SEMINR

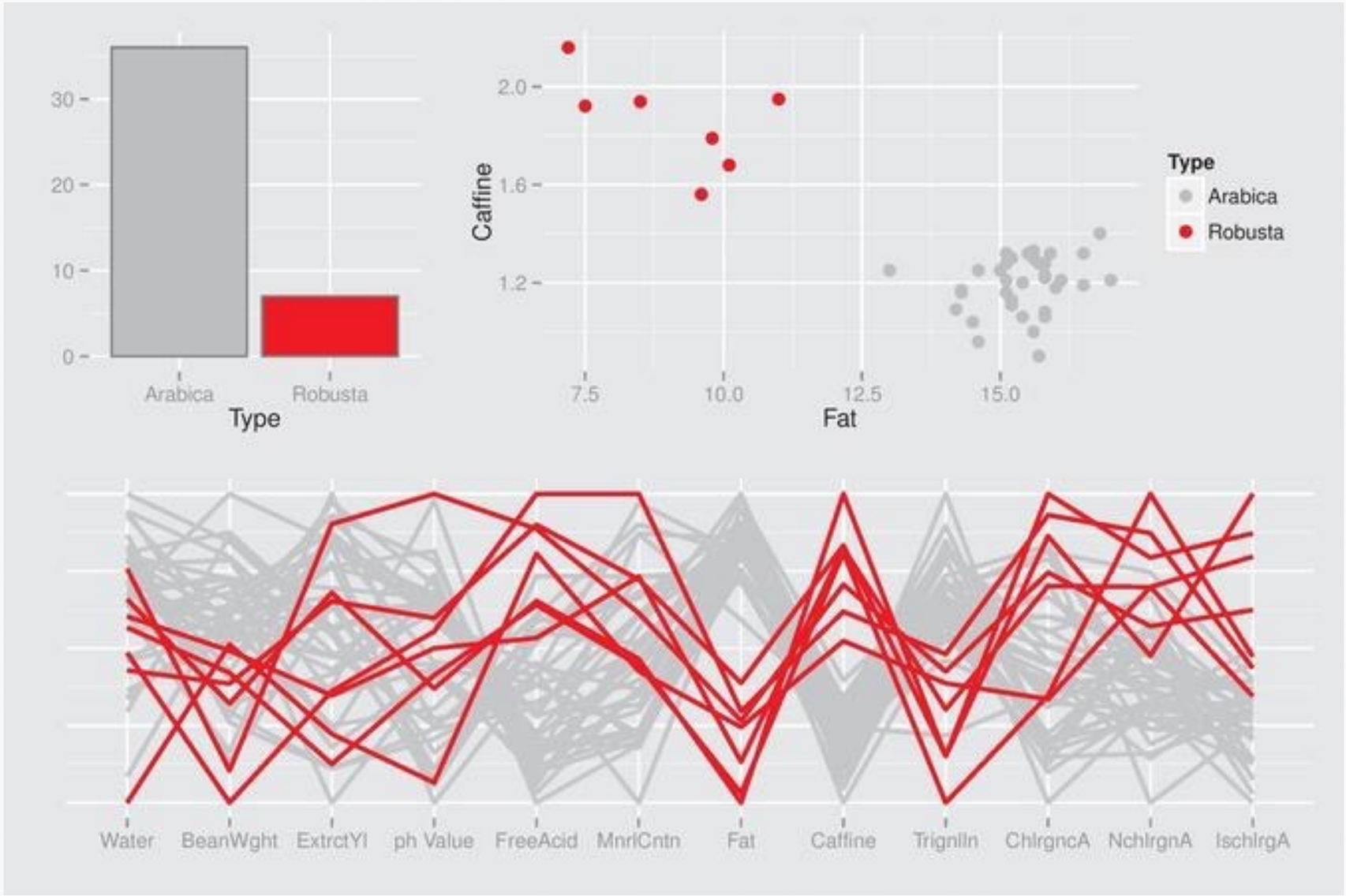
LES GRAPHIQUES AVEC GGPLOT2

CHRISTOPHE LALANNE

www.aliquote.org

“Use excellent graphics, liberally.”

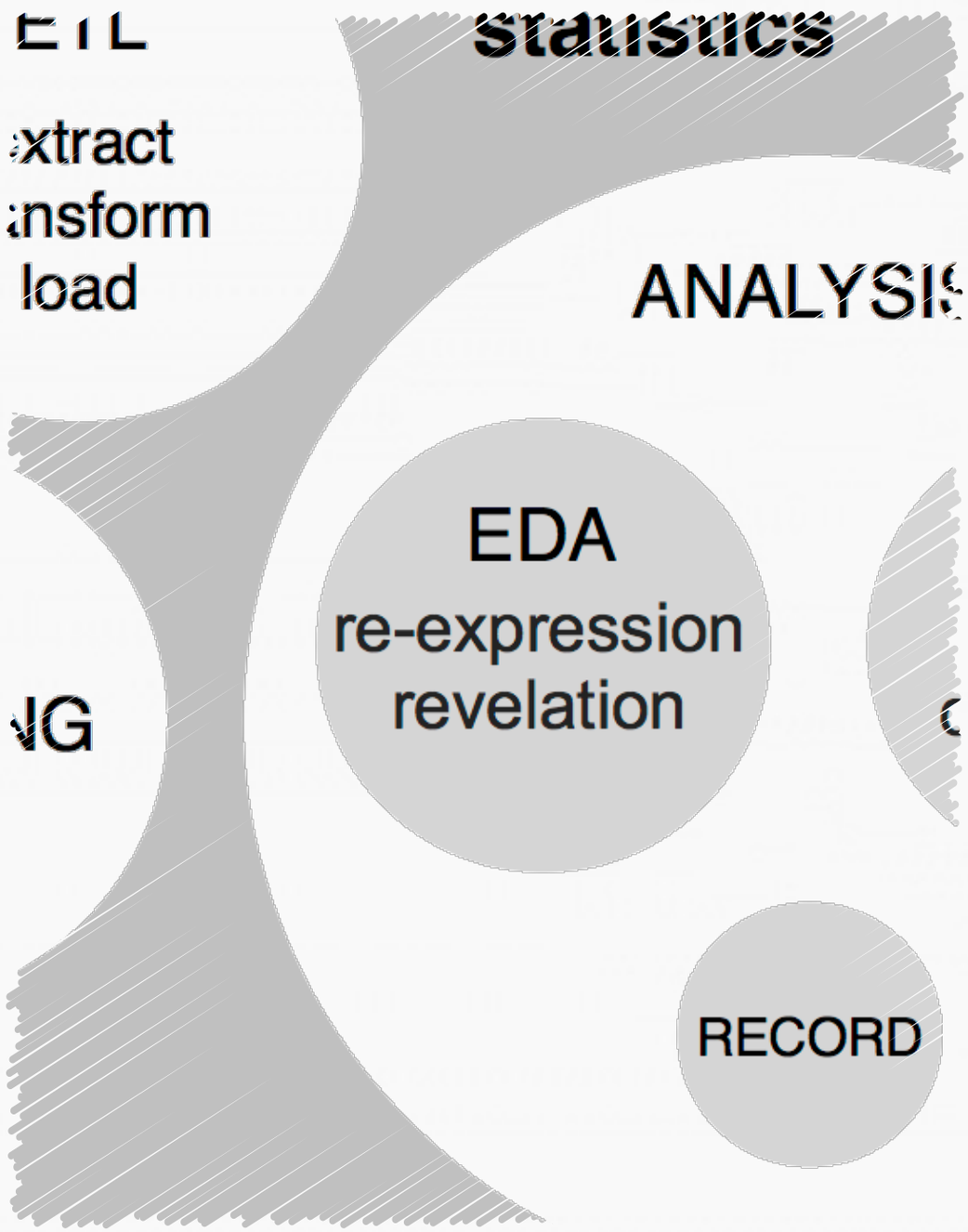
— Frank E. Harrell



www.gradaanwr.net

SYSTEMES GRAPHIQUES SOUS R

UNE APPROCHE INTERACTIVE



READ-EVAL-PRINT-LOOP

The plural of anecdote is (not) data, http://blog.revolutionanalytics.com/2011/04/the-plural-of-anecdote-is-data-after-all.html

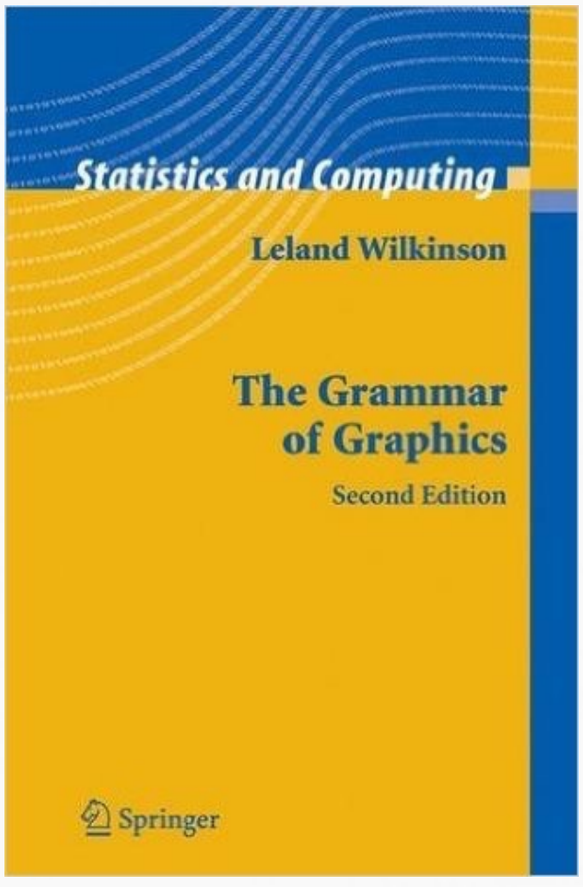
EDA ⇆ MOD

LES 3 SYSTEMES GRAPHIQUES

Commandes de base pour tous graphiques, hautement personnalisables, au prix d'une syntaxe plus lourde.

Commandes unifiées autour de la notion de formules, bien adaptées à la modélisation et avec des options par défaut satisfaisantes.

Couteau suisse de la visualisation de données (« Hadley verse »), reposant sur une syntaxe différente mais avec des capacités plus riches



https://github.com/hadley/ggplot2-book

GRAMMAIRE DES GRAPHIQUES

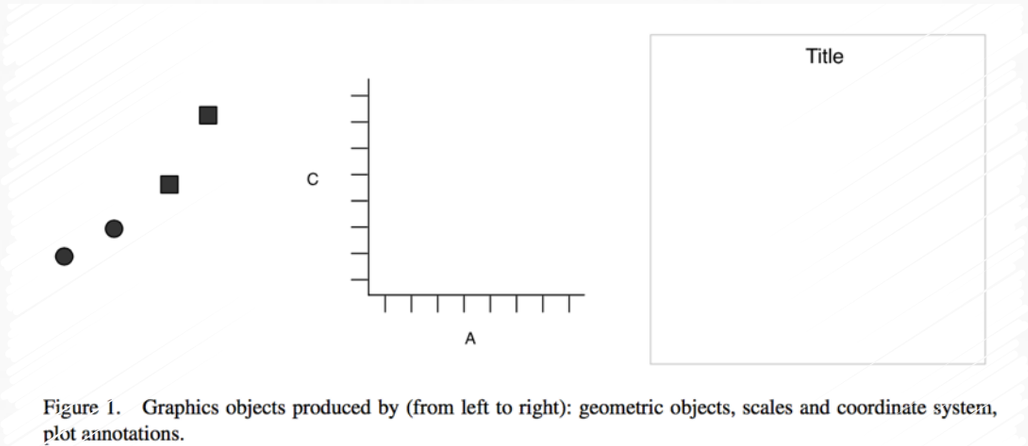


Figure 1. Graphics objects produced by (from left to right): geometric objects, scales and coordinate system, plot annotations.

http://docs.ggplot2.org/current/

PHILOSOPHIE

On arrange des objets géométriques selon des couches superposées et on leur associe des propriétés spatiales : type de forme géométrique, système de coordonnées (métrique), arrangement dans le plan, système d'annotation.

Pas d'usage de formule pour décrire la relation entre les variables, mais possibilité de graphiques en trellis identiques.

Deux commandes : `ggplot()` et `qplot()`.

ELEMENTS CLES

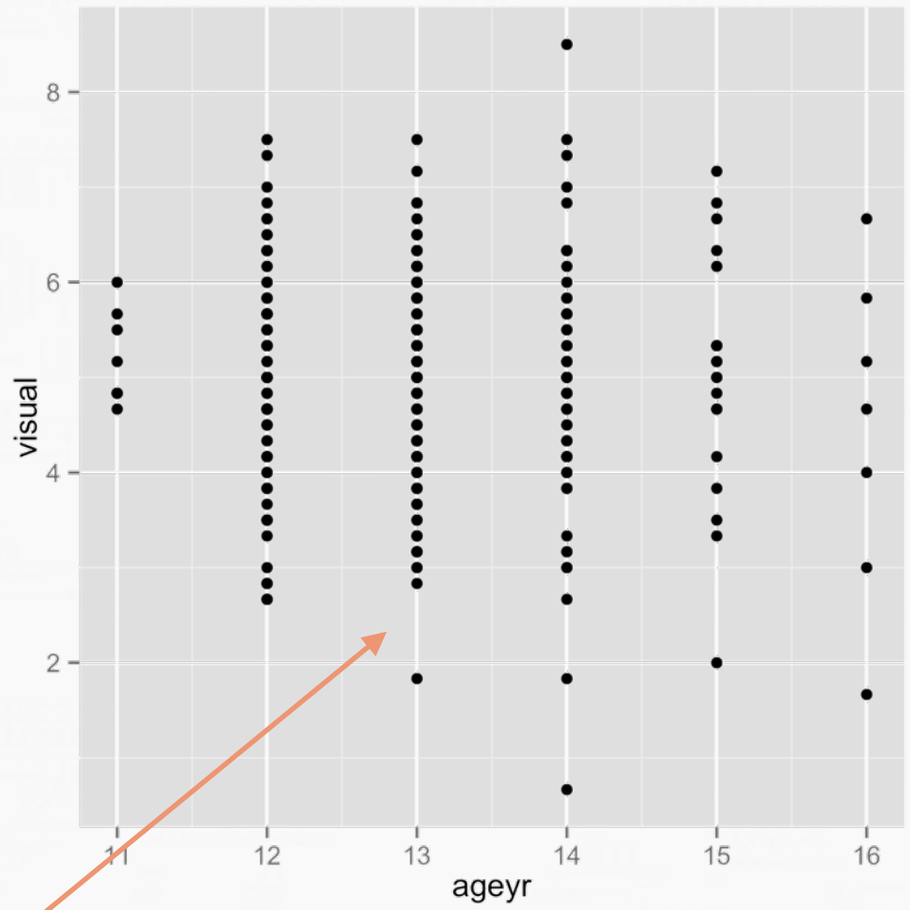
- un data frame et un ensemble de variables mises en correspondance *via* une « esthétique » : `aes(x=, y=)`
- une ou plusieurs couche(s) comprenant : un objet géométrique, une transformation statistique, une position : `geom_*`
- une échelle pour chaque variable : `scale_*`
- un système de coordonnées : `coord_*`
- un système de facettes : `facet_*`

esthétique — objet géométrique — facette

```
ggplot(aes(x = supp, y = len), data = ToothGrowth) + geom_point() + facet_grid(~ dose)
```

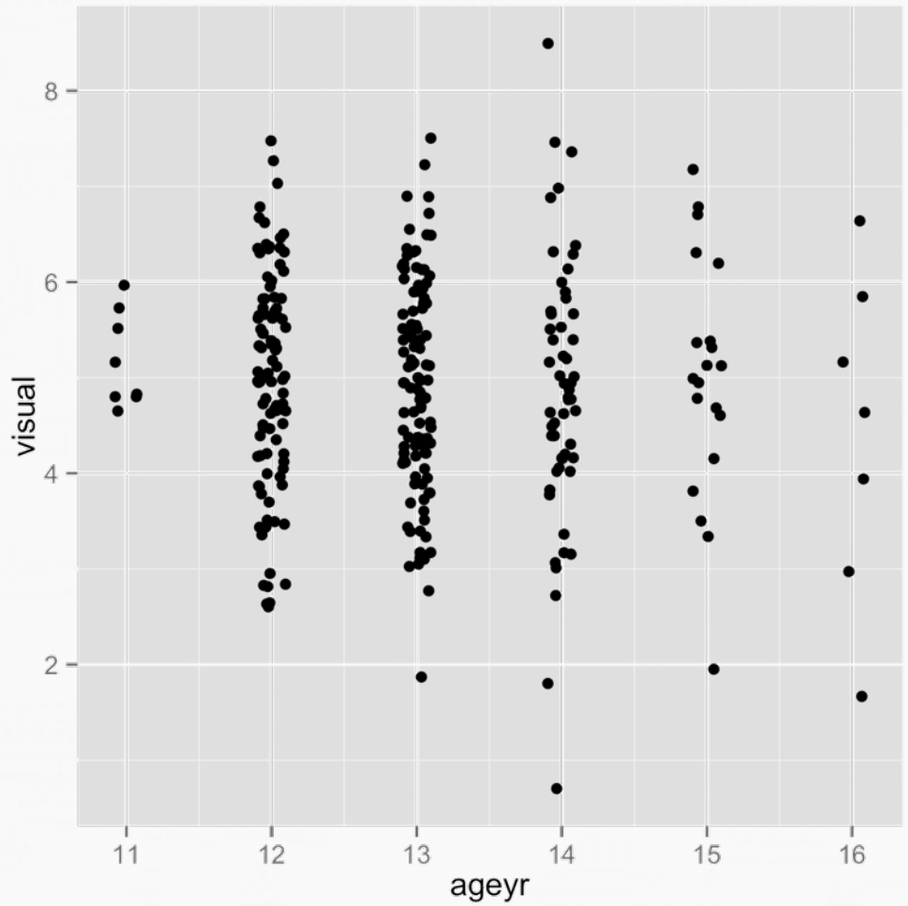
```
qplot(x = supp, y = len, data = ToothGrowth, geom = "point", facets = ~ dose)
```

ILLUSTRATION



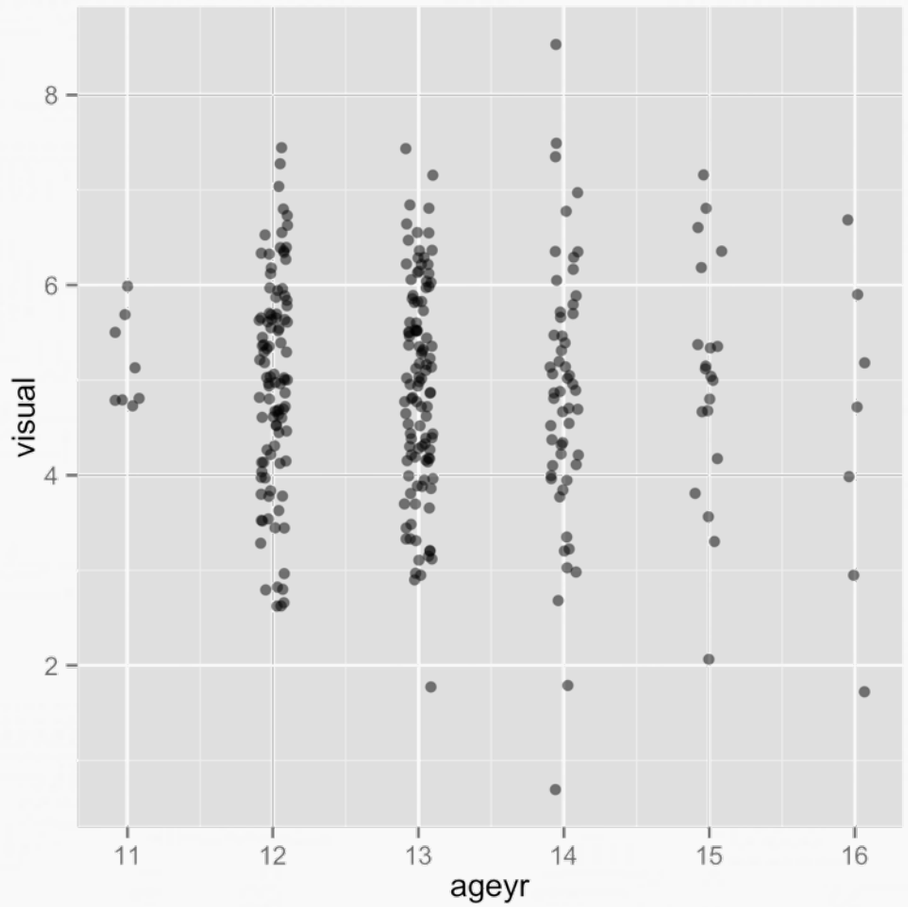

```
1 library(ggplot2)
2
3 p <- ggplot(data = HS, aes(x = ageyr, y = visual))
4 p + geom_point()
```

ILLUSTRATION



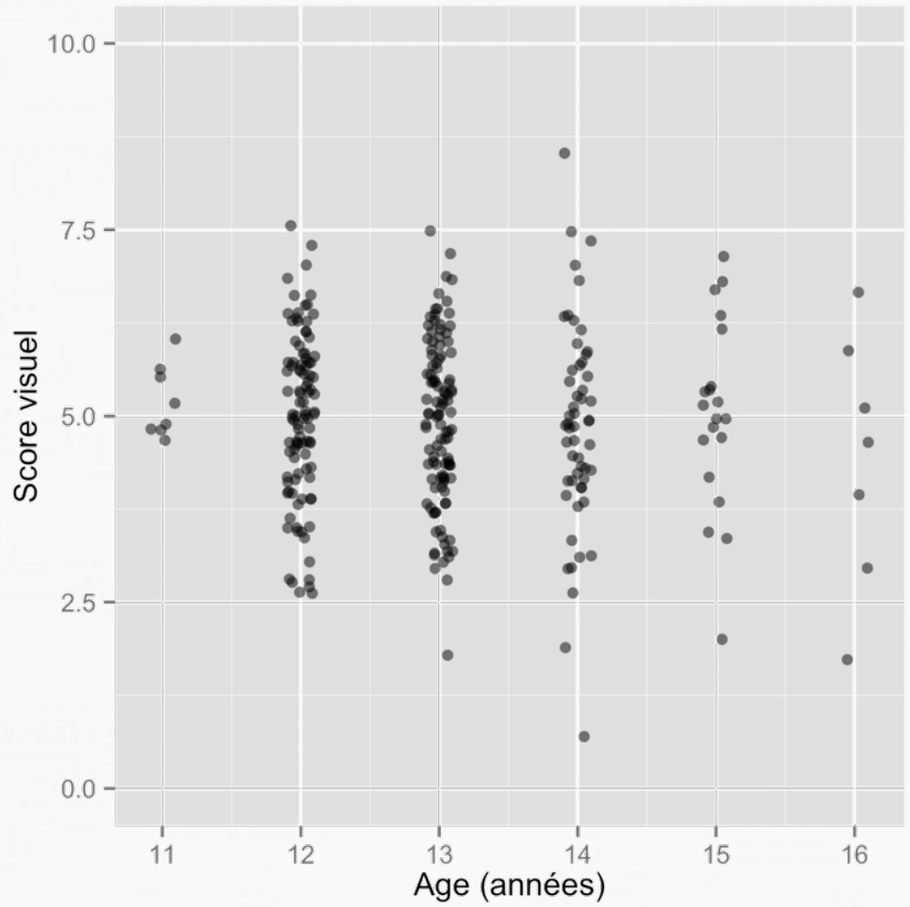

```
1 p + geom_point(position = position_jitter(width = .1))
```

ILLUSTRATION




```
1 p + geom_point(position = position_jitter(width = .1), alpha = .5)
```

ILLUSTRATION




```
p <- p + geom_point(position = position_jitter(width = .1), alpha = .5)
p + scale_y_continuous(limits = c(0,10)) +
  labs(x = "Age (années)", y = "Score visuel")
```

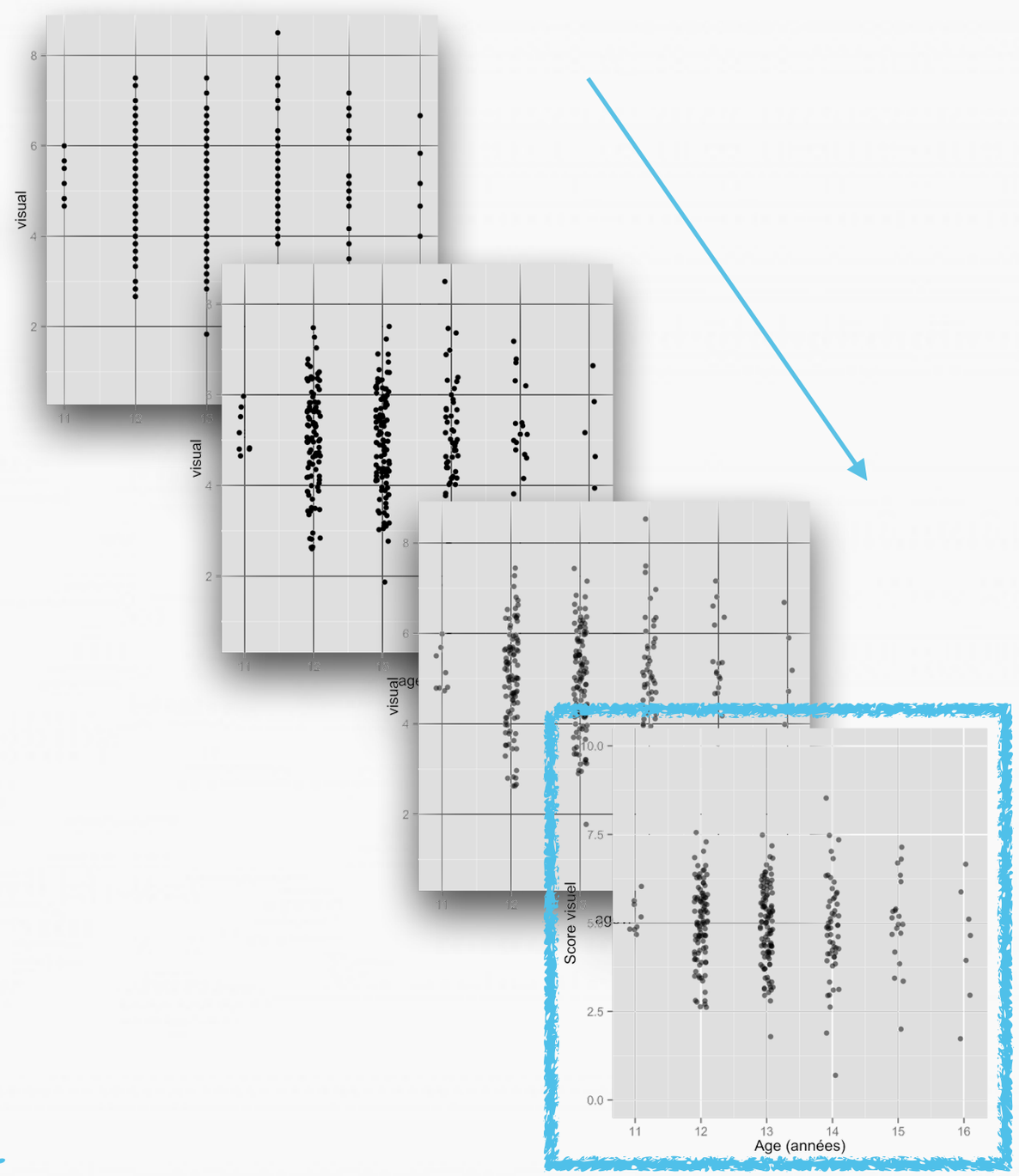
CONSTRUCTION ITERATIVE

1. Coordonnées x (variable ageyr) et y (variable visual)

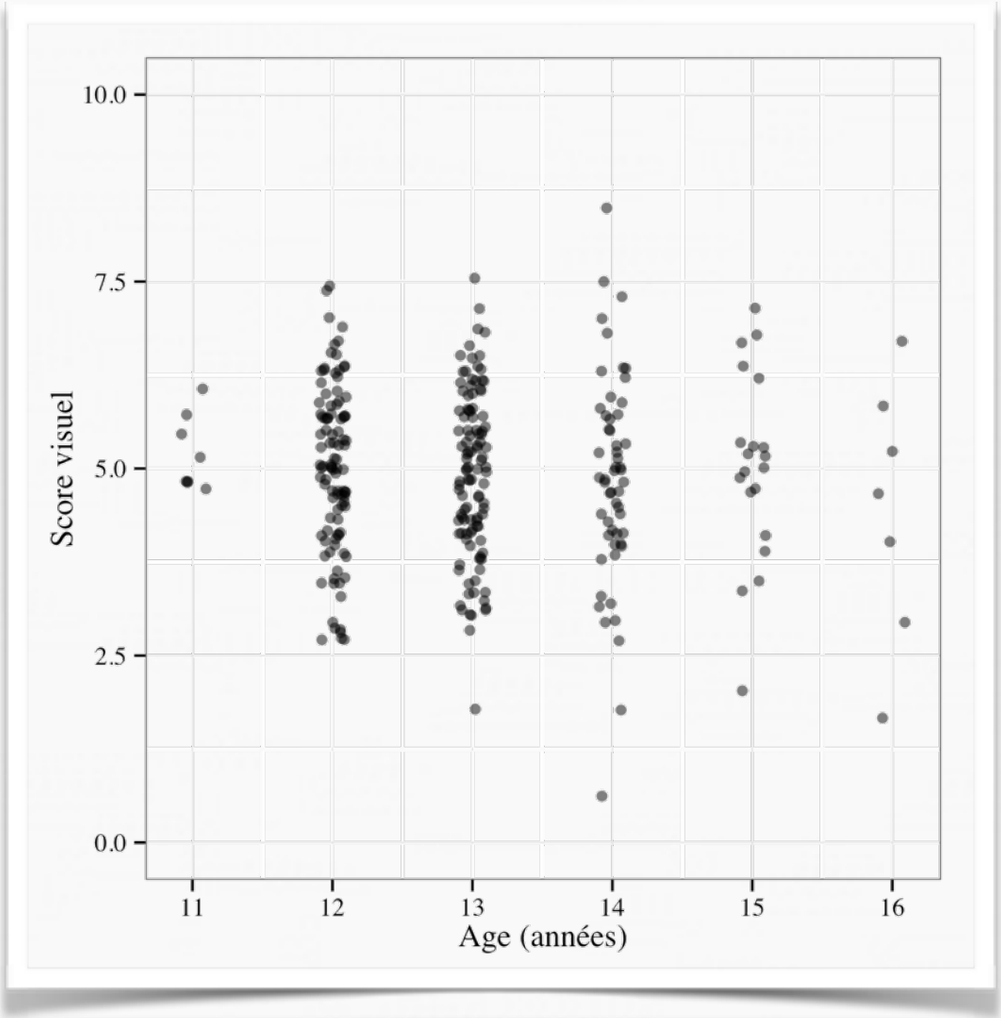
2. forme géométrique = points

3. propriétés spatiales des points = décalage horizontale + transparence

4. système de coordonnées et annotation des axes



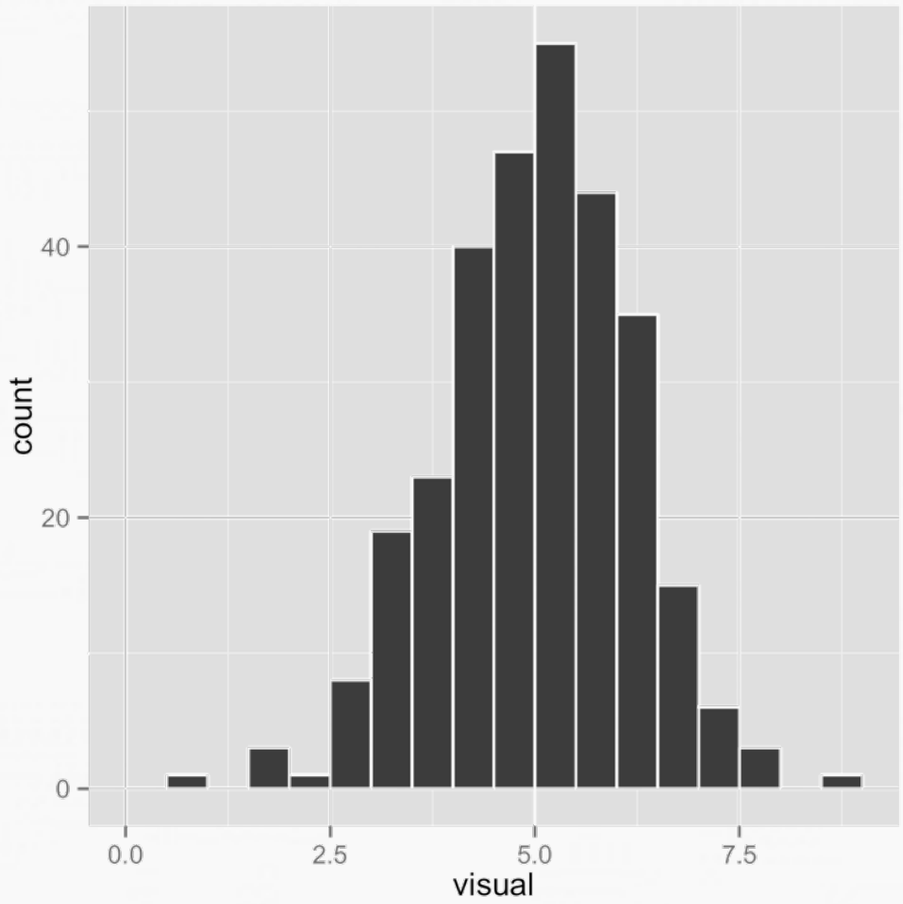
THEME




```
p <- p + scale_y_continuous(limits = c(0,10)) +
  labs(x = "Age (années)", y = "Score visuel")
p + theme_bw(base_size = 12, base_family = "Times")
```

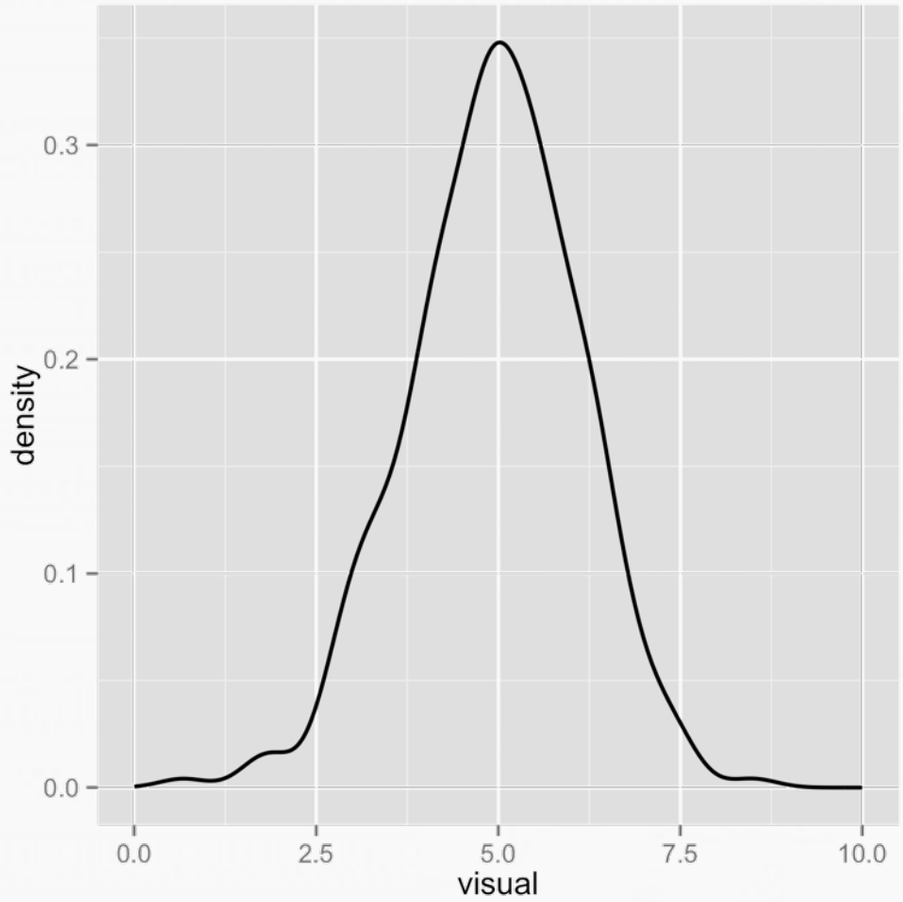
DISTRIBUTIONS UNIVARIEES

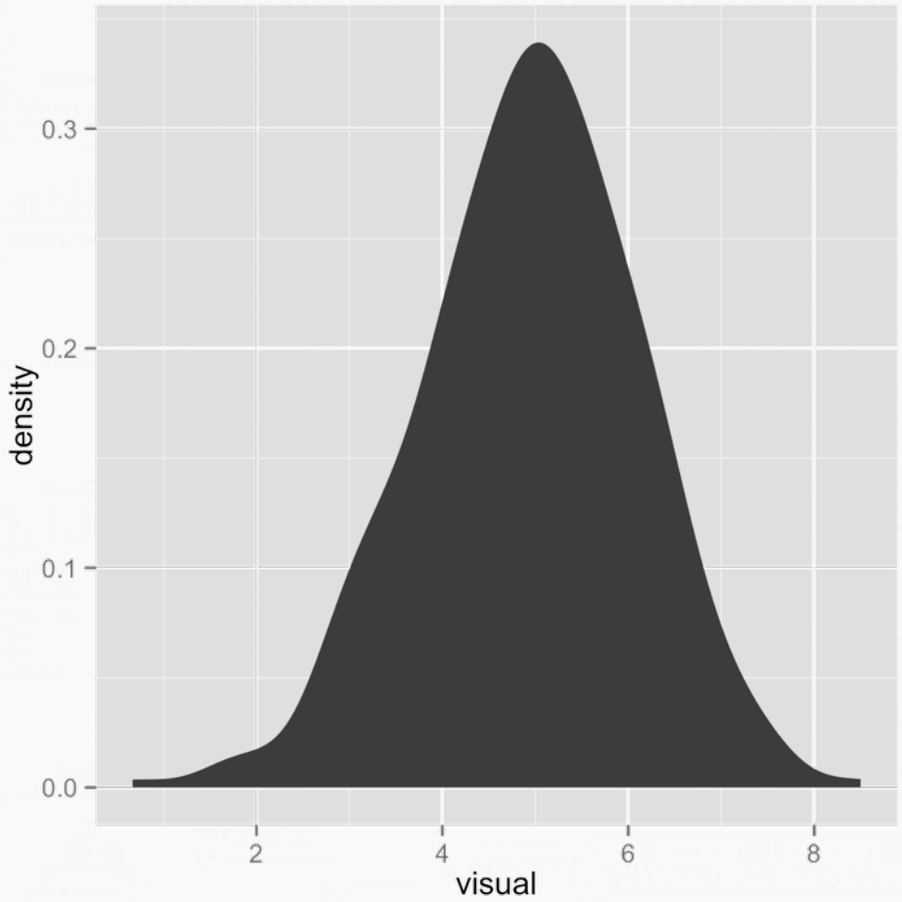
HISTOGRAMMES




```
1 p <- ggplot(data = HS, aes(x = visual))
2 p + geom_histogram(binwidth = .5, fill = "grey30", colour = "white")
3 p + geom_histogram(aes(y = ..density..), binwidth = .5, fill = "grey30", colour = "white")
```

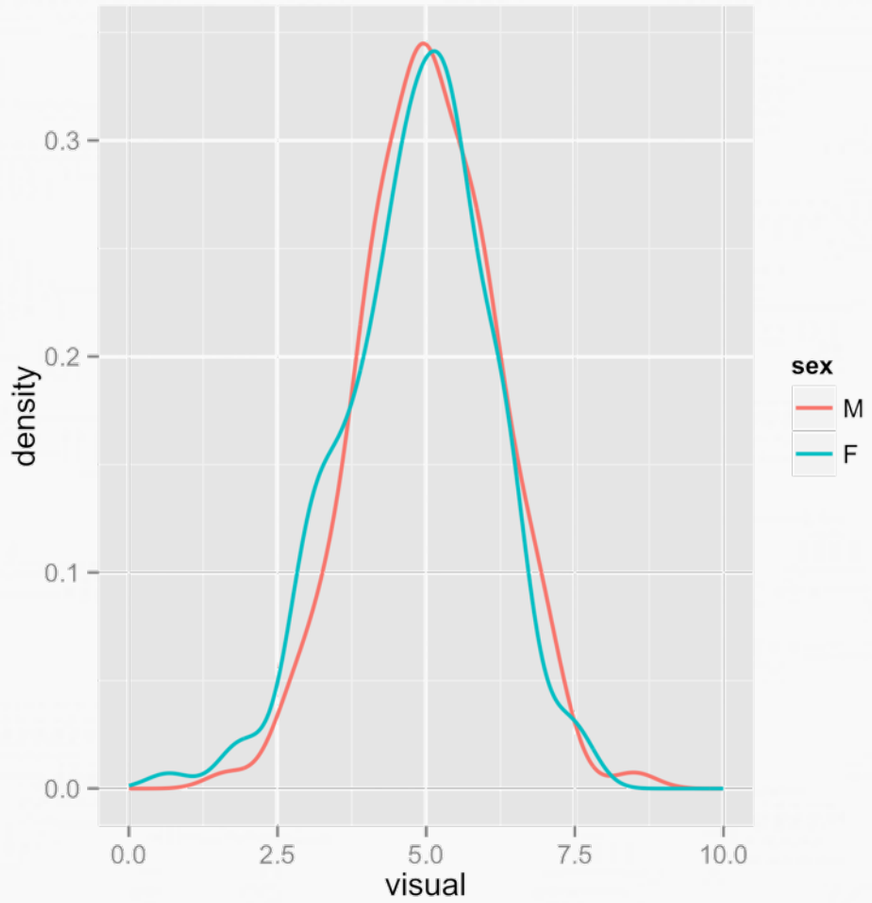
COURBES DE DENSITE





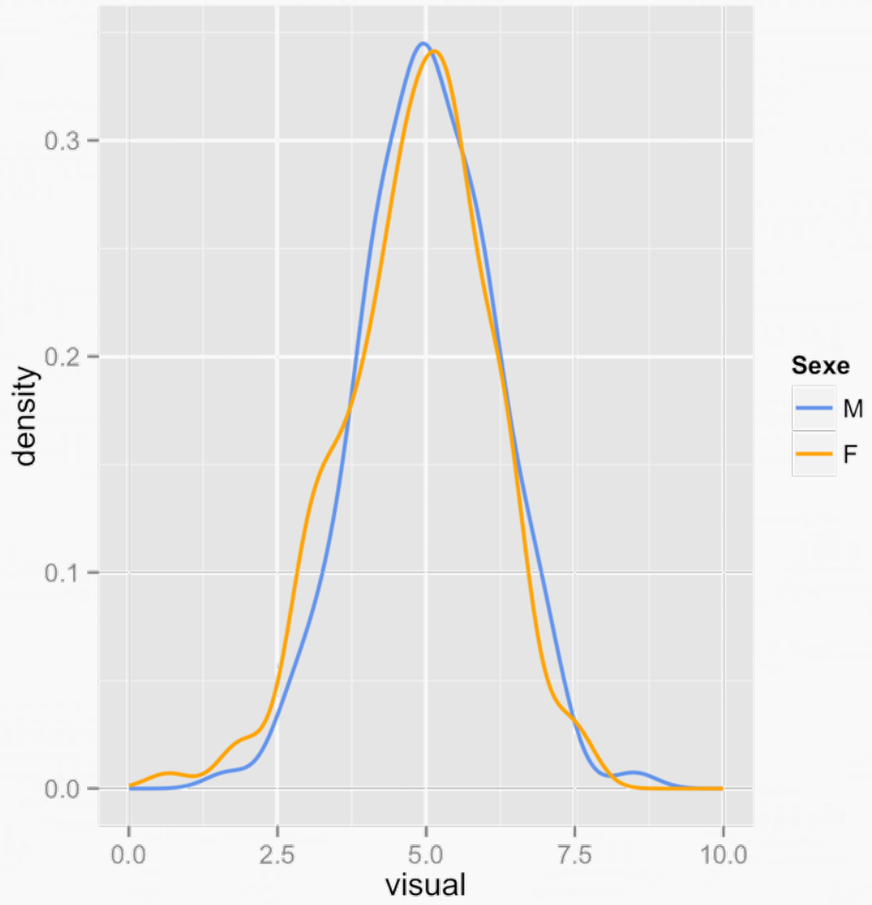

```
1 p + geom_line(stat = "density", size = .7)
2 p + geom_line(stat = "density", size = .7) + expand_limits(x = c(0,10))
3 p + geom_line(stat = "density", adjust = 2, size = .7)
4 p + geom_density(fill = "grey30", adjust = 1, colour = "transparent")
```

COURBES DE DENSITE



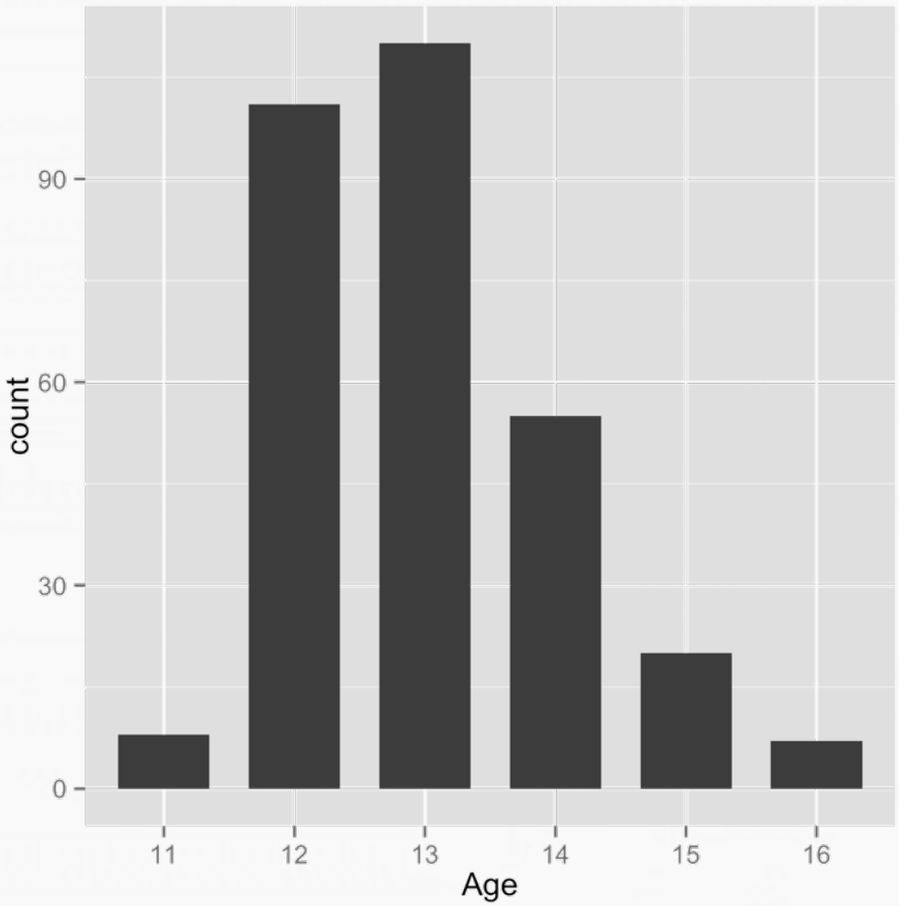

```
1 p <- ggplot(data = HS, aes(x = visual, colour = sex))
2 p + geom_line(stat = "density", size = .7) + expand_limits(x = c(0,10))
```

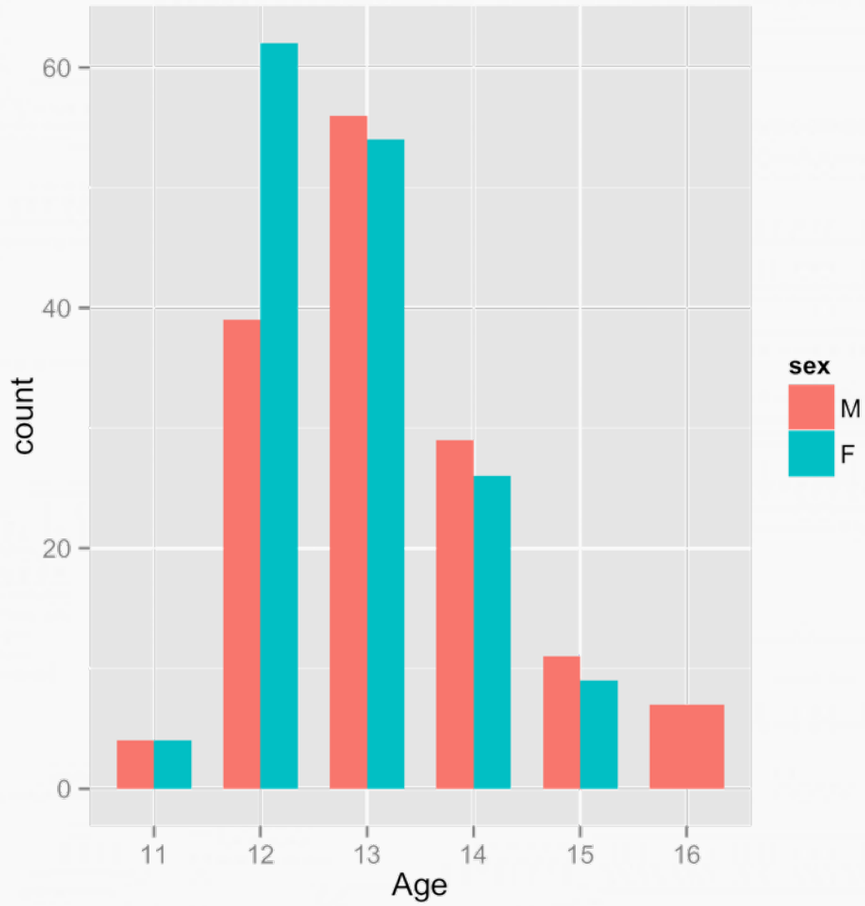
COURBES DE DENSITE




```
1 p <- p + geom_line(stat = "density", size = .7) + expand_limits(x = c(0,10))
2 p + scale_colour_manual(values = c("cornflowerblue", "orange"), name = "Sexe")
```

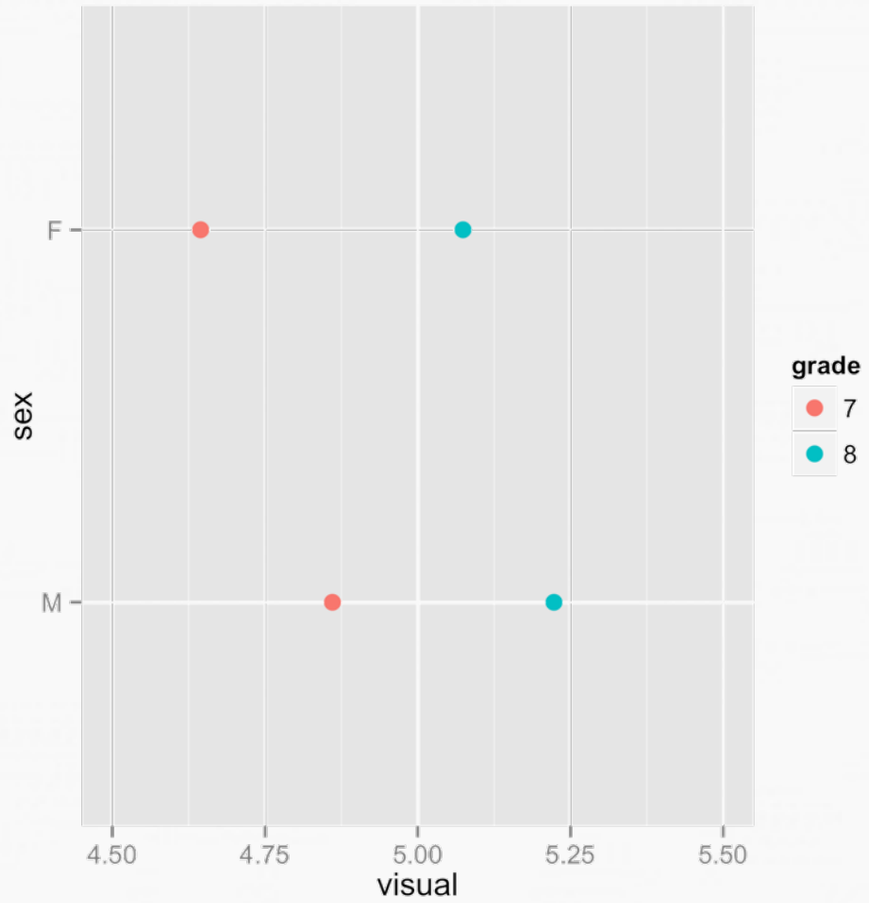
DIAGRAMMES EN BARRES






```
1 p <- ggplot(data = HS, aes(x = factor(ageyr)))
2 p + geom_bar(width = .7, fill = "grey30") + labs(x = "Age")
3
4 p <- ggplot(data = HS, aes(x = factor(ageyr)))
5 p + geom_bar(width = .7, aes(fill = sex), position = position_dodge()) + labs(x = "Age")
```

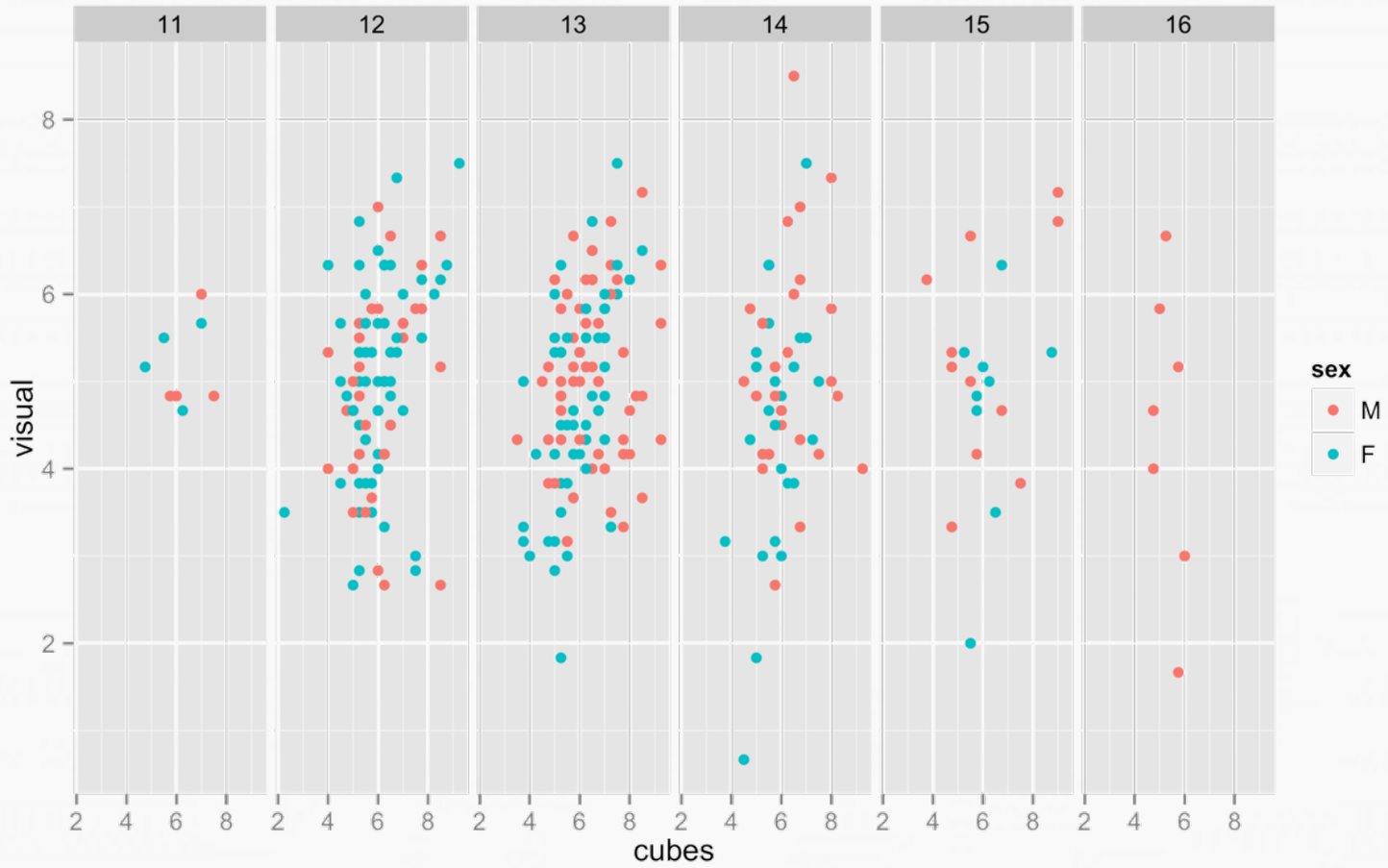
DIAGRAMMES EN POINTS




```
1 d <- aggregate(visual ~ sex + grade, data = HS, mean)
2 p <- ggplot(data = d, aes(x = sex, y = visual, colour = grade))
3 p + geom_point(size = 3) + ylim(c(4.5,5.5)) + coord_flip()
```

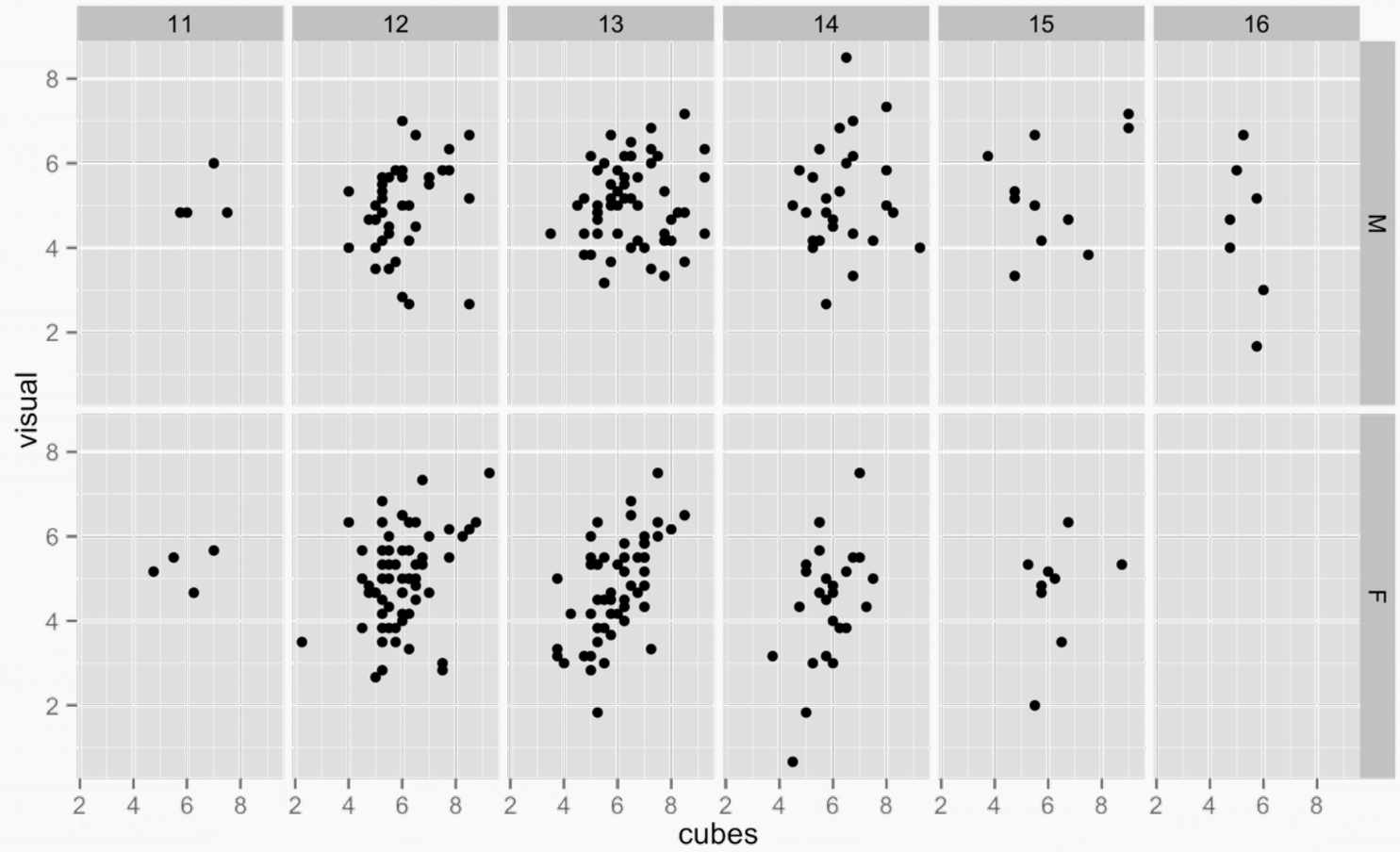
SYSTEME DE FACETTES

DIAGRAMME DE DISPERSION



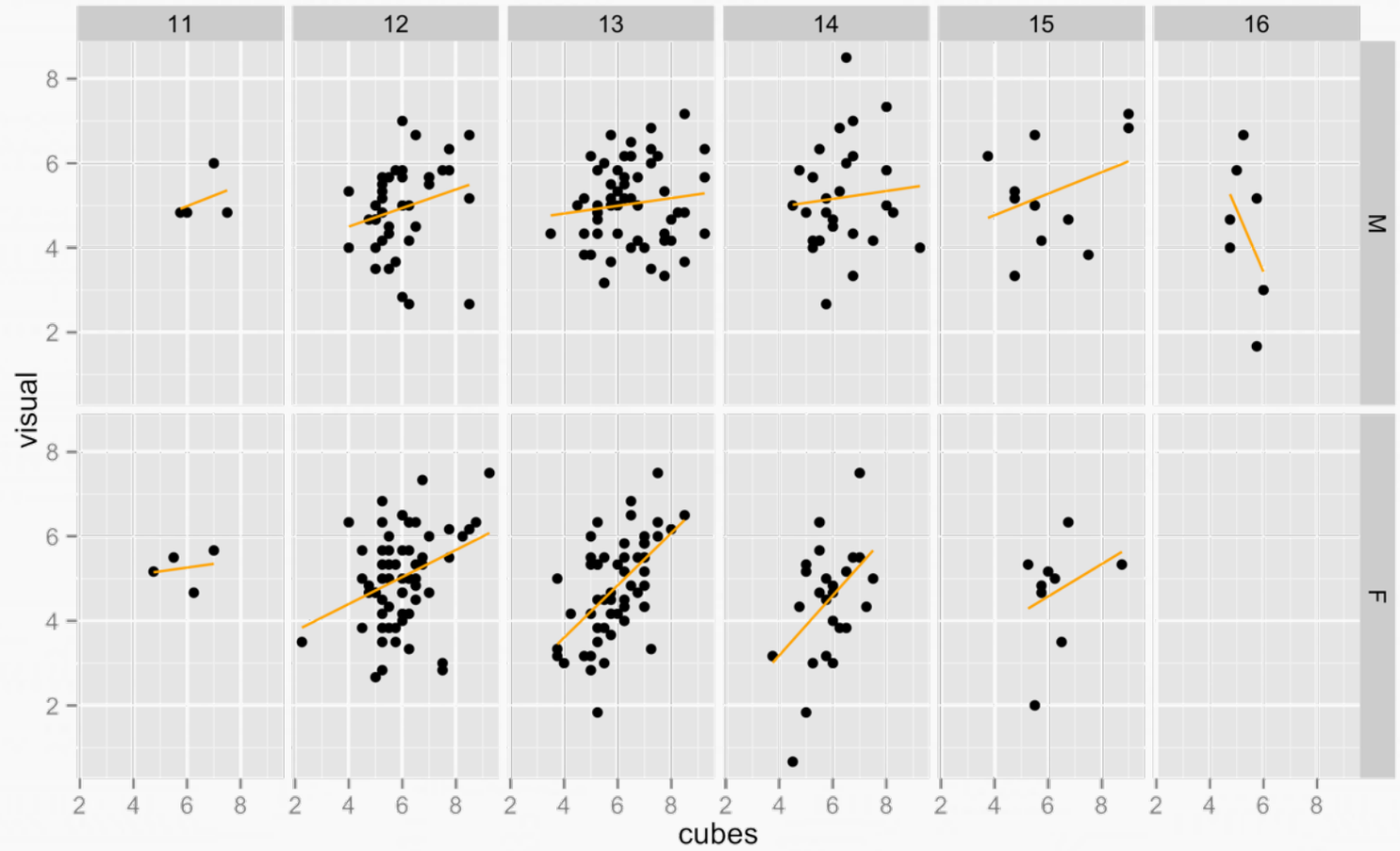

```r
p <- ggplot(data = HS, aes(x = cubes, y = visual, colour = sex))
p + geom_point()
p + geom_point() + facet_grid(~ageyr)
```

DIAGRAMME DE DISPERSION



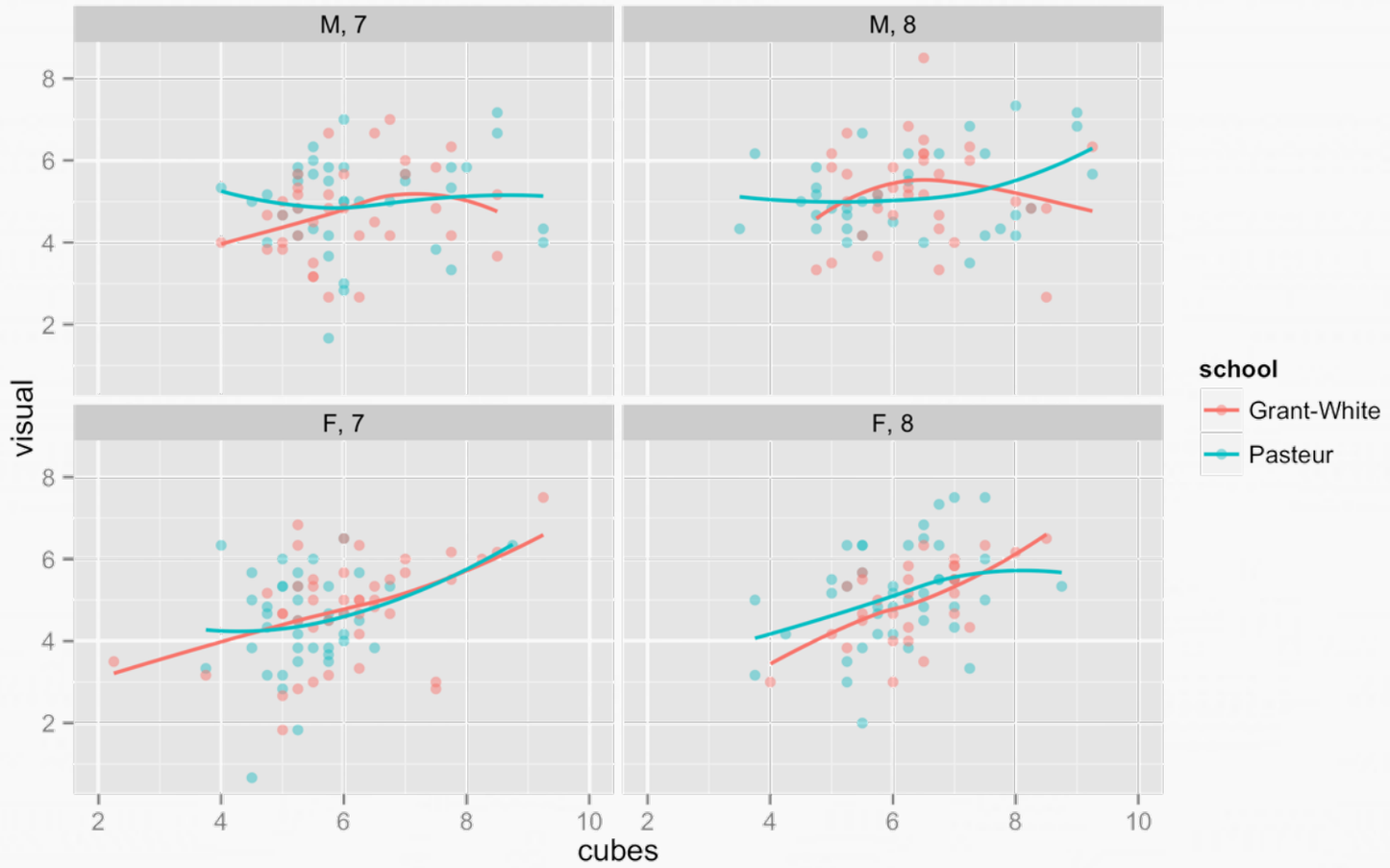

```
1 p <- ggplot(data = HS, aes(x = cubes, y = visual))
2 p + geom_point() + facet_grid(sex ~ ageyr)
```

DIAGRAMME DE DISPERSION




```
1 p <- p + geom_point() + facet_grid(sex ~ ageyr)
2 p + geom_smooth(method = "lm", se = FALSE, colour = "orange")
```

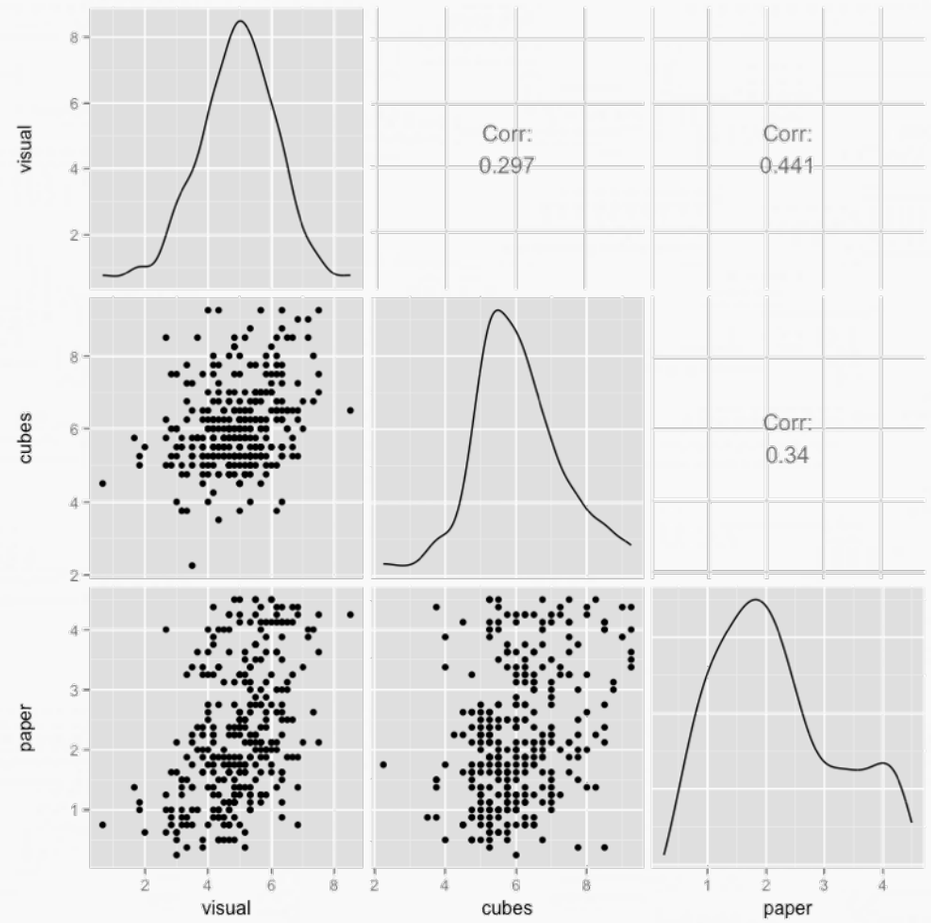
DIAGRAMME DE DISPERSION




```r
p <- ggplot(data = subset(HS, complete.cases(grade)),
            aes(x = cubes, y = visual, colour = school))
p <- p + geom_point(alpha = .5) + geom_smooth(span = 1.5, size = .7, se = FALSE)
p + scale_x_continuous(limits = c(2,10)) + facet_wrap(sex ~ grade)
```

EXTENSIONS

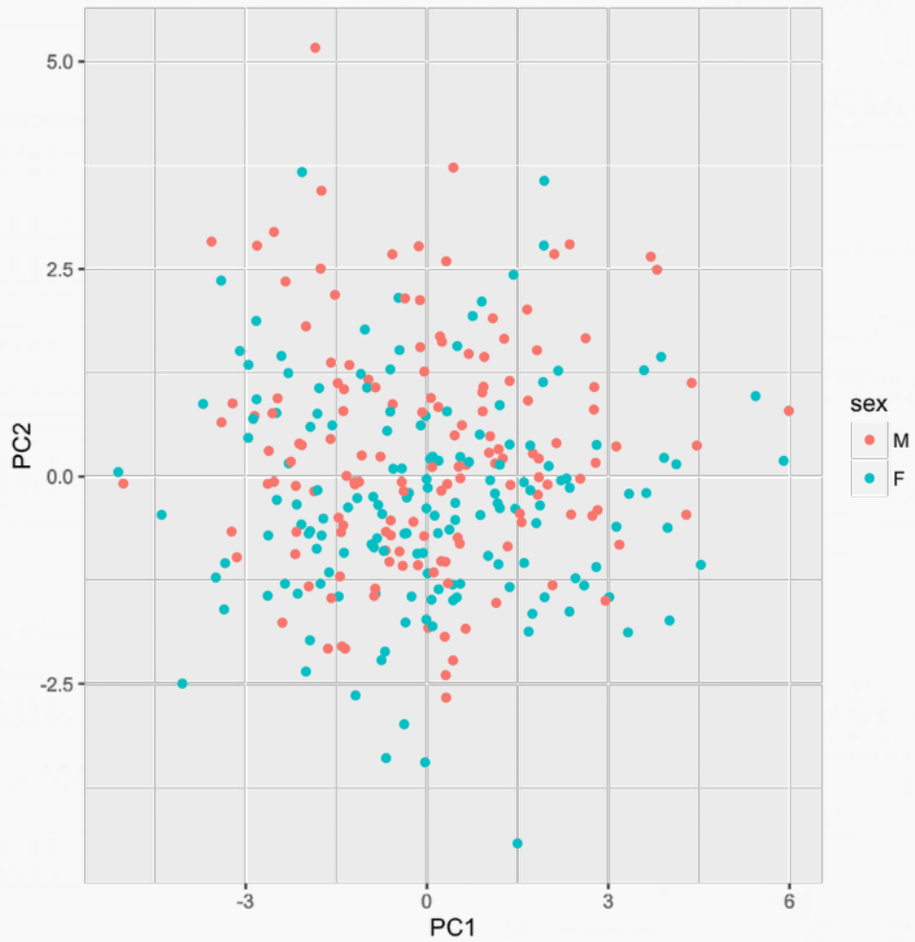
GGALLY

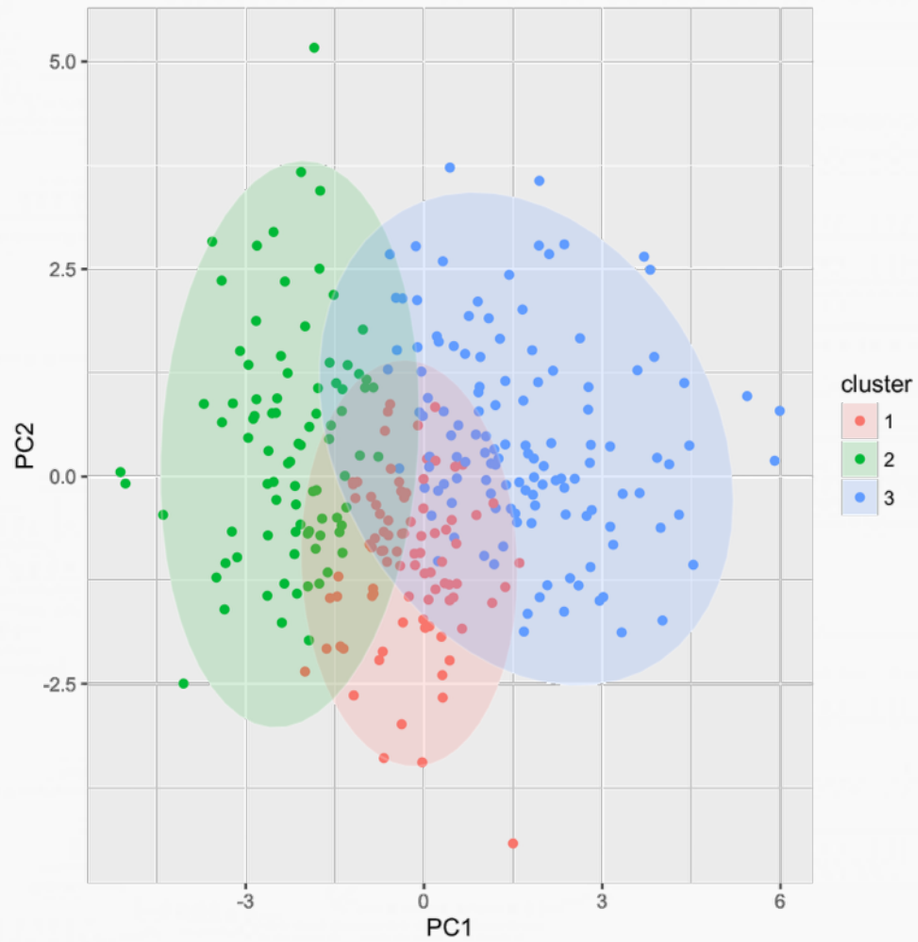



```
1 library(GGally)
2 ggpairs(HS[,c("visual","cubes","paper")])
```

http://ggobi.github.io/ggally/

GGFORTIFY






```
1 library(ggfortify)
2 pca <- prcomp(HS[,7:12])
3 autoplot(pca, data = HS, colour = "sex")
4 autoplot(cluster::pam(HS[,7:12], k = 3), frame = TRUE, frame.type = "norm")
```