Rdata<-read.csv("C:\Users\sandh\Downloads\Rdata.csv")

view(Rdata)

str(Rdata)

# Data cleaning

# Work Class Combining

table(Rdata$workclass)

Rdata$workclass<-as.character(Rdata$workclass)

Rdata$workclass[Rdata$workclass == "without-pay" | Rdata$workclass == "Never-worked"] <-"unemployed"

Rdata$workclass[Rdata$workclass =="State-gov" | Rdata$workclass == "Local-gov"]<- "SL-gov"

Rdata$workclass[Rdata$workclass == "Self-emp-inc" | Rdata$workclass == "Self-emp-not-inc"] <- "Self-employed"

table(Rdata$workclass)

##Martial Status Combining

table(Rdata$martial.status)

Rdata$martial.status <- as.character(Rdata$martial.status)

Rdata$martial.status[Rdata$martial.status == "Married-AF-Spouse" | Rdata$martial.status == "Married-Civ-Spouse" | Rdata$martial.status == "Married- spouse-absent"] <- "Married"

Rdata$martial.status[Rdata$martial.status == "Divorced" | Rdata$martial.status == "Separated" | Rdata$martial.status == "Widowed"] <- "Not-Married"

table(Rdata$martial.status)

## Country Combining

Rdata$native.country <- as.character(Rdata$native.country)

north.america <- c("Canada","Cuba", "Dominican-Republic", "El-Salvador","Guatemala","Haiti","Honduras","Jamaica","Mexico","Nicaragua","Outlying-US(Guam-USVI-etc)","Puerto-Rico","Trinadada&Tobago","United-States")

asia <- c("Cambodia","China","Hong","India","Iran","Japan","Laos","Philippines","Taiwan","Thailand","Vietnam")

south.america <- c("Columbia","Ecuador","Peru")

europe <- c("England","France","Germany","Greece","Holand-Netherlands","Hungary","Ireland","Italy","Poland","Portugal","Scotland","Yugoslavia")

other<-c("South","?")

Rdata$native.country[Rdata$native.country %in% north.america] <- "North America"

Rdata$native.country[Rdata$native.country %in% asia] <- "Asia"

Rdata$native.country[Rdata$native.country %in% south.america] <- "South America"

Rdata$native.country[Rdata$native.country %in% europe] <- "Europe"

Rdata$native.country[Rdata$native.country %in% other] <- "Other"

table(Rdata$native.country)

Rdata$native.country <- as.factor(Rdata$native.country)

Rdata$martial.status <- as.factor(Rdata$martial.status)

Rdata$workclass <- as.factor(Rdata$workclass)

str(Rdata)

## Dealing with Missing Data

table(Rdata$workclass)

Rdata[Rdata =="?"] <-NA

table(Rdata$workclass)

#missmap

install.packages("Amelia")

library(Amelia)

missmap(Rdata, y.at = 1, y.labels = "", col = c("yellow","black"),legend = FALSE)

Rdata <-na.omit(Rdata)

missmap(Rdata, y.at=1,y.label= "", legend = FALSE, col=c("yellow","black"))

install.packages("ggplot2")

library(ggplot2)

ggplot(Rdata, aes(age)) + geom\_histogram(aes(fill=income),color ="black",birwidth=1)

ggplot(Rdata, aes(hours.per.week)) + geom\_histogram()

#by country

install.packages("data.table")

library(data.table)

setNames(Rdata,"native.country", "region")

#Reorder factor levels by country

region.ordered<- reorder(Rdata$region,Rdata$region,length)

region.ordered <-factor(region.ordered, levels = rev(levels(region.ordered)))

ggplot(Rdata,aes(region.ordered)) + geom\_bar(aes(fill = income), color = "black")

#Building the Model

#The purpose of this model is to classify people into two groups, below 50k or above

install.packages("caTools")

library(caTools)

split<-sample.split(Rdata$income, SplitRatio = 0.7)

Rdata <- subset(Rdata, split == TRUE)

Rdata<- subset(Rdata, split == FALSE)

#Training the model

log.model <- glm(income ~ ., family = binomial(), Rdata)

##prediction

prediction <- predict(log.model, Rdata, type = "response")

table(Rdata$income, prediction >= 0.5)

#Accuracy

(9639+2116)/(9639+744+2116+1311)

9649/(9639+1311)

#Precision

9639/(9639+744)