script\_graphique

brunop31

26/05/2020

library(dplyr)  
library(DT)  
library(reshape)  
library(splitstackshape)  
library(ggplot2)  
library(scales)  
library(ineq)  
library(zoo)  
library(cluster)  
library(corrplot)  
library(Hmisc)  
library("ggpubr")  
library(rstatix)  
library(BioStatR)  
library(reshape)  
library(questionr)  
library("FactoMineR")  
library("factoextra")  
library("gplots")  
library(tidyr)  
library(gglorenz)

###Je charge les tableaux creer dans mon script nettoyage  
load("table\_p4")  
  
###Je calcul des mesures de tendances et de dispersions  
  
###Sur les gains par jour avant le 02/10  
summary(filter(nbre\_date, date < "2021-10-02")$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 13792 15241 15700 15830 16277 18792

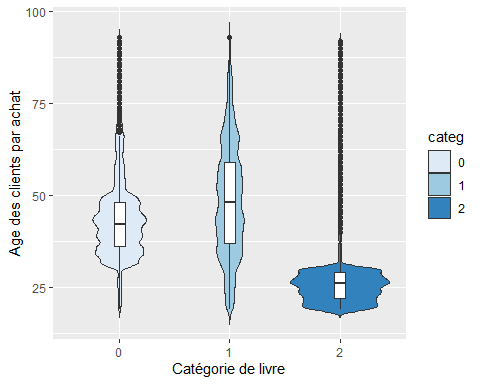
###Sur les gains par jour entre le 02/10 et le 27/10  
summary(filter(nbre\_date, date > "2021-10-02" &  
 date < "2021-10-27")$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8076 8839 9205 9187 9460 10231

###Sur les gain par jour sur toute la période  
summary(nbre\_date$price)

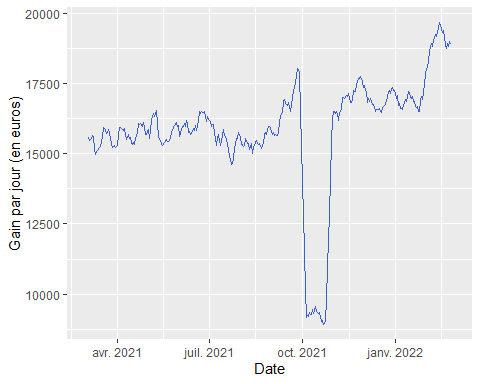
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8076 15280 16066 15884 17063 20313

###Je crée un violin plot pour mesurer la dispersion des aages des client  
###par achat en fonction de leur âge pour chaque catégorie  
  
###Je filtre les gros clients et les clients de 18 ans trop nombreux  
ggplot(filter(transactions,!(client\_id %in% big\_customers),  
 age != 18),  
 aes(x = categ, y = age, fill = categ)) +  
 geom\_violin(trim=FALSE)+  
 geom\_boxplot(width=0.1, fill="white")+  
 labs(x="Catégorie de livre", y = "Age des clients par achat") +  
 scale\_fill\_brewer(palette="Blues")



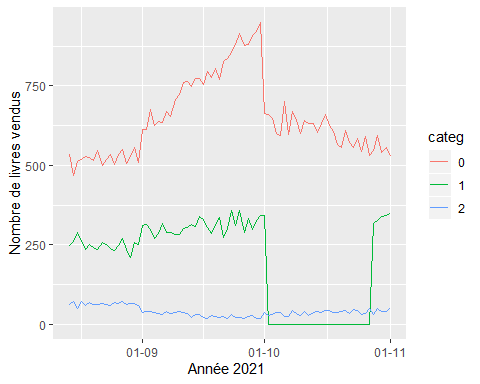
###Je crée une représentation de série temporelle, les gains par jours  
ggplot(nbre\_date, aes(date, roll7)) +  
 geom\_line(col = "royalblue3") + xlab("Date") +   
 ylab("Gain par jour (en euros)")

## Warning: Removed 6 rows containing missing values (geom\_path).

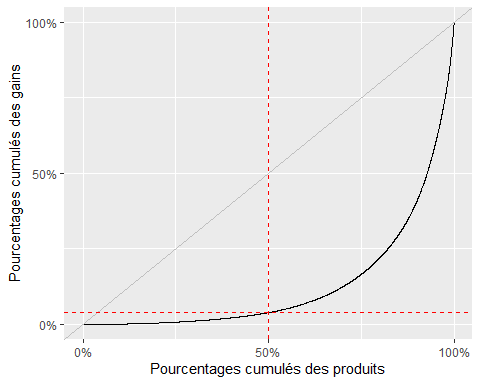


###Je regarde le nombre de livre vendu par jour entre le 14/08 et le01/11  
ggplot(nbre\_date\_categ, aes(date, n)) +  
 geom\_line(aes(colour = categ))+  
 scale\_x\_date(breaks = date\_breaks("month"),   
 limits = c(as.Date("2021-08-14"),as.Date("2021-11-01")),  
 labels = date\_format("%d-%m")) +  
 ylab("Nombre de livres vendus") + xlab("Année 2021 ")

## Warning: Removed 855 rows containing missing values (geom\_path).



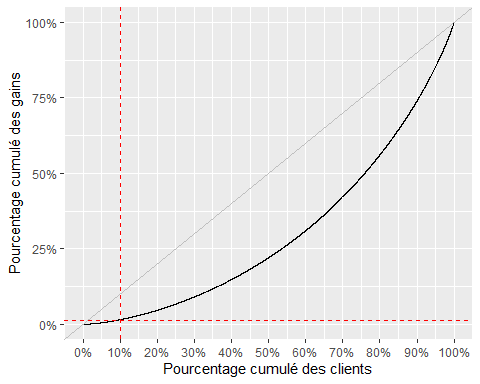
###Je crée une courbe de Lorenz pour étudié la répartition des gains  
###par rapport aux produits vendu  
ggplot(products, aes(n\_price)) +   
 stat\_lorenz() +   
 geom\_abline(color = "grey") +   
 xlab("Pourcentages cumulés des produits") +  
 ylab("Pourcentages cumulés des gains") +  
 scale\_x\_continuous(breaks = round(seq(0, 1, by = 0.5),1),  
 labels = percent) +  
 scale\_y\_continuous(breaks = round(seq(0, 1, by = 0.5),1),  
 labels = percent)+   
 geom\_hline(yintercept=0.04, linetype="dashed", color = "red")+  
 geom\_vline(xintercept=0.5, linetype="dashed", color = "red")



###Je calcul l'indice de Gini de cette courbe  
ineq(products$n\_price,type="Gini")

## [1] 0.7402233

###Je crée une courbe de Lorenz pour étudié la répartition des gains  
###par rapport aux clients  
ggplot(filter(customers,!(client\_id %in% big\_customers)),  
 aes(montant\_total)) +   
 stat\_lorenz() +   
 geom\_abline(color = "grey") +  
 xlab("Pourcentage cumulé des clients") +  
 ylab("Pourcentage cumulé des gains") +  
 scale\_x\_continuous(breaks = round(seq(0, 1, by = 0.1),1),  
 labels = percent) +  
 scale\_y\_continuous(labels = percent)+   
 geom\_hline(yintercept=0.013, linetype="dashed", color = "red")+  
 geom\_vline(xintercept=0.10, linetype="dashed", color = "red")



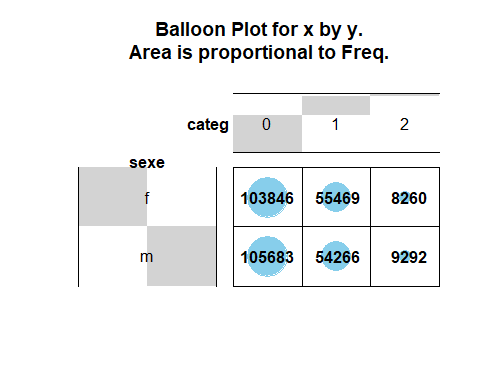
###Je calcul l'indice de Gini de cette courbe  
ineq(customers$montant\_total,type="Gini")

## [1] 0.4397162

###Je crée l'histogramme du nombre de client en fonction de leur âge   
ggplot(customers\_age, aes(x = age, y = n)) +   
 geom\_bar(stat = "identity", width = 0.3) +  
 xlab("age") + ylab("nombre de customers")



###Je réalise un test du chi-2  
  
###Je crée une table de contingence. sexe des clients/catégorie de livre  
contingence\_table<-table(transactions$sex, transactions$categ)  
  
###Je crée un balloonplot  
balloonplot(t(contingence\_table), xlab = "categ", ylab = "sexe",  
 label = TRUE, show.margins = FALSE)



###Je réalise mon test du chi-2  
chisq.test(contingence\_table)

##   
## Pearson's Chi-squared test  
##   
## data: contingence\_table  
## X-squared = 81.733, df = 2, p-value < 2.2e-16

###je calcul le V de cramer (c'est ma statistique de test)  
cramer.v(contingence\_table)

## [1] 0.0155777

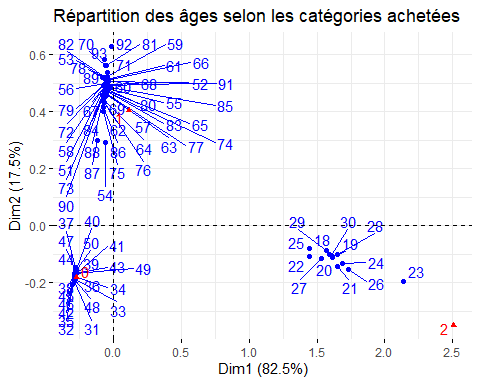
###J'effectue un test ANOVA pour calculer l'effet de l'âge  
###sur les catégories acheté  
res.aov <- aov(age ~ categ , data = transactions)  
summary(res.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## categ 2 7320833 3660416 22713 <2e-16 \*\*\*  
## Residuals 336813 54281447 161   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

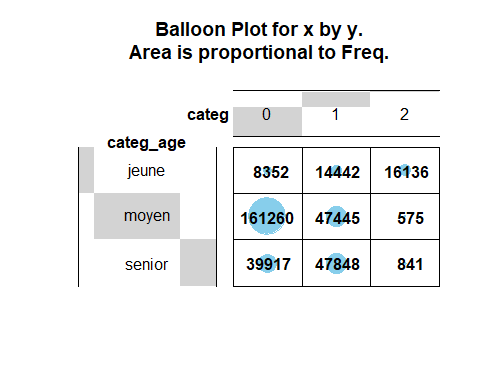
# Effect size  
eta\_squared(res.aov)

## categ   
## 0.1188403

###Je décide de faire une afc pour voire si il est pertinent de regrouper  
###les gens selon des catégories d'âge  
variance\_table<-table(transactions$age, transactions$categ)  
  
res.ca <- CA (variance\_table, graph = FALSE)  
  
fviz\_ca\_biplot (res.ca, repel = TRUE, title =   
 "Répartition des âges selon les catégories achetées")



###Je crée une table de contingence. sexe des clients/catégorie de livre  
contingence\_table<-table(transactions$categ\_age, transactions$categ)  
  
###Je crée un balloonplot  
balloonplot(t(contingence\_table), xlab = "categ", ylab = "categ\_age",  
 label = TRUE, show.margins = FALSE)



variance\_table\_chi2<-  
 table(transactions$categ\_age,transactions$categ)  
  
chisq.test(variance\_table\_chi2)

##   
## Pearson's Chi-squared test  
##   
## data: variance\_table\_chi2  
## X-squared = 152008, df = 4, p-value < 2.2e-16

cramer.v(variance\_table\_chi2)

## [1] 0.4750309

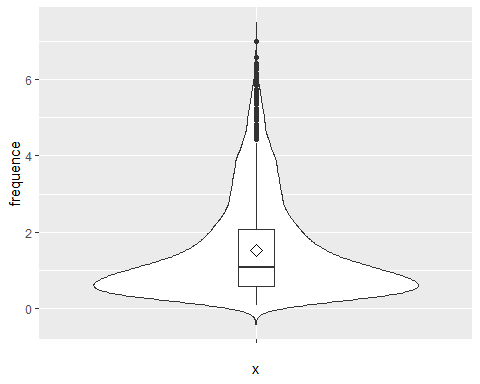
###Je fais un test du chi2 entre la categ d'age et la categ de livre  
###favorite  
variance\_table\_chi2<-  
 table(customers$categ\_age,customers$categ\_favorite)  
  
chisq.test(variance\_table\_chi2)

##   
## Pearson's Chi-squared test  
##   
## data: variance\_table\_chi2  
## X-squared = 12493, df = 4, p-value < 2.2e-16

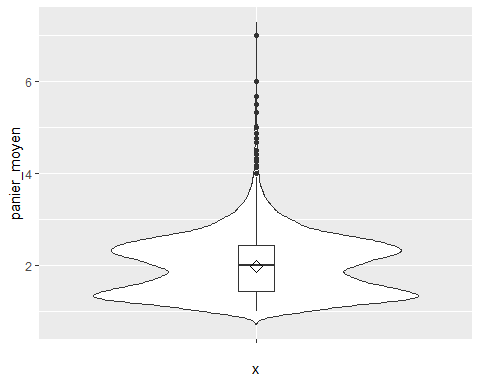
cramer.v(variance\_table\_chi2)

## [1] 0.8522528

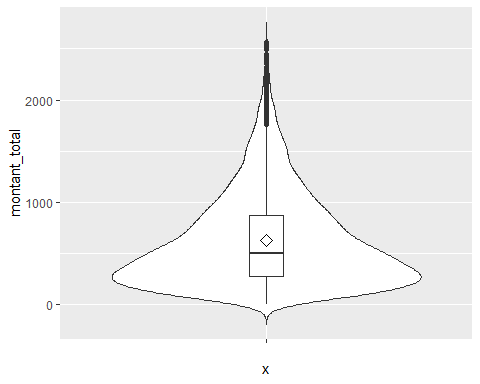
###Analyse de la dispersion des frequence  
ggplot(filter(customers, !(client\_id %in% big\_customers)),  
 aes(x = "",y = frequence))+  
 geom\_violin(trim = FALSE)+  
 geom\_boxplot(width=0.1, fill="white")+  
 stat\_summary(fun.y=mean, geom="point", shape=23, size=3)



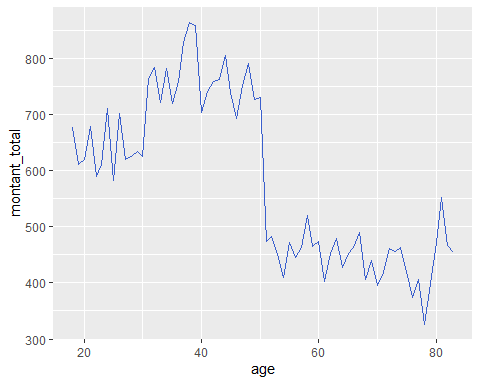
###Analyse de la dispersion des paniers moyens  
ggplot(filter(customers, !(client\_id %in% big\_customers)),  
 aes(x = "",y = panier\_moyen))+  
 geom\_violin(trim = FALSE)+  
 geom\_boxplot(width=0.1, fill="white")+  
 stat\_summary(fun.y=mean, geom="point", shape=23, size=3)



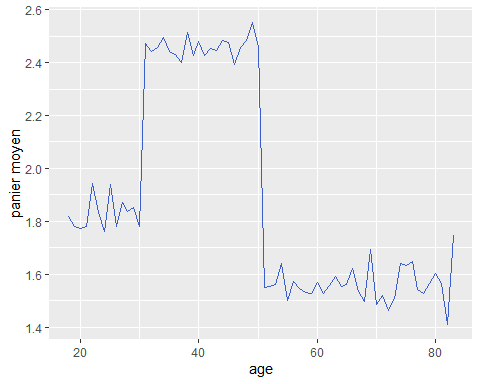
###Analyse de la dispersion des montants totaux  
ggplot(filter(customers, !(client\_id %in% big\_customers)),  
 aes(x = "",y = montant\_total))+  
 geom\_violin(trim = FALSE)+  
 geom\_boxplot(width=0.1, fill="white")+  
 stat\_summary(fun.y=mean, geom="point", shape=23, size=3)



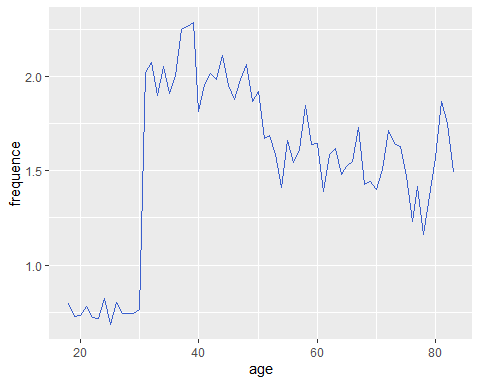
###Au vu de mes analyses je décides de travailler sur ma table  
###où les clients sont regroupés par âge et les variables  
###calculées avec la moyenne.  
###Je perds un peu de précision mais je gagné énormément en lisibilité  
  
###Je trace la courbe de l'âge en fonction du montant\_total  
ggplot(customers\_age, aes(age, montant\_total)) +  
 geom\_line(col = "royalblue3") + xlab("age") + ylab("montant\_total")



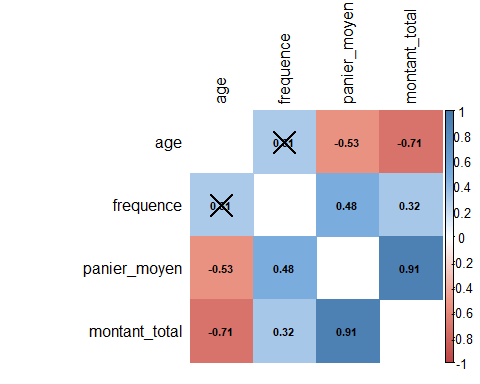
###Je trace la courbe de l'âge en fonction du panier moyen  
ggplot(customers\_age, aes(age, panier\_moyen)) +  
 geom\_line(col = "royalblue3") + xlab("age") + ylab("panier moyen")



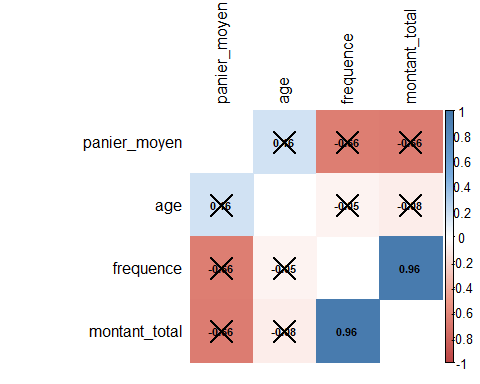
###Je trace la courbe de l'âge en fonction de la fréquence  
ggplot(customers\_age, aes(age, frequence)) +  
 geom\_line(col = "royalblue3") + xlab("age") + ylab("frequence")



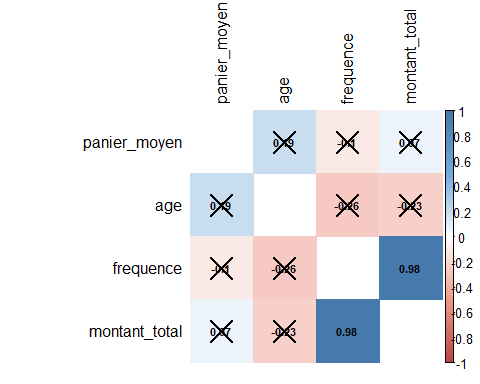
###je réalise la matrice des corrélation pour ces 3 variables par  
###raport à l'âge du client  
  
customers\_cor<-select(customers\_age, -"categ\_age", -"n")  
  
cormat <- cor(customers\_cor, method = "pearson")  
  
p.mat <- cor.mtest(customers\_cor)$p  
  
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD",  
 "#4477AA"))  
  
corrplot(cormat, method = "color", col = col(200),  
 type = "full", order = "FPC", number.cex = .7,  
 addCoef.col = "black", # Add coefficient of correlation  
 tl.col = "black", tl.srt = 90, # Text label color and rotation  
 # Combine with significance  
 p.mat = p.mat, sig.level = 0.01,   
 # hide correlation coefficient on the principal diagonal  
 diag = FALSE)



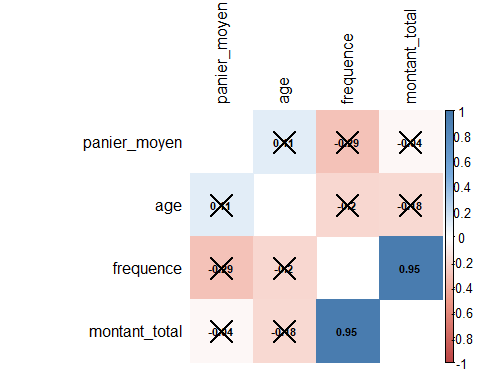
###je réalise la matrice des corrélation pour ces 3 variables par  
###raport à l'âge du client pour les =<30 ans  
  
customers\_cor<-select(customers\_age, -"categ\_age", -"n")%>%  
 filter(age<=30)  
  
cormat <- cor(customers\_cor, method = "pearson")  
  
p.mat <- cor.mtest(customers\_cor)$p  
  
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD",  
 "#4477AA"))  
  
corrplot(cormat, method = "color", col = col(200),  
 type = "full", order = "AOE", number.cex = .7,  
 addCoef.col = "black", # Add coefficient of correlation  
 tl.col = "black", tl.srt = 90, # Text label color and rotation  
 # Combine with significance  
 p.mat = p.mat, sig.level = 0.01,   
 # hide correlation coefficient on the principal diagonal  
 diag = FALSE)



###je réalise la matrice des corrélation pour ces 3 variables par  
###raport à l'âge du client pour les 30-50 ans  
  
customers\_cor<-select(customers\_age, -"categ\_age", -"n")%>%  
 filter(age>30 & age<=50)  
  
cormat <- cor(customers\_cor, method = "pearson")  
  
p.mat <- cor.mtest(customers\_cor)$p  
  
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD",  
 "#4477AA"))  
  
corrplot(cormat, method = "color", col = col(200),  
 type = "full", order = "AOE", number.cex = .7,  
 addCoef.col = "black", # Add coefficient of correlation  
 tl.col = "black", tl.srt = 90, # Text label color and rotation  
 # Combine with significance  
 p.mat = p.mat, sig.level = 0.01,   
 # hide correlation coefficient on the principal diagonal  
 diag = FALSE)



###je réalise la matrice des corrélation pour ces 3 variables par  
###raport à l'âge du client pour les >50 ans  
  
customers\_cor<-select(customers\_age, -"categ\_age", -"n")%>%  
 filter(age>50)  
  
cormat <- cor(customers\_cor, method = "pearson")  
  
p.mat <- cor.mtest(customers\_cor)$p  
  
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD",  
 "#4477AA"))  
  
corrplot(cormat, method = "color", col = col(200),  
 type = "full", order = "AOE", number.cex = .7,  
 addCoef.col = "black", # Add coefficient of correlation  
 tl.col = "black", tl.srt = 90, # Text label color and rotation  
 # Combine with significance  
 p.mat = p.mat, sig.level = 0.01,   
 # hide correlation coefficient on the principal diagonal  
 diag = FALSE)



###J'effectue un test ANOVA pour calculer l'effet des categ d'âge  
###sur le montant total  
res.aov <- aov( categ\_age%>%as.factor()%>%as.numeric()~ montant\_total,  
 data = customers\_age)  
summary(res.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## montant\_total 1 18.08 18.078 53.3 5.97e-10 \*\*\*  
## Residuals 63 21.37 0.339   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Effect size  
eta\_squared(res.aov)

## montant\_total   
## 0.4582948

###J'effectue un test ANOVA pour calculer l'effet des categ d'âge  
###sur le panier moyen  
res.aov <- aov(categ\_age%>%as.factor()%>%as.numeric()   
 ~ panier\_moyen , data = customers\_age)  
summary(res.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## panier\_moyen 1 9.246 9.246 19.29 4.39e-05 \*\*\*  
## Residuals 63 30.200 0.479   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Effect size  
eta\_squared(res.aov)

## panier\_moyen   
## 0.2343986

###J'effectue un test ANOVA pour calculer l'effet des categ d'âge  
###sur la fréquence  
res.aov <- aov(categ\_age%>%as.factor()%>%as.numeric()   
 ~ frequence , data = customers\_age)  
summary(res.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## frequence 1 8.66 8.660 17.72 8.27e-05 \*\*\*  
## Residuals 63 30.79 0.489   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

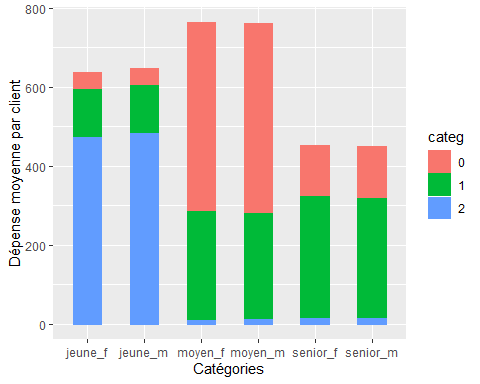
# Effect size  
eta\_squared(res.aov)

## frequence   
## 0.2195515

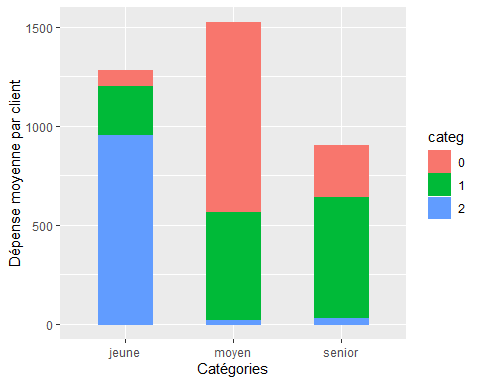
###Je crée un tableau afin de construire des histogrammes  
  
###Je calcul les gains pour chaque catégorie de livre pour chaque  
###catégorie d'âge et sexe  
categ\_age\_sex<-transactions%>%filter(!(client\_id %in% big\_customers))%>%  
 group\_by(categ\_age, sex, categ)%>%  
 summarise(sum\_price = sum(price))%>%  
 transmute(  
 age\_sex = paste(categ\_age, sex, sep="\_"),  
 categ = categ,  
 sum\_price = sum\_price)  
  
x<-transactions%>%  
 group\_by(client\_id, age\_sex)%>%summarise()%>%  
 group\_by(age\_sex)%>%count()  
  
categ\_age\_sex<-inner\_join(categ\_age\_sex, x)

## Joining, by = "age\_sex"

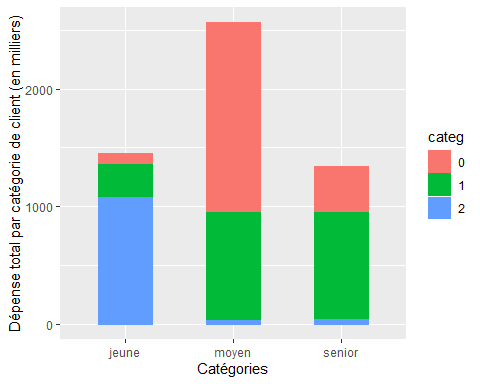
###Je crée une colonne mean pour avoir les moyennes par client pour  
###catégorie  
categ\_age\_sex$mean\_price<-  
 categ\_age\_sex$sum\_price/categ\_age\_sex$n  
  
  
###Je divise par 1000 pour plus de lisibilité  
categ\_age\_sex$sum\_price<-categ\_age\_sex$sum\_price/1000  
  
###Je construit mon histogramme  
ggplot(categ\_age\_sex, aes(x = age\_sex, y = mean\_price)) +   
 geom\_bar(aes(color = categ, fill = categ),  
 stat = "identity", width = 0.5) +  
 xlab("Catégories") +   
 ylab("Dépense moyenne par client")



ggplot(categ\_age\_sex, aes(x = categ\_age, y = mean\_price)) +   
 geom\_bar(aes(color = categ, fill = categ),  
 stat = "identity", width = 0.5) +  
 xlab("Catégories") +   
 ylab("Dépense moyenne par client")



ggplot(categ\_age\_sex, aes(x = categ\_age, y = sum\_price)) +   
 geom\_bar(aes(color = categ, fill = categ),  
 stat = "identity", width = 0.5) +  
 xlab("Catégories") +   
 ylab("Dépense total par catégorie de client (en milliers)")



###J'ai essayé de montrer la corrélation entre la part d'achat  
###dans chaque catégorie et l'âge (ou la categ d'âge)   
###mais je ne sais pas encore trop comment faire  
###Je n'ai pas eu le temps..  
  
###J'effectue un test ANOVA pour calculer l'effet des categ d'âge  
###sur les proportions d'achat par catégorie de livre  
res.aov <- aov(categ\_age%>%as.factor()%>%as.numeric()  
 ~ montant\_2 + montant\_1,  
 data = customers)  
summary(res.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## montant\_2 1 2887.1 2887.1 15925 <2e-16 \*\*\*  
## montant\_1 1 731.3 731.3 4034 <2e-16 \*\*\*  
## Residuals 8597 1558.6 0.2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Effect size  
eta\_squared(res.aov)

## montant\_2 montant\_1   
## 0.5576824 0.1412564