Session 6.2: Visualisation avec ggplot

Cours et Labo

Visseho Adjiwanou, PhD.

Département de Sociologie - UQAM

15 February 2022

Plan

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

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

Les types de graphiques

- Dépend en général du type de variable (qualitative ou quantitative) et du nombre de variables

1. 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)

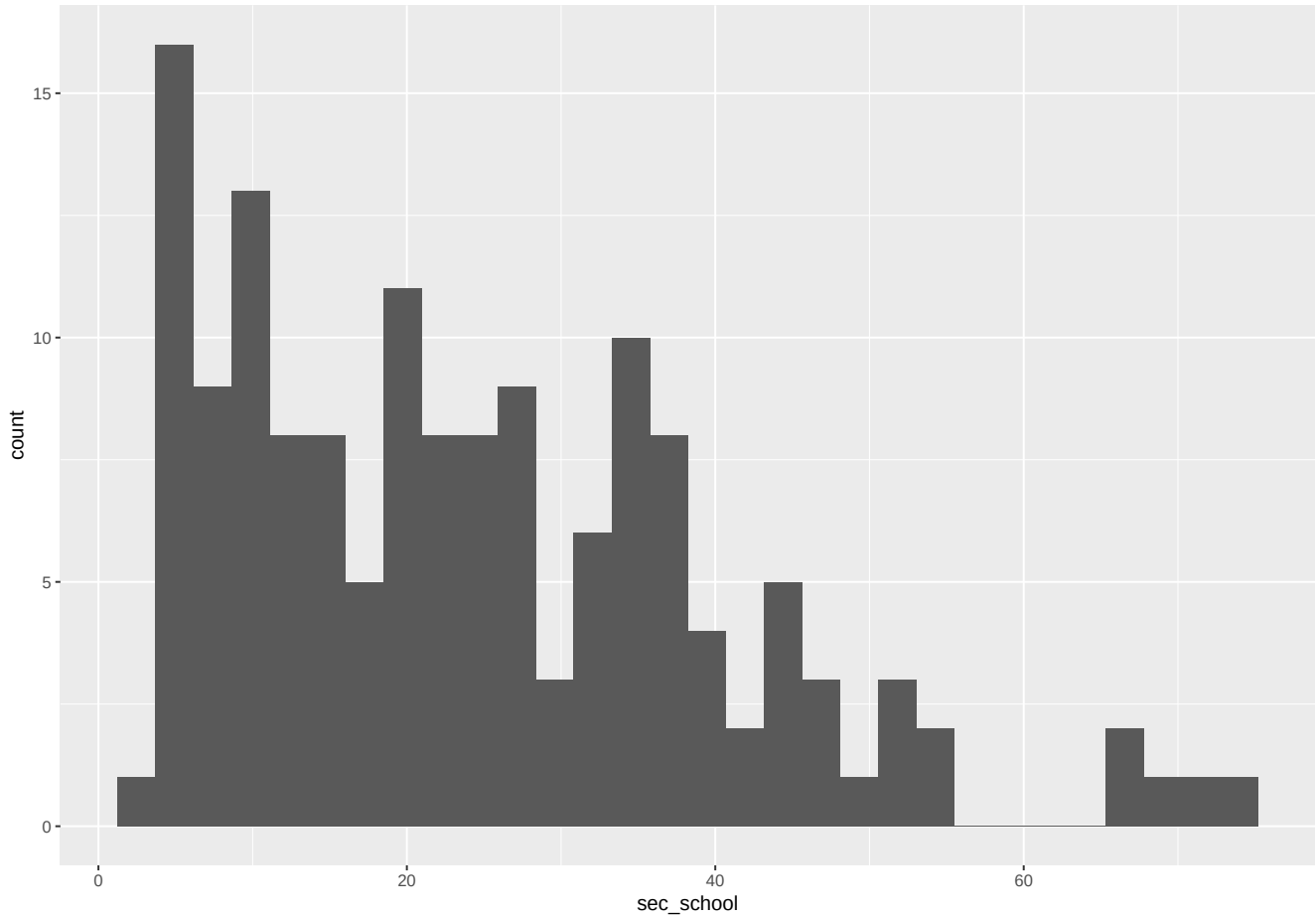
Exemples: Visualiser la distribution univariée

Distribution univariée

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

Chargement du package et de la base de donnée

Histogramme



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.

Diagramme en bâtons ou à barres

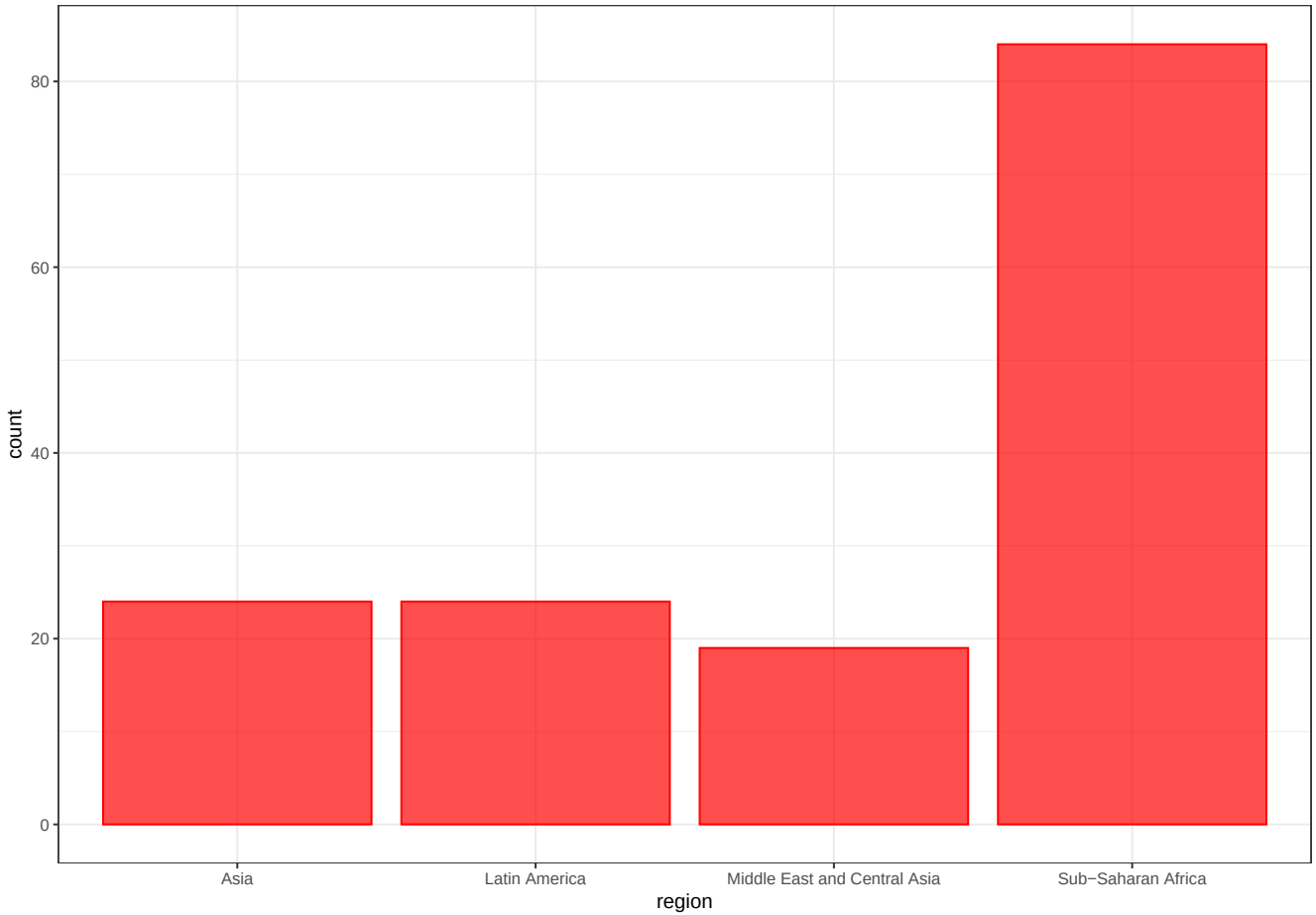


Diagramme en bâtons ou à barres

- 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.

Remarques

Un diagramme à bandes verticales diffère d'un histogramme de par les éléments suivants :

1. Dans un histogramme, la fréquence est mesurée par la surface de la colonne.
2. Dans un diagramme à bandes verticales, la fréquence est mesurée par la hauteur de la barre.

Diagramme circulaire

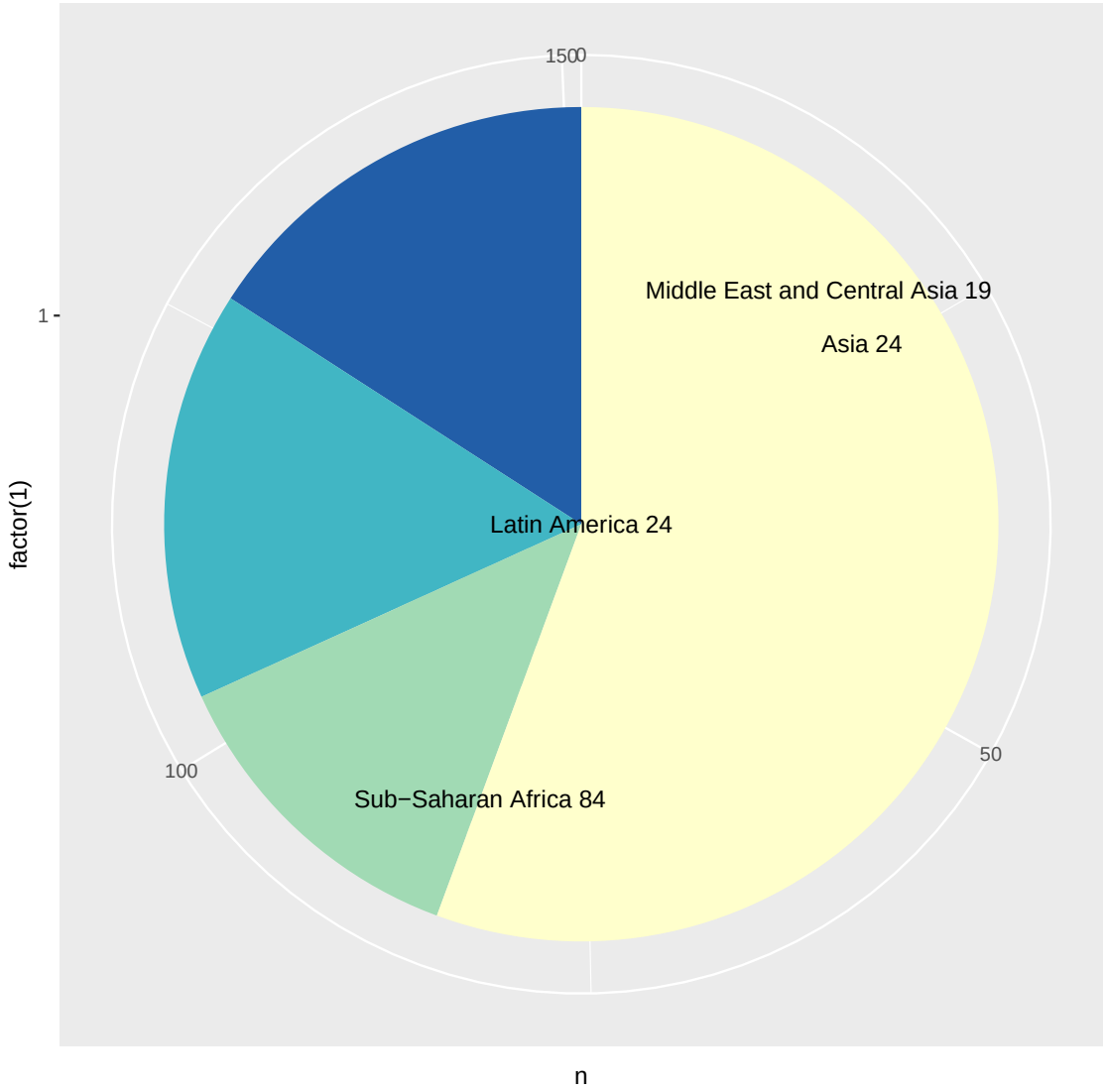


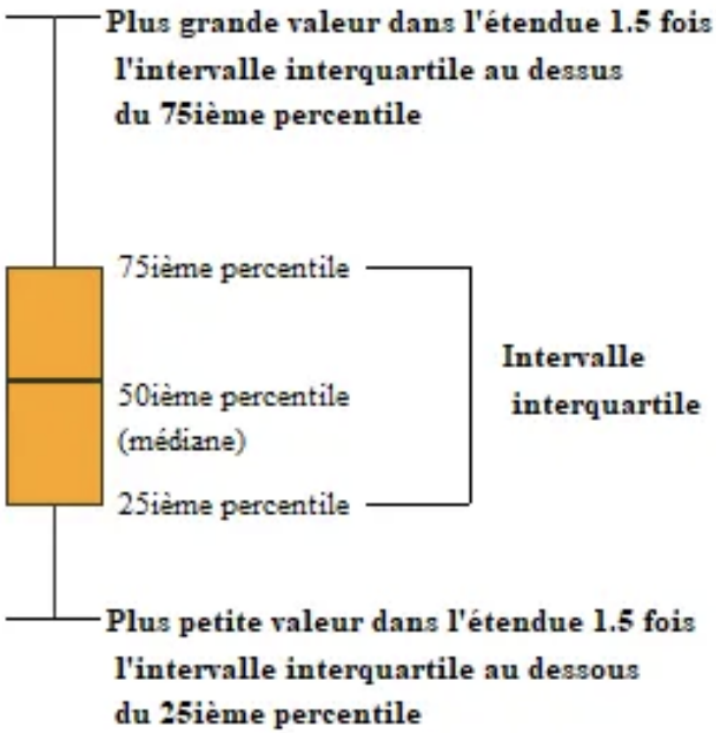
Diagramme circulaire

- Essayer de l'améliorer : https://dataparkblog.wordpress.com/2017/09/24/diagramme-en-camembert-avec-r-et-ggplot/

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.
- C'est particulièrement utile lorsque vous **comparez la distribution de plusieurs variables** en les plaçant côte à côte.

Diagramme de quartile



- **Valeur Outlier** La valeur est >1.5 fois et <3 fois l'intervalle interquartile au delà de chaque coté de la boite

Figure 2:

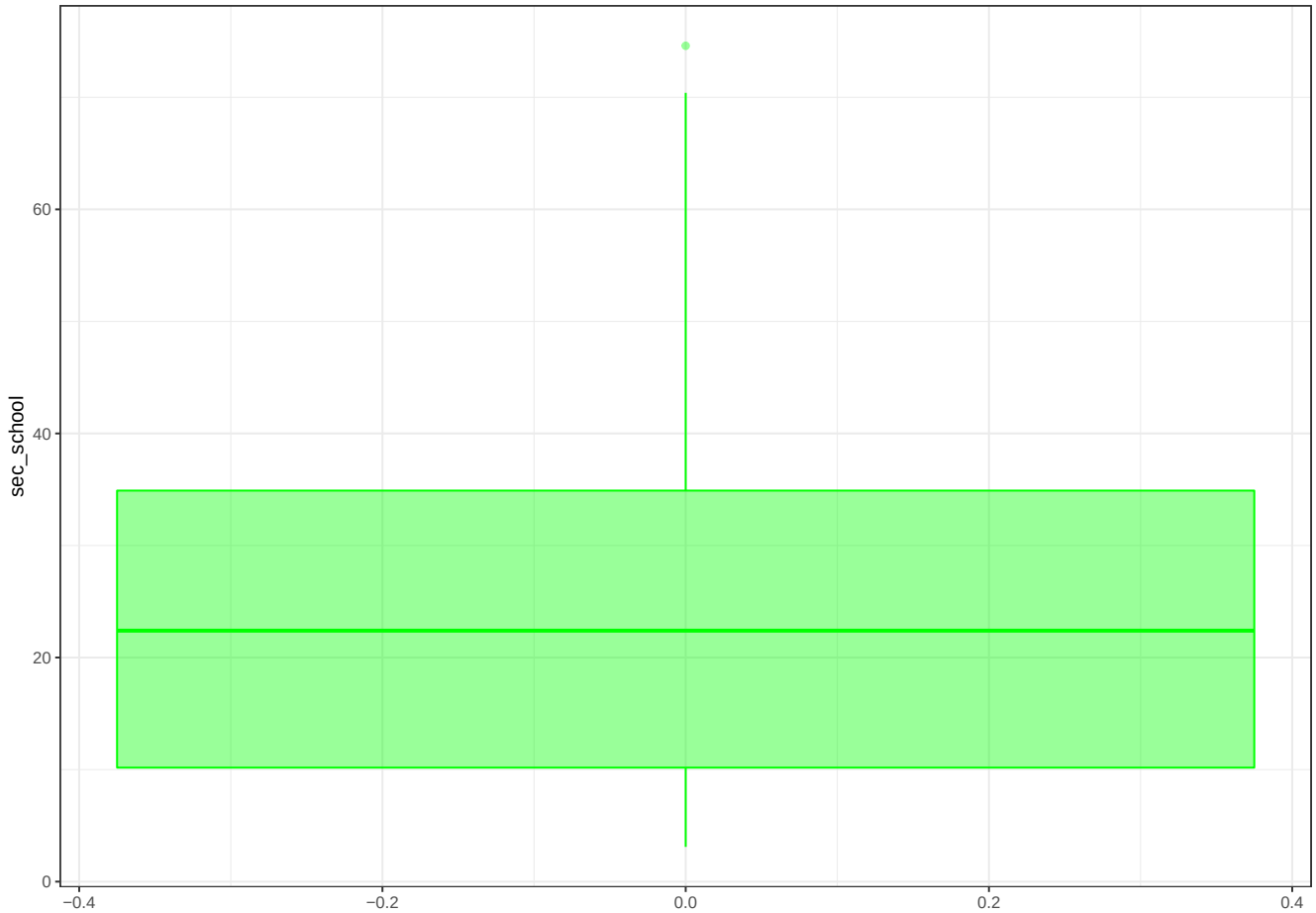
Diagramme de quartile

- Ce graphique permet de visualiser les données aberrantes ou les outliers.
- Un **outlier**, ou **donnée aberrante** est “une valeur ou une observation qui est « distante » des autres observations effectuées sur le même phénomène, c’est-à-dire qu’elle contraste grandement avec les valeurs « normalement » mesurées.
- Une donnée aberrante peut être due à la variabilité inhérente au phénomène observé ou bien elle peut aussi indiquer une erreur expérimentale. Les dernières sont parfois exclues de la série de données”.

Diagramme de quartile

- Comment détecter les outliers? https://statistique-et-logiciel-r.com/comment-detecter-les-outliers-avec-r/

Diagramme de quartile



Présentation de ggplot

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)
```

Forme générale

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

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

Figure 3:

1. **ggplot** spécifie que vous utilisez 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>))
```

Figure 4:

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**,

Forme générale

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

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

Figure 5:

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. Ces paramètres sont principalement liés aux variables.

Remarques

- Chaque graphique comprend principalement deux parties:

1. **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*)

2. **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

Utilisation du package exquisse

- Le package **esquisse** vous permet d'utiliser ggplot d'une manière interactive.

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

#install.packages("esquisse")

#library(esquisse)
```

- Il va ajouter un addins à votre Rstudio

Remarques

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

Pour aller plus loin

1. 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*)

Pour aller plus loin

2. **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

Pour aller plus loin

3. 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.

Pour aller plus loin

4. 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

Exemple

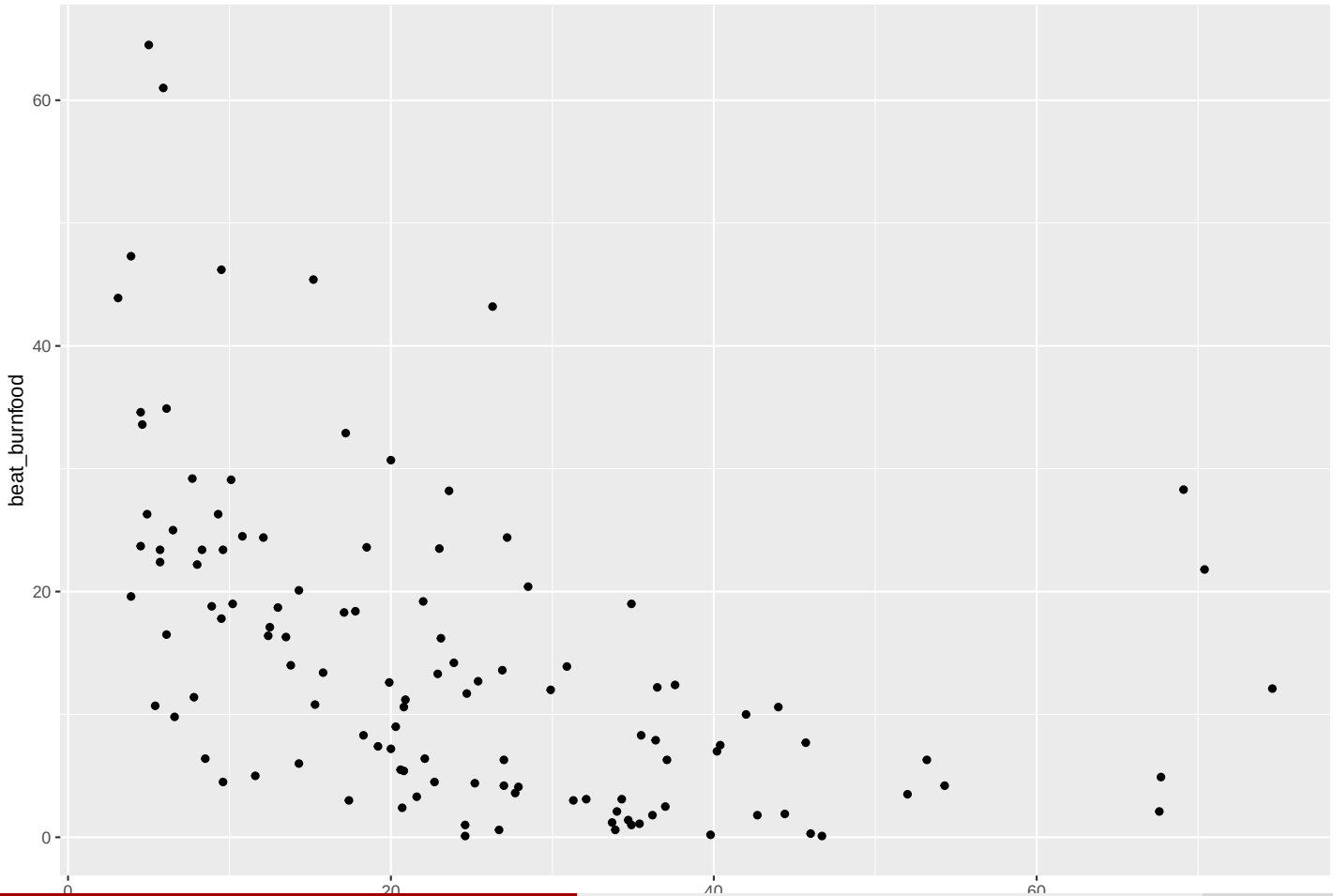
1. **Nuage de points**

```
d1 <- ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood))
```

Exemple

1 Nuage de points

```
d1
```



Exemple

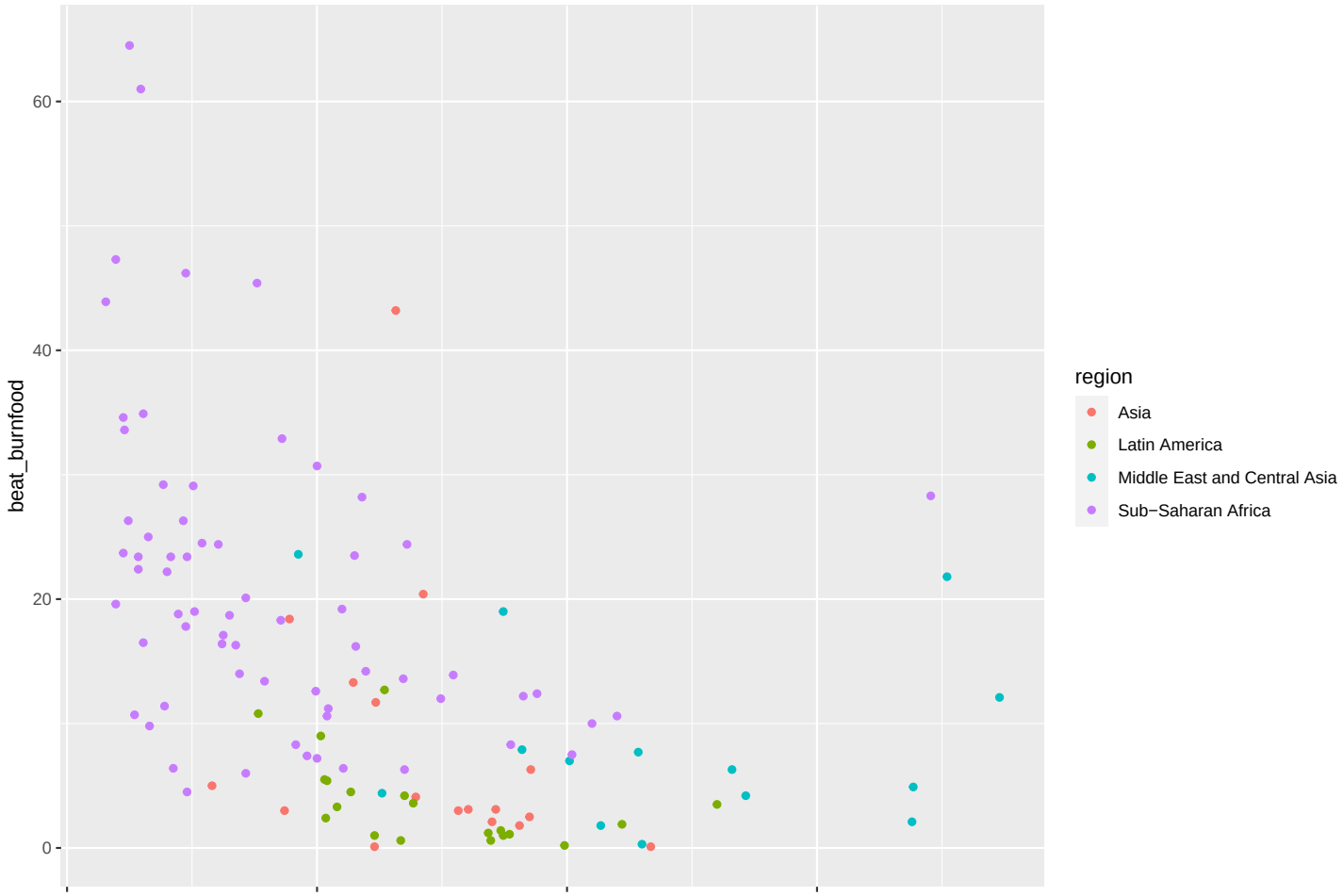
2. **Nuage de points avec couleur pour distinguer les régions**

```
d2 <- ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood, color = re
```

Exemple

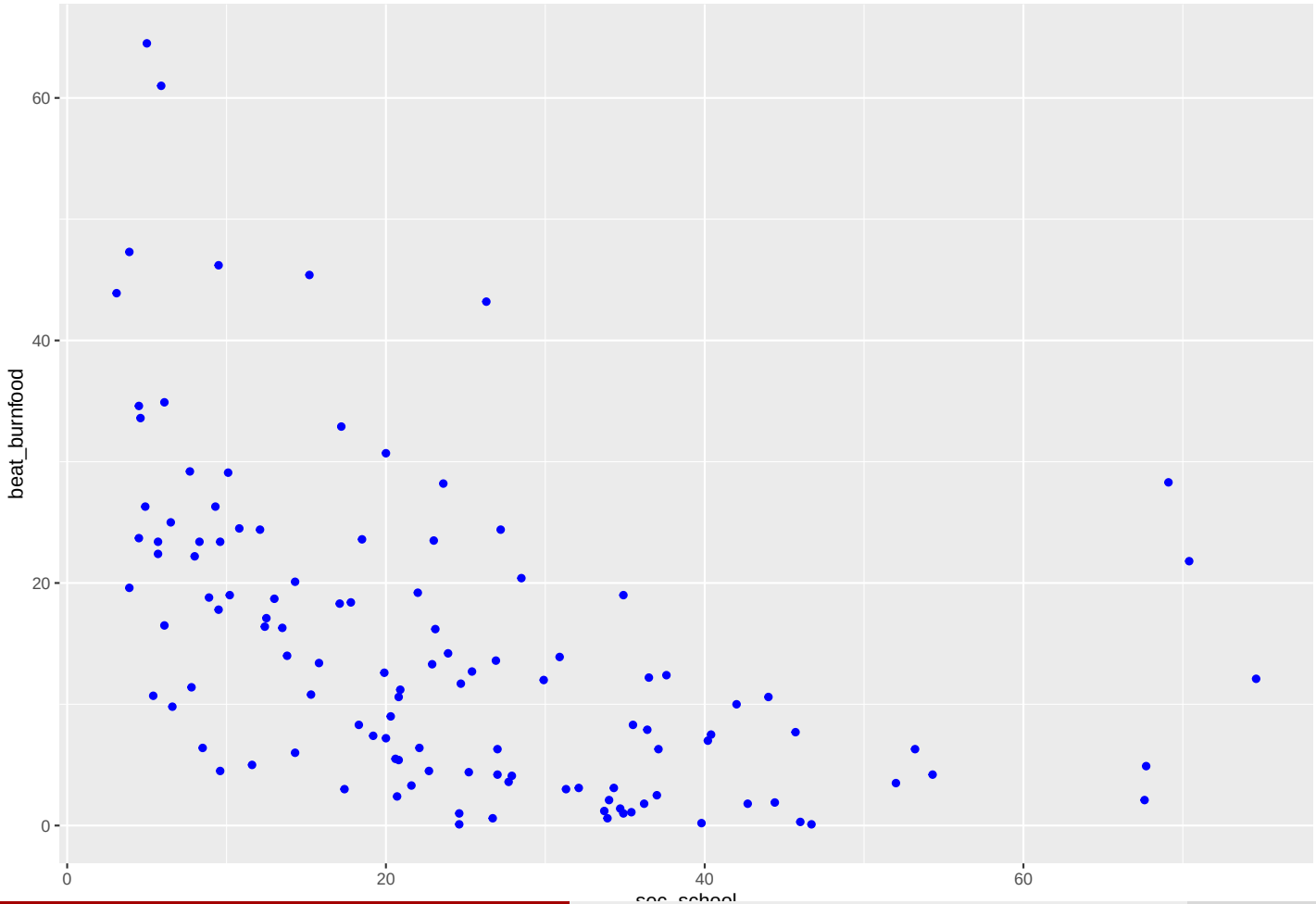
2 Nuage de points avec couleur pour distinguer les régions

```
d2
```



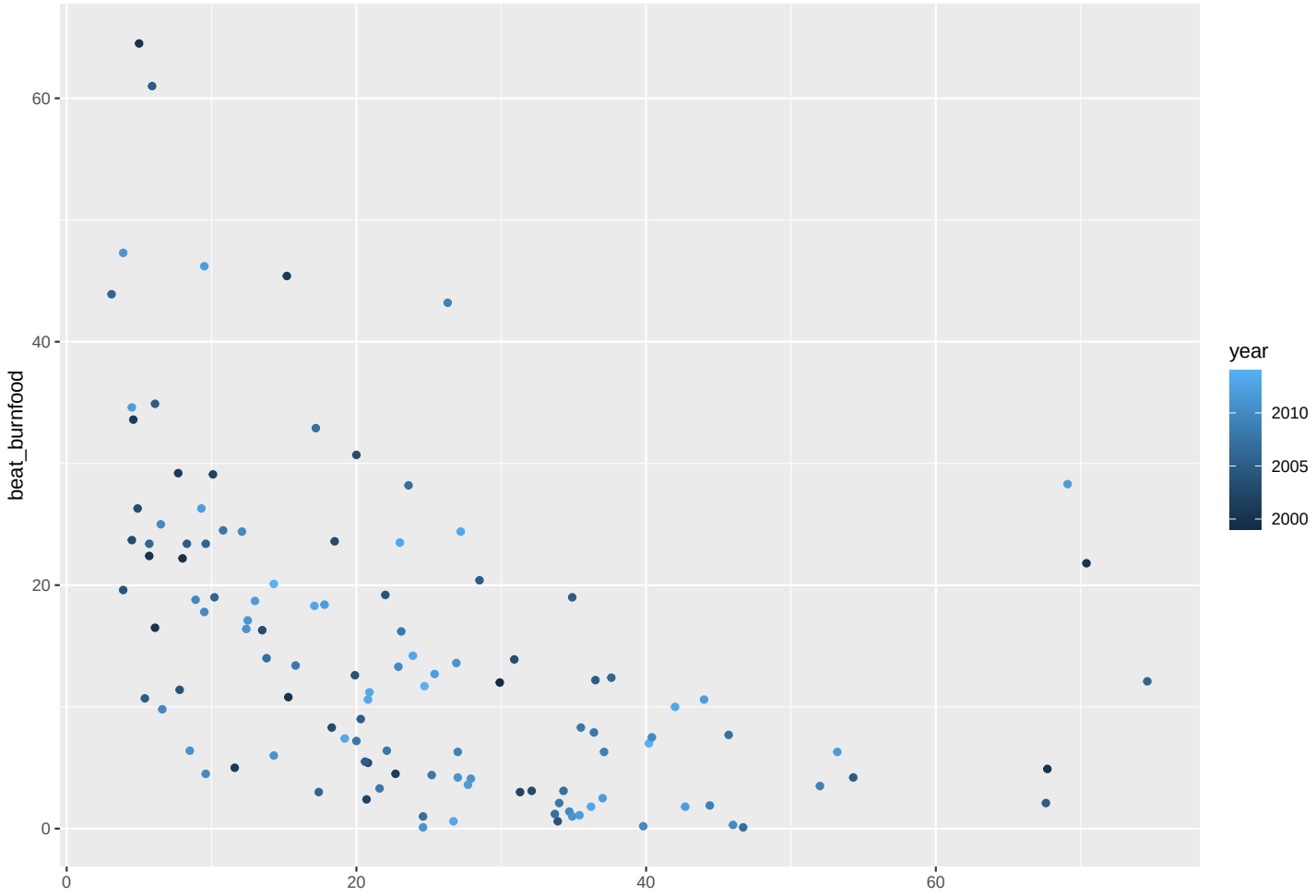
Exemple

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



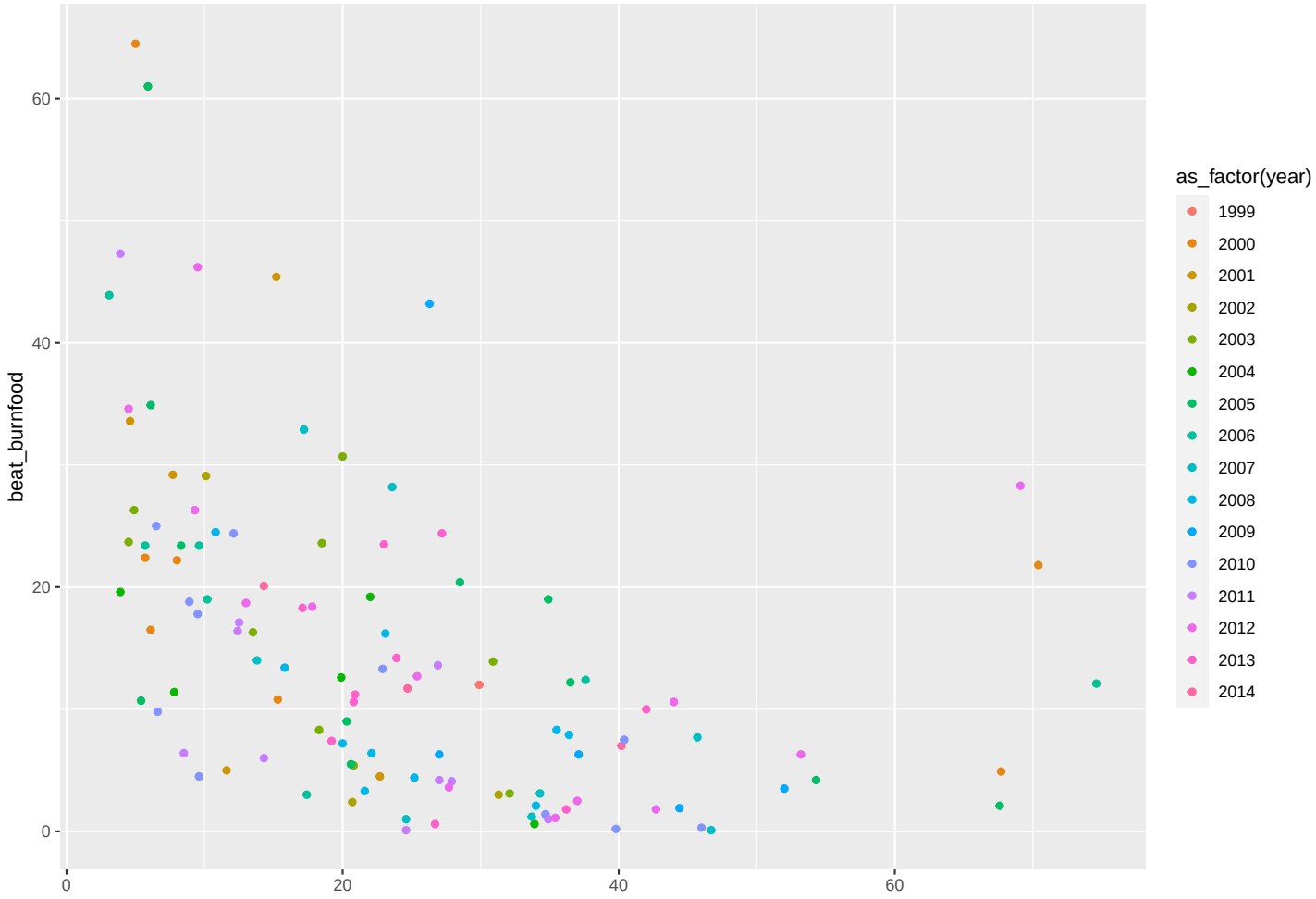
Exemple

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



Exemple

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



Exemple

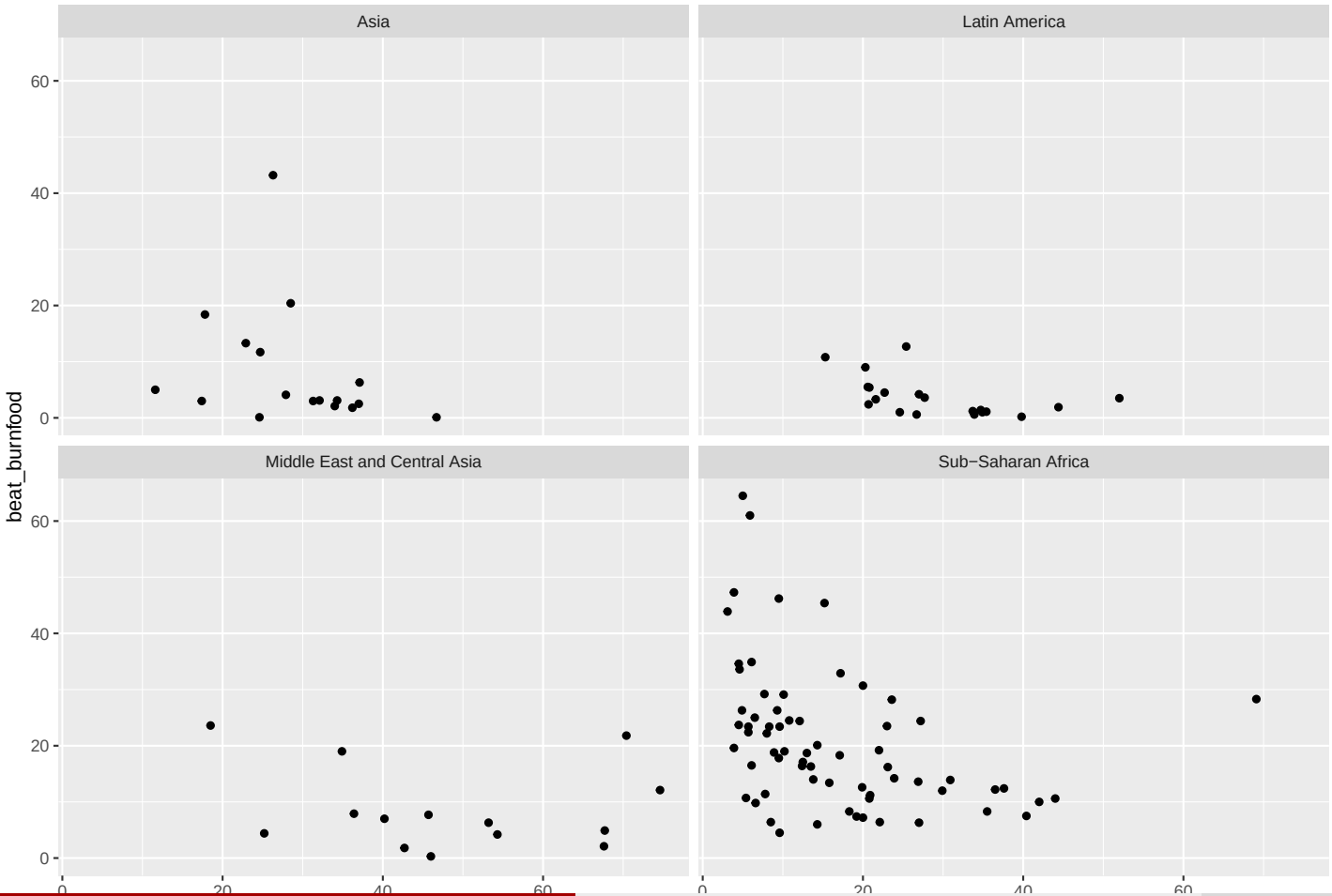
3. **Nuage de points pour chaque region**

```
d3 <- ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood)) +
  facet_wrap(~region)
```

Exemple

3 Nuage de points pour chaque region

```
d3
```



Exemple

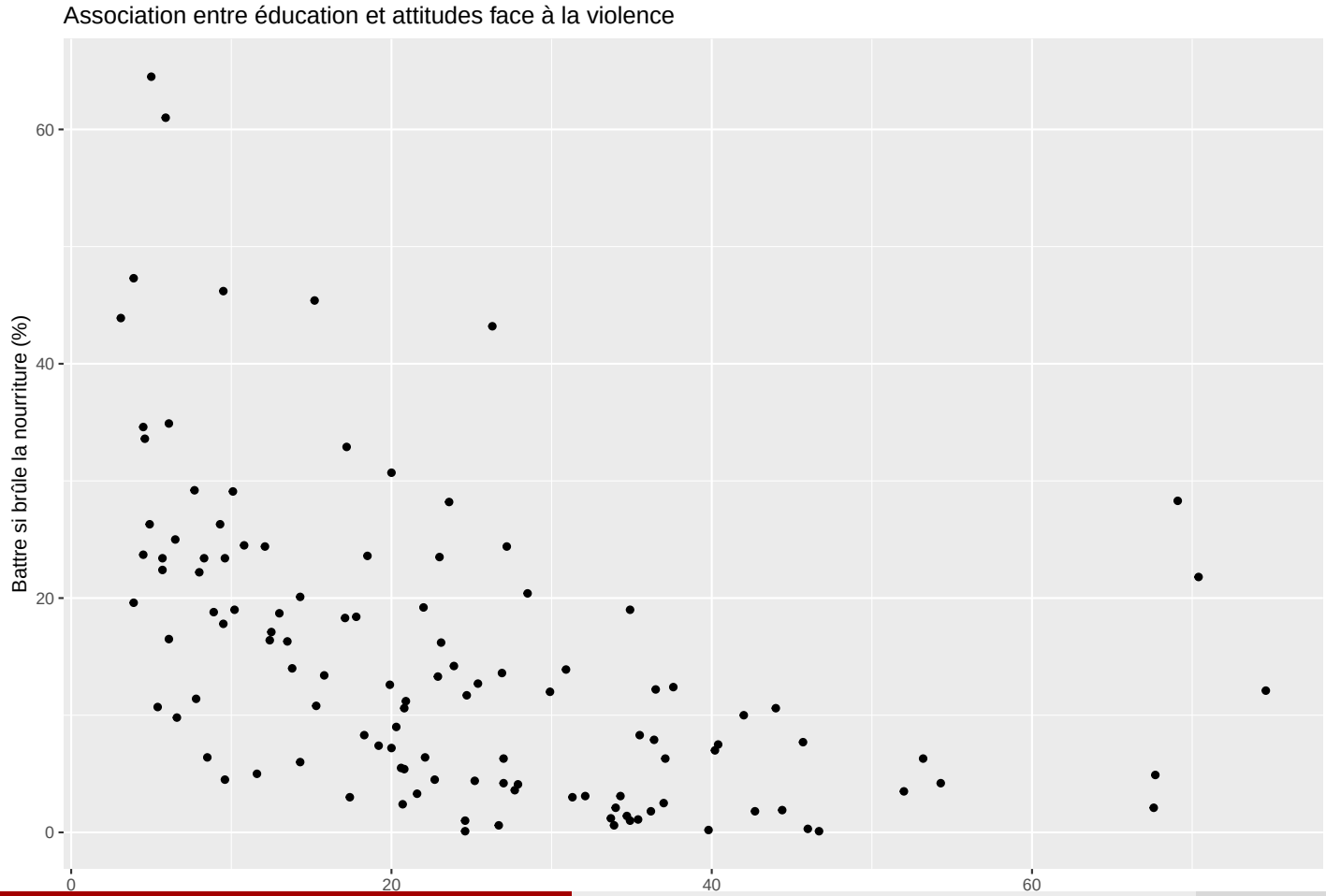
4. **Nuage de points avec indication sur les axes**

```
d4 <- ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood)) +
  labs(title = "Association entre éducation et attitudes face
       x = "% de femmes avec niveau secondaire",
       y = "Battre si brûle la nourriture (%)",
       "region" = "Région")
```

Exemple

4 Nuage de points avec indication sur les axes

```
d4
```



Remarques sur les informations du chunck

- Dans le cadre de cette présentation, je mets des options dans le **chunk**
- **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 donnent des informations, par exemple sur les valeurs manquantes.

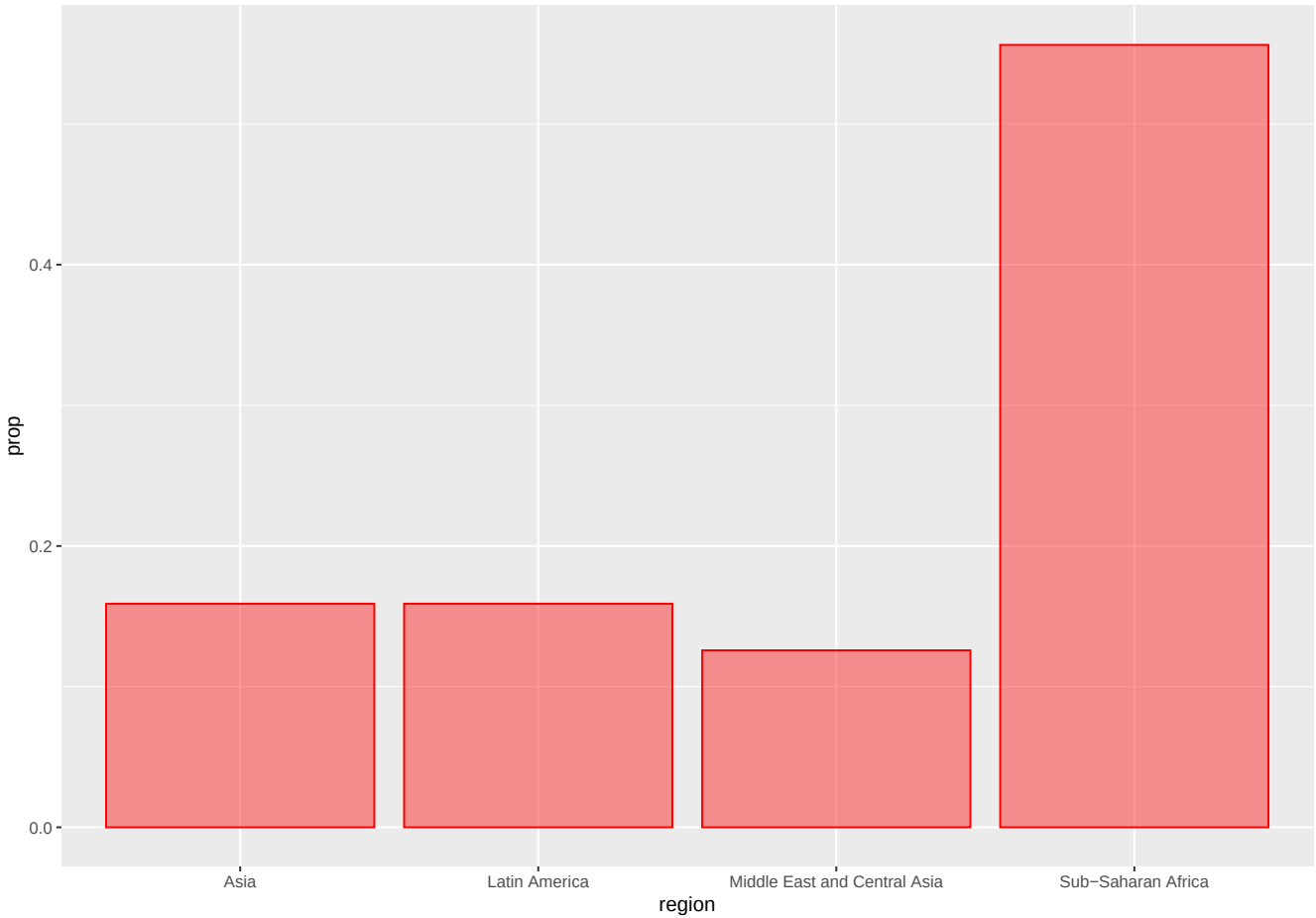
Remarques sur certains graphiques

- Le diagramme de barres avec les effectifs n'est pas utile pour les comparaisons.

```
d5 <- ggplot(dhs_ipv) +
  geom_bar(aes(x = region, y = ..prop.., group = 1), fill = "1
```

Remarques sur certains graphiques

```
d5
```



Remarques générales sur les graphiques

- 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

Type de graphiques pour les distributions bivariées

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

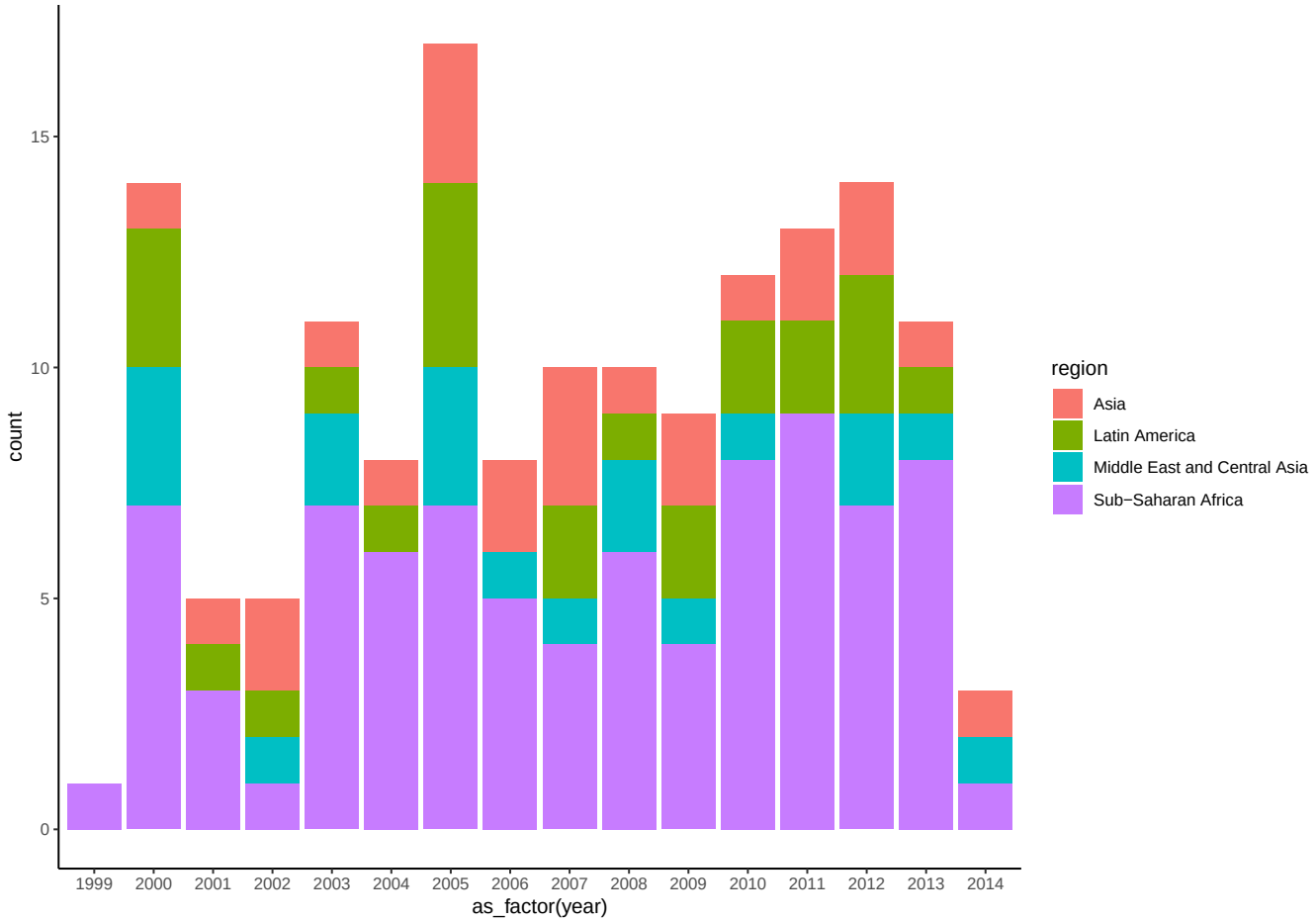
Figure 6:

Exemples: Visualiser la distribution bivariée

Distribution bivariée

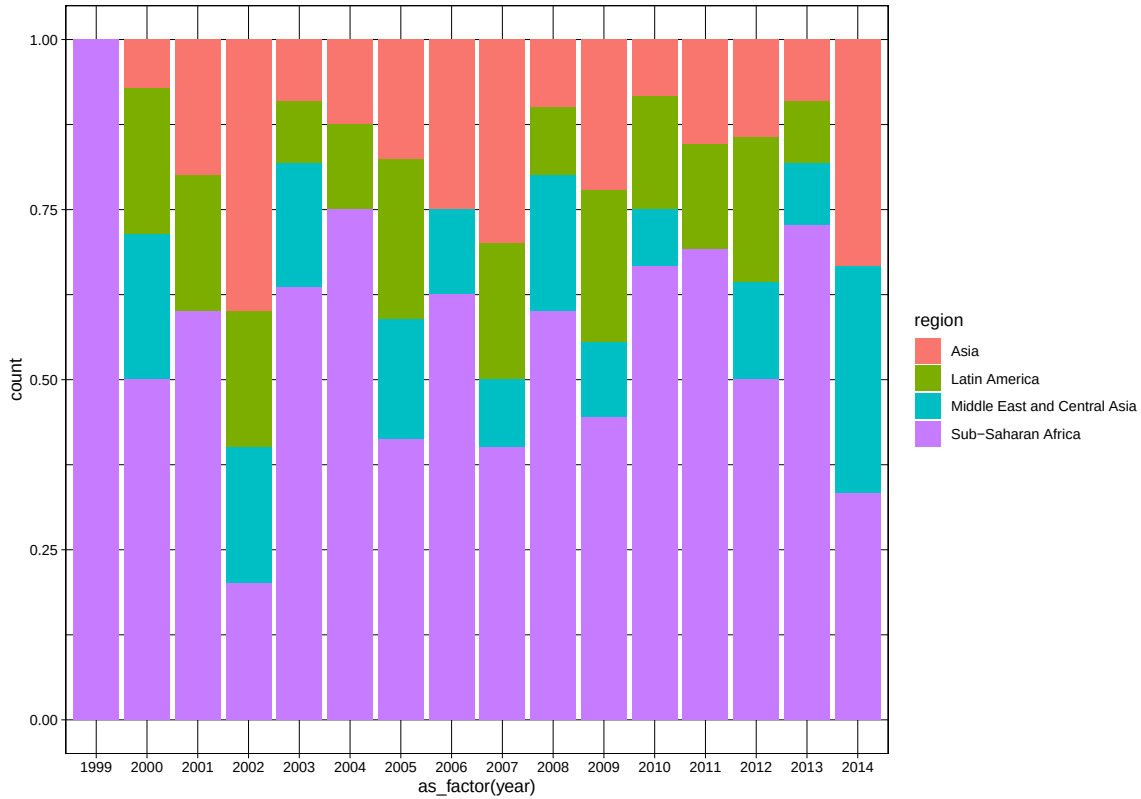
- Croisement de deux variables qualitatives: diagramme de bâton empillé
- Croisement d'une variable qualitative et d'une variable quantitative: diagramme de boîte à moustache ou Boxplot
- Croisement de deux variables quantitative: diagramme de dispersion ou nuage de points

Croisement de deux variables qualitatives



- Ce graphique nous donne pour chaque sexe, le nombre de personne qui sont dans chaque catégorie de la variable dépendante.

Croisement de deux variables qualitatives



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

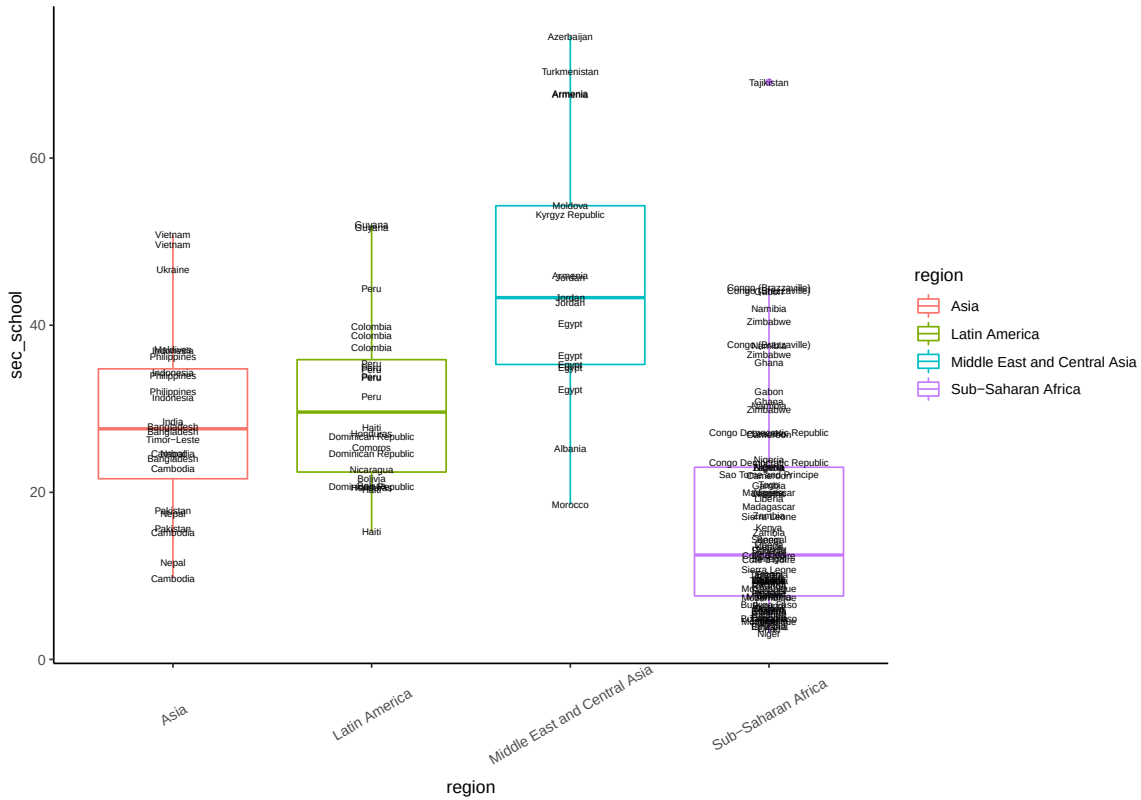
Croisement d'une variable quantitative et d'une variable qualitative

- 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

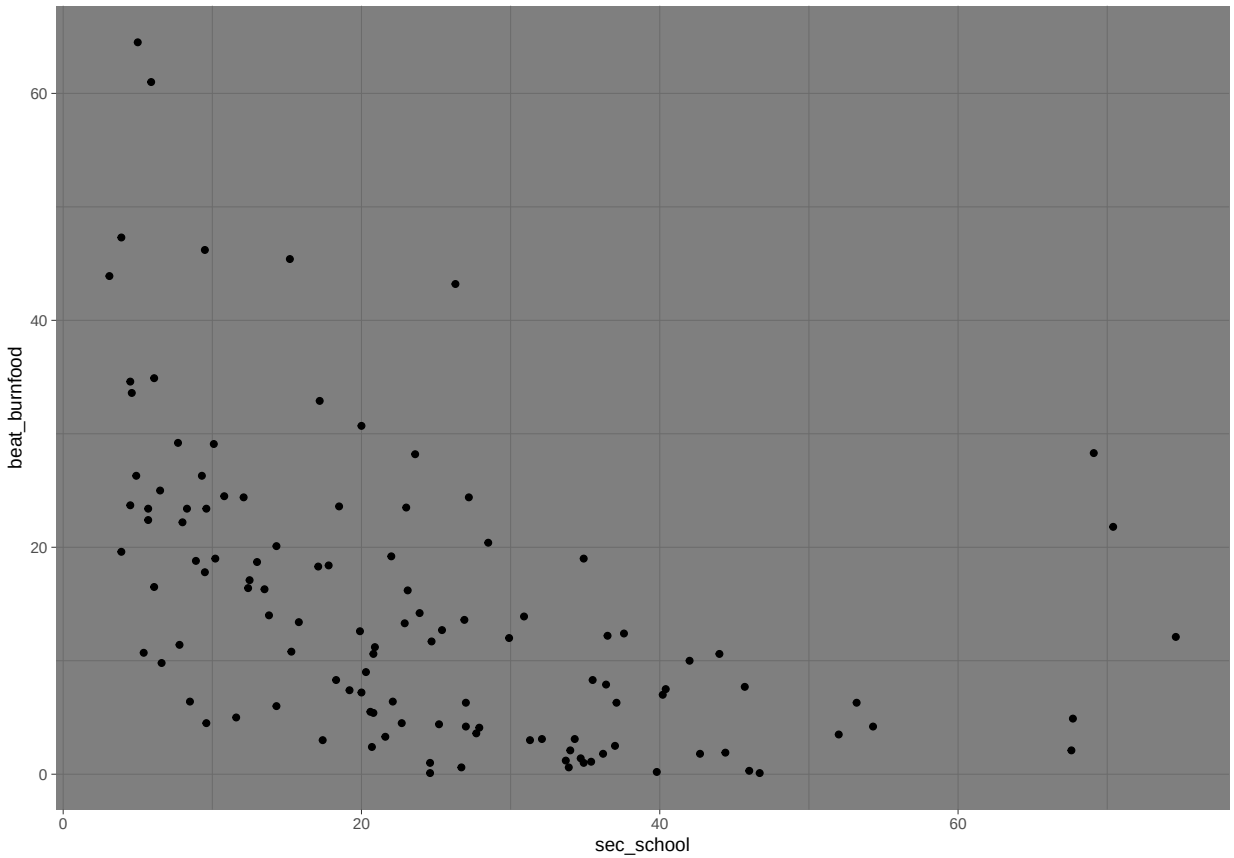
Diagramme de quartile

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



- Note : x doit être une variable qualitative, et y une variable

Corrélation linéaire : Croisement de deux variables quantitatives



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. . .