TP1

Statistique descriptive et tests d'hypothèse avec R

Christophe TIET

1-1: Variables Quantitatives

1-Analyse univariée

Commençons par observer notre jeu de données:

```
##     Nuitees                     Type         Origine      Montant
##  Min.   :1.000   Chambre seule   : 62   AGENCE  :33   Min.   : 50.0
##  1st Qu.:2.000   Demi-Pension    :123   DIRECT  :34   1st Qu.:100.0
##  Median :5.000   Pension Complete: 51   GUIDE   :70   Median :250.0
##  Mean   :4.839                          INTERNET:57   Mean   :266.1
##  3rd Qu.:7.000                          OFFICE  :25   3rd Qu.:400.0
##  Max.   :9.000                          PASSAGE :17   Max.   :657.0
##    Reduction          Age        Sexe              Motif
##  Min.   : 0.00   Min.   :18.00   F: 94   Professionnel:105
##  1st Qu.: 0.00   1st Qu.:32.00   M:142   Tourisme     :131
##  Median : 0.00   Median :50.50
##  Mean   : 4.75   Mean   :49.53
##  3rd Qu.: 5.00   3rd Qu.:64.00
##  Max.   :58.40   Max.   :85.00
##        Chaine
##  Campanule:51
##  Etophotel:57
##  Formule 0:68
##  Navotel  :60
##
##
```

Nous avons ici quatres variables quantitatives: Nuitees, Montant, Reduction et Age.

Tableau des écart-types:

```
##    Nuitees    Montant  Reduction        Age
##   2.781306 163.111979   9.233189  19.033989
```

Tableau des moyennes:

```
##    Nuitees    Montant  Reduction        Age
##   4.838983 266.118644   4.749576  49.529661
```

Tableau des coefficients de variation

```
##   Nuitees   Montant Reduction       Age
## 0.5747709 0.6129295 1.9440028 0.3842948
```

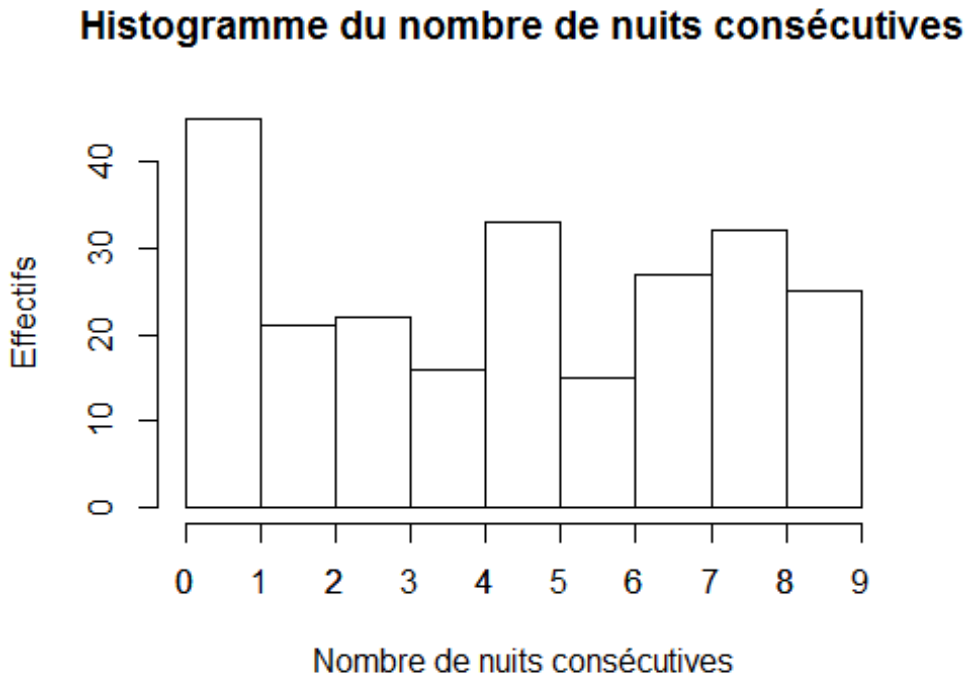
Tableau des coefficients d’asymétrie et d’aplatissement

```
##              skewness   kurtosis   n
## Nuitees   -0.01286046 -1.3786782 236
## Montant    0.36574409 -0.6831866 236
## Reduction  2.38303064  6.6762503 236
## Age        0.15695059 -1.1472946 236
```

On observe que la variable reduction se démarque des autres. Celle-ci est très dispersée,étalée à droite et non aplatie(cv=1.9, skewness=2.4, kurtosis=6.7)

Variable Nuitees

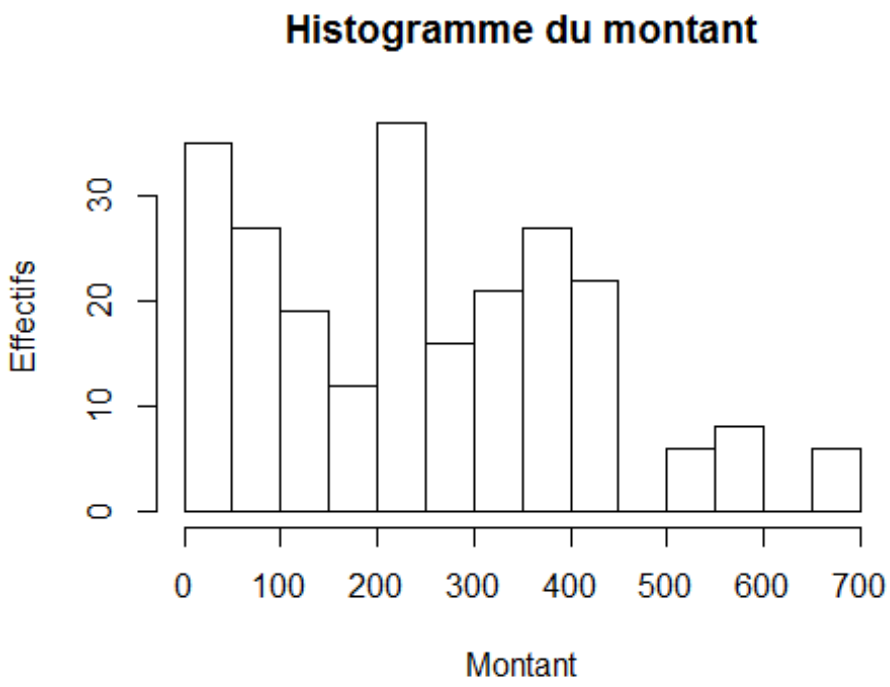
Cette variable correspond au nombre de nuits consécutives passées à l'hôtel par le client. Le temps de séjour moyen est de 4.84 nuits avec un minimum de 1 nuit et un maximum de 9 nuits.

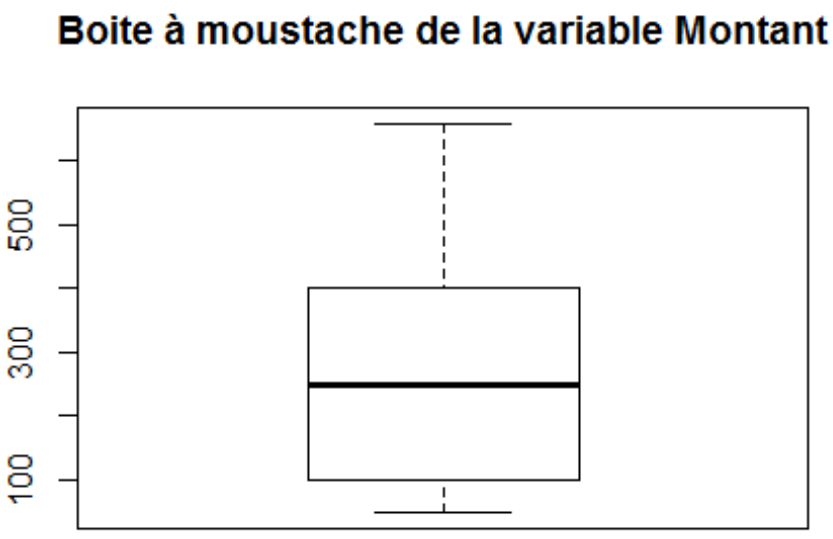


On peut remarquer qu'aucune tendence ne se dégage de cette histogramme. Le temps de séjour est relativement homogène.

Variable Montant

Le panier moyen des clients de ces hotels est de 266€, le panier minimum est de 50€ tandis que le plus gros panier est de 657€.



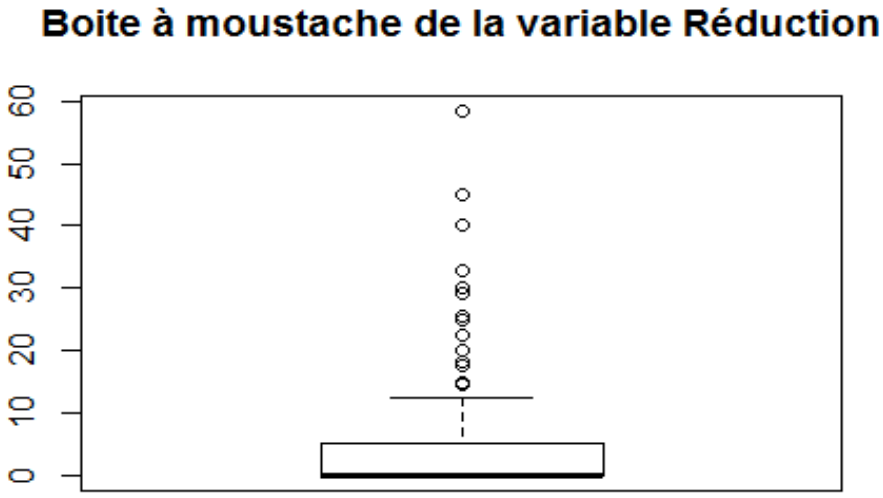


On observe une fréquence relativement élevé des paniers à 50€ ainsi que de paniers entre 200 et 250€. Les montants au-delà de 450 euros sont rares et sont relativement extrêmes comme nous pouvons le remarquer à l'aide d'une boite à moustache.

Variable Reduction

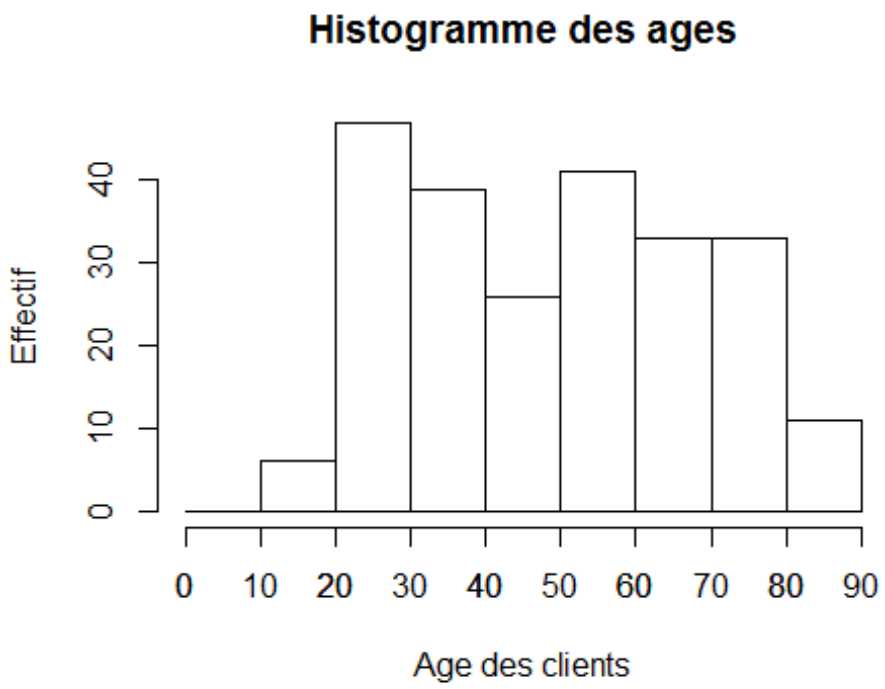
Table de contingence de la variable reduction:

```
##     0   2.5  3.65     5   7.3   7.5    10 10.95  12.5  14.6    15  17.5
##   163     7     4     6     1     6     2     1     9     1     3     5
## 18.25    20  22.5    25 25.55  29.2    30 32.85    40    45  58.4
##     1    13     3     2     1     2     2     1     1     1     1
```



Quasiment 69% de la clientèle ne bénéficie d'aucune réduction. On peut dès lors considérer comme valeur extrême les réductions au dela de 12.5%.

Variable Age



On peut remarquer que l'âge de la clientèle est relativement homogène entre 20 et 80 ans. On peut considérer que les hotels de notre jeu de données couvrent un spectre assez large au niveau de l'âge des clients.

2-Etude de liaison

Nous allons cherché un lien entre les variables Age/Montant, Age/Reduction Nuitées/Montant, Nuitée/reduction.

Tableau de correlation:

```
##           Nuitees Montant Reduction  Age
## Nuitees      1.00    0.94      0.26 0.12
## Montant      0.94    1.00      0.24 0.12
## Reduction    0.26    0.24      1.00 0.09
## Age          0.12    0.12      0.09 1.00
```

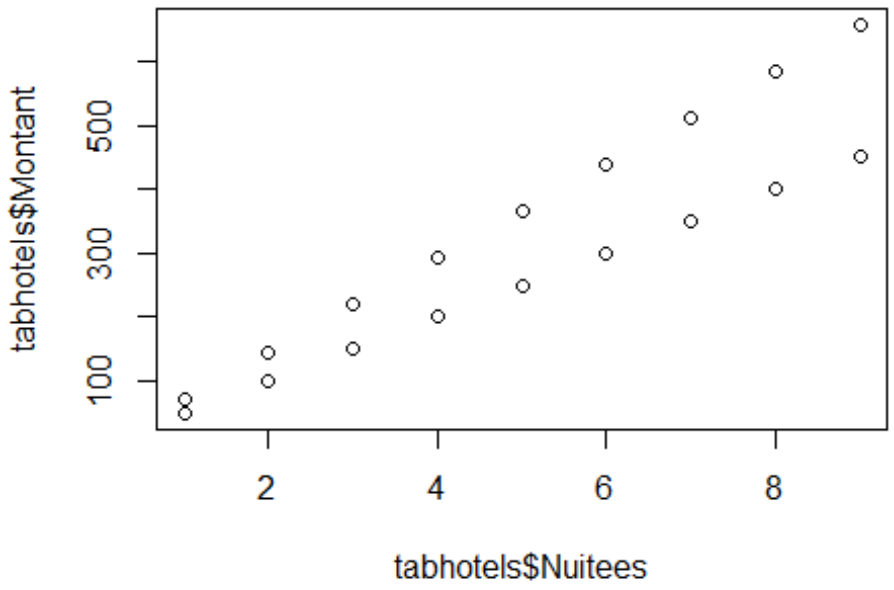
On remarquera ici une grande corrélation positive de 0.94 entre le montant et le nombre de nuitées. Des corrélations positives légères entre Reduction/Nuitées, Reduction/montant. Et enfin on observe une corrélation positive faible entre Age/montant et Age/Reduction

Nuitées/Montant

Nous avons remarqué une forte corrélation positive entre ces deux variables. Il semble évident que cette corrélation est significative puisque plus le nombre de nuitées est elevée et plus le montant payé est élevé.

Test de corrélation

```
##  Pearson's product-moment correlation
##
## data:  tabhotels$Nuitees and tabhotels$Montant
## t = 44.152, df = 234, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9293331 0.9571088
## sample estimates:
##       cor
## 0.9448967
```



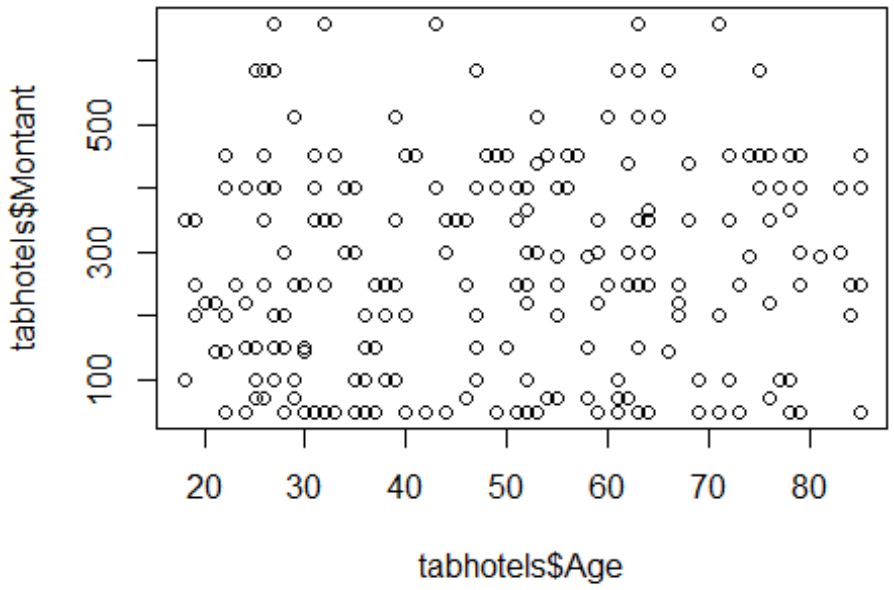
Effectivement, une relation existe entre le nombre de nuit et le montant payé par le client. On remarque qu'il semblerait qu'il y ait deux politiques de prix d'après notre nuage de points pour un même nombre de nuits. En fouillant un peu on peut se rendre compte que cela dépend du type de pension choisi par le client.

Age/Montant et Age/Reduction

Nous voulons voir si l'âge du client et le montant dépensé ou le taux de réduction sont deux variables corrélées. Nous avons remarqué une faible corrélation de 0.12 et 0.09.

```
##
##  Pearson's product-moment correlation
##
```

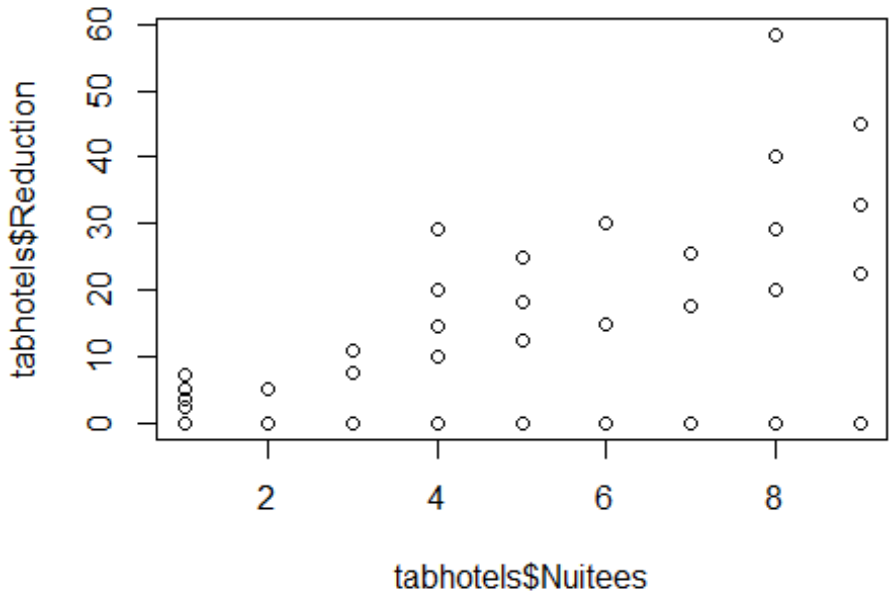
```
## data:  tabhotels$Age and tabhotels$Montant
## t = 1.8197, df = 234, p-value = 0.07009
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.009724325  0.242170345
## sample estimates:
##       cor
## 0.1181228


##  Pearson's product-moment correlation
##
## data:  tabhotels$Age and tabhotels$Reduction
## t = 1.3778, df = 234, p-value = 0.1696
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03843476  0.21494420
## sample estimates:
##        cor
## 0.08970601
```



On ne rejette pas l'hypothèse nulle: Aucune corrélation entre les deux variables. Nos données ne permettent donc pas d'affirmer qu'une corrélation existe entre l'age du client et le montant ou le taux de réduction.

Nuitées/Reduction et Montant/Reduction

```
##
##  Pearson's product-moment correlation
##
## data:  tabhotels$Nuitees and tabhotels$Reduction
## t = 4.1567, df = 234, p-value = 4.532e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1391833 0.3772896
## sample estimates:
##      cor
## 0.262223
```



Le test de Pearson rejette l'hypothèse nulle et nous obtenons une corrélation de 0.26. On peut donc constater une légère corrélation positive entre le taux de reduction et le nombre de nuits. La variable Nuitées et Montant étant fortement corrélées, il est donc assez logique que la corrélation entre Montant et Reduction soit significative.

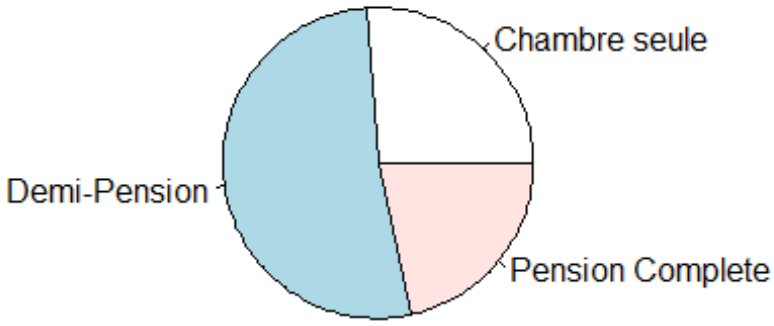
1-2: Variables Qualitatives

1-Tri à plat et graphique

Nous allons nous intéressé aux variables Type, Origine, Motif et Chaine

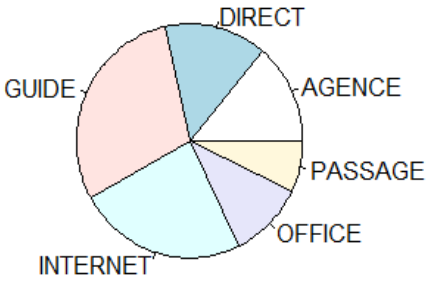
Variable Type

```
##
##    Chambre seule     Demi-Pension Pension Complete
##               62              123               51
```



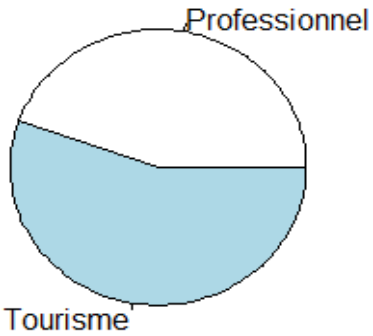
Variable Origine

```
##
##   AGENCE   DIRECT    GUIDE INTERNET   OFFICE  PASSAGE
##       33       34       70       57       25       17
```



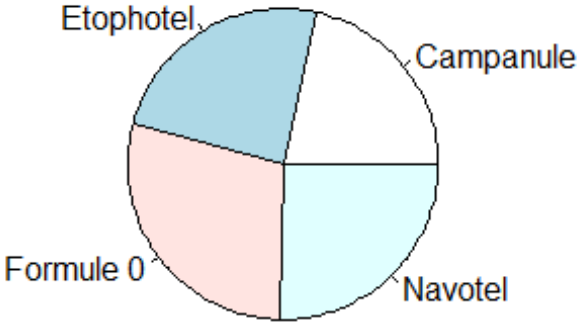
Variable Motif

```
##
## Professionnel      Tourisme
##           105           131
```



Variable Chaine

```
##
## Campanule Etophotel Formule 0   Navotel
##        51        57        68        60
```



2- Liaisons entre deux variables qualitatives

```
##  Pearson's Chi-squared test
##
## data:  tabhotels$Type and tabhotels$Origine
## X-squared = 9.237, df = 10, p-value = 0.5098


##  Pearson's Chi-squared test
##
## data:  tabhotels$Type and tabhotels$Motif
## X-squared = 1.0109, df = 2, p-value = 0.6032


##  Pearson's Chi-squared test
##
## data:  tabhotels$Type and tabhotels$Chaine
## X-squared = 14.326, df = 6, p-value = 0.0262
```

```
##  Pearson's Chi-squared test
##
## data:  tabhotels$Origine and tabhotels$Motif
## X-squared = 4.8856, df = 5, p-value = 0.43


##  Pearson's Chi-squared test
##
## data:  tabhotels$Origine and tabhotels$Chaine
## X-squared = 16.39, df = 15, p-value = 0.3566


##  Pearson's Chi-squared test
##
## data:  tabhotels$Motif and tabhotels$Chaine
## X-squared = 13.056, df = 3, p-value = 0.004517
```

D'après nos tests d'indépendance, les variables Type/Chaine et Motif/Chaine sont liées au seuil significatif de 5%.

2-a- Tableau de tri croisé Type/Chaine

```
##
##             Chambre seule Demi-Pension Pension Complete
##   Campanule            12           24               15
##   Etophotel             7           34               16
##   Formule 0            20           38               10
##   Navotel              23           27               10

##
##             Chambre seule Demi-Pension Pension Complete
##   Campanule          5.08        10.17             6.36
##   Etophotel          2.97        14.41             6.78
##   Formule 0          8.47        16.10             4.24
##   Navotel            9.75        11.44             4.24
```

Tableau de tri croisé Motif/Chaine

```
##
##             Professionnel Tourisme
##   Campanule            20       31
##   Etophotel            34       23
##   Formule 0            34       34
##   Navotel              17       43

##
##             Professionnel Tourisme
##   Campanule          8.47    13.14
##   Etophotel         14.41     9.75
```

```
##   Formule 0         14.41    14.41
##   Navotel            7.20    18.22
```

Tableaux de profils-lignes et profils-colonnes des variables Type/Chaine

```
##
##             Chambre seule Demi-Pension Pension Complete
##   Campanule     0.2352941    0.4705882        0.2941176
##   Etophotel     0.1228070    0.5964912        0.2807018
##   Formule 0     0.2941176    0.5588235        0.1470588
##   Navotel       0.3833333    0.4500000        0.1666667

##
##             Chambre seule Demi-Pension Pension Complete
##   Campanule     0.1935484    0.1951220        0.2941176
##   Etophotel     0.1129032    0.2764228        0.3137255
##   Formule 0     0.3225806    0.3089431        0.1960784
##   Navotel       0.3709677    0.2195122        0.1960784
```

Parmis la population de clients qui choisissent de séjourner en chambre seule:

- 19.3% choisissent Campanule
- 11.3% choisissent EtopHotel
- 32.2% choisissent Formule 0
- 37.1% choisissent Navotel

Parmis la population de clients qui choisissent l'hotel Campanule:

- 23.5% séjournent en Chambre seule
- 47% séjournent en demi-pension
- 29.6% séjournent en pension complète

Tableaux de profils-lignes et profils-colonnes des variables Motif/Chaine

```
##
##             Professionnel  Tourisme
##   Campanule     0.3921569 0.6078431
##   Etophotel     0.5964912 0.4035088
##   Formule 0     0.5000000 0.5000000
##   Navotel       0.2833333 0.7166667


##             Professionnel  Tourisme
##   Campanule     0.1904762 0.2366412
##   Etophotel     0.3238095 0.1755725
##   Formule 0     0.3238095 0.2595420
##   Navotel       0.1619048 0.3282443
```

Parmis les séjours à titre professionnels:

- 19% choisissent Campanule
- 32.4% choisissent EtopHotel
- 32.4% choisissent Formule 0
- 16.2% choisissent Navotel

Parmis les séjours à titre de tourisme:

- 23.7% choisissent Campanule
- 17.5% choisissent EtopHotel
- 25.9% choisissent Formule 0
- 32.8% choisissent Navotel

Parmis les clients de Campanule:

- 39.2% séjournent à titre professionnel
- 60.8% séjournent à titre personnel

Parmis les clients de EtopHotel:

- 59.6% séjournent à titre professionnel
- 40.4% séjournent à titre personnel

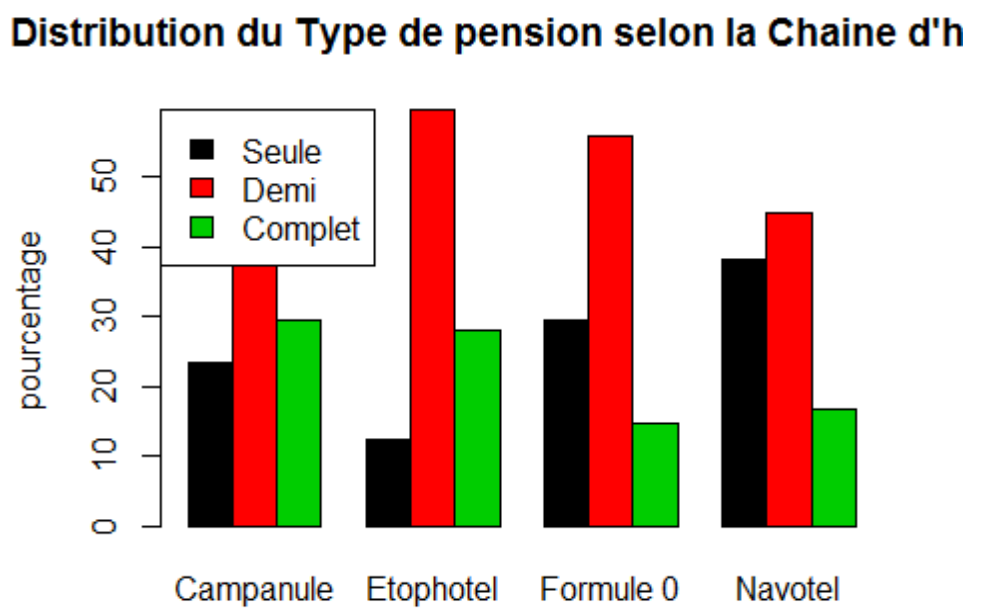
Parmis les clients de Formule 0:

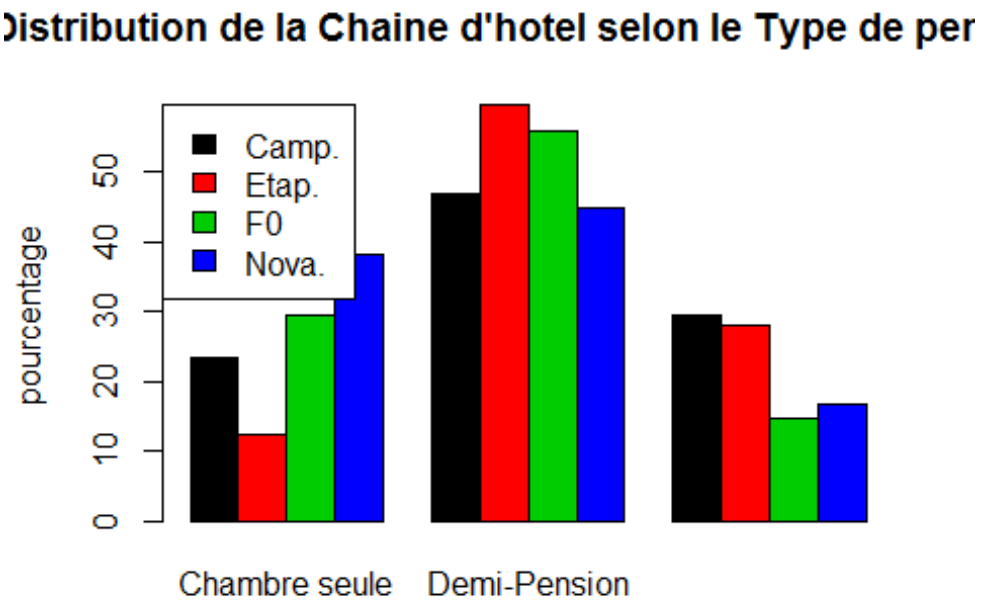
- 50% séjournent à titre professionnel
- 50% séjournent à titre personnel

Parmis les clients de Navotel:

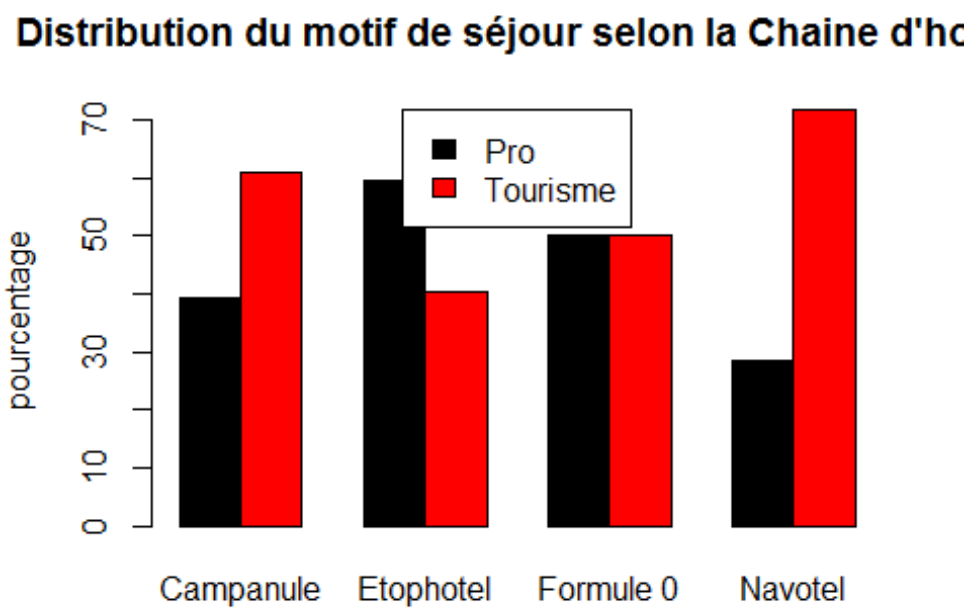
- 28.3% séjournent à titre professionnel
- 71.7% séjournent à titre personnel

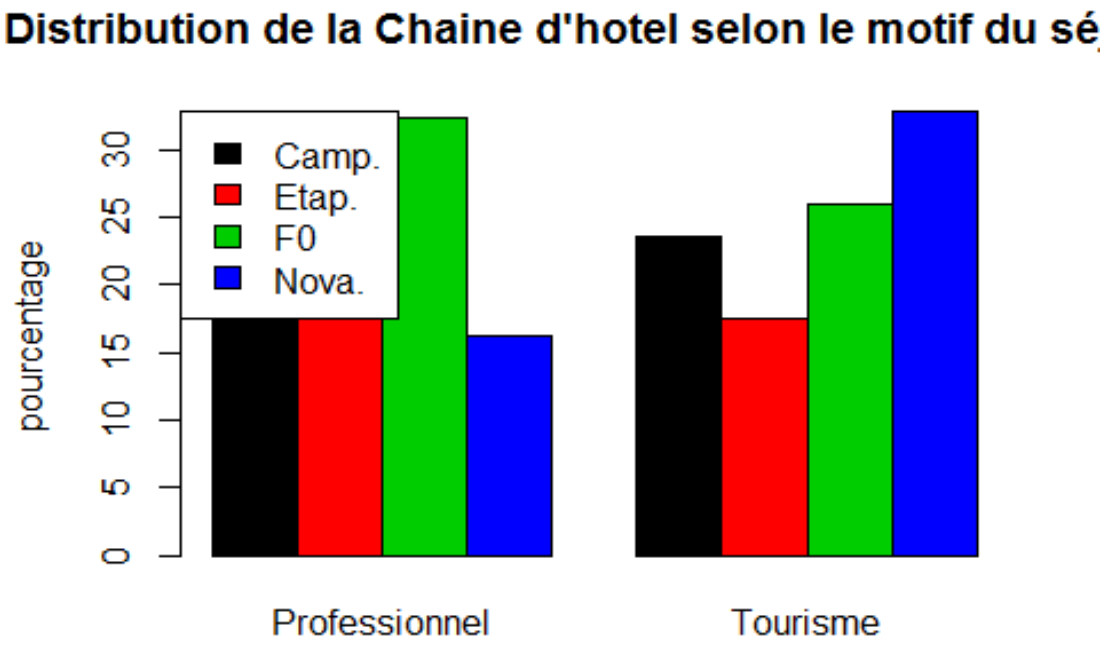
2-b- Graphique de Type/Chaine





Graphique de Motif/Chaine





2-c- Analyse de la liaison entre Type/Chaine

Effectifs observes:

```
##                   tabhotels$Chaine
## tabhotels$Type     Campanule Etophotel Formule 0 Navotel
##   Chambre seule           12         7        20      23
##   Demi-Pension            24        34        38      27
##   Pension Complete        15        16        10      10
```

Effectifs attendus

```
##                   tabhotels$Chaine
## tabhotels$Type     Campanule Etophotel Formule 0  Navotel
##   Chambre seule     13.39831  14.97458  17.86441 15.76271
##   Demi-Pension      26.58051  29.70763  35.44068 31.27119
##   Pension Complete  11.02119  12.31780  14.69492 12.96610
```

```
## X-squared
##  14.32552
```

Contributions aux khi-deux

```
##                   tabhotels$Chaine
## tabhotels$Type     Campanule Etophotel Formule 0   Navotel
##   Chambre seule    0.1459332 4.2467891 0.2552986 3.3229269
##   Demi-Pension     0.2505228 0.6201931 0.1848195 0.5833816
##   Pension Complete 1.4364114 1.1007343 1.4999902 0.6785200
```

Les fortes contributions (>14.32/12=1.19) concernent les couples de variables:

- Chambre seule/EtopHotel avec une sous-représentation

- Chambre seule/Navotel avec une sur-représentation

- Pension complète/Campanule avec une sur-représentation

- Pension complète/Formule 0 avec une sous-représentation

Analyse de la liaison entre Motif/Chaine

Effectifs observés

```
##                tabhotels$Chaine
## tabhotels$Motif Campanule Etophotel Formule 0 Navotel
##   Professionnel        20        34        34      17
##   Tourisme             31        23        34      43
```

Effectifs attendus

```
##                tabhotels$Chaine
## tabhotels$Motif Campanule Etophotel Formule 0  Navotel
##   Professionnel  22.69068  25.36017  30.25424 26.69492
##   Tourisme       28.30932  31.63983  37.74576 33.30508

## X-squared
##  13.05608
```

Contributions aux khi-deux

```
##                tabhotels$Chaine
## tabhotels$Motif Campanule Etophotel Formule 0   Navotel
##   Professionnel 0.3190627 2.9434611 0.4637611 3.5209470
##   Tourisme      0.2557372 2.3592627 0.3717169 2.8221331
```

Les fortes contributions (>13.06/8=1.63) concernent les couples de variables:

- Professionnel/EtopHotel avec une sur-représentation

- Professionnel/Navotel avec une sous-représentation

- Tourisme/EtopHotel avec une sous-représentation

- Tourisme/Navotel avec une sur-représentation

ANNEXE

```
library(Rcmdr)

tabhotels <- read.csv2("C:/Users/Chris/Desktop/Prepa_M2/Data_Mining/Cours_1/hotels.csv")

summary(tabhotels)

#VARIABLE QUANTITATIVE

#Statistiques descriptives

tabquanti=tabhotels[,c(-2,-3,-7,-8,-9)]
tabecarttype=apply(tabquanti,2,sd,na.rm=T)
tabmoy=apply(tabquanti,2,mean,na.rm=T)
tabcv=tabecarttype/tabmoy
tabecarttype

tabmoy

tabcv

numSummary(tabquanti, statistics=c("skewness","kurtosis"),type="1")



#Graphique de la variable nuitées

hist(tabhotels$Nuitees,breaks=seq(0,9,1),main="Histogramme du nombre de nuits consécutives",xlab="Nombre de nuits consécutives",ylab="Effectifs")
axis(side=1,at=seq(0,9,1))

#Graphique de la variable Montant

hist(tabhotels$Montant,breaks=seq(0,700,50),main="Histogramme du montant",xlab="Montant",ylab="Effectifs")

boxplot(tabhotels$Montant,main="Boite à moustache de la variable Montant")

#Table de contingence de la variable Reduction

table(tabhotels$Reduction)

#Boite à moustache de la variable Reduction

boxplot(tabhotels$Reduction,main="Boite à moustache de la variable Réduction")

#Graphique de la variable Age
```

```r
hist(tabhotels$Age,breaks=seq(0,90,10),main="Histogramme des ages",xlab="Age
des clients",ylab="Effectif")
axis(side=1,at=seq(0,90,10))

#Tableau de correlation

tabcor=cor(tabquanti,use="pairwise.complete.obs")
round(tabcor,2)

#Test de correlation

cor.test(tabhotels$Nuitees,tabhotels$Montant)

cor.test(tabhotels$Age,tabhotels$Montant)

cor.test(tabhotels$Age,tabhotels$Reduction)

cor.test(tabhotels$Nuitees,tabhotels$Reduction)


#Nuage de points

plot(tabhotels$Montant~tabhotels$Nuitees)

plot(tabhotels$Montant~tabhotels$Age)

plot(tabhotels$Reduction~tabhotels$Nuitees)


#VARIABLE QUALITATIVE

table(tabhotels$Type)

pie(table(tabhotels$Type))

table(tabhotels$Origine)

pie(table(tabhotels$Origine))

table(tabhotels$Motif)

pie(table(tabhotels$Motif))

table(tabhotels$Chaine)

pie(table(tabhotels$Chaine))


#Test de liaison

chisq.test(tabhotels$Type,tabhotels$Origine)
```

```r
chisq.test(tabhotels$Type,tabhotels$Motif)

chisq.test(tabhotels$Type,tabhotels$Chaine)

chisq.test(tabhotels$Origine,tabhotels$Motif)

chisq.test(tabhotels$Origine,tabhotels$Chaine)

chisq.test(tabhotels$Motif,tabhotels$Chaine)

#Tableau de tri croisé

TC=table(tabhotels$Chaine,tabhotels$Type)
TC

TCP=(TC/236)*100
round(TCP,digits=2)

MC=table(tabhotels$Chaine,tabhotels$Motif)
MC

MCP=(MC/236)*100
round(MCP,digits=2)

#Tableau de Profils lignes et profils-colonnes


TCpl=prop.table(TC,1)
TCpl

TCpc=prop.table(TC,2)
TCpc


MCpl=prop.table(MC,1)
MCpl
MCpc=prop.table(MC,2)
MCpc

#Graphique en baton justaposé

barplot(t(TCpl*100),beside=T,col=1:3,ylab="pourcentage",main="Distribution du
Type de pension selon la Chaine d'hotel")
legend("topleft",legend=c("Seule","Demi","Complet"),fill=1:3)

barplot(TCpl*100,beside=T,col=1:4,ylab="pourcentage",main="Distribution de la
Chaine d'hotel selon le Type de pension")
legend("topleft",legend=c("Camp.","Etap.","F0","Nova."),fill=1:4)



barplot(t(MCpl*100),beside=T,col=1:2,ylab="pourcentage",main="Distribution du
motif de séjour selon la Chaine d'hotel")
legend("top",legend=c("Pro","Tourisme"),fill=1:2)
```

```r
barplot(MCpc*100,beside=T,col=1:4,ylab="pourcentage",main="Distribution de la Chaine d'hotel selon le motif du séjour")
legend("topleft",legend=c("Camp.","Etap.","F0","Nova."),fill=1:4)
```

#Analyse de liaison

```r
resutest1=chisq.test(tabhotels$Type,tabhotels$Chaine)
names(resutest1)
```

```
## [1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
## [7] "expected"  "residuals" "stdres"
```

```r
is.list(resutest1)
```

```
## [1] TRUE
```

```r
resutest1$observed

resutest1$expected

resutest1$statistic

resutest1$residuals^2 #contributions au khi-deux

resutest2=chisq.test(tabhotels$Motif,tabhotels$Chaine)
names(resutest2)
```

```
## [1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
## [7] "expected"  "residuals" "stdres"
```

```r
is.list(resutest2)
```

```
## [1] TRUE
```

```r
resutest2$observed

resutest2$expected

resutest2$statistic

resutest2$residuals^2
```