# Multinomial logistic regression

#install.packages("rstudioapi")

current\_path=dirname(rstudioapi::getSourceEditorContext()$path)

setwd(current\_path)

#install.packages("readxl")

library(readxl)

eretail = read\_excel("Online Retail.xlsx")

dim(eretail)

names(eretail)

summary(eretail)

eretail[1:20,]

eretail = eretail[eretail$Country != "Unspecified",] # remove 'unspecified' country

eretail = eretail[eretail$Quantity > 0,] # remove returns/cancellations

eretail[1:20,]

eretail = eretail[is.na(eretail$CustomerID)==FALSE,] # remove unknown CustomerID

eretail$Amount = eretail$Quantity \* eretail$UnitPrice # compute amount per invoice item

eretail$StockCode=as.factor(eretail$StockCode)

eretail$Description=as.factor(eretail$Description)

eretail$CustomerID=as.factor(eretail$CustomerID)

eretail$Country=as.factor(eretail$Country)

eretail$InvoiceYear = as.factor(substr(eretail$InvoiceDate, 1, 4)) # extract year of invoice

eretail$InvoiceMth = as.factor(substr(eretail$InvoiceDate, 6, 7)) # extract month of invoice

eretail$InvoiceDay = as.factor(substr(eretail$InvoiceDate, 9, 10)) # extract DAy of invoice

eretail=eretail[eretail$InvoiceYear==2011,] #removing 2010 data as they are small compared to 2011 data

eretail=eretail[,-c(1, 5, 10)] # dropping InvoiceNo, InvoiceDate and InvoiceYear columns

summary(eretail)

summary(eretail$InvoiceDay)

levels(eretail$InvoiceDay)

# creating training and test data set

rm(.Random.seed, envir=globalenv()) #resetting random number

set.seed(1)

noobs=dim(eretail)[1] # number of observations

train.seeds=sample(noobs, noobs\*60/100) # 60% of observation for training set

length(train.seeds)

eretail.train=eretail[train.seeds,] # training set

eretail.test=eretail[-train.seeds,] # testing set

eretail[1:10,]

#Logistic regression classifier

# creating data set for Multinomial logistic regression

# Taking Quantity as frequency of purchases for each product name

names(eretail.train)

frequency.table = aggregate(Quantity ~ Country + InvoiceMth + InvoiceDay + StockCode, data = eretail.train, sum) #sum quantities over customerID

frequency.table[1:10,]

#creating case form Table from frequency Table

#install.packages("vcdExtra")

library(vcdExtra)

CaseForm.table = expand.dft(frequency.table, freq="Quantity")

#install.packages("openxlsx")

openxlsx::write.xlsx(CaseForm.table, "Case Form Table.xlsx",sheetName = "Case Form Table", col.names = TRUE, row.names = TRUE)

CaseForm.tab=read\_excel("Case Form Table.xlsx")

contin.table=ftable(CaseForm.tab, row.vars = 1:3, col.vars = 4 )

#flat.table=ftable(CaseForm.tab, row.vars = 1:3, col.vars = 4)

#contingency.table=data.frame(flat.table)

install.packages("vgam") # to fit multinomial logit models

library(VGAM)

LR.fit=vglm(StockCode~Country+InvoiceMth+InvoiceDay, data = contin.table, subset = train.seeds, family=multinomial)

summary(LR.fit)