**Codes R**

Probleme2\_1:

library(readr)

library(dplyr)

data <- read\_csv("Online\_Retail.csv")

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

### Nettoyage de donnees

# 1- Enlever "C" ET "A" au debut de InvoiceNo, car il doit etre un numero

a=grep("^[A-Z]", data$InvoiceNo)

data$tempo=data$InvoiceNo

data$tempo[a]=substr(data$tempo[a], 2, 8)

data$InvoiceNo=data$tempo

data=data[-9]

# 2- enlever les quantites negative

# j'ai remarque que les InvoiceID qui commencetn par C obt les quantite negative et A sont les objet abandonnes.

# donc, les 2 sont a enlever anyways => on enleve les quantites negatives pour nettoyer les donnees

data=subset(data, Quantity>=0)

# 3- to lower case

data$Description=tolower(data$Description)

# On peut garder CustomerID=NA, car on n'apas besoin de CustomerID pour faire notre analyse

# 4 enlever discount

unique(data$StockCode)

data=subset(data, StockCode!="D")

# a) Quel est le produit qui se retrouve le plus au sein des transactions?

B=summarise(group\_by(data, StockCode), Count=length(StockCode))

B=B[order(B$Count, decreasing = TRUE),]

C=summarise(group\_by(data, StockCode, Description), Count=length(Description))

subset(C, C$StockCode==B$StockCode[1])

# b) Quelle est la provenance engendrant le plus de transactions ?

B=summarise(group\_by(data, Country), Count=length(InvoiceNo))

B=B[order(B$Count, decreasing = TRUE),]

B$Country[1]

# c) Quel est le produit le plus rentable au sein des transactions ?

data$paid=data$Quantity\*data$UnitPrice

B=summarise(group\_by(data, StockCode),times\_perchased=length(StockCode), rentabilite=sum(paid))

B=B[order(B$rentabilite, decreasing = TRUE),]

subset(B, B$StockCode==B$StockCode[c(1,2)])[c(1,2)]

# d) Quel est le client ayant fait le plus de visites ?

B=summarise(group\_by(data, CustomerID), Count=length(unique(InvoiceDate)))

B=B[order(B$Count, decreasing = TRUE),]

B$CustomerID[1]

# enregistrer le fichier nettoye

saveRDS(data, file = "clean\_data.rds")

Probleme2\_2:

library(readr)

library(arules)

library(tidyr)

# Loader le fichier nettoye enregistre

data=readRDS("clean\_data.rds")

# Enlever les columnes qui ne sont pas necessaires pour cette analyse

data=data[,c(1,2)]

# combien d'item existe-il au magasin?

n\_item=length(unique(data$StockCode))

# rename coloumn names

names(data)[2]="item"

write.csv(data\_1,

"donneesTransac.csv",

row.names = F)

trans <- read.transactions(

file = "donneesTransac.csv",

format = "single",

sep = ",", header = T,

cols=c("InvoiceNo","item"),

rm.duplicates = T

)

rm(data)

inspect(trans[1:5])

# les 25 permieres items qui ont la plus grande frequence (support)

itemFrequencyPlot(trans, topN=25)

# 2eme solution pour probleme2\_1 a)

sort(itemFrequency(trans), decreasing = TRUE)[1]

# Rule d'association

assoc\_rules <- apriori(trans, parameter = list(supp=0.024, conf=0.7,maxlen=5))

Apriori\_df = inspect(assoc\_rules)

inspect(sort(assoc\_rules,by="lift"))

summary(assoc\_rules)