Codes used in data cleaning

1. To detect Percentage of missing values in the variables in a dataset:

mis\_detect <- function(x)

{

mat <- matrix(NA,ncol(x),1)

for(i in 1:ncol(x))

{

mat[i,1] <- sum(is.na(x[,i]))/nrow(x)\*100

}

rownames(mat) <- names(x)

colnames(mat) <- "%missing"

return(round(mat,3))

}

mis\_detect(co2)

1. **Replacing 0 with o:**

cln.mk <- read.csv("GoogleDrive-Sheet1.csv", header = T)

cln.mk <- edit(cln.mk)

cln.mk$Var1

## replacing 0 with o:

rep0 <- function(data)

{

for(i in 1:ncol(data))

{

a <- grep("^0",data)

for(i in a[i])

{

a <- gsub("^0","^o",a[i],ignore.case = F)

}

}

return(a)

}

rep0(cln.mk)

1. Handling space in front:

grep("^0",cln.mk$Var1)

1. Replacing numbers with NA:

co2$Mk[co2$Mk == 270] <- NA

co2$Mk[co2$Mk == 0] <- NA

co2$Mk[co2$Mk == "706/2007"] <- NA

1. Remove unnecessary (non-alpha numeric) symbols:

code <- read.csv("co2.cars.csv", header = FALSE)

head(code)

code <- code[,1]

code <- unique(code) #Remove repetition

length(code)

head(code)

class(code)

code <- code[!is.na(code)]

code <- read.csv()

code <- unique(code)

x <- "î-î¥undai"

str\_replace\_all(x, "[^[:alnum:]]", "")

trim <- function (x) gsub("^\\s+|\\s+$", "", x)

y <- " msjn h"

trim(y)

str\_trim(y)

library(stringr)

co2$Mk <- str\_replace\_all(co2$Mk, "[^[:alnum:]]", " ")

head(co2$Mk)

1. Create a vector of codes:

code <- read.csv("co2.cars.csv", header = FALSE)

head(code)

code <- code[,1]

code <- unique(code) #Remove repetition

length(code)

head(code)

class(code)

code <- code[!is.na(code)]

1. Calculate the distance matrix:

D <- adist(co2$Mk, code)

colnames(D) <- code

rownames(D) <- co2$Mk

D[1:5,1:10]

1. Code for each of the rows of the matrix return the minimum distance:

i <- apply(X=D, MARGIN = 1, FUN = which.min)

class(i)

i <- as.integer(i)

1. compare the actual and coded:

dat <- data.frame(rawtext = co2$Mk, coded = code[i])

dat[1:100,]

1. replace the original variable y the coded vector:

co2$Mk <- code[i]

co2$Mk <- as.factor(co2$Mk)

length(levels(co2$Mk))

1. Dealing the variable ‘m’:

work\_1 <- na.omit(work[-c(1,2,3,4,5,11,12,15,6)])

median(work\_1$m)

median <- numeric(6)

for(i in 1:6){median[i] = median(work\_1[,i] , na.rm = T)}

median

wd1 <- read.csv("Working\_data\_CO2\_00.csv")

wd1$m[wd1$m == "NA"] <- 1503

#wd1$m[wd1$m == NaN] = 1503

wd1$m[wd1$m == 0] <- 1503

1. Cleaning The MS:

summary(co2$MS)

co2$MS <- tolower(co2$MS)

co2$MS[445:455]

class(co2$MS)

co2$MS <- as.factor(co2$MS)

length(levels(co2$MS))

1. Cleaning Ft:

summary(co2$Ft)

library(stringr)

co2$Ft[co2$Ft == ""] <- NA

co2$Ft <- str\_replace\_all(co2$Ft, "[^[:alnum:]]", " ")

head(co2$Ft)

co2$Ft <- tolower(co2$Ft)

co2$Ft[1:20]

class(co2$Ft)

co2$Ft <- as.factor(co2$Ft)

length(levels(co2$Ft))

levels(co2$Ft)

co2$Ft[25:45]