CKME 136 Capstone - Retirement Income Adequacy

Melissa Cortina

## Import data and select attributes

labourincomedata <- read.csv(file="C:/Users/cortinm/Documents/original\_data.csv", header = TRUE)

There are a total of 147 attributs in this dataset. After reading through the associated literature for the dataset, 26 relevant were chosen.

Select only the 26 columns:

attributes <- c(7,8,10,11,14,17,23,26,29,50,51,55,57,58,61,69,72,73,74,76,79,89,114,120,122,140)  
data <- labourincomedata[attributes]  
#write.csv(data, file="C:/Users/huhmyman/Documents/selected\_data")

## Observe dataset

head(data)

## ecage26 ecsex99 marst26 mjacg26 immst15 pvreg25 dwtenr25 mortg25 multj28  
## 1 53 2 1 1 7 10 1 2 2  
## 2 53 1 1 1 7 10 1 2 2  
## 3 26 2 1 4 7 10 1 1 2  
## 4 55 2 1 1 7 10 1 1 2  
## 5 57 1 1 1 7 10 1 1 2  
## 6 21 1 6 2 7 10 1 1 2  
## alhrp28 mtlswk28 yrxfte11 clwkr1 prmjb1 fllprt1 nocg2e6 imphwe1 penpln1  
## 1 2086 999 32 1 1 1 5 24.00 2  
## 2 3129 999 97 1 1 1 3 37.44 2  
## 3 697 999 97 1 1 7 7 33.00 1  
## 4 2086 999 35 1 1 1 12 15.00 2  
## 5 2086 999 35 1 1 1 6 25.24 1  
## 6 156 999 1 1 2 2 11 10.00 2  
## uncoll1 n07c3g10 pubpv10 cqpc42 rppc42 udpd42 wgsal42 hleveg18  
## 1 3 12 1 1750 2800 450 40000 9  
## 2 3 15 2 2300 0 0 105000 6  
## 3 1 12 1 1500 2200 400 32000 11  
## 4 3 6 2 1350 550 125 33000 10  
## 5 1 16 1 2100 3800 825 60000 9  
## 6 3 13 1 675 0 0 17000 8

tail(data)

## ecage26 ecsex99 marst26 mjacg26 immst15 pvreg25 dwtenr25 mortg25  
## 47700 63 1 4 3 7 24 1 1  
## 47701 44 1 6 1 7 24 1 1  
## 47702 34 1 4 1 7 24 2 9  
## 47703 21 1 6 2 2 24 2 9  
## 47704 34 1 1 1 7 24 2 9  
## 47705 44 2 4 1 7 47 1 1  
## multj28 alhrp28 mtlswk28 yrxfte11 clwkr1 prmjb1 fllprt1 nocg2e6  
## 47700 2 0 24 21 99 9 9 99  
## 47701 2 1040 999 22 1 1 1 8  
## 47702 2 2086 999 14 1 7 1 2  
## 47703 2 701 4 0 1 1 1 17  
## 47704 2 2466 999 18 4 7 1 3  
## 47705 1 1668 999 15 1 7 2 8  
## imphwe1 penpln1 uncoll1 n07c3g10 pubpv10 cqpc42 rppc42 udpd42  
## 47700 999.99 9 9 99 9 0 0 0  
## 47701 12.00 2 3 12 2 400 0 0  
## 47702 26.40 2 3 9 2 875 0 25  
## 47703 9.50 2 7 14 2 0 0 0  
## 47704 999.99 9 9 10 2 0 0 1050  
## 47705 21.00 2 3 12 2 1100 0 0  
## wgsal42 hleveg18  
## 47700 0 6  
## 47701 12500 9  
## 47702 20000 9  
## 47703 1750 7  
## 47704 0 97  
## 47705 18000 9

summary(data)

## ecage26 ecsex99 marst26 mjacg26   
## Min. :16.00 Min. :1.00 Min. : 1.000 Min. :1.000   
## 1st Qu.:33.00 1st Qu.:1.00 1st Qu.: 1.000 1st Qu.:1.000   
## Median :49.00 Median :2.00 Median : 1.000 Median :1.000   
## Mean :48.13 Mean :1.52 Mean : 2.917 Mean :2.011   
## 3rd Qu.:62.00 3rd Qu.:2.00 3rd Qu.: 5.000 3rd Qu.:3.000   
## Max. :80.00 Max. :2.00 Max. :97.000 Max. :4.000   
## immst15 pvreg25 dwtenr25 mortg25   
## Min. :1.000 Min. :10.00 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:24.00 1st Qu.:1.000 1st Qu.:1.000   
## Median :7.000 Median :35.00 Median :1.000 Median :2.000   
## Mean :5.585 Mean :34.56 Mean :1.203 Mean :2.968   
## 3rd Qu.:7.000 3rd Qu.:47.00 3rd Qu.:1.000 3rd Qu.:2.000   
## Max. :7.000 Max. :97.00 Max. :7.000 Max. :9.000   
## multj28 alhrp28 mtlswk28 yrxfte11 clwkr1   
## Min. :1.000 Min. : 0 Min. : 0 Min. : 0 Min. : 1.00   
## 1st Qu.:2.000 1st Qu.: 521 1st Qu.:997 1st Qu.: 6 1st Qu.: 1.00   
## Median :2.000 Median :1892 Median :999 Median :21 Median : 1.00   
## Mean :2.925 Mean :2642 Mean :810 Mean :27 Mean :33.63   
## 3rd Qu.:2.000 3rd Qu.:2174 3rd Qu.:999 3rd Qu.:36 3rd Qu.:99.00   
## Max. :9.000 Max. :9999 Max. :999 Max. :98 Max. :99.00   
## prmjb1 fllprt1 nocg2e6 imphwe1   
## Min. :1.000 Min. :1.000 Min. : 1.00 Min. : 6.00   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.: 9.00 1st Qu.: 18.07   
## Median :2.000 Median :1.000 Median :19.00 Median : 36.00   
## Mean :4.819 Mean :3.872 Mean :43.21 Mean : 430.70   
## 3rd Qu.:9.000 3rd Qu.:9.000 3rd Qu.:99.00 3rd Qu.: 999.99   
## Max. :9.000 Max. :9.000 Max. :99.00 Max. : 999.99   
## penpln1 uncoll1 n07c3g10 pubpv10   
## Min. :1.000 Min. :1.000 Min. : 1.00 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:3.000 1st Qu.: 6.00 1st Qu.:2.000   
## Median :2.000 Median :3.000 Median :12.00 Median :2.000   
## Mean :4.719 Mean :5.164 Mean :39.27 Mean :4.146   
## 3rd Qu.:9.000 3rd Qu.:9.000 3rd Qu.:99.00 3rd Qu.:9.000   
## Max. :9.000 Max. :9.000 Max. :99.00 Max. :9.000   
## cqpc42 rppc42 udpd42 wgsal42   
## Min. : 0.0 Min. : 0 Min. : 0.0 Min. : 0   
## 1st Qu.: 0.0 1st Qu.: 0 1st Qu.: 0.0 1st Qu.: 0   
## Median : 450.0 Median : 0 Median : 0.0 Median : 12500   
## Mean : 908.3 Mean : 745 Mean : 152.3 Mean : 27638   
## 3rd Qu.:2100.0 3rd Qu.: 0 3rd Qu.: 0.0 3rd Qu.: 44000   
## Max. :4600.0 Max. :18000 Max. :6250.0 Max. :1400000   
## hleveg18   
## Min. : 1.000   
## 1st Qu.: 6.000   
## Median : 9.000   
## Mean : 7.914   
## 3rd Qu.: 9.000   
## Max. :97.000

str(data)

## 'data.frame': 47705 obs. of 26 variables:  
## $ ecage26 : int 53 53 26 55 57 21 37 38 48 46 ...  
## $ ecsex99 : int 2 1 2 2 1 1 2 1 1 2 ...  
## $ marst26 : int 1 1 1 1 1 6 1 1 1 1 ...  
## $ mjacg26 : int 1 1 4 1 1 2 1 4 1 1 ...  
## $ immst15 : int 7 7 7 7 7 7 7 7 7 7 ...  
## $ pvreg25 : int 10 10 10 10 10 10 10 10 10 10 ...  
## $ dwtenr25: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ mortg25 : int 2 2 1 1 1 1 1 1 1 1 ...  
## $ multj28 : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ alhrp28 : int 2086 3129 697 2086 2086 156 1943 0 1955 1955 ...  
## $ mtlswk28: int 999 999 999 999 999 999 999 21 999 999 ...  
## $ yrxfte11: int 32 97 97 35 35 1 15 97 30 29 ...  
## $ clwkr1 : int 1 1 1 1 1 1 1 99 1 1 ...  
## $ prmjb1 : int 1 1 1 1 1 2 1 9 1 1 ...  
## $ fllprt1 : int 1 1 7 1 1 2 1 9 1 1 ...  
## $ nocg2e6 : int 5 3 7 12 6 11 4 99 9 12 ...  
## $ imphwe1 : num 24 37.4 33 15 25.2 ...  
## $ penpln1 : int 2 2 1 2 1 2 1 9 1 1 ...  
## $ uncoll1 : int 3 3 1 3 1 3 3 9 1 3 ...  
## $ n07c3g10: int 12 15 12 6 16 13 9 99 12 6 ...  
## $ pubpv10 : int 1 2 1 2 1 1 2 9 1 2 ...  
## $ cqpc42 : int 1750 2300 1500 1350 2100 675 1500 0 2200 2600 ...  
## $ rppc42 : int 2800 0 2200 550 3800 0 525 0 4200 0 ...  
## $ udpd42 : int 450 0 400 125 825 0 0 0 625 300 ...  
## $ wgsal42 : int 40000 105000 32000 33000 60000 17000 34000 0 57500 135000 ...  
## $ hleveg18: int 9 6 11 10 9 8 9 9 10 11 ...

Check if there are any missing entries:

apply(data, 2, function(x) sum(is.na(x)))

## ecage26 ecsex99 marst26 mjacg26 immst15 pvreg25 dwtenr25 mortg25   
## 0 0 0 0 0 0 0 0   
## multj28 alhrp28 mtlswk28 yrxfte11 clwkr1 prmjb1 fllprt1 nocg2e6   
## 0 0 0 0 0 0 0 0   
## imphwe1 penpln1 uncoll1 n07c3g10 pubpv10 cqpc42 rppc42 udpd42   
## 0 0 0 0 0 0 0 0   
## wgsal42 hleveg18   
## 0 0

Since all results sum to 0, then there are no null entries in the dataset.

## Analysis of unanswered surveys

If an individual left an answer blank or refused to answer while filling out the survey, then the answers were “dummied” and were entered as “98”, “99”, etc. Using the dataset’s literature, each attribute was analyzed to determine if the dummied entries would impact the results. If there are too many dummied entries, then the attribute could be removed. If the dummied entries are low, then the dummied entries were replaced with the mode of that particular attribute if it is a qualitative attribute. If it was a quantitative attribute, the dummied entries were replaced with the mean.

1. ecage26 - Person’s age as of December 31  
   No dummied fields
2. ecsex99 - Sex of respondent  
   No dummied fields
3. marst26 - Marital status of person as of December 31  
   71 dummied fields with “97”  
   Mode = 1

data$marst26[data$marst26==97]=1  
summary(data$marst26)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 2.775 5.000 6.000

1. mjacg26 - Person’s major activity at the end of reference year (working, retired, etc.)  
   No dummied fields
2. immst15 - Flag-Person is an immigrant  
   34,905 dummied fields with “7”  
   Attribute to be removed
3. pvreg25 - Province of residence  
   387 dummied fields with “97”  
   Mode = 35

data$pvreg25[data$pvreg25==97]=35  
summary(data$pvreg25)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.00 24.00 35.00 34.05 47.00 59.00

1. dwtenr25 - Ownership of dwelling  
   7 dummied fields with “7”  
   Mode = 1

data$dwtenr25[data$dwtenr25==7]=1  
summary(data$dwtenr25)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 1.202 1.000 2.000

1. mortg25 - Flag-A mortgage on the dwelling  
   9,654 dummied fields with “9”  
   Mode = 1

data$mortg25[data$mortg25==9]=1  
summary(data$mortg25)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 1.349 2.000 2.000

1. multj28 - Flag-Multiple job holder, any mth refyr  
   6,861 dummied fields with “9”  
   Mode = 2

data$multj28[data$multj28==9]=2  
summary(data$multj28)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 2.000 1.918 2.000 2.000

1. alhrp28 - Total hours paid all jobs during refyr  
   290 dummied fields with “9997”  
   6,861 dummied fields with “9999”

data$alhrp28[data$alhrp28==9997]=9999  
mean.alhrp28 <- mean(data$alhrp28[!data$alhrp28==9999])  
data$alhrp28[data$alhrp28==9999]=mean.alhrp28  
summary(data$alhrp28)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 521 1344 1344 2086 5200

1. mtlswk28 - Number of months since person last worked  
   2,213 dummied fields with “997”  
   35,700 dummied fields wtih “999”  
   Attribute to be removed
2. yrxfte11 - Number of years of experience, full-yr full-time  
   25 dummied fields with “98”  
   4,685 dummied fields with “97”

data$yrxfte11[data$yrxfte11==98]=97  
mean.yrxfte11 <- mean(data$yrxfte11[!data$yrxfte11==97])  
data$yrxfte11[data$yrxfte11==97]=mean.yrxfte11  
summary(data$yrxfte11)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 6.00 19.33 19.33 30.00 50.00

1. clwkr1 - Class of worker for this job  
   15,719 dummied fields with “99”  
   Mode = 1

data$clwkr1[data$clwkr1==99]=1  
summary(data$clwkr1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 1.335 1.000 6.000

1. prmjb1 - Flag to indicate if job is permanent  
   4,579 dummied fields with “7”  
   Change all to “9” = Not Applicable

data$prmjb1[data$prmjb1==7]=9  
summary(data$prmjb1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 2.000 5.011 9.000 9.000

1. fllprt1 - Flag-If job was full-time  
   854 dummied fields with “7”  
   20 dummied fields with “8”  
   Change all to “9” = Not Applicable

data$fllprt1[data$fllprt1==7]=9  
data$fllprt1[data$fllprt1==8]=9  
summary(data$fllprt1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 3.908 9.000 9.000

1. nocg2e6 - Occupational code  
   1,465 dummied fields with “97”  
   Change all to “99” = Not Applicable

data$nocg2e6[data$nocg2e6==97]=99  
summary(data$nocg2e6)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 9.00 19.00 43.27 99.00 99.00

1. imphwe1 - Implicit hourly wage for this paid worker job  
   19,893 dummied fields with “999.99”  
   Kept unchanged
2. penpln1 - If person is covered by a pension plan connected with this job  
   688 dummied fields with “7”  
   11 dummied fields with “8”  
   19,855 dummied fields with “9”  
   Mode = 2

data$penpln1[data$penpln1==7]=2  
data$penpln1[data$penpln1==8]=2  
data$penpln1[data$penpln1==9]=2  
summary(data$penpln1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 2.000 1.732 2.000 2.000

1. uncoll1 - If person was a member of a union/collective agreement  
   434 dummied fields with “7”  
   5 dummied fields with “8”  
   19,855 dummied fields with “9”  
   Mode = 3

data$uncoll1[data$uncoll1==7]=3  
data$uncoll1[data$uncoll1==8]=3  
data$uncoll1[data$uncoll1==9]=3  
summary(data$uncoll1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 3.00 3.00 2.63 3.00 3.00

1. n07c3g10 - Industry code of employer  
   331 dummied fields with “97”  
   Change to “99” = Not Applicable

data$n07c3g10[data$n07c3g10==97]=99  
summary(data$n07c3g10)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 6.00 12.00 39.28 99.00 99.00

1. pubpv10 - If employer is in the public or private sector  
   No dummied fields
2. cqpc42 - CPP/QPP contributions  
   No dummied fields
3. rppc42 - Registered pension plan (RPP) contributions / employer-sponsored  
   No dummied fields
4. udpd42 - Union dues (oth professional premiums)  
   No dummied fields
5. wgsal42 - Wages and salaries, before deductions (employment income)  
   No dummied fields
6. hleveg18 - Highest level of education  
   106 dummied fields with “97” Change to “1” = Never attended school

data$hleveg18[data$hleveg18==97]=1  
summary(data$hleveg18)

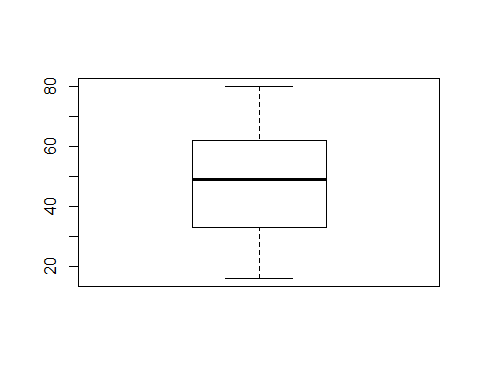
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 6.0 9.0 7.7 9.0 12.0

## Boxplot analysis

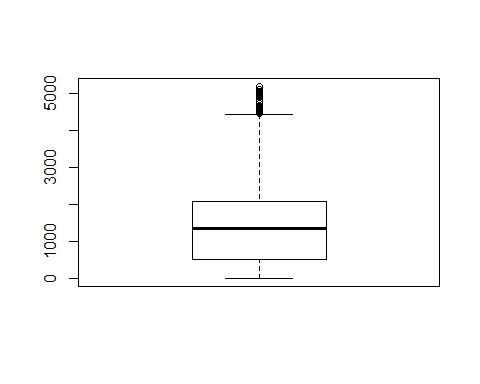
The quantitative attributes are in the follwowing positions: 1, 10, 11, 12, 17, 22, 23, 24, 25

Boxplots for each attribute:

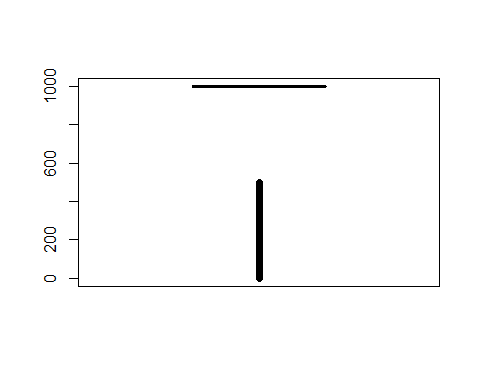
quant <- c(1,10,11,12,17,22,23,24,25)  
boxplot(data[1])



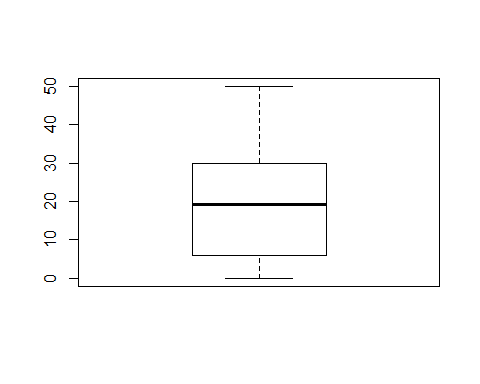
boxplot(data[10])



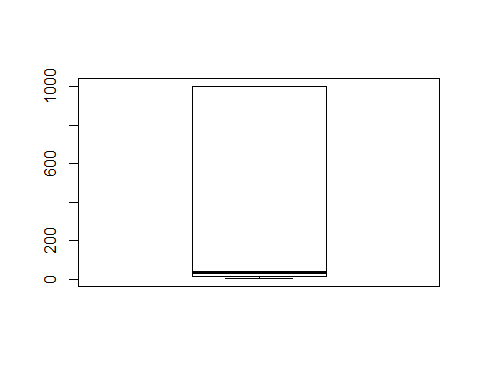
boxplot(data[11])



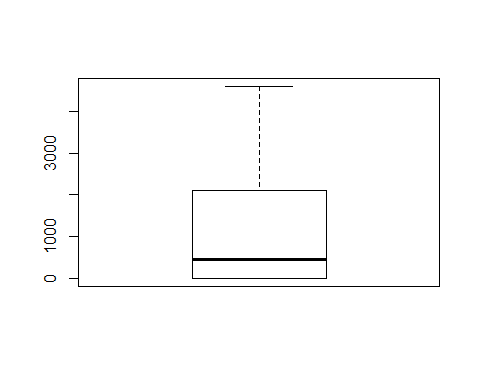
boxplot(data[12])



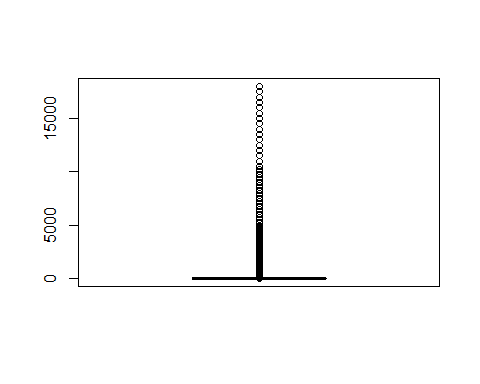
boxplot(data[17])



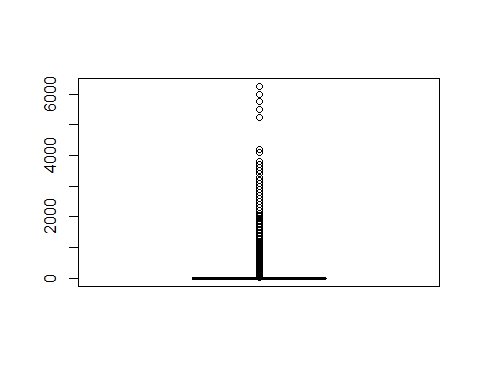
boxplot(data[22])



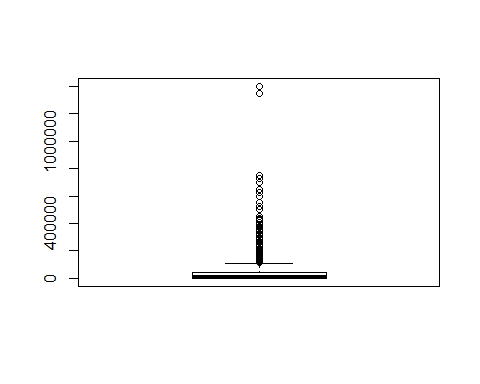
boxplot(data[23])



boxplot(data[24])

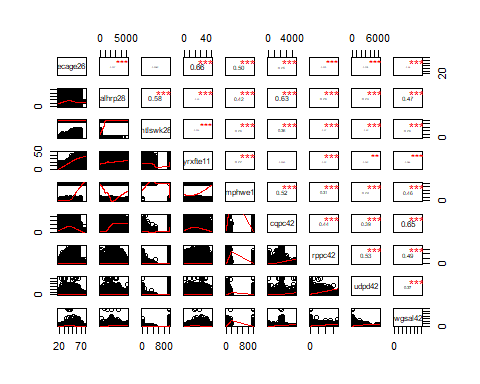


boxplot(data[25])



## Correlation analysis - quantitative

dataquant <- data[quant]  
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)   
{  
 usr <- par("usr"); on.exit(par(usr))   
 par(usr = c(0, 1, 0, 1))   
 r <- abs(cor(x, y))   
 txt <- format(c(r, 0.123456789), digits=digits)[1]   
 txt <- paste(prefix, txt, sep="")   
 if(missing(cex.cor)) cex <- 0.8/strwidth(txt)   
 test <- cor.test(x,y)   
 Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,   
 cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),  
 symbols = c("\*\*\*", "\*\*", "\*", ".", " "))   
 text(0.5, 0.5, txt, cex = cex \* r)   
 text(.8, .8, Signif, cex=cex, col=2)   
}  
pairs(dataquant, lower.panel=panel.smooth, upper.panel=panel.cor)



## Removal of attributes

Results from the above “dummy” analysis conclude that the following attributes should be removed from the dataset:  
immst15  
mtlswk28

From the correlation analysis for the quantitative attributes, the correlation between any two attributes is not strong. Therefore, no further removal of attributes is needed.

The elimination of these attributes will be performed using WEKA.

## Export dataset

write.csv(data, file="C:/Users/cortinm/Documents/selecteddata.csv")