Untitled

##Load file:

adult<-read.csv(file= "https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data", sep = ",", header=FALSE)  
  
colnames(adult) <- c("Age","workclass","fnlwgt","education","education-num","marital-status","occupation","relationship","race","sex","capital-gain","capital-loss", "hours-per-week", "native-country", "class")  
  
str(adult)

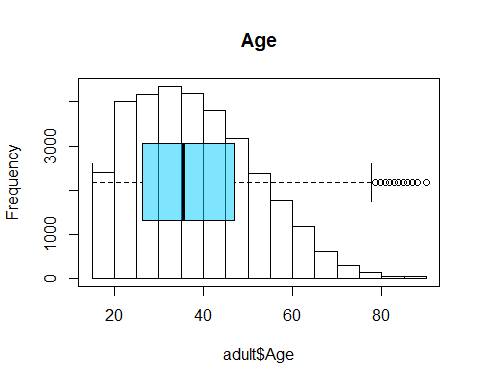
## 'data.frame': 32561 obs. of 15 variables:  
## $ Age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ workclass : chr " State-gov" " Self-emp-not-inc" " Private" " Private" ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : chr " Bachelors" " Bachelors" " HS-grad" " 11th" ...  
## $ education-num : int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital-status: chr " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...  
## $ occupation : chr " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...  
## $ relationship : chr " Not-in-family" " Husband" " Not-in-family" " Husband" ...  
## $ race : chr " White" " White" " White" " Black" ...  
## $ sex : chr " Male" " Male" " Male" " Male" ...  
## $ capital-gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital-loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours-per-week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native-country: chr " United-States" " United-States" " United-States" " United-States" ...  
## $ class : chr " <=50K" " <=50K" " <=50K" " <=50K" ...

summary(adult)

## Age workclass fnlwgt education   
## Min. :17.00 Length:32561 Min. : 12285 Length:32561   
## 1st Qu.:28.00 Class :character 1st Qu.: 117827 Class :character   
## Median :37.00 Mode :character Median : 178356 Mode :character   
## Mean :38.58 Mean : 189778   
## 3rd Qu.:48.00 3rd Qu.: 237051   
## Max. :90.00 Max. :1484705   
## education-num marital-status occupation relationship   
## Min. : 1.00 Length:32561 Length:32561 Length:32561   
## 1st Qu.: 9.00 Class :character Class :character Class :character   
## Median :10.00 Mode :character Mode :character Mode :character   
## Mean :10.08   
## 3rd Qu.:12.00   
## Max. :16.00   
## race sex capital-gain capital-loss   
## Length:32561 Length:32561 Min. : 0 Min. : 0.0   
## Class :character Class :character 1st Qu.: 0 1st Qu.: 0.0   
## Mode :character Mode :character Median : 0 Median : 0.0   
## Mean : 1078 Mean : 87.3   
## 3rd Qu.: 0 3rd Qu.: 0.0   
## Max. :99999 Max. :4356.0   
## hours-per-week native-country class   
## Min. : 1.00 Length:32561 Length:32561   
## 1st Qu.:40.00 Class :character Class :character   
## Median :40.00 Mode :character Mode :character   
## Mean :40.44   
## 3rd Qu.:45.00   
## Max. :99.00

##Age Variable

# Histogram  
hist(adult$Age, freq = TRUE, col = "white", main = "Age")  
# Add new plot  
par(new = TRUE)  
# Box plot  
boxplot(adult$Age, horizontal = TRUE, axes = FALSE, col = rgb(0, 0.8, 1, alpha = 0.5))  
# Box around the plots  
box()



sd(adult$Age)

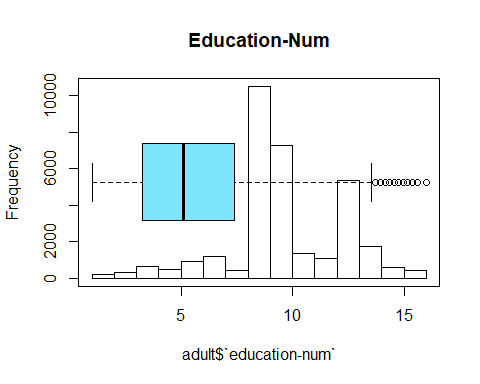
## [1] 13.64043

var(adult$Age)

## [1] 186.0614

#Education Num Variable

# Histogram  
hist(adult$`education-num`, freq = TRUE, col = "white", main = "Education-Num")  
# Add new plot  
par(new = TRUE)  
# Box plot  
boxplot(adult$Age, horizontal = TRUE, axes = FALSE, col = rgb(0, 0.8, 1, alpha = 0.5))  
# Box around the plots  
box()



sd(adult$`education-num`)

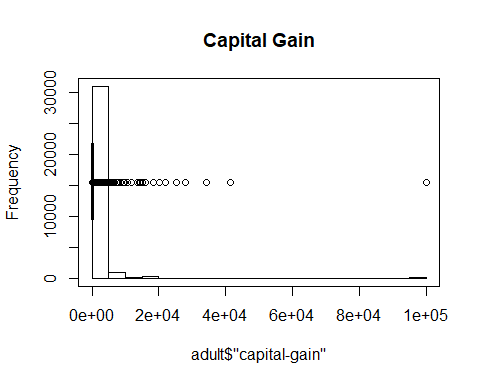
## [1] 2.57272

var(adult$`education-num`)

## [1] 6.61889

##Capital Gain Variable

# Histogram  
hist(adult$'capital-gain', freq = TRUE, col = "white", main = "Capital Gain")  
# Add new plot  
par(new = TRUE)  
# Box plot  
boxplot(adult$'capital-gain', horizontal = TRUE, axes = FALSE, col = rgb(0, 0.8, 1, alpha = 0.5))  
# Box around the plots  
box()



sd(adult$`capital-gain`)

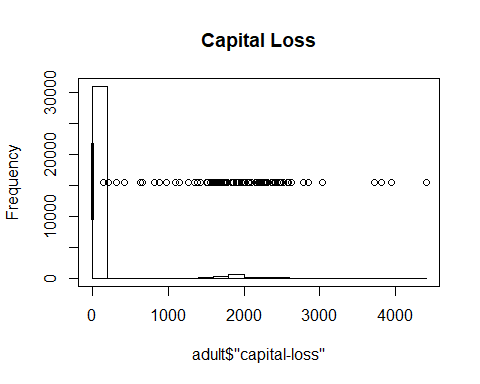
## [1] 7385.292

var(adult$`capital-gain`)

## [1] 54542539

#Capital Loss Variable

# Histogram  
hist(adult$'capital-loss', freq = TRUE, col = "white", main = "Capital Loss")  
# Add new plot  
par(new = TRUE)  
# Box plot  
boxplot(adult$'capital-loss', horizontal = TRUE, axes = FALSE, col = rgb(0, 0.8, 1, alpha = 0.5))  
# Box around the plots  
box()



sd(adult$`capital-loss`)

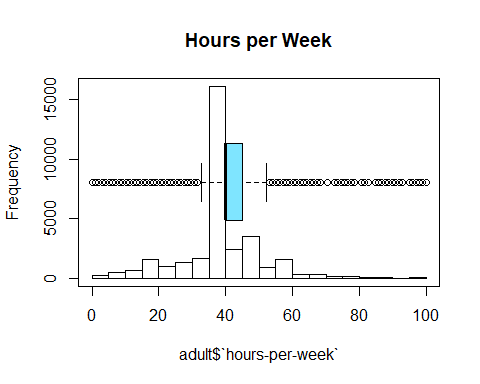
## [1] 402.9602

var(adult$`capital-loss`)

## [1] 162376.9

#Hours per week Variable

# Histogram  
hist(adult$`hours-per-week`, freq = TRUE, col = "white", main = "Hours per Week")  
# Add new plot  
par(new = TRUE)  
# Box plot  
boxplot(adult$`hours-per-week`, horizontal = TRUE, axes = FALSE, col = rgb(0, 0.8, 1, alpha = 0.5))  
# Box around the plots  
box()



sd(adult$`hours-per-week`)

## [1] 12.34743

var(adult$`hours-per-week`)

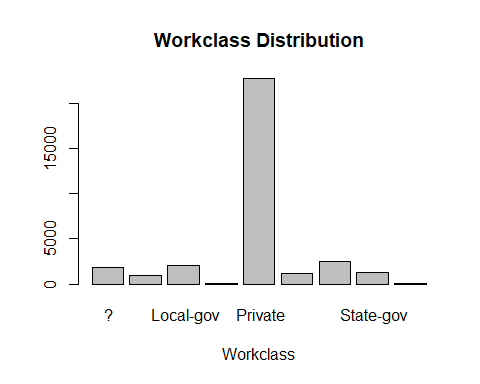
## [1] 152.459

#Workclass Variable

table(adult$workclass)

##   
## ? Federal-gov Local-gov Never-worked   
## 1836 960 2093 7   
## Private Self-emp-inc Self-emp-not-inc State-gov   
## 22696 1116 2541 1298   
## Without-pay   
## 14

countsworkclass <- table(adult$workclass)  
barplot(countsworkclass, main="Workclass Distribution",  
 xlab="Workclass")

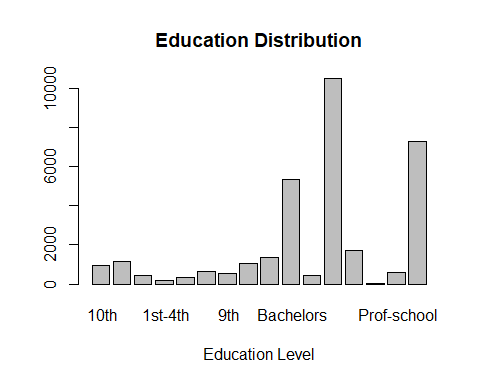


#Education Variable

table(adult$education)

##   
## 10th 11th 12th 1st-4th 5th-6th   
## 933 1175 433 168 333   
## 7th-8th 9th Assoc-acdm Assoc-voc Bachelors   
## 646 514 1067 1382 5355   
## Doctorate HS-grad Masters Preschool Prof-school   
## 413 10501 1723 51 576   
## Some-college   
## 7291

countsedu <- table(adult$education)  
barplot(countsedu, main="Education Distribution",  
 xlab="Education Level")

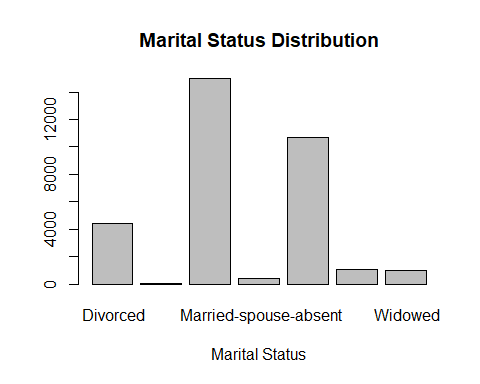


#Marital Status Variable

table(adult$`marital-status`)

##   
## Divorced Married-AF-spouse Married-civ-spouse   
## 4443 23 14976   
## Married-spouse-absent Never-married Separated   
## 418 10683 1025   
## Widowed   
## 993

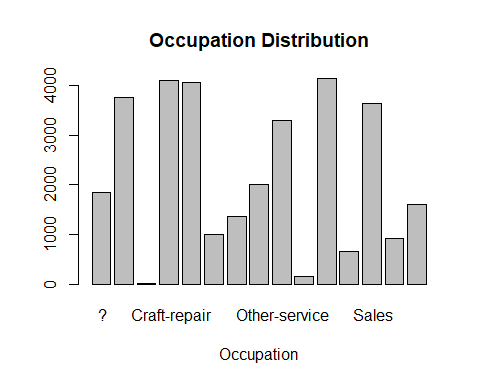
countsmarital <- table(adult$`marital-status`)  
barplot(countsmarital, main="Marital Status Distribution",  
 xlab="Marital Status")

 #Occupation Variable

table(adult$occupation)

##   
## ? Adm-clerical Armed-Forces Craft-repair   
## 1843 3770 9 4099   
## Exec-managerial Farming-fishing Handlers-cleaners Machine-op-inspct   
## 4066 994 1370 2002   
## Other-service Priv-house-serv Prof-specialty Protective-serv   
## 3295 149 4140 649   
## Sales Tech-support Transport-moving   
## 3650 928 1597

countsoccu <- table(adult$occupation)  
barplot(countsoccu, main="Occupation Distribution",  
 xlab="Occupation")

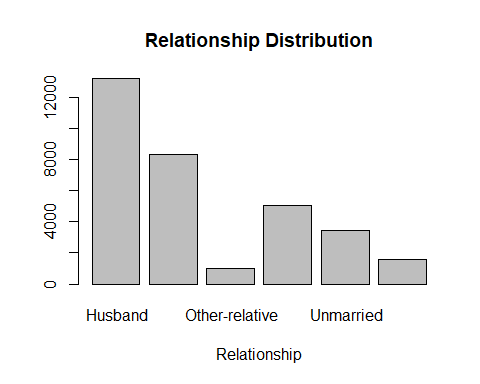


#Relationship Variable

table(adult$relationship)

##   
## Husband Not-in-family Other-relative Own-child Unmarried   
## 13193 8305 981 5068 3446   
## Wife   
## 1568

countsrel <- table(adult$relationship)  
barplot(countsrel, main="Relationship Distribution",  
 xlab="Relationship")

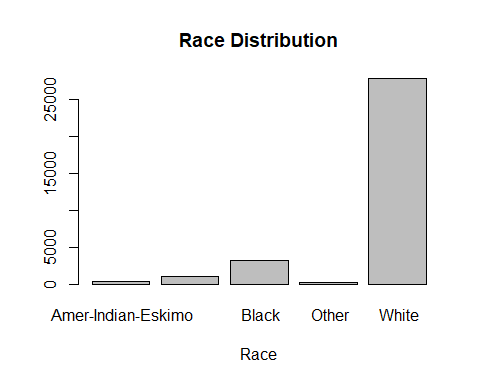


#Race Variable

table(adult$race)

##   
## Amer-Indian-Eskimo Asian-Pac-Islander Black Other   
## 311 1039 3124 271   
## White   
## 27816

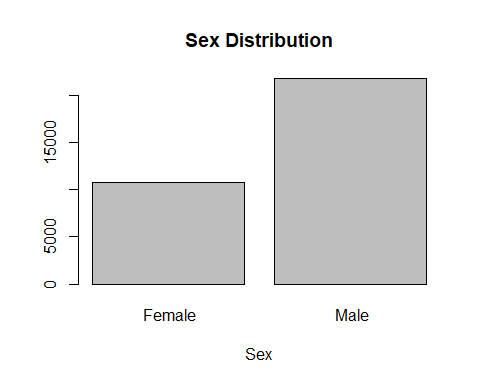
countsrace <- table(adult$race)  
barplot(countsrace, main="Race Distribution",  
 xlab="Race")

 #Sex Variable

table(adult$sex)

##   
## Female Male   
## 10771 21790

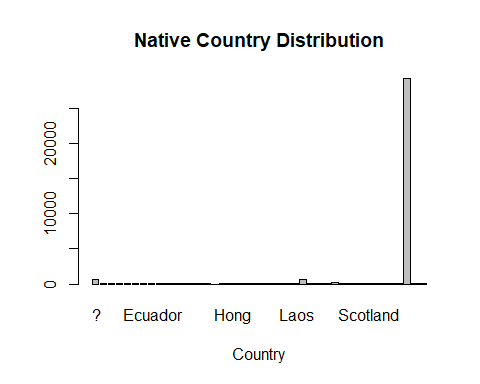
countssex <- table(adult$sex)  
barplot(countssex, main="Sex Distribution",  
 xlab="Sex")

 #Native Country Variable

table(adult$`native-country`)

##   
## ? Cambodia   
## 583 19   
## Canada China   
## 121 75   
## Columbia Cuba   
## 59 95   
## Dominican-Republic Ecuador   
## 70 28   
## El-Salvador England   
## 106 90   
## France Germany   
## 29 137   
## Greece Guatemala   
## 29 64   
## Haiti Holand-Netherlands   
## 44 1   
## Honduras Hong   
## 13 20   
## Hungary India   
## 13 100   
## Iran Ireland   
## 43 24   
## Italy Jamaica   
## 73 81   
## Japan Laos   
## 62 18   
## Mexico Nicaragua   
## 643 34   
## Outlying-US(Guam-USVI-etc) Peru   
## 14 31   
## Philippines Poland   
## 198 60   
## Portugal Puerto-Rico   
## 37 114   
## Scotland South   
## 12 80   
## Taiwan Thailand   
## 51 18   
## Trinadad&Tobago United-States   
## 19 29170   
## Vietnam Yugoslavia   
## 67 16

countsnative <- table(adult$`native-country`)  
barplot(countsnative, main="Native Country Distribution",  
 xlab="Country")

 #Create duplicate file for cleaning

adultclean<-adult  
  
str(adultclean)

## 'data.frame': 32561 obs. of 15 variables:  
## $ Age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ workclass : chr " State-gov" " Self-emp-not-inc" " Private" " Private" ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : chr " Bachelors" " Bachelors" " HS-grad" " 11th" ...  
## $ education-num : int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital-status: chr " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...  
## $ occupation : chr " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...  
## $ relationship : chr " Not-in-family" " Husband" " Not-in-family" " Husband" ...  
## $ race : chr " White" " White" " White" " Black" ...  
## $ sex : chr " Male" " Male" " Male" " Male" ...  
## $ capital-gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital-loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours-per-week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native-country: chr " United-States" " United-States" " United-States" " United-States" ...  
## $ class : chr " <=50K" " <=50K" " <=50K" " <=50K" ...

##FIND MISSING VALUES

# find elements  
replace <- adultclean == " ?"  
# replace elements with NA  
is.na(adultclean) <- replace

#Missing values per attribute  
sapply(adultclean, function(x) sum(is.na(x)))

## Age workclass fnlwgt education education-num   
## 0 1836 0 0 0   
## marital-status occupation relationship race sex   
## 0 1843 0 0 0   
## capital-gain capital-loss hours-per-week native-country class   
## 0 0 0 583 0

#subset of missing values  
new\_DF <- adultclean[rowSums(is.na(adultclean)) > 0,]

##Clean white space

adultclean <- as.data.frame(apply(adultclean,2, function(x) gsub("\\s+", "", x)))

#Fill Missing Values with Mode

#Replace Workclass missing values with mode - Private  
A<-is.na(adultclean$workclass)  
adultclean$workclass[A]<-"Private"  
  
#Replace occupation missing values with mode - Prof Specialty  
B<-is.na(adultclean$occupation)  
adultclean$occupation[B]<-"Prof-specialty"  
  
  
#Replace Native Country missing values with mode - United-States  
C<-is.na(adultclean$`native-country`)  
adultclean$`native-country`[C]<-"United-States"

# Redo counts after missing values

#Workclass Variable  
table(adultclean$workclass)

##   
## Federal-gov Local-gov Never-worked Private   
## 960 2093 7 24532   
## Self-emp-inc Self-emp-not-inc State-gov Without-pay   
## 1116 2541 1298 14

#Occupation Variable  
table(adultclean$occupation)

##   
## Adm-clerical Armed-Forces Craft-repair Exec-managerial   
## 3770 9 4099 4066   
## Farming-fishing Handlers-cleaners Machine-op-inspct Other-service   
## 994 1370 2002 3295   
## Priv-house-serv Prof-specialty Protective-serv Sales   
## 149 5983 649 3650   
## Tech-support Transport-moving   
## 928 1597

#Native Country Variable  
table(adultclean$`native-country`)

##   
## Cambodia Canada   
## 19 121   
## China Columbia   
## 75 59   
## Cuba Dominican-Republic   
## 95 70   
## Ecuador El-Salvador   
## 28 106   
## England France   
## 90 29   
## Germany Greece   
## 137 29   
## Guatemala Haiti   
## 64 44   
## Holand-Netherlands Honduras   
## 1 13   
## Hong Hungary   
## 20 13   
## India Iran   
## 100 43   
## Ireland Italy   
## 24 73   
## Jamaica Japan   
## 81 62   
## Laos Mexico   
## 18 643   
## Nicaragua Outlying-US(Guam-USVI-etc)   
## 34 14   
## Peru Philippines   
## 31 198   
## Poland Portugal   
## 60 37   
## Puerto-Rico Scotland   
## 114 12   
## South Taiwan   
## 80 51   
## Thailand Trinadad&Tobago   
## 18 19   
## United-States Vietnam   
## 29753 67   
## Yugoslavia   
## 16

#Look at Class Variable

table(adult$class)

##   
## <=50K >50K   
## 24720 7841

###Compare Education and Education-num

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

distinct(adult,education,`education-num`)

## education education-num  
## 1 Bachelors 13  
## 2 HS-grad 9  
## 3 11th 7  
## 4 Masters 14  
## 5 9th 5  
## 6 Some-college 10  
## 7 Assoc-acdm 12  
## 8 Assoc-voc 11  
## 9 7th-8th 4  
## 10 Doctorate 16  
## 11 Prof-school 15  
## 12 5th-6th 3  
## 13 10th 6  
## 14 1st-4th 2  
## 15 Preschool 1  
## 16 12th 8

#Drop Education-num

adultclean$`education-num`<-NULL

#Drop Fnlwgt

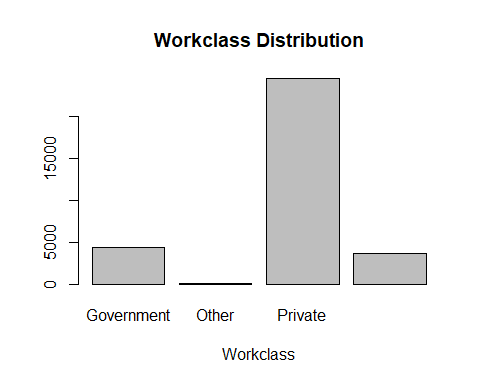
adultclean$fnlwgt<-NULL

#Update Workclass Categories

adultclean["workclass"][adultclean["workclass"]=='Federal-gov' |adultclean["workclass"]=='Local-gov' |adultclean["workclass"]=='State-gov']<-"Government"  
  
adultclean["workclass"][adultclean["workclass"]=='Never-worked' |adultclean["workclass"]=='Without-pay']<-"Other"  
  
adultclean["workclass"][adultclean["workclass"]=='Self-emp-inc' |adultclean["workclass"]=='Self-emp-not-inc']<-"Self Employed"  
  
table(adultclean$workclass)

##   
## Government Other Private Self Employed   
## 4351 21 24532 3657

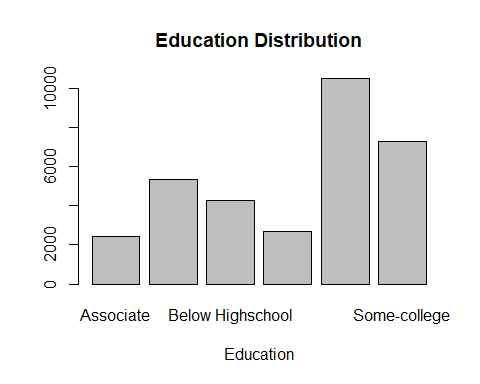
countswc <- table(adultclean$workclas)  
barplot(countswc, main="Workclass Distribution",  
 xlab="Workclass")

 #Update Education Categories

adultclean["education"][adultclean["education"]=='Preschool' |adultclean["education"]=='1st-4th' |adultclean["education"]=='5th-6th' |adultclean["education"]=='7th-8th' |adultclean["education"]=='9th' |adultclean["education"]=='10th' |adultclean["education"]=='11th' |adultclean["education"]=='12th']<-"Below Highschool"  
  
adultclean["education"][adultclean["education"]=='HS-grad']<-"Highschool Grad"  
  
adultclean["education"][adultclean["education"]=='Assoc-acdm' |adultclean["education"]=='Assoc-voc']<-"Associate"  
  
adultclean["education"][adultclean["education"]=='Prof-school' |adultclean["education"]=='Doctorate' |adultclean["education"]=='Masters']<-"Grad School"  
  
table(adultclean$education)

##   
## Associate Bachelors Below Highschool Grad School   
## 2449 5355 4253 2712   
## Highschool Grad Some-college   
## 10501 7291

countsed <- table(adultclean$education)  
barplot(countsed, main="Education Distribution",  
 xlab="Education")

 #Clean Marital Status

adultclean["marital-status"][adultclean["marital-status"]=='Married-civ-spouse' |adultclean["marital-status"]=='Married-spouse-absent' |adultclean["marital-status"]=='Married-AF-spouse']<-"Married"  
  
table(adultclean$`marital-status`)

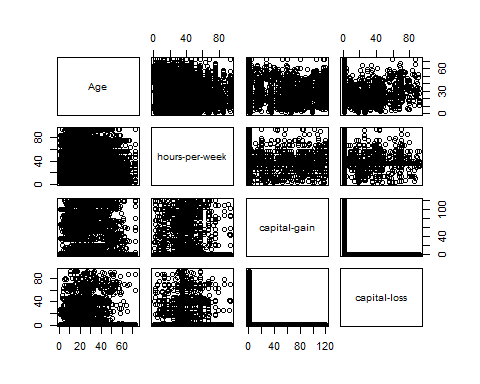
##   
## Divorced Married Never-married Separated Widowed   
## 4443 15417 10683 1025 993

#Clean Native Country

adultclean["native-country"][adultclean["native-country"]=='Cuba' |adultclean["native-country"]=='Jamaica' |adultclean["native-country"]=='India' |adultclean["native-country"]=='Mexico' |adultclean["native-country"]=='South' |adultclean["native-country"]=='Puerto-Rico' |adultclean["native-country"]=='Honduras' |adultclean["native-country"]=='England'|adultclean["native-country"]=='Canada'|adultclean["native-country"]=='Germany'|adultclean["native-country"]=='Iran'|adultclean["native-country"]=='Philippines'|adultclean["native-country"]=='Italy'|adultclean["native-country"]=='Poland'|adultclean["native-country"]=='Columbia'|adultclean["native-country"]=='Cambodia'|adultclean["native-country"]=='Thailand'|adultclean["native-country"]=='Ecuador'|adultclean["native-country"]=='Laos'|adultclean["native-country"]=='Taiwan'|adultclean["native-country"]=='Haiti'|adultclean["native-country"]=='Portugal'|adultclean["native-country"]=='Dominican-Republic'|adultclean["native-country"]=='El-Salvador'|adultclean["native-country"]=='France'|adultclean["native-country"]=='Guatemala'|adultclean["native-country"]=='China'|adultclean["native-country"]=='Japan'|adultclean["native-country"]=='Yugoslavia'|adultclean["native-country"]=='Peru'|adultclean["native-country"]=='Outlying-US(Guam-USVI-etc)'|adultclean["native-country"]=='Scotland'|adultclean["native-country"]=='Trinadad&Tobago'|adultclean["native-country"]=='Greece'|adultclean["native-country"]=='Nicaragua'|adultclean["native-country"]=='Vietnam'|adultclean["native-country"]=='Hong'|adultclean["native-country"]=='Ireland'|adultclean["native-country"]=='Hungary'|adultclean["native-country"]=='Holand-Netherlands']<-"Outside US"

#Correlation

subset\_adult<-adultclean[,c("Age","hours-per-week","capital-gain","capital-loss")]  
plot(subset\_adult)



subset\_adult$Age<-as.integer(subset\_adult$Age)  
subset\_adult$`hours-per-week`<-as.integer(subset\_adult$`hours-per-week`)  
subset\_adult$`capital-gain`<-as.integer(subset\_adult$`capital-gain`)  
subset\_adult$`capital-loss`<-as.integer(subset\_adult$`capital-loss`)  
str(subset\_adult)

## 'data.frame': 32561 obs. of 4 variables:  
## $ Age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ hours-per-week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ capital-gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital-loss : int 0 0 0 0 0 0 0 0 0 0 ...

cor(subset\_adult,method="pearson")

## Age hours-per-week capital-gain capital-loss  
## Age 1.00000000 0.06875571 0.07767450 0.05777454  
## hours-per-week 0.06875571 1.00000000 0.07840862 0.05425636  
## capital-gain 0.07767450 0.07840862 1.00000000 -0.03161506  
## capital-loss 0.05777454 0.05425636 -0.03161506 1.00000000

cor(subset\_adult,method="spearman")

## Age hours-per-week capital-gain capital-loss  
## Age 1.00000000 0.14290681 0.12494799 0.05848388  
## hours-per-week 0.14290681 1.00000000 0.09332205 0.05985243  
## capital-gain 0.12494799 0.09332205 1.00000000 -0.06656945  
## capital-loss 0.05848388 0.05985243 -0.06656945 1.00000000

subset\_adult$Age<-as.integer(subset\_adult$Age)  
subset\_adult$`hours-per-week`<-as.integer(subset\_adult$`hours-per-week`)  
subset\_adult$`capital-gain`<-as.integer(subset\_adult$`capital-gain`)  
subset\_adult$`capital-loss`<-as.integer(subset\_adult$`capital-loss`)  
str(subset\_adult)

## 'data.frame': 32561 obs. of 4 variables:  
## $ Age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ hours-per-week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ capital-gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital-loss : int 0 0 0 0 0 0 0 0 0 0 ...

adultclean$Age<-as.integer(adultclean$Age)  
adultclean$`hours-per-week`<-as.integer(adultclean$`hours-per-week`)  
adultclean$`capital-gain`<-as.integer(adultclean$`capital-gain`)  
adultclean$`capital-loss`<-as.integer(adultclean$`capital-loss`)  
str(adultclean)

## 'data.frame': 32561 obs. of 13 variables:  
## $ Age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ workclass : chr "Government" "Self Employed" "Private" "Private" ...  
## $ education : chr "Bachelors" "Bachelors" "Highschool Grad" "Below Highschool" ...  
## $ marital-status: chr "Never-married" "Married" "Divorced" "Married" ...  
## $ occupation : chr "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Handlers-cleaners" ...  
## $ relationship : chr "Not-in-family" "Husband" "Not-in-family" "Husband" ...  
## $ race : chr "White" "White" "White" "Black" ...  
## $ sex : chr "Male" "Male" "Male" "Male" ...  
## $ capital-gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital-loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours-per-week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native-country: chr "United-States" "United-States" "United-States" "United-States" ...  
## $ class : chr "<=50K" "<=50K" "<=50K" "<=50K" ...

#Update Class Variable #<=50K is 0 and >50K is 1

adultclean["class"][adultclean["class"]=='<=50K']<-0  
adultclean["class"][adultclean["class"]=='>50K']<-1

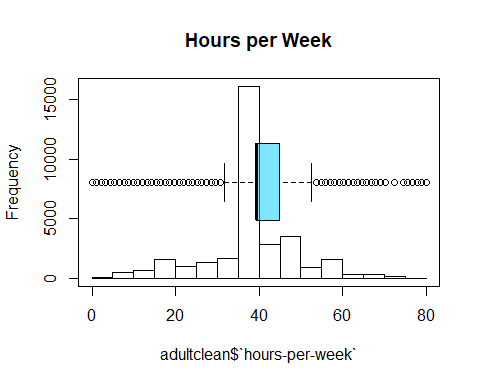
##Check Outliers for Hours per Week

adultclean$`hours-per-week`<-as.integer(adultclean$`hours-per-week`)  
adultclean$zscorehours<-scale(adultclean$`hours-per-week`)  
length(which(abs(adultclean$zscorehours)>3))

## [1] 440

#Replace outliers with mean value

adultclean$`hours-per-week`[which(abs(adultclean$zscorehours)>3)] <- mean(adultclean$`hours-per-week`, na.rm = TRUE)  
  
  
# Histogram  
hist(adultclean$`hours-per-week`, freq = TRUE, col = "white", main = "Hours per Week")  
# Add new plot  
par(new = TRUE)  
# Box plot  
boxplot(adultclean$`hours-per-week`, horizontal = TRUE, axes = FALSE, col = rgb(0, 0.8, 1, alpha = 0.5))  
# Box around the plots  
box()

 #Standardize Numeric Variables

normalize <- function(x, na.rm = TRUE) {  
 return((x- min(x)) /(max(x)-min(x)))  
}  
  
adultnorm<-adultclean  
  
adultnorm$`capital-gain`<-as.integer(adultnorm$`capital-gain`)  
adultnorm$`capital-loss`<-as.integer(adultnorm$`capital-loss`)  
adultnorm$Age<-as.integer(adultnorm$Age)  
adultnorm$class<-as.integer(adultnorm$class)  
#str(adultnorm)  
adultnorm$`capital-gain`<-normalize(adultnorm$`capital-gain`)  
adultnorm$`capital-loss`<-normalize(adultnorm$`capital-loss`)  
adultnorm$Age<-normalize(adultnorm$Age)  
adultnorm$`hours-per-week`<-normalize(adultnorm$`hours-per-week`)  
  
adultnorm$zscorehours<-NULL

#Split into Test and Train

set.seed(13)  
train\_index=sample(1:nrow(adultnorm),0.7\*nrow(adultnorm))  
train<-adultnorm[train\_index,]  
test<-adultnorm[-train\_index,]  
  
head(train)

## Age workclass education marital-status occupation  
## 13784 0.41095890 Private Some-college Divorced Adm-clerical  
## 16131 0.04109589 Private Highschool Grad Never-married Prof-specialty  
## 25536 0.10958904 Private Some-college Never-married Sales  
## 5989 0.35616438 Private Some-college Divorced Machine-op-inspct  
## 30282 0.19178082 Private Grad School Never-married Tech-support  
## 717 0.21917808 Private Highschool Grad Married Machine-op-inspct  
## relationship race sex capital-gain capital-loss hours-per-week  
## 13784 Other-relative White Female 0 0 0.4931507  
## 16131 Not-in-family Other Male 0 0 0.5342466  
## 25536 Not-in-family White Male 0 0 0.3561644  
## 5989 Not-in-family White Male 0 0 0.4931507  
## 30282 Not-in-family White Male 0 0 0.6986301  
## 717 Husband White Male 0 0 0.4931507  
## native-country class  
## 13784 United-States 0  
## 16131 United-States 0  
## 25536 United-States 0  
## 5989 United-States 0  
## 30282 United-States 0  
## 717 United-States 0

head(test)

## Age workclass education marital-status occupation  
## 1 0.30136986 Government Bachelors Never-married Adm-clerical  
## 9 0.19178082 Private Grad School Never-married Prof-specialty  
## 11 0.27397260 Private Some-college Married Exec-managerial  
## 13 0.08219178 Private Bachelors Never-married Adm-clerical  
## 22 0.50684932 Private Highschool Grad Separated Other-service  
## 24 0.35616438 Private Below Highschool Married Transport-moving  
## relationship race sex capital-gain capital-loss hours-per-week  
## 1 Not-in-family White Male 0.02174022 0.0000000 0.4931507  
## 9 Not-in-family White Female 0.14084141 0.0000000 0.6301370  
## 11 Husband Black Male 0.00000000 0.0000000 0.4991432  
## 13 Own-child White Female 0.00000000 0.0000000 0.3561644  
## 22 Unmarried Black Female 0.00000000 0.0000000 0.2191781  
## 24 Husband White Male 0.00000000 0.4687787 0.4931507  
## native-country class  
## 1 United-States 0  
## 9 United-States 1  
## 11 United-States 1  
## 13 United-States 0  
## 22 United-States 0  
## 24 United-States 0

#Logistic Regression

glm\_model<-glm(class~.,train,family="binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(glm\_model)

##   
## Call:  
## glm(formula = class ~ ., family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.0848 -0.5266 -0.1967 -0.0302 3.3753   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.757e+00 3.706e-01 -12.837 < 2e-16 \*\*\*  
## Age 1.553e+00 1.388e-01 11.189 < 2e-16 \*\*\*  
## workclassOther -1.144e+01 1.187e+02 -0.096 0.923247   
## workclassPrivate -8.492e-04 6.240e-02 -0.014 0.989141   
## workclassSelf Employed -1.963e-01 8.127e-02 -2.415 0.015728 \*   
## educationBachelors 5.960e-01 8.267e-02 7.209 5.65e-13 \*\*\*  
## educationBelow Highschool -1.477e+00 1.113e-01 -13.264 < 2e-16 \*\*\*  
## educationGrad School 1.146e+00 9.810e-02 11.685 < 2e-16 \*\*\*  
## educationHighschool Grad -6.269e-01 7.912e-02 -7.923 2.32e-15 \*\*\*  
## educationSome-college -2.441e-01 8.177e-02 -2.985 0.002839 \*\*   
## `marital-status`Married 7.537e-01 1.968e-01 3.830 0.000128 \*\*\*  
## `marital-status`Never-married -5.041e-01 1.024e-01 -4.921 8.63e-07 \*\*\*  
## `marital-status`Separated -8.466e-02 1.897e-01 -0.446 0.655378   
## `marital-status`Widowed -4.584e-02 1.888e-01 -0.243 0.808170   
## occupationArmed-Forces -6.090e-01 1.674e+00 -0.364 0.715960   
## occupationCraft-repair -4.448e-02 9.340e-02 -0.476 0.633933   
## occupationExec-managerial 6.424e-01 9.012e-02 7.129 1.01e-12 \*\*\*  
## occupationFarming-fishing -1.219e+00 1.662e-01 -7.332 2.26e-13 \*\*\*  
## occupationHandlers-cleaners -7.320e-01 1.669e-01 -4.385 1.16e-05 \*\*\*  
## occupationMachine-op-inspct -3.695e-01 1.191e-01 -3.102 0.001920 \*\*   
## occupationOther-service -9.737e-01 1.384e-01 -7.038 1.95e-12 \*\*\*  
## occupationPriv-house-serv -3.509e+00 1.892e+00 -1.855 0.063666 .   
## occupationProf-specialty 1.365e-01 8.903e-02 1.533 0.125221   
## occupationProtective-serv 3.300e-01 1.480e-01 2.230 0.025764 \*   
## occupationSales 1.344e-01 9.671e-02 1.390 0.164498   
## occupationTech-support 5.028e-01 1.305e-01 3.852 0.000117 \*\*\*  
## occupationTransport-moving -2.102e-01 1.162e-01 -1.808 0.070584 .   
## relationshipNot-in-family -8.343e-01 1.901e-01 -4.390 1.13e-05 \*\*\*  
## relationshipOther-relative -1.215e+00 2.629e-01 -4.623 3.78e-06 \*\*\*  
## relationshipOwn-child -2.035e+00 2.454e-01 -8.290 < 2e-16 \*\*\*  
## relationshipUnmarried -1.021e+00 2.145e-01 -4.760 1.94e-06 \*\*\*  
## relationshipWife 1.248e+00 1.208e-01 10.325 < 2e-16 \*\*\*  
## raceAsian-Pac-Islander 5.261e-01 2.796e-01 1.882 0.059853 .   
## raceBlack 4.210e-01 2.595e-01 1.622 0.104816   
## raceOther 2.232e-01 3.950e-01 0.565 0.571964   
## raceWhite 5.735e-01 2.470e-01 2.322 0.020255 \*   
## sexMale 8.704e-01 9.380e-02 9.280 < 2e-16 \*\*\*  
## `capital-gain` 3.110e+01 1.187e+00 26.194 < 2e-16 \*\*\*  
## `capital-loss` 2.763e+00 1.915e-01 14.427 < 2e-16 \*\*\*  
## `hours-per-week` 2.676e+00 1.604e-01 16.687 < 2e-16 \*\*\*  
## `native-country`United-States 2.448e-01 8.961e-02 2.732 0.006290 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 25201 on 22791 degrees of freedom  
## Residual deviance: 14828 on 22751 degrees of freedom  
## AIC: 14910  
##   
## Number of Fisher Scoring iterations: 12

#Logistic Regression Confusion Matrix

lrpredicted <- predict(glm\_model,test,type="response")  
predicted\_income<-ifelse(lrpredicted>=0.5,1,0)  
confusionmatrix<-table(actual=test$class,predicted=predicted\_income)  
print(confusionmatrix)

## predicted  
## actual 0 1  
## 0 6957 476  
## 1 957 1379

Accuracy<-sum(diag(confusionmatrix))/nrow(test)  
print(Accuracy)

## [1] 0.8533115

#####Dummy Variables

#install.packages('fastDummies')  
library('fastDummies')

## Warning: package 'fastDummies' was built under R version 4.0.5

adultdum<-adultnorm  
adultdum <- dummy\_cols(adultdum, select\_columns = 'workclass')  
adultdum <- dummy\_cols(adultdum, select\_columns = 'education')  
adultdum <- dummy\_cols(adultdum, select\_columns = 'marital-status')  
adultdum <- dummy\_cols(adultdum, select\_columns = 'occupation')  
adultdum <- dummy\_cols(adultdum, select\_columns = 'relationship')  
adultdum <- dummy\_cols(adultdum, select\_columns = 'race')  
adultdum <- dummy\_cols(adultdum, select\_columns = 'sex')  
  
adultdum <- dummy\_cols(adultdum, select\_columns = 'native-country')  
  
adultdum$workclass<-NULL  
adultdum$education<-NULL  
adultdum$`marital-status`<-NULL  
adultdum$occupation<-NULL  
adultdum$relationship<-NULL  
adultdum$race<-NULL  
adultdum$sex<-NULL  
adultdum$`native-country`<-NULL

#Convert all to numeric  
adultdum2 <- data.frame(lapply(adultdum, function(x) as.numeric(as.character(x))))

##KNN

#install.packages("class")  
#install.packages("gmodels")  
library(class)

## Warning: package 'class' was built under R version 4.0.5

library(gmodels)

## Warning: package 'gmodels' was built under R version 4.0.5

set.seed(13)  
train\_indexdum=sample(1:nrow(adultdum2),0.7\*nrow(adultdum2))  
traindum<-adultdum2[train\_indexdum,]  
testdum<-adultdum2[-train\_indexdum,]  
  
train.set.knn<-traindum[-5]  
test.set.knn<-testdum[-5]  
  
  
  
train.set\_labels<-traindum$class  
test.set\_labels<-testdum$class  
  
  
  
test\_pred <- knn(train = train.set.knn, test = test.set.knn, cl = train.set\_labels, k=10)  
test\_pred2<-knn(train = train.set.knn, test = test.set.knn, cl = train.set\_labels, k=180)

#k=10  
CrossTable(x=test.set\_labels, y=test\_pred, prop.chisq=FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 9769   
##   
##   
## | test\_pred   
## test.set\_labels | 0 | 1 | Row Total |   
## ----------------|-----------|-----------|-----------|  
## 0 | 6801 | 632 | 7433 |   
## | 0.915 | 0.085 | 0.761 |   
## | 0.869 | 0.326 | |   
## | 0.696 | 0.065 | |   
## ----------------|-----------|-----------|-----------|  
## 1 | 1028 | 1308 | 2336 |   
## | 0.440 | 0.560 | 0.239 |   
## | 0.131 | 0.674 | |   
## | 0.105 | 0.134 | |   
## ----------------|-----------|-----------|-----------|  
## Column Total | 7829 | 1940 | 9769 |   
## | 0.801 | 0.199 | |   
## ----------------|-----------|-----------|-----------|  
##   
##

table(Actual=test.set\_labels, Predicted=test\_pred)

## Predicted  
## Actual 0 1  
## 0 6801 632  
## 1 1028 1308

#k=180  
CrossTable(x=test.set\_labels, y=test\_pred2, prop.chisq=FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 9769   
##   
##   
## | test\_pred2   
## test.set\_labels | 0 | 1 | Row Total |   
## ----------------|-----------|-----------|-----------|  
## 0 | 6930 | 503 | 7433 |   
## | 0.932 | 0.068 | 0.761 |   
## | 0.862 | 0.290 | |   
## | 0.709 | 0.051 | |   
## ----------------|-----------|-----------|-----------|  
## 1 | 1106 | 1230 | 2336 |   
## | 0.473 | 0.527 | 0.239 |   
## | 0.138 | 0.710 | |   
## | 0.113 | 0.126 | |   
## ----------------|-----------|-----------|-----------|  
## Column Total | 8036 | 1733 | 9769 |   
## | 0.823 | 0.177 | |   
## ----------------|-----------|-----------|-----------|  
##   
##

table(Actual=test.set\_labels, Predicted=test\_pred2)

## Predicted  
## Actual 0 1  
## 0 6930 503  
## 1 1106 1230

##Decision Tree

library("party")

## Warning: package 'party' was built under R version 4.0.5

## Loading required package: grid

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.0.5

## Loading required package: modeltools

## Warning: package 'modeltools' was built under R version 4.0.3

## Loading required package: stats4

## Loading required package: strucchange

## Warning: package 'strucchange' was built under R version 4.0.5

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.5

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: sandwich

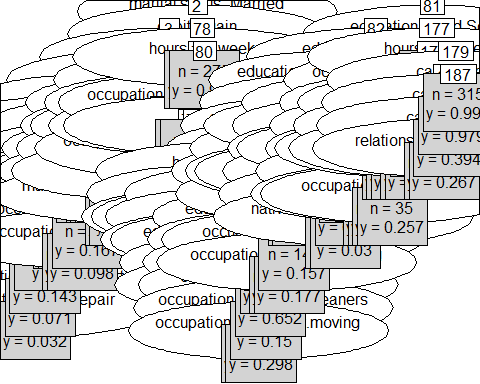
## Warning: package 'sandwich' was built under R version 4.0.5

adult\_ctree <- ctree(class ~ ., data=adultdum2)

print(adult\_ctree)

##   
## Conditional inference tree with 94 terminal nodes  
##   
## Response: class   
## Inputs: Age, capital.gain, capital.loss, hours.per.week, workclass\_Government, workclass\_Other, workclass\_Private, workclass\_Self.Employed, education\_Associate, education\_Bachelors, education\_Below.Highschool, education\_Grad.School, education\_Highschool.Grad, education\_Some.college, marital.status\_Divorced, marital.status\_Married, marital.status\_Never.married, marital.status\_Separated, marital.status\_Widowed, occupation\_Adm.clerical, occupation\_Armed.Forces, occupation\_Craft.repair, occupation\_Exec.managerial, occupation\_Farming.fishing, occupation\_Handlers.cleaners, occupation\_Machine.op.inspct, occupation\_Other.service, occupation\_Priv.house.serv, occupation\_Prof.specialty, occupation\_Protective.serv, occupation\_Sales, occupation\_Tech.support, occupation\_Transport.moving, relationship\_Husband, relationship\_Not.in.family, relationship\_Other.relative, relationship\_Own.child, relationship\_Unmarried, relationship\_Wife, race\_Amer.Indian.Eskimo, race\_Asian.Pac.Islander, race\_Black, race\_Other, race\_White, sex\_Female, sex\_Male, native.country\_Outside.US, native.country\_United.States   
## Number of observations: 32561   
##   
## 1) marital.status\_Married <= 0; criterion = 1, statistic = 6159.569  
## 2) capital.gain <= 0.06849068; criterion = 1, statistic = 1797.481  
## 3) education\_Grad.School <= 0; criterion = 1, statistic = 896.03  
## 4) education\_Bachelors <= 0; criterion = 1, statistic = 394.929  
## 5) hours.per.week <= 0.4931507; criterion = 1, statistic = 179.643  
## 6) Age <= 0.2191781; criterion = 1, statistic = 113.412  
## 7) occupation\_Protective.serv <= 0; criterion = 1, statistic = 35.14  
## 8) education\_Associate <= 0; criterion = 1, statistic = 29.338  
## 9) Age <= 0.2054795; criterion = 0.991, statistic = 14.035  
## 10) workclass\_Private <= 0; criterion = 0.968, statistic = 11.528  
## 11) race\_Amer.Indian.Eskimo <= 0; criterion = 0.985, statistic = 13.002  
## 12)\* weights = 518   
## 11) race\_Amer.Indian.Eskimo > 0  
## 13)\* weights = 9   
## 10) workclass\_Private > 0  
## 14) marital.status\_Never.married <= 0; criterion = 0.999, statistic = 19.044  
## 15) capital.loss <= 0.365932; criterion = 1, statistic = 24.206  
## 16) occupation\_Tech.support <= 0; criterion = 0.975, statistic = 11.98  
## 17) occupation\_Craft.repair <= 0; criterion = 0.997, statistic = 15.964  
## 18)\* weights = 494   
## 17) occupation\_Craft.repair > 0  
## 19)\* weights = 62   
## 16) occupation\_Tech.support > 0  
## 20)\* weights = 14   
## 15) capital.loss > 0.365932  
## 21)\* weights = 7   
## 14) marital.status\_Never.married > 0  
## 22)\* weights = 4676   
## 9) Age > 0.2054795  
## 23)\* weights = 223   
## 8) education\_Associate > 0  
## 24)\* weights = 475   
## 7) occupation\_Protective.serv > 0  
## 25)\* weights = 100   
## 6) Age > 0.2191781  
## 26) occupation\_Exec.managerial <= 0; criterion = 1, statistic = 40.768  
## 27) occupation\_Protective.serv <= 0; criterion = 1, statistic = 25.471  
## 28) sex\_Male <= 0; criterion = 1, statistic = 21.768  
## 29) occupation\_Prof.specialty <= 0; criterion = 0.98, statistic = 12.421  
## 30)\* weights = 2255   
## 29) occupation\_Prof.specialty > 0  
## 31)\* weights = 403   
## 28) sex\_Male > 0  
## 32) occupation\_Tech.support <= 0; criterion = 1, statistic = 19.522  
## 33) occupation\_Prof.specialty <= 0; criterion = 0.999, statistic = 18.032  
## 34)\* weights = 1243   
## 33) occupation\_Prof.specialty > 0  
## 35)\* weights = 133   
## 32) occupation\_Tech.support > 0  
## 36)\* weights = 42   
## 27) occupation\_Protective.serv > 0  
## 37)\* weights = 67   
## 26) occupation\_Exec.managerial > 0  
## 38)\* weights = 340   
## 5) hours.per.week > 0.4931507  
## 39) Age <= 0.2876712; criterion = 1, statistic = 55.391  
## 40) relationship\_Not.in.family <= 0; criterion = 1, statistic = 26.345  
## 41) race\_Asian.Pac.Islander <= 0; criterion = 1, statistic = 22.492  
## 42)\* weights = 658   
## 41) race\_Asian.Pac.Islander > 0  
## 43)\* weights = 15   
## 40) relationship\_Not.in.family > 0  
## 44) workclass\_Private <= 0; criterion = 0.999, statistic = 17.822  
## 45)\* weights = 144   
## 44) workclass\_Private > 0  
## 46)\* weights = 751   
## 39) Age > 0.2876712  
## 47) capital.loss <= 0.5064279; criterion = 1, statistic = 25.713  
## 48) sex\_Female <= 0; criterion = 1, statistic = 20.991  
## 49) marital.status\_Never.married <= 0; criterion = 0.988, statistic = 13.427  
## 50)\* weights = 318   
## 49) marital.status\_Never.married > 0  
## 51)\* weights = 133   
## 48) sex\_Female > 0  
## 52)\* weights = 429   
## 47) capital.loss > 0.5064279  
## 53)\* weights = 15   
## 4) education\_Bachelors > 0  
## 54) hours.per.week <= 0.5205479; criterion = 1, statistic = 99.634  
## 55) Age <= 0.2876712; criterion = 1, statistic = 60.115  
## 56) capital.loss <= 0.3232323; criterion = 1, statistic = 25.242  
## 57) Age <= 0.1643836; criterion = 1, statistic = 29.304  
## 58)\* weights = 811   
## 57) Age > 0.1643836  
## 59)\* weights = 362   
## 56) capital.loss > 0.3232323  
## 60)\* weights = 30   
## 55) Age > 0.2876712  
## 61)\* weights = 487   
## 54) hours.per.week > 0.5205479  
## 62) Age <= 0.1369863; criterion = 1, statistic = 30.187  
## 63)\* weights = 181   
## 62) Age > 0.1369863  
## 64) occupation\_Exec.managerial <= 0; criterion = 0.98, statistic = 12.421  
## 65)\* weights = 403   
## 64) occupation\_Exec.managerial > 0  
## 66)\* weights = 150   
## 3) education\_Grad.School > 0  
## 67) hours.per.week <= 0.5342466; criterion = 1, statistic = 47.559  
## 68) Age <= 0.2054795; criterion = 0.977, statistic = 12.136  
## 69)\* weights = 148   
## 68) Age > 0.2054795  
## 70) sex\_Female <= 0; criterion = 0.996, statistic = 15.581  
## 71)\* weights = 156   
## 70) sex\_Female > 0  
## 72)\* weights = 270   
## 67) hours.per.week > 0.5342466  
## 73) occupation\_Exec.managerial <= 0; criterion = 0.965, statistic = 11.365  
## 74)\* weights = 251   
## 73) occupation\_Exec.managerial > 0  
## 75) Age <= 0.3013699; criterion = 0.961, statistic = 11.189  
## 76)\* weights = 26   
## 75) Age > 0.3013699  
## 77)\* weights = 43   
## 2) capital.gain > 0.06849068  
## 78) hours.per.week <= 0.4246575; criterion = 0.986, statistic = 26.455  
## 79)\* weights = 27   
## 78) hours.per.week > 0.4246575  
## 80)\* weights = 275   
## 1) marital.status\_Married > 0  
## 81) education\_Grad.School <= 0; criterion = 1, statistic = 995.423  
## 82) education\_Bachelors <= 0; criterion = 1, statistic = 1071.757  
## 83) education\_Below.Highschool <= 0; criterion = 1, statistic = 426.724  
## 84) capital.gain <= 0.0501305; criterion = 1, statistic = 239.026  
## 85) occupation\_Exec.managerial <= 0; criterion = 1, statistic = 164.263  
## 86) capital.loss <= 0.399449; criterion = 1, statistic = 113.888  
## 87) hours.per.week <= 0.4246575; criterion = 1, statistic = 105.922  
## 88) relationship\_Wife <= 0; criterion = 1, statistic = 35.142  
## 89) education\_Highschool.Grad <= 0; criterion = 0.999, statistic = 17.525  
## 90)\* weights = 280   
## 89) education\_Highschool.Grad > 0  
## 91)\* weights = 403   
## 88) relationship\_Wife > 0  
## 92) workclass\_Self.Employed <= 0; criterion = 0.985, statistic = 12.988  
## 93)\* weights = 226   
## 92) workclass\_Self.Employed > 0  
## 94)\* weights = 34   
## 87) hours.per.week > 0.4246575  
## 95) Age <= 0.2465753; criterion = 1, statistic = 150.329  
## 96) Age <= 0.1506849; criterion = 1, statistic = 42.756  
## 97) education\_Highschool.Grad <= 0; criterion = 0.999, statistic = 17.597  
## 98)\* weights = 343   
## 97) education\_Highschool.Grad > 0  
## 99)\* weights = 481   
## 96) Age > 0.1506849  
## 100) education\_Highschool.Grad <= 0; criterion = 0.996, statistic = 15.333  
## 101)\* weights = 634   
## 100) education\_Highschool.Grad > 0  
## 102) hours.per.week <= 0.6575342; criterion = 0.955, statistic = 10.898  
## 103) relationship\_Wife <= 0; criterion = 0.971, statistic = 11.725  
## 104)\* weights = 662   
## 103) relationship\_Wife > 0  
## 105)\* weights = 59   
## 102) hours.per.week > 0.6575342  
## 106)\* weights = 105   
## 95) Age > 0.2465753  
## 107) education\_Highschool.Grad <= 0; criterion = 1, statistic = 51.317  
## 108) occupation\_Farming.fishing <= 0; criterion = 1, statistic = 26.657  
## 109) occupation\_Other.service <= 0; criterion = 1, statistic = 19.663  
## 110) occupation\_Tech.support <= 0; criterion = 0.998, statistic = 16.835  
## 111) occupation\_Handlers.cleaners <= 0; criterion = 0.995, statistic = 15.04  
## 112) occupation\_Transport.moving <= 0; criterion = 0.985, statistic = 12.996  
## 113)\* weights = 1283   
## 112) occupation\_Transport.moving > 0  
## 114)\* weights = 114   
## 111) occupation\_Handlers.cleaners > 0  
## 115)\* weights = 40   
## 110) occupation\_Tech.support > 0  
## 116)\* weights = 112   
## 109) occupation\_Other.service > 0  
## 117)\* weights = 74   
## 108) occupation\_Farming.fishing > 0  
## 118)\* weights = 84   
## 107) education\_Highschool.Grad > 0  
## 119) occupation\_Other.service <= 0; criterion = 1, statistic = 19.897  
## 120) occupation\_Farming.fishing <= 0; criterion = 0.999, statistic = 17.736  
## 121)\* weights = 2057   
## 120) occupation\_Farming.fishing > 0  
## 122)\* weights = 141   
## 119) occupation\_Other.service > 0  
## 123)\* weights = 140   
## 86) capital.loss > 0.399449  
## 124) capital.loss <= 0.4538567; criterion = 1, statistic = 74.373  
## 125)\* weights = 192   
## 124) capital.loss > 0.4538567  
## 126) capital.loss <= 0.4933425; criterion = 0.982, statistic = 12.596  
## 127)\* weights = 44   
## 126) capital.loss > 0.4933425  
## 128)\* weights = 39   
## 85) occupation\_Exec.managerial > 0  
## 129) capital.loss <= 0.399449; criterion = 1, statistic = 30.993  
## 130) education\_Associate <= 0; criterion = 0.994, statistic = 14.784  
## 131)\* weights = 815   
## 130) education\_Associate > 0  
## 132)\* weights = 140   
## 129) capital.loss > 0.399449  
## 133) capital.loss <= 0.4538567; criterion = 0.954, statistic = 10.86  
## 134)\* weights = 49   
## 133) capital.loss > 0.4538567  
## 135)\* weights = 11   
## 84) capital.gain > 0.0501305  
## 136) Age <= 0.5890411; criterion = 1, statistic = 37.342  
## 137)\* weights = 437   
## 136) Age > 0.5890411  
## 138)\* weights = 54   
## 83) education\_Below.Highschool > 0  
## 139) capital.gain <= 0.0501305; criterion = 1, statistic = 64.864  
## 140) capital.loss <= 0.4044995; criterion = 1, statistic = 24.148  
## 141) occupation\_Exec.managerial <= 0; criterion = 0.998, statistic = 16.796  
## 142) occupation\_Sales <= 0; criterion = 0.996, statistic = 15.537  
## 143) hours.per.week <= 0.6164384; criterion = 0.989, statistic = 13.583  
## 144) native.country\_Outside.US <= 0; criterion = 0.961, statistic = 11.186  
## 145)\* weights = 995   
## 144) native.country\_Outside.US > 0  
## 146)\* weights = 371   
## 143) hours.per.week > 0.6164384  
## 147)\* weights = 205   
## 142) occupation\_Sales > 0  
## 148)\* weights = 86   
## 141) occupation\_Exec.managerial > 0  
## 149)\* weights = 65   
## 140) capital.loss > 0.4044995  
## 150) capital.loss <= 0.4538567; criterion = 0.993, statistic = 14.352  
## 151)\* weights = 17   
## 150) capital.loss > 0.4538567  
## 152)\* weights = 18   
## 139) capital.gain > 0.0501305  
## 153)\* weights = 39   
## 82) education\_Bachelors > 0  
## 154) occupation\_Exec.managerial <= 0; criterion = 1, statistic = 75.315  
## 155) capital.gain <= 0.0501305; criterion = 1, statistic = 44.527  
## 156) capital.loss <= 0.399449; criterion = 1, statistic = 59.046  
## 157) relationship\_Not.in.family <= 0; criterion = 1, statistic = 26.238  
## 158) hours.per.week <= 0.4109589; criterion = 1, statistic = 24.388  
## 159) relationship\_Wife <= 0; criterion = 1, statistic = 22.863  
## 160)\* weights = 124   
## 159) relationship\_Wife > 0  
## 161)\* weights = 57   
## 158) hours.per.week > 0.4109589  
## 162) occupation\_Other.service <= 0; criterion = 0.992, statistic = 14.213  
## 163)\* weights = 1380   
## 162) occupation\_Other.service > 0  
## 164)\* weights = 35   
## 157) relationship\_Not.in.family > 0  
## 165)\* weights = 36   
## 156) capital.loss > 0.399449  
## 166) race\_Asian.Pac.Islander <= 0; criterion = 0.988, statistic = 14.375  
## 167)\* weights = 115   
## 166) race\_Asian.Pac.Islander > 0  
## 168)\* weights = 8   
## 155) capital.gain > 0.0501305  
## 169)\* weights = 223   
## 154) occupation\_Exec.managerial > 0  
## 170) capital.gain <= 0.0501305; criterion = 0.992, statistic = 14.098  
## 171) capital.loss <= 0.399449; criterion = 0.995, statistic = 14.98  
## 172) capital.loss <= 0; criterion = 0.986, statistic = 13.109  
## 173)\* weights = 628   
## 172) capital.loss > 0  
## 174)\* weights = 8   
## 171) capital.loss > 0.399449  
## 175)\* weights = 81   
## 170) capital.gain > 0.0501305  
## 176)\* weights = 145   
## 81) education\_Grad.School > 0  
## 177) hours.per.week <= 0.3287671; criterion = 1, statistic = 43.621  
## 178)\* weights = 93   
## 177) hours.per.week > 0.3287671  
## 179) capital.gain <= 0.0501305; criterion = 1, statistic = 31.136  
## 180) capital.loss <= 0.399449; criterion = 1, statistic = 45.63  
## 181) capital.gain <= 0; criterion = 1, statistic = 19.563  
## 182) relationship\_Not.in.family <= 0; criterion = 0.999, statistic = 17.372  
## 183)\* weights = 1085   
## 182) relationship\_Not.in.family > 0  
## 184)\* weights = 15   
## 181) capital.gain > 0  
## 185)\* weights = 33   
## 180) capital.loss > 0.399449  
## 186)\* weights = 187   
## 179) capital.gain > 0.0501305  
## 187)\* weights = 315

plot(adult\_ctree, type="simple")



traindec<-traindum  
testdec<-testdum  
  
traindec["class"][traindec["class"]==0]<-'<=50K'  
traindec["class"][traindec["class"]==1]<-'>50K'  
  
testdec["class"][testdec["class"]==0]<-'<=50K'  
testdec["class"][testdec["class"]==1]<-'>50K'  
  
testdec$class<-as.factor(testdec$class)  
traindec$class<-as.factor(traindec$class)

#Model Decision Tree

adult\_ctree\_model <- ctree(class ~ ., data=traindec)

adult\_ctree\_prediction <- predict(adult\_ctree\_model, newdata=testdec)   
# gives the probability for each class  
head(adult\_ctree\_prediction)

## [1] <=50K >50K <=50K <=50K <=50K <=50K  
## Levels: <=50K >50K

table(adult\_ctree\_prediction, testdec$class)

##   
## adult\_ctree\_prediction <=50K >50K  
## <=50K 7099 1097  
## >50K 334 1239

###Naive Bayes

library(e1071)

## Warning: package 'e1071' was built under R version 4.0.5

nb\_model<-naiveBayes(class~.,data=train)

nb\_predict<-predict(nb\_model,newdata=test)

table(nb\_predict,test$class)

##   
## nb\_predict 0 1  
## 0 7045 1373  
## 1 388 963