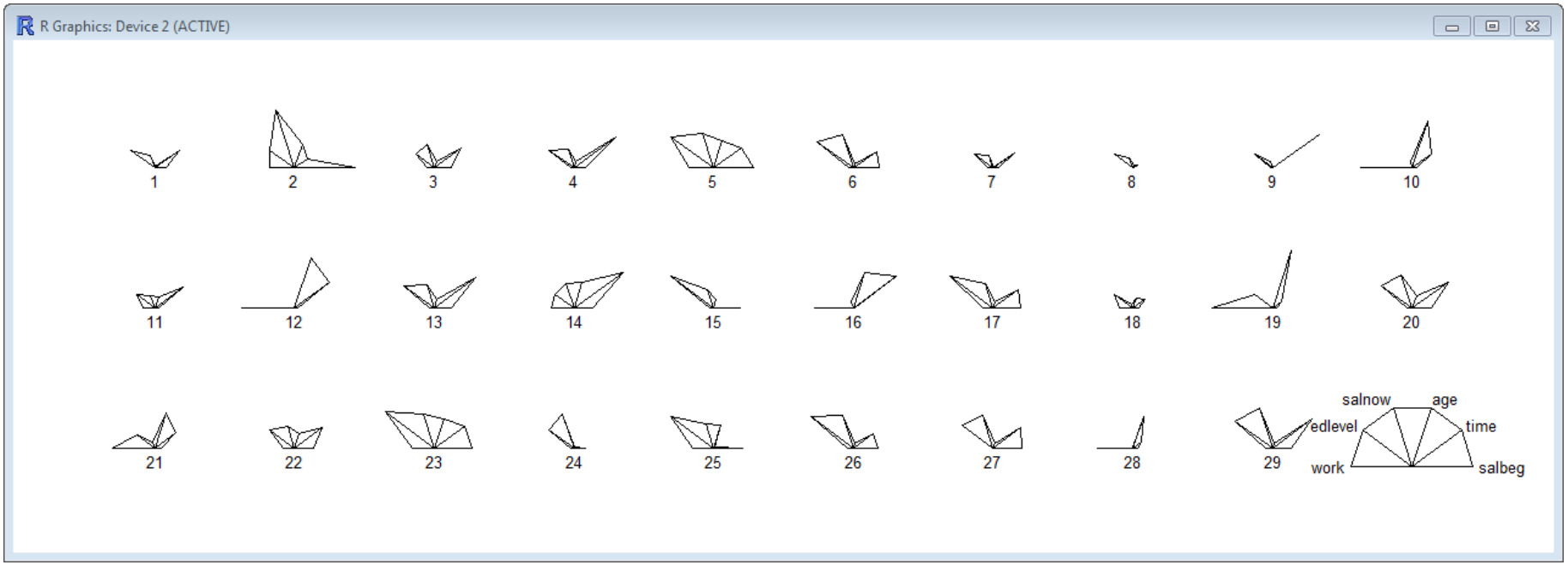
```
library(TeachingDemos)
faces2(t(sal3))
```

Star plots for observations



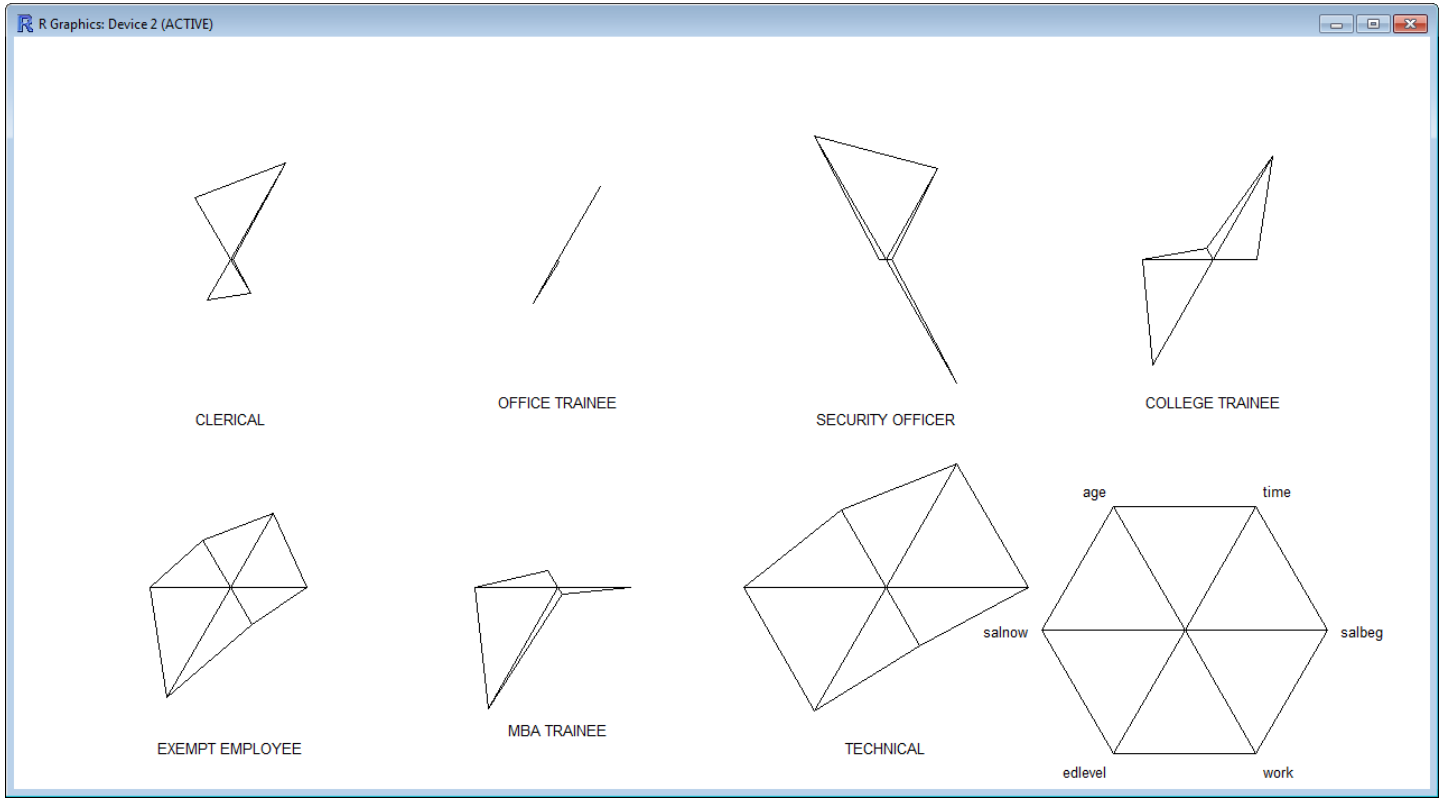

```
stars(sal2[1:29,], key.loc=c(23,2), nrow=3, ncol=10)
```

Half star plots for observations




```
stars(sal2[1:29,], key.loc=c(23,2), nrow=3, ncol=10, full=FALSE)
```

Star plots for means per Job category




```
stars(t(sal3), key.loc=c(9,2), ncol=4, nrow=2)
```