Statistique Descriptive

Kossi Tonyi Wobubey ABOTSI

default

```
library(readxl)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Importation des données :

```
# install.packages("readxl")

# Load the readxl package


# Read data from the Excel file
data <- read_excel("données_complètes_9_classes_MEFG_stagiaire_tatistique.xlsx")

#Selection des colonnes utile
data = data  %>% dplyr::select(collège:classe,`taille cm`: gender,sb:pmvpa,time,CA:`CSP mère`)

#Renommage des colonnes
colnames(data)[3:4] = c("taille_cm","weight_kg")
colnames(data)[22:23] = c("CSP_père","CSP_mère")

#Ajout de colonne des IMC
data$IMC_kg_m2 <- data$weight_kg / (data$taille_cm * 10^-2)^2

# Ajout d'une nouvelle colonne "IPS_categorie"
data$IPS_categorie <- ifelse(data$IPS < 89, "Faible",
                             ifelse(data$IPS >= 90 & data$IPS <= 114, "Moyenne", "Élevée"))

# Print the first few rows of the data to verify
head(data)
```

```
## # A tibble: 6 x 25
##   collège classe taille_cm weight_kg   age gender    sb   lpa   mpa   vpa psb
```

```
##   <chr>   <chr>      <dbl>     <dbl> <dbl> <chr>  <dbl> <dbl> <dbl> <dbl> <chr>
## 1 aigle   3P           157        55    15 F       26.3  4.83  21.2  2.33 47,88~
## 2 aigle   3P           178        61    14 M       14    8.33  28    4    25,45~
## 3 aigle   3P           170        75    15 M       20.3  7.33  21.8  5    36,97~
## 4 aigle   3P           153        68    15 F       26.2  7.33  18.7  2.83 47,58~
## 5 aigle   3P           181        95    15 M       12.2 12.3   22.3  6.17 22,12~
## 6 aigle   3P           164        51    15 F       20.5  6.5   20.3  4.83 37,27~
## # i 14 more variables: pla <chr>, pmpa <chr>, pvpa <chr>, mvpa <dbl>,
## #   pmvpa <chr>, time <chr>, CA <dbl>, Activités <chr>, IPS <dbl>,
## #   Géographie <chr>, CSP_père <chr>, CSP_mère <chr>, IMC_kg_m2 <dbl>,
## #   IPS_categorie <chr>
```

Statistique descriptive de la Population

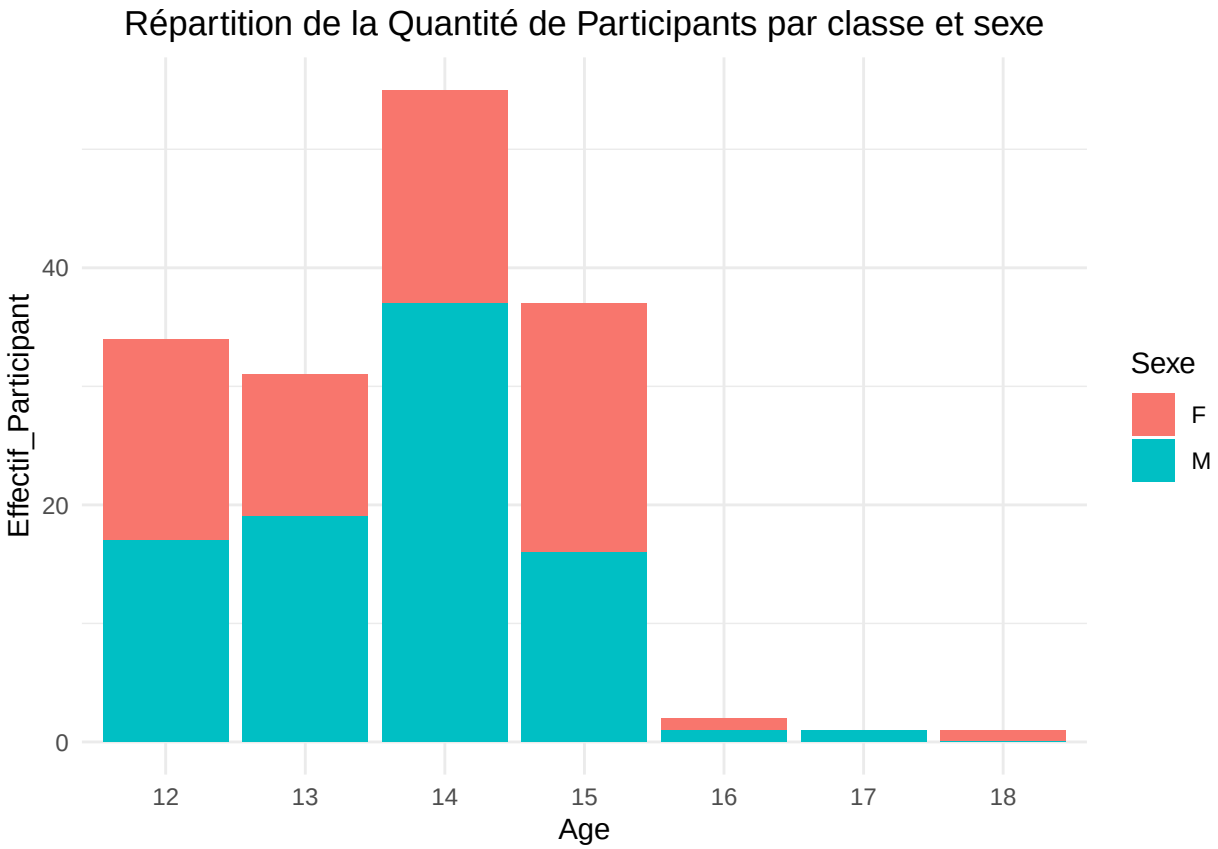
- **Age selon le sexe**

```
data_1 = as.data.frame(table(data$age,data$gender))

names(data_1) = c("age","sexe","Effectif_Participant")

data_1
```

```
##    age sexe Effectif_Participant
## 1   12    F                   17
## 2   13    F                   12
## 3   14    F                   18
## 4   15    F                   21
## 5   16    F                    1
## 6   17    F                    0
## 7   18    F                    1
## 8   12    M                   17
## 9   13    M                   19
## 10  14    M                   37
## 11  15    M                   16
## 12  16    M                    1
## 13  17    M                    1
## 14  18    M                    0
```

Mise en place d'un barplot pour visualiser le nombre de participant par age.

```
# Créer le diagramme en barres empilées
ggplot(data_1, aes(x = age, y = Effectif_Participant, fill = sexe)) +
  geom_bar(stat = "identity") +
  labs(x = "Age", y = "Effectif_Participant", fill = "Sexe", title = "Répartition de la Quantité de Par
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))  # Centrer le titre
```



Dans cette observation, il est noté une prédominance de garçons âgés de 13 et 14 ans par rapport aux filles, tandis qu'il y a pratiquement autant de filles que de garçons âgés de 12 et 16 ans. De plus, il y a davantage de filles de 15 ans, tandis qu'il n'y a que des garçons de 17 ans et seulement des filles de 18 ans. On a une observation total de **161**.

Calculons maintenant l'age moyens des filles et garçons et l'age moyen des participants.

1. **Age moyen des Participants**

```
age_sexe_data=data %>%
  group_by(gender) %>%
  summarise(age_total = sum(age),effectif = n())

age_moyen = sum(age_sexe_data$age_total)/sum(age_sexe_data$effectif)

age_moyen
```

```
## [1] 13.68323
```

Donc l'age moyen des participants est **13.68**.

2. **L'age moyen des filles et garçons**

```
age_sexe_data$age_moyen = age_sexe_data$age_total/age_sexe_data$effectif


age_sexe_data
```

```
## # A tibble: 2 x 4
##   gender age_total effectif age_moyen
```

```
##   <chr>      <dbl>    <int>     <dbl>
## 1 F            961       70      13.7
## 2 M           1242       91      13.6
```

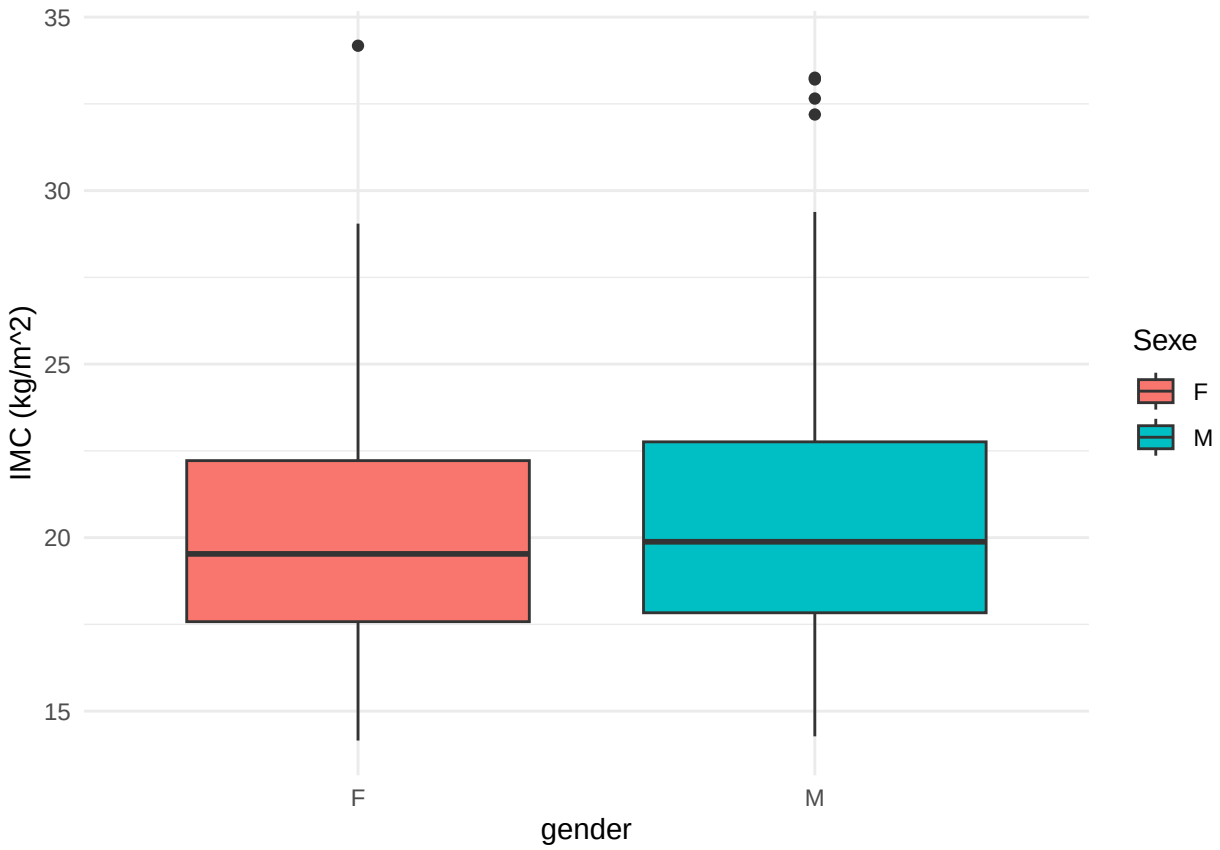
Sexe	F	M	Participant(les deux sexes)
Age Moyen	13.73	13.65	13.68

- **IMC selon le sexe**

```
data %>%
  group_by(gender) %>%
  summarise(IMC_tot = mean(na.omit(IMC_kg_m2)))
```

```
## # A tibble: 2 x 2
##   gender IMC_tot
##   <chr>    <dbl>
## 1 F         20.2
## 2 M         20.8
```

```
# Créer le diagramme en boîte pour l'IMC par classe et sexe
ggplot(data, aes(x = gender, y = IMC_kg_m2, fill = gender)) +
  geom_boxplot() +
  labs(x = "gender", y = "IMC (kg/m^2)", fill = "Sexe") +
  theme_minimal()
```



En moyenne l'IMC des garçons est légèrement plus grand que celui des filles. Récapitulatif dans le tableau suivant :

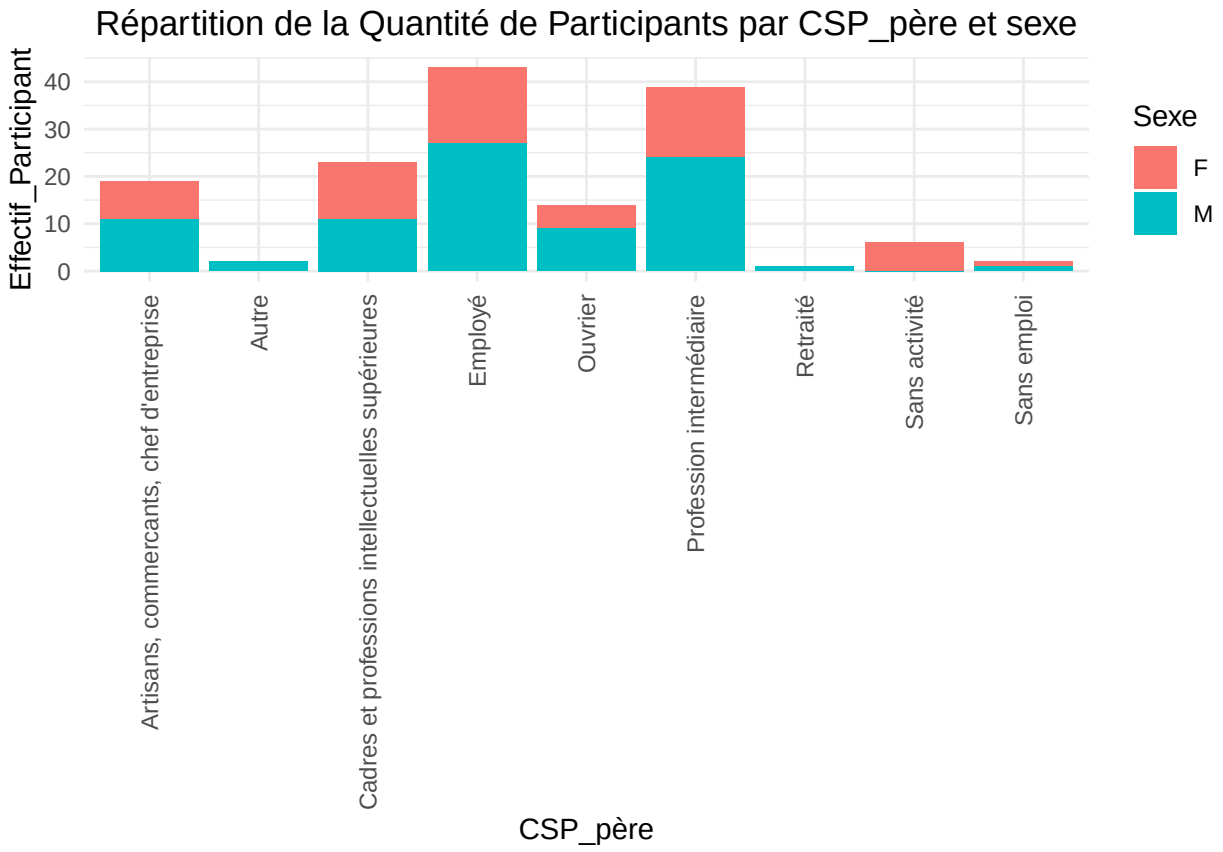
	Population globale	Filles	Garçons
IMC	20.56	20.23	20.83

- **CSP du père**

```
data_1 = as.data.frame(table(data$CSP_père,data$gender))

names(data_1) = c("CSP","sexe","Effectif_Participant")

data_1
```

```
##                                              CSP sexe Effectif_Participant
## 1           Artisans, commercants, chef d'entreprise    F                  8
## 2                                           Autre    F                  0
## 3  Cadres et professions intellectuelles supérieures    F                 12
## 4                                         Employé    F                 16
## 5                                         Ouvrier    F                  5
## 6                          Profession intermédiaire    F                 15
## 7                                        Retraité    F                  0
## 8                                     Sans activité    F                  6
## 9                                       Sans emploi    F                  1
## 10          Artisans, commercants, chef d'entreprise    M                 11
## 11                                          Autre    M                  2
## 12 Cadres et professions intellectuelles supérieures    M                 11
## 13                                        Employé    M                 27
## 14                                        Ouvrier    M                  9
## 15                         Profession intermédiaire    M                 24
## 16                                       Retraité    M                  1
## 17                                    Sans activité    M                  0
## 18                                      Sans emploi    M                  1
```

```
# Créer le diagramme en barres empilées avec les modalités en abscisses affichées verticalement
ggplot(data_1, aes(x = CSP, y = Effectif_Participant, fill = sexe)) +
  geom_bar(stat = "identity") +
  labs(x = "CSP_père", y = "Effectif_Participant", fill = "Sexe", title = "Répartition de la Quantité d
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, vjust = 0.5, hju
```



Avec un total de **149** observation,Il est observé une prédominance des garçons dont le père exerce les professions d'artisans, commerçants, chefs d'entreprise, employés ou ouvriers par rapport aux filles. On constate également qu'il y a presque autant de filles que de garçons dont le père occupe des postes de cadres et professions intellectuelles supérieures, ainsi que chez ceux sans emploi. En revanche, on ne retrouve que des filles parmi les enfants dont le père est sans activité, et exclusivement des garçons parmi ceux dont le père est retraité ou exerce une autre profession.

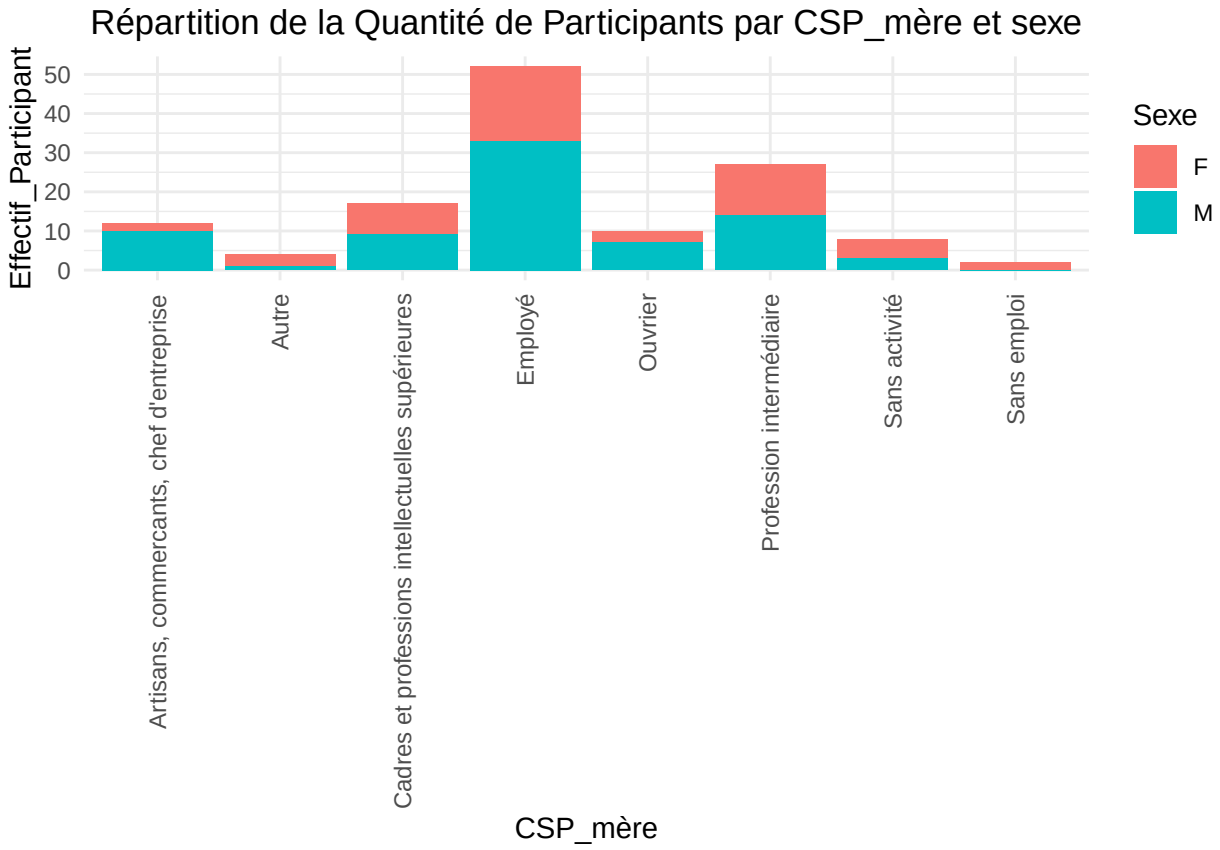
- **CSP de la mère**

```
data_2 = as.data.frame(table(data$CSP_mère,data$gender))
names(data_2) = c("CSP","sexe","Effectif_Participant")

data_2
```

```
##                                          CSP sexe Effectif_Participant
## 1          Artisans, commercants, chef d'entreprise    F                   2
## 2                                         Autre    F                   3
## 3  Cadres et professions intellectuelles supérieures    F                   8
## 4                                       Employé    F                  19
## 5                                       Ouvrier    F                   3
## 6                        Profession intermédiaire    F                  13
## 7                                  Sans activité    F                   5
## 8                                    Sans emploi    F                   2
## 9          Artisans, commercants, chef d'entreprise    M                  10
## 10                                        Autre    M                   1
## 11 Cadres et professions intellectuelles supérieures    M                   9
## 12                                      Employé    M                  33
## 13                                      Ouvrier    M                   7
```

```
## 14                          Profession intermédiaire    M                  14
## 15                                    Sans activité    M                   3
## 16                                      Sans emploi    M                   0
```

```r
# Créer le diagramme en barres empilées avec les modalités en abscisses affichées verticalement
ggplot(data_2, aes(x = CSP, y = Effectif_Participant, fill = sexe)) +
  geom_bar(stat = "identity") +
  labs(x = "CSP_mère", y = "Effectif_Participant", fill = "Sexe", title = "Répartition de la Quantité d
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, vjust = 0.5, hju
```



Sur un échantillon total de **132** observations, il est remarqué que les garçons sont majoritaires lorsque leur mère exerce des professions telles que artisans, commerçants, chefs d'entreprise, employés ou ouvriers, en comparaison avec les filles. D'autre part, il est observé qu'il y a presque autant de filles que de garçons lorsque la mère occupe des postes de cadres et professions intellectuelles supérieures, profession intermédiaire, est sans activité ou occupe d'autres fonctions. En revanche, tous les enfants dont la mère est infirmière et sans emploi sont des filles.

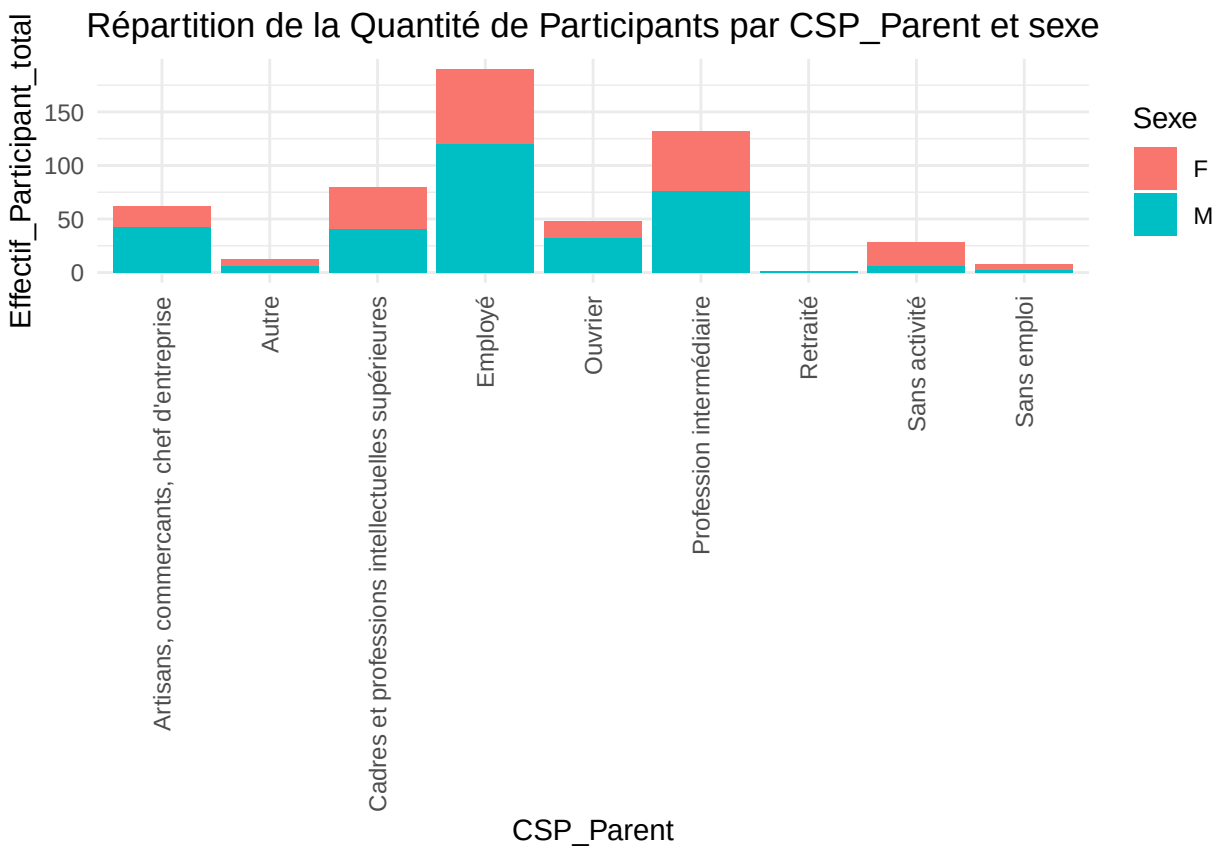
- **CSP des parents**

```r
CSP_data = rbind(data_1,data_2)

colnames(CSP_data)[1] = "CSP_Parent"

#Groupage par IPS du parent et Sexe
CSP_data = CSP_data %>%
  group_by(CSP_Parent,sexe)  %>%
  mutate(Effectif_Participant_total=sum(Effectif_Participant))  %>%
  dplyr::select(-Effectif_Participant)
```

```
CSP_data
```

```
## # A tibble: 34 x 3
## # Groups:   CSP_Parent, sexe [18]
##    CSP_Parent                                          sexe  Effectif_Participant~1
##    <fct>                                               <fct>                  <int>
##  1 Artisans, commercants, chef d'entreprise            F                         10
##  2 Autre                                               F                          3
##  3 Cadres et professions intellectuelles supérieur~ F                         20
##  4 Employé                                             F                         35
##  5 Ouvrier                                             F                          8
##  6 Profession intermédiaire                            F                         28
##  7 Retraité                                            F                          0
##  8 Sans activité                                       F                         11
##  9 Sans emploi                                         F                          3
## 10 Artisans, commercants, chef d'entreprise            M                         21
## # i 24 more rows
## # i abbreviated name: 1: Effectif_Participant_total
```

```
# Créer le diagramme en barres empilées avec les modalités en abscisses affichées verticalement
ggplot(CSP_data, aes(x = CSP_Parent, y = Effectif_Participant_total, fill = sexe)) +
  geom_bar(stat = "identity") +
  labs(x = "CSP_Parent", y = "Effectif_Participant_total", fill = "Sexe", title = "Répartition de la Qua
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, vjust = 0.5, hjus
```



Sur un échantillon total de 149 observations, on constate une prédominance des garçons lorsque l'un de leurs parents exerce les professions telles que artisan, commerçant, chef d'entreprise, employé, ouvrier ou retraité. En revanche, une majorité de filles est observée lorsque l'un des parents est sans activité, sans emploi ou infirmier. Par ailleurs, on remarque une répartition presque égale entre les filles et les garçons lorsque l'un des parents exerce une autre activité, est cadre et une professionnel intellectuelle supérieure ou professionnel intermédiaire.

Statistique descriptive pour le lieu d'étude :

- **Proportion de l'échantillon global de la population selon le CA**

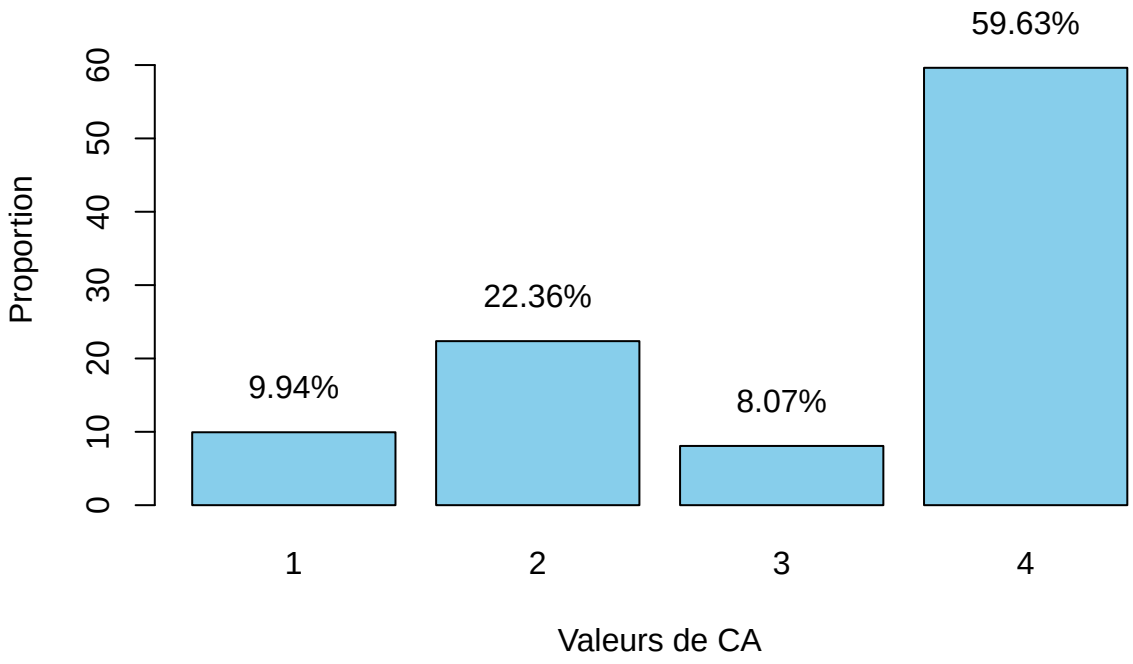
```
prop.table(table(data$CA))*100
```

```
##
##         1         2         3         4
##  9.937888 22.360248  8.074534 59.627329
```

CA	**1**	**2**	**3**	**4**
Proportion(%)	9.938	22.36	8.01	59.63

Illustration avec un barplot :

```
# Calculer les proportions
proportions <- prop.table(table(data$CA)) * 100

# Créer le barplot
barplot(proportions, main = "Proportion des valeurs dans la colonne CA",
        xlab = "Valeurs de CA", ylab = "Proportion", col = "skyblue",
        ylim=c(0, max(proportions) + 10))  # Ajuster ylim pour éviter le chevauchement du texte

# Ajouter les proportions sur les barres
text(x = barplot(proportions, plot = FALSE),  # Obtenir les positions en x des barres
     y = proportions+2,  # Décaler légèrement le texte au-dessus des barres
     labels = sprintf("%.2f%%", proportions),  # Formater les proportions avec deux décimales
     pos = 3)  # Positionner le texte au-dessus des barres (3 = au-dessus)
```



- **Proportion de l'échantillon global de la population selon l'IPS**

-IPS faible inférieur à 89

-IPS moyenne entre 90 et 114

-IPS élevé supérieur à 115

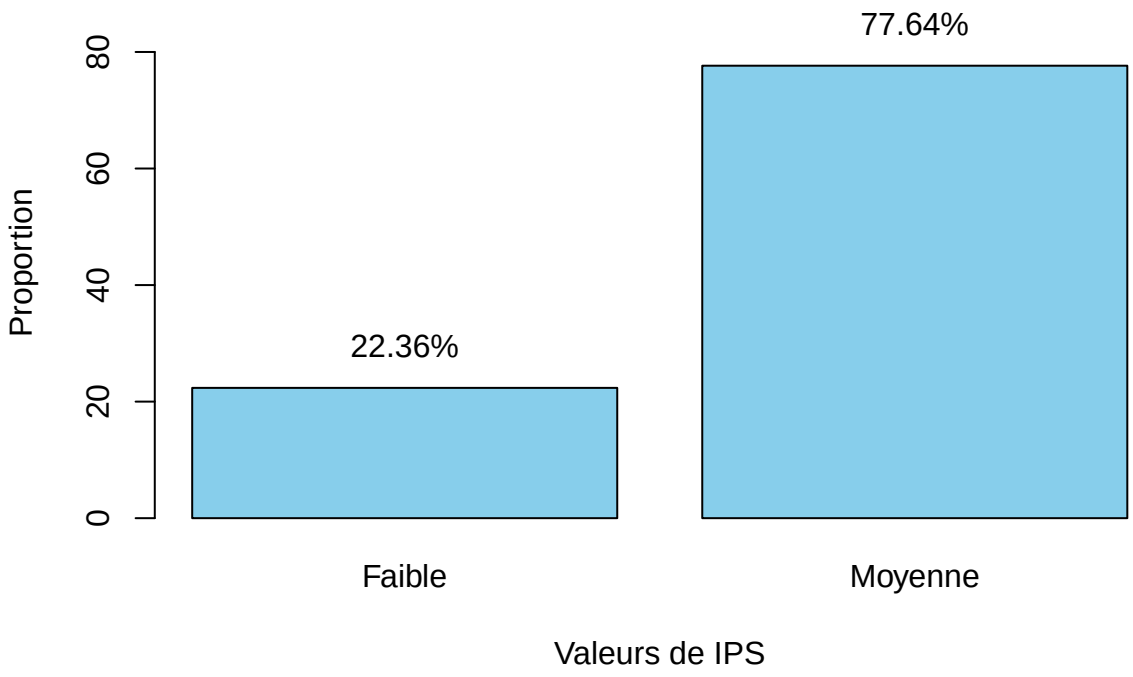
```
prop.table(table(data$IPS_categorie))*100
```

```
##
##   Faible  Moyenne
## 22.36025 77.63975
```

IPS	Faible	Moyenne	Elevé
Proportion(%)	22.36	77.64	0

Illustration avec un barplot

```
# Calculer les proportions pour la colonne IPS_categorie
proportions_ips = prop.table(table(data$IPS_categorie)) * 100

# Créer le barplot pour IPS_categorie
bp <- barplot(proportions_ips, main = "Proportion des valeurs dans la colonne IPS",
              xlab = "Valeurs de IPS", ylab = "Proportion", col = "skyblue",
              ylim = c(0, max(proportions_ips) + 10))  # Ajuster ylim pour éviter le chevauchement du t

# Ajouter les proportions sur les barres
text(x = bp,  # Positions en x des barres, retournées par barplot
     y = proportions_ips + 2,  # Ajouter un petit espace au-dessus de chaque barre pour le texte
     labels = sprintf("%.2f%%", proportions_ips),  # Formater les proportions avec deux décimales
     pos = 3)  # Poser le texte au-dessus des barres
```



- **Proportion de l'échantillon global de la population selon le milieu géographique**

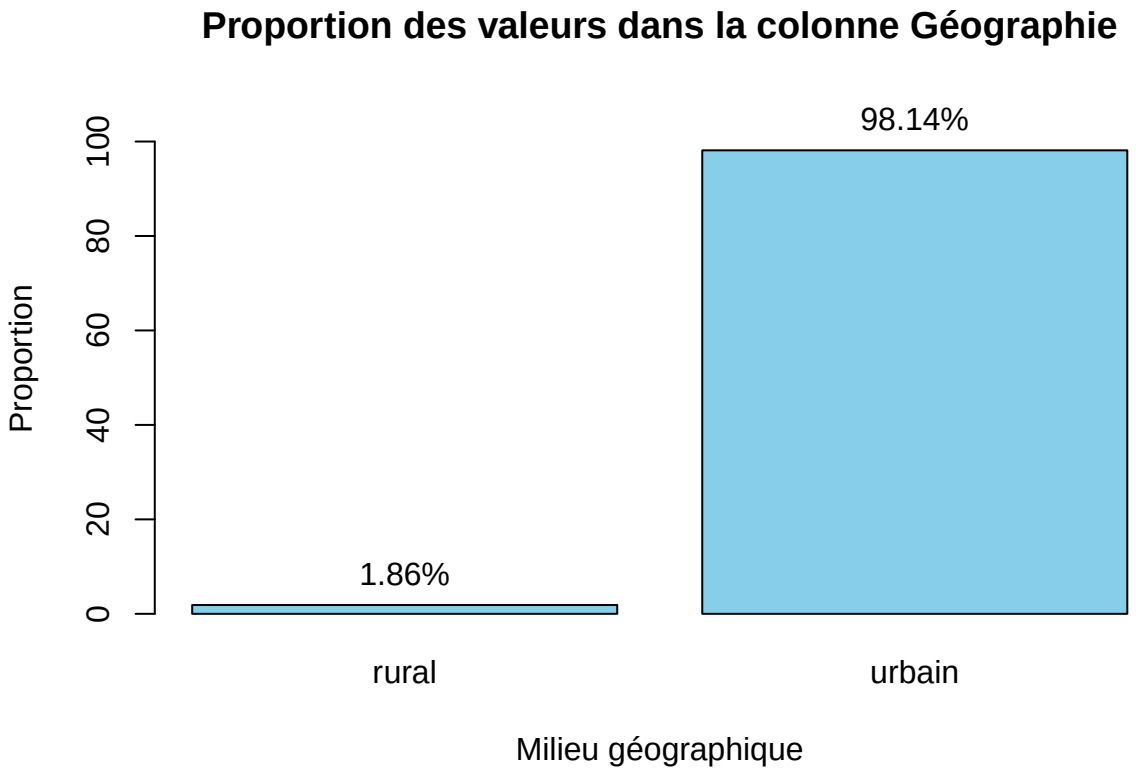
```
prop.table(table(data$Géographie))*100
```

```
##
##     rural    urbain
##  1.863354 98.136646
```

Milieu géographique	**urbain**	**rural**
Proportion(%)	98.14	1.86

Illustration avec un barplot

```
# Calculer les proportions pour la colonne Géographie
proportions_geo = prop.table(table(data$Géographie)) * 100

# Créer le barplot pour Géographie
bp_geo <- barplot(proportions_geo, main = "Proportion des valeurs dans la colonne Géographie",
                  xlab = "Milieu géographique", ylab = "Proportion", col = "skyblue",
                  ylim = c(0, max(proportions_geo) + 10))  # Ajuster ylim pour éviter le chevauchement

# Ajouter les proportions sur les barres
text(x = bp_geo,  # Positions en x des barres, retournées par barplot
     y = proportions_geo + 0.25,  # Ajouter un petit espace au-dessus de chaque barre pour le texte
     labels = sprintf("%.2f%%", proportions_geo),  # Formater les proportions avec deux décimales
     pos = 3)  # Poser le texte au-dessus des barres
```


Proportion des valeurs dans la colonne Géographie
98.14%
1.86%
Proportion
0
20
40
60
80
100
rural
urbain
Milieu géographique