Session3: Visualisation avec ggplot

Visseho Adjiwanou, PhD.

SICSS - Montréal

09 June 2021

Plan

- Introduction
- Type de graphiques pour les distributions univariées et bivariées
- Présentation de ggplot de tidyverse
- Visualisation de distribution univariée
- Visualisation de distribution bivariée
- Remarques
- Ressources

Introduction

Introduction

- R dispose de plusieurs systèmes pour créer des graphiques, mais ggplot2 est l'un des plus élégants et des plus polyvalents.
- Avec ggplot2, vous pouvez faire plus rapidement en apprenant un système et en l'appliquant à de nombreux graphiques.
- Parce qu'il fait partie de `tidyverse`:
- Il sera chargé automatiquement une fois que vous chargez tidyverse;
- Il va fonctionner sur les bases de données ou les `tribbles`

Introduction

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6
```

```
## Warning: package 'tibble' was built under R version 3.6.
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

Introduction

- Les graphiques nous permettent de répondre à plusieurs types de questions :
  - Quelle est la distribution d'une variable?
  - Est-ce que les filles ont plus tendances à vivre dans un type particulier de structure familiale?
  - Comment est-ce que la structure de la famille affecte la santé des enfants?
  - Est-ce qu'il existe une association entre les attitudes envers la violence conjugale et le niveau de scolarisation (données dhs_ipv)
  - Cette relation est-elle positive? négative? ou nulle?

Type de graphiques pour les distributions univariées et bivariées

Les types de graphiques

- Dépend en général du type de variable (qualitative ou quantitative) et du nombre de variable
- Graphiques pour representer une seule variable:

Type de variables	Une seule variable
Qualitative	Diagramme de barre (diagramme en bâton)
	Diagramme circulaire
	Carte (map)
Quantitative	Histogramme (geom_histogram)
	Diagramme de quartile (boîte à moustaches)

Les types de graphiques

- Graphiques pour representer l'association entre deux variables

		Variable dépendante	
	Type de variables	Qualitative	Quantitative
Variable indépendante	Qualitative	Diagramme en bâtons divisés geom_bar	Diagramme de quartile ou boîte à moustaches geom_boxplot
	Quantitative	Transformer la variable en qualitative	Nuage de points geom_point

ggplot

Forme générale

- La forme générale d'un code de graphique est le suivant:

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

1. **ggplot** spécifie que vous utiliser la commande ggplot. C'est à ce niveau que vous spécifier les données que vous voulez utiliser.

- Ce n'est pas toujours obligatoire si vous utilisez plus d'une base de données.

Forme générale

- La forme générale d'un code de graphique est le suivant:

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

2. **geom_function**, contient plusieurs fonctions pour spécifier le type de graphique que vous voulez faire. Le type de graphique indique le nombre de paramètres à inclure.

- Exemples: geom_histogram() pour les **histogrammes**
- geom_point() pour les **diagrammes de dispertions**,
- geom_barplot() pour les **diagrammes de barre**.
- La liste complète est ici: https://ggplot2.tidyverse.org/reference/

Forme générale

- La forme générale d'un code de graphique est le suivant:

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

3. **aes** pour aestetics indique le nombre de paramètres à passer à la fonction **geom_function**. Il permet également de spécifier des informations sur le graphique.

Exemples: Visualiser la distribution univariée

Introduction

- Nous allons utiliser les données dhs_ipv

```
library(tidyverse)

dhs_ipv <- read_csv("dhs_ipv.csv")
dhs_ipv <-
  dhs_ipv %>%
  mutate(beat_burnfood_cat = factor(ntile(beat_burnfood, 4)
                                    labels = c('très faible
         beat_goesout_cat = factor(ntile(beat_goesout, 4),
                                   labels = c('très faible
         sec_school_cat = factor(ntile(sec_school, 3),
                                 labels = c('pauvre', 'moye
         no_media_cat = factor(ntile(no_media, 3),
                               labels = c('riche', 'moyen'
```

Introduction

- et les données crsc_small, que vous connaissez bien.

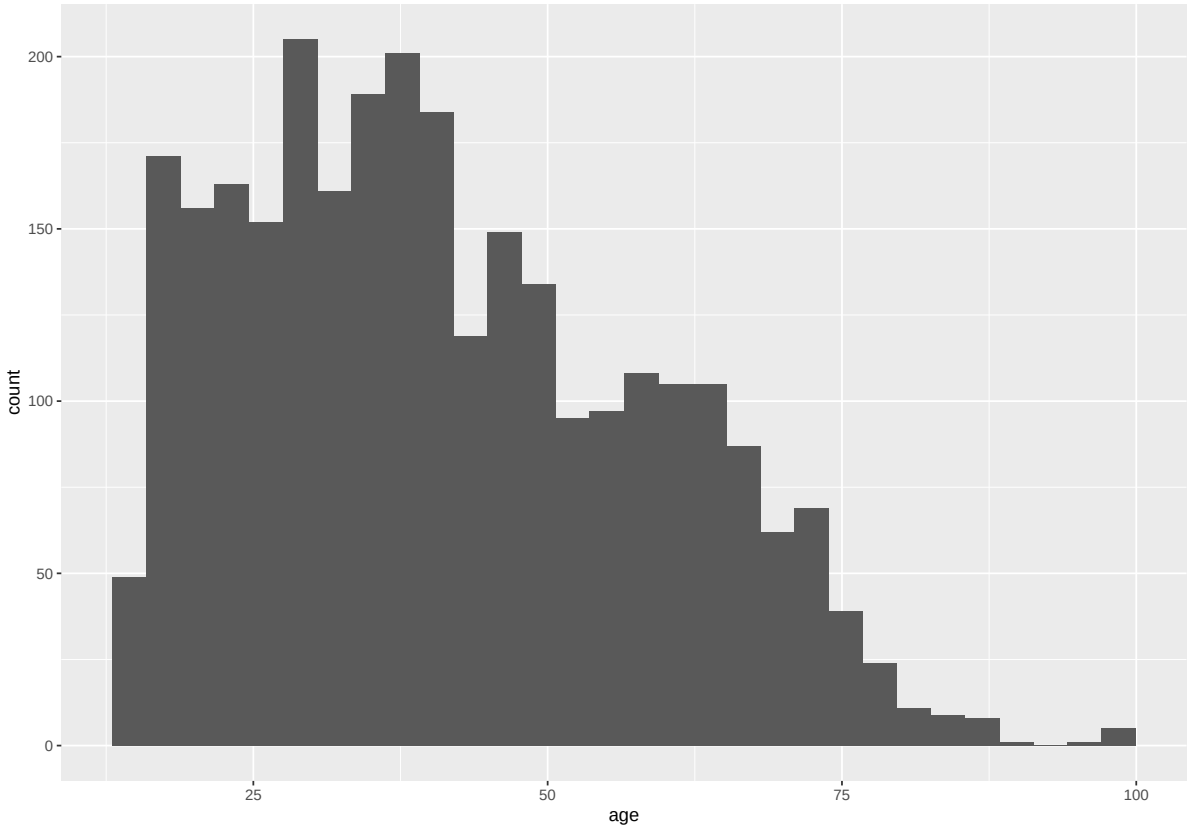
```
crsc96 <- read_csv("cora-crsc1996-E-1996_F1.csv")
crsc96_small <-
  crsc96 %>%
  select(sexq, region, age, ageq, q1, q2, q3, q4, q44, q95)
  mutate(sexe = factor(sexq, labels = c("Homme", "Femme")),
         q2_new = factor(q2,
                         labels = c("totally agree", "agree
                                    "DK/NA", "disagree some
```

Distribution univariée

- Histogramme : pour variable continue
- Diagramme de barre : pour variable catégorielle
- Diagramme de quartile qui résume cinq indicateurs
- Diagramme circulaire

Exemple: histogramme

```
ggplot(crsc96_small) +
  geom_histogram(aes(x = age))
```



Exemple: histogramme

- L'histogramme est une méthode courante pour visualiser la distribution d'une variable numérique plutôt que d'une variable factorielle.
- Un histogramme divise les données en champs
- L'aire de chaque domaine représente la proportion d'observations qui y sont classées.
- La hauteur de chaque case représente la densité, qui est égale à la proportion d'observations dans chaque case divisée par la largeur de la case.
- Un histogramme se rapproche de la distribution d'une variable.

Exemple: histogramme

- Dans le cadre de cette présentation, je mets des options dans le **chunk** (Vous ne les voyez pas dans la présentation regarder le fichier .rmd)
- **out.width** pour préciser la largeur du graphique
- **message = FALSE** : pour ne pas afficher des messages
- **warning = FALSe** : pour ne pas afficher des messages d'avertissement.
- Il faut les utiliser avec précaution. Les messages et les warning nous donne des informations ar exemple sur les valeurs manquantes.

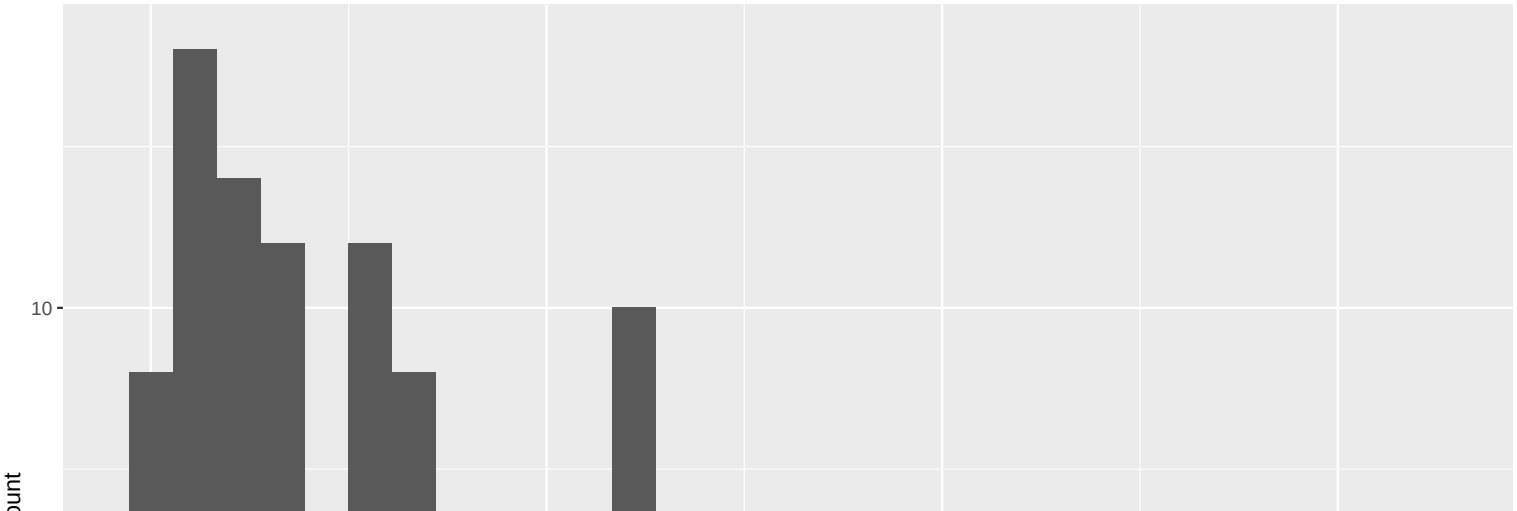
Exemple: histogramme

- Voici ce que vous obtenez si je ne les mets pas.

```
ggplot(dhs_ipv) +
  geom_histogram(aes(x = beat_burnfood))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `
```

```
## Warning: Removed 31 rows containing non-finite values (s
```

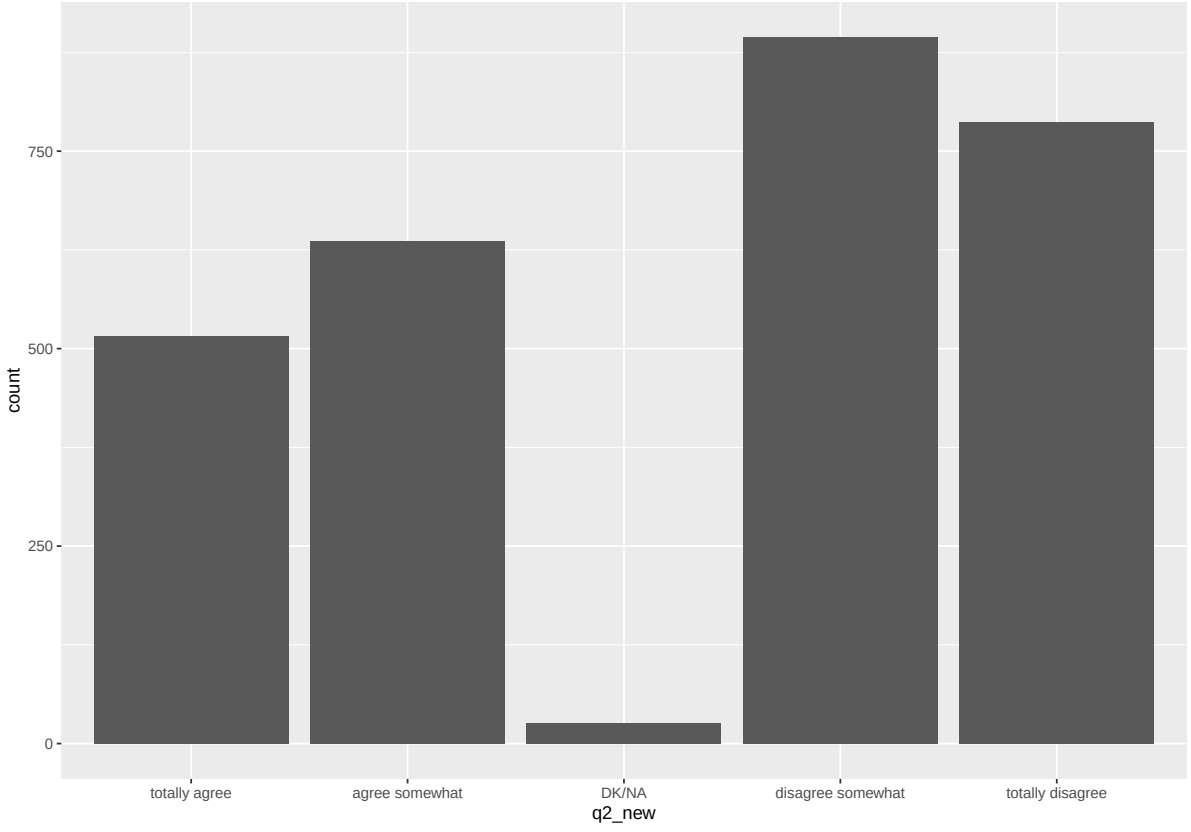


Exemple: Diagramme en bâtons ou à barres

- Pour résumer la distribution d'une variable **facteur** ou d'une **variable factorielle** (ou d'une variable catégorielle ou qualitative) avec plusieurs catégories, un simple tableau avec des comptes ou des proportions est souvent suffisant.
- Cependant, il est également possible d'utiliser un graphique en barres pour visualiser la distribution.

Exemple: Diagramme en bâtons

```
ggplot(crsc96_small) +
  geom_bar(aes(x = q2_new))
```

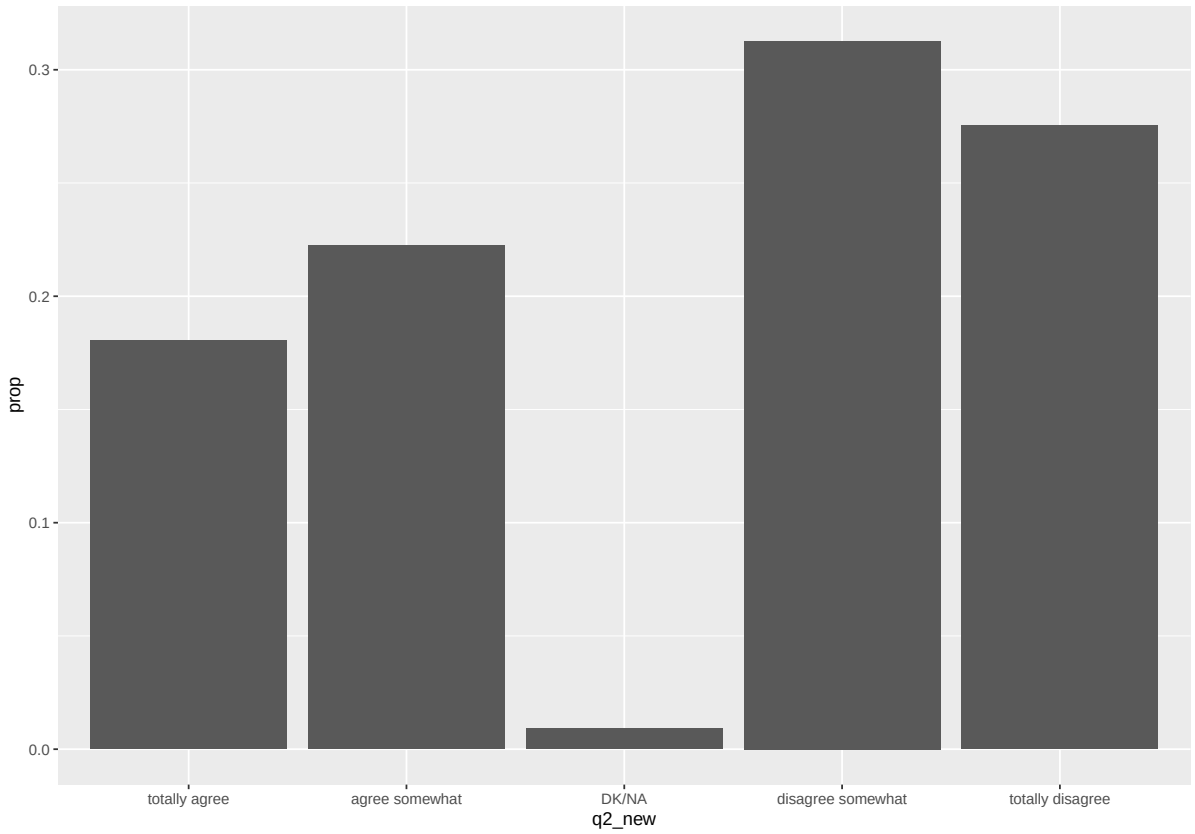


Exemple: Diagramme en bâtons

- Dans les graphiques précédents, l'ordonnée (y) est indiqué en effectif. Ceci pose un problème de comparaison entre différents échantillons.
- Dans ce cas, il faut plutôt utiliser les proportions.

Exemple: Diagramme en bâtons

```
ggplot(crsc96_small) +
  geom_bar(aes(x = q2_new, y = ..prop.., group = 1))
```



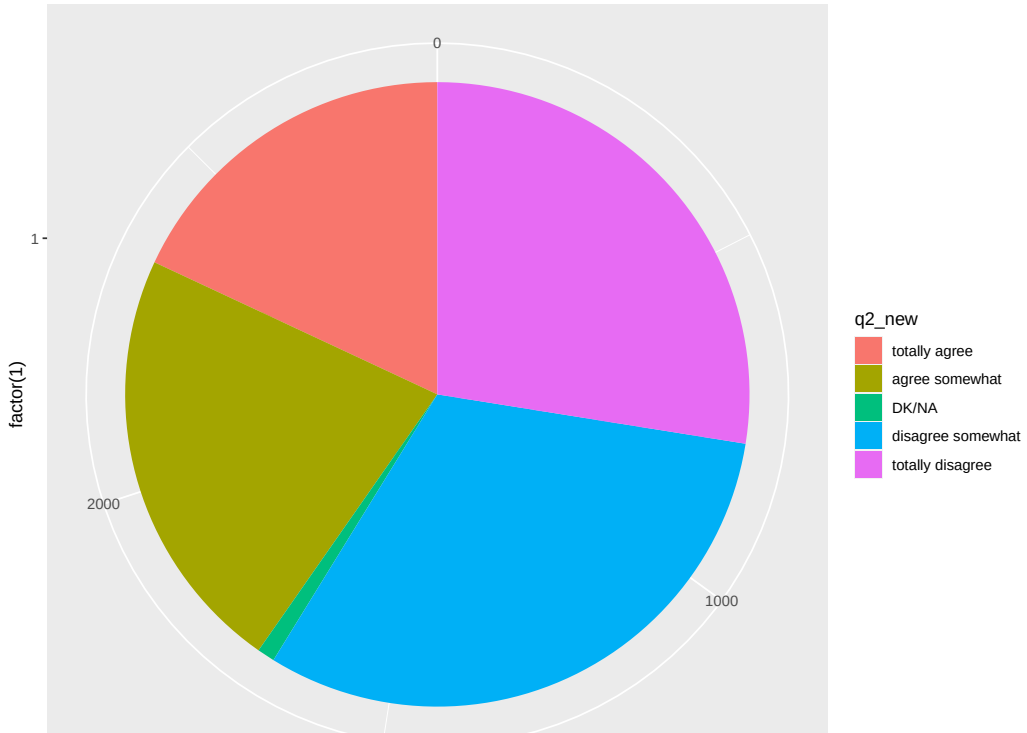
Exemple: Diagramme en bâtons

- Que se passe-t-il si vous ne mettez pas **group = 1**
- On peut représenter ce diagramme par un diagramme circulaire. comment créer un diagramme circulaire?

Diagramme circulaire

https://ggplot2.tidyverse.org/reference/coord_polar.html

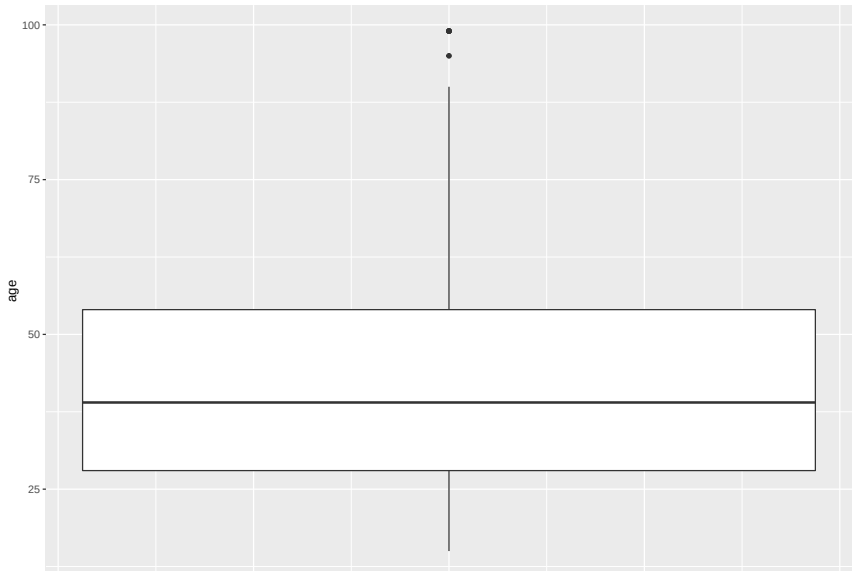
```
ggplot(crsc96_small) +
  geom_bar(aes(x = factor(1), fill = q2_new), width = 1) +
  coord_polar("y", start = 0)
```



Exemple: Diagramme de quartile

- La boîte à moustaches représente un autre moyen de visualiser les distributions d'une variable numérique.
- Une boîte à moustaches visualise **la médiane**, **les quartiles** et **l'écart-interquartile** sous la forme d'un seul objet.

```
ggplot(crsc96_small) +
  geom_boxplot(aes(y = age))
```



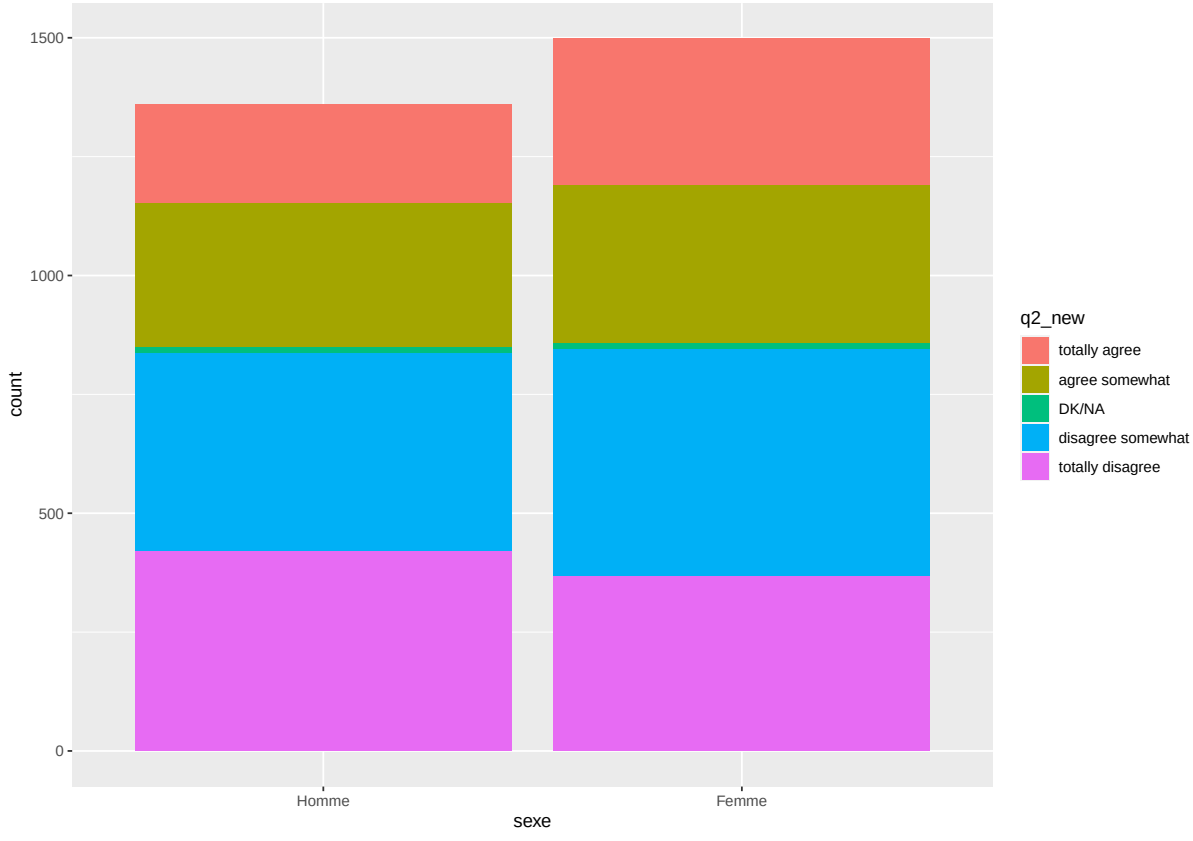
Exemple: Diagramme de quartile ou boîte à moustaches

- C'est particulièrement utile lorsque vous **comparez la distribution de plusieurs variables** en les plaçant côte à côte.
- `geom_boxplot` permet de représenter des boîtes à moustaches. On lui passe en **y** (axe des ordonnées) la variable dont on veut étudier la répartition (variable quantitative), et en **x** (axe des abscisses) la variable contenant les classes qu'on souhaite comparer (variable qualitative).

Exemples: Visualiser la distribution bivariée

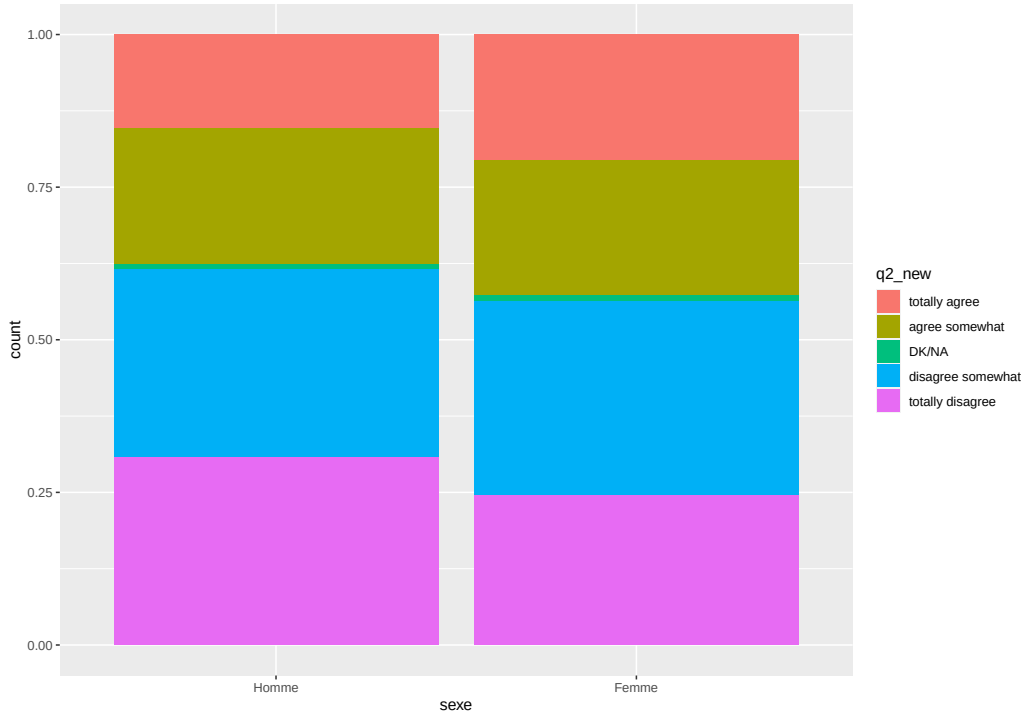
Croisement de deux variables qualitatives

```
ggplot(crsc96_small) +
  geom_bar(aes(x = sexe, fill = q2_new))
```



Croisement de deux variables qualitatives

```
ggplot(crsc96_small) +
  geom_bar(aes(x = sexe, fill = q2_new), position = "fill")
```



- On voit clairement la différence d'opinion entres les hommes et les femmes

Croisement de deux variables qualitatives

- On peut changer les couleurs, on verra cela plus loin.
- http://www.sthda.com/french/wiki/couleurs-dans-r

```
ggplot(crsc96_small) +
  geom_bar(aes(x = sexe, fill = q2_new), position = "fill") +
  scale_fill_brewer(palette="PRGn")
```

Croisement d'une variable quantitative et d'une variable qualitative

- Croiser une variable quantitative et une variable qualitative, c'est essayé de voir si les valeurs de la variable quantitative se répartissent différemment selon les catégories d'appartenance de la variable qualitative.

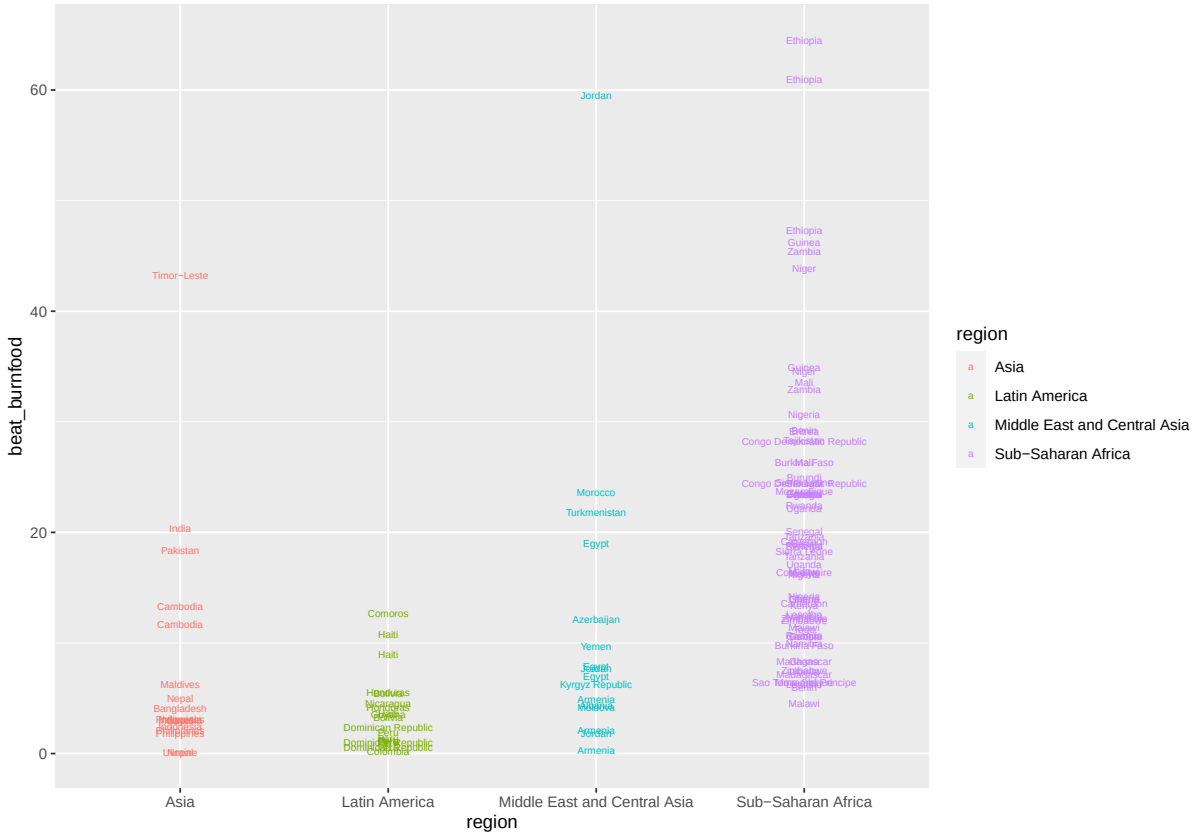
Croisement d'une variable quantitative et d'une variable qualitative

- Avant de présenter ce diagramme, regardons un peu la distribution de la variable beat_burnfood par région.

```
a <-
ggplot(dhs_ipv) +
  geom_text(aes(x = region, y = beat_burnfood,
                label = country, color = region), size = 2)
```

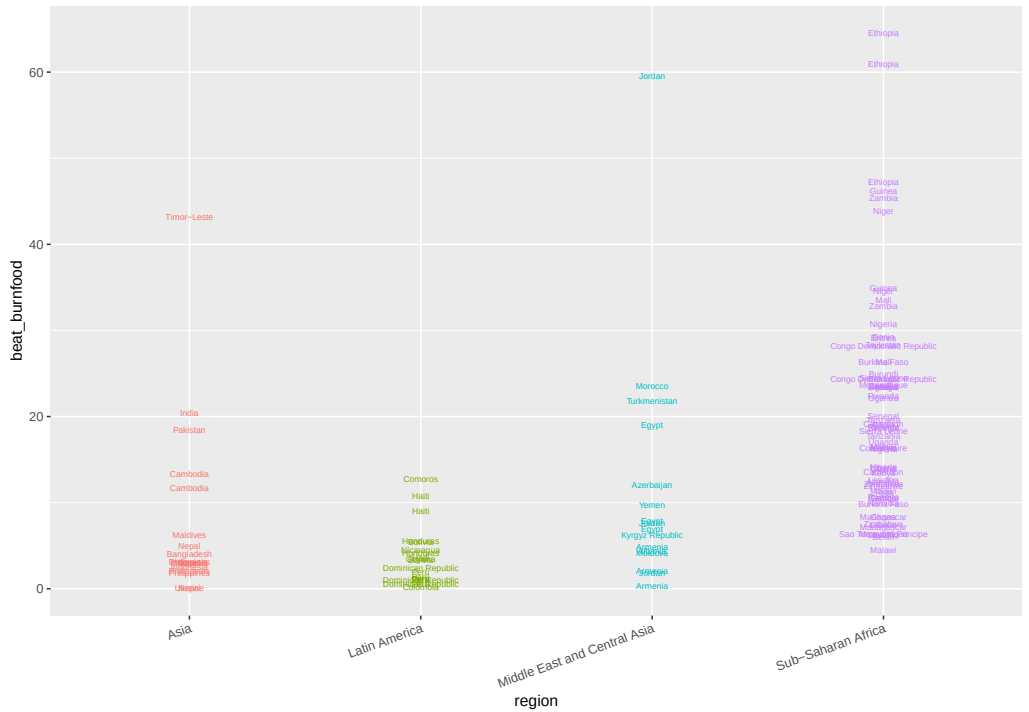
Croisement d'une variable quantitative et d'une variable qualitative

a



Croisement d'une variable quantitative et d'une variable qualitative

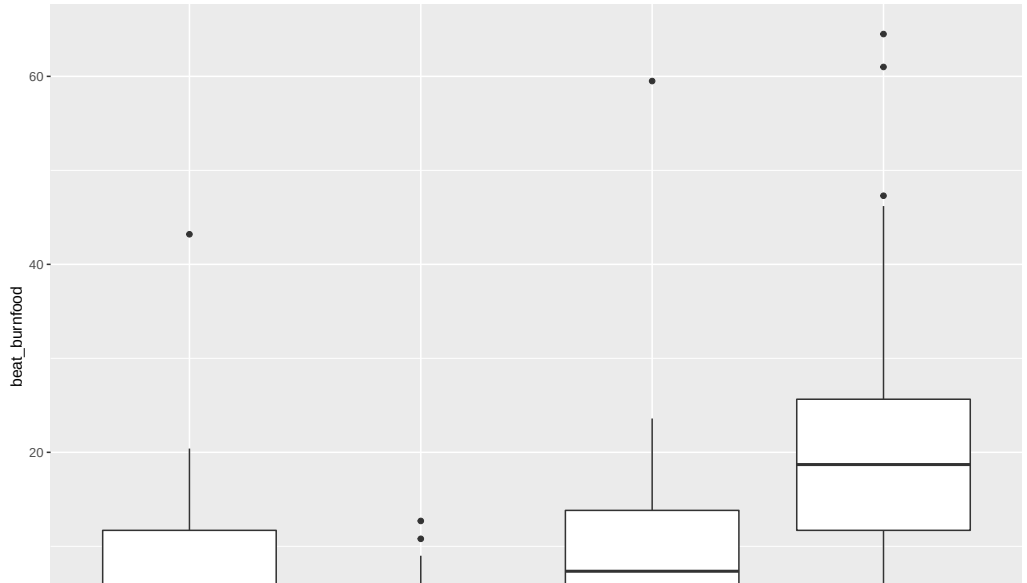
```
a +
  theme(axis.text.x = element_text(angle = 20, hjust = 1),
        legend.position = "none")
```



Exemple: Diagramme de quartile

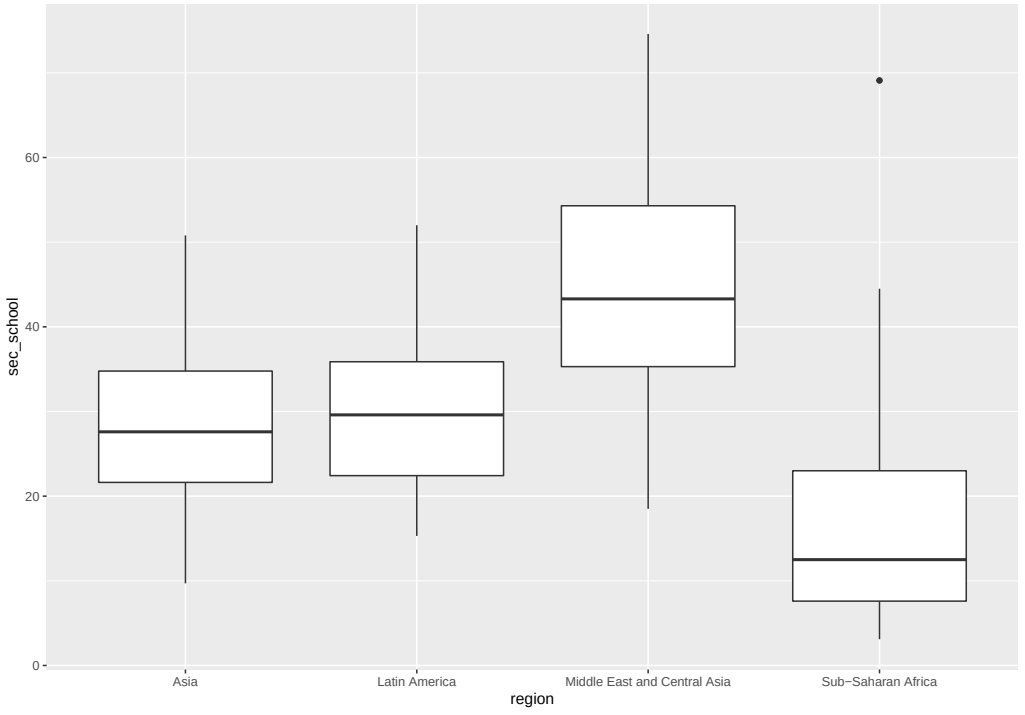
- Le diagramme de quartile permet de synthétiser l'information contenue dans ce nuage de point pour une comparaison plus efficiente.

```
b <- ggplot(dhs_ipv) +
  geom_boxplot(aes(x = region, y = beat_burnfood))
b
```



Exemple: Diagramme de quartile

```
c <- ggplot(dhs_ipv) +
  geom_boxplot(aes(x = region, y = sec_school))
c
```

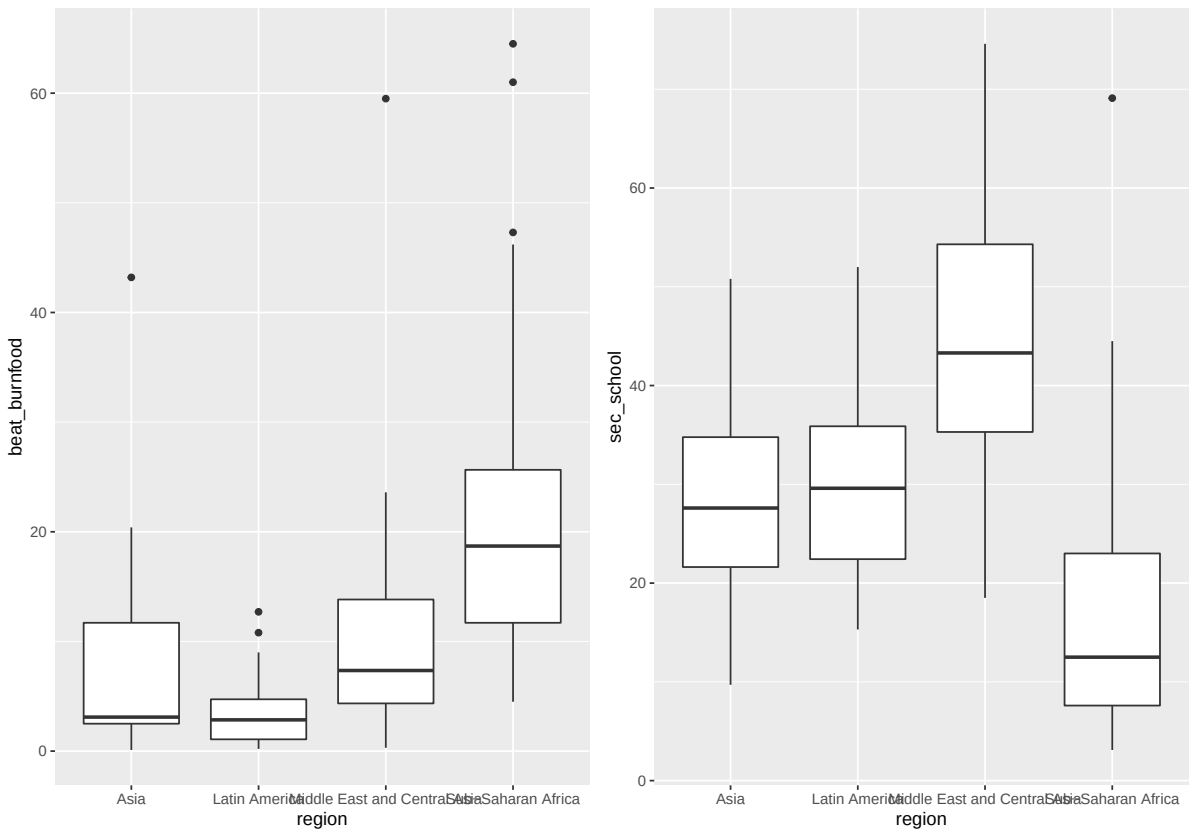


Exemple: Diagramme de quartile

- Commencez-vous par tirer une petite conclusion ici?
- Pour bien visualiser le tout, il faut les mettre dans un même graphique. La commande **ggarrange** du package **ggpubr** vous permet de faire cela.

Exemple: Diagramme de quartile

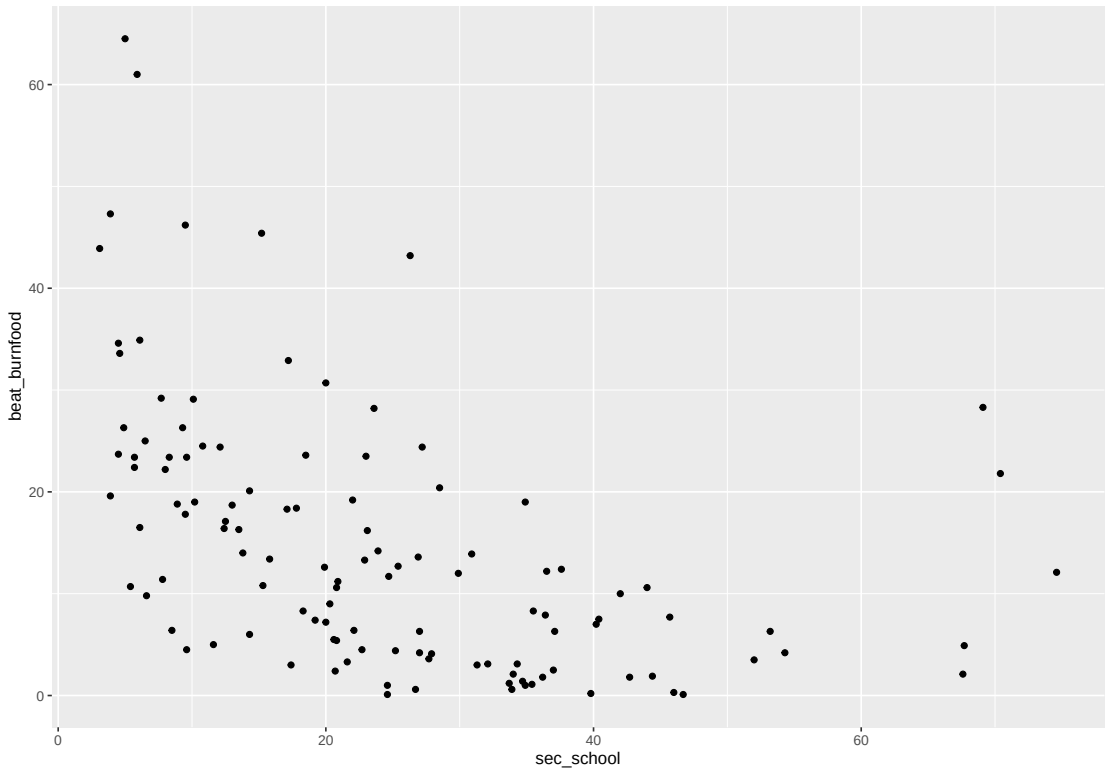
```
#install.packages("ggpubr")
library(ggpubr)
ggarrange(b, c, ncol = 2)
```



Association entre deux variables quantitatives

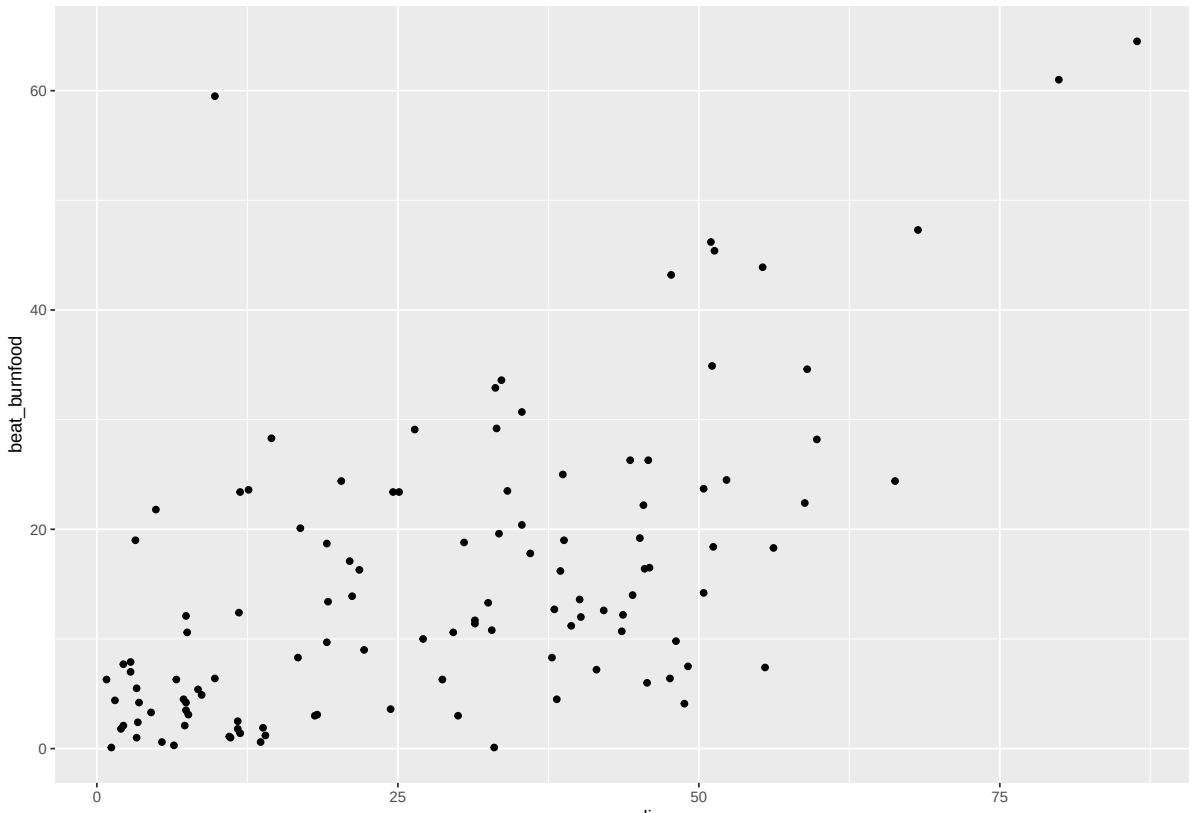
Corrélation linéaire : Croisement de deux variables quantitatives

```
ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood))
```



Corrélation linéaire : Croisement de deux variables quantitatives

```
ggplot(dhs_ipv) +
  geom_point(aes(x = no_media, y = beat_burnfood))
```



Remarques

Remarques

- Les annotations graphiques sont très utiles pour mettre en évidence les messages clés.
- Dans un bulletin ou un rapport statistique, tous les graphiques doivent être étiquetés comme des figures et numérotés, en fonction de leur ordre d'apparition.
- Ecrire clairement les titres : préciser la région et la période.
- Soyez concis, en nommant les principaux axes du graphique.
- Le texte du graphique doit être horizontal.
- Si les étiquettes ne tiennent pas dans l'espace requis, transposez le graphique ou convertissez les unités.
- Elles doivent être concises et pertinentes.
- Placez les sur le graphique aussi près que possible des points de données qui vous intéressent.
- Indiquer la source

Remarques

https://slideplayer.fr/slide/10114066/

Pour aler plus loin

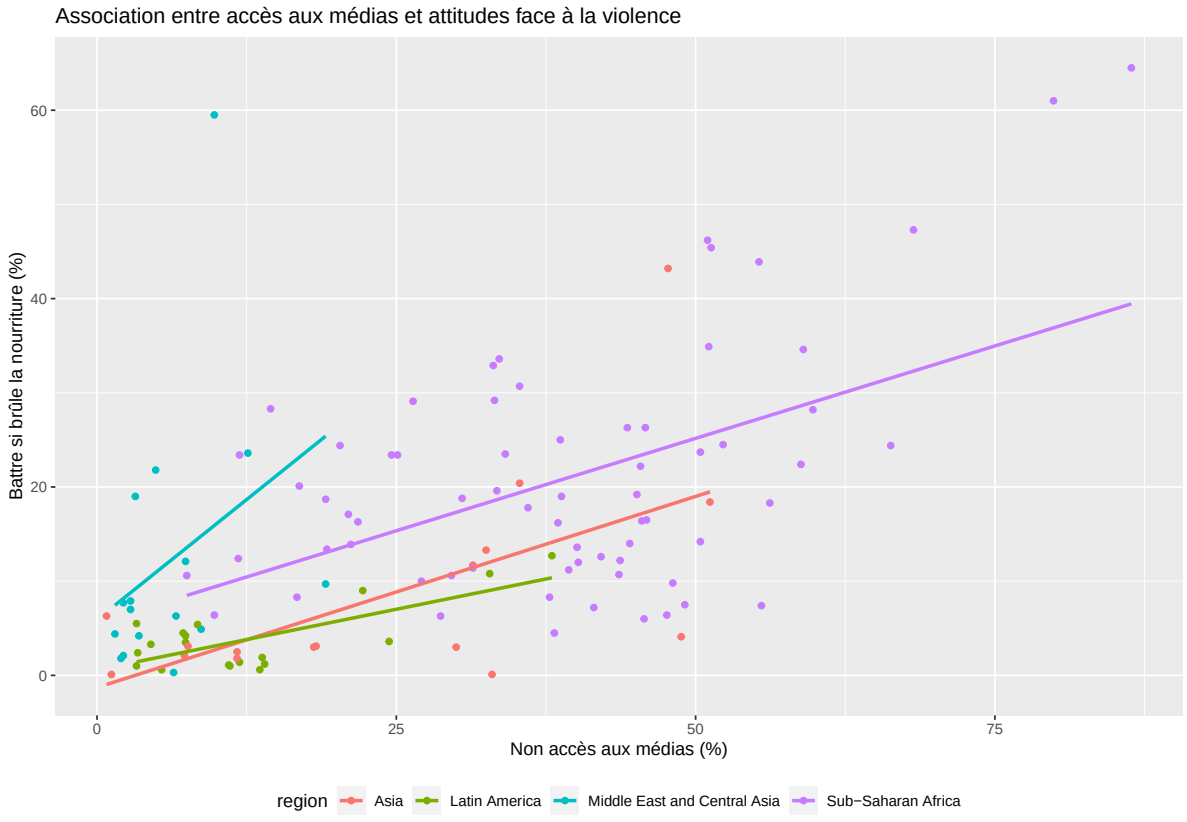
- Plus dans aes : **mappage** : * c'est une mise en relation entre un **attribut graphique du geom** et une variable du tableau de données.
- Changer les couleurs (*color*), la taille (*size*), la position (*position*), la transparence (*alpha*), le remplissage (*fill*)
- **Facets** : le **faceting** permet d'effectuer plusieurs fois le même graphique selon les valeurs d'une ou plusieurs variables qualitatives (notre *group_by*): facet_wrap, facet_grid
- Les **scales** : ils permettent de modifier la manière dont un attribut graphique va être relié aux valeurs d'une variable, et dont la légende correspondante va être affichée.
- Les **thèmes** : ils permettent de contrôler l'affichage de tous les éléments du graphique qui ne sont pas reliés aux données : **titres**, **grilles**, **fonds**, etc.
  https://ggplot2.tidyverse.org/reference/theme.html

Pour aller plus loin : exemple 1

```
d <- ggplot(dhs_ipv) +
  geom_point(aes(x = no_media, y = beat_burnfood,
                 color = region)) +
  geom_smooth(aes(x = no_media, y = beat_burnfood,
                  color = region),
              method = lm, se = FALSE,  formula = y ~ x) +
  labs(title = "Association entre accès aux médias et atti
       x = "Non accès aux médias (%)",
       y = "Battre si brûle la nourriture (%)") +
  theme(legend.position = "bottom", legend.direction = "ho
```

Pour aller plus loin : exemple 1

```
d
```

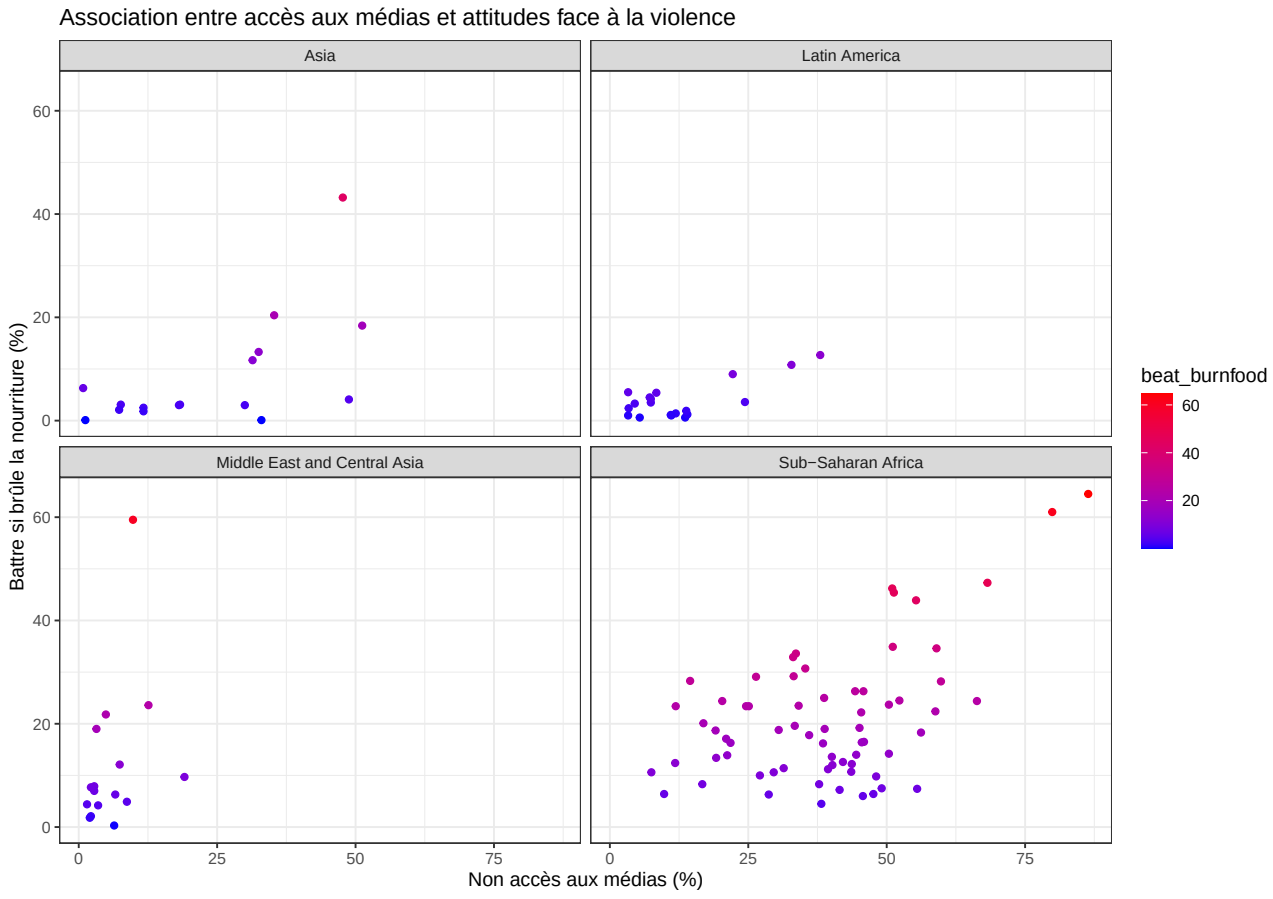


Pour aller plus loin : exemple 2

```
e <- ggplot(dhs_ipv) +
  geom_point(aes(x = no_media, y = beat_burnfood, color = b
  scale_color_gradient("beat_burnfood", low = "blue", high
# scale_size(range = c(0,4), breaks = c(15, 25, 35, 45)) +
  facet_wrap(~region) +
  labs(title = "Association entre accès aux médias et attit
       x = "Non accès aux médias (%)",
       y = "Battre si brûle la nourriture (%)",
       "region" = "Région")  +
  theme_bw()
```

Pour aller plus loin : exemple 2

```
e
```



Ressources

Ressources

- https://www.google.com/search?q=ggplot+theme%2C+dont+show+legend&oq=ggplot+theme%2C+dont+show+legend&aqs=chrome..69i57j0.7717j0j4&sourceid=chrome&ie=UTF-8
- https://juba.github.io/tidyverse/08-ggplot2.html#
- Fortement recommandé
- https://www.rstudio.com/resources/cheatsheets/
- http://r4ds.had.co.nz/data-visualisation.html#aesthetic-mappings
- http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html
- http://www.cookbook-r.com/Graphs/
- http://www.ggplot2-exts.org/gallery/
- Si vous y trouver de la passion...