Analyse de données agroalimentaire

KIOYE Togo Jean Yves, Amadou Sakho NDIAYE, SAGHIR Saleck, GUEYE Babacar

17/03/2022

SUJET 5 : ANALYSE DE DONNEES AGROALIMENTAIRE

I. INTRODUCTION

Allant de la production agricole aux denrées alimentaires élaborées à partir de ressources naturelles et des techniques industrielles. L'agroalimentaire englobe les phases de transformation, conservation et de commercialisation des produits se situant entre le stade agricole et celui de la consommation.

Par ailleurs, ce secteur dispose d'un marché fort concurrentiel, dynamique. Ainsi, être réactif et performant est une des priorités des acteurs de ce secteur. Cependant, pour faire face à ces défis, maîtriser la gestion, le traitement et l'analyse des données pourrait être déterminant dans le sens où cela permet d'assurer la productivité, de garantir une meilleure qualité de leurs produits et de faire des modèles de prédictions afin d' améliorer le processus de transformation. Tenant compte de tous ces enjeux, nous comprenons maintenant pourquoi l'analyse de données intéresse les industriels de l'agroalimentaire.

Dans le cadre de notre projet annuel, nous sommes amenés à faire une étude des données sur de la découpe de fruits (ananas) sur deux années 2020 et 2021 pour une entreprise agroalimentaire.

L'objectif de cette étude est principalement d'expliquer la productivité en fonction des caractéristiques du fruit,mettre en évidence des typologies de fruits et mettre en place des modèles prédictifs. Pour cela, nous essayerons de :

- Comparer des données de 2020 à celles de 2021 par catégorie.
- Mettre en évidence des liaisons entre variables quantitatives et variables qualitatives.

Nous serons amenés à faire des analyses descriptives, exploratoires et à utiliser des outils statistiques nécessaires (classification, ACP, ACM) pour répondre à la problématique.

II. Présentation des données

La première étape serait de charger notre base de données `BDD_agro_2020_2021` .Elle comporte 113 observations et 29 variables.

```
#Chargement des données
mydata=read.csv("C:/Users/21412149/Desktop/ProjetS2/BDD_agro_2020_2021.csv",header = T,sep=
";",dec=",",stringsAsFactors = TRUE)
```

Nous disposons d'une base contenant toutes les informations nécessaires pour répondre aux questions pouvant se poser. Les données s'étalent sur deux années 2020 et 2021. Nous pouvons noter des variables relatives aux caractéristiques du fruit:

- `Colo` : la coloration de la chair à sa réception.
- `Brix` : le taux de sucre en %.
- `Acidité` : l'acidité du produit.
- `Taches` : Présence ou non de taches d'eau dans la chair.

- `Forme` : La forme est le rapport de la hauteur `Hauteur` sur le `diamètre` .
- `Hauteur` : hauteur des fruits (cm).
- `Diametre` : diamètre des fruits (cm)
- `Poids` : poids brut des fruits en kg.
- `Age` : son age à la réception.
- `Fournisseurs` : fournisseurs du produit MI ou KE.

Ensuite il y' a des variables d'évaluation du processus de transformation du fruit (pendant et après):

- `delais_transf` : délais entre la réception et la transformation en nombre de jours.
- `age_7j` : son age à la réception à 7 jours.
- `Note_7j` : notation organoleptique à 7 jours de vie (DVP-1)de 0 à 20.
- `Conf_7j` : conformité à 7 jours avec trois modalités).
- `Homo_7j` : homogénéité à la dégustation à DVP-1 (oui/non).
- `Carac_7j` : caractéristiques à 7 jours avec des modalités. Puis ces mêmes variables à 7jours sont étudiées au jour 8.
- `Com_recep` : commentaire à la réception avec deux nouvelles modalités présentes en 2021

Enfin, des variables de mesures de la qualité :

- `Productivité` : productivité de l'OF en kg/h/personne
- `Rendement` : rendement du produit semi-fini (OF).

1. Correction, conversion et recodage des variables

a. Correction

Deuxième étape: Corriger la variable `Rendement` pour l'année 2021 car elles ne sont pas comprises dans l'intervalle 0 et 1.

```
#Correction de la variable rendement. pour 2021 on divise par 100
mydata[mydata$Annee==2021,"Rendement"]<-round(((mydata[mydata$Annee==2021,"Rendement"])/100),
2)
```

b. Conversion

Seront converties les variables `Note_recep` , `Annee` , `Colo` , `Colo_7j` , `Note_7j` , `delais_transf` en facteurs car dans le cadre de notre étude, nous allons considérer que ces variables prennent peu de valeurs

```
mydata$Annee=as.factor(mydata$Annee)
mydata$Colo=as.factor(mydata$Colo)
mydata$Colo_7j=as.factor(mydata$Colo_7j)
mydata$Note_recep = as.factor(mydata$Note_recep)
mydata$Note_7j = as.factor(mydata$Note_7j)
mydata$delais_transf=as.factor(mydata$delais_transf)
mydata$delais_transf=factor(mydata$delais_transf,levels=c(0:12))
```

c. Recodage

Dans notre jeu de données les modalités des variables diffèrent d'une année à l'autre. Pour rendre comparables ces variables nous procéderons à un recodage des modalités afin d'avoir des modalités uniformes pour les deux années.

- Pour la variable `Com_recep` : En 2021 nous avons regroupé les modalités "acides" et "très acides" puis "doux et"très doux". Commentaire à la réception : TD -> D (Tres Doux à Doux) Commentaire à la réception : TA -> A (Tres acide à acide)

- Pour la variable `Carac_7j` : Nous avons regroupé “lb” et “bru” puis “fer” et “lev” Les `lb` et `fer`, faiblement représentées, sont intégrées respectivement dans `levure` et `brunissement` car elles dominent les données.
- Pour la variable `Note_recep` : Sont regroupés “3.5”et “3” puis “3.8” et “4”

Note à la reception : 3.5 ->3 Note à la reception : 3.8-> 4

```
#creation d'une copie de Com_rec
mydata$reco_Com_rec<-mydata$Com_rec

#on recode Com_rec en reco_Com_rec avec 2 modalités A=TA, D=TD
mydata$reco_Com_rec[mydata$reco_Com_rec=="TD"]<-"D"
mydata$reco_Com_rec[mydata$reco_Com_rec=="TA"]<-"A"
mydata$reco_Com_rec <- factor(mydata$reco_Com_rec, levels = c("A","E","D"))


# Pas de modalité lb et fer en 2020
mydata$reco_caract_7j<-mydata$Carac_7j
mydata$reco_caract_7j[mydata$reco_caract_7j=="lb"]<-"bru"
mydata$reco_caract_7j[mydata$reco_caract_7j=="fer"]<-"lev"

#Pas de note 3.5 et 3.8 en 2020
mydata$reco_Note_recep<-mydata$Note_recep
mydata$reco_Note_recep[mydata$reco_Note_recep=="3.5"]<-"3"
mydata$reco_Note_recep[mydata$reco_Note_recep=="3.8"]<-"4"
factor(mydata$reco_Note_recep, levels=c(2,3,4,5))
```

```
##   [1] 4 3 4 3 3 3 3 3 3 3 4 3 3 3 3 4 3 4 4 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 3 4 4 4
##  [38] 3 5 5 4 4 3 3 4 5 3 3 3 3 4 4 4 4 4 5 4 4 3 3 5 5 5 5 3 2 3 3 3 4 3 4 4 3
##  [75] 3 3 3 3 4 4 4 3 4 4 4 4 4 5 4 4 4 4 3 5 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 4 5
## [112] 3 4
## Levels: 2 3 4 5
```

2. Comparaison de certaines variables par année

Dans cette section, nous allons d’abord classer les variables par type (quantitative et qualitative) dans le but de faire des analyses pour une meilleure organisation, ensuite créer des bases en fonction des années. Ces étapes déjà faites, nous pourrons comparer les variables selon les caractéristiques du fruit, les variables d’évaluation qui dépendent des critères liés au fruit et des variables de performance de l’entreprise.

a. Création des bases de données pour les années 2020 et 2021

```
#creation d'une base avec les données de 2020
mydata2020<-mydata[mydata$Annee==2020,]
#creation d'une base avec les données de 2021
mydata2021<-mydata[mydata$Annee==2021,]
```

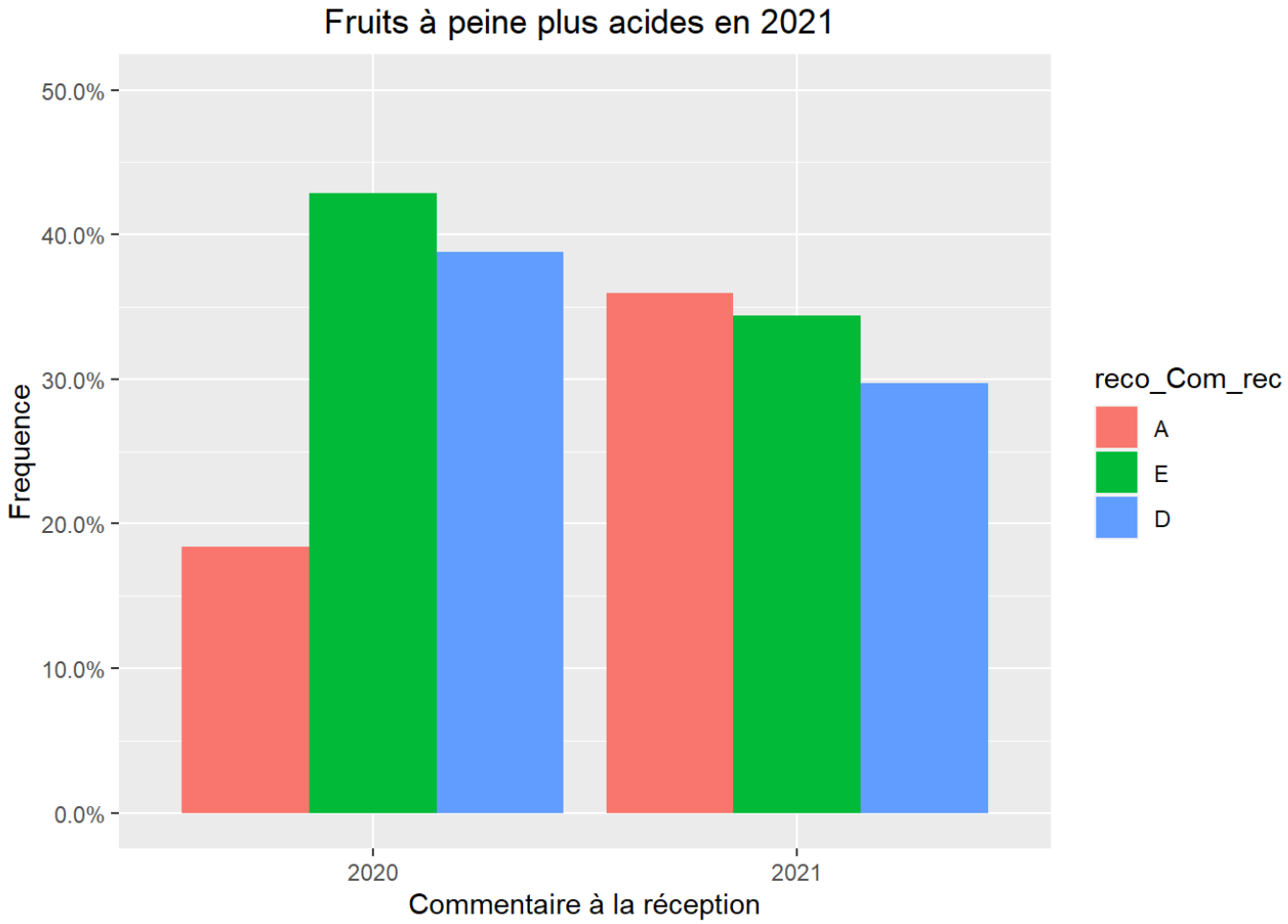
b. Variables qualitatives

ii. `Com_recep` commentaire à la réception avec des deux nouvelles modalités présentes en 2021

```
Com2020=(table(mydata2020$reco_Com_rec)/length(mydata2020$reco_Com_rec))*100
Com2021=(table(mydata2021$reco_Com_rec)/length(mydata2021$reco_Com_rec))*100
Compr=round(rbind(Com2020,Com2021),2)
colnames(Compr)=c("Acide %", "Doux %", "Equilibré %")
rownames(Compr)=c("Année 2020","Année 2021")
Compr
```

```
##            Acide % Doux % Equilibré %
## Année 2020   18.37  42.86       38.78
## Année 2021   35.94  34.38       29.69
```

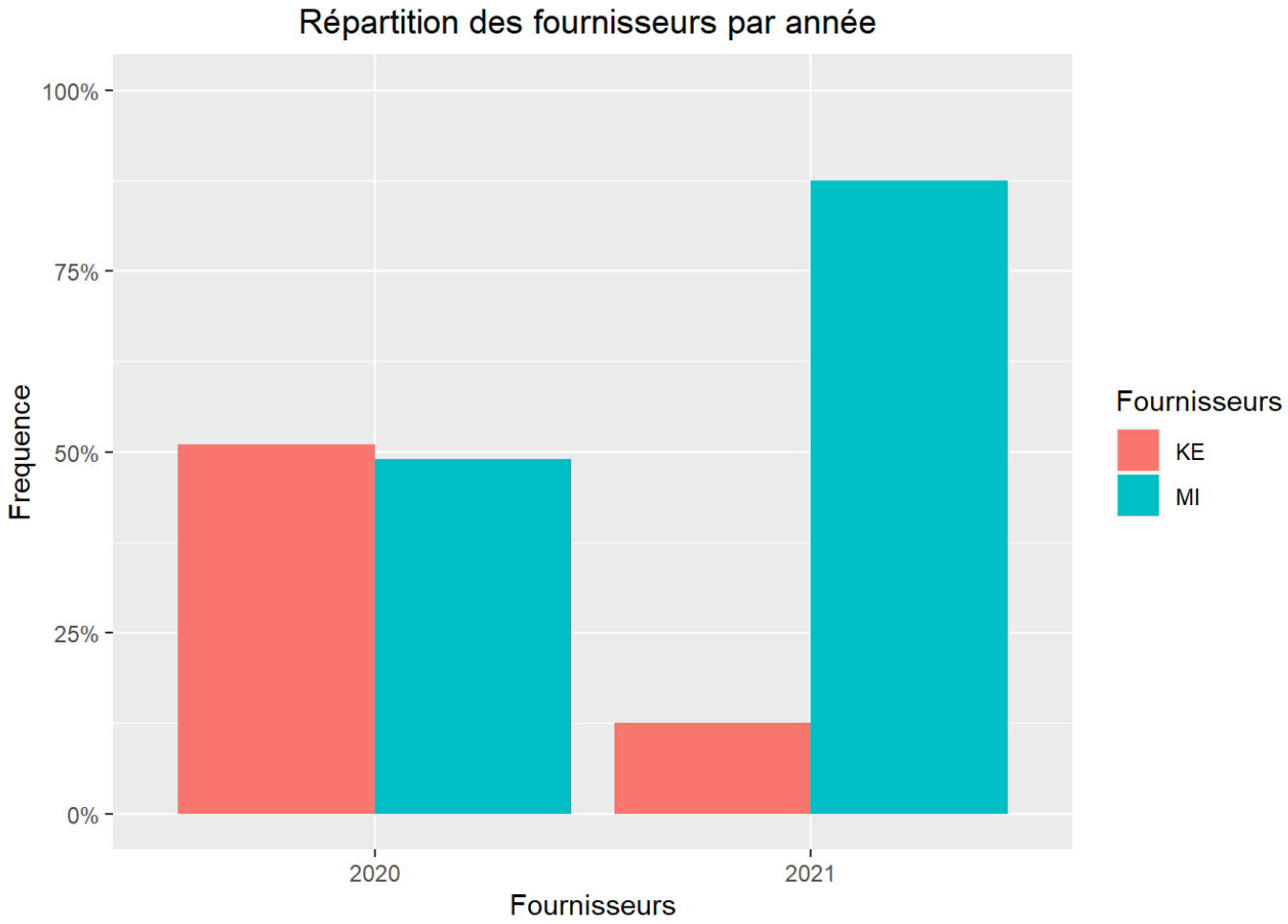
```
freqreco_Com_rec <- mydata %>%   group_by(Annee,reco_Com_rec)  %>% summarise(n=n(),.groups =
"drop_last")   %>% mutate(freq=n/sum(n))

ggplot(freqreco_Com_rec,  aes(x=Annee, fill = reco_Com_rec
                 , group = reco_Com_rec)) + geom_bar(aes(y=freq), stat="identity", position =
"dodge")+ coord_cartesian(ylim=c(0,0.5))+
  xlab("Commentaire à la réception") +
  ylab("Frequence")+ scale_y_continuous(labels = percent)+
  ggtitle("Fruits à peine plus acides en 2021 ")+theme(
  plot.title = element_text(hjust = 0.5))
```



Cette variable représente le commentaire à la réception des fruits. Le graphe montre que la proportion des fruits doux (D) et équilibrés (E) 39% est sensiblement la même dans les deux années et on remarque qu'en 2021, la proportion des fruits jugés acides (A) 36% est la plus importante,et plus élevée qu'en 2020.

iii. Variable `Fournisseurs`

```
##            KE en % MI en %
## Année 2020   51.02   48.98
## Année 2021   12.50   87.50
```



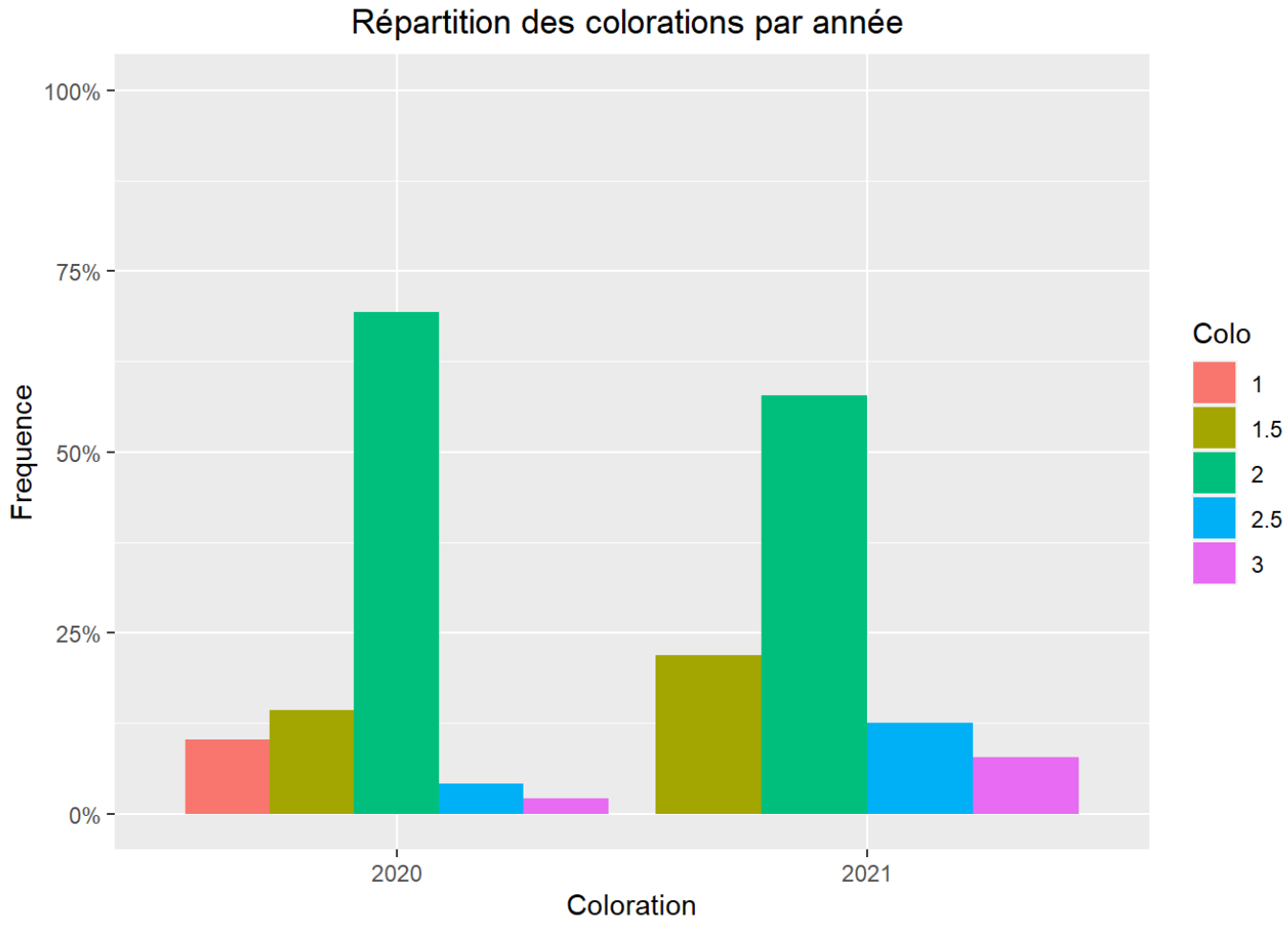
Une analyse comparative par année sur les fournisseurs de fruits nous permet de conclure à travers le graphique ci-dessous que KE et MI fournissaient sensiblement les mêmes quantités de fruits en 2020 respectivement réparties en 51% et 49%. Tandis qu'en 2021 MI passe en fournisseur principal avec plus de 87%.

iv. Variable `Colo` : la coloration de la chair à sa réception

```
colo2020=(table(mydata2020$Colo)/length(mydata2020$Colo))*100
colo2021=(table(mydata2021$Colo)/length(mydata2021$Colo))*100
colopr=round(rbind(colo2020,colo2021),2)

colnames(colopr)=c("1","1.5","2","2.5","3")
rownames(colopr)=c("Année 2020","Année 2021")
colopr
```

```
##               1   1.5     2   2.5    3
## Année 2020 10.2 14.29 69.39  4.08 2.04
## Année 2021  0.0 21.88 57.81 12.50 7.81
```

```
freqColo <- mydata %>%   group_by(Annee,Colo)  %>% summarise(n=n(),.groups = "drop_last")   %
>% mutate(freq=n/sum(n))


ggplot(freqColo,  aes(x=Annee, fill = Colo
                , group = Colo)) + geom_bar(aes(y=freq), stat="identity", position = "dodge"
)+ coord_cartesian(ylim=c(0,1))+
  xlab("Coloration") +
  ylab("Frequence")+ scale_y_continuous(labels = percent)+
  ggtitle("Répartition des colorations par année ")+theme(
  plot.title = element_text(hjust = 0.5))
```



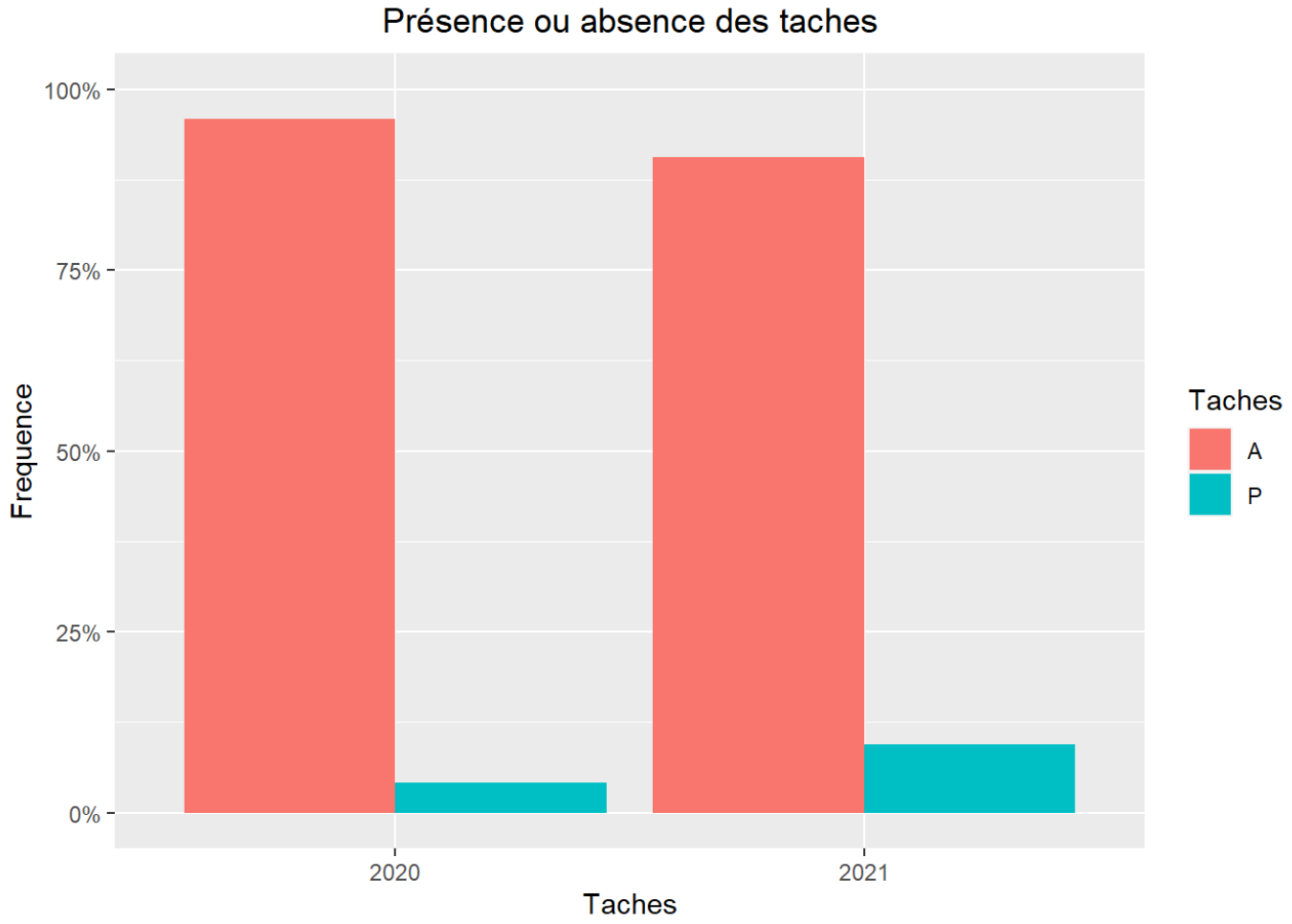
Pour les deux années, la coloration 2 est largement dominante par rapport aux autres colorations. En 2021 on constate une absence de la coloration 1.

v. Variable `Taches` : Présence ou non de taches d'eau dans la chair

```
Tache2020=(table(mydata2020$Taches)/length(mydata2020$Taches))*100
Tache2021=(table(mydata2021$Taches)/length(mydata2021$Taches))*100
Tachepr=round(rbind(Tache2020,Tache2021),2)
colnames(Tachepr)=c("Abscence %", "Presence %")
rownames(Tachepr)=c("Année 2020","Année 2021")
Tachepr
```

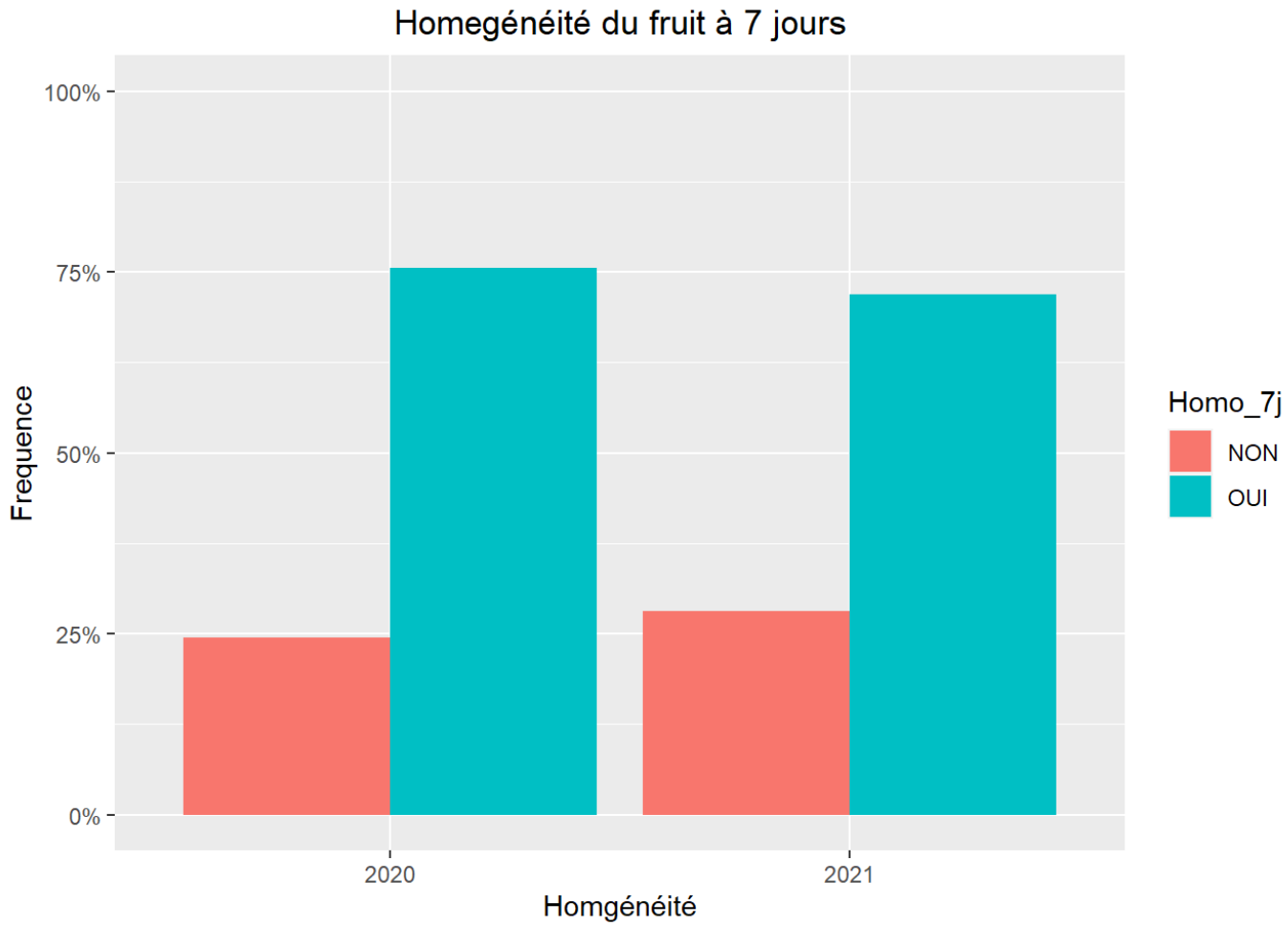
```
##            Abscence % Presence %
## Année 2020      95.92       4.08
## Année 2021      90.62       9.38
```

```
#multi=Tachepr[2,2]/Tachepr[1,2] #calcul du  quotient de proportion entre les taches 2020 et
 2021
#multi

freqTaches <- mydata %>%   group_by(Annee,Taches)  %>% summarise(n=n(),.groups = "drop_last")
%>% mutate(freq=n/sum(n))


ggplot(freqTaches,  aes(x=Annee, fill = Taches
                , group = Taches)) + geom_bar(aes(y=freq), stat="identity", position = "dodg
e")+ coord_cartesian(ylim=c(0,1))+
  ylab("Frequence")+
  xlab("Taches") +scale_y_continuous(labels = percent)+
  ggtitle("Présence ou absence des taches")+theme(plot.title = element_text(hjust = 0.5))
```



Une analyse comparative par année indique que plus de 90% des fruits n'avaient pas de tâches. La présence de tâches en 2021 est 2.30 fois supérieur à celui obtenue en 2020.

vi. Variable `Homo_7j` :



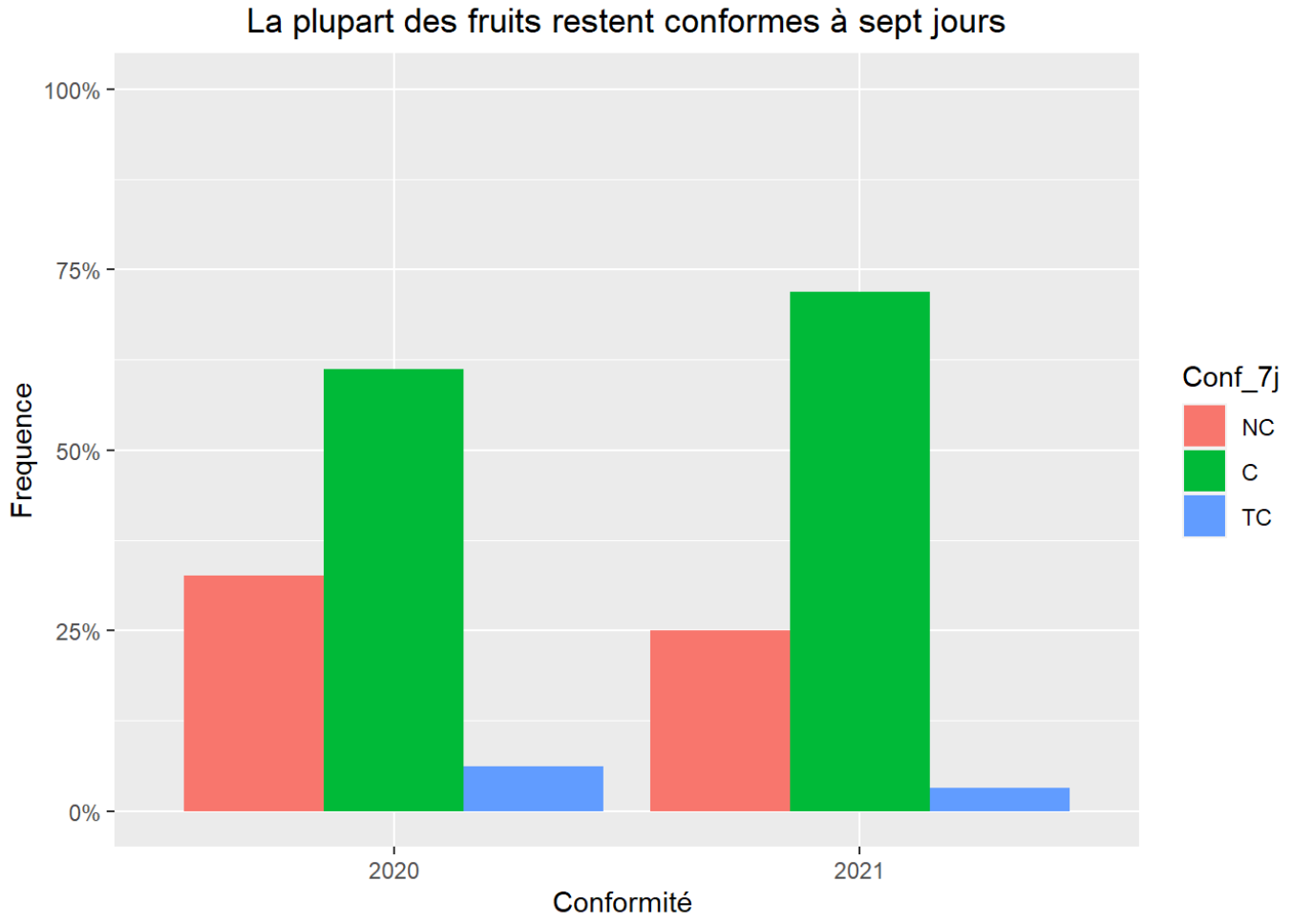
Le graphique nous montre que durant ces deux années la majorité du fruit sont homogène à 7 jours. Pendant la première année, le taux de fruits homogène était de 75.5% contre 24.49% de fruits non homogène. En 2021, on a une homogénéité de 71.8%.

vii. Variable `Conf_7j`

```
mydata$Conf_7j = factor(mydata$Conf_7j,c("NC","C","TC"))
pr2020=(table(mydata2020$Conf_7j)/length(mydata2020$Conf_7j))*100
pr2021=(table(mydata2021$Conf_7j)/length(mydata2021$Conf_7j))*100
Conf=round(rbind(pr2020,pr2021),2)
colnames(Conf)=c("Conforme en %", "Non Conforme en %" ,"Très Conforme en %")
rownames(Conf)=c("Année 2020","Année 2021")
Conf
```

```
##            Conforme en % Non Conforme en % Très Conforme en %
## Année 2020         61.22             32.65               6.12
## Année 2021         71.88             25.00               3.12
```

```
t=((pr2021[1]-pr2020[1])/pr2020[1])*100
Taux_evolution=round(t,2)

freqConf_7j <- mydata %>%   group_by(Annee,Conf_7j)  %>% summarise(n=n(),.groups = "drop_las
t")   %>% mutate(freq=n/sum(n))


ggplot(freqConf_7j,  aes(x=Annee, fill = Conf_7j
                 , group = Conf_7j)) + geom_bar(aes(y=freq), stat="identity", position = "dod
ge")+ coord_cartesian(ylim=c(0,1))+
  ylab("Frequence")+
  xlab("Conformité ") +scale_y_continuous(labels = percent)+
  ggtitle("La plupart des fruits restent conformes à sept jours")+theme(plot.title = element_
text(hjust = 0.5))
```



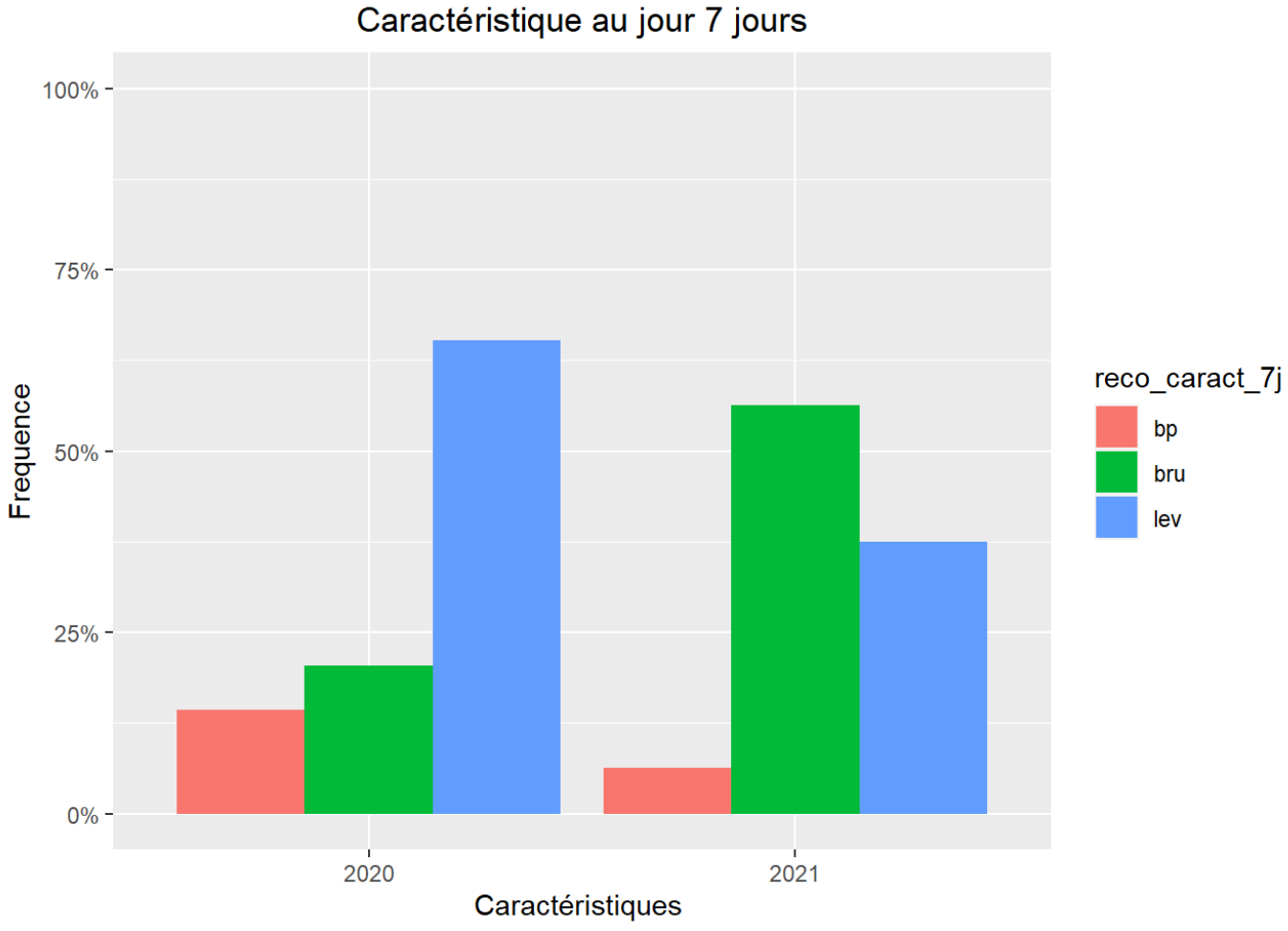
Sur les deux ans la majorité des fruits réceptionnés sont conformes. On constate une augmentation de la conformité des fruits de près de 17% en 2021. La proportion de fruits non conforme en 2020 reste supérieure à la proportion de fruit en 2021.

viii. Variable `carac_7j` :

```
pr2020=(table(mydata2020$reco_caract_7j)/length(mydata2020$reco_caract_7j))*100
pr2021=(table(mydata2021$reco_caract_7j)/length(mydata2021$reco_caract_7j))*100
vi=round(rbind(pr2020,pr2021),2)

Carac=vi[,-c(3:4)]
colnames(Carac)=c("bp en %", "bru en %","lev en %")
rownames(Carac)=c("Année 2020","Année 2021")
Carac
```

```
##            bp en % bru en % lev en %
## Année 2020   14.29    20.41    65.31
## Année 2021    6.25    56.25    37.50
```

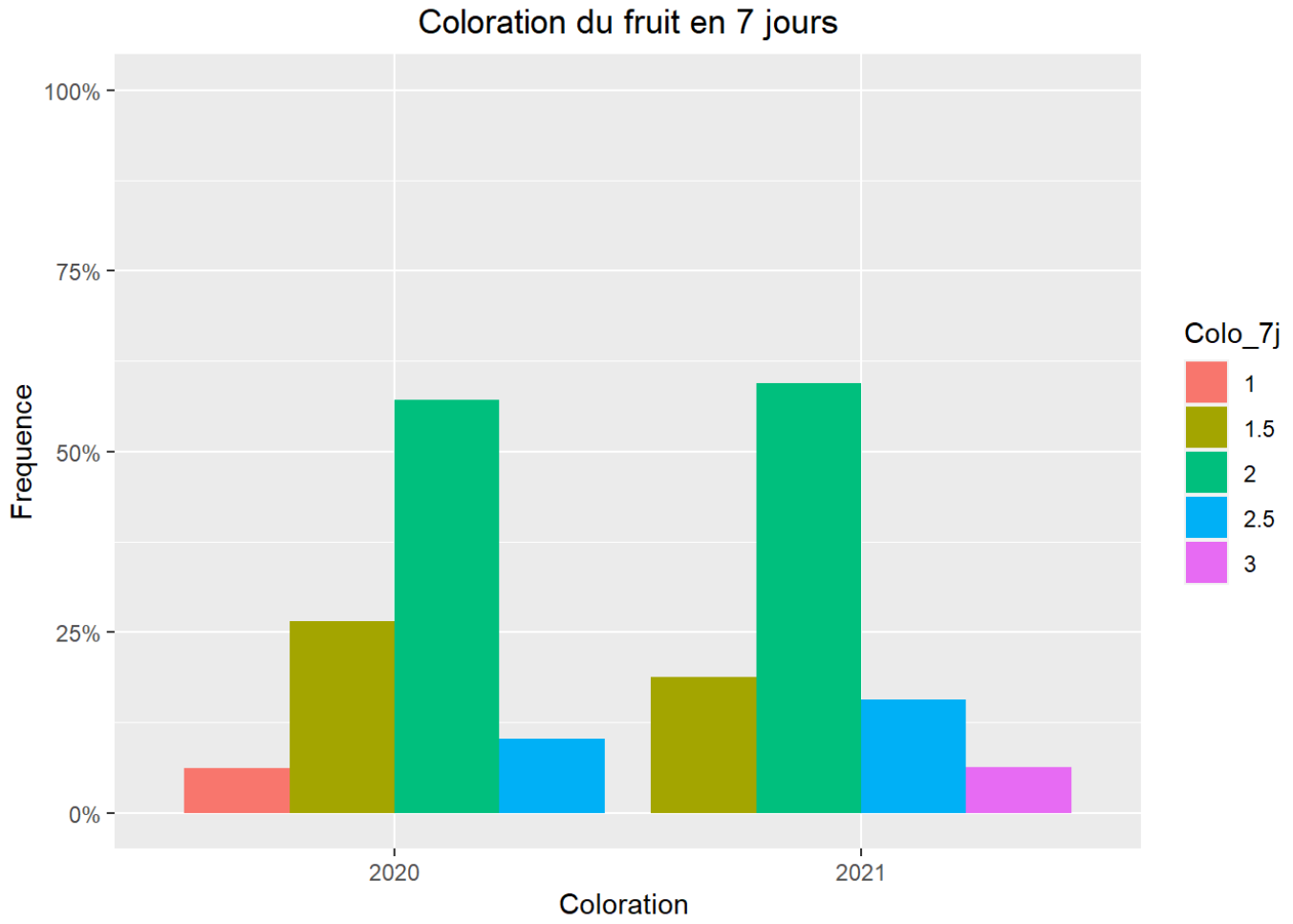
```
freqreco_caract_7j <- mydata %>%   group_by(Annee,reco_caract_7j)  %>% summarise(n=n(),.group
s = "drop_last")   %>% mutate(freq=n/sum(n))


ggplot(freqreco_caract_7j,  aes(x=Annee, fill = reco_caract_7j
                 , group = reco_caract_7j)) + geom_bar(aes(y=freq), stat="identity", position
= "dodge")+ coord_cartesian(ylim=c(0,1))+
  ylab("Frequence")+
  xlab("Caractéristiques ") +scale_y_continuous(labels = percent)+
  ggtitle("Caractéristique au jour 7 jours")+theme(plot.title = element_text(hjust = 0.5))
```



En 2020 au bout de sept jours près de 29% des fruits présentent des traces de levure. En 2021, le brunissement et la levure sont les plus présentes. On remarque une meilleure qualité de fruit en 2021.

ix. Variable `Colo_7j`

```
freqColo_7j <- mydata %>%   group_by(Annee,Colo_7j)  %>% summarise(n=n(),.groups = "drop_las
t")   %>% mutate(freq=n/sum(n))


ggplot(freqColo_7j,  aes(x=Annee, fill = Colo_7j
                 , group = Colo_7j)) + geom_bar(aes(y=freq), stat="identity", position = "dod
ge")+ coord_cartesian(ylim=c(0,1))+
  ylab("Frequence")+
  xlab("Coloration ") +scale_y_continuous(labels = percent)+
  ggtitle("Coloration du fruit en 7 jours")+theme(plot.title = element_text(hjust = 0.5))
```



On constate une absence de fruit ayant la coloration 3 en 2020 et une absence de fruit ayant la coloration 1 en 2021. La grande majorité des fruits avaient la coloration 2. La proportion de fruit prenant une coloration 2 évolue peu de 2020 à 2021.

c. Variables quantitatives continues

i. Variable `Hauteur` :

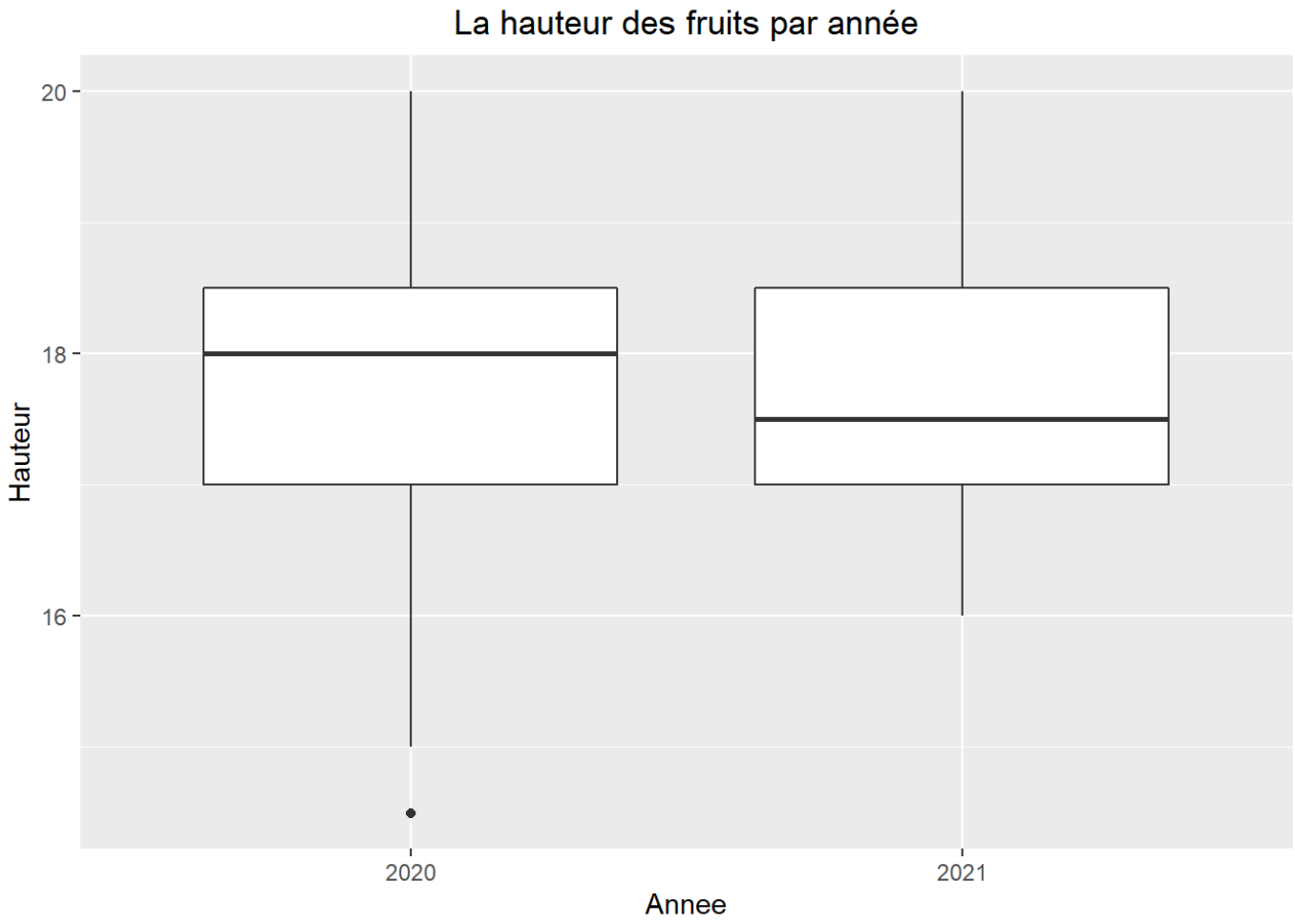
```
summary(mydata2020$Hauteur)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    14.5    17.0    18.0    17.6    18.5    20.0
```

```
summary(mydata2021$Hauteur)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   17.00   17.50   17.74   18.50   20.00
```

```
ggplot(mydata, aes(x=Annee, y=Hauteur)) +
  geom_boxplot()+ggtitle("La hauteur des fruits par année")+theme(
  plot.title = element_text(hjust = 0.5))
```



En 2020, un quart de la hauteur est inférieur à 17, la moitié de la hauteur est en dessous de 17.6 et les trois quart sont en dessous de 18.5. Le point en dessous de la boite à moustache représente une valeur atypique.

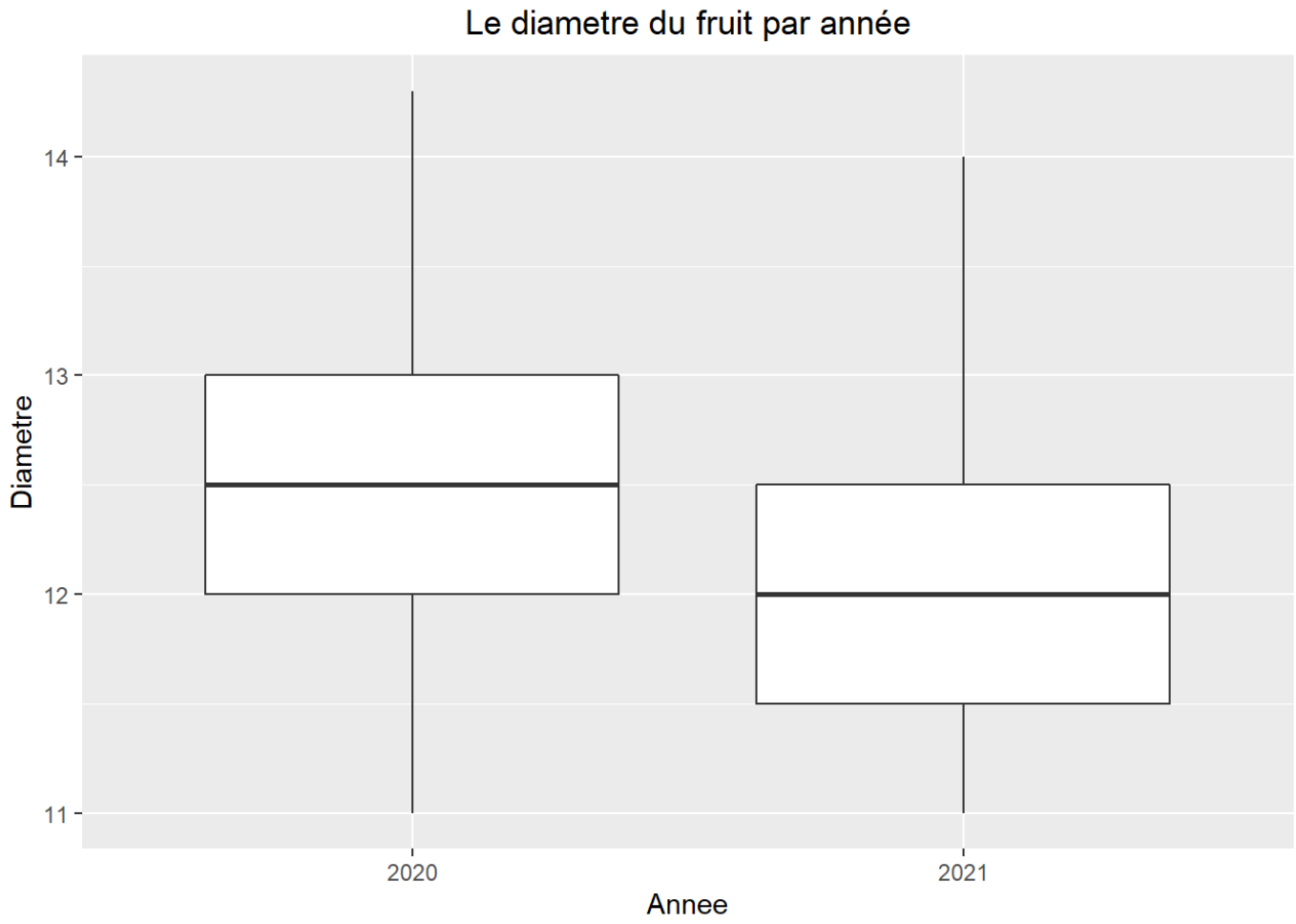
En 2021, un quart des de la hauteur est de valeur inférieure à 17, la moitié de la hauteur est en dessous de 17.74 et les trois quart sont en dessous de 18.5. On remarque que les deux boites à moustache n'ont pas la même allure avec la hauteur et le positionnement des diagrammes qui diffèrent et les dispersions des valeurs semblent légèrement différer.

Globalement, en 2020 la dispersion de la hauteur des fruits est plus grande pour celle inférieure au deuxième quartile 18, par contre en 2021 la dispersion de la hauteur des fruits est plus grande pour celle supérieur au deuxième quartile 17.5.

ii. Variable `Diametre`

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   12.00   12.50   12.53   13.00   14.30
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   11.50   12.00   12.06   12.50   14.00
```



Le diamètre des fruits en centimètres est compris entre 11 et 14.3. En 2020, 25% des diamètres ont des valeurs inférieures à 12, 50% des diamètres ont en dessous de 12.5 et les 75% des diamètres des fruits sont en dessous de 13.

En 2021, 25% des de la Diamètre ont de valeurs inférieures à 11.5, 50% des valeurs sont en dessous de 12 et les 75% des diamètres sont en dessous de 12.5. On remarque que les deux boites à moustaches n'ont pas la même allure avec le positionnement des diagrammes qui diffèrent.

Les dispersions des valeurs semblent différer. Notamment les médianes qui sont représentées par les traits horizontaux dans les boites, ne sont pas au meme niveau.

iii. Variable `Rendement` :

```
summary(mydata2020$Rendement)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2800  0.4300  0.4500  0.4478  0.4900  0.5500
```

```
summary(mydata2021$Rendement)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3500  0.4500  0.4700  0.4805  0.5000  0.6500
```

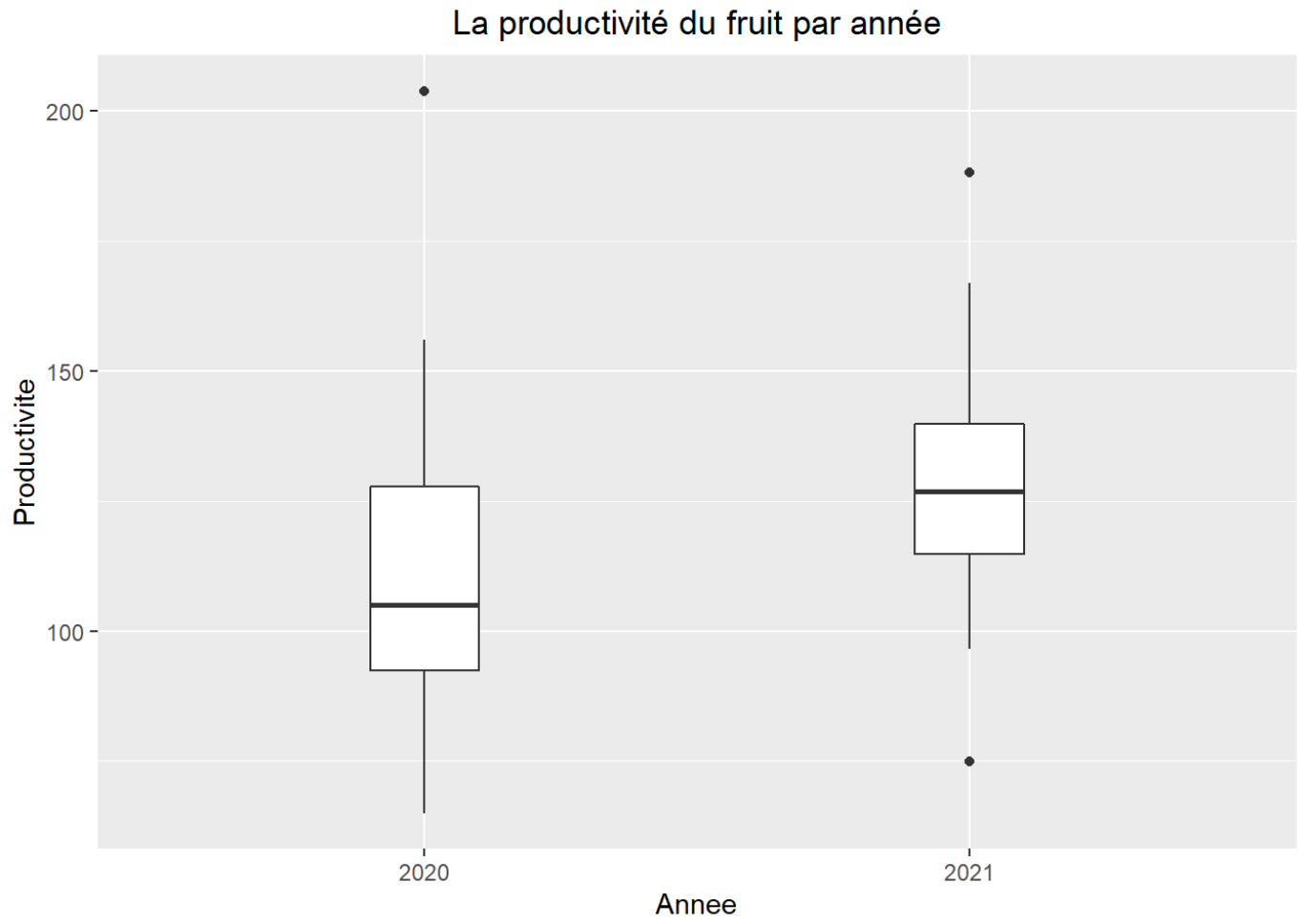
```
mydata1 = rbind(mydata2020,mydata2021)
ggplot(mydata1, aes(x=Annee, y=Rendement)) +
  geom_boxplot()+ggtitle("Rendement des fruits par année")+theme(
  plot.title = element_text(hjust = 0.5))
```

Le rendement varie entre 0.28 et 0.65. En 2021 la boite à moustache est très étalée alors que celle de 2020 est beaucoup plus resserrée. On peut également distinguer d'importantes différences sur les quartiles ainsi que la présence de valeurs aberrantes dans les deux groupes.

iv. Variable `Productivité`

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    65.0    92.4   105.0   110.8   127.8   203.9
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    74.8   114.9   126.8   129.2   139.8   188.2
```



En 2020, 25% des valeurs de la productivité sont inférieures à 92.4, 50% des valeurs sont en dessous de 105 et les 75% des productivités des fruits sont en dessous de 127.8.

En 2021, 25% des de la productivité est de valeur inférieure à 114.9, 50% des valeurs est en dessous de 126.8 et 75% de la productivité est en dessous de 139.8.

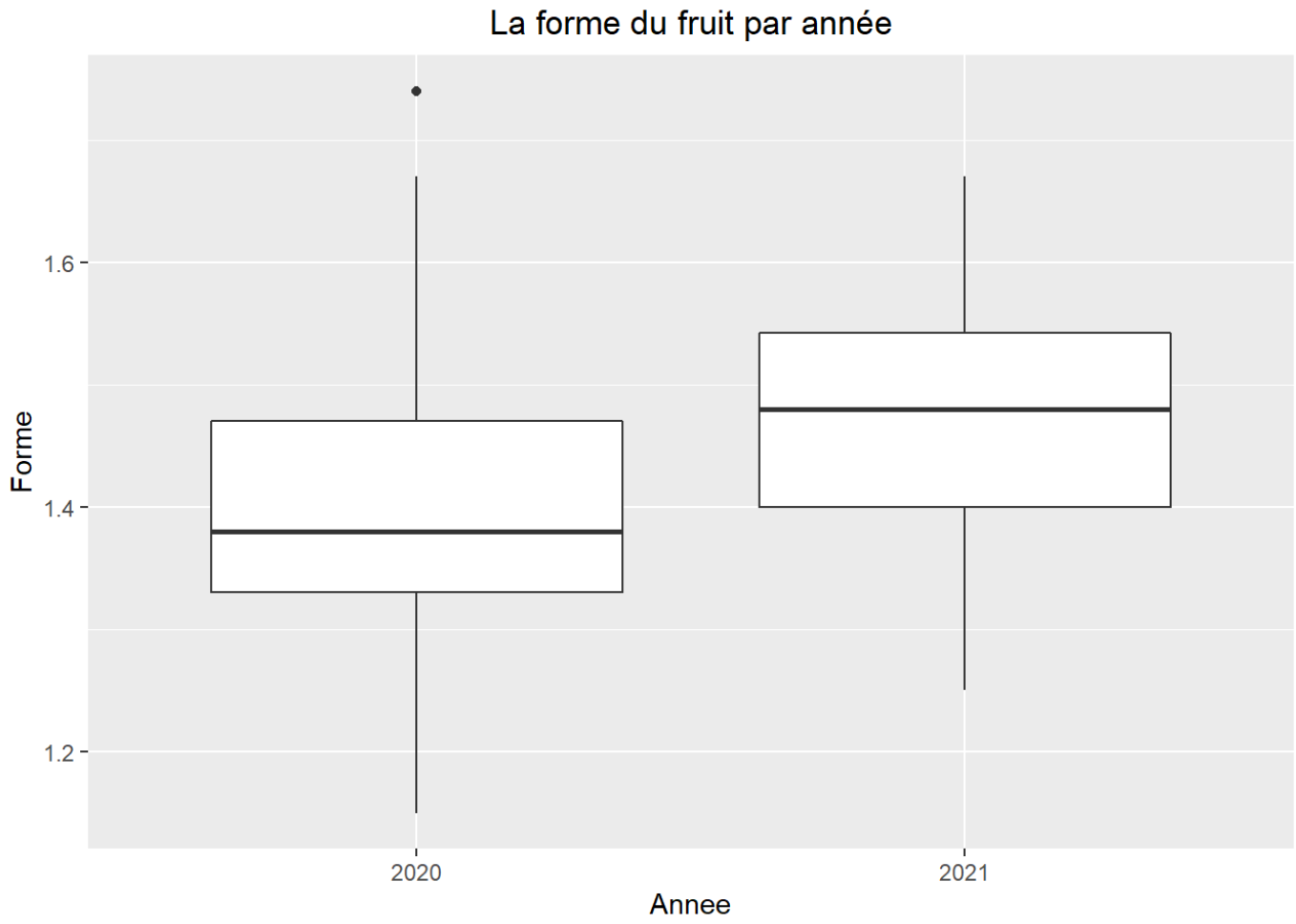
On remarque que les deux boites à moustache n'ont pas la même allure avec le positionnement des diagrammes qui diffèrent. Les dispersions des valeurs semblent différer. De plus les médianes qui sont représentées par les traits horizontaux dans les boites à moustache ne sont pas au meme niveau.

v. Variable `Forme`

La forme c'est rapport de la hauteur sur le diamètre. Elle dépend de la hauteur et du diamètre des fruits et elle varie entre 1.15 et 1.74.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.150   1.330   1.380   1.407   1.470   1.740
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.250   1.400   1.480   1.474   1.542   1.670
```



En 2020, 25% des valeurs de la Forme sont inférieures à 1.33, 50% des valeurs sont en dessous de 1.38 et les 75% de la forme des fruits sont en dessous de 1.74. En 2021, 25% des formes sont de valeurs inférieures à 1.4, 50% des valeurs sont en dessous de 1.48 et 75% des formes sont en dessous de 1.67.

Dans le graphique ci dessus, on remarque que le positionnement des deux boites à moustache diffèrent selon les années. De plus les médianes qui sont représentées par les traits horizontaux dans les boites à moustaches, ne sont pas au meme niveaux. On note la présence d'une valeur aberrante en 2020.

vi. Variable `Brix`

C'est le taux de sucre en pourcentage

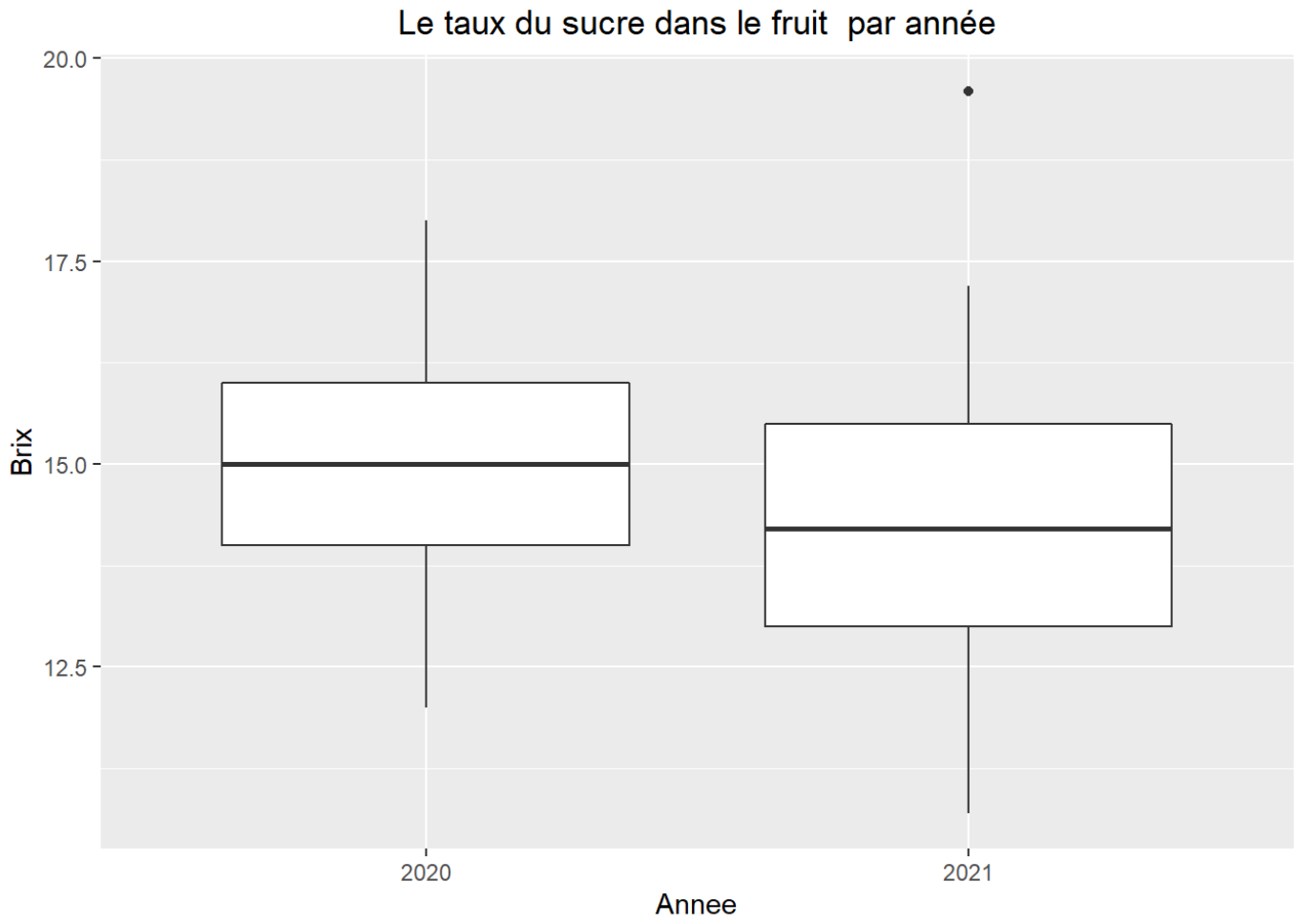
```
summary(mydata2020$Brix)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   14.00   15.00   14.74   16.00   18.00
```

```
summary(mydata2021$Brix)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.70   13.00   14.20   14.26   15.50   19.60
```

```
ggplot(mydata, aes(x=Annee, y=Brix)) +
  geom_boxplot()+ggtitle("Le taux du sucre dans le fruit  par année")+theme(
  plot.title = element_text(hjust = 0.5))
```



Le centre de la distribution du taux de sucre en 2021 est plus petit que la distribution en 2020. La distribution des données en 2021 est asymétrique car la portion en bas de la moustache est plus longue que du côté haut. On constate la présence de valeur aberrante en 2021.

d. Variables quantitatives discrètes

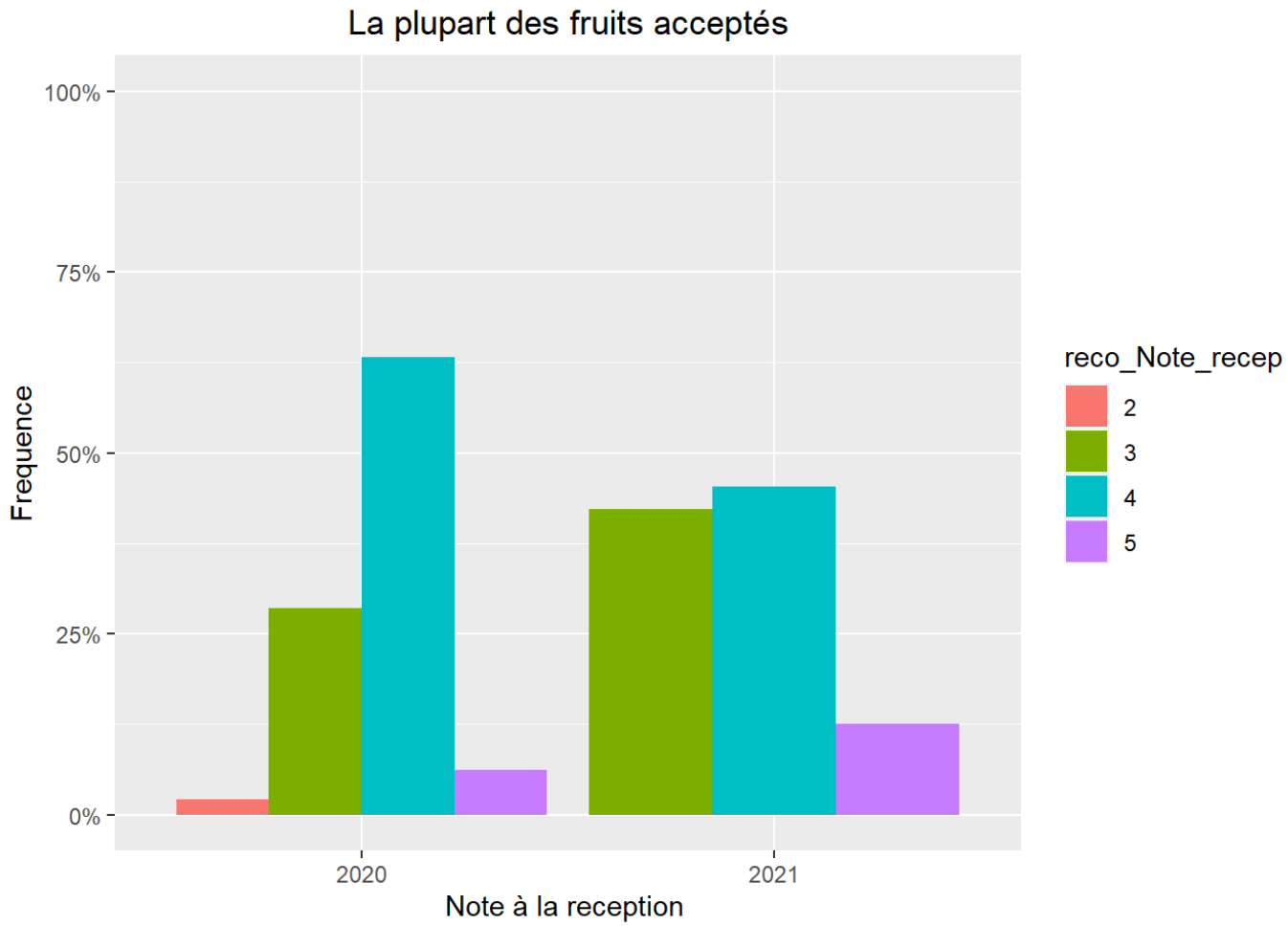
i. Variable `Note_recep` :

Cette variable représente la note à la réception de la palette. Elle varie entre 1 et 5. Elle permet de juger de la recevabilité du produit.

```
(table(mydata2020$reco_Note_recep)/length(mydata2020$reco_Note_recep))*100
```

```
##
##         2         3       3.5       3.8         4         5
##  2.040816 28.571429  0.000000  0.000000 63.265306  6.122449
```

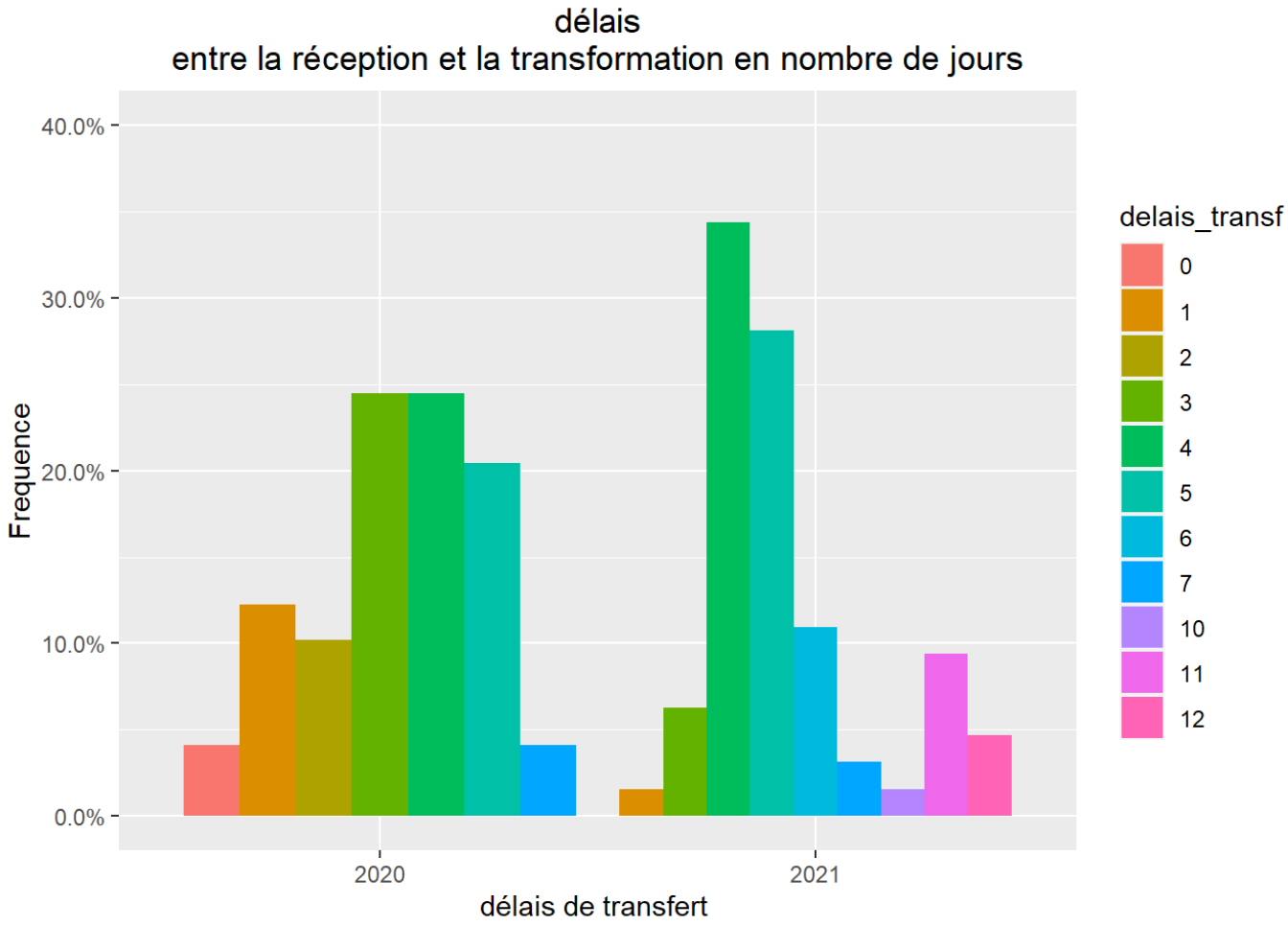
```
freqreco_Note_recep <- mydata %>%   group_by(Annee,reco_Note_recep)  %>% summarise(n=n(),.gro
ups = "drop_last")   %>% mutate(freq=n/sum(n))


ggplot(freqreco_Note_recep,  aes(x=Annee, fill = reco_Note_recep
                 , group = reco_Note_recep)) + geom_bar(aes(y=freq), stat="identity", positio
n = "dodge")+ coord_cartesian(ylim=c(0,1))+
  ylab("Frequence")+
  xlab("Note à la reception ") +scale_y_continuous(labels = percent)+
  ggtitle("La plupart des fruits acceptés")+theme(plot.title = element_text(hjust = 0.5))
```



Un produit n'est accepté si et seulement cette note est supérieure ou égale à 3. Tous les fruits réceptionnés en 2021 sont acceptables, près de 2% des fruits ont été jugé inacceptable en 2020.

ii. Variable `delais_transf` :

Elle représente le délai entre la réception et la transformation en nombre de jours.



Pour l'année 2020 : le nombre de jours max avant transformation est de 7 jours et la majorité les fruits sont transformés en 4-5 jours.

Pour l'année 2021 : le nombre de jours max avant transformation est de 12 jours et la majorité des fruits est transformée en 4-5 jours.

e. Variables quantitatives continues

III. Etude de la normalité et Tests de comparaison :

1. Etude de la normalité

un test de normalité Un test de shapiro-wilk qui nous permettra d'avoir une position plus tranchée sur la normalité des données. Si la p-value obtenue est supérieur à 0.05 on peut admettre que les données suivent une loi normale.

```
shapiro.test(mydata$Brix[mydata$Annee==2020])$p.value
```

```
## [1] 0.08420841
```

```
shapiro.test(mydata$Brix[mydata$Annee==2021])$p.value
```

```
## [1] 0.09875654
```

Ainsi on pour l'année 2020 une p-value = 0.08420841. Comme p-value est > a 0.05, on peut admettre que le taux de sucre pour l'année 2020 suit une loi normale. De façon similaire le taux de sucre pour l'année 2021 suit une loi normale car on a une p-value=0.09875654 qui est supérieur à 0.05. L'hypothèse de normalité étant

vérifiée, on peut comparer les distributions 2020 et 2021 avec le test de Student.

```
shapiro.test(mydata$Hauteur[mydata$Annee==2020])$p.value
```

```
## [1] 0.2554412
```

```
shapiro.test(mydata$Hauteur[mydata$Annee==2021])$p.value
```

```
## [1] 0.0004889163
```

Ainsi on pour l'année 2020 une p-value = 0.2554412. Comme p-value est > a 0.05, on peut admettre que le taux de sucre pour l'année 2020 suit une loi normale. Cependant l'hypothèse de normalité n'est pas vérifiée pour l'année 2021. Un test de Mann-Whitney est préconisé pour comparer ses deux distributions.

```
shapiro.test(mydata$Diametre[mydata$Annee==2020])$p.value
```

```
## [1] 0.02046019
```

```
shapiro.test(mydata$Diametre[mydata$Annee==2021])$p.value
```

```
## [1] 0.0002104291
```

Nous avons les mêmes conclusions que pour la Hauteur.

```
shapiro.test(mydata$Age_recep[mydata$Annee==2020])$p.value
```

```
## [1] 0.008720714
```

```
shapiro.test(mydata$Age_recep[mydata$Annee==2021])$p.value
```

```
## [1] 1.457076e-07
```

Comme p-value est inférieur à 0.05 pour les deux échantillons on poursuivra avec un test de Mann-Whitney pour comparer la distribution des données.

```
shapiro.test(mydata$Rendement[mydata$Annee==2020])$p.value
```

```
## [1] 0.0001204789
```

```
shapiro.test(mydata$Rendement[mydata$Annee==2021])$p.value
```

```
## [1] 3.365752e-05
```

Comme p-value est inférieure à 0.05 pour les deux échantillons on poursuivra avec un test de Mann-Whitney pour comparer la distribution des données.

2. Test de comparaison de moyennes pour certaines variables quantitatives

Sous l'hypothèse de normalité, nous utiliserons le test de student pour la comparaison des moyennes. En dehors de cette condition, nous ferons un test non paramétrique comme le test de Mann-Whitney. Le test U de Mann-Whitney (aussi appelé test de la somme des rangs de Wilcoxon ou plus simplement test de Wilcoxon) sert à tester l'hypothèse selon laquelle la distribution des données est la même pour deux groupes.

Lorsque la p-value est supérieur a 5%, on ne peut pas prétendre à une différence significative entre les 2 années.

```
b = c(t.test(mydata$Brix[mydata$Annee==2020],mydata$Brix[mydata$Annee==2021])$p.value,
      wilcox.test(mydata$Hauteur[mydata$Annee==2020],mydata$Hauteur[mydata$Annee==2021])$p.va
lue,
      wilcox.test(mydata$Diametre[mydata$Annee==2020],mydata$Diametre[mydata$Annee==2021])$p.
value,
      wilcox.test(mydata$Age_recep[mydata$Annee==2020],mydata$Age_recep[mydata$Annee==2021])
$p.value,
      wilcox.test(mydata$Rendement[mydata$Annee==2020],mydata$Rendement[mydata$Annee==2021])
$p.value)

b2=as.numeric(b)
a =c("Brix", "Hauteur", "Diametre", "Age_recep", "Rendement")


ddd=data.frame(a,round(b,4))
colnames(ddd)=c("Variables","P-value")
ddd
```

```
##   Variables P-value
## 1      Brix  0.1310
## 2   Hauteur  0.7868
## 3  Diametre  0.0003
## 4 Age_recep  0.0217
## 5 Rendement  0.1018
```

On peut donc affirmer avec un faible risque de se tromper que le `Brix` , la `Hauteur` et le `Rendement` qu'il n'y a pas une différence significative entre les moyennes données 2021. Pour les autres variables, la p-value est inférieure à 5%. On dira qu'il y'a une différence significative entre les moyennes données 2020 et 2021.

3. Comparaison par année 2020 et 2021 pour les variables qualitatives (ou quanti discretes)

i. Variable Discrete `Note_Recep` :

Afin d'avoir un résultat significatif, nous allons recoder les données comme suit : - Pour 2020 : Les notes 2 et 5 seront affectées aux notes 3 et 4 respectivement - Pour 2021 : Les notes 3.5, 3.8 et 5 seront affectées aux notes 3, 4 et 4 respectivement.

```
##
##  3  4
## 15 34
```

```
##
##  3  4
## 27 37
```

```
##
##  Chi-squared test for given probabilities
##
## data:  TableauContegence2021
## X-squared = 4.5268, df = 1, p-value = 0.03337
```

```
##
##  3  4
## 27 37
```

```
##    3    4
## 19.2 44.8
```

Le resultat du test de chi-deux nous donne une p-value inférieur a 5% (p-value = 0.03337), donc on peut affirmer avec un faible risque de se tromper qu'il y'a une différence significative entre les distributions de 2020 et 2021 concernant la Note à la réception.

ii. Variable Qualitative : `Taches`

```
table(mydata2020$Taches);
```

```
##
##  A  P
## 47  2
```

```
table(mydata2021$Taches)
```

```
##
##  A  P
## 58  6
```

Il faut un effectif supérieur 5 pour que le test détecte des différences pratiques avec une probabilité élevée.

On ne peut donc pas trancher quant à la différence des distributions des Taches entre 2020 et 2021.

iii. Variable Qualitative : `Conf_7j`

```
mydata2020$reco_Conf_7j<-mydata2020$Conf_7j
mydata2020$reco_Conf_7j[mydata2020$reco_Conf_7j=="TC"]<-"C"


mydata2021$reco_Conf_7j<-mydata2021$Conf_7j
mydata2021$reco_Conf_7j[mydata2021$reco_Conf_7j=="TC"]<-"C"

mydata2020$reco_Conf_7j=as.character(mydata2020$reco_Conf_7j)
mydata2021$reco_Conf_7j=as.character(mydata2021$reco_Conf_7j)

table(mydata2020$reco_Conf_7j);table(mydata2021$reco_Conf_7j)
```

```
##
##  C NC
## 33 16
```

```
##
##  C NC
## 48 16
```

```
resConf <- chisq.test(x=table(mydata$Taches), p=c(0.67,0.33))
resConf
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(mydata$Taches)
## X-squared = 34.338, df = 1, p-value = 4.633e-09
```

Le test de Chi-deux renvoie une p-value très petite. Il y a une différence significative entre les distributions de Conformité à 7 jours de 2020 et 2021.

IV. Etude de la Liaison entre les variables qualitatives

Il y'a beaucoup de données manquantes pour 2020, nous travaillerons avec les données de 2021. Nous allons créer une base contenant toutes les variables qualitatives

```
qlNA<-select(mydata2021,Colo,Taches,reco_Com_rec,Conf_7j,Colo_7j,Homo_7j,Carac_7j,Conf_8j,Colo_8j,Homo_8j,Carac_8j)
```

Dans cette section, on cherche à savoir comment la `couleur des fruits` et les `taches` sur ces derniers influent sur le `commentaire à la reception` ou la `note à la reception` . Ensuite on cherchera à voir comment l' `homogeneité` et le `caracteristique à 7j` influent sur le commentaire à 7j.

Pour cela nous procéderons comme suit : - visualiser les fréquences absolues(à travers un tableau de contingence par exemple) - tester le lien entre les 2 variables (liaison).

Pour tester le lien entre deux variables qualitatives on utilise le test du Khi-2 (test paramétrique), ou, si les conditions du diagnostic de régression ne sont pas remplies, on utilise le test exact de Fisher (test non-paramétrique). Ces deux tests sont pratique pour comparer des pourcentages.

Condition de validité du Khi-2 - Les valeurs de toutes les cases du tableau des effectifs doivent être supérieur à 5 - Les pourcentages ne sont pas trop proche de zéro ou de 100%

i. Liaison entre la coloration et le commentaire à la reception

Regardons le tableau de contingence des deux variables après avoir recoder afin de remplir les conditions pour un futur test de khi-deux

```
#recodage avant le test de khi*deux pour avoir des classes avec n>5
mydata2021$reco_Colo2<-mydata2021$Colo
mydata2021$reco_Colo2[mydata2021$reco_Colo2=="1.5"]<-"2"
mydata2021$reco_Colo2[mydata2021$reco_Colo2=="3"]<-"2.5"

mydata2021$reco_Colo2=as.character(mydata2021$reco_Colo2)


Tc1<-table(mydata2021$reco_Colo2,qlNA$reco_Com_rec, deparse.level=2)

ColoCom_rec<-prop.table(Tc1)*100
lprop(ColoCom_rec)
```

```
##                      qlNA$reco_Com_rec
## mydata2021$reco_Colo2 A     E     D     Total
##              2         43.1  31.4  25.5 100.0
##              2.5        7.7  46.2  46.2 100.0
##              Ensemble  35.9  34.4  29.7 100.0
```

`Com_recep` : commentaire à la réception - A : acide, - D : doux, - E : équilibré

`Colo` : coloration de la chair à réception échelle 1/1,5/2/2,5/3

En 2021, Nous constatons que la majorité des fruits sont `acide` et appartiennent majoritairement à la coloration 2.

```
mydata2021$reco_Colo2=as.character(mydata2021$reco_Colo2)
qlNA$reco_Com_rec=as.character(qlNA$reco_Com_rec)
qlNA$reco_Com_rec
```

```
##  [1] "D" "D" "E" "A" "A" "A" "A" "A" "A" "A" "E" "E" "A" "A" "A" "A" "D" "D" "D"
## [20] "A" "A" "A" "A" "E" "E" "E" "E" "E" "E" "D" "D" "D" "E" "D" "E" "E" "E" "D"
## [39] "D" "D" "E" "E" "A" "A" "E" "E" "A" "A" "A" "A" "D" "D" "E" "D" "E" "E" "E"
## [58] "E" "A" "A" "D" "D" "D" "D"
```

```
chisq.test(mydata2021$reco_Colo2,qlNA$reco_Com_rec)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mydata2021$reco_Colo2 and qlNA$reco_Com_rec
## X-squared = 5.7699, df = 2, p-value = 0.05586
```

La p-value du test est supérieur à 5. on peut donc admettre avec un faible risque de se tromper que les deux variables Taches et Commentaires sont liées.

ii. Liaison entre les taches et les commentaires à la reception

```
##            qlNA$reco_Com_rec
## qlNA$Taches A     D     E     Total
##    A         36.2  29.3  34.5 100.0
##    P         33.3  33.3  33.3 100.0
##    Ensemble  35.9  29.7  34.4 100.0
```

```
##
##  Pearson's Chi-squared test
##
## data:  qlNA$Taches and qlNA$reco_Com_rec
## X-squared = 0.044227, df = 2, p-value = 0.9781
```

La p-value du test étant 0.9781, on peut donc admettre avec un faible risque de se tromper que les deux variables Taches et Commentaires sont liées.

Nous constatons que moins il y' a de `Taches` , plus le fruit a tendance à être noté `Doux` grace au tableau de contingence.

iii. Liaison entre l'Homogénéité et les commentaires à la reception

```
##             qlNA$reco_Com_rec
## qlNA$Homo_7j A     D     E     Total
##     NON       38.9  22.2  38.9 100.0
##     OUI       34.8  32.6  32.6 100.0
##     Ensemble  35.9  29.7  34.4 100.0
```

```
##
##  Pearson's Chi-squared test
##
## data:  qlNA$Homo_7j and qlNA$reco_Com_rec
## X-squared = 0.67927, df = 2, p-value = 0.712
```

La p-value du test étant 0.712, on peut donc admettre avec un faible risque de se tromper que les deux variables Taches et Commentaires sont liées.

iv. Liaison entre du caractère à 7 jours et les commentaires à la reception

```
##              qlNA$reco_Com_rec
## qlNA$Carac_7j A     D     E     Total
##      bp        25.0   0.0  75.0 100.0
##      bru       44.4  33.3  22.2 100.0
##      fer       37.5  18.8  43.8 100.0
##      lb        22.2  33.3  44.4 100.0
##      lev       50.0  50.0   0.0 100.0
##      Ensemble  35.9  29.7  34.4 100.0
```

```
##
##  Pearson's Chi-squared test
##
## data:  qlNA$Carac_7j and qlNA$reco_Com_rec
## X-squared = 11.378, df = 8, p-value = 0.1812
```

La p-value du test étant 0.1812, on peut donc admettre que les deux variables Taches et Commentaires sont corrélées. Cependant, la significativitè de lien est ici plutôt faible.

4. Etude de la Correlation entre les variables quantitatives

Pour l'étude des liaisons entre variables quantitatives, nous utiliserons le coefficient de corrélation de pearson ainsi que le niveau de significativité pour toutes les paires possibles de variables quantitatifs ainsi que le niveau de significativité qui est le résultat des tests de corrélation entre chaque paire de variables.

Les hypothèses du test sont les suivants:

H0 : “les caractères X et Y sont indépendants” contre H1 : “les caractères X et Y ne sont pas indépendants”.

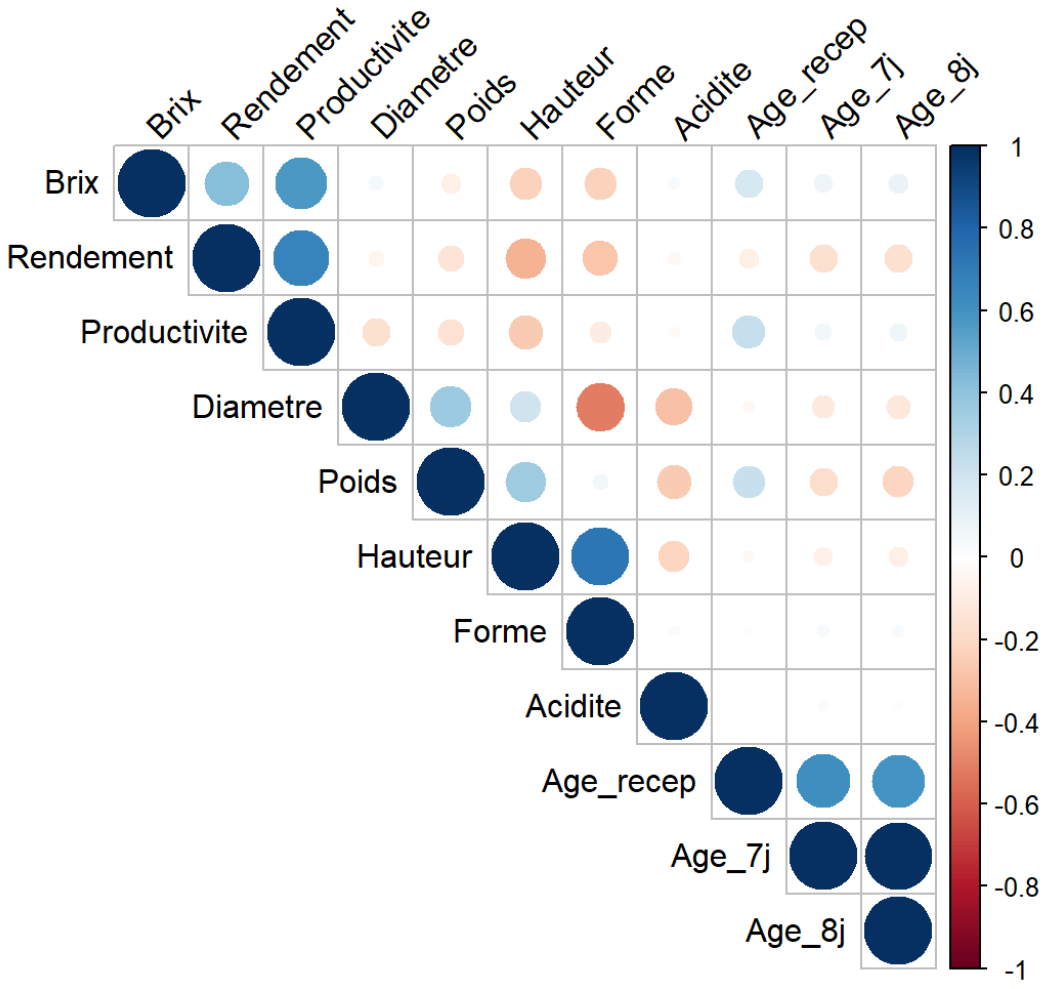
Si la p-valeur obtenue du test de corrélation est inférieur à 0.05 on rejette H0 et ce réjet est hautement significatif. Ainsi on peut dire que X et Y on une dépendance.

```
#création d'une base avec des données quantitatives
qtNA<-select(mydata2021, Rendement,Brix,Acidite,Hauteur,Diametre,Forme,Poids,Productivite,Age
_7j,Age_8j,Age_recep)
#calcul du coefficient de corrélation
mcor<-cor(qtNA)
#mcor
#test de significativité de la correlation(p-value)
a=rcorr(as.matrix(qtNA),type="pearson")
```

i. Diagramme de corrélation : Corrélogramme

Dans le graphique ci dessous les corrélations positives sont affichées en bleu et les corrélations négatives en rouge. L’intensité de la couleur et la taille des cercles sont proportionnelles aux coefficients de corrélation. A droite du correllèlogramme, la légende de couleurs montre les coefficients de corrélation et les couleurs correspondantes.

```
#utiliser ggplot après
corrplot(mcor, type="upper", order="hclust", tl.col="black",tl.srt=45,insig="blank")
```




```
#corrplot(mcor, type="upper", order="hclust", p.mat = a$P, sig.level = 0.05)
```

L'étude des corrélations de `Rendement` avec les autres variables nous indiquent l'existence d'une correlation faible et positive avec `Brix` et `Productivite` . Cette corrélation est hautement significative entre `Rendement` avec les variables `Brix` et `Productivite` .

La corrélation entre `Rendement` avec `Hauteur` et `Forme` est négative et faible. D'après la probabilité de test de corrélation l'existence d'un lien linéaire du `Rendement` est très significatif avec `Hauteur` , et significatif avec `Forme` .

Une liaison linéaire hautement significative est aussi observée entre `Brix` et les variables `Rendement` , `Productivite` .

quant à `Acidite` , elle est significativement corrélé au `Diametre` , `Poids` .

L'analyse de la corrélation entre `Hauteur` avec les autres variables montre l'existence d'une corrélation hautement significative entre `Hauteur` et `Forme` , Très significative entre `Hauteur` avec les variables `Rendement` , `Poids` . cette corrélation est significative entre `Hauteur` et `Productivite` .

La Corrélation entre `Diamètre` est négative et hautement significative avec `Forme` ,positive et très significative avec `Poids` , négative et significative avec `Acidite` .

La corrélation entre `Forme` et les autres variables indiquent Une corrélation hautement significative avec `Hauteur` et `Diametre` cela est normale car la forme est le rapport de la hauteur sur le diamètre. Cette corrélation significative avec `Rendement` .

`Poids` quant à lui est faiblement et positivement corrélé avec `Hauteur` et `Diamètre` , sa corrélation est négative avec `Acidité` et `delais_transf` . Cette corrélation est significative avec `Acidite` et `delais_transf` , très significative avec `Hauteur` et `Diamètre` .

De même `Productivité` est corrélé positivement avec le `Rendement` et `Brix` , par contre cette corrélation est négative avec `Hauteur` . La corrélation est hautement significative entre `Productivite` avec `Rendement` , `Brix` , elle est significative avec `Hauteur` .

Certaines liaisons parmis nos variables sont liées à une dépendance temporelle. Par exemple la corrélation entre `Age_recep` , `Age_7j` et `Age_8j` . En effet `Age_7j` est égale à `Age_recep` plus sept jours et `Age_8j` est égale à `Age_7j` plus 1 jour cela explique la forte liaison de ces variables deux à deux.

Deux variables qui ont un coefficient de corrélation très élevés sont des variables qui on tendance à apporter la même information.

Pour résoudre ce problème nous allons construire des indicateurs qui résument beaucoup de variables, autrement dit à partir de l'ensemble des variables, nous chercherons à avoir une vision de l'ensemble des liaisons sans passer en revue chaque couple de variables.

L'analyse en composantes principales(ACP) est la méthode qui sera utilisée pour extraire et visualiser les informations importantes contenues dans une table de données multivariées. L'ACP synthétise cette information en seulement quelques nouvelles variables appelées composantes principales.

Ces nouvelles variables correspondent à une combinaison linéaire des variables originels. Le nombre de composantes principales est inférieur ou égal au nombre de variables d'origine.

V. Analyse En Composante Principale et Typologie de fruit

L'objectif de cette partie du document est de considérer simultanément les caractéristiques des fruits afin de mettre en évidence:

Premièrement le profil des fruits qui se ressemble qui s'oppose, déterminer les principales dimensions de variabilité des fruits et déterminer les typologies de fruits qui ressemblent ou qui s'opposent.

Deuxièmement il s'agira de déterminer un bilan des liaisons entre les différentes variables.

Troisièmement associer à un groupe de fruit qui se ressemblent un profil particulier.

Pour la mise en oeuvre de l'ACP nous allons considérer les variables `Diamètre` , `Poids` , `Hauteur` , `Forme` , `Acidite` , et `Brix` comme variables actives car ces variables sont des caractéristiques quantitatives qui définissent mieux le profil de chaque fruit à partir d'elles d'autres variables telles que le rendement et la productivité ont été déterminés. Dans notre ACP nous considérerons comme variables quantitatives supplémentaires `Age_recep` , `Delais_transf` , `Age_7j` , `Note_8j` , `Note_7j` , `Age_8j` , `Rendement` et `Poductivite` .

```
ACPdata<-select(mydata2021, Brix,Acidite,Hauteur,Diametre,Forme,Poids,Productivite,Rendement,
delais_transf,Age_7j,Note_7j,Age_8j,Age_recep,Note_8j,Colo,Conf_7j, Com_rec,Note_recep)
res.pca=PCA(ACPdata,scale.unit=TRUE,quanti.sup = 7:14,quali.sup = 15:18, graph = FALSE)
#resultats des variables et individus pour l'acp
res.var<-get_pca_var(res.pca)
res.ind<-get_pca_ind(res.pca)
```

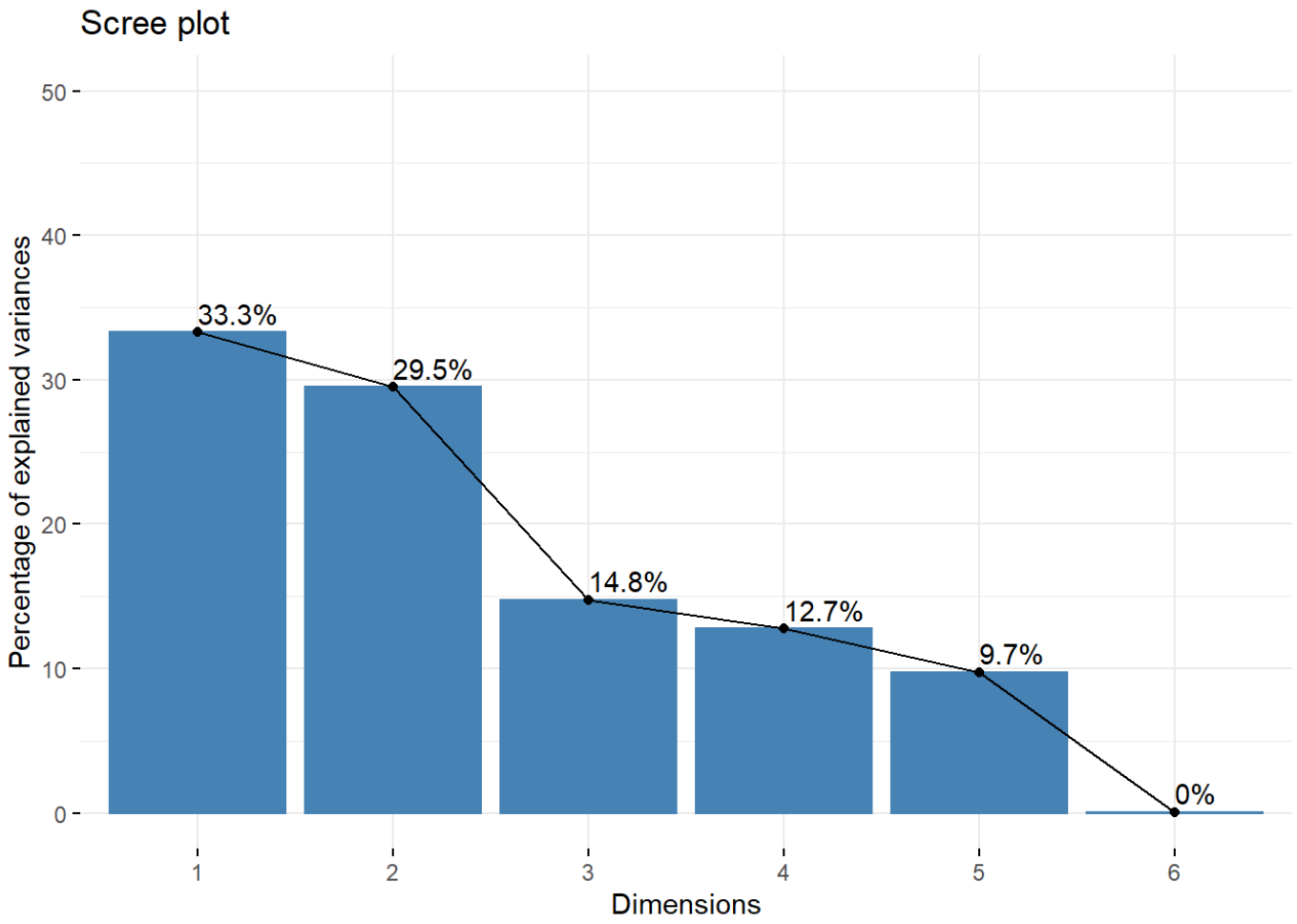
Choix du nombre de dimensions

Pour déterminer le nombre de dimension à retenir nous allons utilisés la règle de Kaiser. Selon cette règle une valeur propre supérieur à 1 indique une composante qui représente plus de variance par rapport à une seule variable.

```
#valeurs propres
val_propr=get_eigenvalue(res.pca)
val_propr
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1 1.997082465      33.28470775                    33.28471
## Dim.2 1.768838726      29.48064543                    62.76535
## Dim.3 0.885289116      14.75481859                    77.52017
## Dim.4 0.764695117      12.74491862                    90.26509
## Dim.5 0.582483854       9.70806424                    99.97315
## Dim.6 0.001610722       0.02684537                   100.00000
```

```
#histogramme pour le coude
scree.plot=fviz_eig(res.pca,addlabels = T,ylim=c(0,50))
scree.plot
```



Ainsi on retiendra 2 dimensions cela est confirmé par le graphiques des valeurs propres qui nous indiques que les 2 premières dimensions expliques 62.77% des variations.

La première dimension exprime à elle seule 33.3% de la variabilité des données. La seconde dimension exprime 29.5% de l'inertie de la variabilité.

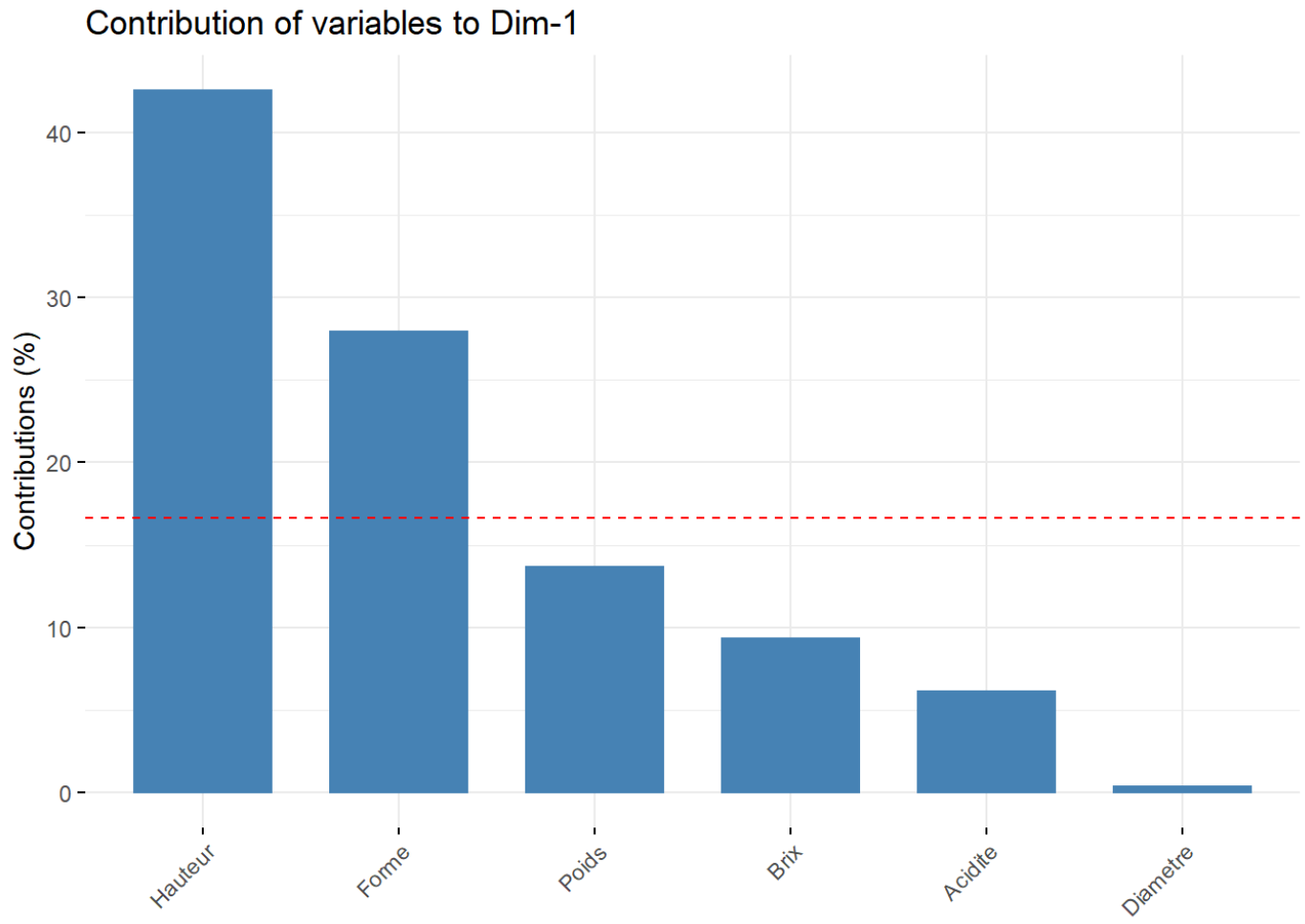
1. Etude de Variables

2. Contribution des variables aux deux principales dimensions

```
contrib_attend=(1/6)*100
res.var$contrib[,1:2]
```

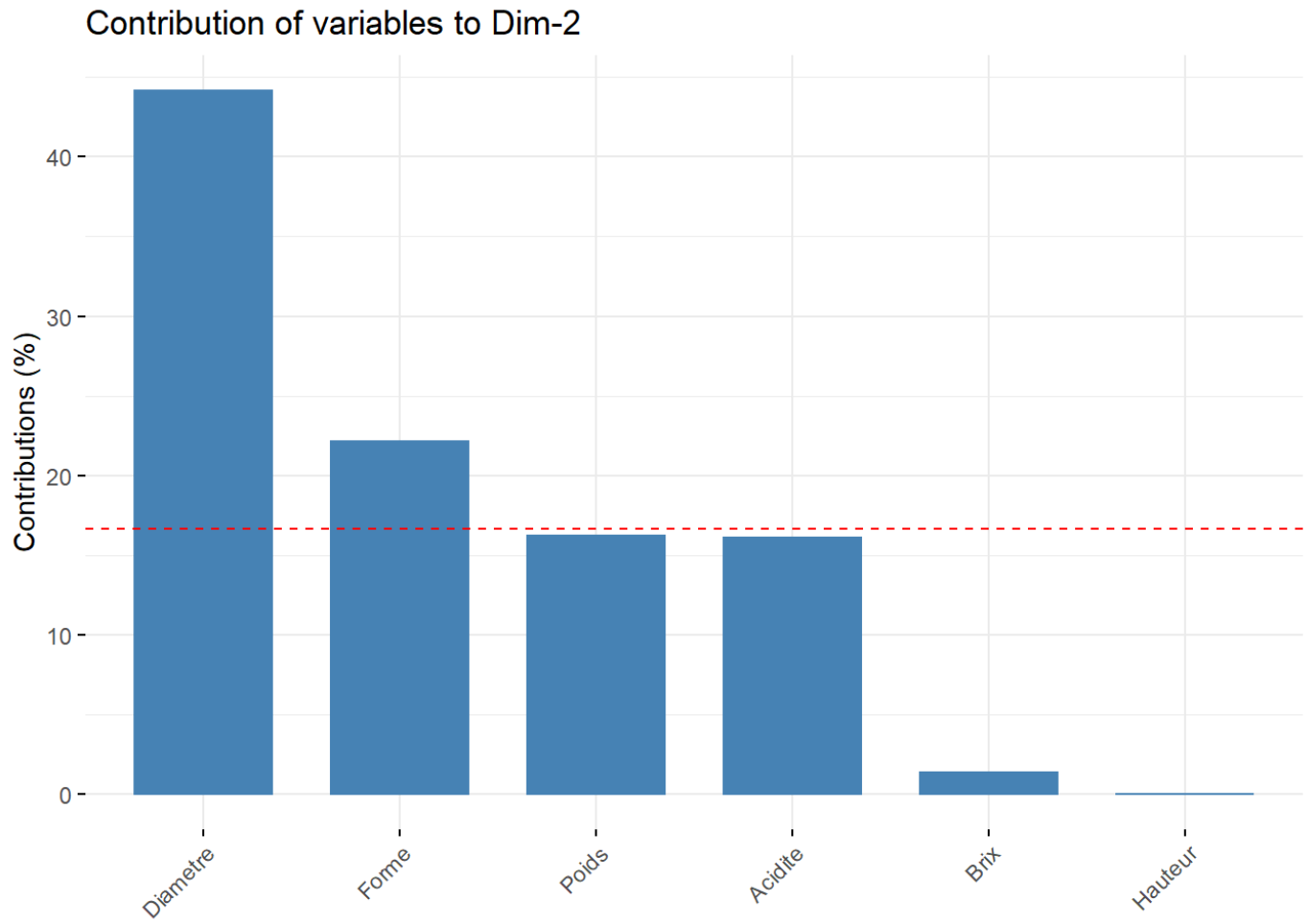
```
##               Dim.1        Dim.2
## Brix      9.3274763  1.367731648
## Acidite   6.1181238 16.105777781
## Hauteur  42.5845902  0.002559624
## Diametre  0.3556347 44.184259048
## Forme    27.9313112 22.149048208
## Poids    13.6828637 16.190623691
```

```
# Contributions des variables à PC1
fviz_contrib(res.pca, choice = "var", axes = 1)
```



Le graphe de la contribution des variables nous montre que les variables qui contribuent le plus à la première dimension sont `Hauteur` (42.5%) et `Forme` (27%). La contribution des autres variables tels que `Brix`, `Acidite`, `Diamètre` et `Poids` sont inférieur à la contribution moyenne attendue sur cette dimension(16.6%) représente par la ligne en pointillé rouge.

```
# Contributions des variables à PC2
fviz_contrib(res.pca, choice = "var", axes = 2)
```

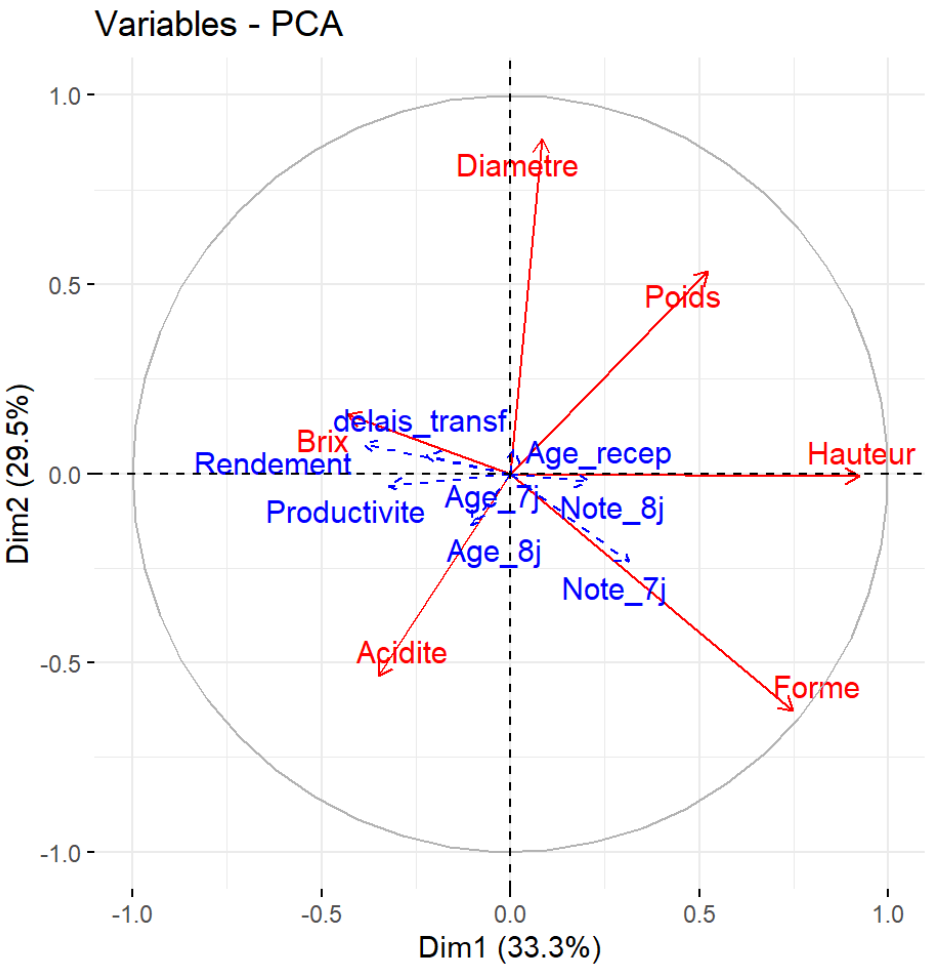


De même `Diamètre` (44.1%) et `Forme` (22.1%) contribuent plus à la deuxième dimension. `Acidite` (16.1%) et `Poids` (16.19%) on une contribution presque moyenne (16.6%). Quant à `Brix` (1.36%) et `Hauteur` (0.0025%) leurs contributions ne sont pas significatives.

3. Cercle de corrélation

Sur le cercle de corrélation, les variables en rouge sont les variables actives, celles en bleu sont les variables quantitatives supplémentaires.

```
#cercle de corrélation
fviz_pca_var(res.pca,col.var = "red",col.quanti.sup = "blue",repel = T)
```



A partir du graphique de corrélation des variables nous pouvons remarqués que les variables `Diametre` , `Hauteur` et `Forme` sont bien représentes sur les deux axes principaux.

La corrélation entre `Diamètre` et `Hauteur` est faible (r= 0.21), la corrélation entre `Diamètre` et `Forme` est négative (-0.51), la corrélation entre `Hauteur` et `Forme` est forte et positive (0.71).

En effet `Forme` est le rapport entre `Hauteur` et `Diamètre` ainsi une hauteur plus grande aura tendance à augmenter la forme du fruit et un grand diamètre divisera la forme du fruit cela entraînera donc une opposition de `Forme` et `Diamètre` sur la deuxième dimension.

Une description des dimensions est obtenue ci dessous

```
#pour voir les variables qui décrivent chaque dimension
res.dim<-dimdesc(res.pca,axes = 1:2)
res.dim$Dim.1
```

```
## $quanti
##              correlation      p.value
## Hauteur        0.9221981 2.860340e-27
## Forme          0.7468677 1.378347e-12
## Poids          0.5227409 9.375126e-06
## Note_7j        0.3140877 1.148799e-02
## Productivite  -0.3215898 9.563010e-03
## Acidite       -0.3495482 4.637753e-03
## Rendement     -0.3874196 1.562733e-03
## Brix          -0.4315986 3.696791e-04
##
## $quali
##                 R2    p.value
## Conf_7j 0.1078122 0.03082594
##
## $category
##               Estimate     p.value
## Conf_7j=C   0.4252633 0.012199422
## Conf_7j=NC -0.6516680 0.008262207
##
## attr(,"class")
## [1] "condes" "list"
```

```
res.dim$Dim.2
```

```
## $quanti
##          correlation      p.value
## Diametre   0.8840522 3.764066e-22
## Poids      0.5351505 5.215891e-06
## Acidite   -0.5337464 5.580173e-06
## Forme     -0.6259241 3.186789e-08
##
## $category
##             Estimate    p.value
## Colo=Colo_3 1.083056 0.02129057
##
## attr(,"class")
## [1] "condes" "list"
```

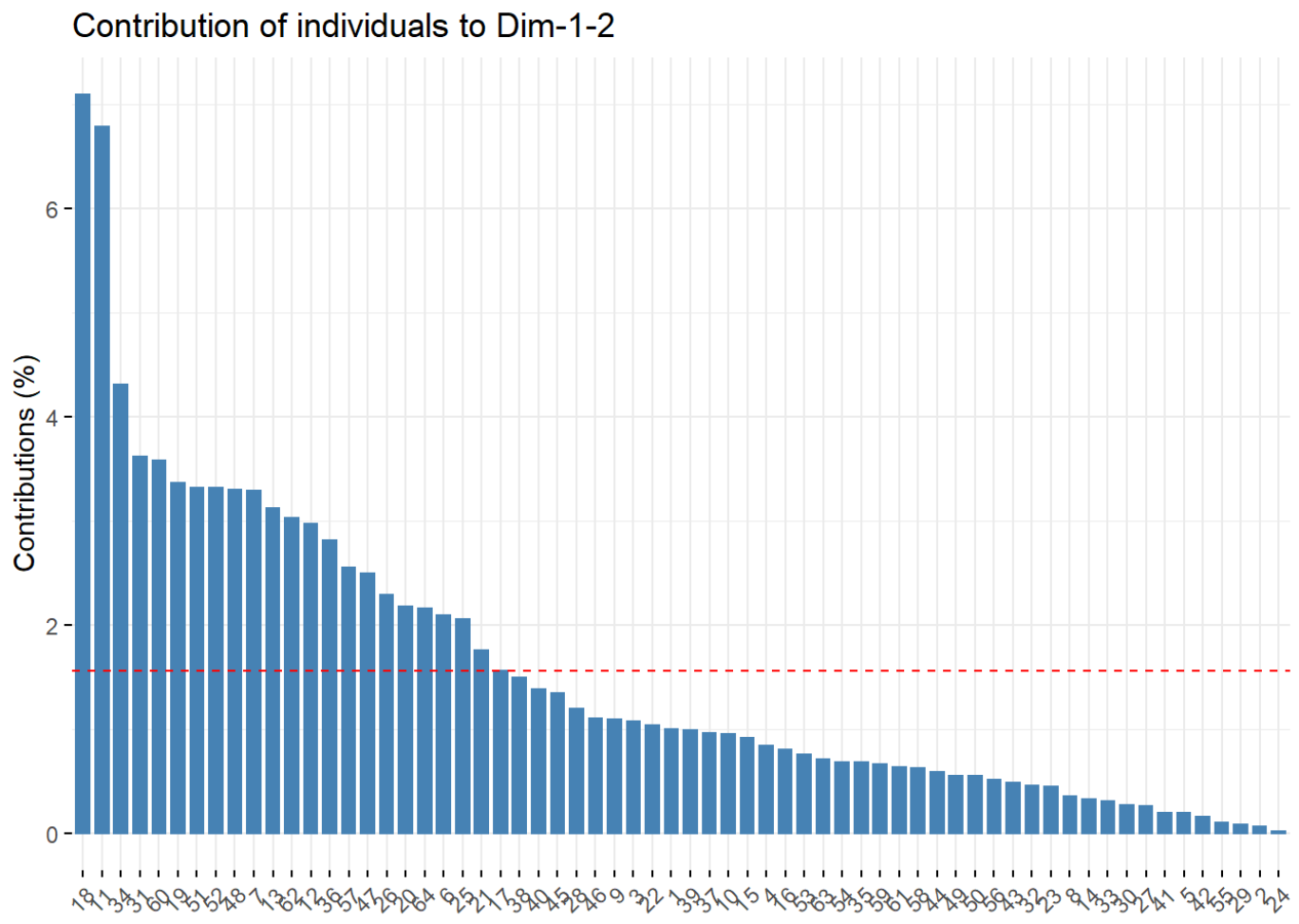
La première dimension est plus liée aux variables `Hauteur` (0.92), `Forme` (0.74).

Avec un niveau de de confiance de 95%, la variable supplémentaire `Conf_7j` (p-valeur=0.03082594) caractérise la première dimension; les fruits Conformes (C) à 7 jours ont des coordonnées significativement plus élevées que la moyenne sur la première dimension tandis que les fruits non conformes (NC) à 7 jours ont des coordonnées inférieures à la moyenne.

La deuxième dimension est caractérisée par `Diamètre` (0.88) et `Forme` (-0.625). La modalité 3 de la coloration `Colo_3` (p-valeur=0.02) est aussi liée à la deuxième dimension avec un niveau de confiance de 95%.

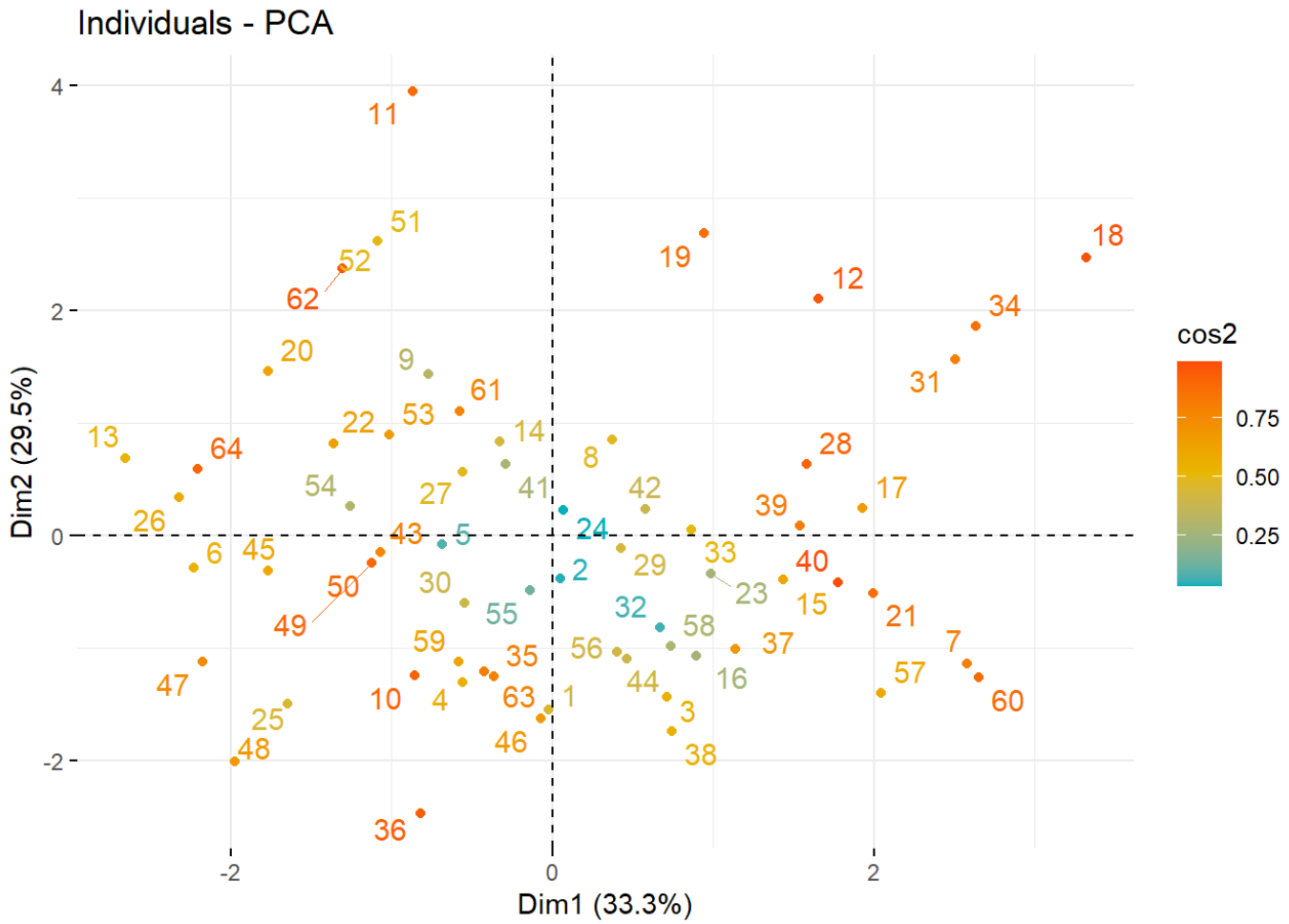
4. Etude sur les Fruits

23 fruits contribuent fortement aux deux principales dimensions.



5. La représentation des fruits sur les deux principales est ci dessous:

```
fviz_pca_ind (res.pca, col.ind = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE # Évite le chevauchement de texte
              )
```

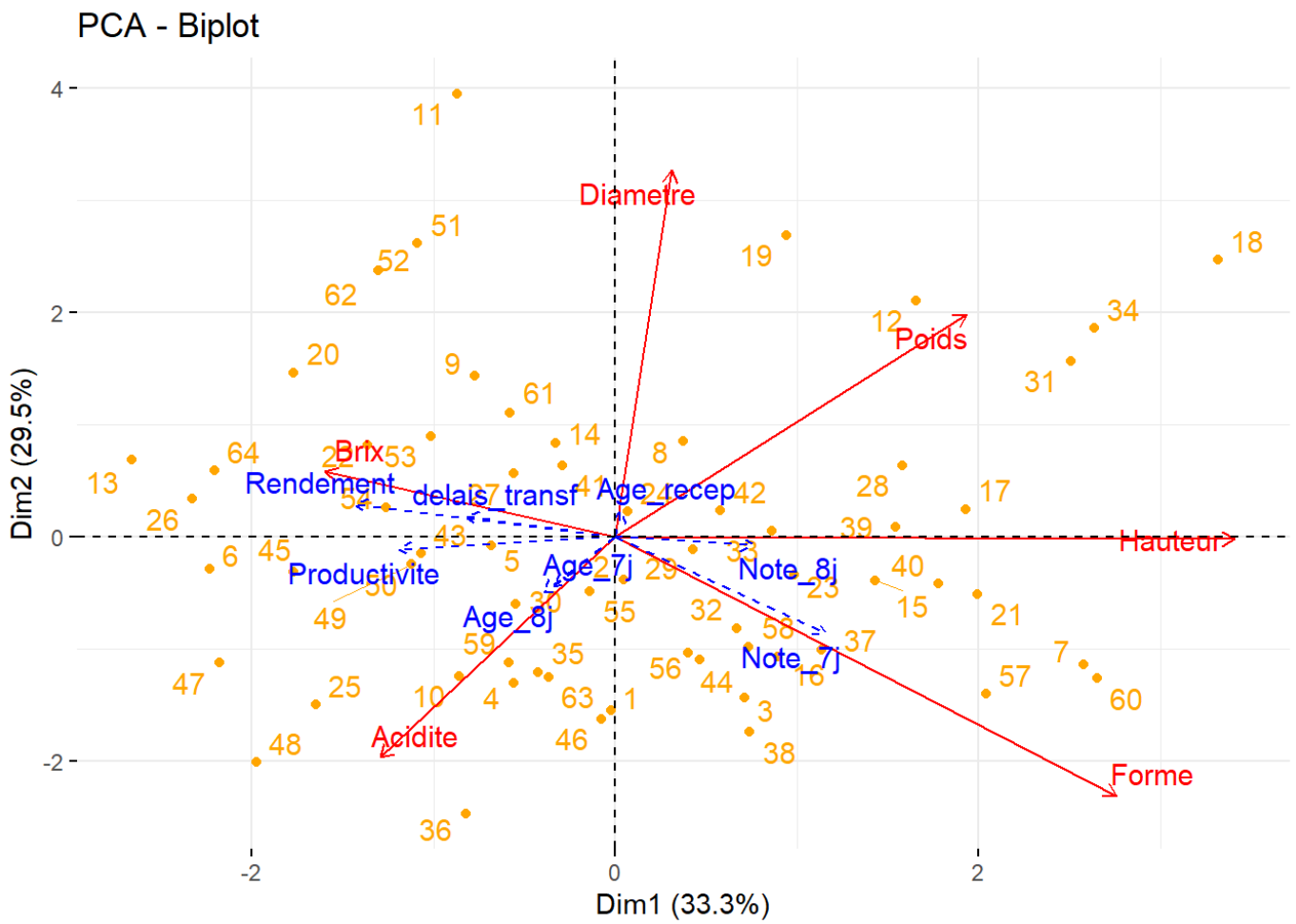


A partir du graphique des individus nous remarquons la dimension 1 oppose des fruits caractérisés par une coordonnée fortement positive par exemple les individus 60,7,21,40,19,34,31,18,12 à des individus caractérisés par une coordonnée fortement négative On peut citer par exemple les individus 11,62,64, 47,18.

La dimension 2 oppose les fruits tels que 18,62,12 64, 11, 19, 34 et 31 à des fruits comme 36, 47, 47 et 48.

Pour associer à un groupe de fruit qui se ressemble un profil particulier nous utiliserons le biplot qui est graphique présentant les variables et les individus.

```
fviz_pca_biplot(res.pca, repel = TRUE,
                col.var = "red", # Couleur des variables
                col.ind = "orange"  # Couleur des individus
                )
```



L'analyse du biplot sur les différentes dimensions laissent apparaître plusieurs groupes.

Ainsi sur la première dimensions on a:

- Le groupe constitué par les individus (7,21 et 60) se distingue par de fortes valeurs pour les variables `Forme` , `Hauteur` et `Note_j7` et de faibles valeurs le taux de sucre `Brix` .
- Un autre groupe représenter par les fruits (34,31,12,18,19) avec de fortes valeurs pour les variables `Diametre` , `Poids` et `Hauteur`, et de faibles valeurs pour les variables `Acidite` et `Rendement` .
- Le groupe auquel les fruits (36, 48 et 47) appartiennent on de fortes valeurs pour les variables `Acidite` et `Brix` , et de faibles valeurs pour les variables `Poids` , `Diametre` , `Hauteur` et `Note_8j` .

On peut également observé un groupe qui partage de fortes valeurs pour les variables `Rendement` et `Diametre` , et de faibles valeurs pour les variables `Forme` , `Note_7j` et `Hauteur` auquel les fruits (62, 64 et 11) appartiennent.

Pour la deuxième dimension on peu mettre en évidence :

Le groupe représenté par les individus (11, 62 et 64) qui ont des valeurs élevées pour les variables `Rendement` et `Diamètre` , et des valeurs faibles pour les variables `Forme` , `Note_7j` et `Hauteur` .

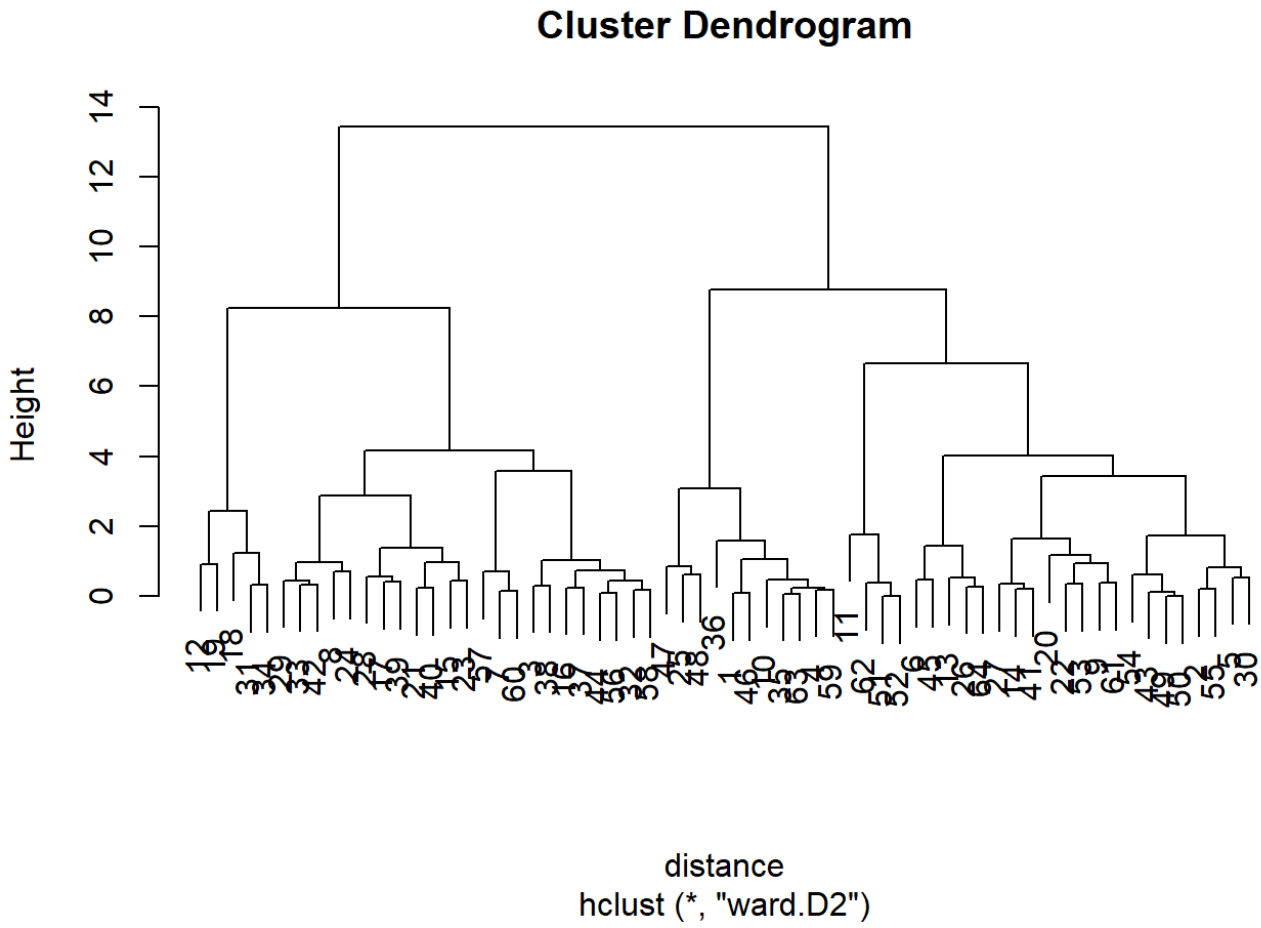
Un autre groupe avec de fortes valeurs pour les variables `Poid` , `Hauteur` et `Diamètre` , et de faibles valeurs pour les variables `Acidite` et `Rendement` s'observe également avec les fruits (19, 34, 18, 31).

Finalement nous pouvons aussi distigué Le groupe auquel les individus (36, 47 et 48) appartiennent caractérisé par de fortes valeurs pour les variables `Acidite` et `Brix` , et de faibles valeurs pour les variables `Poids` , `Diametre` , `Hauteur` et `Note_8j` .

VI. Classification

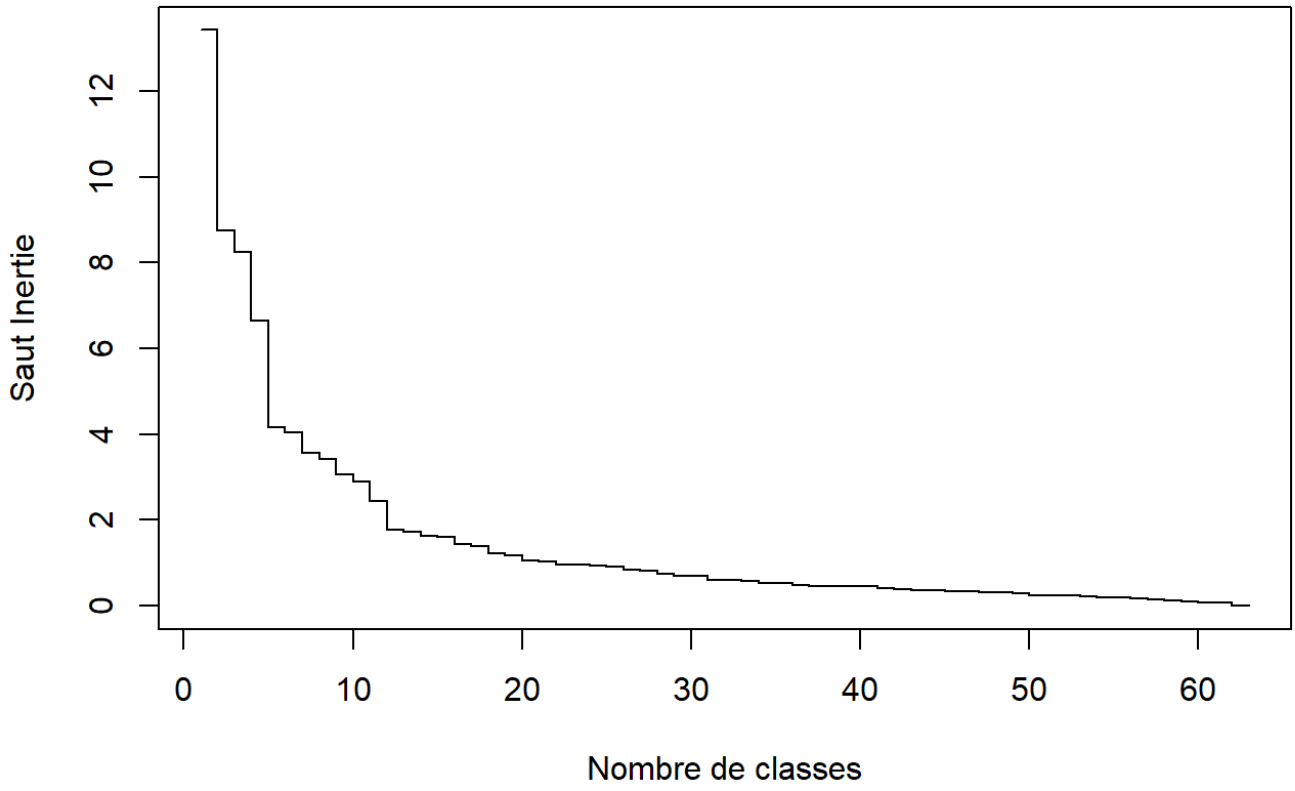
Au vu des analyses précédentes, on a réalisé une classification ascendante hiérarchique des individus.

```
#calcul de la distance
distance<-dist(res.ind$coord[,1:2])
#HCPC(res.ind$coord[,1:2], nbclust=-1)

#classification cash
classif<-hclust(distance,method="ward.D2")
plot(classif)
```



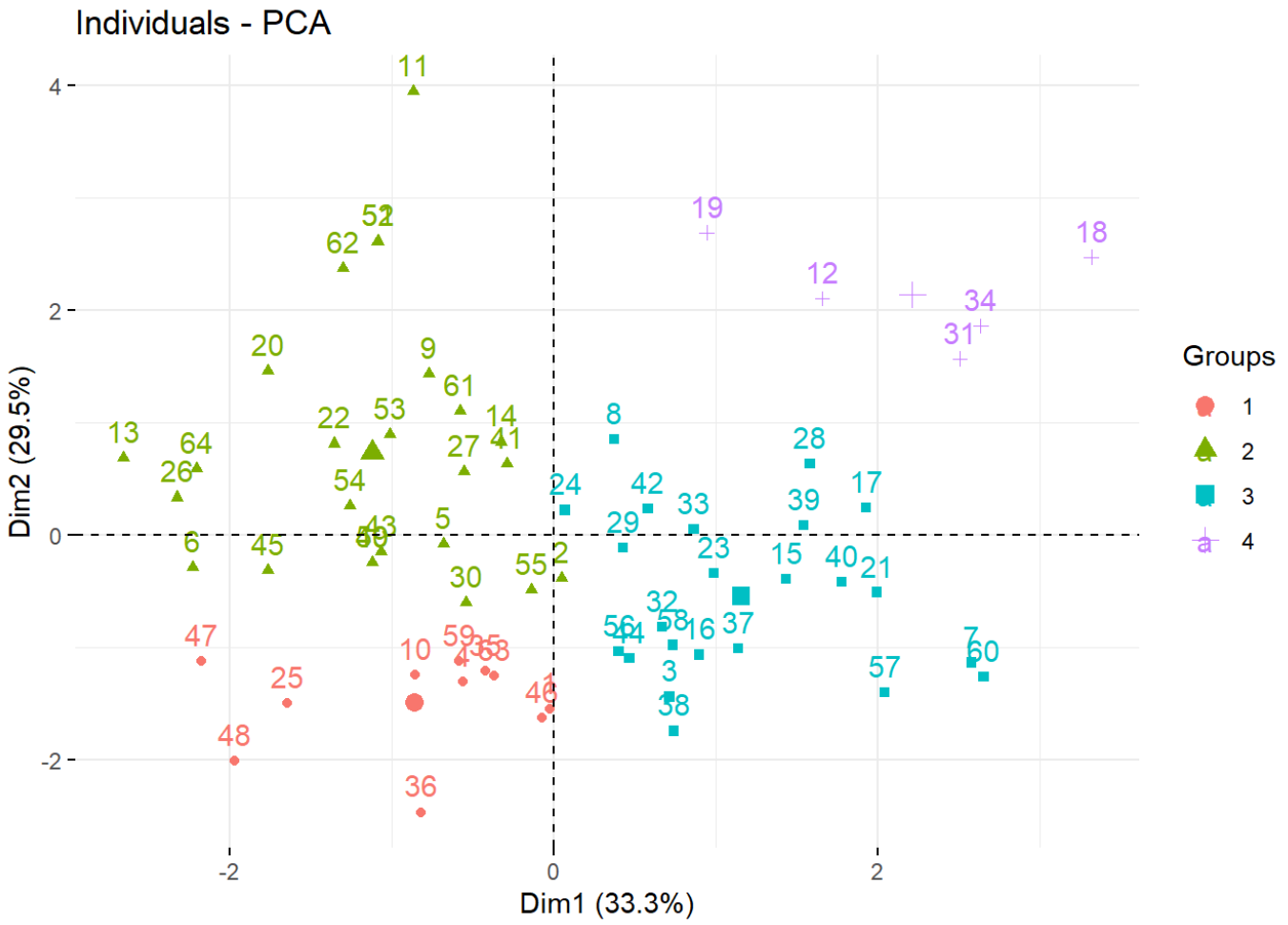
Pour obtenir une partition des fruits, nous allons regrouper les fruits dont le regroupement entraîne une perte minimale de l'inertie intra-classes. Pour cela nous utiliserons les sauts d'inertie du dendrogramme selon le nombre de classes retenu.

```
# trie des inertie par ordre décroissant
saut_inert <- sort(classif$height, decreasing = TRUE)
plot(saut_inert, type = "s", xlab = "Nombre de classes", ylab = "Saut Inertie")
```



Quatre groupes ressortent de cette classification.(k=4)

```
#découpage en 4 groupes
groupes.cah <- cutree(classif,k=4)
#représentation des classes sur les deux premières dimensions
fviz_pca_ind(res.pca,axes=c(1,2),habillage=as.factor(groupes.cah))
```



- Le premier groupe (les individus en rouge) est caractérisé par des fruits très acides. Ces fruits ont de fortes valeurs pour la variable `Acidite` , et de faibles valeurs pour les variables `Diamètre` et `Hauteur` et `Poids` .
- Le deuxième groupe est caractérisé par des fruits riches en sucre. Ces fruits ont de fortes valeurs pour les variables , `Diamètre` , `Brix` et `Rendement` , et de faibles valeurs pour les variables `Forme` et `Hauteur` et `Note_7j`
- Le troisième groupe est constitué par des fruits pauvres en sucre. Ces fruits ont de fortes valeurs pour les variables `Forme` , `Hauteur` et `Note_7j` , et des valeurs moins grandes pour les variables `Diamètre` et `Brix` .
- Le quatrième groupe est constitué par des fruits moins acides. Ces fruits ont de fortes valeurs pour les variables `Poids` , `Hauteur` et `Diametre` , et des valeurs moins grandes pour les variables `Acidite` et `Poids` .

VII. Analyses des correspondances multiples (ACM)

Afin de faire une exploration des données qualitatives et d'appréhender des liaisons entre variables, l'outil statistique le plus pertinent serait l'ACM. Pour cela nous allons utiliser que la base sur l'année 2021 vu que les données sur celle de 2020 sont incomplètes.

L'objectif est d'identifier une typologie de fruits et de détecter les variables qui peuvent jouer sur le rendement et la productivité. Le choix des variables se justifie par : les variables qualitatives à 7jours sont choisies à la place de celles à 8 jours en raison de leur forte corrélation positive ensuite nous prendrons comme variable illustrative quantitative : `Productivité` et comme qualitative `Conf_7j` .

C’est en fonction de cette dernière que nous tenterons de faire une typologie de fruits. Pour faire l’ACM, nous passerons par les étapes suivantes :

- Choix du nombre de dimension
- Etude des individus(fruits)
- Etude des variables
- Conclusion

Notons que les données peuvent être étudiées à partir des individus, des variables et des modalités, ceci améne à se poser plusieurs types de questions relatives à ces objets de nature différentes.

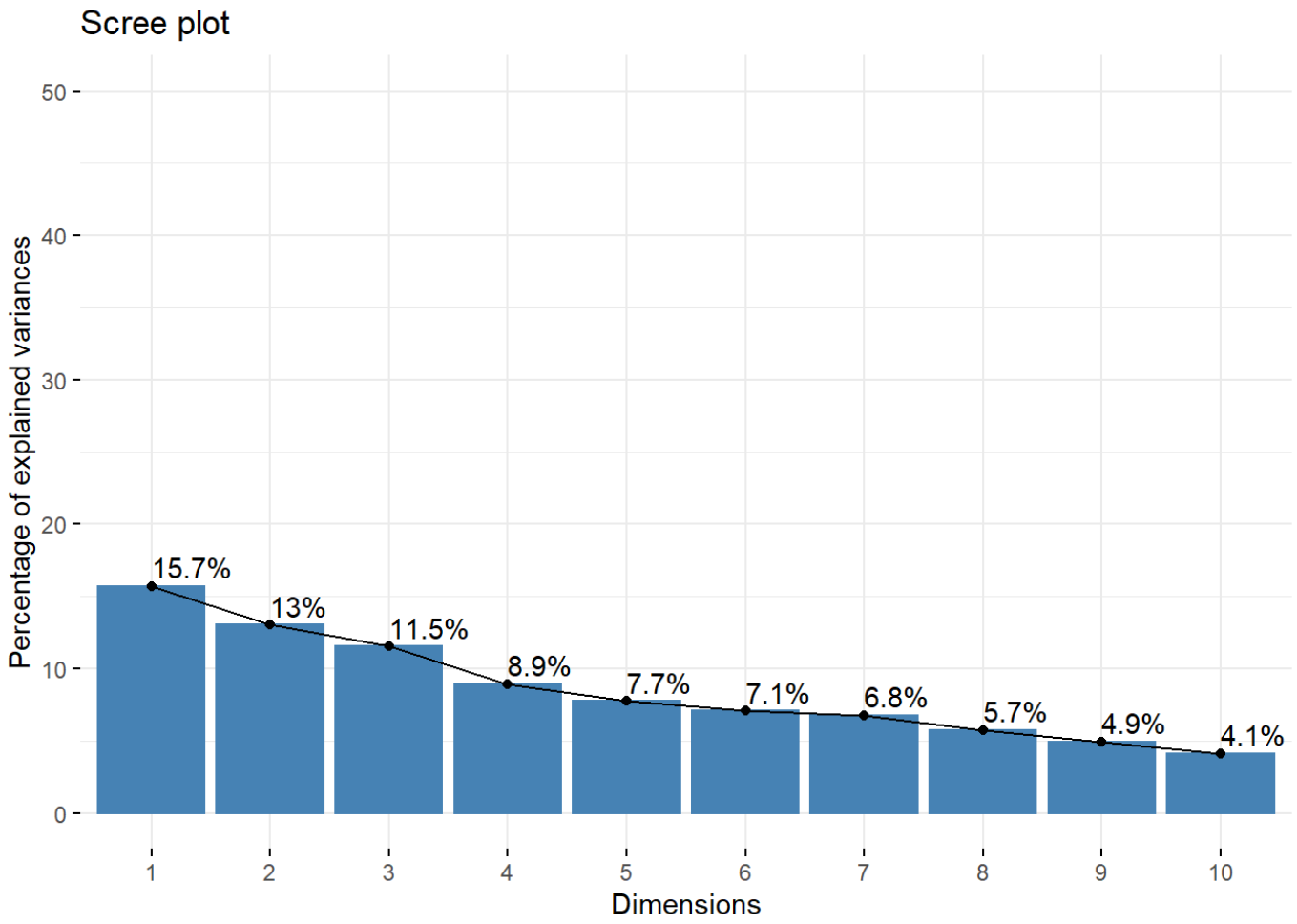
```
#création d'une base avec uniquement les données de 2021
qlNA<-select(mydata2021,Fournisseurs,Colo,Taches,Com_rec,Colo_7j,Conf_7j,Homo_7j,Carac_7j,Pro
ductivite)
#Mise en oeuvre de l'ACM
res.mca <- MCA(qlNA,  quanti.sup = 9, quali.sup=6, graph =F)
```

1. choix du nombre de dimensions

```
val_propr=get_eigenvalue(res.mca)
val_propr
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.38125978       15.6989320                    15.69893
## Dim.2  0.31618546       13.0194014                    28.71833
## Dim.3  0.28003091       11.5306844                    40.24902
## Dim.4  0.21681621        8.9277261                    49.17674
## Dim.5  0.18773533        7.7302781                    56.90702
## Dim.6  0.17145957        7.0601000                    63.96712
## Dim.7  0.16393699        6.7503468                    70.71747
## Dim.8  0.13905643        5.7258529                    76.44332
## Dim.9  0.11920760        4.9085483                    81.35187
## Dim.10 0.09913132        4.0818777                    85.43375
## Dim.11 0.08399379        3.4585677                    88.89232
## Dim.12 0.07544354        3.1064987                    91.99881
## Dim.13 0.05638702        2.3218184                    94.32063
## Dim.14 0.04598056        1.8933173                    96.21395
## Dim.15 0.03984629        1.6407298                    97.85468
## Dim.16 0.03415113        1.4062230                    99.26090
## Dim.17 0.01794951        0.7390973                   100.00000
```

```
#histogramme pour le coude
scree.plot=fviz_eig(res.mca,addlabels = T,ylim=c(0,50))
scree.plot
```



La décroissance des valeurs propres est un peu régulière, nous n'observons pas de décrochage flagrant donc il sera difficile de choisir le nombre d'axes graphiquement. Par contre, on sait que les trois dimensions expriment 40 % de l'inertie totale c'est à dire 40% de l'information du tableau est résumée par les trois premières dimensions donc il serait intéressant d'en retenir 3.

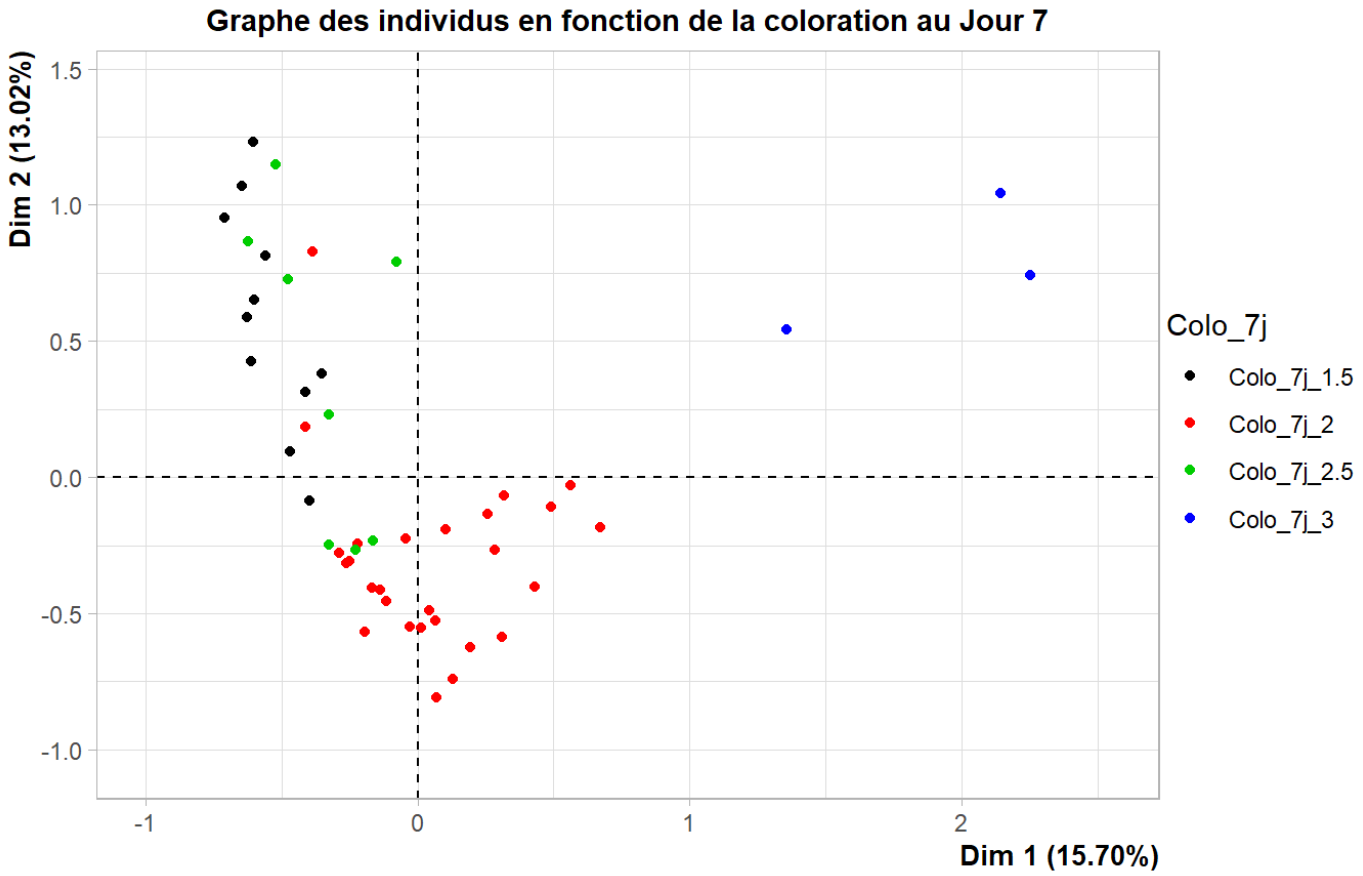
2. Analyse des résultats

3. Etude des individus

L'étude des individus consiste de comprendre les ressemblances entre individus du point de vue de l'ensemble des variables, à dresser une typologie des individus en se basant sur les modalités d'une variable.

Dans cette partie, le but est de représenter les individus (fruits) afin d'apprécier l'allure générale du nuage de points et de voir s'il y'a des formations de groupes. Ici on décide de choisir la coloration du fruit au jour 7 et les commentaires à la réception pour voir si on peut classer les individus en fonction de ces deux variables. Nous représenterons deux graphes où nous colorierons les individus en fonction de la variable `Colo_7j` et en fonction de la variable `Com_recep` .

```
plot(res.mca,invisible = c("var"),habillage=5,title = "Graphe des individus en fonction de la coloration au Jour 7", label="none")
```

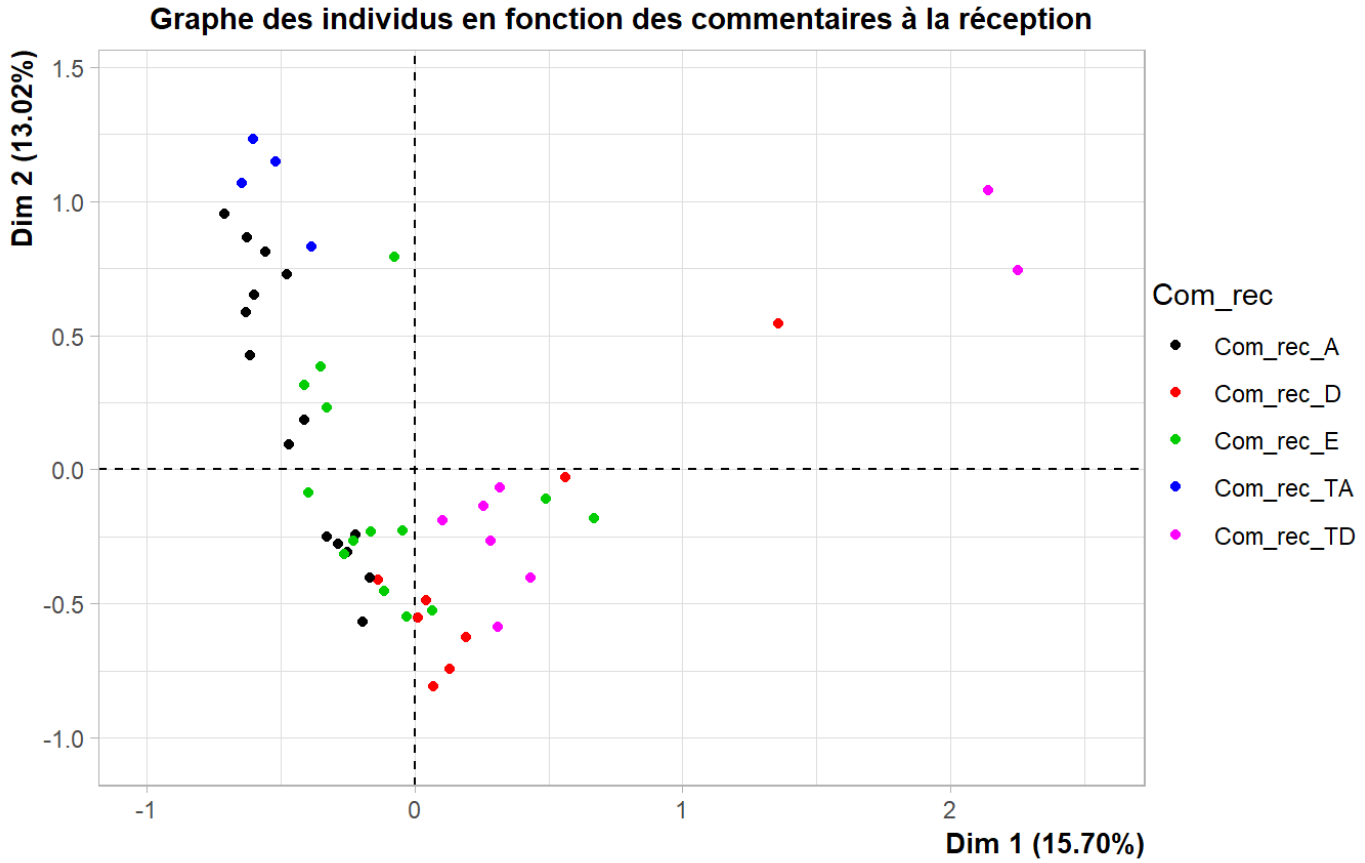


Le graphe des individus (ici fruits) nous permet de voir que des groupes particuliers se dégagent en fonction de la variable `COlo_7j` .

le second axe sépare les fruits à coloration 1.5 et ceux à coloration 3 puis le premier axe sépare les fruits à coloration 2 et ceux à coloration 3.

Le interprétations et les conclusions pourront se faire après avoir fait la représentation des variables. A partir de là, nous aurons une vue globale des tendances afin de faire des typologies de fruits.

```
plot(res.mca,invisible = c("var"),habillage=4,title = "Graphe des individus en fonction des c
ommentaires à la réception", label="none")
```



Vu le grand nombre de modalités, il serait très difficile d'identifier les groupes. Dans ce cas, pour une analyse plus simpliste nous allons regrouper les variables `A` (Acides) et `TA` (très acides), ensuite regrouper `D` (doux) et `TD` (très doux). Nous remarquons que l'axe 1 sépare les fruits acides à droite et ceux doux à gauche.

Ces informations combinées aux analyses qui seront tirées sur le graphe des variables seront déterminantes.

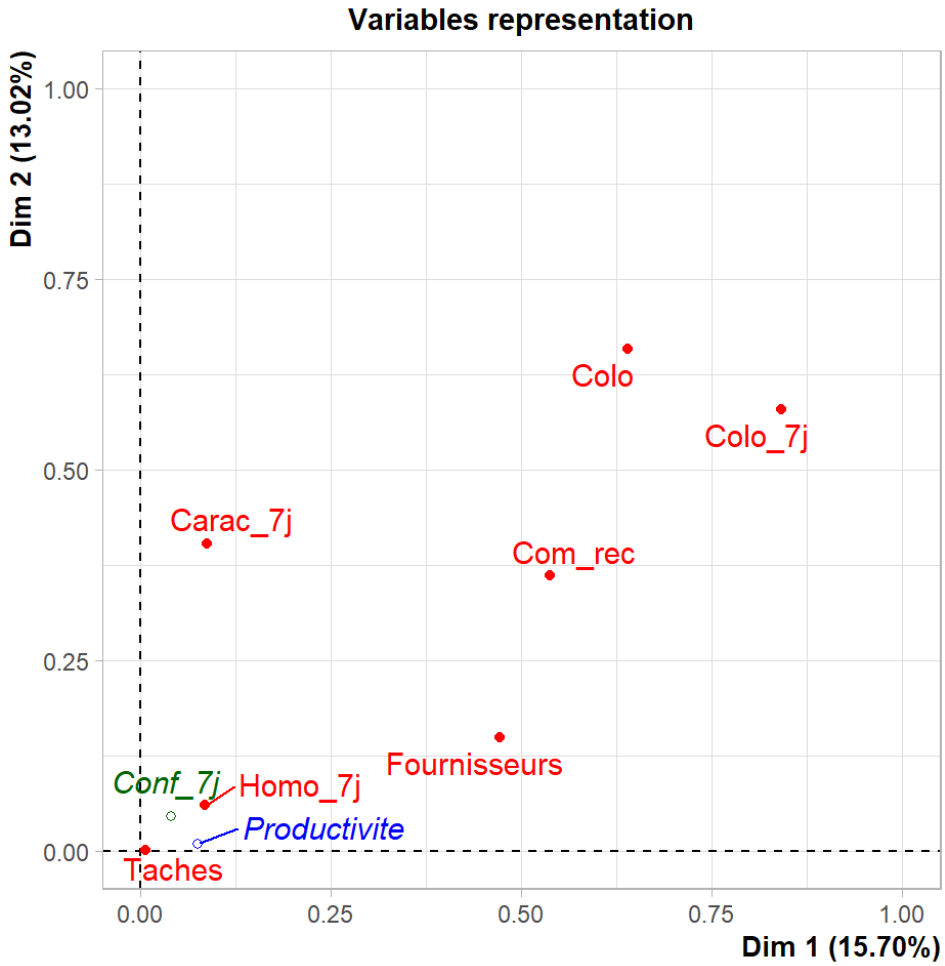
4. Etude des variables

Les variables peuvent être représentées en calculant le rapport de corrélation entre les individus sur un axe et chacune des variables qualitatives. Si le rapport de corrélation entre la variable j et l'axe s est proche de 1 ,les individus possédant la même modalité (pour cette variable qualitative) ont des coordonnées voisines pour l'axe s.

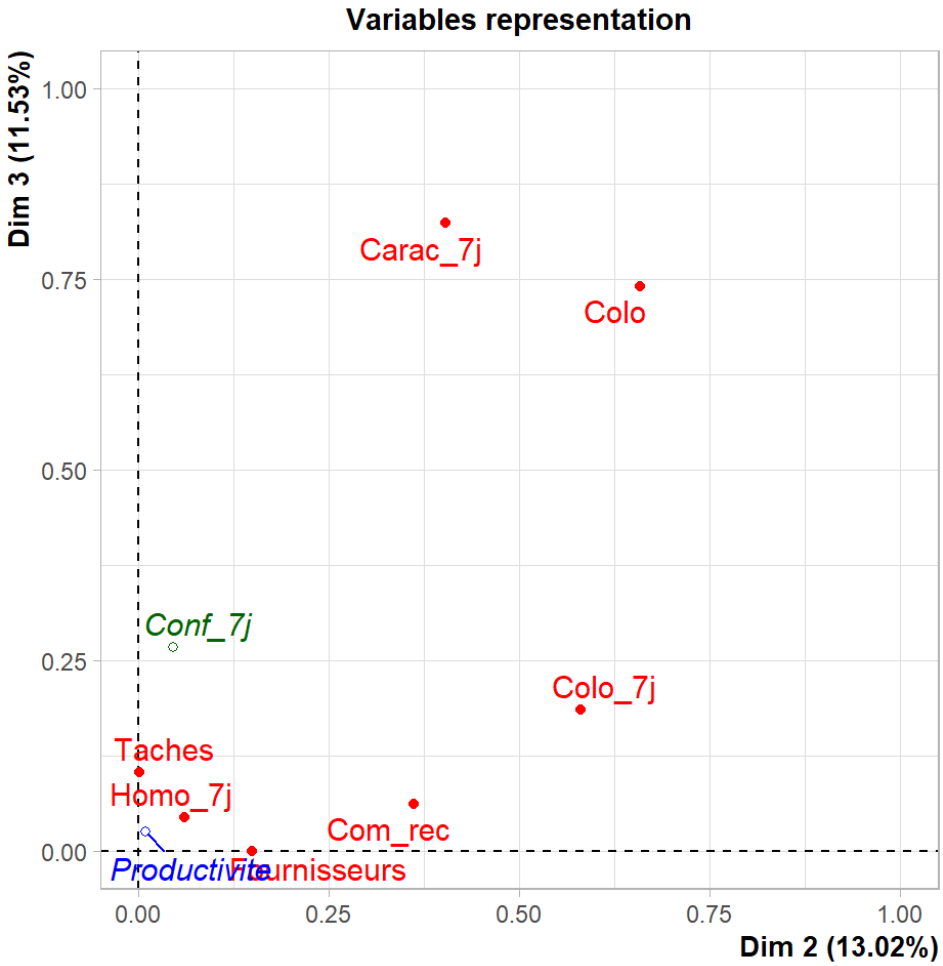
5. Graphe des variables

Nous allons essayer d'identifier les variables qui contribuent le plus à la formation des trois dimensions.

```
# Représentation des variables aux axes 1 et 2
plot(res.mca,choix = "var")
```




```
# Représentation des variables aux axes 2 et 3
plot(res.mca,choix = "var",axes = 2:3)
```



Ainsi, nous voyons que la variable `Colo_7j` participe le plus à la formation de l'axe 1, la variable `Colo` à celle de l'axe 2 et enfin la variable `Carac_7` celle de l'axe 3. Cette information résume l'influence globale de toutes les modalités de chacune des variables sur la construction des axes.

Par rapport aux variables supplémentaires : la variable quantitative `Productivité` est liée à la dimension 1 .Puis,on constate que la variable qualitative `Conf_7` est très liée à la dimension 3.

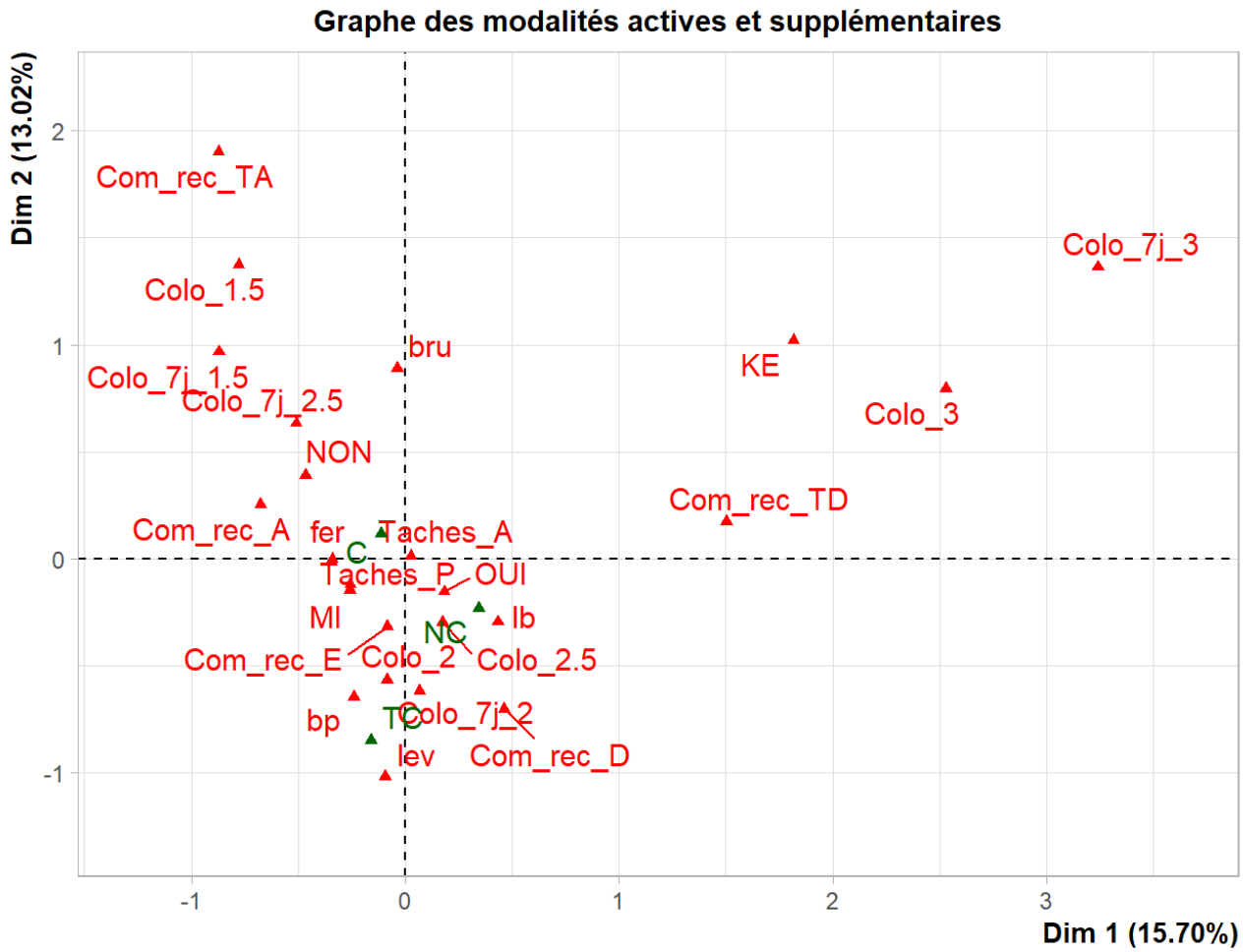
Pour une interprétation plus détaillée nous allons représenter le graphe des modalités de variables

6. Graphe des modalités de variables

Poussons à présent notre analyse en essayant de savoir les grandes tendances en représentant uniquement les modalités de variable.

i. Analyse sur l'axe 1 et 2

```
plot(res.mca,invisible = "ind",title = "Graphe des modalités actives et supplémentaires")
```

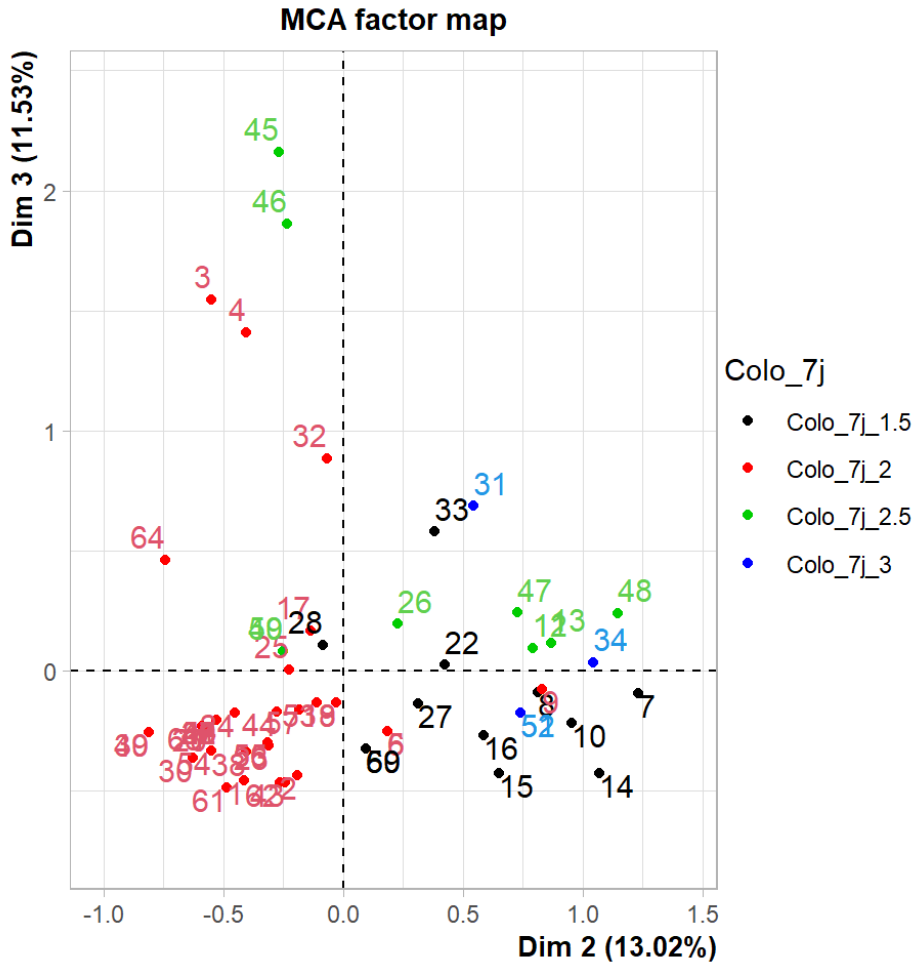


En se basant sur les règles générales d'interprétation de proximité entre modalités de deux variables différentes ou non . Nous pouvons classer les fruits en fonction de la conformité à 7jours :

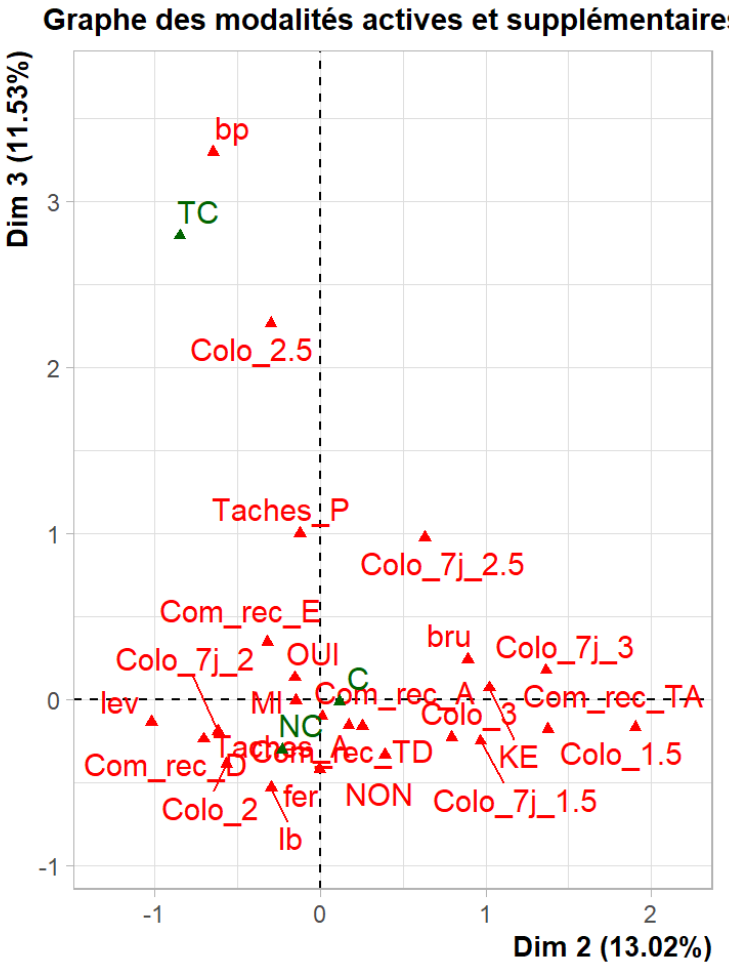
- Les fruits avec comme caractéristiques : de la levure au 7éme jour `lev` , `bp` beau produit ont tendance à être très conformes `TC`
- Les fruits fermentés au 7ème jour en l'absence de taches sont classés conformes `C`
- Enfin les fruits caractérisés par une présence de levures et de brunissement et avec des taches ont tendance à ne pas être conformes `NC`

ii. 8. Analyse sur l'axe 2 et 3

```
options(ggrepel.max.overlaps = Inf)
plot(res.mca,invisible = c("var","quali.sup"),axes=2:3,habillage=5)
```




```
plot(res.mca,invisible = "ind",axes=2:3,title = "Graphe des modalités actives et supplémentai
res")
```



Ce graphe nous permet d'ajouter à l'ancienne typologie de fruits que :

Les fruits classés "Conformes" ont tendance à être aussi homogènes.

iii. Description automatique des axes

Les axes fournis par l'ACM peuvent être décrits de façon automatique par l'ensemble des variables, qu'elles soient quantitatives ou qualitatives, actives ou supplémentaires. Pour cela, nous utiliserons la commande `dimdesc` qui nous permettra de voir à quelles variables les axes sont le plus liés, quelles variables continues sont les plus corrélées à chaque axe et quelles variables qualitatives et quelles modalités le mieux chaque axe.

```
dimdesc(res.mca)$'Dim 1'
```

```
## $quanti
##              correlation    p.value
## Productivite   0.2729246 0.02911165
##
## $quali
##                       R2      p.value
## Colo_7j      0.84159179 5.669939e-24
## Colo         0.64002162 2.464128e-13
## Fournisseurs 0.47164153 3.721100e-10
## Com_rec      0.53680977 2.323475e-09
## Homo_7j      0.08488725 1.950068e-02
##
## $category
##                        Estimate      p.value
## Colo_7j=Colo_7j_3    1.7031288 7.445301e-18
## Colo=Colo_3          1.2772982 4.168380e-12
## Fournisseurs=KE      0.6411025 3.721100e-10
## Com_rec=Com_rec_TD   0.8863764 1.004562e-07
## Homo_7j=OUI          0.2000629 1.950068e-02
## Carac_7j=lb          0.3024120 2.975303e-02
## Homo_7j=NON         -0.2000629 1.950068e-02
## Colo=Colo_1.5       -0.7644964 7.264360e-04
## Colo_7j=Colo_7j_1.5 -0.8349326 5.875329e-04
## Com_rec=Com_rec_A   -0.4589656 2.796703e-04
## Fournisseurs=MI     -0.6411025 3.721100e-10
##
## attr(,"class")
## [1] "condes" "list"
```

```
dimdesc(res.mca)$'Dim 2'
```

```
## $quali
##                     R2      p.value
## Colo         0.6582347 5.260275e-14
## Colo_7j      0.5796712 2.458355e-11
## Carac_7j     0.4033649 3.118703e-06
## Com_rec      0.3612979 2.103468e-05
## Fournisseurs 0.1490601 1.627515e-03
##
## $category
##                         Estimate      p.value
## Colo=Colo_1.5        0.58955503 9.314850e-12
## Carac_7j=bru         0.62116014 1.712618e-06
## Com_rec=Com_rec_TA   0.92255972 3.756641e-05
## Colo_7j=Colo_7j_1.5  0.21404899 1.084405e-04
## Fournisseurs=KE      0.32821817 1.627515e-03
## Colo_7j=Colo_7j_3    0.43666949 4.323271e-03
## Colo_7j=Colo_7j_2.5  0.02575112 2.942703e-02
## Com_rec=Com_rec_D   -0.54269678 1.509055e-02
## Carac_7j=lev        -0.45272195 1.680956e-03
## Fournisseurs=MI     -0.32821817 1.627515e-03
## Colo=Colo_2         -0.50112011 2.997606e-09
## Colo_7j=Colo_7j_2   -0.67646960 1.775893e-12
##
## attr(,"class")
## [1] "condes" "list"
```

```
dimdesc(res.mca)$'Dim 3'
```

```
## $quali
##                 R2      p.value
## Carac_7j 0.8234337 1.537288e-21
## Colo     0.7409420 1.366527e-17
## Conf_7j  0.2673107 7.586729e-05
## Colo_7j  0.1850391 6.197905e-03
## Taches   0.1034181 9.563729e-03
##
## $category
##                         Estimate      p.value
## Colo=Colo_2.5        1.0049468 1.928644e-19
## Carac_7j=bp          1.4867198 4.961546e-19
## Conf_7j=TC           1.0420848 2.400444e-05
## Colo_7j=Colo_7j_2.5  0.4231528 5.577736e-04
## Taches=Taches_P      0.2919185 9.563729e-03
## Com_rec=Com_rec_E    0.2225342 4.741068e-02
## Taches=Taches_A     -0.2919185 9.563729e-03
## Carac_7j=lb         -0.5411335 6.919256e-03
## Colo=Colo_2         -0.4006199 1.444664e-04
##
## attr(,"class")
## [1] "condes" "list"
```

La description des axes obtenue à partir des données qualitatives confirme la description faite des axes.

La commande dimdesc nous permet de mesurer la significativité de la représentation des variables sur les axes :s'il s'agit des variables qualitatives nous nous baserons sur la colonne "R2" puis s'il s'agit des variables quantitatives ,sur la colonne "corrélation".

Nous voyons que la variable quantitative supplémentaire `Productivité` est corrélée positivement avec l'axe 1 à 28%. Malgré cette faible corrélation, nous pouvons déduire que : les fruits à coloration de niveau 3 (à la réception et à 7jours) ont tendance à augmenter la productivité puis les fruits acides à la réception, non homogènes et à coloration de niveau 1.5 ont tendance à la réduire.

Notons que la plupart des fruits qui ont tendance à augmenter la productivité ont été fournis par KE.

iv. Conclusion

Le bilan des liaisons entre les variables qualitatives met en évidence des corrélations positives entre les variables `Colo_7j` , `Colo` , `Homo_7j` , `Com_rec` et elles sont liées à la productivité même si c'est faible. Plus finement, ça voudrait dire que les fruits à coloration 3, homogènes et doux ont tendance à augmenter la productivité. Ensuite, on pourrait proposer une typologie des fruits en fonction de leur conformité, paramètre qui influence notamment sur la productivité.

- Les fruits à coloration 2.5/3, jugés "beau produit" sont classés "très conformes".
- Les fruits qui n'ont pas de taches et fermentés au 7éme jour sont classés "conformes"
- Enfin, les fruits caractérisés par un brunissement de la chair comportant de la levure au jour 7, avec une présence de taches sont classés "non conformes".

On pourrait dire aussi que les fruits jugés "très conformes" peuvent augmenter la productivité et sont généralement fournis par KE.

VIII. Contruction de Modèle de prédiction

Dans cette section nous allons essayer de construire des modèles statistiques afin de prédire la productivité, le rendement et la conformité au jour 7.

Dans premier temps nous ferons une représentation du nuage de point entre chaque couple de variable quantitative.

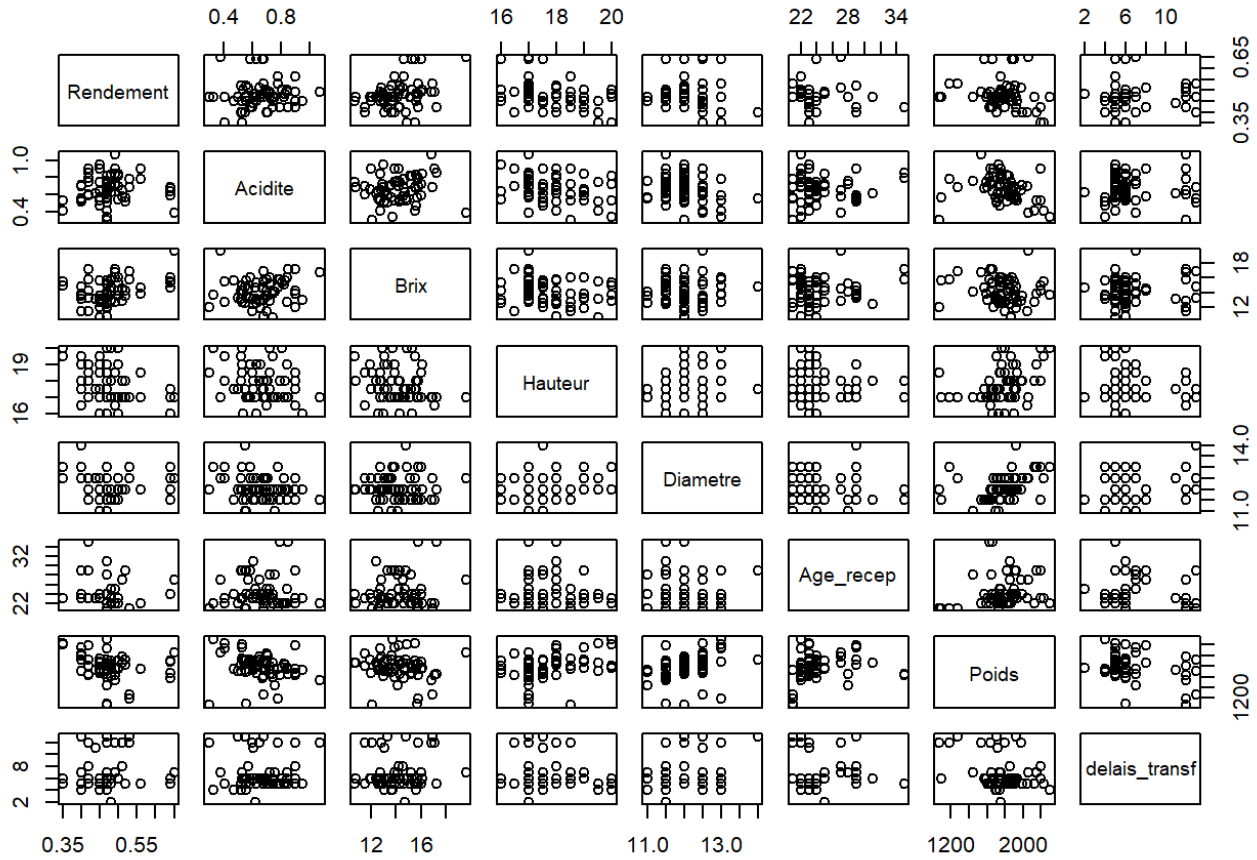
1. Choix des variables du modèle

Pour la construction de nos modèles de régression et de classification nous retiendrons les variables suivantes : `Acidite` , `Brix` , `Tache` , `Hauteur` , `Diametre` , `Colo` , `Age_recep`

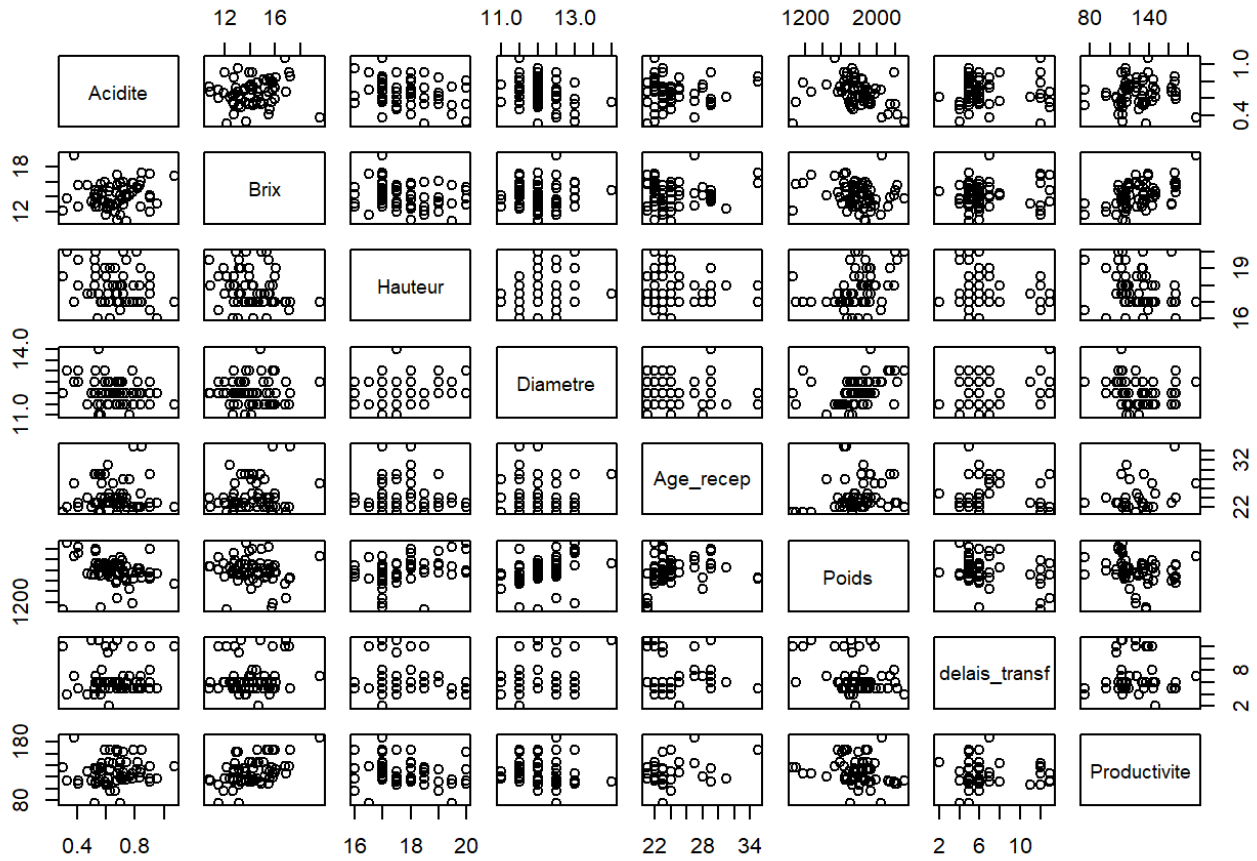
Dans un premier temps nous avons exclu les variables qui ont été mesurées après le phénomène que nous cherchons à expliquer car ces deux mesures sont éloignées dans le temps. Nous pouvons cité par exemple les variables dont les valeurs et modalités ont été déterminées au jour 7 et 8 (`Age_7j` , `Note_7j` , `Conf_7j` , `Colo_7j` , `Homo_7j` , `Carac_7j` , `Age_8j` , `Note_8j` , `Conf_8j` , `Colo_8j` , `Homo_8j` et `Carac_8j`)

Dans un deuxième temps nous avons utilisé les liaisons entre les variables pour déterminer les autres variables explicatives à retenir. Deux variables liées ont tendance à apporter les mêmes informations et peux entraîner des problèmes de multi-colinéarité par exemple. Ainsi lorsque deux variables sont liées on garde l'une des deux.

```
regdatarp<-select(mydata2021,c(Rendement,Acidite, Brix,Hauteur, Diametre,Age_recep, Poids,del
ais_transf, Productivite,Taches, Colo))
plot(regdatarp[1:8])
```




```
plot(regdatarp[2:9])
```



Nous pouvons remarquer une absence de l'existence de lien linéaire entre Rendement et les autres variables il en est de même pour pour `Productivite` .

De ce fait l'utilisation de modèle linéaire n'est pas pertinente. Nous utiliserons les arbres de régression pour la construction des modèles permettant de prédire la productivité et le rendement car pour la construction des arbres de régression on accorde peu d'importance à la structure liant les variables explicatives aux variables expliquées, il n'y a pas d'hypothèse sous-jacente sur les variables et sur le modèle et les dépendances entre les variables explicatives ne posent pas de problème.

Afin de minimiser le taux d'erreur global et d'avoir un traitement efficace nous utiliserons des forêts d'arbres de régression qui est un ensemble d'arbre de régression. La valeur prédite sachant les variables expliquées pour un fruit est données par la moyenne des valeurs prédite des arbres de notre forêt.

Pour évaluer la qualité de notre modèle nous avons séparé notre jeux de données en deux bases que sont train et test. Le train sera utilisé pour la construction de notre modèle et test pour évaluer la qualité du modèle.

2. Creation du jeu de données Train et test

Train : Est le jeu de données sur lequel on va entraîner le modèle et qui correspond à 70% de nos données.

Test : Correspond à 30% de nos données. Nous allons l'utiliser pour tester le modèle créé sur le jeu de données train afin d'évaluer la performance de notre modèle.

```
regdata<-select(mydata2021,c(Rendement,Brix,Acidite,Age_recep,Hauteur,Diametre,Poids,Colo,Tac
hes,Com_rec,delais_transf,Productivite))
set.seed(123)
n = nrow(regdata)
s = sample(2, n, prob = c(0.7,0.3),replace = T)
#creation du train
train = regdata[s==1, ]
#creation du test
test = regdata[s==2, ]
```

i. Creation du modèle pour la prédiction du Rendement

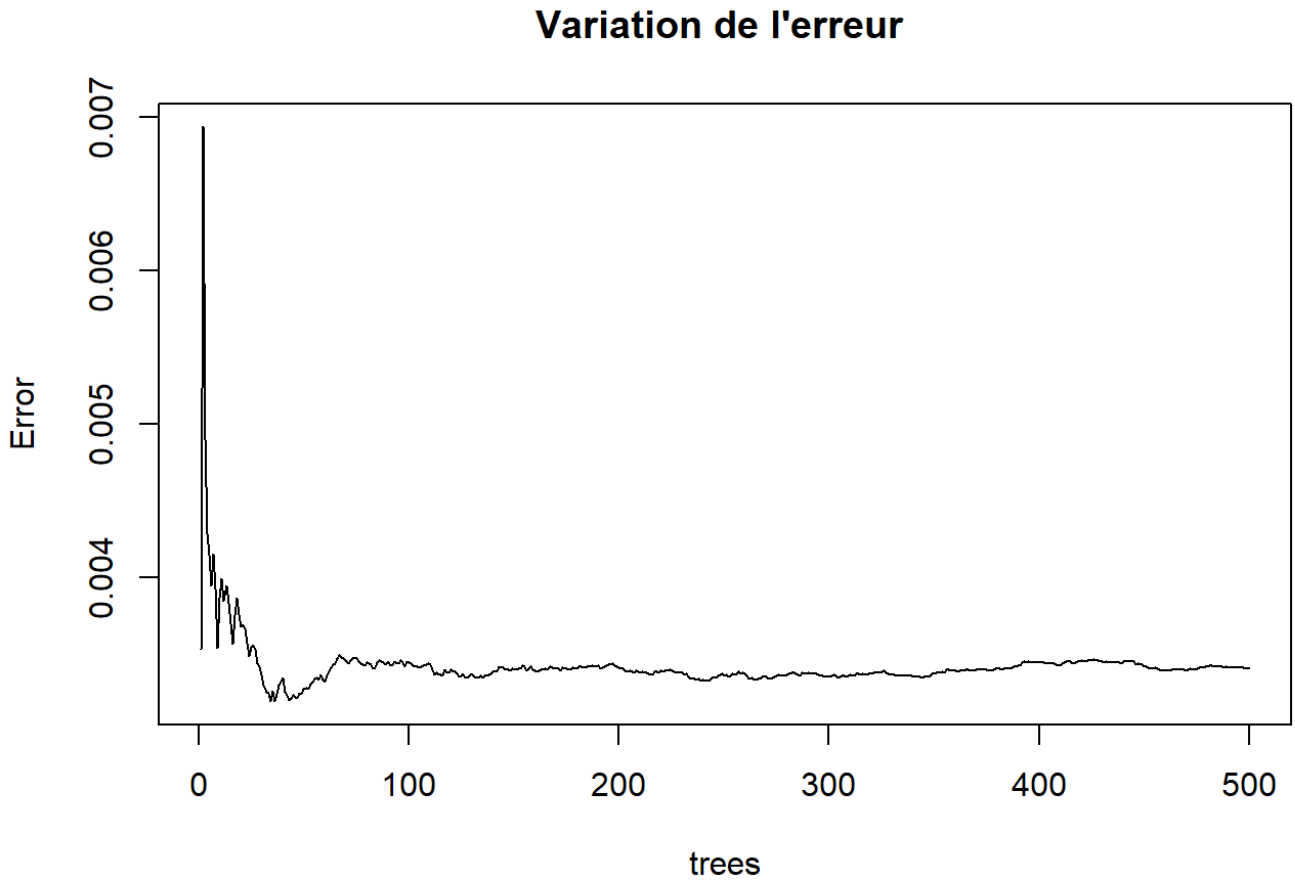
On cherche à modéliser le rendement qui est une variable quantitative, on réalise donc une random forest de type régression.

```
set.seed(123) # permet de fixer les paramètres aléatoires de la randf
#construction de la forêt aléatoire sur toute les variables
randf=randomForest(Rendement~Brix+Acidite+Age_recep+Hauteur+Diametre+Poids+Colo+Taches+Com_re
c+delais_transf,importance = T, ntree = 500,data = train)
randf
```

```
##
## Call:
##  randomForest(formula = Rendement ~ Brix + Acidite + Age_recep +      Hauteur + Diametre +
Poids + Colo + Taches + Com_rec + delais_transf,      data = train, importance = T, ntree = 5
00)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 0.003411018
##                     % Var explained: 28.18
```

On observe ici que la variance expliquée par le modèle créé est d'environs 28.18% et l'erreur résiduelle de 0.003 environs. cette variance expliquée est en dessous de 50%, notre modèle n'est pas pertinent.

```
plot(randf,main = "Variation de l'erreur ")
```



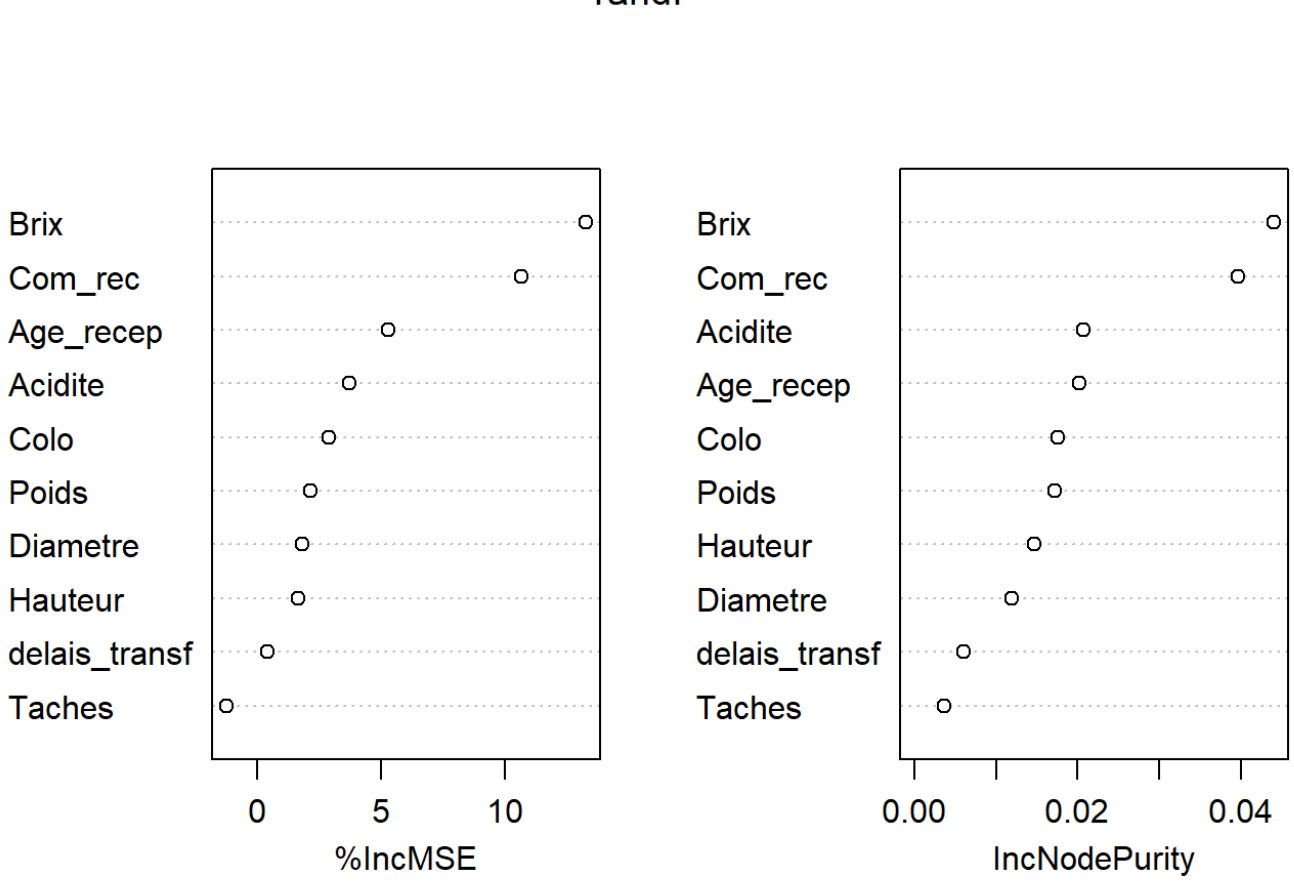
Le graphique ci-dessus nous indique l'évolution l'erreur en fonction du nombre d'arbre. On remarque qu'autour de 150 arbres l'erreur varie peu.

De ce fait, nous pouvons réduire le nombre d'arbre à 150 arbres pour gagner en temps de calcul et sans dégrader la qualité du modèle.

i. Classement des variables explicatives

Afin de visualiser les variables par ordre d'importance classées par le random forest, on utilise la commande `varImPlot` .

```
varImpPlot(randf)
```



Dans le graphique ci-dessus les variables explicatives sont ordonnées par ordre décroissant en fonction de la capacité à prédire le rendement. Ce classement est basé sur la diminution moyenne des erreurs apporter par chaque variable(%IncMSE). par exemple, les 4 premiers critères qui ont utilisé pour prédire le rendement sont: `Brix` , `Com_rec` `Age_recep` , `Acidite` et `Poids` .

i. Validation du modèle

En appliquant notre modèle entrainè au jeu de données test nous pouvons voir les écarts entre la valeur prédite et la valeur réelle.

```
pred <- predict(randf,test[2:12])
ecart_model1=round(abs(test$Rendement-pred),2)
results <- data.frame(actual = test$Rendement, prediction = round(pred,2), erreur=ecart_model
1)
head(results)
```

```
##    actual prediction erreur
## 2    0.45       0.56   0.11
## 4    0.49       0.48   0.01
## 5    0.45       0.45   0.00
## 8    0.40       0.46   0.06
## 11   0.40       0.50   0.10
## 16   0.42       0.47   0.05
```

Pour la validation de notre modèle nous nous servirons de la racine carrée de l'erreur quadratique moyenne (RMSE). On pourra ainsi mesurer la capacité de notre modèle sur des données qui n'ont pas servi à son entrainement c'est à dire nos données de test.

```
evaluation_model<-function(model){
  #prédiction sur le train
  train_predict<-predict(model,train[2:11])
  # calcul de l'erreur sur le train
  train_error<-train_predict-train$Rendement
    #prédiction sur le test
  test_predict<-predict(model,test[2:11])
  # calcul de l'erreur sur le test
  test_error<-test_predict-test$Rendement
  #RMSE sur le train
  rmse_train<-sqrt(mean(train_error**2))
  #RMSE sur le train
  rmse_test<-sqrt(mean(test_error**2))
  #affichage
  print(paste("rmse train",round(rmse_train,2)))
  print(paste("rmse test",round(rmse_test,2)))

    }
```

La fonction ci dessus permet obtenir le RMSE pour le train et le test. Les résultats du RMSE sur les données d'apprentissage est d'environ 0.03.

Cette valeur est inférieure au RMSE des données de test qui est de 0.08. Le RMSE des données d'apprentissage s'explique par le fait que notre modèle est optimale pour les données qui ont servi à son entrainement. On peut remarquer également un écart faible (0.04) entre les deux RMSE.

```
 evaluation_model(randf)
```

```
## [1] "rmse train 0.03"
## [1] "rmse test 0.08"
```

3. Creation du modèle pour la prédiction de la productivité

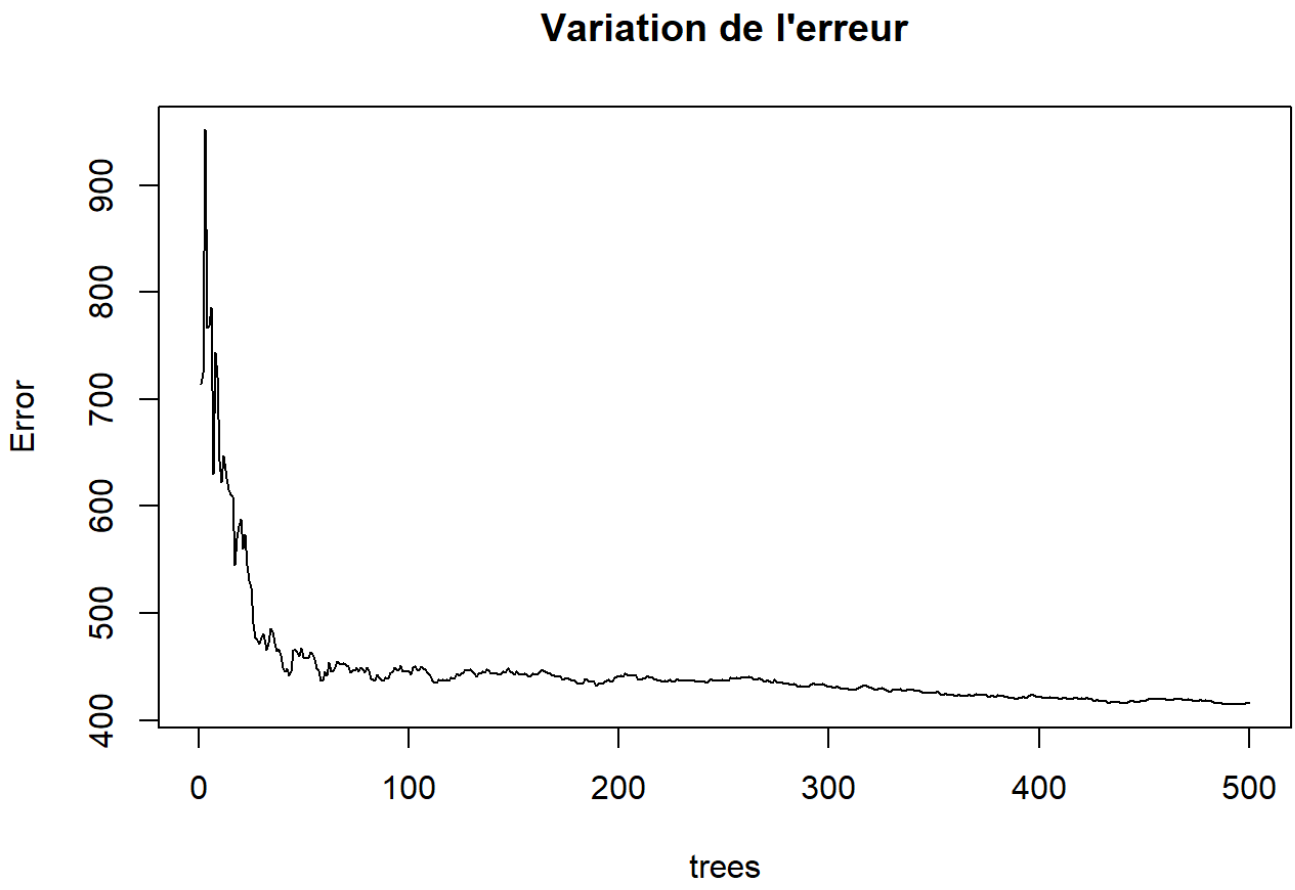
Comme pour le rendement on cherche ici à modéliser la productivité qui est une variable quantitative, on réalise donc une random forest de type régression. On observe ici que la variance expliquée par le modèle créé est d'environs 29.93% et l'erreur résiduelle moyenne est de 415.97 environ.

```
set.seed(123) # permet de fixer les paramètres aléatoires de la randf
randf2=randomForest(Productivite~Brix+Acidite+Age_recep+Hauteur+Diametre+Poids+Colo+Taches+Co
m_rec+delais_transf,importance = T, ntree = 500,data = train)
randf2
```

```
##
## Call:
##  randomForest(formula = Productivite ~ Brix + Acidite + Age_recep +      Hauteur + Diametr
e + Poids + Colo + Taches + Com_rec + delais_transf,      data = train, importance = T, ntree
= 500)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 415.9744
##                     % Var explained: 29.93
```

Le graphique ci dessous nous indique l'évolution de l'erreur en fonction du nombre d'arbres.

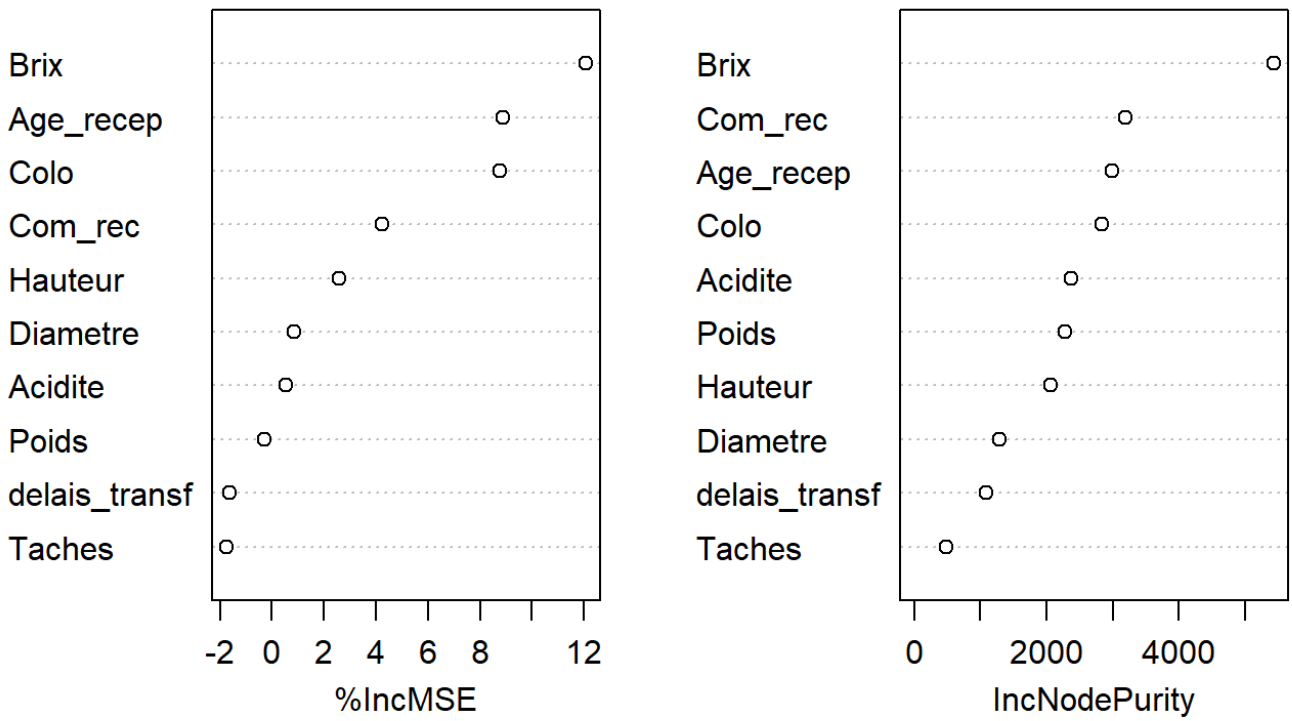
```
plot(randf2,main = "Variation de l'erreur")
```



On remarque qu'autour de 100 arbres l'erreur varie peu. De ce fait, nous pouvons réduire le nombre d'arbre à 100 arbres pour gagner en temps de calcul et sans dégrader la qualité du modèle.

i. Importance des variables

```
varImpPlot(randf2)
```



Dans le graphique ci-dessus les variables explicatives sont ordonnées par ordre décroissant de leurs pouvoirs prédictives. Ainsi en se basant sur la contribution de chaque variable à la diminution de l'erreur moyenne (%IncMSE), les 4 critères qui comptent le plus pour prédire le rendement sont : `Brix` , `Age_recep` , `Colo` , `Com_rec` .

i. Validation du modèle

En appliquant notre modèle entrainé au jeux de données test nous pouvons voir les écarts entre la valeur prédite et la valeur réelle de `Productivite` .

```
pred2 <- predict(randf2,test[2:11])
ecart=round(abs(test$Productivite-pred2),2)
results <- data.frame(actual = test$Productivite, prediction = round(pred2,2), erreur=ecart)
head(results)
```

```
##    actual prediction erreur
## 2   129.6     144.20  14.60
## 4   127.6     136.91   9.31
## 5   116.5     114.74   1.76
## 8   115.7     126.10  10.40
## 11  111.8     132.28  20.48
## 16  111.7     128.17  16.47
```

```
results <- data.frame(actual = test$Productivite, prediction = pred2)
head(results)
```

```
##    actual prediction
## 2   129.6   144.2046
## 4   127.6   136.9055
## 5   116.5   114.7389
## 8   115.7   126.0964
## 11  111.8   132.2786
## 16  111.7   128.1745
```

```
evaluation_model(randf2)
```

```
## [1] "rmse train 133.24"
## [1] "rmse test 129.47"
```

Notre fonction evaluation_model implémentée dans la section validation du modèle de `Rendement` permet obtenir le RMSE pour le train et le test. Les résultats du RMSE sur les données d'apprentissage est d'environ 133.24. Cette valeur est légèrement plus élevée que le RMSE des données de test qui est de 129.47. On peut remarquer également un écart faible (3.77) entre les deux RMSE.

En somme les modèles prédictifs pour le rendement et la productivité ne sont pas pertinent car les variances expliqué par nos modèles sont inférieurs à 50%.

3. Modèle de regression pour la Conformité à 7 jours

On s'intèresse à la conformité des fruits. L'évaluation de la conformité est classée selon trois modalités à savoir `C` pour `conforme` , `NC` pour `non-Conforme` et `TC` pour `très-Conforme` . Pour un fruit donnée on souhaite prédire la modalité de la conformité au septième jour en fonction des variables `Acidite` , `Brix` , `Tache` , `Hauteur` , `Diametre` , `Colo` , `Age_rec` . Avec trois modalités nous utiliserons un modèle de régression multinomiale.

```
#création d'une base avec les variables pour la régression
datmul<-select(mydata2021,c(Conf_7j,Hauteur,Brix,Colo,Taches,Acidite,Diametre,Age_recep))
#creation du train et test pour la regression 70% train & 30% test
set.seed(123)
n = nrow(datmul)
s = sample(2, n, prob = c(0.7,0.3),replace = T)
#creation du train
trainreg = datmul[s==1, ]
#creation du test
testreg = datmul[s==2, ]
#modèle
regmul = multinom(Conf_7j~Hauteur+Brix+Colo+Taches+Acidite+Diametre+Age_recep,data = trainre
g)
```

```
## # weights:  36 (22 variable)
## initial  value 50.536165
## iter  10 value 19.706675
## iter  20 value 18.352556
## iter  30 value 18.238965
## iter  40 value 18.238003
## iter  50 value 18.236329
## iter  60 value 18.236236
## final  value 18.236235
## converged
```

On affiche le résumé des estimations des coefficients de régression en faisant:

```
summary(regmul)
```

```
## Call:
## multinom(formula = Conf_7j ~ Hauteur + Brix + Colo + Taches +
##     Acidite + Diametre + Age_recep, data = trainreg)
##
## Coefficients:
##    (Intercept)   Hauteur      Brix   Colo1.5    Colo2     Colo2.5      Colo3
## NC   1.3137905 -1.213246 0.5897114 -1.428036 1.490869  0.02937191   1.221585
## TC  -0.2173783  7.767987 2.4007516 -6.679206 2.352533 32.19953986 -28.090245
##       TachesP    Acidite   Diametre   Age_recep
## NC -0.8022394 -0.8045364  0.9875452 -0.06704456
## TC  0.8272768 -9.0526498 -8.7221997 -3.87684181
##
## Std. Errors:
##    (Intercept)    Hauteur       Brix      Colo1.5     Colo2   Colo2.5
## NC  13.2823955  0.6264232  0.4024747 3.557060e+00  3.094496  3.520088
## TC   0.4897549 44.0522815 26.7870075 2.949498e-08 15.795680 15.314882
##           Colo3    TachesP  Acidite   Diametre  Age_recep
## NC 4.879295e+00 1.76052192 4.266841  0.9341038  0.1535582
## TC 1.403296e-09 0.03329603 3.080415 19.9561628 26.6711312
##
## Residual Deviance: 36.47247
## AIC: 76.47247
```

i. Significativité du modèle

Pour mesurer l'influence des coefficients sur notre variable expliquée nous avons effectué le test de Wald. Le test de Wald est un test de significativité individuelle des coefficients. Il est basé sur les hypothèses suivantes:

H0: Il n'y a pas de variables influentes

H1: Au moins une variable des variables est influente

```
 z= summary(regmul)$coeff / summary(regmul)$standard.errors
pvaleur = 2 * (1 - pnorm(abs(z), 0, 1))
pvaleur
```

```
##    (Intercept)    Hauteur      Brix   Colo1.5     Colo2   Colo2.5     Colo3
## NC   0.9212080 0.05277176 0.1428627 0.6880776 0.6299616 0.9933425 0.8023082
## TC   0.6571502 0.86003027 0.9285862 0.0000000 0.8816047 0.0355095 0.0000000
##      TachesP     Acidite  Diametre Age_recep
## NC 0.6486181 0.850441194 0.2904151 0.6623964
## TC 0.0000000 0.003295116 0.6620621 0.8844288
```

A partir des résultats du test de significativité des coefficients On remarque que :

Pour la non-conformité (NC), les coefficients significativement différents de 0 sont: `Hauteur` (significatif), la modalité 3 de la coloration `Colo3` (significatif).

Pour très Conforme(TC), les coefficients significativement différents de 0 sont: la modalité 2.5 de la coloration `Colo2.5` (significatif), la modalité 3 de la coloration `Colo3` (hautement significatif), la présence de taches `TacheP` (hautement significatif) et `Acidité` (très significatif).

On peut aussi faire le test du rapport de vraisemblance (lr) pour évaluer la significativité globale des coefficients de notre modèle.

```
reg0 = multinom(Conf_7j~ 1,data = trainreg)
```

```
## # weights:  6 (2 variable)
## initial  value 50.536165
## iter  10 value 32.805624
## final  value 32.717098
## converged
```

```
rv = reg0$deviance - regmul$deviance
ddl = regmul$edf - reg0$edf
pvaleur = 1 - pchisq(rv, ddl)
pvaleur
```

```
## [1] 0.04884883
```

Cela renvoie une p-valeur=0.0488 Ainsi, la régression multinomiale est significative avec la p-valeur est inférieure à 0.05.

ii. Taux d'erreurs

La matrice de confusion est une synthèse des résultats des classes prédite par notre modèle. Elle compare les données réelles de la conformité au jour 7 à celles prédites par un modèle. Le nombres de bonne prédiction pour chaque groupe de fruits est sur la diagonale. les autres valeurs sont des erreurs de prédictions pour chaque modèle.

```
#matrice confusion
#2:8 c'est les variables du test.. le 1 c'est l'étiquette
pr = predict(regmul,testreg[,2:8])
mc = table(testreg$Conf_7j, pr)
mc
```

```
##     pr
##       C NC TC
##   C  10  3  2
##   NC  0  2  0
##   TC  1  0  0
```

Notre matrice de confusion indique nous montre que 3 fruits conforme(C) ont été considérer non conforme(NC) par notre modèle et 2 fruits conforme on été considérer non conforme(NC). Les fruits non conforme(NC) ont été bien classifier. Un fruit très conforme(TC) a été classifier conforme.

Pour évaluer la qualité prédictive de notre modèle nous interpréterons le taux d'erreur. Ce taux est la proportion des modalités prédites qui diffèrent des modalités observées.

```
t=(sum(mc)-sum(diag(mc)))/sum(mc)
t
```

```
## [1] 0.3333333
```

On obtient un taux d'erreur de 0.33 cela est inférieur à 0.5, de ce fait notre modèle à une bonne qualité prédictive.

Nous pouvons essayer d'améliorer ou de simplifier notre modèle en effectuant une sélection de variables. En effet en observant les résultats du test de wald, on peut remarquer que certains coefficients sont non significatifs. Mais l'approche qui consiste à éliminer d'un seul coup les variables non significatives n'est pas bonne ; certaines variables peuvent être corrélées à d'autres, ce qui peut masquer leur réelle influence sur la conformité au septième jour .

Ainsi en effectuant une sélection de variables par la méthode "stepwise" en utilisant comme critère de performance le AIC.

```
reg1= step(regmul, direction = "both", k = 2)
```

```
## initial  value 50.536165
## iter  10 value 24.580249
## iter  20 value 22.487311
## iter  30 value 22.080325
## iter  40 value 22.053436
## iter  50 value 22.049279
## iter  60 value 22.033722
## iter  70 value 22.005999
## iter  80 value 21.995694
## iter  90 value 21.994687
## iter 100 value 21.993644
## final  value 21.993644
## stopped after 100 iterations
##              Df      AIC
## <none>        8 60.24377
## + +Acidite   10 61.33168
## - Diametre    6 63.82767
## + +Taches    10 63.92142
## + +Age_recep 10 63.98729
## + +Colo      14 65.15771
## - Brix        6 66.55958
## - Hauteur     6 66.61947
```

```
summary(reg1)
```

```
## Call:
## multinom(formula = Conf_7j ~ Hauteur + Brix + Diametre, data = trainreg)
##
## Coefficients:
##     (Intercept)    Hauteur       Brix    Diametre
## NC -0.189058718 -0.7856596  0.4310384   0.5682312
## TC  0.003793494 22.3228124 23.6167827 -68.1202976
##
## Std. Errors:
##    (Intercept)    Hauteur     Brix  Diametre
## NC  11.4164848  0.4628562  0.23700 0.7150534
## TC   0.1452023 20.3485802 21.07079 2.3351075
##
## Residual Deviance: 44.24377
## AIC: 60.24377
```

Après la sélection des variables les variables qui ont été retenue pour la prédiction de la conformité sont: `Hauteur` , `Brix` , et `Diamètre` . notre nouveau modèle obtenu à un AIC de 60.24377.

iii. Etude de la significativté du deuxième modèle

Le test de Wald pour de la significativité des variables explicatives donne:

```
 z= summary(reg1)$coeff / summary(reg1)$standard.errors
pvaleur = 2 * (1 - pnorm(abs(z), 0, 1))
pvaleur
```

```
##    (Intercept)    Hauteur       Brix  Diametre
## NC   0.9867875 0.08961802 0.06895308 0.4268057
## TC   0.9791572 0.27263237 0.26236011 0.0000000
```

Pour la classe de la non-conformité (NC) il n'y a pas de coefficient significativement différent de 0 au seuil de 5%

Pour la classe très Conforme(TC), seul le diamètre est hautement significatif.

Le test de significativité globale du deuxième modele donne :

```
reg0 = multinom(Conf_7j~ 1,data = trainreg)
```

```
## # weights:  6 (2 variable)
## initial  value 50.536165
## iter  10 value 32.805624
## final  value 32.717098
## converged
```

```
rv = reg0$deviance - reg1$deviance
ddl = reg1$edf - reg0$edf
pvaleur = 1 - pchisq(rv, ddl)
pvaleur
```

```
## [1] 0.001695524
```

Après le test de significativité globale on obtient une p-valeur = 0.0016 nous conclure que notre deuxième régression est très significative.

iv. Taux d'erreur

```
pr = predict(reg1,testreg[,2:8])
#matrice confusion
mc = table(testreg$Conf_7j, pr)
mc
```

```
##     pr
##       C NC TC
##   C  13  2  0
##   NC  2  0  0
##   TC  1  0  0
```

```
t1=(sum(mc)-sum(diag(mc)))/sum(mc)
t1
```

```
## [1] 0.2777778
```

On obtient un taux d'erreur de 0.277 cela est inférieur à 0.5, de ce fait notre modèle a une bonne qualité prédictive.

```
a=paste("AIC modèle 1 : ",round(AIC(regmul),2))
b=paste("AIC modèle 2 : ",round(AIC(reg1),2))
d=paste("Taux d'erreur modèle 1 : ", round(t,3))
e=paste("Taux d'erreur modèle 2 : ", round(t1,3))
rbind(a,b,d,e)
```

```
##   [,1]
## a "AIC modèle 1 :  76.47"
## b "AIC modèle 2 :  60.24"
## d "Taux d'erreur modèle 1 :  0.333"
## e "Taux d'erreur modèle 2 :  0.278"
```

Une analyse comparative de nos deux modèles indique que le deuxième modèle a le plus petit taux d'erreur. de plus son AIC est le plus petit. Pour de futures prédictions, il est préférable d'utiliser le deuxième qui est plus simple (moins de variables) et plus performant que le modèle initial en termes de AIC et du taux d'erreur.

IV. Conclusion et Discussion

L'agro-alimentaire est un secteur dynamique. Les différents acteurs cherchent de nouvelles informations pour améliorer la qualité de leurs produits et de maximiser leurs profits. A travers cette étude nous voulons fournir quelques outils d'aides à la décision pour une entreprise agro-alimentaire sur la découpe de fruit. Deux séries de données sur l'année 2020 et 2021 ont été mises à notre disposition.

Pour y arriver, nous nous sommes d'abord penchés sur une analyse comparative annuelle de certaines données. Nous y avons utiliséss des approches graphiques puis des approches numérique à travers les tests statistiques.

Par la suite nous avons mis en évidence les éventuelles liaisons entre variables. Il a aussi été le lieux d'expliquer la variabilité des données à travers de nouvelles variables synthétiques et de détecter les variables explicatives qui ont servit à la construction de nos modèles de régression et de classification.

A la lumière de ces différentes analyses mise en oeuvre nous avons obtenues des résultats qui mettent en évidence des liasons positives hautement significatives entre le taux de sucre des fruits avec la productivité et le rendement. A l'issue de l'ACP et de l'ACM nous avons trouver que la hauteur, la forme, le diamètre, la coloration et la caractéristique au septième jours pourraient une grande variabilité dans les données. Des résultats de l'ACP quatre typologies de fruit se sont distingués.

Nous allons à présent discuter sur quelques résultats obtenus au cours de notre étude.

Le critère d'évaluation de certaines variables ont changé d'une année à l'autre cela nous a conduit à faire du recodage afin d'avoir les mêmes modalités qui nous ont permis de faire une comparaison par année. Cela peut masquer l'effet réelle d'une ou de plusieurs modalités.

La présence de valeurs manquantes en 2020 nous conduit à considérer uniquement les données de l'année 2021 pour la suite de nos analyses, ainsi les éventuelles liaisons de variables et typologies de fruits en 2020 n'ont pas été prises en compte dans notre étude.

On peut affirmer que les fruits sans taches ont tendance à etre Doux et font une meilleure productivité et un meilleur rendement. Il est aussi préférable de selectionner des fruits de coloration 2.5 et 3 car ils sont souvent plus doux. Les plus gros fruits (Poids, Diametre et hauteur) sont les meilleurs quand on cherche à atteindre un caracteristique tres conforme à 7 jours mais pas quand l'objectif est de trouver des fruits sucrés.

Pour la construction des modèles régressions pour le `Rendement` et `Productivité` les variances expliquées sont inférieures à 50%, ces deux modèles ne sont pas satisfaisants car ils n'expliquent pas les variabilités présentes dans nos données. Cela pourrait être lié à la taille des données qui ont été utilisées pour entraîner

nos modèles.

V. Bliographie

François Husson, Sébastien Lê, Jérôme Pagès, Analyse des données avec R, 2e édition revue et augmentée.

François Husson et Jérôme Pagès, Statistiques générales pour les utilisateurs, 2-Exercices et corrigés.

François Husson et al, R pour la statistique et la science des données.

Christophe Chesneau, Introduction aux arbres de decision https://chesneau.users.lmno.cnrs.fr/arbres.pdf (https://chesneau.users.lmno.cnrs.fr/arbres.pdf)

Christophe Chesneau, Modèles de régression https://chesneau.users.lmno.cnrs.fr/Reg-M2.pdf (https://chesneau.users.lmno.cnrs.fr/Reg-M2.pdf)

Christophe Chesneau, Element de classification https://chesneau.users.lmno.cnrs.fr/classif-cours.pdf (https://chesneau.users.lmno.cnrs.fr/classif-cours.pdf)