INPT
المعهد الوطني للبريد والمواصلات
ⵓⵙⵜⵉⵜⵓ ⵓⵏⵓⵔⵉⵙ ⵏ ⵜⵙⵡⵉⵍⵜ ⴰ ⵜⵎⴰⵡⴰⵙⵜ
Institut National des Postes et Télécommunications

RAPPORT PROJET SCORING

Réalisé par :

TAIBI Fadoua

EL MAGUI Fatiha

ZAZA Zakaria

EL AFFANE Nouamane

- **Chargement des données :**

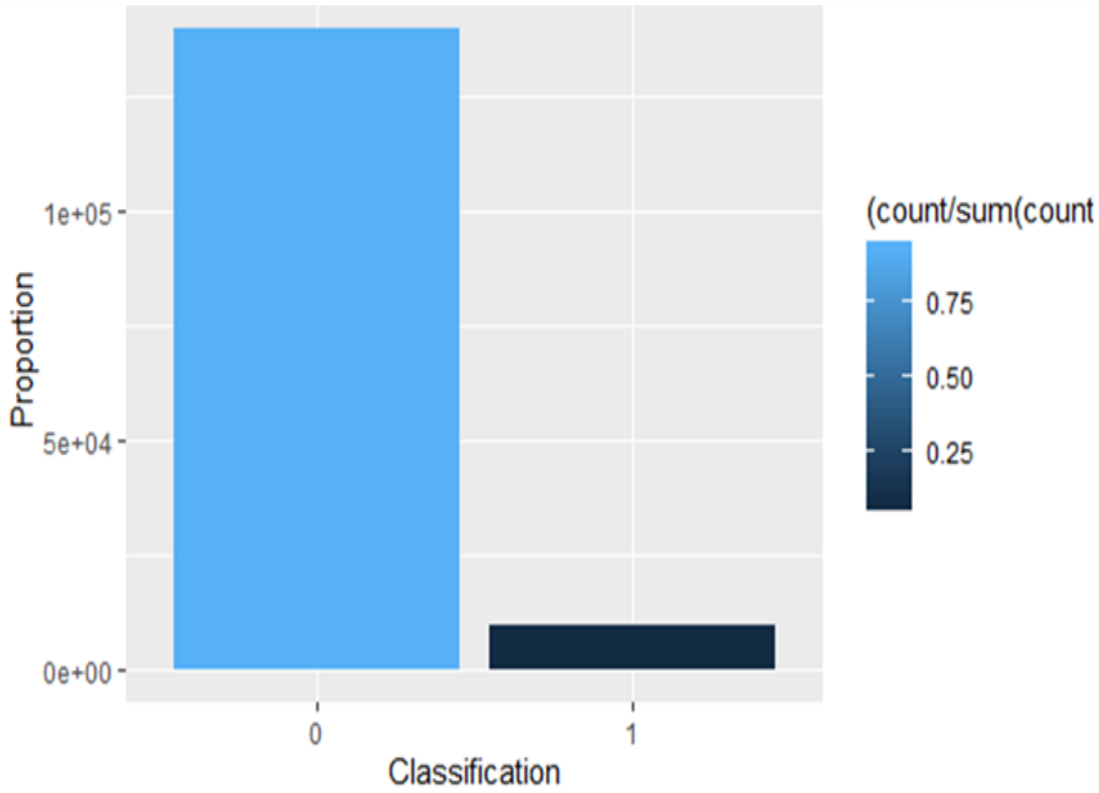
Tout d'abord, on charge notre jeu de données à l'aide de la fonction read_csv du package readr en enlevant la première colonne du dataset qui va ne servit à rien dans notre projet.

```
library(readr)
read_csv("C:/Users/hp/Downloads/scoring/Projet/ScoringTraining.csv")
ScoringTraining=ScoringTraining[,-1]
```

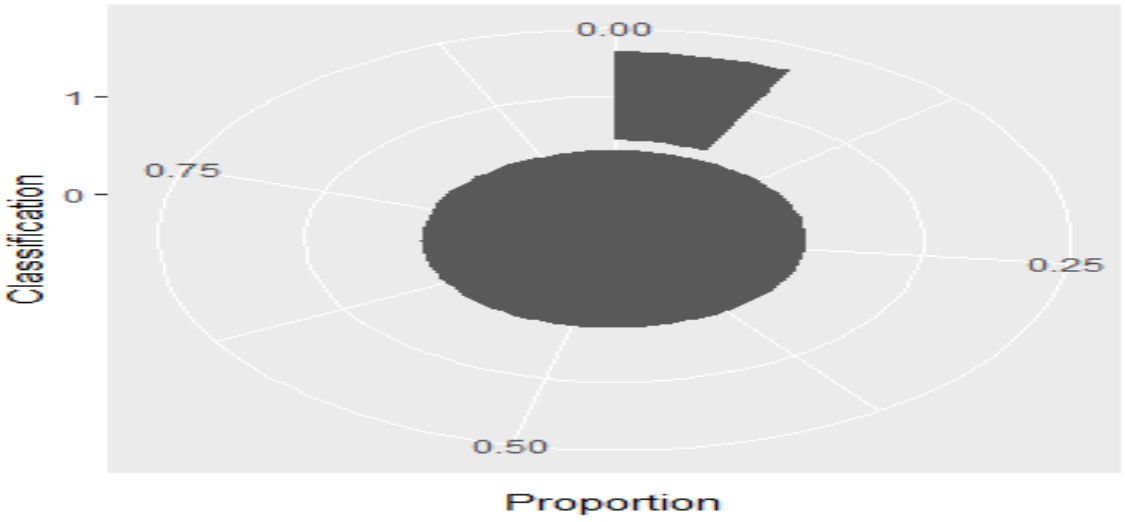
I-Phase de prétraitement :

1-La proportion de défaut :représente les clients qui ne seront pas capables de rembourser leurs dettes pendant les deux années à venir.

```
p2<- ggplot(ScoringTraining) + aes(x =factor(SeriousDlqin2yrs) , fill = (..count../sum(..count..))) + geom_bar()+
  ylab("Proportion") + xlab("Classification")
```




```
>   p<-pie <-p1 + coord_polar("y", start=0)
```



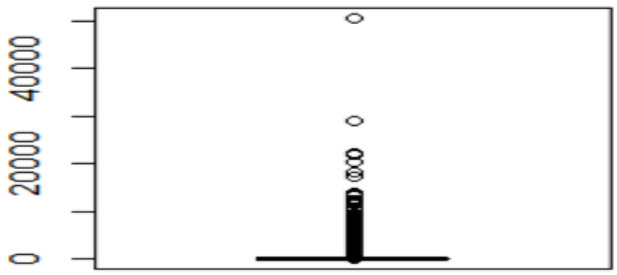
A partir des histogrammes et des diagrammes en camembert ,la proportion de défaut représente à peu près 5 % des personnes qui ne peuvent pas payer leur crédits dans les deux années à venir .

2-Les données extrêmes : Les outliers ou les valeurs aberrantes ce sont des valeurs distantes des observations et présentent des erreurs. On utilise les boites à moustaches :

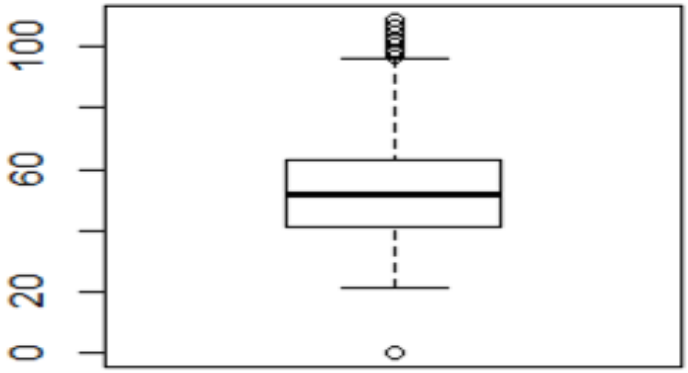
```
boxplot(ScoringTraining$RevolvingUtilizationOfUnsecuredLines)
boxplot(ScoringTraining$age)
boxplot(ScoringTraining$'NumberOfTime30-59DaysPastDueNotWorse')
boxplot(ScoringTraining$NumberOfTimes90DaysLate)
boxplot(ScoringTraining$'NumberOfTime60-89DaysPastDueNotWorse')
```

On remaque que :

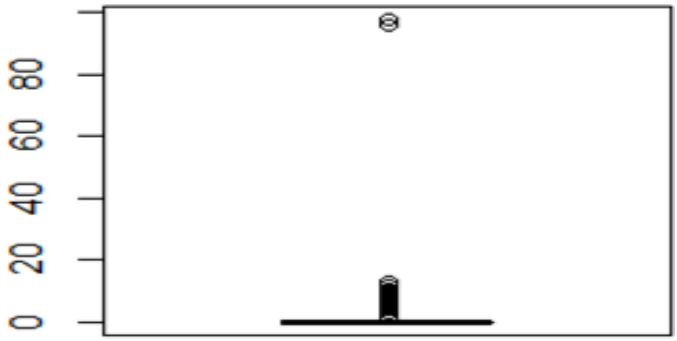
- ✓ pour la variable RevolvingUtilizationOfUnsecuredLines, on trouve qu'il y a des valeurs supérieures à 1 ce qui est qui impossible puisque la variable est en pourcentage.



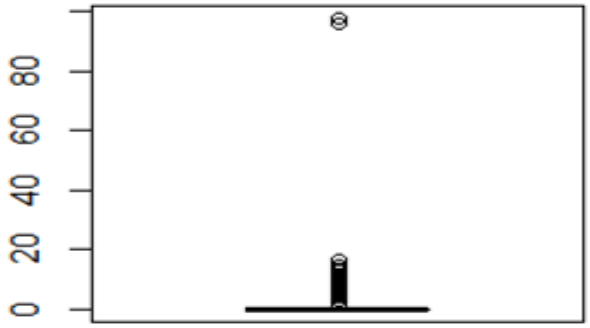
- ✓ Pour la variable age, , On remarque q'un client a l'age 0.



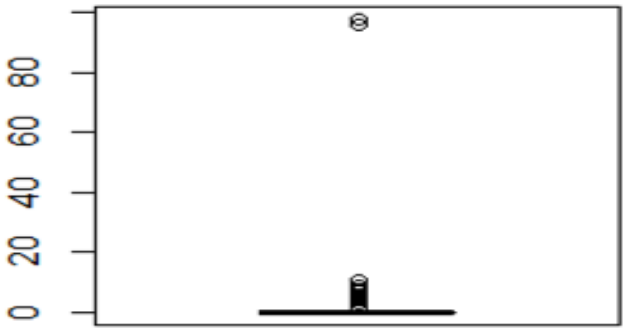
- ✓ Pour la variable NumberOfTime30-59DaysPastDueNotWorse, on remarque qu'elles y des valeurs très élevées ce qui illogique car la durée maximum d'emprunt est de 2 mois.



- ✓ Pour la variable NumberOfTimes90DaysLate , c'est la même remarque.



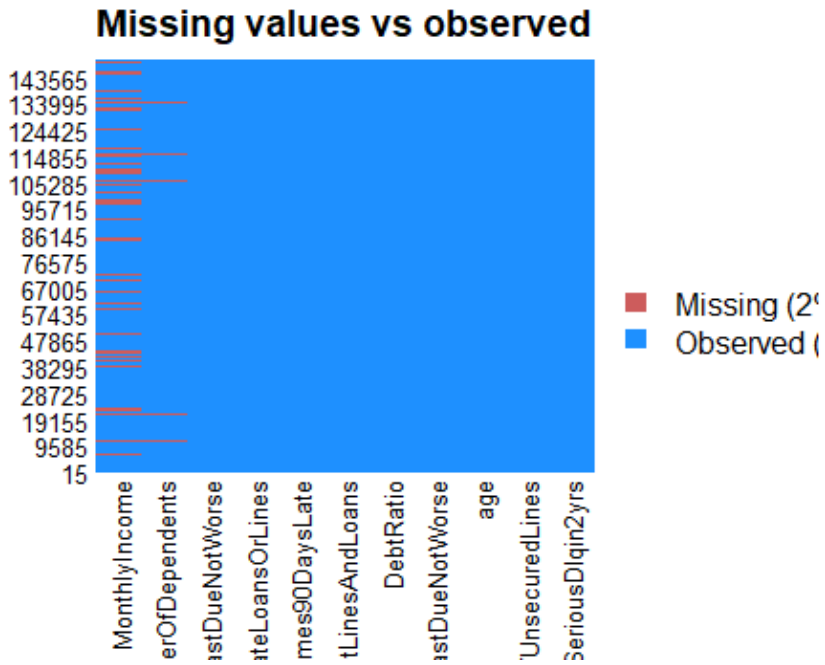
- ✓ Du même pour la variable NumberOfTime60-89DaysPastDueNotWorse,



3-les valeurs manquantes :

En utilisant la fonction missmap() du package Amelia :

```
library(Amelia)
missmap(ScoringTraining, main = "Missing values vs observed")
```



on trouve les valeurs manquantes : On trouve comme résultat qu'il existe des valeurs manquantes pour deux variables :MothlyIncome et NumberOfDependents.

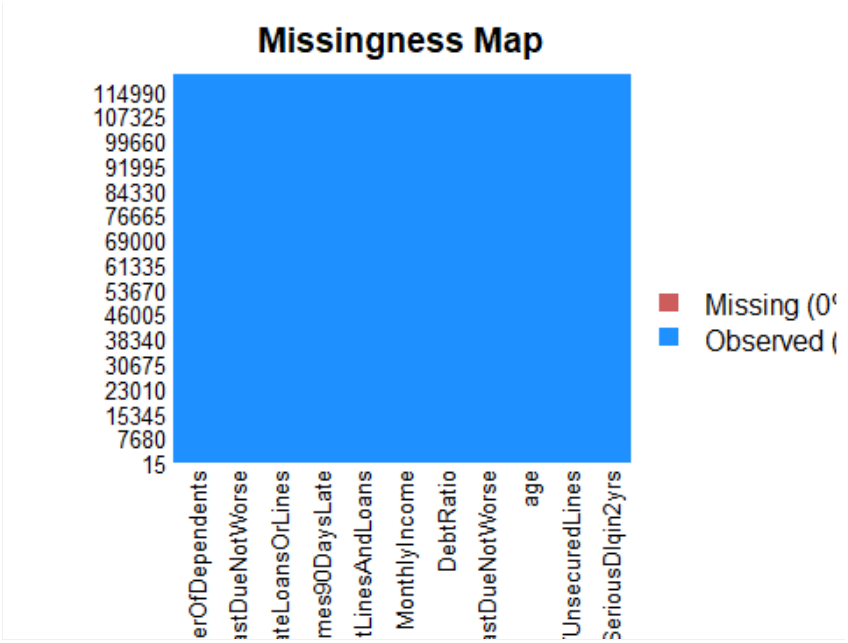
4-la manière de gérer ces cas :

Premièrement, on construit à nouveau le jeu de données par la fonction rbind qui permet de combiner le dataframe par ligne. Après, on remplace les outliers par des NA et en fin on supprime toutes les lignes qui contiennent des valeurs manquantes par la fonction omit().

```
>data<-rbind(ScoringTraining)

>data$age<-ifelse(data$age > 0,data$age,NA)
>data$RevolvingUtilizationOfUnsecuredLines<ifelse(data$RevolvingUtilizationOfUnsecuredLines<=1,data$RevolvingUtilizationOfUnsecuredLines,NA)

>data$`NumberOfTime30-59DaysPastDueNotWorse`<-ifelse(data$`NumberOfTime30-59DaysPastDueNotWorse` < 60,data$`NumberOfTime30-59DaysPastDueNotWorse`,NA)
>data$NumberOfTimes90DaysLate<-ifelse(data$NumberOfTimes90DaysLate < 60,data$NumberOfTimes90DaysLate,NA)

>data$`NumberOfTime60-89DaysPastDueNotWorse`<-ifelse(data$`NumberOfTime60-89DaysPastDueNotWorse` < 60,data$`NumberOfTime60-89DaysPastDueNotWorse`,NA)
>data$DebtRatio<-ifelse(is.na(data$MonthlyIncome),NA,data$DebtRatio)

>newdata<-na.omit(data)
```

On verifie que notre jeu de données ne contiennent pas les valeurs manquantes :

```
missmap(newdata)
dim(newdata)
```



(le dataset contient d'abord 120120 lignes)

Autre façon : On peut aussi, à la place de supprimer les lignes contenant les NA, remplacer les valeurs manquantes par une moyenne ou une médiane de la variable :

```
library(tidyr)
>data_rep <- data_1 %>%
 mutate(MonthlyIncome=replace_na(MonthlyIncome, mean(MonthlyIncome, na.rm=TRUE)))
summary(data_rep$MonthlyIncome)
>data_rep1 <- data_1 %>%
 mutate(NumberOfDependents=replace_na(NumberOfDependents, mean(NumberOfDependents, na.rm=TRUE)))
summary(data_rep$NumberOfDependents)
missmap(data_rep)
missmap(data_rep1)
newdata<-na.omit(data_1)
```

5-d'équilibrer les données : Afin d'équilibrer les données,On cherche en premier temps le nombre de 0 et de 1 dans la variable SeriousDlqin2yrs:

```
>   table(newdata$SeriousDlqin2yrs)

     0      1
111847   8273
```

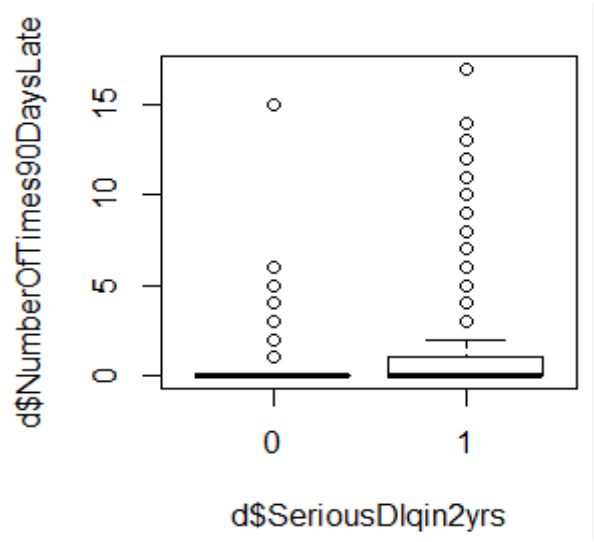
On utilise la fonction Down Simple() du package caret pour avoir la même proportion que la classe minoritaire qui est la classe 1 dans notre cas. Tout d'abord, on a convertit la variable SeriousDlqin2yrs en facteur, après on a appliqué le downSimpling :

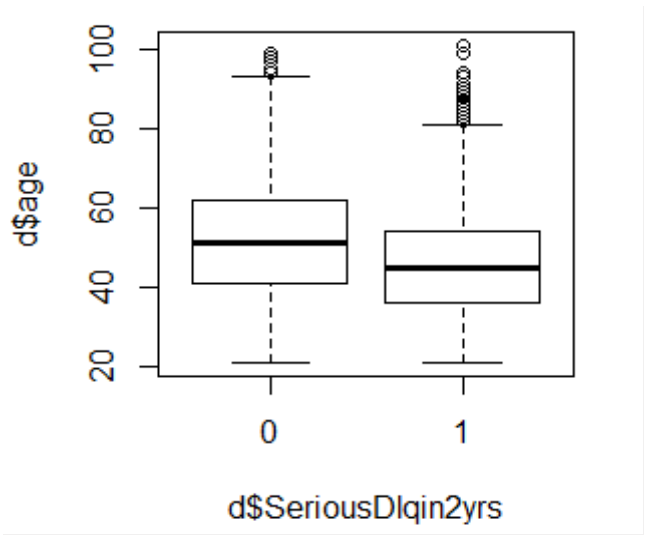
```
>    library(caret)
> fac <- factor(newdata$SeriousDlqin2yrs)
>    d<-downSample(newdata,fac)
>    table(d$SeriousDlqin2yrs)

   0    1
8273 8273
```

6- Les meilleures variables de prédictions :

Essayons maintenant d'identifier les meilleures variables pour la prédiction, pour ce faire on trace les boites à moustaches des deux classes pour chaque variable :

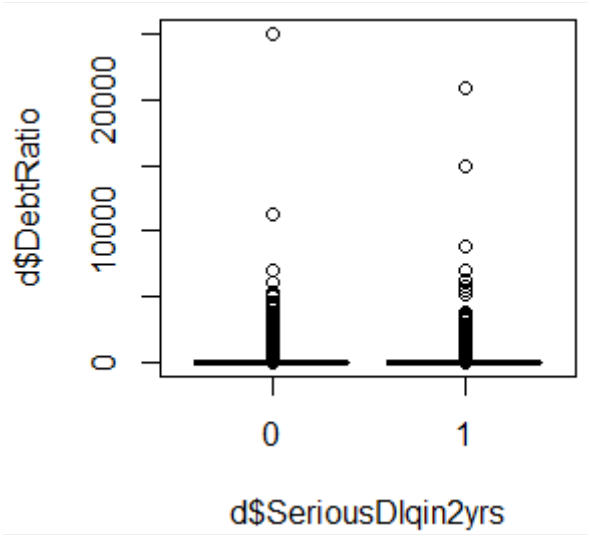
```
d$SeriousDlqin2yrs<-factor(d$SeriousDlqin2yrs, levels=c(1,0))
boxplot(d$age~d$SeriousDlqin2yrs,d)
boxplot(d$RevolvingUtilizationOfUnsecuredLines~d$SeriousDlqin2yrs,d)
boxplot(d$`NumberOfTime30-59DaysPastDueNotWorse`~d$SeriousDlqin2yrs,d)
boxplot(d$DebtRatio~d$SeriousDlqin2yrs,d)
boxplot(d$MonthlyIncome~d$SeriousDlqin2yrs,d)
boxplot(d$NumberOfOpenCreditLinesAndLoans~d$SeriousDlqin2yrs,d)
boxplot(d$NumberOfTimes90DaysLate~d$SeriousDlqin2yrs,d)
boxplot(d$NumberRealEstateLoansOrLines~d$SeriousDlqin2yrs,d)
boxplot(d$`NumberOfTime60-89DaysPastDueNotWorse`~d$SeriousDlqin2yrs,d)
boxplot(d$NumberOfDependents~d$SeriousDlqin2yrs,d)
```

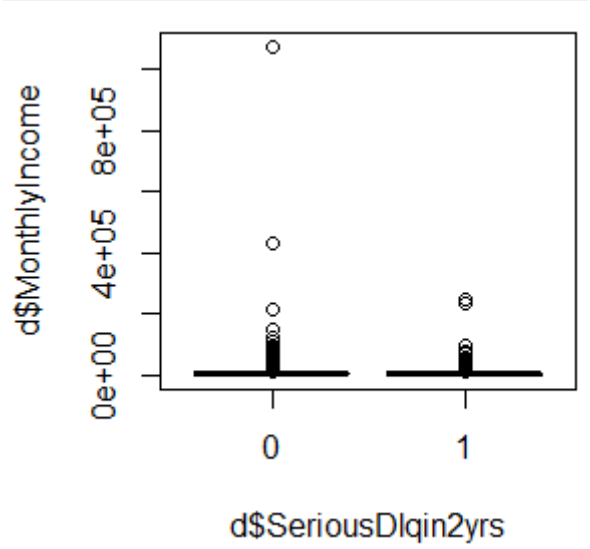


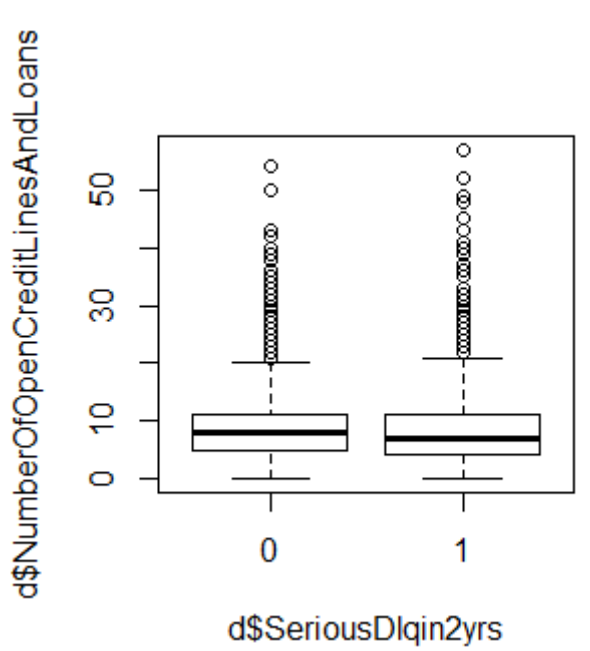


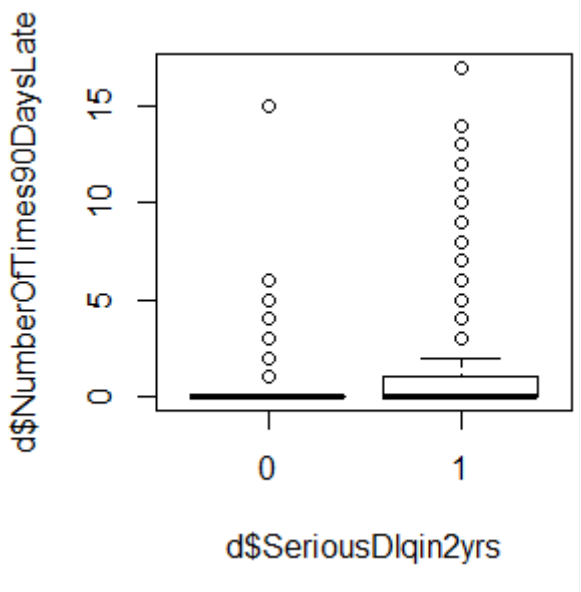
d$RevolvingUtilizationOfUnsecuredLines
0
2000
6000
0
1
d$SeriousDlqin2yrs

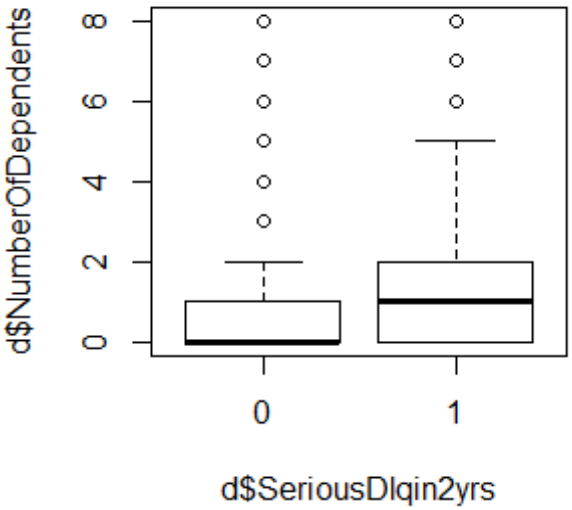
d$`NumberOfTime30-59DaysPastDueNotWorse
0
2
4
6
8
12
0
1
d$SeriousDlqin2yrs


d$DebtRatio
0
10000
20000
0
1
d$SeriousDlqin2yrs


d$MonthlyIncome
0e+00
4e+05
8e+05
0
1
d$SeriousDlqin2yrs


d$NumberOfOpenCreditLinesAndLoans
0
10
30
50
0
1
d$SeriousDlqin2yrs





En analysant les boites à moustaches obtenues, on remarque pour la majorité des variables que la représentation des boites à moustaches se ressemble, ceci ne nous permet pas de distinguer les groupes des individus. Cependant, nous observons que les deux variables (age et NumberOfDependents) se caractérisent par des représentations différentes, ce qui nous permet de classifier les individus selon des groupes. Donc les meilleures variables à mettre en considération dans le modèle de prédiction sont : age et NumberOfDependents.