```
# Transformation de la variable quantitative taux de pauvrete en variable qualitative
# selon 4 classes: inferieur à 12, entre 12 et 18, entre 18 et 24 et superieur à 24.
# Respectivement pour: faible, intermediaire, eleve et tres-eleve.

base$taux_de_pauvrete <- ifelse(base$taux_de_pauvrete <
12,"faible",base$taux_de_pauvrete)
base$taux_de_pauvrete <- ifelse(base$taux_de_pauvrete <
18,"intermediaire",base$taux_de_pauvrete)
base$taux_de_pauvrete <- ifelse(base$taux_de_pauvrete <
24,"eleve",base$taux_de_pauvrete)
base$taux_de_pauvrete <- ifelse(base$taux_de_pauvrete < 50,"tres-
eleve",base$taux_de_pauvrete)


base$taux_de_pauvrete <- factor(base$taux_de_pauvrete, levels = c( "faible",
"intermediaire", "eleve", "tres-eleve"))


# Resume globale
head(base)
tail(base)
summary(base)

# Resume variable Quantitative

# 1)Etude de la variable Niveau de vie Median

summary(base$`Mediane_niveau_de_vie_(en euros)`)
var(base$`Mediane_niveau_de_vie_(en euros)`)
n=length(base$`Mediane_niveau_de_vie_(en euros)`)
(n-1)/n*var(base$`Mediane_niveau_de_vie_(en euros)`)
sd(base$`Mediane_niveau_de_vie_(en euros)`)
sqrt((n-1)/n)*sd(base$`Mediane_niveau_de_vie_(en euros)`)
sqrt((n-1)/n)*sd(base$`Mediane_niveau_de_vie_(en
euros)`)/mean(base$`Mediane_niveau_de_vie_(en euros)`)
range(base$`Mediane_niveau_de_vie_(en euros)`)
diff(range(base$`Mediane_niveau_de_vie_(en euros)`))
IQR(base$`Mediane_niveau_de_vie_(en euros)`)
quantile(base$`Mediane_niveau_de_vie_(en euros)`)
quantile(base$`Mediane_niveau_de_vie_(en euros)`,probs=seq(0.1,1,by=0.1))
hist(base$`Mediane_niveau_de_vie_(en euros)`,freq=TRUE,xlab="Mediane revenu
disponible par UC (en euros)",ylab="Densité de fréquence",
    main="Histogramme de la variable Mediane niveau de vie ")
boxplot(base$`Mediane_niveau_de_vie_(en euros)`,ylab="Mediane niveau de vie (en
euros)",
```

```
      main="Boîte-à-moustaches de la variable Mediane nveau de vie ")
plot(ecdf(base$`Mediane_niveau_de_vie_(en euros)`), ylab="F(x)", main="Fonction de
répartition empirique de la mediane du revenu disponible")

median(base$`Mediane_niveau_de_vie_(en euros)`)

#2) Etude la variable Rapport interdecile:

summary(base$Rapport_interdecile)
var((base$Rapport_interdecile))
n=length(base$Rapport_interdecile)
(n-1)/n*var(base$Rapport_interdecile)
sd(base$Rapport_interdecile)
sqrt((n-1)/n)*sd(base$Rapport_interdecile)
(sqrt((n-1)/n)*sd(base$Rapport_interdecile))/mean(base$Rapport_interdecile)
range(base$Rapport_interdecile)
diff(range(base$Rapport_interdecile))
IQR(base$Rapport_interdecile)
quantile(base$Rapport_interdecile)
quantile(base$Rapport_interdecile,probs=seq(0.1,1,by=0.1))

hist(base$Rapport_interdecile,freq=TRUE,xlab="Rapport interdecile",ylab="Densité de
fréquence",
    main="Histogramme de la variable Rapport interdecile")

boxplot(base$Rapport_interdecile,xlab="Rapport interdecile",ylab="Intensité",
      main="Boîte-à-moustaches de la variable Rapport interdecile")
plot(ecdf(base$Rapport_interdecile),xlab="Rapport interdecile",ylab="F(x)",main="Fonction
de répartition empirique du rapport interdecile")
median(base$Rapport_interdecile)

# Resume variable qualitative

# 1) Etude de la variable taux de pauvrete

# Diagramme circulaire

table(base$taux_de_pauvrete)
prop.table(table(base$taux_de_pauvrete))
tableau=table(base$taux_de_pauvrete)
pie(tableau)

# Methode plus logique mais même resultat

pie(prop.table(table(base$taux_de_pauvrete)))
main="Diagramme circulaire du taux de pauvrete"
```

```
# Diagramme tuyaux d'orgue

barplot(table(base$taux_de_pauvrete),
      main="Diagramme en tuyaux d'orgue de la variable taux de pauvrete",
      ylab="effectif", xlab="taux de pauvrete")


# Etude de la liaison entre deux variables quantitatives

#1) Entre mediane_niveau_de_vie_(en euros) et rapport_interdecile

summary(base$`Mediane_niveau_de_vie_(en euros)`)
summary(base$Rapport_interdecile)
# la mediane des niveaux de vie median est de 20093€ tandis que sa moyenne est de 20392€
# la mediane du rapport interdecile est de 3.096 tandis que sa moyenne est de 3.244
# La mediane < moyenne pour les deux series donc il y a une sur-representation des petites
valeurs.
n=length(base$`Mediane_niveau_de_vie_(en euros)`)
(n-1)/n*var(base$`Mediane_niveau_de_vie_(en euros)`)
(n-1)/n*var(base$Rapport_interdecile)
par(mfrow=c(2,2))
hist(base$`Mediane_niveau_de_vie_(en euros)`,freq=FALSE,main="Histogramme du niveau
de vie Median", ylab="densité", xlab="niveau de vie median")
hist(base$Rapport_interdecile,freq=FALSE,main="Histogramme du rapport interdecile",
ylab="densité", xlab="rapport interdecile")
boxplot(base$`Mediane_niveau_de_vie_(en euros)`,main="Boite à moustache du niveau de
vie median", ylab="niveau de vie median")
boxplot(base$Rapport_interdecile,main="Boite à moustache du rapport
interdecile",ylab="rapport interdecile")
# Sur-representation des faibles valeurs pour le rapport interdecile et
# sur-representations des valeurs intermediaires pour le niveau de vie median
par(mfrow=c(1,1)) # Espace graphique rétablie 1/1

#2) Nuages de points

# on choisit le rapport interdecile comme variable à expliquer (=Y, en ordonnées)
# et le niveau de vie median comme variable explicative (=X, en abscisses)

plot(base$`Mediane_niveau_de_vie_(en euros)`,base$Rapport_interdecile,main="Nuage de
points", xlab="niveau de vie median ", ylab="rapport interdecile")
#le nuage de points est concentré avec une legere correlation positive: quand le niveau de
vie median augmente, le rapport interdercile augmente legerement)
# on remarque une dispersion pour les valeurs des niveau de vie median inferieur à 18000
ou superieur à 24000
# onr remarque un point isolé avec un niveau de vie médian très élevé qui peut être un point
influent
```

```
cov(base$`Mediane_niveau_de_vie_(en euros)`,base$Rapport_interdecile)

# cov>0 donc liaison positive entre les deux variables

cor(base$`Mediane_niveau_de_vie_(en euros)`,base$Rapport_interdecile)
# il est positif mais proche de 0.4, il y a donc une faible corrélation linéaire positive entre les deux variables

cov(base$`Mediane_niveau_de_vie_(en
euros)`,base$Rapport_interdecile)/sqrt(var(base$`Mediane_niveau_de_vie_(en
euros)`)*var(base$Rapport_interdecile))
# même valeur

regression=lm(base$Rapport_interdecile~base$`Mediane_niveau_de_vie_(en euros)`)

# equation de la droite de regession (D): rapport_interdecile=
0,000126*mediane_niveau_de_vie+0.681201

#coefficiants
regression$coefficients

#coefficiants arrondis

round(regression$coefficients,3)

#valeurs ajustés

regression$fitted.values

#résidus

regression$residuals

# Test pour le departement 14 Calvados

regression$coefficients[2]*base$`Mediane_niveau_de_vie_(en
euros)`[14]+regression$coefficients[1]
regression$fitted.values[14]

# Verrification réussit

# résidu e_14 chapeau

base$`Mediane_niveau_de_vie_(en euros)`[14]-regression$fitted.values[14]
regression$residuals[14]
summary(regression)
# R2= Multiple R-squared: 0.172
```

```
# même valeur avec
cor(base$`Mediane_niveau_de_vie_(en euros)`,base$Rapport_interdecile)^2

# Verification réussit

# 17% de la variation de la variable rapport interdecile est expliqué par la regression donc,
de la variable niveau de vie median.

# le R2 est assez proche de 0 donc le modèle est plutot de mauvaise qualité


# on refait le nuage de points
plot(base$`Mediane_niveau_de_vie_(en euros)`,base$Rapport_interdecile,main="Nuage de
points", xlab="niveau de vie median",ylab="rapport inter-decile ")

# on ajoute la droite de régression (en rouge)
abline(regression,col="red")

# on ajoute le barycentre du nuage de points = point de coordonnées (x_barre,y_barre) en
bleu
points(mean(base$`Mediane_niveau_de_vie_(en
euros)`),mean(base$Rapport_interdecile),pch="+",col="blue")
mean(base$`Mediane_niveau_de_vie_(en euros)`)
mean(base$Rapport_interdecile)
# la droite de régression passe bien par le barycentre du nuage de points (vu en cours)

## 1) Résidus en fonction de SE
plot(base$`Mediane_niveau_de_vie_(en euros)`,regression$residuals,
    main="Résidus en fonction de niveau de vie median",
    xlab="variable explicative",ylab="résidus")
# il ne doit pas y avoir de liaison entre les résidus et la variable explicative

## 2) Moyenne des résidus
mean(regression$residuals)
# en théorie la moyenne des résidus est nulle, sa valeur est très proche de 0
sum(regression$residuals)
# en théorie la somme des résidus est également nulle, sa valeur est aussi très proche de 0


## 3) Identification de résidus sur le nuage de points

# Il est important de repérer les departements ayant de forts résidus car leur rapport
interdecile
# a été mal prédit par la régression (erreur de saisie ? comportement particulier ? point
# influent = qui a un fort impact dans l'estimation des coef. de la régression ?)
```

```
plot(base$`Mediane_niveau_de_vie_(en euros)`,base$Rapport_interdecile,main="Nuage de points", xlab="niveau de vie median à l'embauche",ylab="rapport inter decile")
abline(regression,col="red")

# identify(base$`Mediane_niveau_de_vie_(en euros)`,base$Rapport_interdecile) # cliquer sur les points dont on veut obtenir l'indice
# puis sur le bouton Finish pour obtenir les valeurs


# graphiquement, le departement ayant le plus fort résidu est le 76
# valeur du résidu de cet individu
regression$residuals[76] # 2.29642
max(regression$residuals) # c'est bien le résidu le plus élevé

# on peut repérer graphiquement que les departements 98,76eme valeur donc departement 75=PARIS,97ème valeur donc departement 972=Martinique,94ème valeur donc dep93 seine-Saint dennis ont de forts résidus
# valeur des résidus de ces individus
regression$residuals[c(98,76,97,94)]


# pour une analyse plus fine, on peut déterminer les plus forts résidus par le calcul :

# les 10 plus forts négatifs
head(sort(regression$residuals),10)

# les 10 plus forts positifs
tail(sort(regression$residuals),10)


# les 10 plus grands en valeur absolue
tail(sort(abs(regression$residuals)),10) # mais on perd le signe

## Etude sans le point de salaire à l'embauche maximum


# valeur du niveau de vie median maximum
max(base$`Mediane_niveau_de_vie_(en euros)`) # 26808

# identification de l'individu
which(base$`Mediane_niveau_de_vie_(en euros)`==max(base$`Mediane_niveau_de_vie_(en euros)`)) # 76ème valeur donc 75=Paris


# On crée de nouveaux vecteurs de données sans cet individu
SE2=base$`Mediane_niveau_de_vie_(en euros)`[-76]
SA2=base$Rapport_interdecile[-76]
```

```
# coefficient de corrélation
cor(SE2,SA2) # 0,2474673 (au lieu de 0,4147707), il est nettement inférieur

# régression
regression2=lm(SA2~SE2)
summary(regression2) # le R2 vaut 0,7605, il est lui aussi légèrement inférieur

abline(regression,col="red")

# ajout de cette nouvelle droite de régression sur le nuage de points
abline(regression2,col="green")
# la droite verte est assez éloigné de la rouge : on  peut donc
# considérer que le point 76, Paris est un point influent


## Etude sans le point de plus fort résidu (numéro 76)

# on pourrait faire cette analyse en excluant tous les individus que l'on a repérés comme
# ayant de forts résidus (numéros : 98,76,97,94)
SE3=base$`Mediane_niveau_de_vie_(en euros)`[-76]
SA3=base$Rapport_interdecile[-76]

cor(SE3,SA3) # 0,2474673 (au lieu de 0,0.4147707), mon departement ayant le plus grand
residu est
# le meme que celui ayant le niveau de vie median le plus elevé, 75 PARIS

#même resultat
regression3=lm(SA3~SE3)
summary(regression3)
# on voit que le R2 a diminué : 6% au lieu de 17%
# les coefficients de régression ont évolué

abline(regression3,col="purple")
# la droite violette est confondue avec la vert, car c'est la même valeur qui est écarté
# donc le point 75 est toujours un point très influent

## 1. Problématique


# on cherche à étudier la liaison entre le niveau de vie median et le taux de pauvrete
# est-ce que le niveau de vie median a une influence sur le taux de pauvrete ? en particulier,
# est-ce que le taux de pauvrete de PARIS 75 est plus faible que celui des autres
departements ?
```

```
##############################################################################
##
##  3. Etude de la liaison entre le salaire et le statut professionnel
##############################################################################
##

# 3.1 Représentation graphique des distributions conditionnelles
#####################################################################

boxplot(base$`Mediane_niveau_de_vie_(en euros)`~base$taux_de_pauvrete,xlab="taux de
pauvrete",ylab="niveau de vie median",
      main="Boites à moustaches juxtaposées des distributions conditionnels /n  du taux de
pauvreté sachant le niveau de vie median")
# on remarque de grandes différences de niveau de vie median selon le taux de pauvrete
# les departements au faible taux de pauvrete ont des revenues median plus élevés que les
autres departements
# les departements au taux de pauvreté elevé ont des niveau de vie median intermédiaires,
avec une très faible dispersion
# alors que les niveau de vie median des trois autres catégories sont assez dispersés


# 3.2 Résumés numériques des distributions conditionnelles
#############################################################

n=length(base$`Mediane_niveau_de_vie_(en euros)`)
# moyennes conditionnelles
mcond=tapply(base$`Mediane_niveau_de_vie_(en euros)`,base$taux_de_pauvrete,mean)
mcond
# la moyenne des niveau de vie median des departements au taux de pauvrete faible est
beaucoup plus élevé que celui des autres departements
# la moyenne du  niveau de vie median des departements au taux de pauvreté elevee é est
légèrement supérieur à celui des departements au taux de pauvrete tres eleve
# il semble donc exister une liaison entre le niveau de vie median et le taux de pauvrete

# on ajoute les moyennes conditionnelles aux b-à-m juxtaposées
points(mcond,col="red",pch="*",cex=2)
# option pch pour changer le marqueur et option cex pour augmenter la taille

# variances conditionnelles (définition de la variance de R)
varcondR=tapply(base$`Mediane_niveau_de_vie_(en euros)`,base$taux_de_pauvrete,var)

# Effectifs n_i
table(base$taux_de_pauvrete)

# variances conditionnelles (sigma_i^2) avec la formule du cours
varcond=(table(base$taux_de_pauvrete)-1)*varcondR/table(base$taux_de_pauvrete)
varcond
```

```
# on utilise les effectifs n_i grâce à table(base$taux_de_pauvrete)

# remarque : résumés numériques des distributions conditionnelles
tapply(base$`Mediane_niveau_de_vie_(en euros)`,base$taux_de_pauvrete,summary)


# 3.3 Calcul du rapport de corrélation
########################################


# variance intra (moyenne pondérée des sigma_i^2)
varintra=sum(table(base$taux_de_pauvrete)*varcond)/n
varintra

# variance inter (d'après l'équation d'analyse de la variance)
vartot=var(base$`Mediane_niveau_de_vie_(en euros)`)
varinter=vartot-varintra
varinter

# calcul direct de varinter
1/n*(mcond-rep(mean(base$`Mediane_niveau_de_vie_(en
euros)`)*sum(table(base$taux_de_pauvrete),3)))^2
# calcul du rapport de corrélation eta2
varinter/vartot

# eta2=0,377657 est assez proche de 0.4. Il existe donc une liaison faible mais quand même
notable entre
# le taux de pauvrete et le niveau de vie median, avec la moyenne des niveau de vie median
# des departements au taux de pauvrete faible qui est environ 50% plus élevée que de la
moyenne de ceux
# au taux de pauvreté tres elevé.
```