# turn a date into a 'monthnumber' relative to an origin

monnb <- function(d)

{

lt <- as.POSIXlt(as.Date(paste("01-", d, sep = ""), "%d-%b-%Y"))

return (lt$year\*12 + lt$mon)

}

# compute a month difference as a difference between two monnb's

mondf <- function(d1, d2)

{

return (monnb(d2) - monnb(d1))

}

clean <- function(data\_in)

{

#Interest-Rate Convert from factor to string, replace % and convert to numeric

data\_in$int\_rate <- as.numeric(sub("%","",sapply(data\_in$int\_rate, as.character)))

#data\_in$int\_rate <- NULL

#Enploylength -- Keep only the digits and convert to digit

data\_in$emp\_length <- as.numeric(gsub("\\D","",sapply(data\_in$emp\_length,as.character)))

#data\_in$emp\_length <- NULL

# Zip code replacing xx

#print (data\_in$zip\_code)

data\_in$zip\_code <- as.character(sub("xx","",sapply(data\_in$zip\_code,as.character)))

#data\_in$zip\_code <- NULL

# Revol\_util -- Replace % and convert to float

data\_in$revol\_util <- as.numeric(sub("%","",sapply(data\_in$revol\_util, as.character)))

#data\_in$revol\_util <- NULL

# Date Transformation

data\_in$earliest\_cr\_line <- mondf(data\_in$earliest\_cr\_line, "Dec-2016")

# Remove all the columns that have more than 85% of NA

col\_names <- names(data\_in)[colSums(is.na(data\_in)) >= 0.85 \* dim(data\_in)[1]]

return (drop\_columns(data\_in, col\_names))

}

drop\_columns <- function(data\_in, col\_names)

{

for (cn in col\_names) {

data\_in[,paste(cn)]<-NULL

}

return(data\_in)

}

var\_cleaner = function(data\_in)

{

data\_in$loan\_status.new <- data\_in$loan\_status == 'Fully Paid'

data\_in['loan\_status.new'] <- lapply(data\_in['loan\_status.new'], as.integer)

data\_in$annual\_inc <- as.integer(data\_in$annual\_inc)

data\_in$dti <- as.numeric(data\_in$dti)

data\_in$installment <- as.numeric(data\_in$installment)

data\_in$last\_pymnt\_amnt <- as.numeric(data\_in$last\_pymnt\_amnt)

data\_in$installment <- as.numeric(data\_in$installment)

data\_in$pymnt\_plan <- as.numeric(data\_in$pymnt\_plan)

data\_in$out\_prncp <- as.numeric(data\_in$out\_prncp)

data\_in$out\_prncp\_inv <- as.numeric(data\_in$out\_prncp\_inv)

data\_in$recoveries <- as.numeric(data\_in$recoveries)

data\_in$total\_pymnt <- as.numeric(data\_in$total\_pymnt)

data\_in$total\_pymnt\_inv <- as.numeric(data\_in$total\_pymnt\_inv)

data\_in$total\_rec\_int <- as.numeric(data\_in$total\_rec\_int)

data\_in$total\_rec\_late\_fee <- as.numeric(data\_in$total\_rec\_late\_fee)

data\_in$total\_rec\_prncp <- as.numeric(data\_in$total\_rec\_prncp)

return(data\_in)

}