

ggplot2

a whistlestop tour from basics to extensions

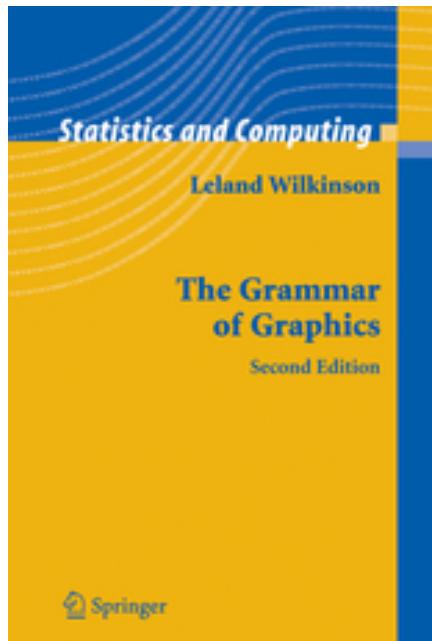


Constantin Ahlmann-Eltze
 @const-ae

Outline

- ggplot's place in the tidyverse
- How to use ggplot
- Data visualization best practices
- The Grammar of Graphics
- Creating a ggplot extension

What does that name even mean?



+

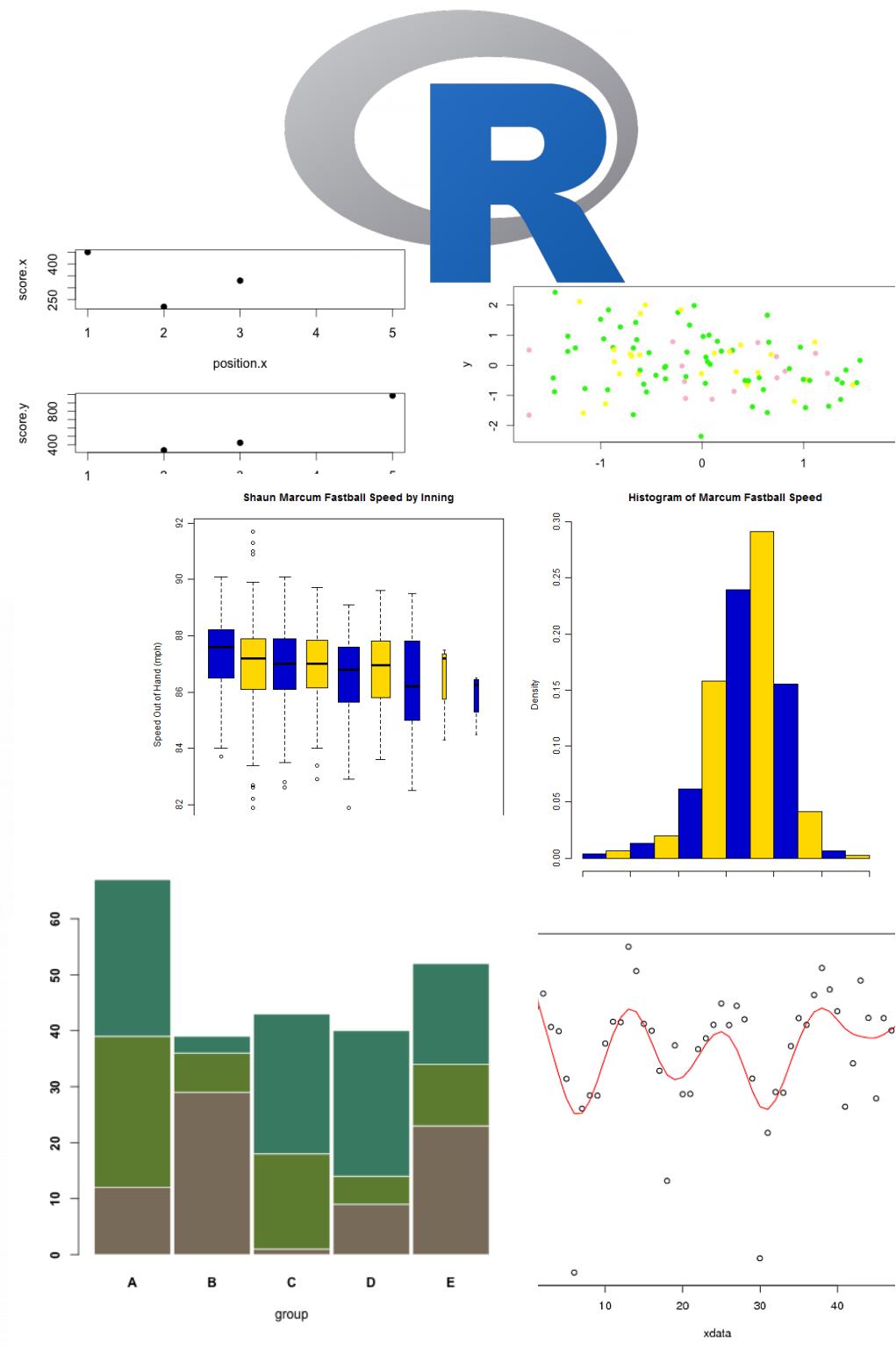
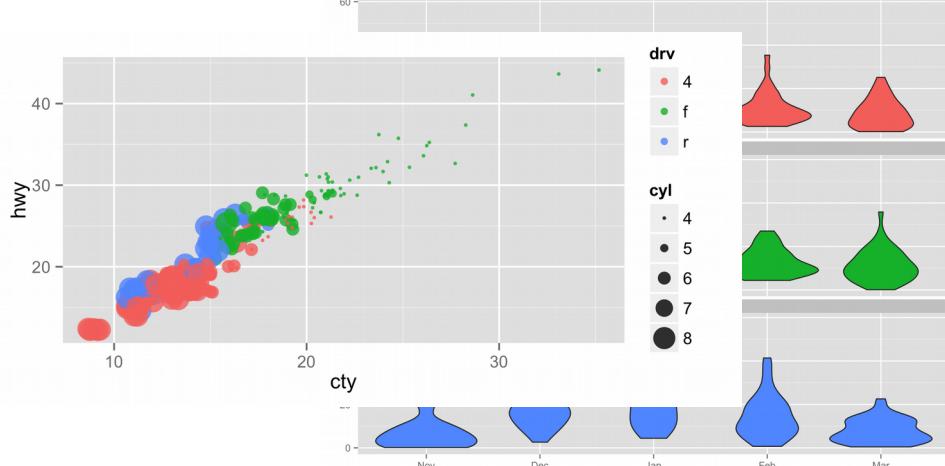
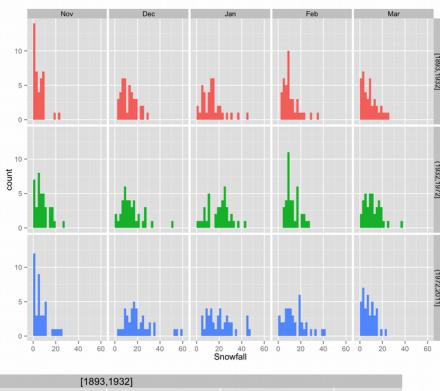
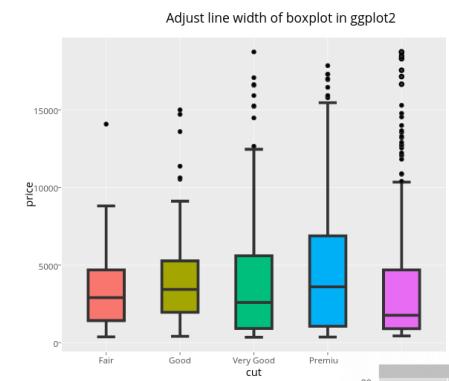
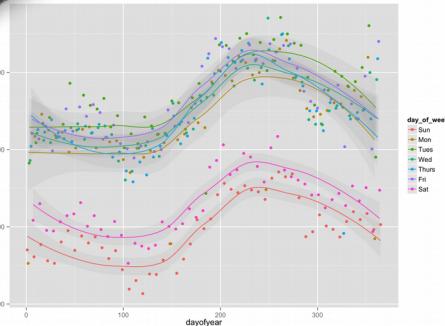
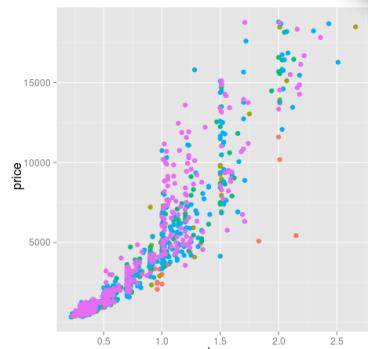


=



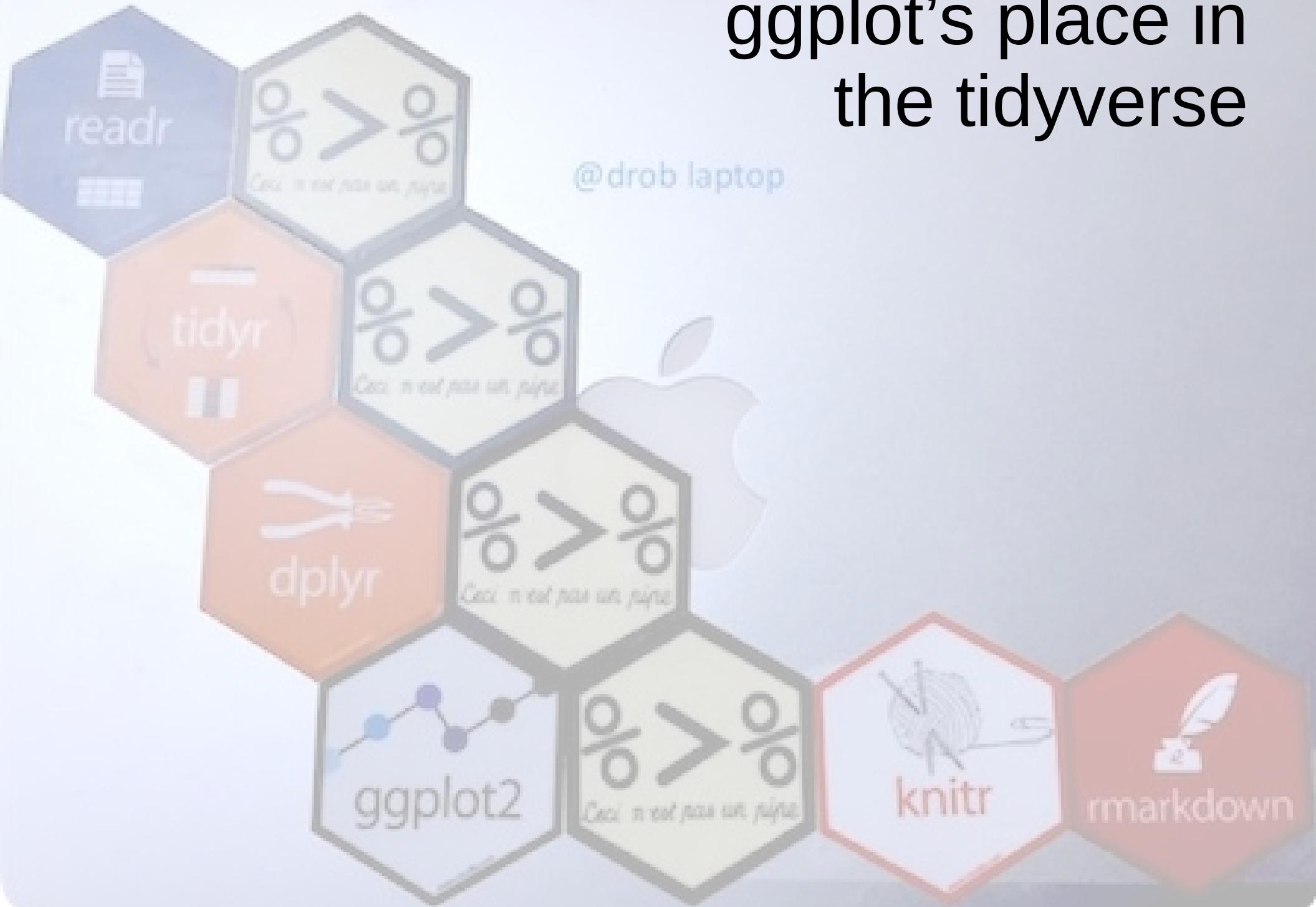
Grammar of Graphics

Plotter

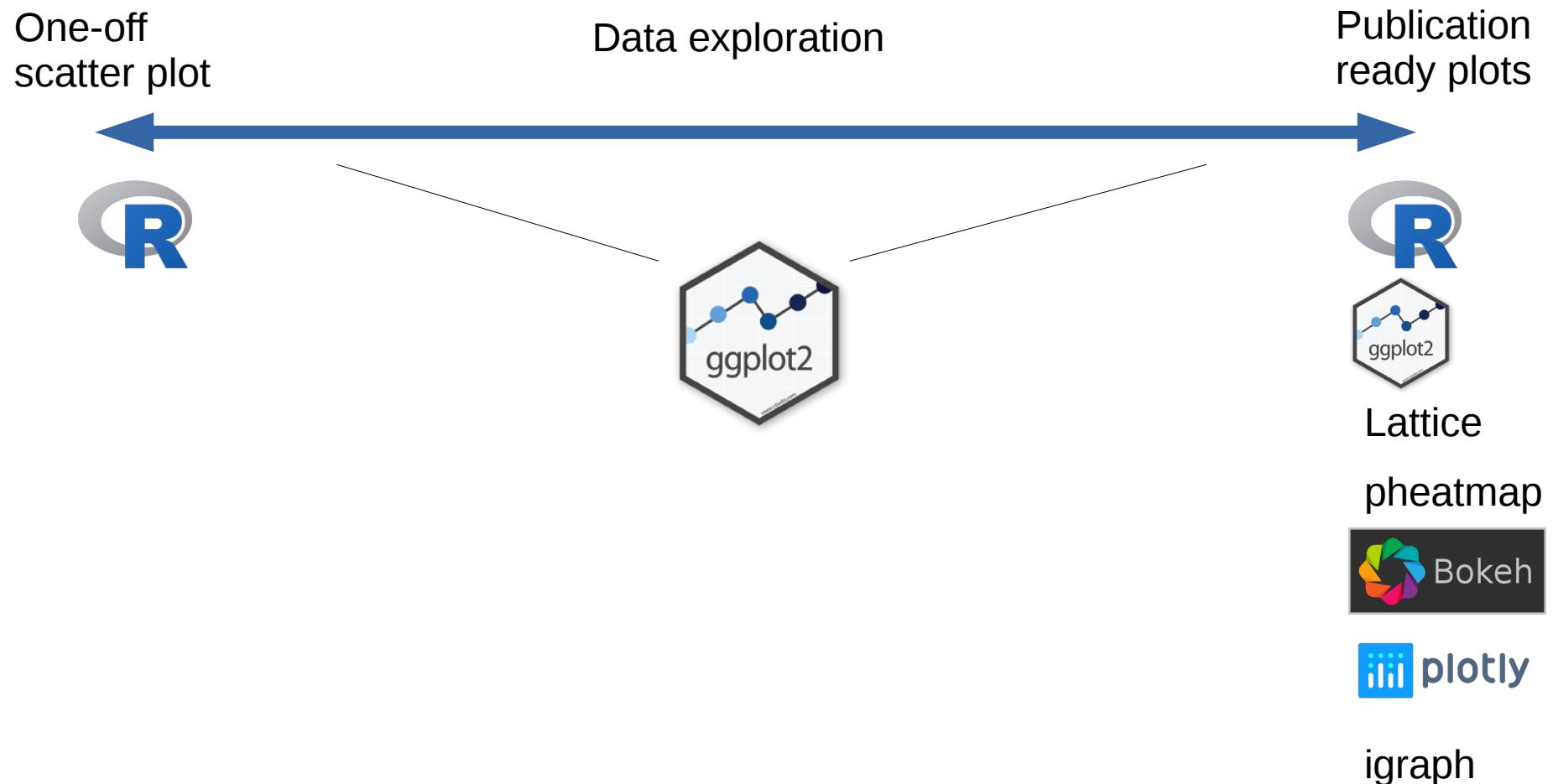


ggplot's place in the tidyverse

@drob laptop



Quickly iterate with ggplot



The Data Science Workflow

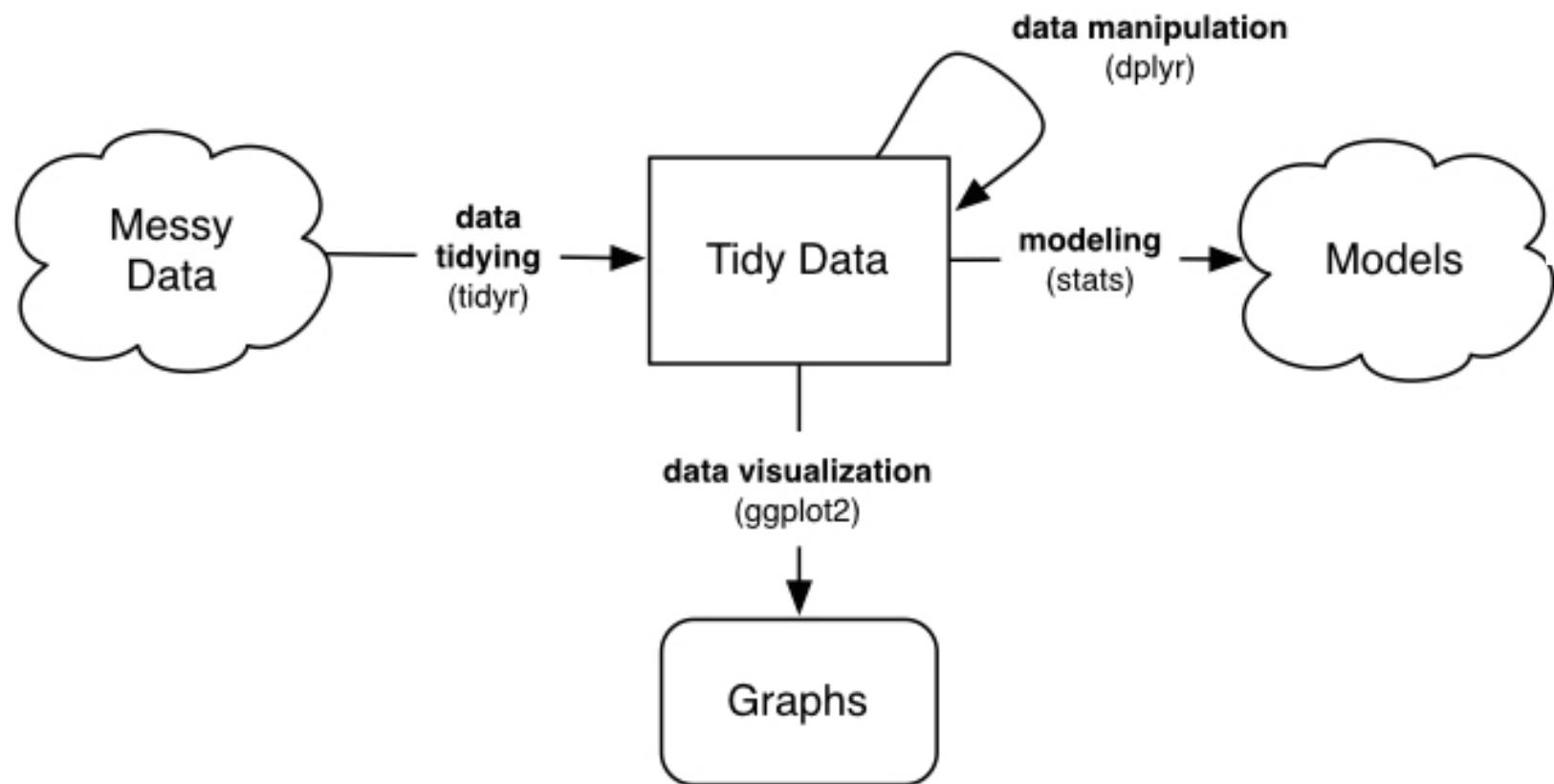


Figure adapted from <http://varianceexplained.org/r/tidy-genomics-biobroom/>

Untidy Data

SeasonStart	SeasonEnd	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
1893	1894	0	0	0	0	8	24.9	12.5	6.8	4.8	2	0	0
1894	1895	0	0	0	0	7.5	5.3	21.5	8	22.5	0	0	0
1895	1896	0	0	0	0.4	23.2	15	NA	8.5	2	0	0	0
1896	1897	0	0	0	0.2	8	8	4.9	11.2	12	0	0	0
1897	1898	0	0	0	0	1.4	8	15.5	29.5	0	0	0	0
1898	1899	0	0	0	0	18.5	18	20	3.4	16	0	0	0
1899	1900	0	0	0	0	1.5	7.2	12	35.5	19	3.5	0	0
1900	1901	0	0	0	0	8	5.5	22.5	23	5.5	0	0	0
1901	1902	0	0	0	0	0.5	9	4	11	3.5	0.3	5.5	0
1902	1903	0	0	0	0	2.5	5.1	7	5.3	NA	0.3	0	0
1903	1904	0	0	0	0	5	22.1	15.6	11.9	20	3.2	0	0
1904	1905	0	0	0	0	0.6	15.7	14.5	18.8	2.4	1.7	0	0

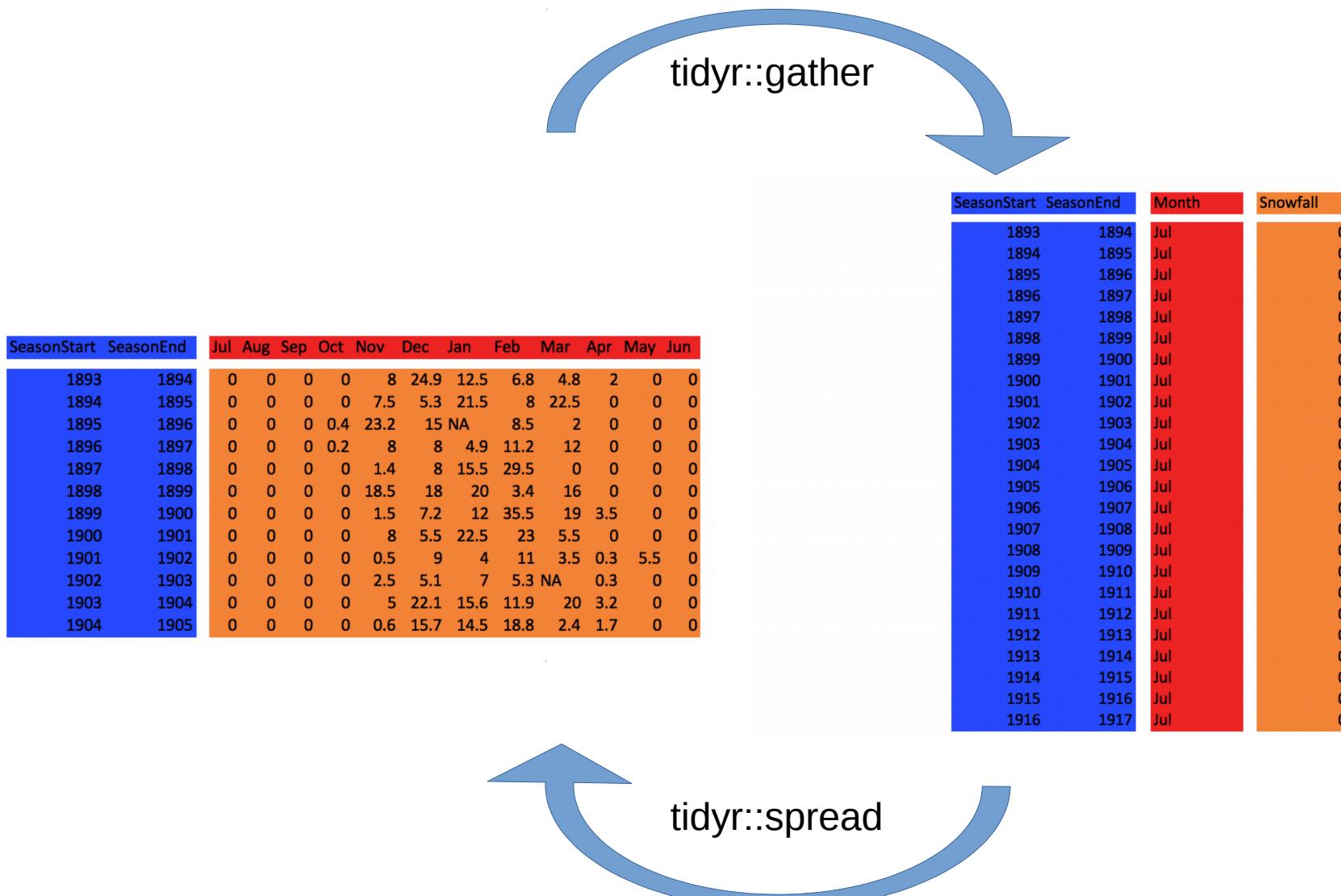
Wide format data

Tidy Data

SeasonStart	SeasonEnd	Month	Snowfall
1893	1894	Jul	0
1894	1895	Jul	0
1895	1896	Jul	0
1896	1897	Jul	0
1897	1898	Jul	0
1898	1899	Jul	0
1899	1900	Jul	0
1900	1901	Jul	0
1901	1902	Jul	0
1902	1903	Jul	0
1903	1904	Jul	0
1904	1905	Jul	0
1905	1906	Jul	0
1906	1907	Jul	0
1907	1908	Jul	0
1908	1909	Jul	0
1909	1910	Jul	0
1910	1911	Jul	0
1911	1912	Jul	0

Long format data

Convert wide to long format



How to use ggplot

```
1  
2  
3 library(ggplot2)  
4  
5 - test_that("the plotting works", {  
6   print(  
7     ggplot(mpg, aes(x=manufacturer, y=displ), .  
8       geom_boxplot() +  
9       stat_signif(comparisons=list(c("audi", "ford"), c("hyundai", "nissan")),  
10          map_signif_level=TRUE,  
11          test="wilcox.test", test.args=list(alternative="two.sided"),  
12          margin_top=0.02, step_increase=0, tip_length=0.01) +  
13        theme(axis.text.x = element_text(angle = 90, hjust = 1)) +  
14        facet_wrap(~ as.factor(year), scale="free")  
15    )  
16 })  
17  
18 - test_that("geom works as well as stat works", {  
19   print(  
20     ggplot(mpg, aes(x=manufacturer, y=displ)) +  
21     geom_boxplot() +  
22     geom_signif(comparisons=list(c("audi", "ford"), c("hyundai", "nissan")),  
23                   annotations=c("Interesting", "Too far apart"),  
24                   # map_signif_level=TRUE,  
25                   test="wilcox.test", test.args=list(alternative="two.sided"),  
26                   margin_top=0.02, step_increase=0, tip_length=0.01) +  
27     theme(axis.text.x = element_text(angle = 90, hjust = 1)) +  
28     facet_wrap(~ as.factor(year), scale="free")  
29   )  
30 })  
31
```

Carprice Data

	Type <fctr>	Price <dbl>	MPG.city <dbl>	MPG.highway <dbl>
6	Midsize	15.7	22	31
7	Large	20.8	19	28
8	Large	23.7	16	25
9	Midsize	26.3	19	27
10	Large	34.7	16	25
11	Midsize	40.1	16	25
12	Compact	13.4	25	36
13	Compact	11.4	25	34
14	Sporty	15.1	19	28
15	Midsize	15.9	21	29

1-10 of 48 rows

Previous 1 2 3 4 5 Next

ggplot Call

```
ggplot(carprize, aes(x=Price, y=MPG.city, color=Type)) +  
  geom_point()
```

The data in a data.frame

Mapping columns to **aesthetics**

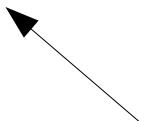
Step by Step adding elements

Layer to draw points

```
graph TD; A[The data in a data.frame] --> B[carprize]; C[Mapping columns to aesthetics] --> D[aes(x=Price, y=MPG.city, color=Type)]; E[Step by Step adding elements] --> F["+"]; G[Layer to draw points] --> H[geom_point()];
```

ggplot Call

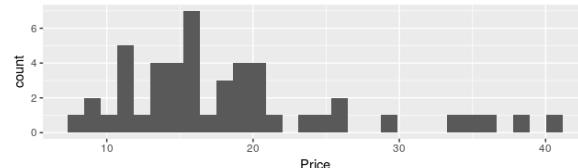
```
ggplot(carprize, aes(x=Price, y=MPG.city, color=Type) +  
  geom_point() +  
  geom_smooth(method="lm")
```



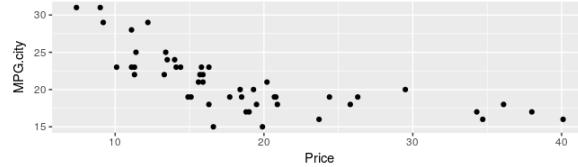
Plot linear
trend line

Layer Overview

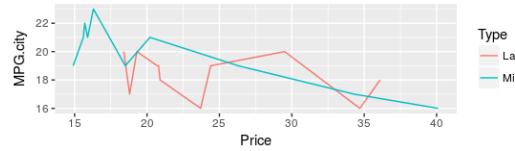
geom_histogram



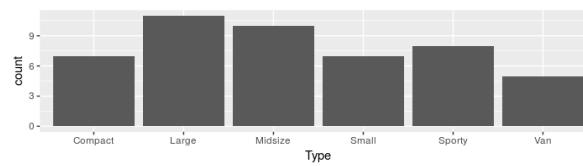
geom_point



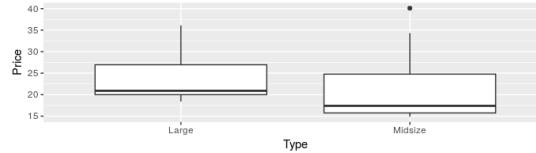
geom_line



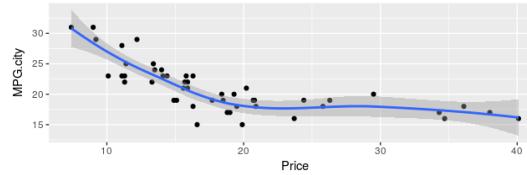
geom_bar



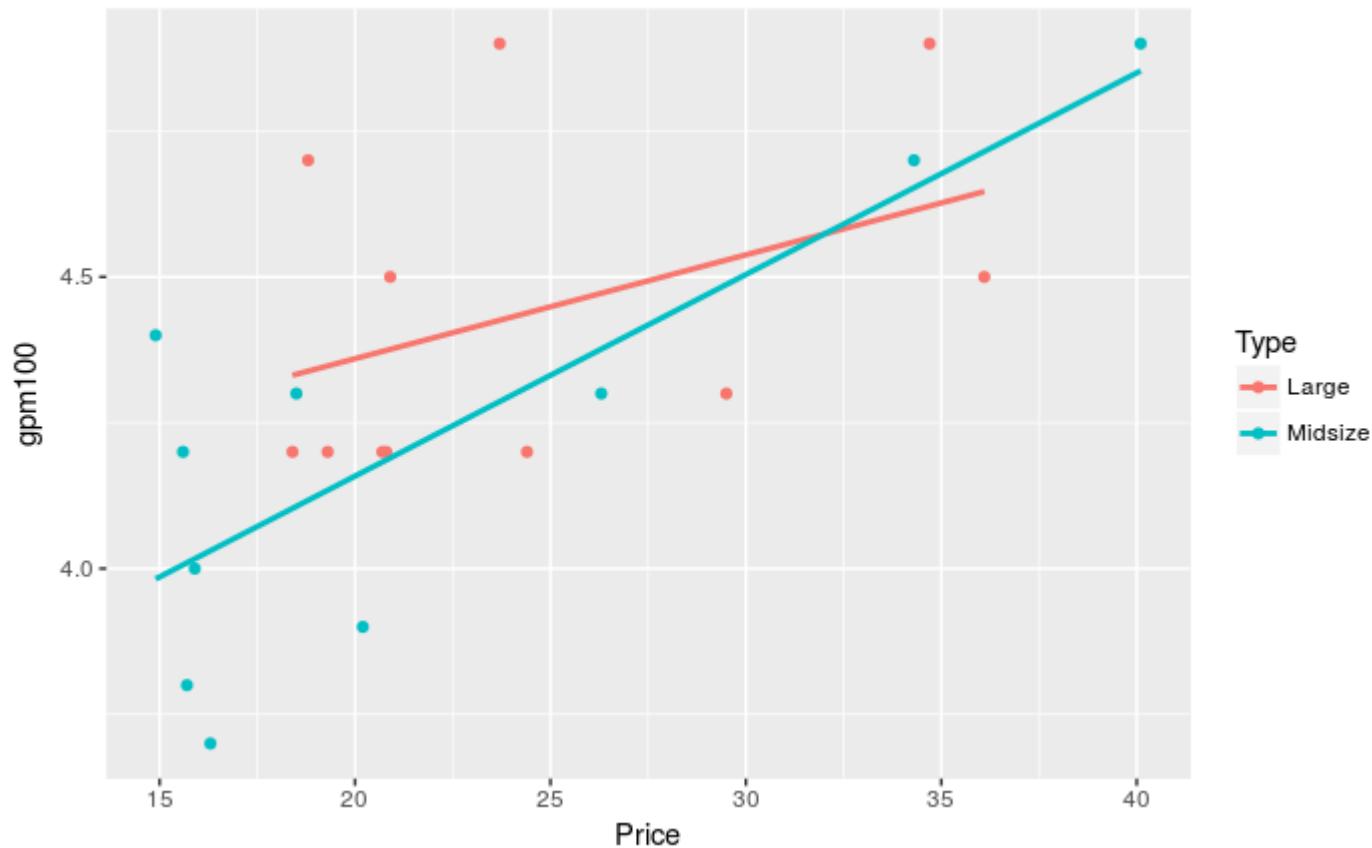
geom_boxplot



geom_smooth



Demo



ggplot_demo.Rmd

Visualization Best Practices

Carte Figurative des pertes d'hommes en bataille de l'Armée Française dans la Campagne de Russie 1812-1813.

Dessiné par M. Minard, Inspecteur Général des Ponts et Chaussées en octobre.

Paris, le 20 Novembre 1869.

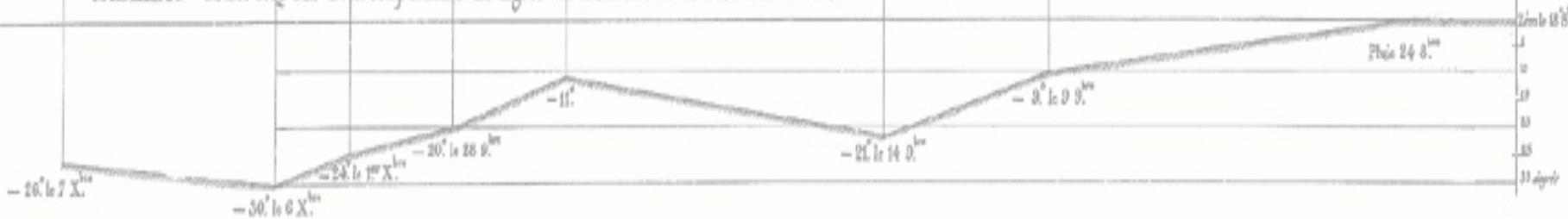
Les nombres d'hommes perdus sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres sur ces zones. Le rouge distingue les hommes qui ont été tués en Russie, le noir ceux qui en sont revenus. — Les renseignements qui ont servi à l'établissement de ce tableau sont tirés des rapports de l'Inspection générale des ouvrages de M. M. Chiers, de Segur, de Tézot, de Chambray et le journal médical de Jacob, pharmacien de l'Armée depuis le 1^{er} juillet 1812 jusqu'à l'arriver à l'armée; j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été défaits à la bataille de Wilschau et à celle de Witebsk avaient toujours marché avec l'armée.

Longtemps faire juger à l'aïe l'immensité de l'armée; j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été défaits à la bataille de Wilschau et à celle de Witebsk, avaient toujours marché avec l'armée.

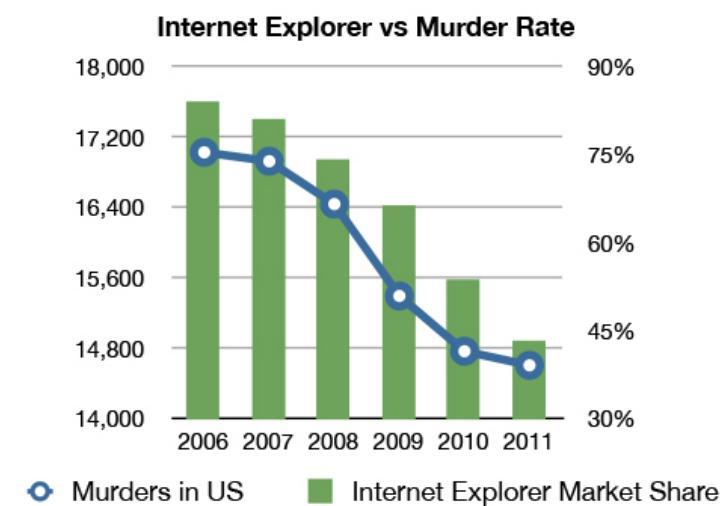
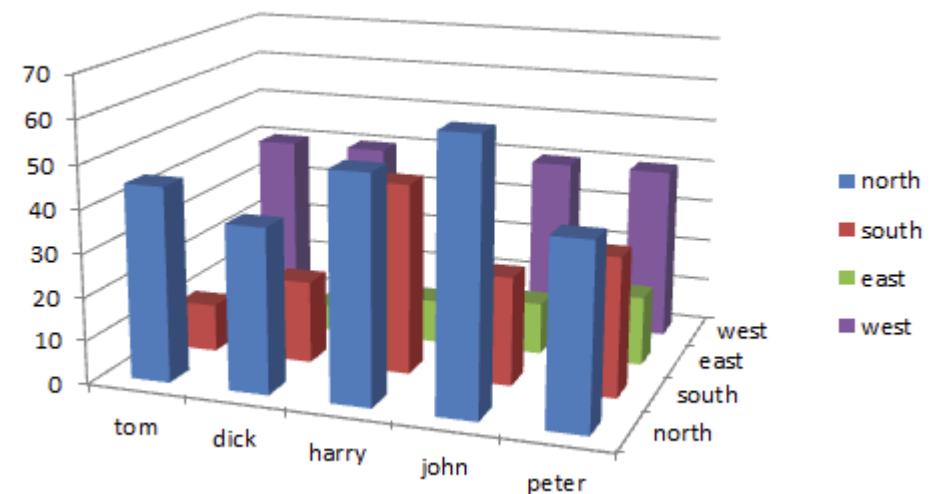
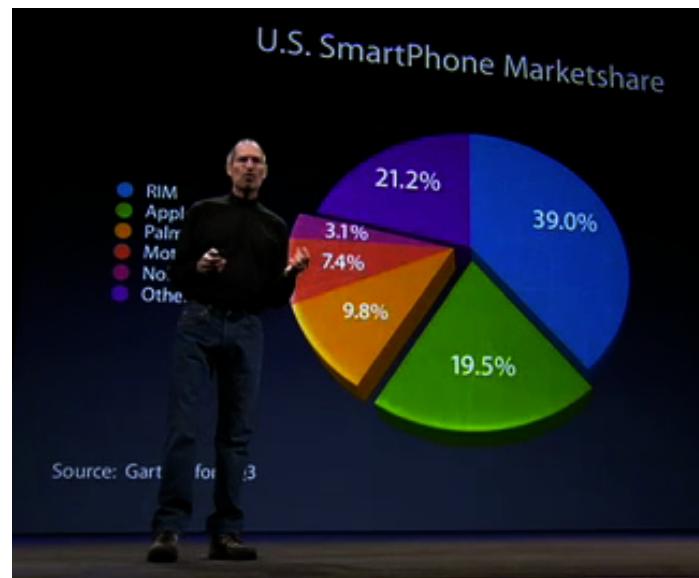


TABLEAU CRATIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

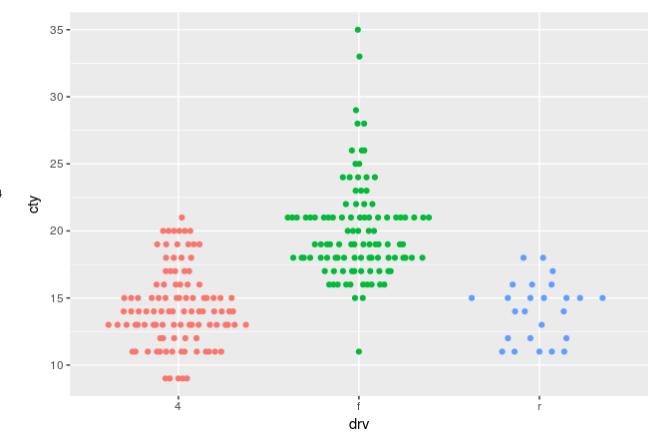
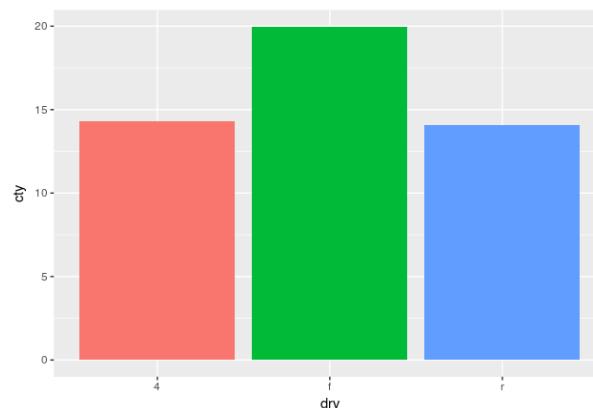
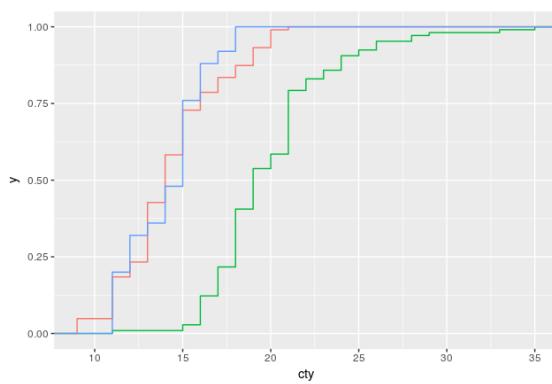
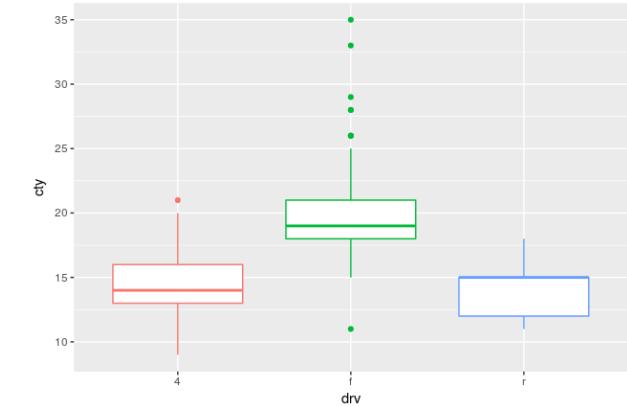
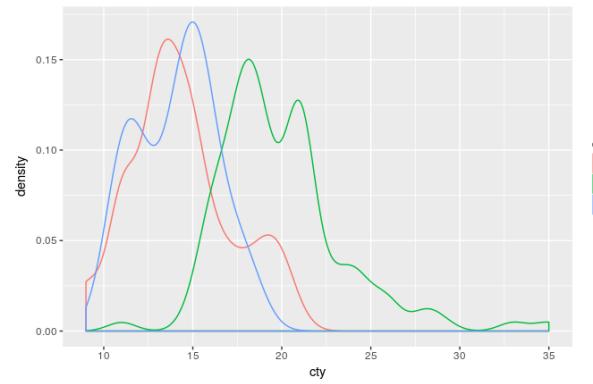
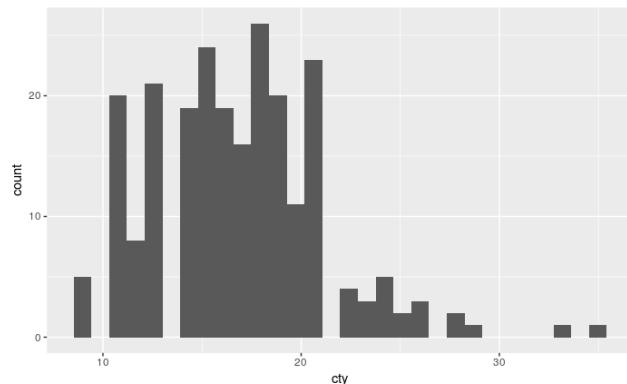
Les troupes passent au gelé le Nivôze, 1^{er}.



How not to do data visualization



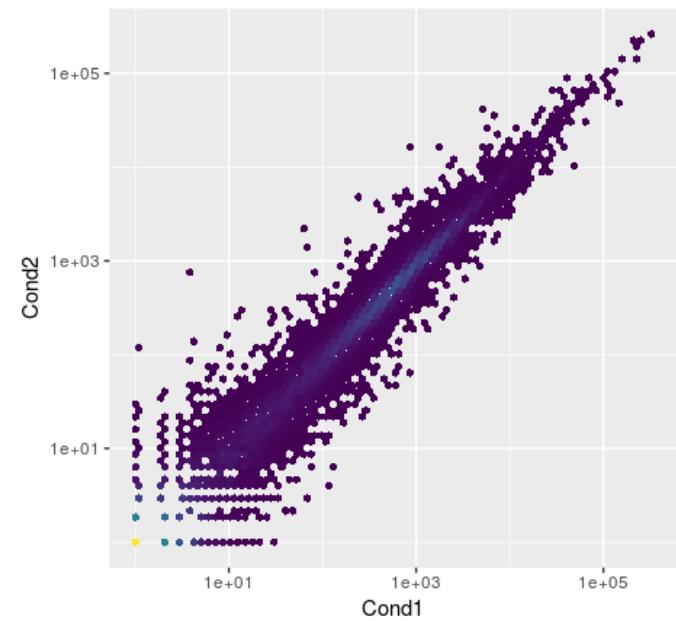
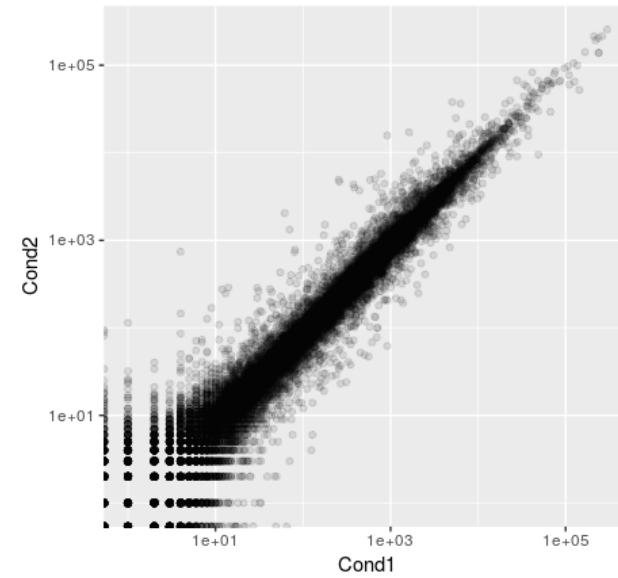
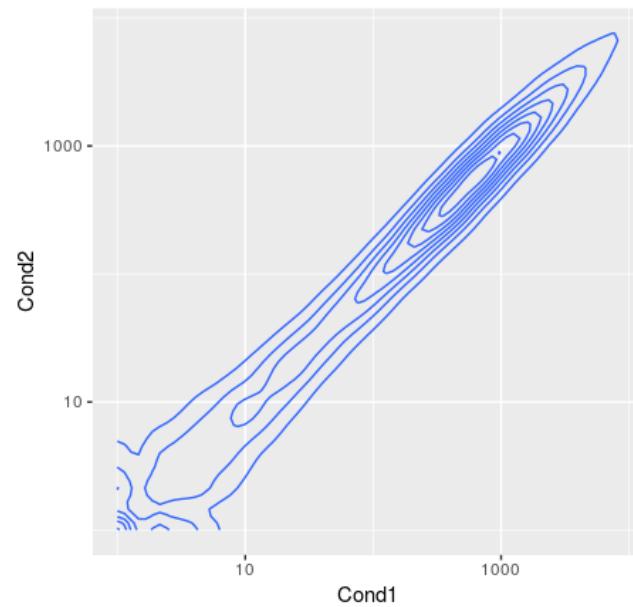
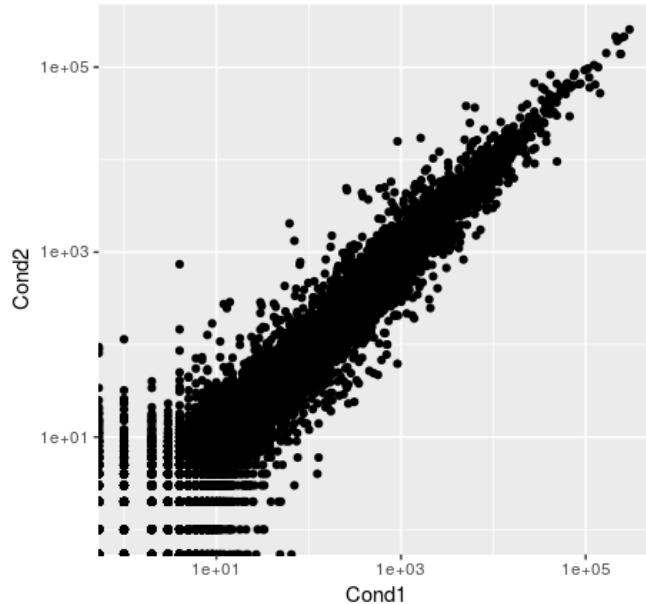
How to show 1D data?



How to show 1D data?

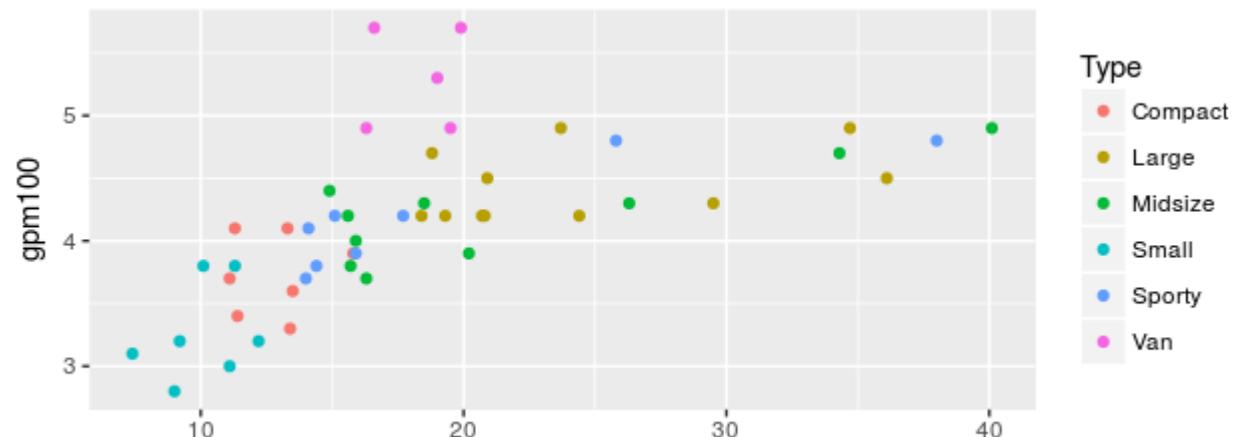
- Histogram
 - Nice and easy, but binning can be misleading
- Density
 - Choice of bandwidth can be difficult
- Empirical distribution function
 - No additional parameter necessary, but less intuitive
- Bar plot
 - Hides most of the data, but is intuitive
- Boxplot
 - Only makes sense if data is unimodal
- Beeswarm
 - Shows data directly, only possible for a medium number of points

How to show 2D data?

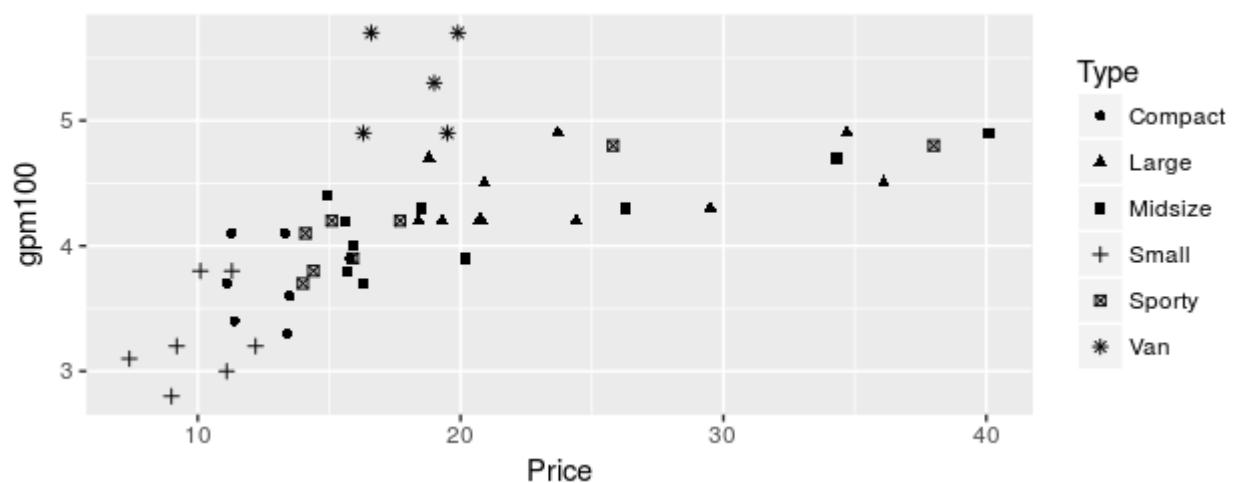


Higher dimensional Data

- Color



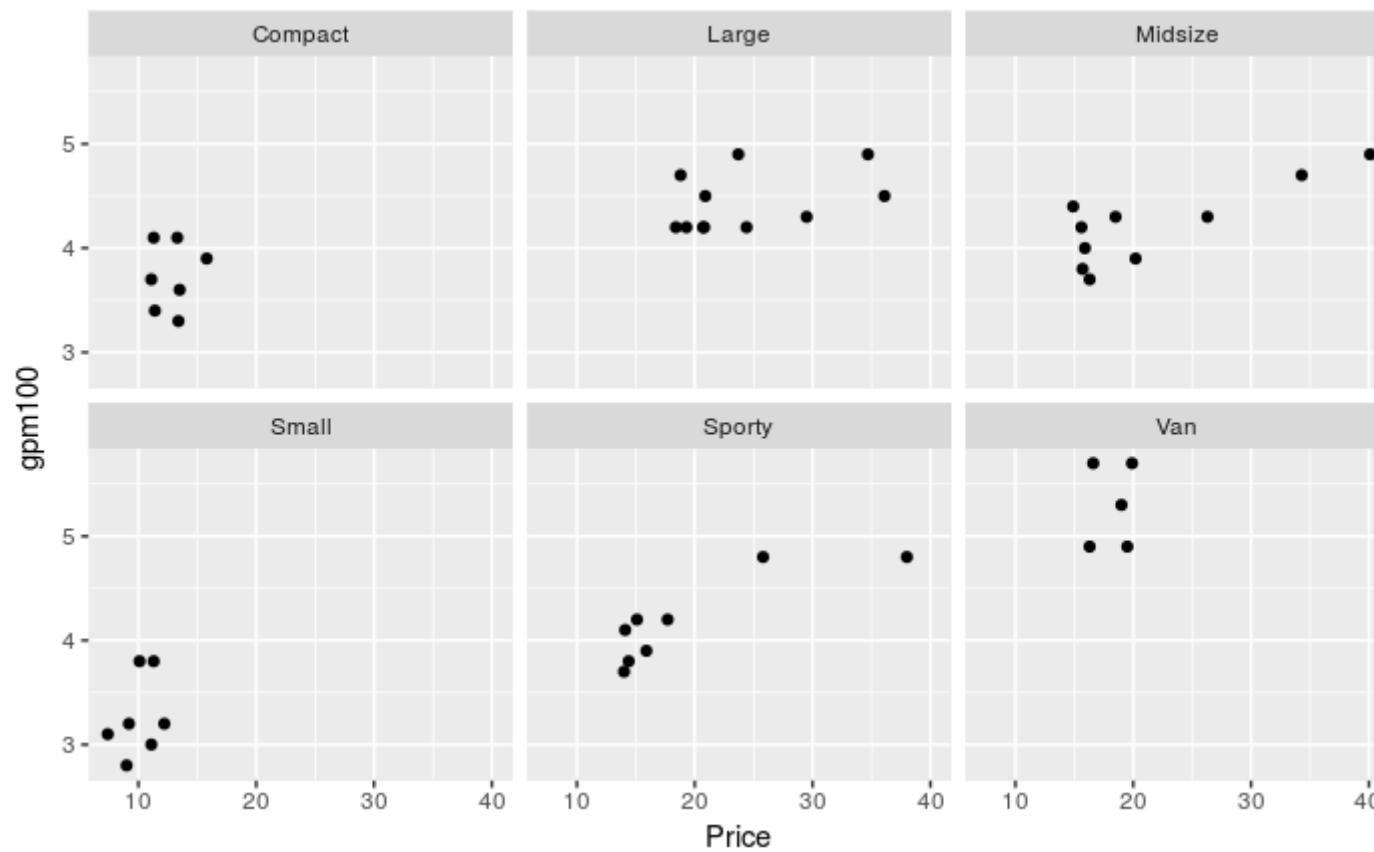
- Shape



- Facets

Higher dimensional Data

```
ggplot(carprize, aes(x=Price, y=MPG.city) +  
  geom_point() +  
  facet_wrap(~ Type)
```



Grammar of Graphics

Statistics and Computing

Leland Wilkinson

The Grammar
of Graphics

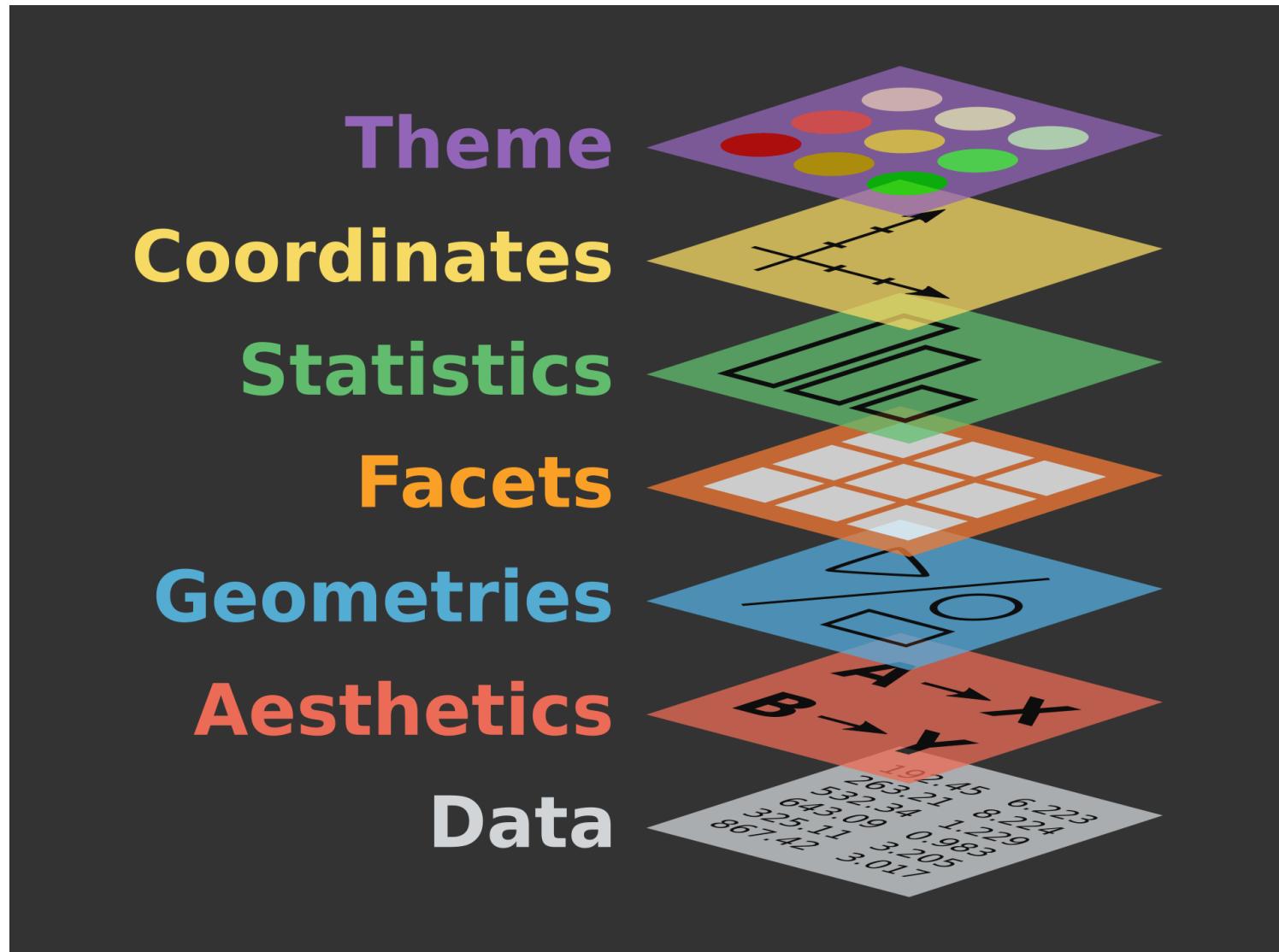
Second Edition

Grammar of Graphics

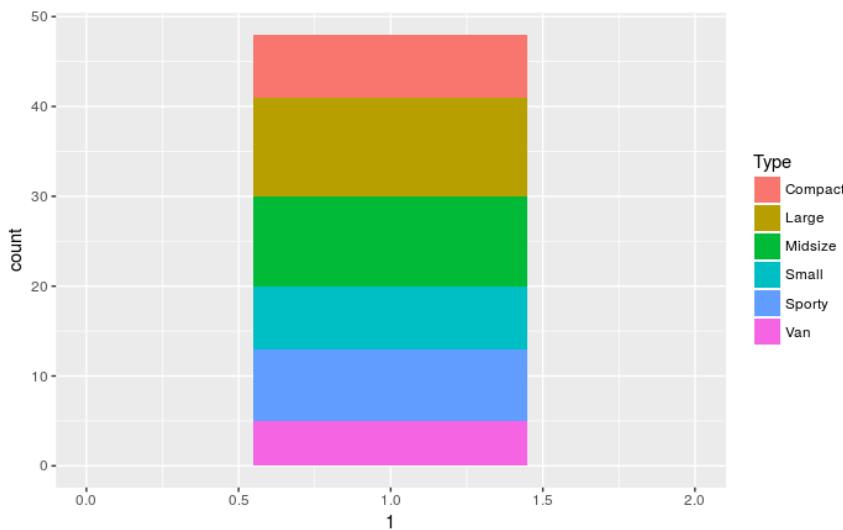
thinking about how we can build on the grammar to learn how to create graphical “poems.”

Wickham, H. A Layered Grammar of Graphics. J. Comput. Graph. Stat. 19, 3–28 (2010).

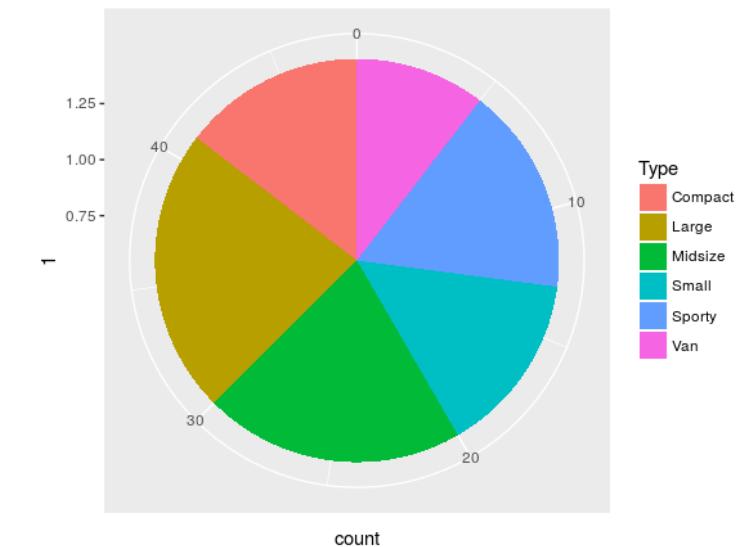
Grammar of Graphics



Grammar of Graphics: Coordinate Systems



```
ggplot(carprize,  
       aes(x=1, fill=Type) +  
       geom_bar() +  
       xlim(0,2))
```

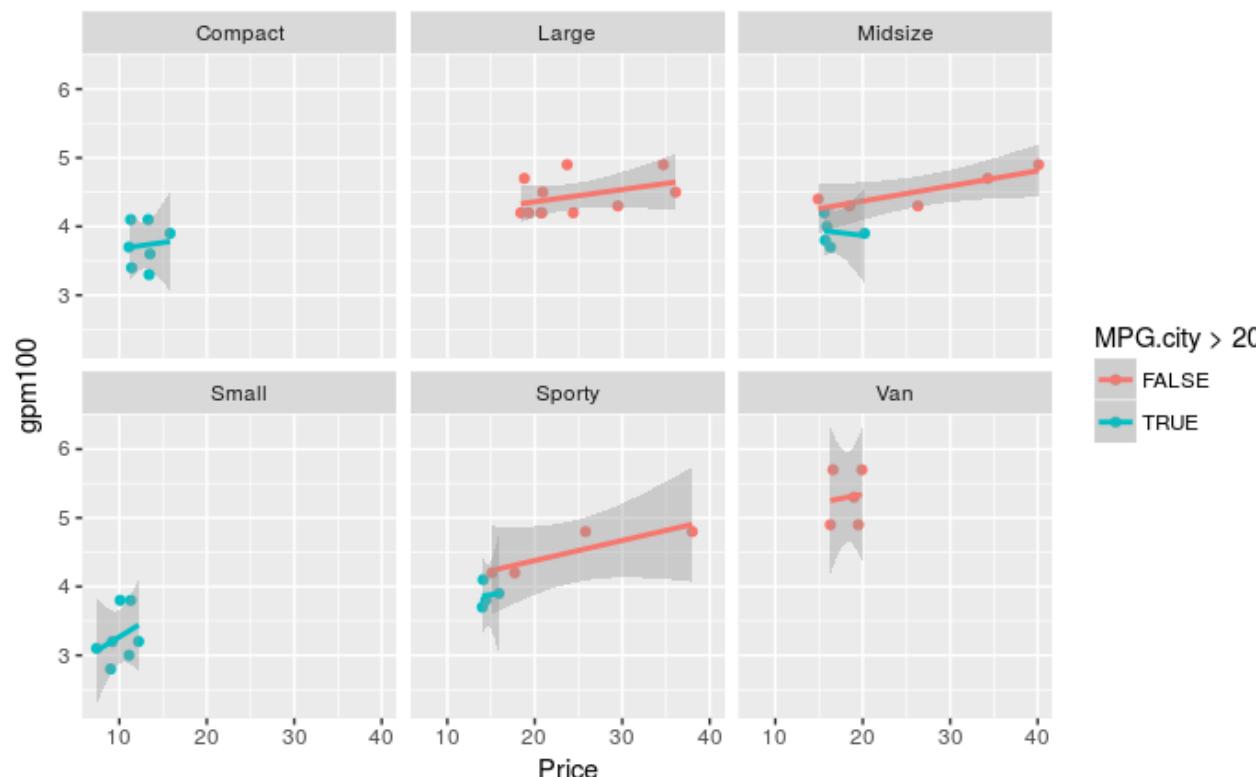


```
ggplot(carprize,  
       aes(x=1, fill=Type) +  
       geom_bar() +  
       coord_polar("y"))
```

A pie chart is just a stacked bar-chart in polar coordinates

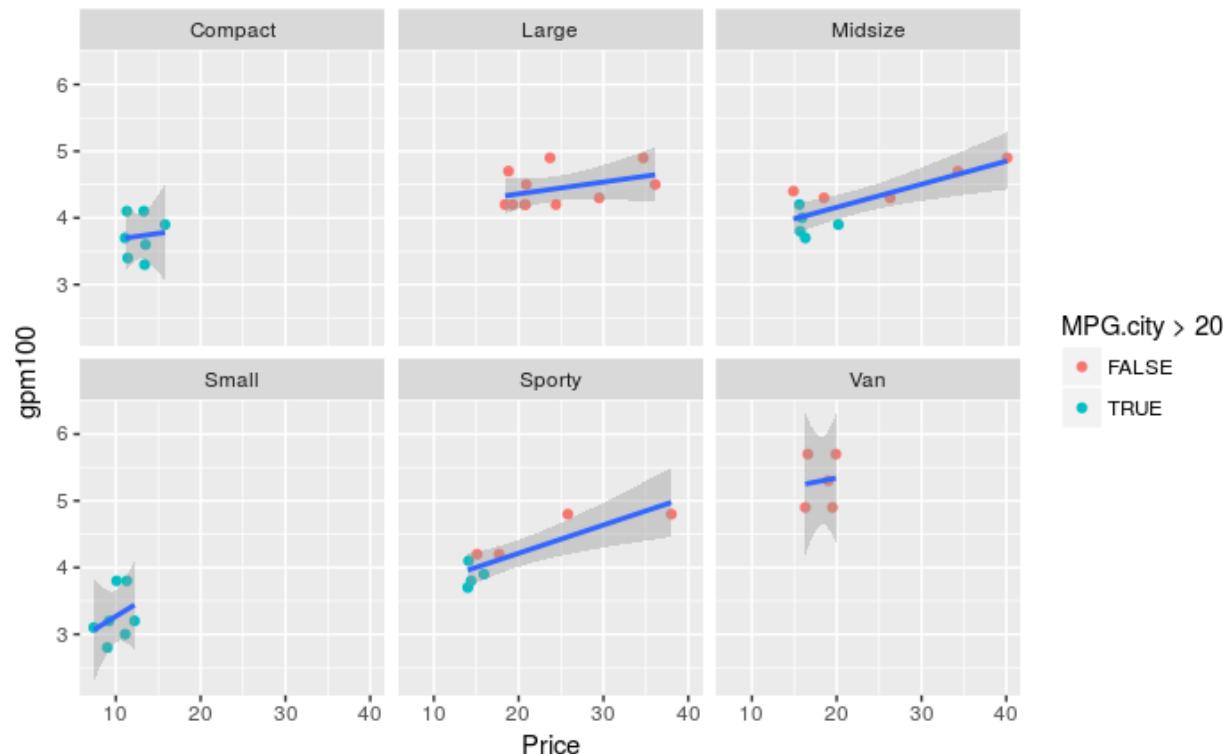
Statistics are computed by group and facet

```
ggplot(aes(x=Price, y=gpm100, color=MPG.city > 20)) +  
  geom_point() +  
  geom_smooth(method = "lm") +  
  facet_wrap(~ Type)
```

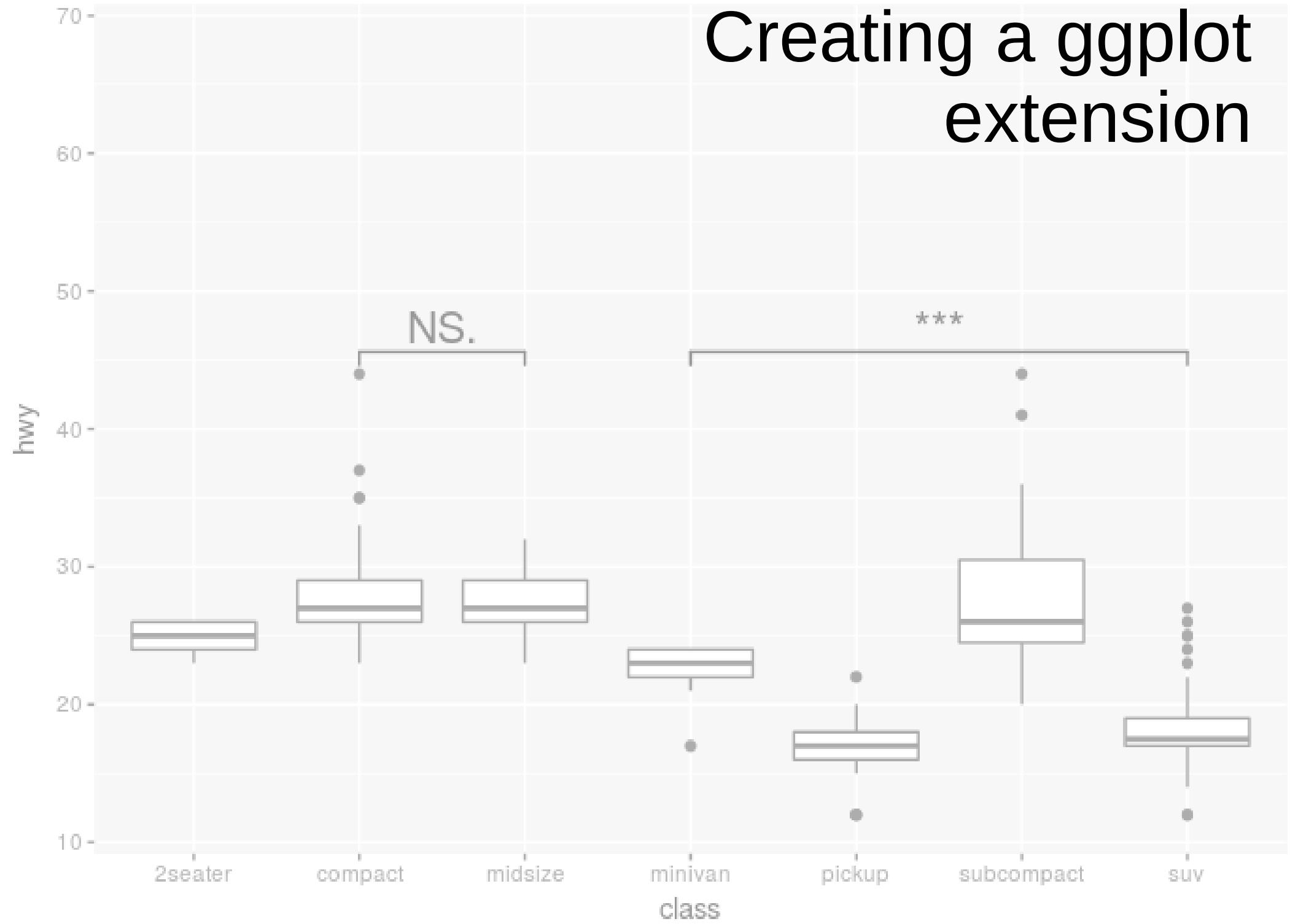


How to get one fit per facet?

```
ggplot(carprize, aes(x=Price, y=gpm100)) +  
  geom_point(aes(color=MPG.city > 20)) +  
  geom_smooth(method = "lm") +  
  facet_wrap(~ Type)
```



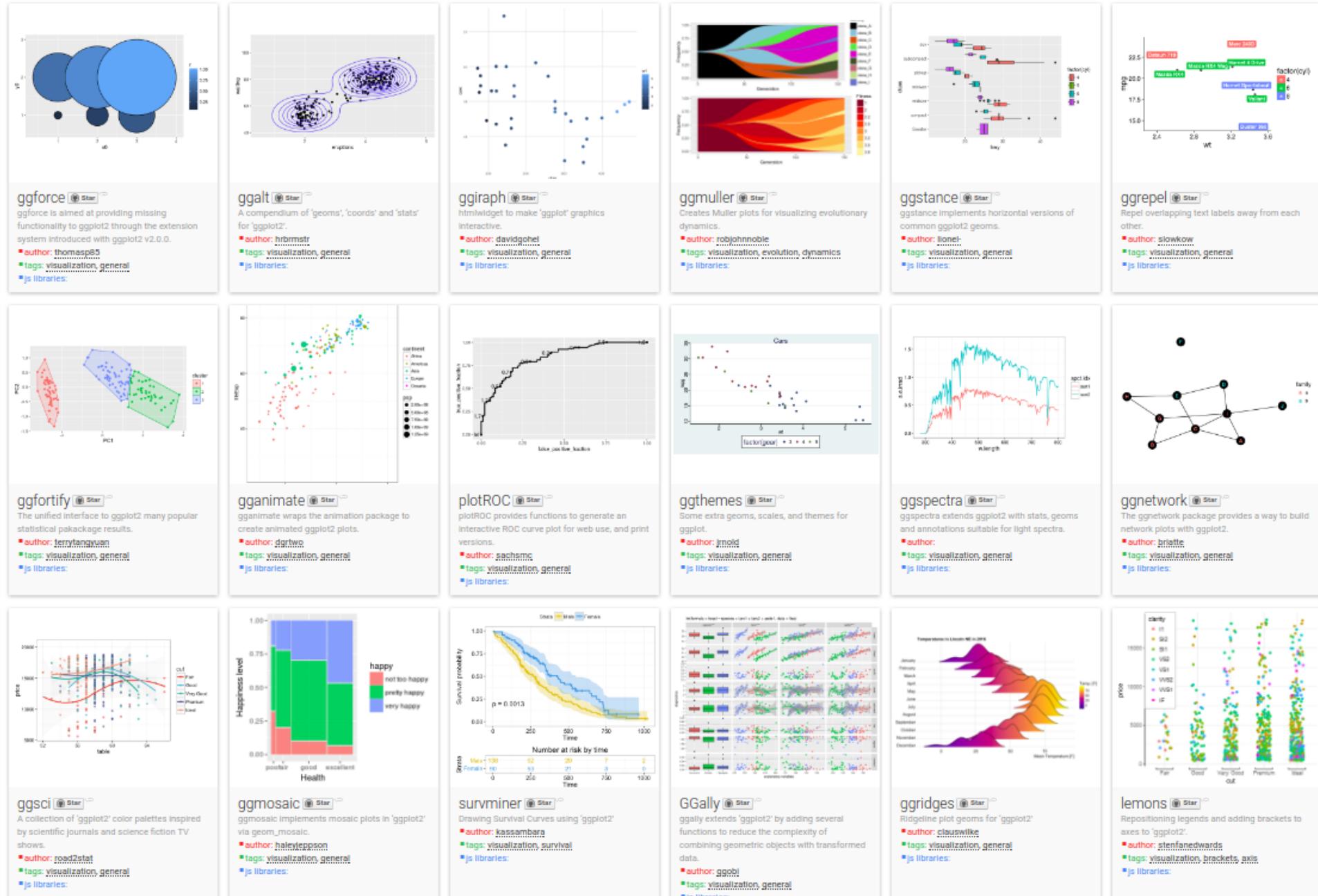
Creating a ggplot extension



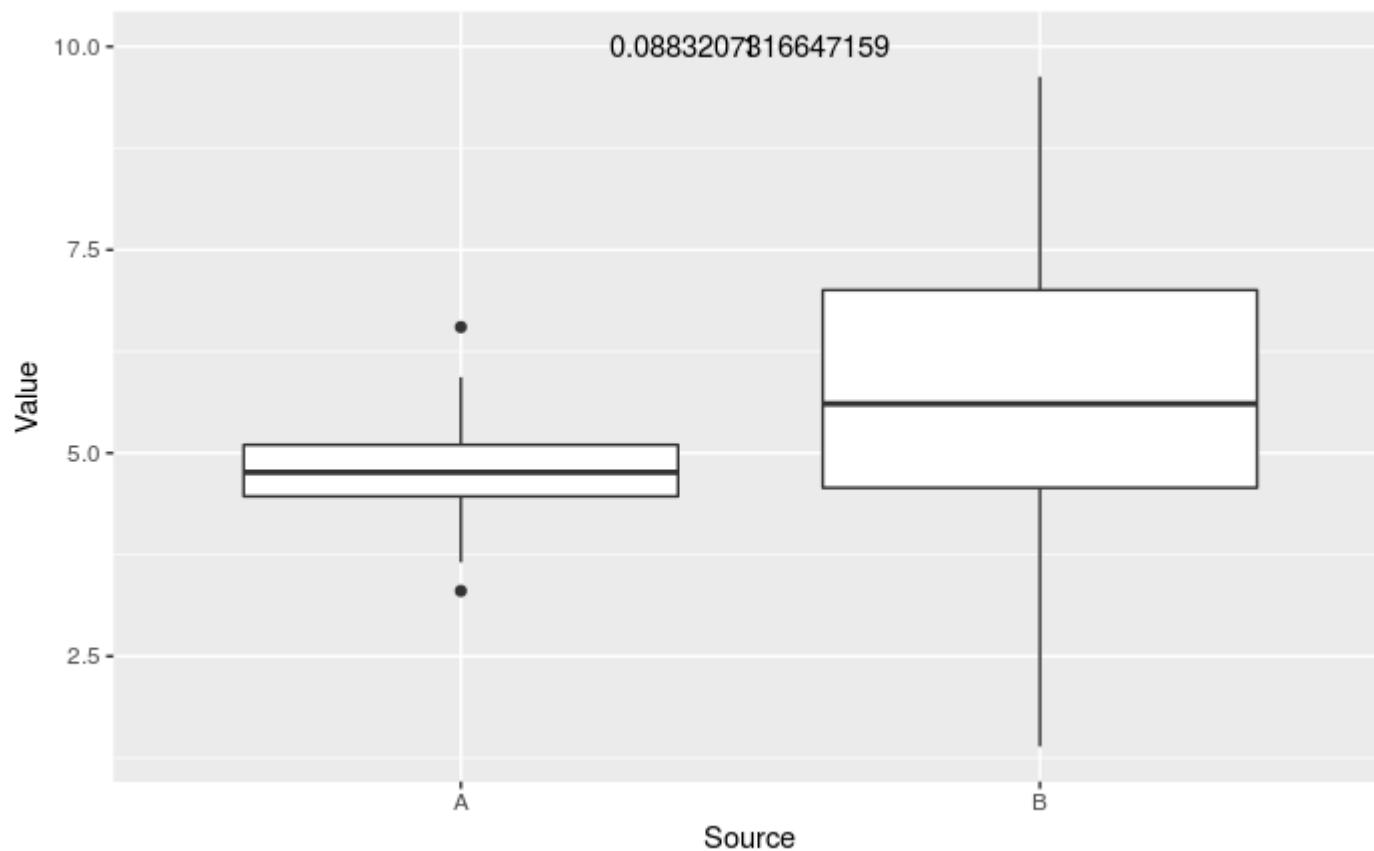
ggplot Extensions

- Since the v.2.0.0 release there is a common interface for creating extensions
- Based on ggproto: a custom object system

Gallery: ggplot2-exts.org



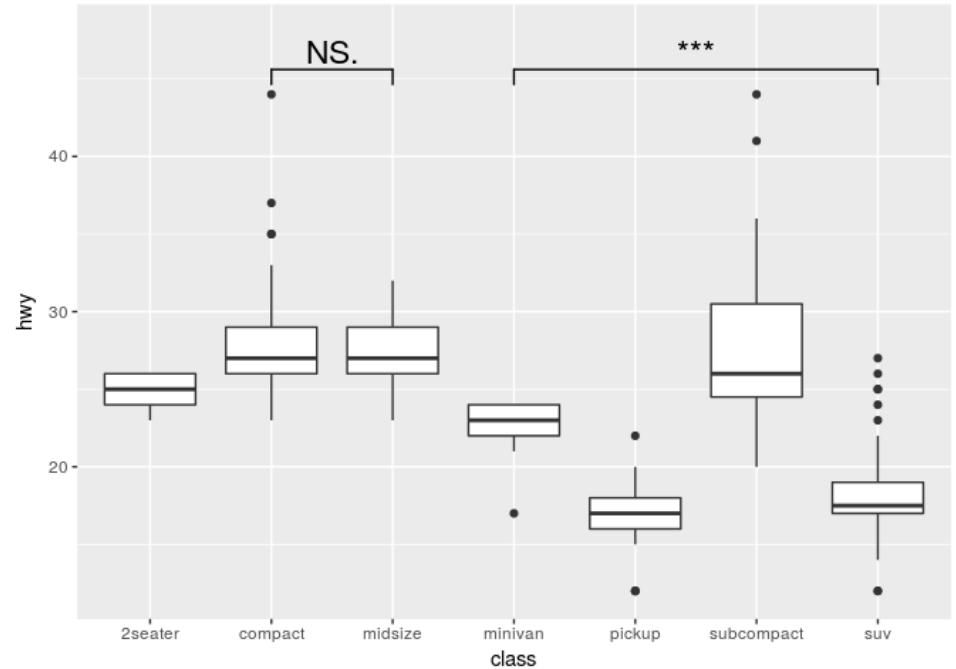
Demo



chull_to_ggsignif.Rmd

ggsignif

- ggplot2 extension to add significance annotations
- Available at CRAN and github.com/const-ae/ggsignif
- Can automatically calculate p-values



```
library(ggsignif)

ggplot(mpg, aes(class, hwy)) +
  geom_boxplot() +
  geom_signif(comparisons = list(c("compact", "midsize"),
                                 c("minivan", "suv")),
              map_signif_level = TRUE, textsize=6) +
  ylim(NA, 48)
```

Thanks for your attention

Constantin Ahlmann-Eltze
 @const-ae