**Data Clean Code**

**# dataset uses the data “ml-100k” folder, which inturn is obtained from http://grouplens.org/datasets/movielens/**

**Libraries Used**

library(plyr)

**Code Doc**

**#initialize libraries**

library(plyr)

**##preparing dataTbl**

dataTbl <- read.table(file="ml-100k/u.data")

**#assign names to columns for ease of processing**

colnames(dataTbl) <- c("user\_id", "item\_id", "rating", "timestamp")

**#check import was okay and table is as expected**

# str(dataTbl)

# head(dataTbl)

# tail(dataTbl)

**## preparing userTbl**

userTbl <- read.table(file="ml-100k/u.user", sep ="|")

**#assign names to columns for ease of processing**

colnames(userTbl) <- c("user\_id", "age" , "gender","occupation", "zip\_code")

**#dropping zip code - decided not useful for me**

userTbl <- subset(userTbl, select = -c(zip\_code) )

**#check import was okay and table is as expected**

# str(userTbl)

genreTbl <- read.table(file="ml-100k/u.genre", sep = "|", quote = "" )

**#preparing itemTbl**

itemTbl <- read.table(file="ml-100k/u.item", sep ="|", quote = "")

**#create vector used to assign names in for columns:**

genreVect <- as.vector(genreTbl[["V1"]])

**#change Childern's to childerns to prevent errors**

genreVect[genreVect %in% "Children\'s"] <- "Childrens"

**#assign names to columns for ease of processing - important**

colnames(itemTbl) <- c(c("item\_id", "movie\_title", "release\_date", "video\_release\_date",

"IMDb\_URL"), genreVect)

**#drop video\_release\_date (seems to not be filled) and URL from table**

itemTbl <- subset(itemTbl, select = -c(IMDb\_URL, video\_release\_date) )

**#fix dates field**

itemTbl$release\_date <- as.Date(itemTbl$release\_date, "%d-%b-%Y")

**##create unified table**

unifiedTbl <- merge(dataTbl, itemTbl)

unifiedTbl <- subset(unifiedTbl, select = -c(item\_id) )

unifiedTbl <- merge(unifiedTbl, userTbl)

**# duplicate rows (rows with a unique user\_id and movie title)**

unifiedTbl <- unique(unifiedTbl)

**#function to create a single genre of field, applying "multiple" to movies with**

**#multiple genres because I may want the genre fields as a single variable for**

**#ease of processing/programming**

createGenreFieldSingle <- function(x){

#temporarally remove variables to make looping easier,

tempDat <- subset(x, select = -c(user\_id, rating, movie\_title,

release\_date, age, gender, occupation))

count <- 0

genre <- "unknown" #unknown genre is default

**#some movies are have multiple rating from same user... !**

**#check if there are muliple rows in x**

if(nrow(x) > 1){

#set tempDat to only have one row

tempDat <- head(tempDat, n = 1)

}

for (i in names(tempDat)){

if(tempDat[i] == 1){

count <- count + 1

genre <- i

}

}

if(count > 1){

genre <- "multiple"

}

names(genre) <- "genre"

return(genre)

}

**#will remove elements where user had voted twice for the same movie – important and slow**

genreDat <- ddply(unifiedTbl, ~user\_id + movie\_title, createGenreFieldSingle)

unifiedTblSingle <- merge(genreDat, unifiedTbl)

unifiedTblSingle <- subset(unifiedTblSingle, select =

c(user\_id, movie\_title, rating, genre,

release\_date, age, gender, occupation) )

**#output table as csv file**

write.csv(unifiedTblSingle, "Results/unifiedMLData\_time.csv", row.names = FALSE)

**#create file with possible muliple files added**

**#function to create potentially multiple values or rows**

createGenreFieldMultiple <- function(x){

#temporarally remove variables to make looping easier,

tempDat <- subset(x, select = -c(user\_id, rating, movie\_title,

release\_date, age, gender, occupation))

genreItem <- data.frame()

if(nrow(x) > 1){

#set tempDat to only have one row

tempDat <- head(tempDat, n = 1)

}

for (i in names(tempDat)){

if(tempDat[i] == 1){

genreItem<- rbind(genreItem,i)

}

genreItem<- rbind(genreItem, NA)

}

names(genreItem) <- "genre"

return(genreItem)

}

**#create multiple based on number of genres**

unifiedTblMulti <- ddply(unifiedTbl, ~movie\_title , createGenreFieldMultiple)

**#remove all NA**

unifiedTblMulti <- na.omit(unifiedTblMulti)

**#clean and remerge**

unifiedTblMulti <- merge(unifiedTblMulti, ddply(unifiedTbl, ~movie\_title + user\_id,

summarize, release\_date))

unifiedTblMulti <- merge(unifiedTblMulti, unifiedTbl)

**#some element re-added have no genre at all - removing from data**

unifiedTblMulti <- unifiedTblMulti[unifiedTblMulti$genre != "1",]

unifiedTblMulti <- subset(unifiedTblMulti,

select = c(user\_id, movie\_title, genre, rating,

release\_date, age, gender, occupation) )

**#write output**

write.csv(unifiedTblMulti, "Results/unifiedMLDataMulti\_time.csv", row.names = FALSE)

**References**

[**https://cran.r-project.org/doc/contrib/de\_Jonge+van\_der\_Loo-Introduction\_to\_data\_cleaning\_with\_R.pdf**](https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf)

[**http://tutorials.iq.harvard.edu/R/Rintro/Rintro.html**](http://tutorials.iq.harvard.edu/R/Rintro/Rintro.html)