

Visualiser des données

Camille Besse

Département d'Informatique et de Génie Logiciel
Université Laval, Québec, Canada

camille.besse@ift.ulaval.ca

May 6, 2019



Présentation originale de John Rauser

"How Humans See Data"

Si vous préférez écouter la version originale
Original material in R :

https://github.com/jrauser/writing/blob/master/how_humans_see_data/

Complété un peu avec de trucs de <http://www.perceptualedge.com>
et de <http://ieg.ifs.tuwien.ac.at/~aigner/>

Outline

- Pourquoi visualiser des données ?
- Perception et schémas
 - ▶ Détection
 - ▶ Montage
 - ▶ Estimation
- Quelques autres trucs. . .

Visualisation

- Pourquoi visualiser ?
- La visualisation est de la communication
- Le but est de permettre à des humains de résoudre un problème analytique rapidement et précisément

Visualisation

- Pourquoi visualiser ?
- La visualisation est de la communication
- Le but est de permettre à des humains de résoudre un problème analytique rapidement et précisément

Des données ...

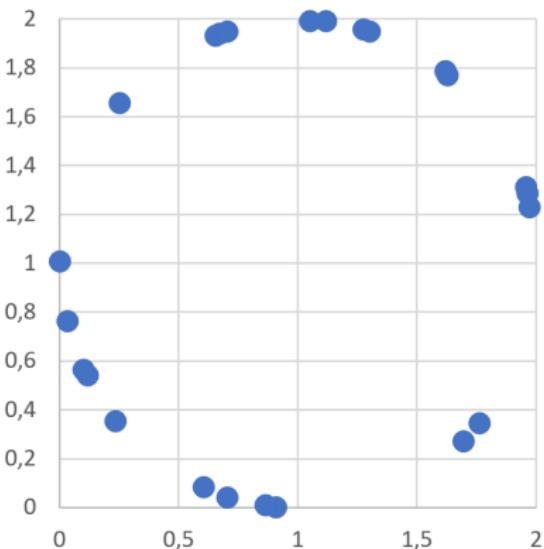
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	id_client	id_vld_policy	id_ue	pol_tpol_coverage	pol_tpolt_pay	frpol_tpolt_usage	pol_insee_code	drv_cdrv_ag_drv	drv_s_drv_drvv_hvh	vhh_cy vh	drv_vh	fuel	vh_make	vh_model	vh_salvh	vh_savh	vh_type	vh_valvh	vh_valvh	vh_valvh	vh_valvh	vh_valvh	vh_valvh	vh_valvh	vh_valvh	vh_valvh	vh_valvh	vh_valvh	
2	A000000001	V01	A000000001-V01	Year 0	0.5	Maxi	29	9 Biannual	No	Retired	36233 No	85 0 M	62 0 10	1587	98	Gasoline	PEUGEOT	306	10	9	182	Tourism	20700						
3	A000000002	V01	A000000002-V01	Year 0	0.5	Maxi	3	1 Biannual	No	Retired	92073 No	69 0 M	39 0 4	2149	170	Diesel	MERCEDES	C220	4	2	229	Tourism	34250						
4	A000000003	V01	A000000003-V01	Year 0	0.5	Maxi	2	2 Yearly	No	WorkPrivate	92026 No	37 0 M	18 0 11	1991	150	Gasoline	BMW	Z3	12	11	210	Tourism	28661						
5	A000000004	V01	A000000004-V01	Year 0	0.5	Median2	22	1 Yearly	No	WorkPrivate	78537 Yes	81 21 M	F 54 3 16	1781	90	Gasoline	POHLWAG	GOLF	18	15	180	Tourism	14407						
6	A000000005	V01	A000000005-V01	Year 0	0.5	Maxi	16	4 Biannual	No	Retired	38544 Yes	62 68 F	M 37 48 11	1598	108	Gasoline	RENAULT	LAGUNA	13	11	195	Tourism	16770						
7	A000000006	V01	A000000006-V01	Year 0	0.5	Median2	5	1 Monthly	No	WorkPrivate	76259 No	68 0 M	40 0 14	1769	60	Diesel	PEUGEOT	205	28	18	155	Tourism	11564						
8	A000000007	V01	A000000007-V01	Year 0	0.5	Maxi	5	3 Biannual	No	Retired	38547 No	77 0 M	55 0 7	1870	108	Diesel	RENAULT	LAGUNA	10	6	193	Tourism	22450						
9	A000000008	V01	A000000008-V01	Year 0	0.5	Maxi	2	2 Biannual	No	Retired	37122 No	64 0 M	37 0 11	1595	101	Gasoline	AUDI	A4	16	13	191	Tourism	20535						
10	A000000009	V01	A000000009-V01	Year 0	0.64	Median2	5	1 Monthly	No	WorkPrivate	83050 No	38 0 M	19 0 9	1997	109	Diesel	PEUGEOT	307	9	7	183	Tourism	23460						
11	A000000010	V01	A00000010-V01	Year 0	0.5	Maxi	26	6 Biannual	No	WorkPrivate	74123 Yes	59 33 M	F 41 15 6	1997	90	Diesel	PEUGEOT	PARTNER	9	7	163	Tourism	18550						
12	A000000011	V01	A00000011-V01	Year 0	0.5	Maxi	8	1 Yearly	No	Retired	31557 Yes	66 32 M	M 45 14 4	1560	90	Diesel	PEUGEOT	308	4	3	180	Tourism	20300						
13	A000000012	V01	A00000012-V01	Year 0	0.5	Maxi	4	4 Yearly	No	Retired	28386 No	61 0 M	43 0 5	1798	127	Gasoline	MAZDA	MX5	6	3	196	Tourism	22450						
14	A000000013	V01	A00000013-V01	Year 0	0.5	Maxi	21	1 Biannual	Yes	Retired	63063 No	65 0 F	43 0 5	999	62	Gasoline	KIA	PICANTO	7	4	150	Tourism	7990						
15	A000000014	V01	A00000014-V01	Year 0	0.5	Maxi	25	8 Monthly	No	Retired	62688 No	72 0 M	46 0 13	1905	68	Diesel	CITROEN	XSARA	14	13	162	Tourism	14773						
16	A000000015	V01	A00000015-V01	Year 0	0.5	Maxi	9	1 Biannual	No	WorkPrivate	44003 Yes	38 34 M	F 19 14 1	1560	109	Diesel	CITROEN	C4 PICASSO	3	1	180	Tourism	27100						
17	A000000016	V01	A00000016-V01	Year 0	0.5	Maxi	6	2 Monthly	No	WorkPrivate	86183 Yes	54 56 M	F 32 37 2	1248	75	Diesel	OPEL	CORSA	5	4	163	Tourism	14900						
18	A000000017	V01	A00000017-V01	Year 0	0.5	Maxi	6	3 Quarterly	No	WorkPrivate	56121 No	41 0 F	22 0 8	1870	81	Diesel	RENAULT	MEGANE	10	8	170	Tourism	16300						
19	A000000018	V01	A00000018-V01	Year 0	0.5	Median2	8	3 Monthly	No	WorkPrivate	42284 No	38 0 M	20 0 29 1124	51	51	Gasoline	PEUGEOT	104	31	23	138	Tourism	6933						
20	A000000019	V01	A00000019-V01	Year 0	0.5	Maxi	19	4 Biannual	No	Retired	1418 No	83 0 M	58 0 5	1398	68	Diesel	CITROEN	C3	9	1	158	Tourism	18150						
21	A000000020	V01	A00000020-V01	Year 0	0.5	Median2	23	15 Monthly	No	WorkPrivate	60103 No	56 0 F	32 0 22	1905	71	Diesel	PEUGEOT	405	23	19	165	Tourism	15535						
22	A000000021	V01	A00000021-V01	Year 0	0.5	Median2	1	1 Yearly	No	WorkPrivate	88209 Yes	44 44 F	M 20 26 8	1997	136	Gasoline	PEUGEOT	307	8	6	207	Tourism	25350						
23	A000000022	V01	A00000022-V01	Year 0	0.68	Median2	11	2 Biannual	No	WorkPrivate	1140 Yes	40 41 M	F 20 23 16	1794	96	Gasoline	RENAULT	R19	19	17	181	Tourism	15931						
24	A000000023	V01	A00000023-V01	Year 0	0.5	Maxi	12	4 Biannual	No	Retired	64102 No	74 0 M	45 0 7	1896	111	Diesel	FORD	GALAXY	11	5	181	Tourism	32900						
25	A000000024	V01	A00000024-V01	Year 0	0.5	Maxi	11	2 Biannual	No	Professional	91345 Yes	59 36 F	F 39 18 11	1689	90	Diesel	MERCEDES	A170	13	10	172	Tourism	21724						
26	A000000025	V01	A00000025-V01	Year 0	0.65	Maxi	16	1 Yearly	No	Retired	38318 No	71 0 F	41 0 2	1560	90	Diesel	CITROEN	C3 PICASSO	2	1	173	Tourism	19350						
27	A000000026	V01	A00000026-V01	Year 0	0.57	Median2	1	1 Monthly	No	WorkPrivate	63180 No	33 0 M	13 0 10	2497	164	Diesel	BMW	525	11	10	219	Tourism	41162						
28	A000000027	V01	A00000027-V01	Year 0	0.5	Median2	11	3 Yearly	No	WorkPrivate	83069 Yes	60 55 M	F 40 35 10	1995	116	Diesel	TOYOTA	AVENSIS	10	10	180	Tourism	25901						
29	A000000028	V01	A00000028-V01	Year 0	0.6	Mini	1	1 Monthly	No	WorkPrivate	17300 No	31 0 F	12 0 16	1896	90	Diesel	AUDI	A4	16	13	183	Tourism	23066						
30	A000000029	V01	A00000029-V01	Year 0	0.5	Mini	4	1 Biannual	No	Professional	83282 No	40 0 M	22 0 16	2446	86	Diesel	CITROEN	JUMPER	17	9	132	Commerci	24615						
31	A000000030	V01	A00000030-V01	Year 0	0.5	Maxi	7	1 Biannual	No	WorkPrivate	78003 Yes	50 23 F	M 30 4 6	1398	68	Diesel	FORD	FUSION	9	1	158	Commerci	14640						
32	A000000031	V01	A00000031-V01	Year 0	0.5	Maxi	2	2 Monthly	No	WorkPrivate	78372 Yes	46 41 M	F 26 14 2	1598	120	Gasoline	PEUGEOT	207	4	2	200	Tourism	19550						
33	A000000032	V01	A00000032-V01	Year 0	0.5	Maxi	4	2 Biannual	No	WorkPrivate	83138 No	42 0 F	24 0 2	1461	86	Diesel	RENAULT	MEGANE	3	1	175	Tourism	19900						
34	A000000033	V01	A00000033-V01	Year 0	0.5	Maxi	24	2 Yearly	No	Retired	10297 No	72 0 M	47 0 3	1461	86	Diesel	RENAULT	CLIO	6	4	174	Tourism	17450						
35	A000000034	V01	A00000034-V01	Year 0	0.54	Mini	1	1 Monthly	No	WorkPrivate	57751 No	33 0 M	13 0 13	2494	170	Gasoline	BMW	323 I	13	11	231	Tourism	31253						
36	A000000035	V01	A00000035-V01	Year 0	0.5	Maxi	4	3 Monthly	No	WorkPrivate	50602 Yes	41 31 M	F 19 6 4	1461	68	Diesel	DACIA	LOGAN	5	3	150	Tourism	12240						
37	A000000036	V01	A00000036-V01	Year 0	0.5	Median2	19	1 Biannual	Yes	Retired	62043 No	79 0 F	47 0 14	1905	68	Diesel	CITROEN	XSARA	14	13	162	Tourism	14773						
38	A000000037	V01	A00000037-V01	Year 0	0.5	Median2	20	3 Yearly	No	Retired	59286 No	59 0 M	41 0 9	1870	82	Diesel	RENAULT	MASTER	10	8	131	Commerci	21050						
39	A000000038	V01	A00000038-V01	Year 0	0.5	Maxi	27	2 Monthly	No	Retired	69264 No	76 0 F	45 0 11	1242	60	Gasoline	FIAT	PUNTO	12	9	155	Commerci	9940						
40	A000000039	V01	A00000039-V01	Year 0	0.5	Maxi	3	3 Yearly	No	WorkPrivate	25056 Yes	49 51 F	M 31 33 10	1870	98	Diesel	RENAULT	ESPACE	12	11	167	Tourism	27441						
41	A00000040	V01	A00000040-V01	Year 0	0.64	Maxi	2	2 Yearly	No	WorkPrivate	78455 No	32 0 F	11 0 7	1398	68	Diesel	CITROEN	C3	9	1	158	Tourism	18150						
42	A000000401	V01	A000000401-V01	Year 0	0.64	Maxi	8	3 Quarterly	No	WorkPrivate	71176 No	27 0 F	8 0 3 10	1686	66	Gasoline	KIA	PICANTO	7	4	154	Tourism	10050						
43	A000000402	V01	A000000402-V01	Year 0	0.5	Mini	4	3 Yearly	No	Retired	38543 No	75 0 M	57 0 10	2496	180	Diesel	AUDI	A6	11	7	219	Tourism	44380						
44	A000000402	V02	A000000402-V02	Year 0	0.5	Mini	4	3 Yearly	No	Retired	38543 No	75 0 M	57 0 25	2445	73	Diesel	RENAULT	B70	25	21	112	Commerci	21552						
45	A000000402	V03	A000000402-V03	Year 0	0.5	Mini	4	3 Yearly	No	Retired	38543 No	75 0 M	57 0 24	1324	67	Gasoline	SUZUKI	SJ 413	26	22	135	Commerci	12591						
46	A000000403	V01	A000000403-V01	Year 0	0.5	Maxi	5	1 Monthly	No	WorkPrivate	79195 No	36 0 F	17 0 3	1560	109	Diesel	CITROEN	C4 PICASSO	4	3	180	Tourism	27000						
47	A000000403	V02	A000000403-V02	Year 0	0.5	Median2	5	1 Monthly	No	WorkPrivate	79195 Yes	37 36 M	F 18 17 15	1905	70	Diesel	PEUGEOT	EXPERT	16	14	138	Commerci	17356						

... brutes.

x	y	x	y
1.972	1.236	0.111	0.542
1.112	1.994	0.902	0.005
0.000	1.009	0.598	0.085
0.665	1.942	1.613	1.790
0.235	0.356	1.298	1.955
0.247	1.658	0.651	1.937
1.275	1.961	1.949	1.316
0.702	0.045	0.099	0.567
1.760	0.350	0.862	0.010
1.691	0.277	0.027	0.768
1.628	1.778	0.706	1.956
1.957	1.290	1.042	1.999

justes brutes mais ...

x	y	x	y
1.972	1.236	0.111	0.542
1.112	1.994	0.902	0.005
0.000	1.009	0.598	0.085
0.665	1.942	1.613	1.790
0.235	0.356	1.298	1.955
0.247	1.658	0.651	1.937
1.275	1.961	1.949	1.316
0.702	0.045	0.099	0.567
1.760	0.350	0.862	0.010
1.691	0.277	0.027	0.768
1.628	1.778	0.706	1.956
1.957	1.290	1.042	1.999



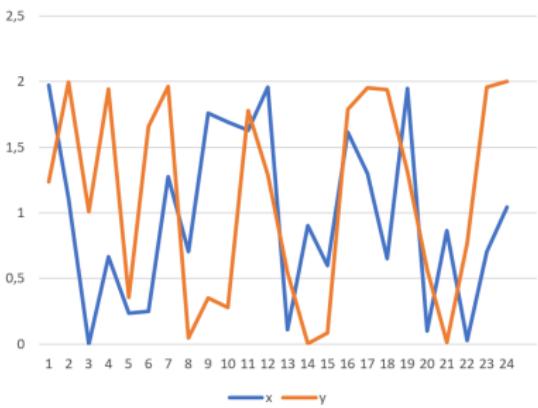
Un graphique est un
encodage
de données.

Et si ...

n	x	y	n	x	y
1	1.972	1.236	13	0.111	0.542
2	1.112	1.994	14	0.902	0.005
3	0.000	1.009	15	0.598	0.085
4	0.665	1.942	16	1.613	1.790
5	0.235	0.356	17	1.298	1.955
6	0.247	1.658	18	0.651	1.937
7	1.275	1.961	19	1.949	1.316
8	0.702	0.045	20	0.099	0.567
9	1.760	0.350	21	0.862	0.010
10	1.691	0.277	22	0.027	0.768
11	1.628	1.778	23	0.706	1.956
12	1.957	1.290	24	1.042	1.999

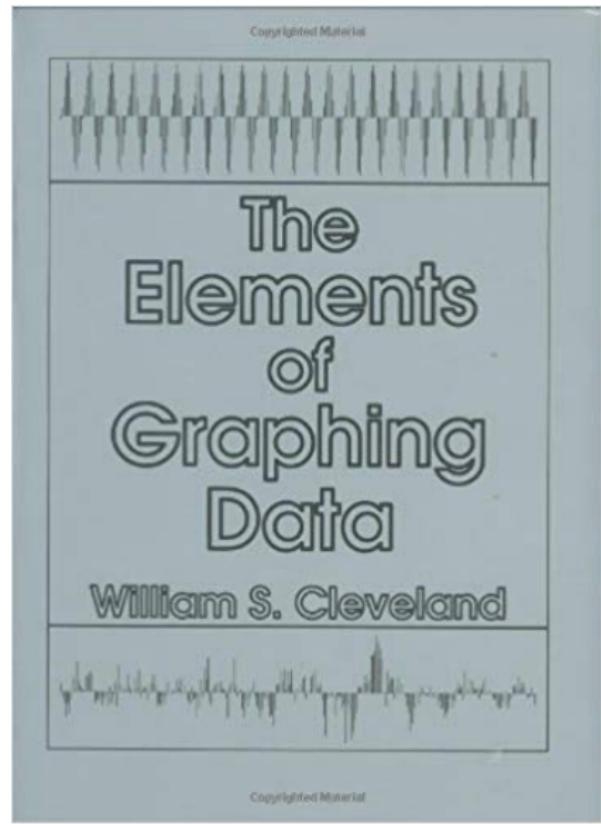
... on fait pas attention ...

n	x	y	n	x	y
1	1.972	1.236	13	0.111	0.542
2	1.112	1.994	14	0.902	0.005
3	0.000	1.009	15	0.598	0.085
4	0.665	1.942	16	1.613	1.790
5	0.235	0.356	17	1.298	1.955
6	0.247	1.658	18	0.651	1.937
7	1.275	1.961	19	1.949	1.316
8	0.702	0.045	20	0.099	0.567
9	1.760	0.350	21	0.862	0.010
10	1.691	0.277	22	0.027	0.768
11	1.628	1.778	23	0.706	1.956
12	1.957	1.290	24	1.042	1.999



**Les bons visuels optimisent
le système cognitif humain.**

Mais comment l'être humain décide t'il un graphique ?



Outline

Trois opérations visuelles dans la perception de schémas :

- 1** Détection
- 2** Construction
- 3** Estimation

Estimation

Trois niveau d'estimation :

- | | | |
|---|----------------|-------------------------|
| 1 | Discrimination | $X = Y$ ou $X \neq Y$? |
| 2 | Ordre | $X < Y$ ou $X > Y$? |
| 3 | Ratio | $X/Y = ?$ |

"Au cœur du raisonnement quantitatif existe une seule question : Comparé à quoi ?"

- Tufte, *Envisioning Information*

Graphical Perception and Graphical Methods for Analyzing Scientific Data

William S. Cleveland and Robert McGill

Science 30 Aug 1985
Vol. 229, Issue 4716, pp. 828-833
DOI: 10.1126/science.229.4716.828

Table 1. Ordering elementary tasks by accuracy, according to theoretical arguments and experimental results. Graphs should exploit tasks as high in the ordering as possible. The tasks are ordered from most accurate to least.

Rank	Aspect judged
1	Position along a common scale
2	Position on identical but nonaligned scales
3	Length
4	Angle
	Slope (with θ not too close to 0, $\pi/2$, or π radians)
5	Area
6	Volume
	Density
	Color saturation
7	Color hue

La chose la plus importante

Table 1. Ordering elementary tasks by accuracy, according to theoretical arguments and experimental results. Graphs should exploit tasks as high in the ordering as possible. The tasks are ordered from most accurate to least.

Rank	Aspect judged
1	Position along a common scale
2	Position on identical but nonaligned scales
3	Length
4	Angle
	Slope (with θ not too close to 0, $\pi/2$, or π radians)
5	Area
6	Volume
	Density
	Color saturation
7	Color hue

Ordre d'encodage

La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune
- 2 Position sur une échelle identique non alignée
- 3 Longueur
- 4 Angle ou pente
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 Teinte de couleur

Ordre d'encodage

La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune
- 2 Position sur une échelle identique non alignée
- 3 Longueur
- 4 Angle ou pente
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 **Teinte de couleur**

“Première règle de la couleur :
On ne parle pas de la couleur !”

- *Tamara Munzner*

Couleur : 3 composantes

Luminosité



Saturation



Teinte

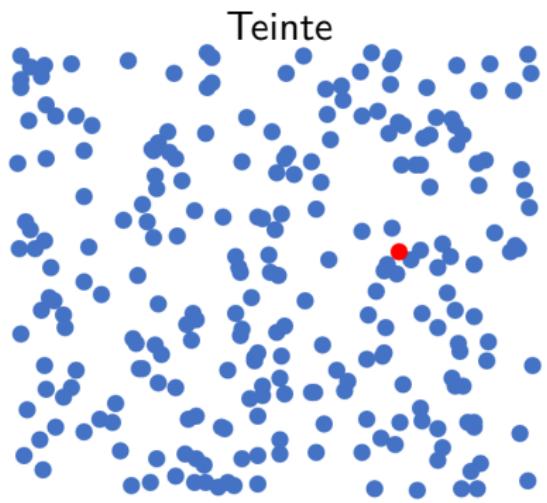
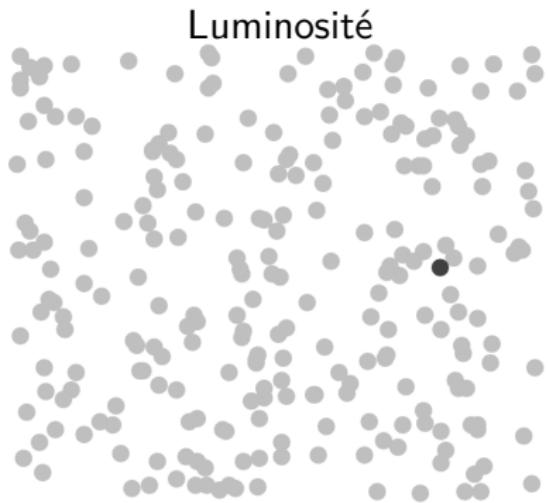


Exemple de saturation : Nombres

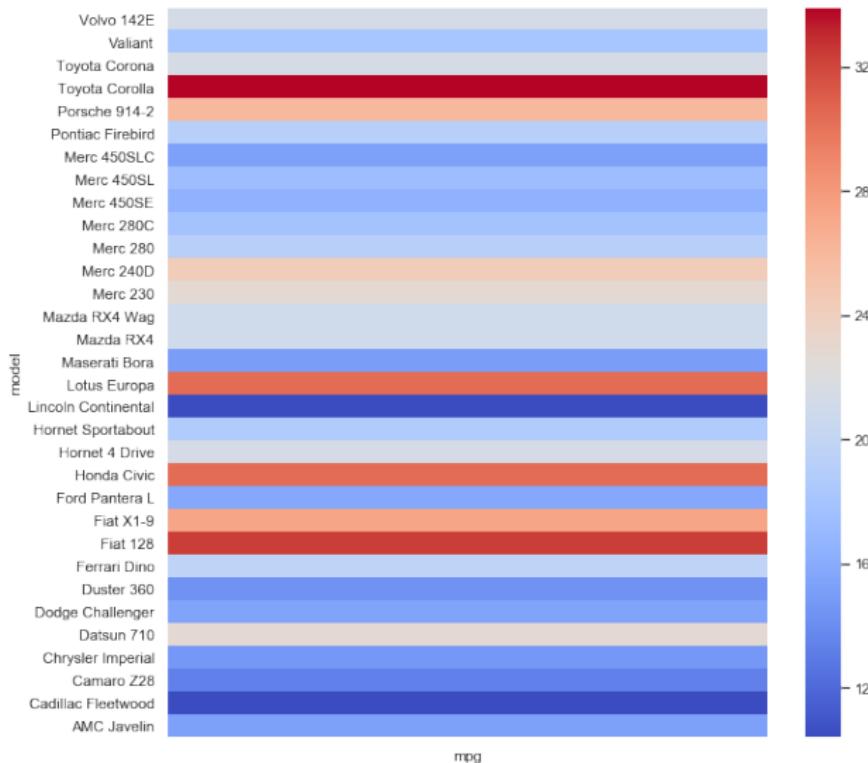
1561321203658413076510374627
4173127527327592732990709742
1703707774179527931749270973
4019743217909370945179279417

1561321203658413076510374627
4173127527327592732990709742
1703707774179527931749270973
4019743217909370945179279417

Exemple de Luminosité et teintes : scatter

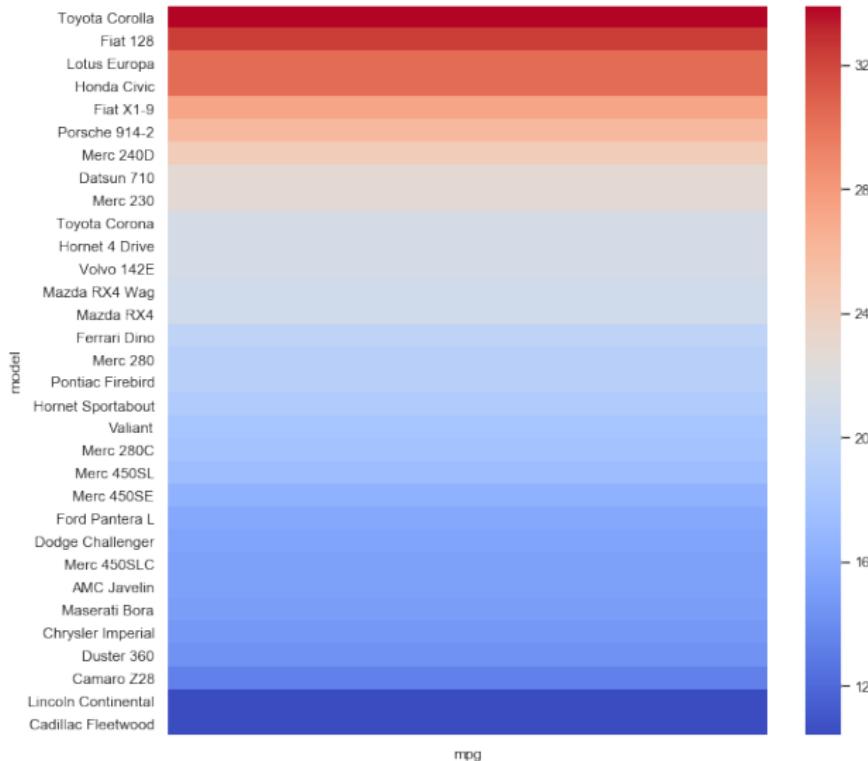


Exemple de teintes : dataset autos



L'ordre alphabétique est quasiment jamais le bon ordre pour une variable catégorique.

Exemple de teintes ordonnées : dataset autos

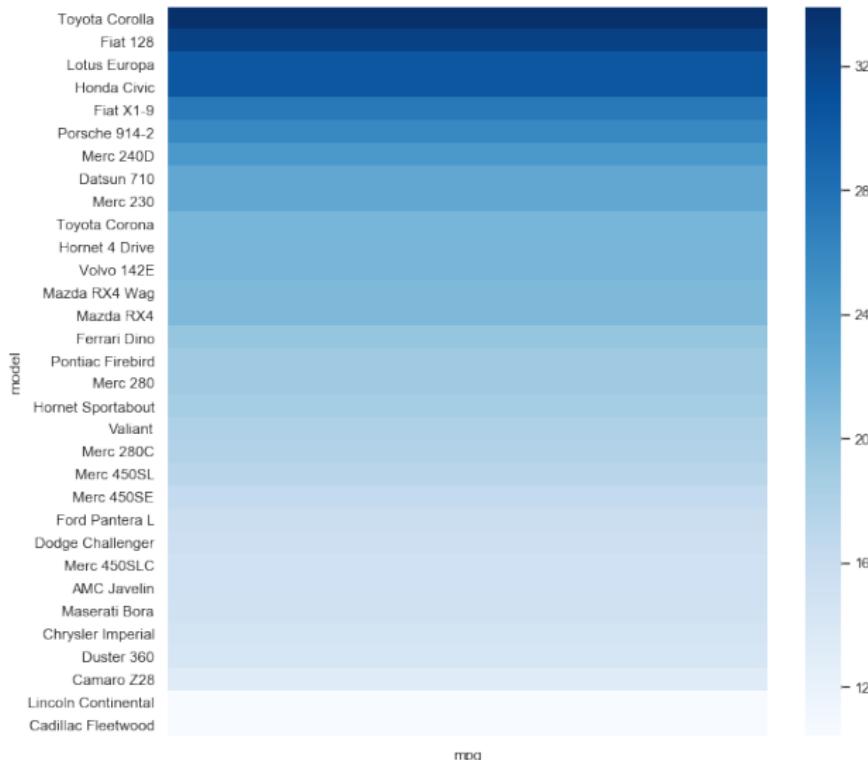


Ordre d'encodage

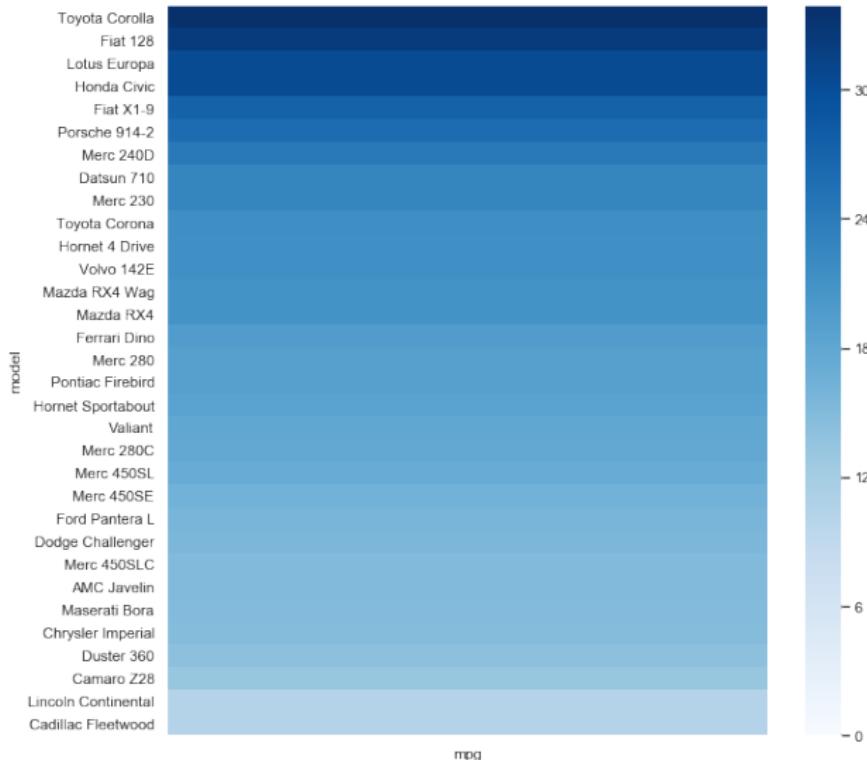
La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune
- 2 Position sur une échelle identique non alignée
- 3 Longueur
- 4 Angle ou pente
- 5 Aire
- 6 Volume ou Densité ou **Saturation de couleur**
- 7 Teinte de couleur

Exemple de saturations ordonnées : dataset autos



Exemple de teintes ordonnées : dataset autos



Ordre d'encodage

La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune**
- 2 Position sur une échelle identique non alignée**
- 3 Longueur**
- 4 Angle ou pente**
- 5 Aire**
- 6 Volume ou Densité ou Saturation de couleur**
- 7 Teinte de couleur**

Exemple d'aire : dataset autos

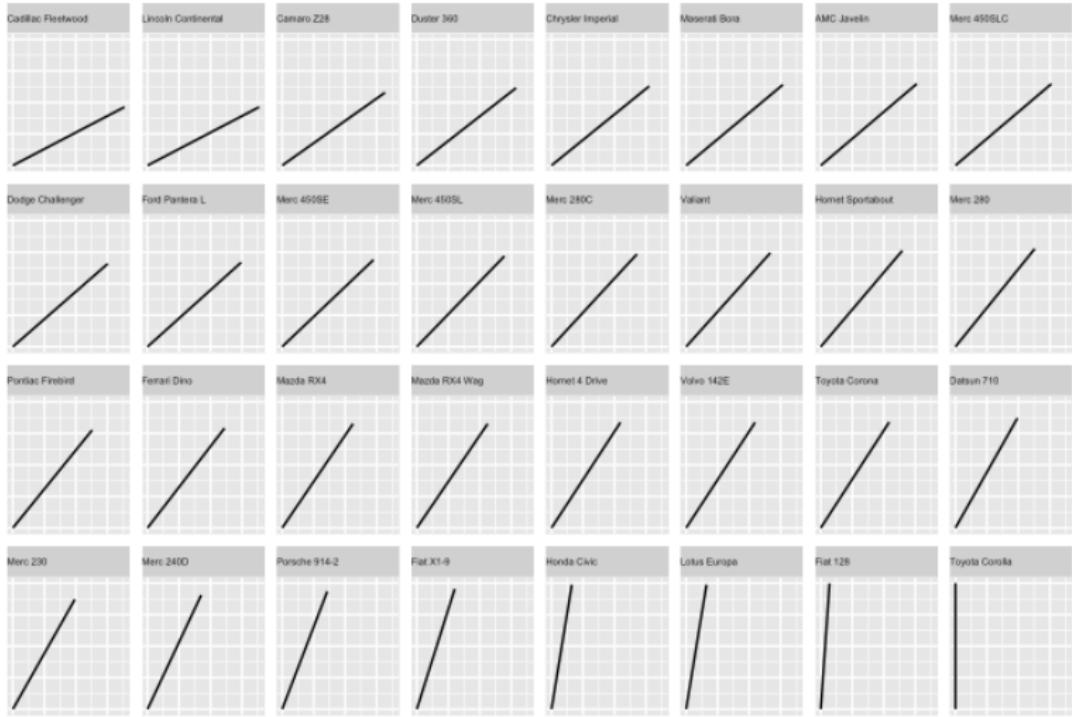


Ordre d'encodage

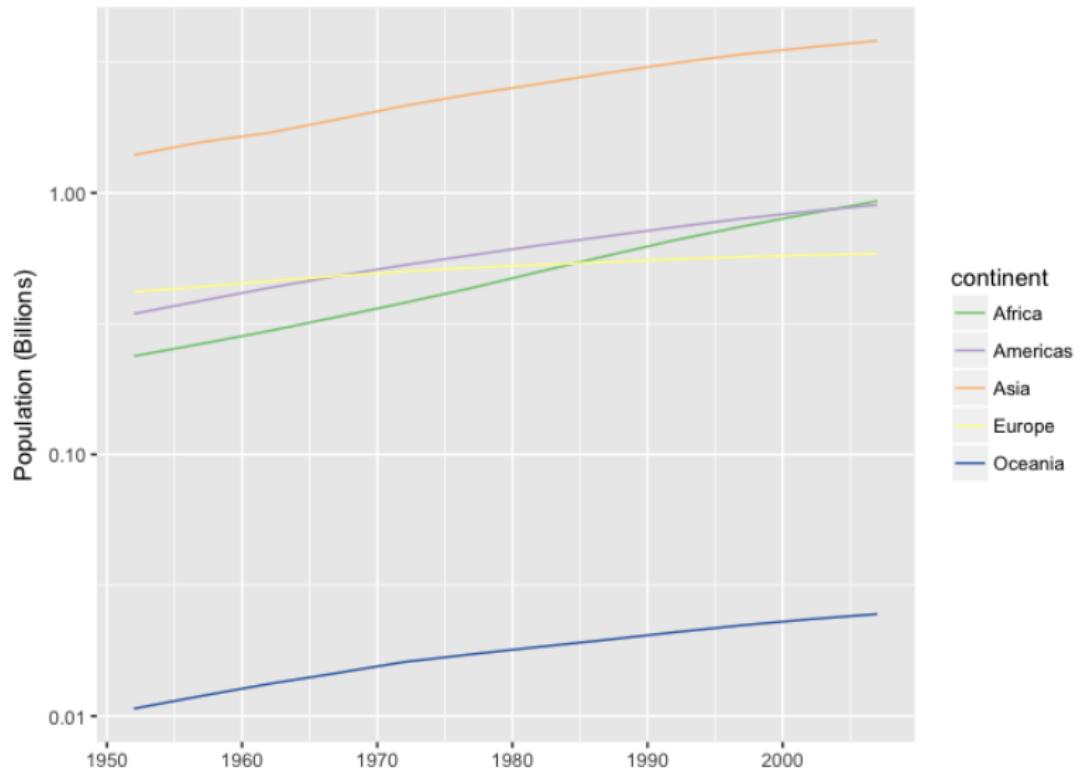
La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune
- 2 Position sur une échelle identique non alignée
- 3 Longueur
- 4 **Angle ou pente**
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 Teinte de couleur

Exemple d'angle : dataset autos



Exemple d'angle : Pop. mondiale



Ordre d'encodage

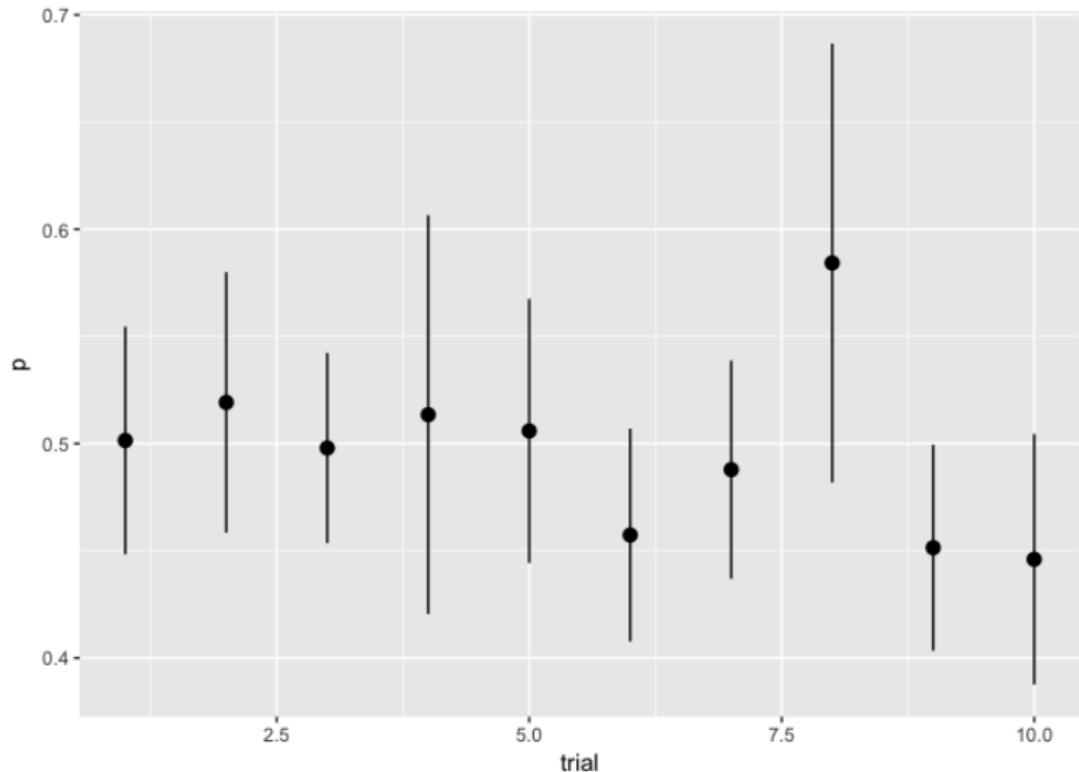
La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune
- 2 Position sur une échelle identique non alignée
- 3 **Longueur**
- 4 Angle ou pente
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 Teinte de couleur

Exemple de longueur : dataset autos



Exemple de longueur : quartiles

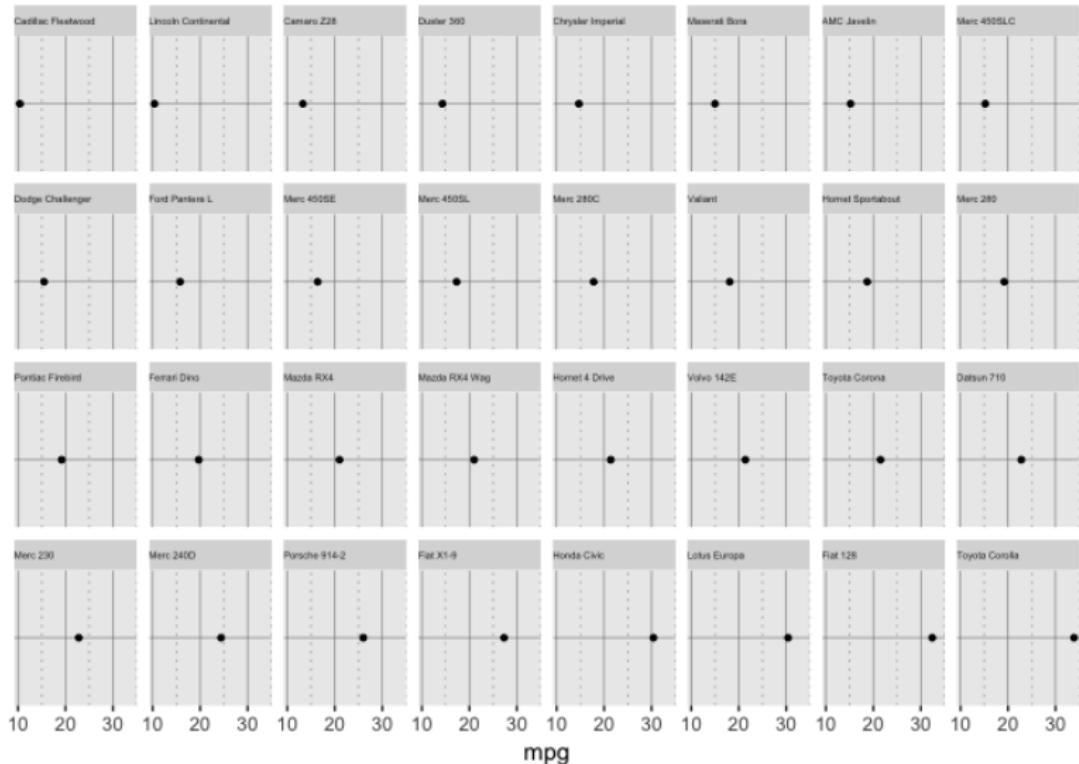


Ordre d'encodage

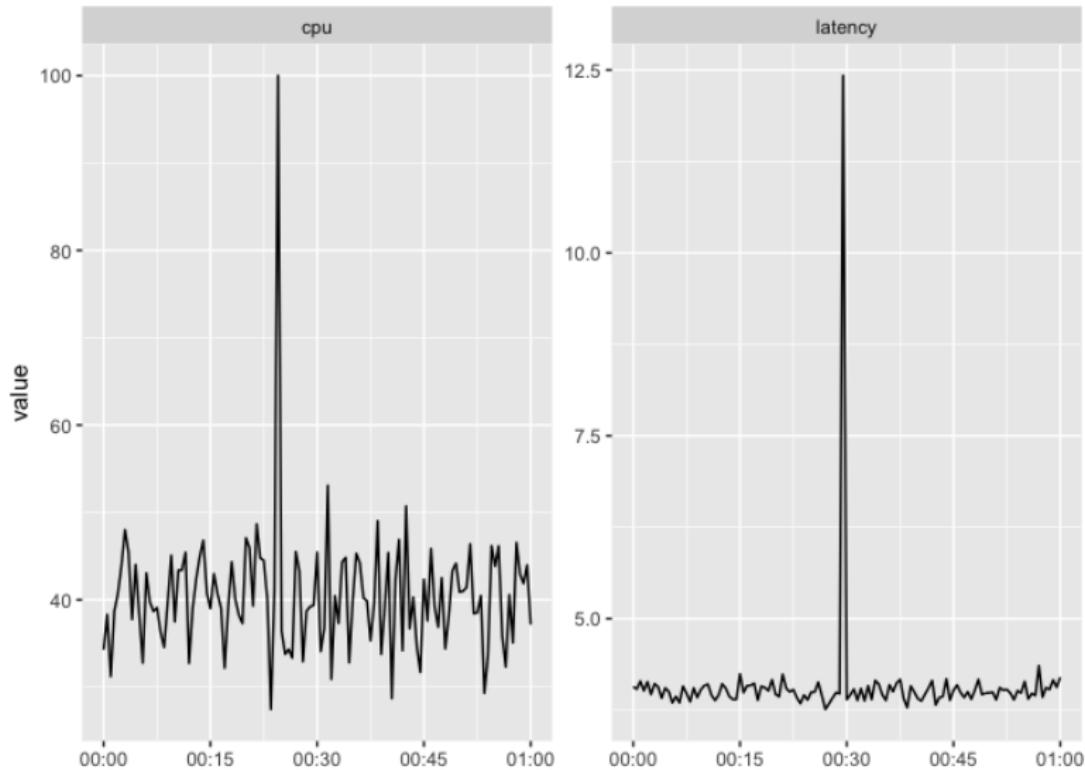
La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune
- 2 **Position sur une échelle identique non alignée**
- 3 Longueur
- 4 Angle ou pente
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 Teinte de couleur

Exemple de position non alignée : autos



Exemple de position non alignée : latency

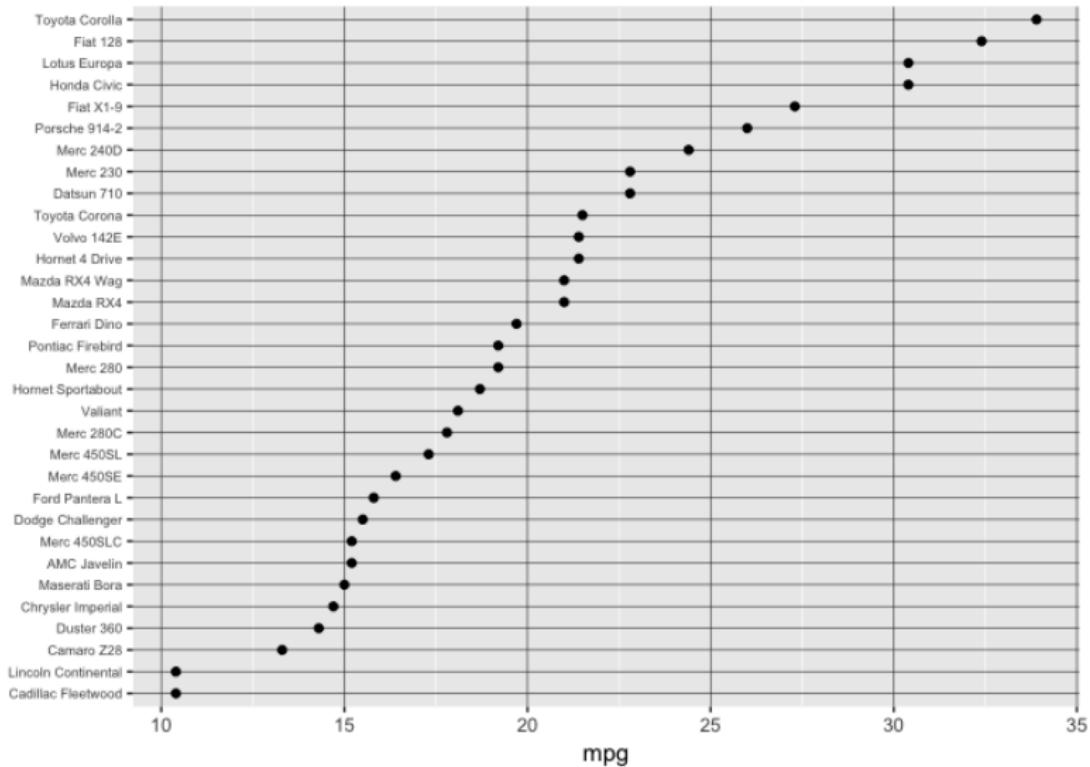


Ordre d'encodage

La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune**
- 2 Position sur une échelle identique non alignée
- 3 Longueur
- 4 Angle ou pente
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 Teinte de couleur

Exemple de position alignée : autos



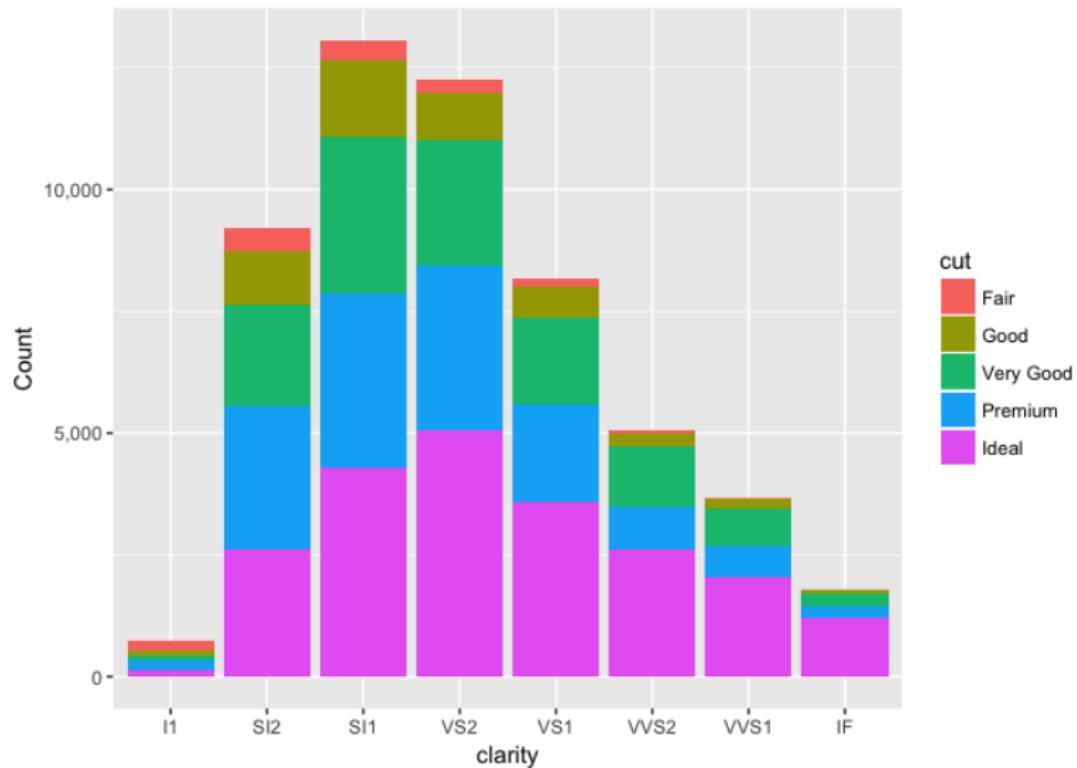
Ordre d'encodage

La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

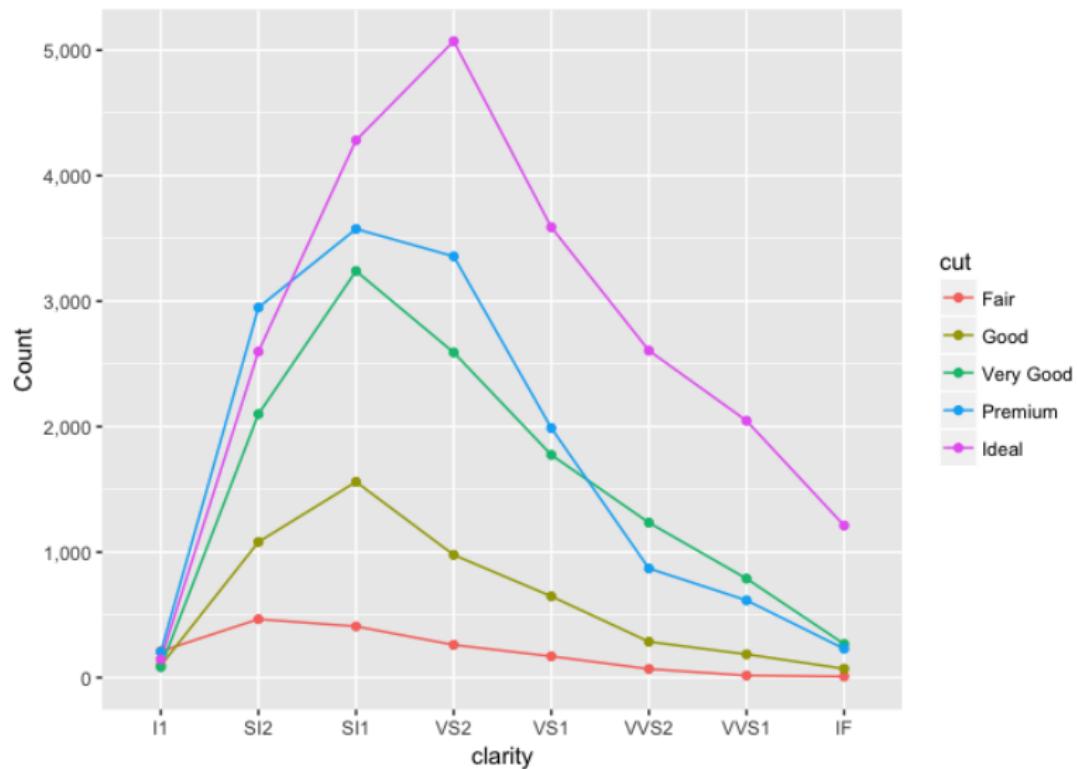
- 1 Position sur une échelle commune
- 2 Position sur une échelle identique non alignée
- 3 Longueur
- 4 Angle ou pente
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 Teinte de couleur

Empiler n'importe quoi est quasiment toujours une erreur.

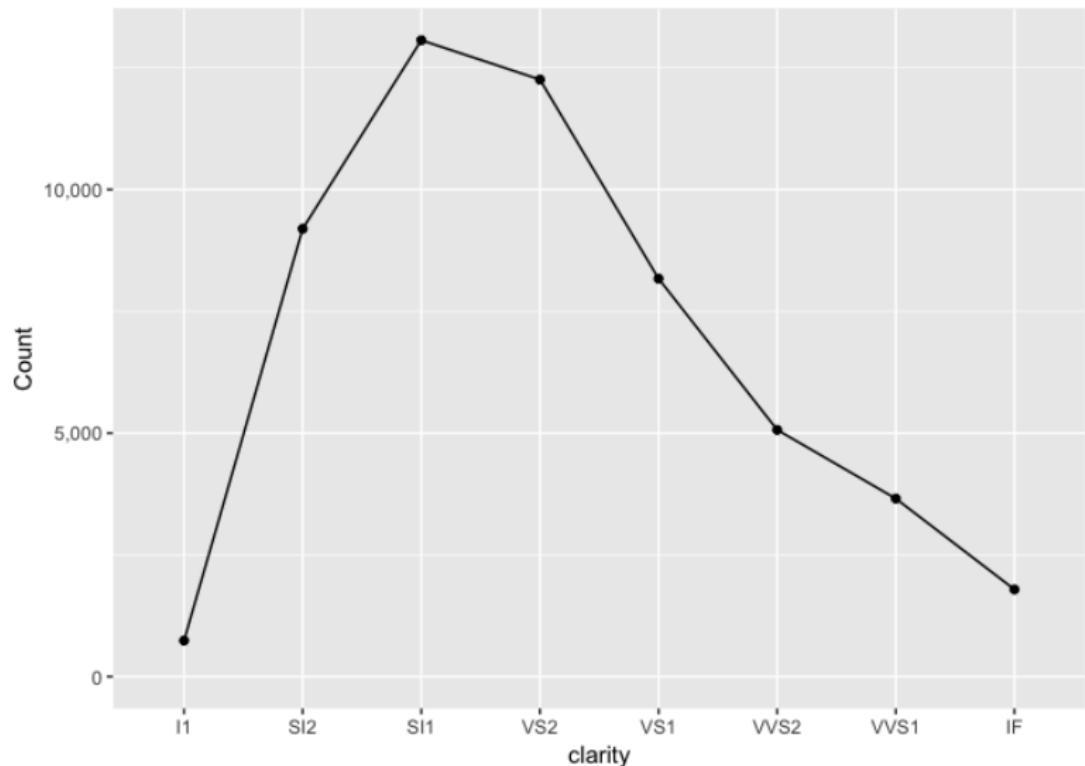
Exemple empilé : dataset diamants



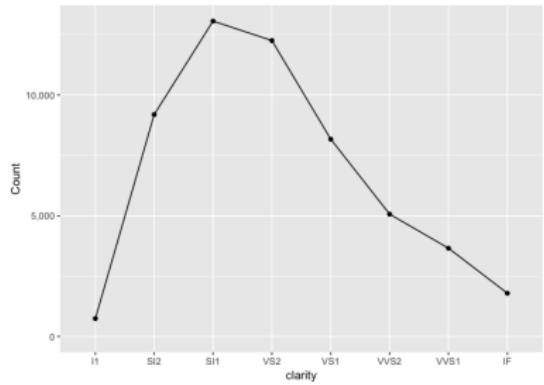
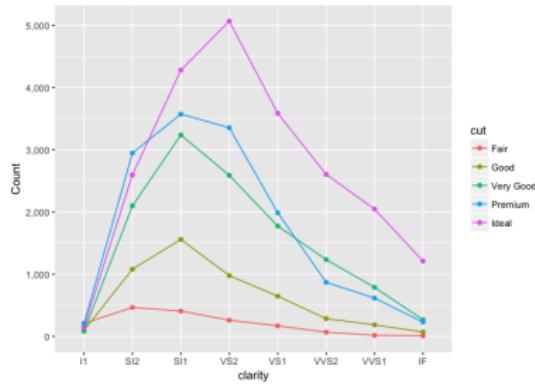
Exemple empilé : dataset diamants



Exemple empilé : dataset diamants



Exemple empilé : dataset diamants



Empiler provoque une interprétation des **longueurs** pas des positions sur une échelle commune.

**Empiler n'importe quoi est
quasiment toujours une
erreur.**

Exemple empilé : dataset OS market



Les *pie charts* sont
TOUJOURS une erreur.

"Pie charts are the information visualization equivalent of a roofing hammer to the frontal lobe."

Les camemberts sont l'équivalent visuel d'un coup de marteau sur le lobe frontal.

- Coda Hale

Pie Charts

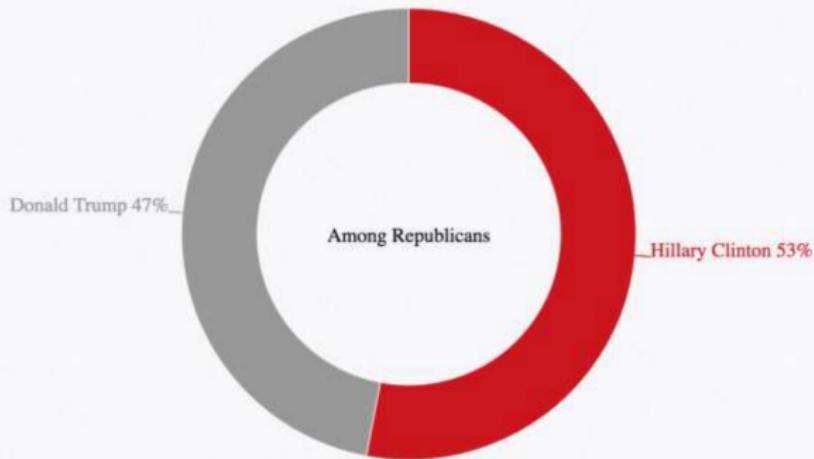
La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune
- 2 Position sur une échelle identique non alignée
- 3 Longueur
- 4 **Angle ou pente**
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 Teinte de couleur

Who do you think did a better job in tonight's debate?

Among Republicans

Among Democrats



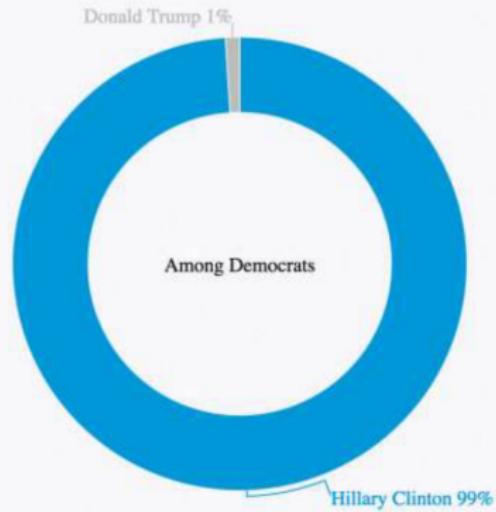
Share

POLITICO

Who do you think did a better job in tonight's debate?

Among Republicans

Among Democrats



Share

POLITICO

Tables are preferable to graphics for many small data sets.

A table is nearly always better than a dumb pie chart; the only thing worse than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies. . . Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.

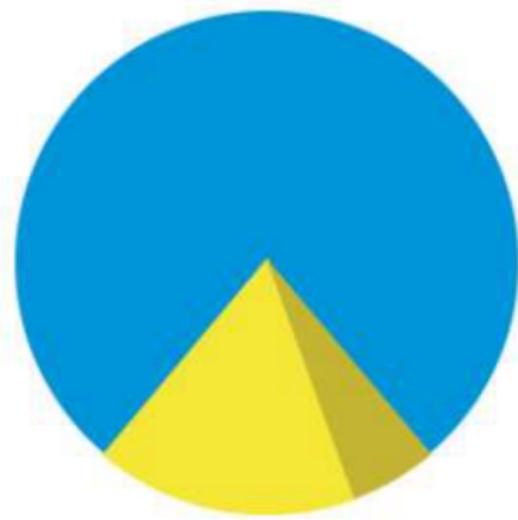
-Edward Tufte, The Visual Display of Quantitative Information

Qui pensez vous a fait un meilleur débat ?

	Clinton	Trump
Démocrates	99%	1%
Républicains	53%	47%

Les bons *pie charts* sont
des **blagues**.

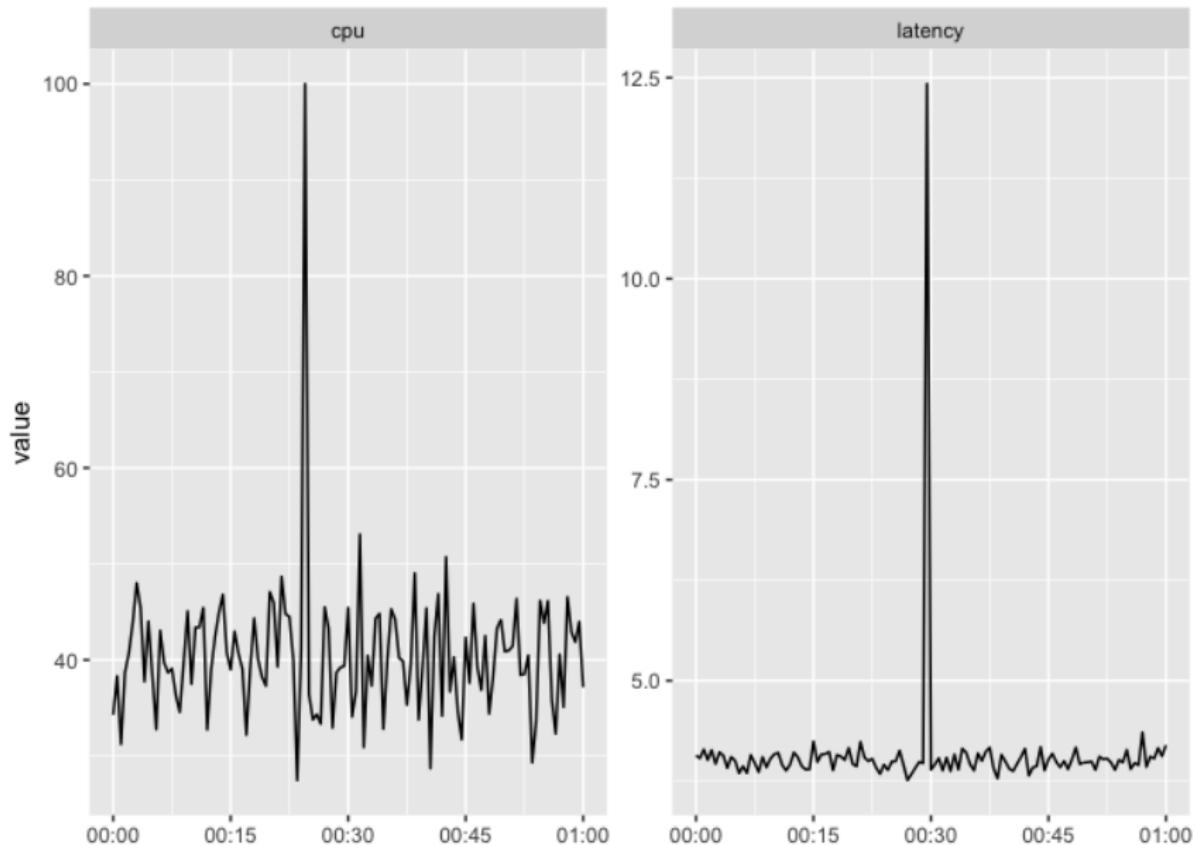
Exemple de pie chart égyptien



- Sky
- Sunny side of pyramid
- Shady side of pyramid

**La comparaison est triviale
sur une échelle commune**

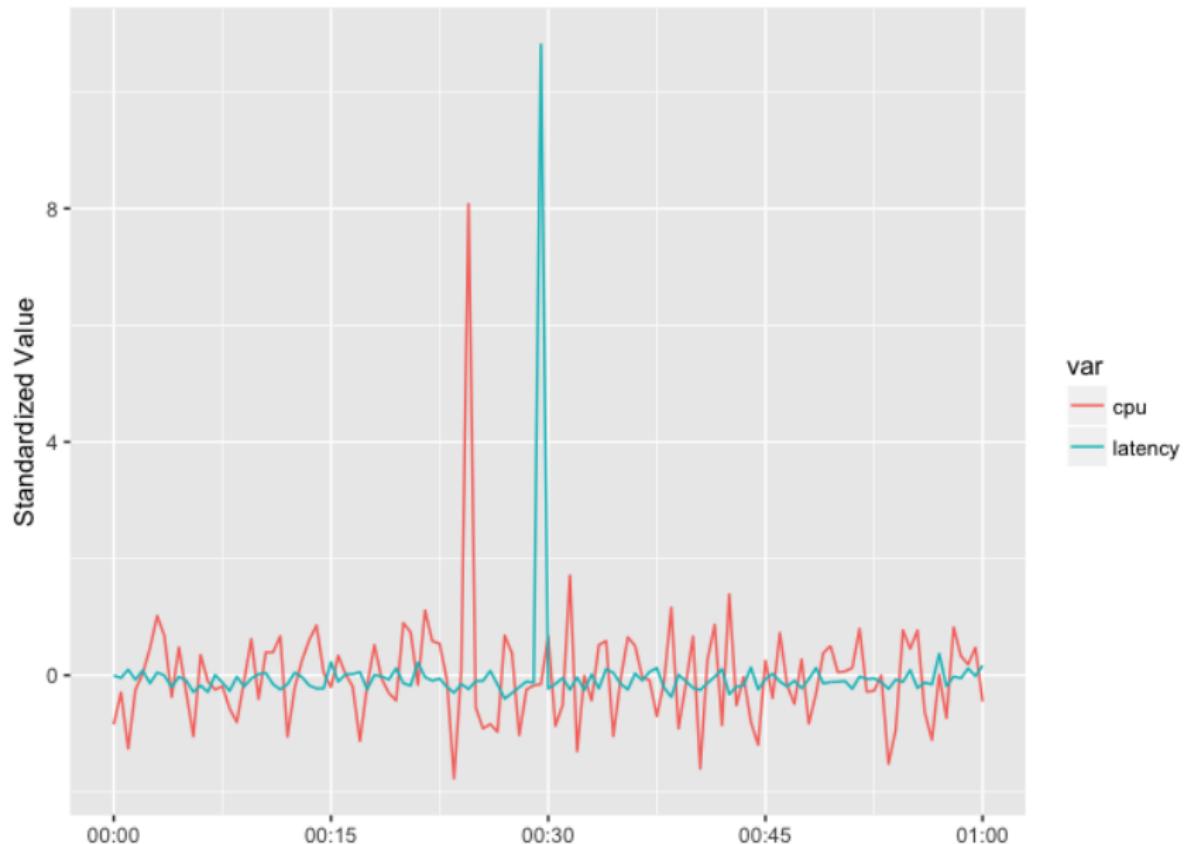
Exemple de comparaison : latence



Exemple de comparaison : latency

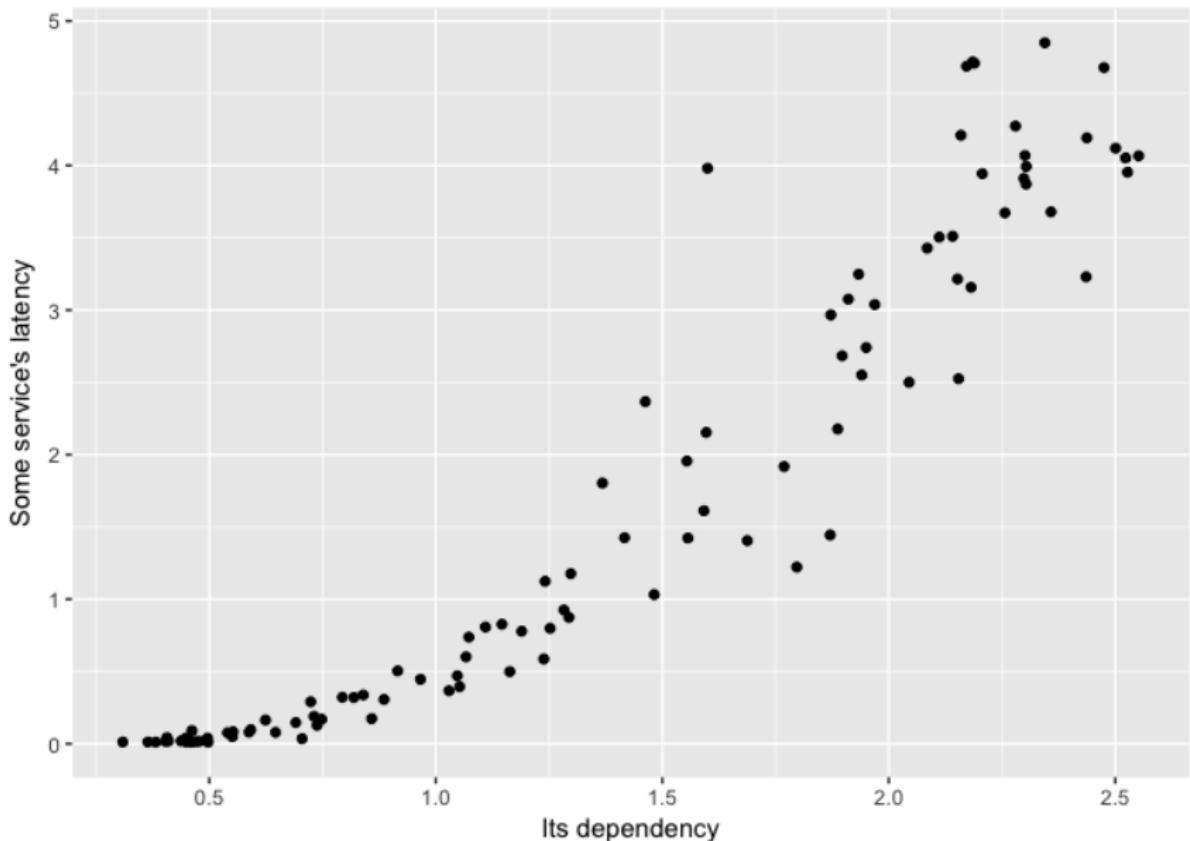


Exemple de comparaison : latency

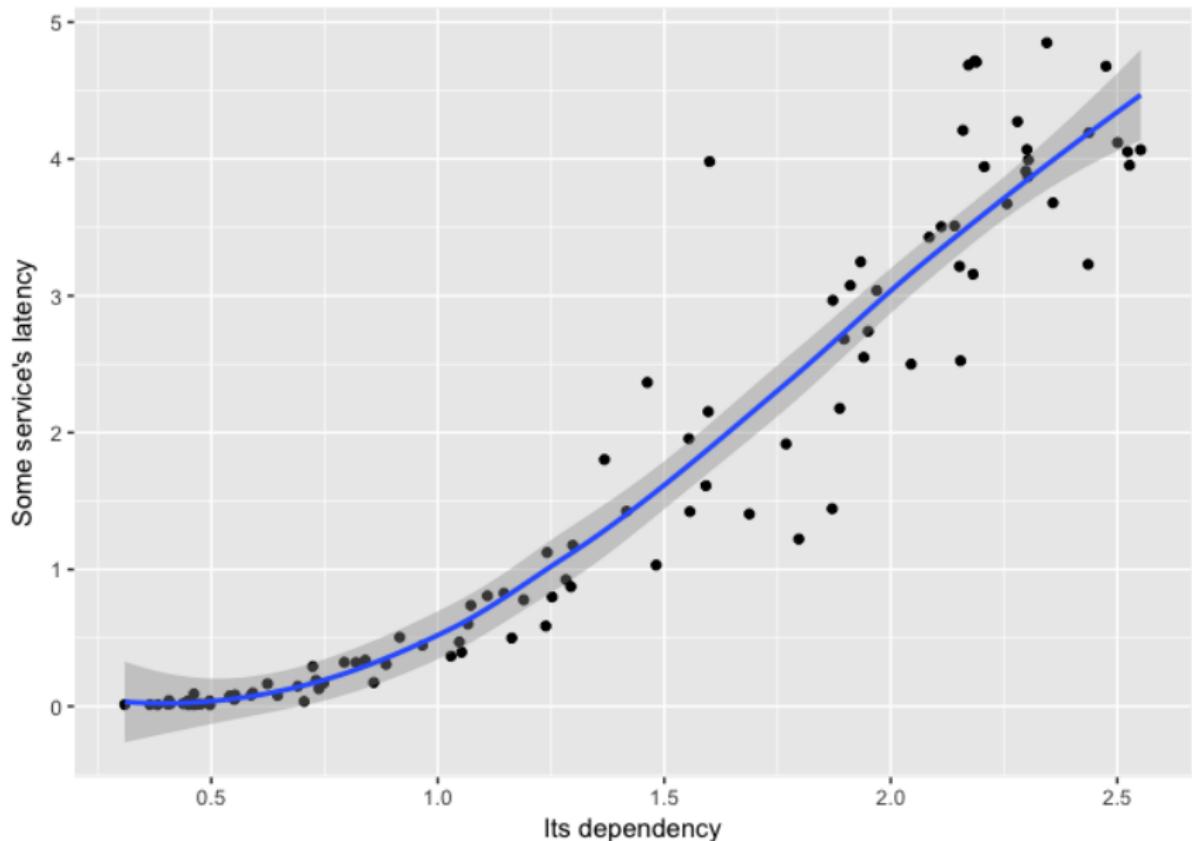


Les *scatters plots* montrent les **relations** directement.

Exemple de *scatters plots* : latence

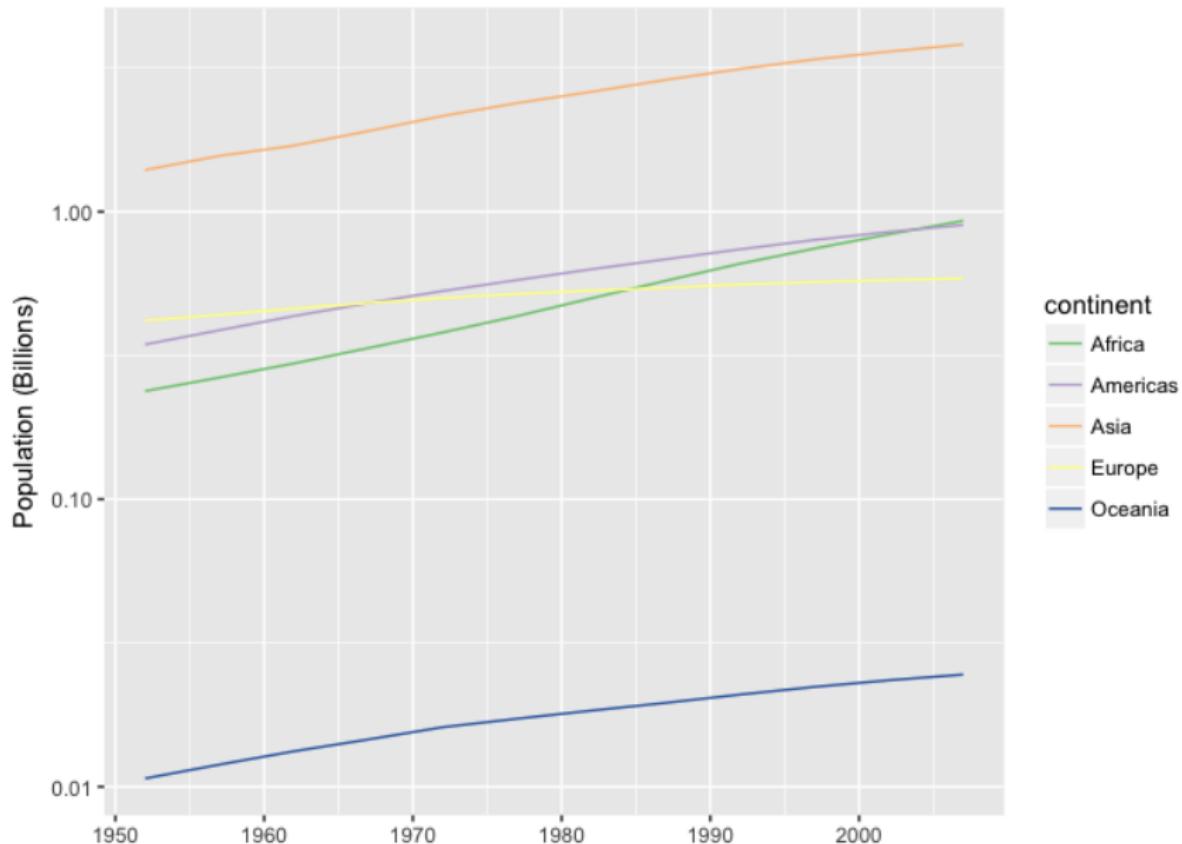


Exemple de *scatters plots* : latence



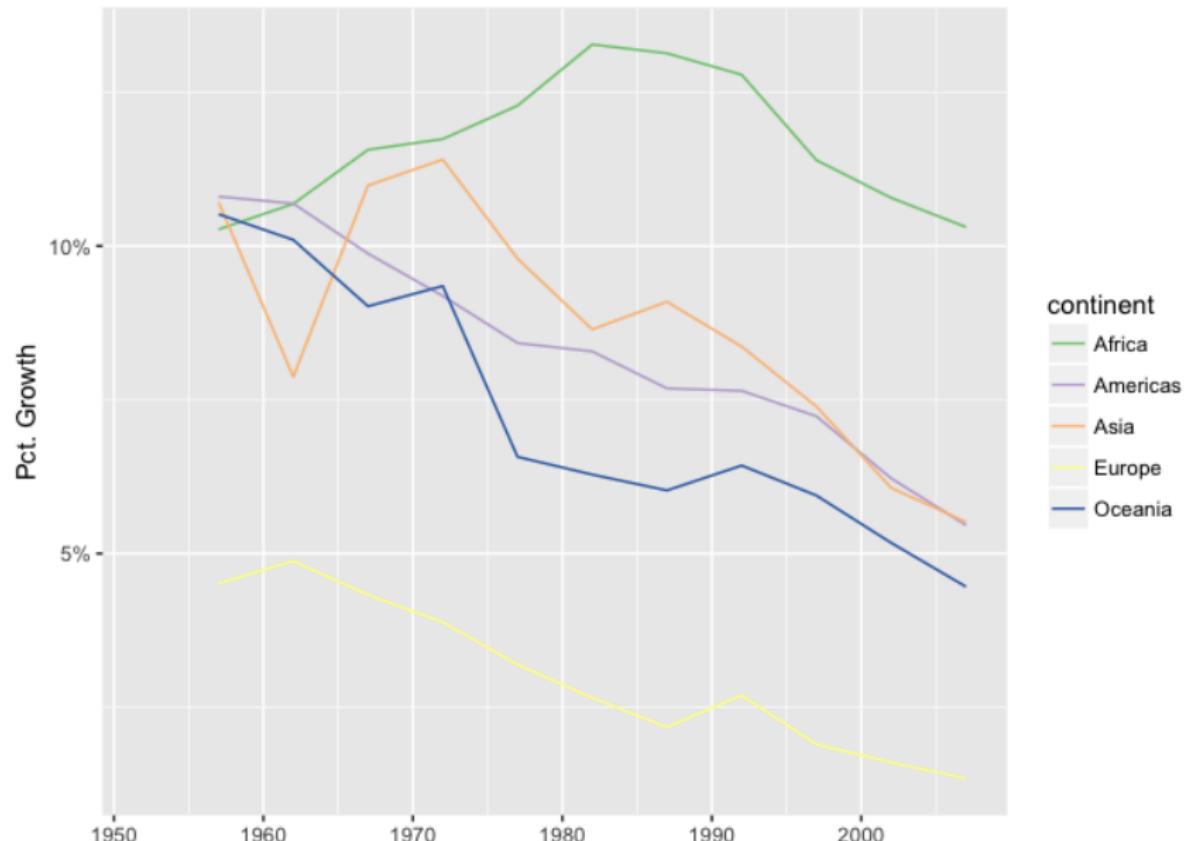
**Les graphes de croissance
n'en sont pas
(usuellement).**

Exemple de *growth chart* : Pop. mondiale



**Si la croissance
(pente/dérivée)
est importante,
affichez-la directement**

Exemple de *growth chart* : Pop. mondiale



La mesure la plus importante devrait exploiter l'encodage ayant le rang le plus haut :

- 1 Position sur une échelle commune
- 2 Position sur une échelle identique non alignée
- 3 Longueur
- 4 Angle ou pente
- 5 Aire
- 6 Volume ou Densité ou Saturation de couleur
- 7 Teinte de couleur

Outline

Trois opérations visuelles dans la perception de schémas :

- 1** Détection
- 2** **Construction**
- 3** Estimation

Principes d'organisation implicite :

- 1 Réification
- 2 Émrgence
- 3 Prägnanz : Concis et plein de sens
 - 1 Fermeture
 - 2 Continuation
 - 3 Proximité
 - 4 Similarité

Réification

A



B



C



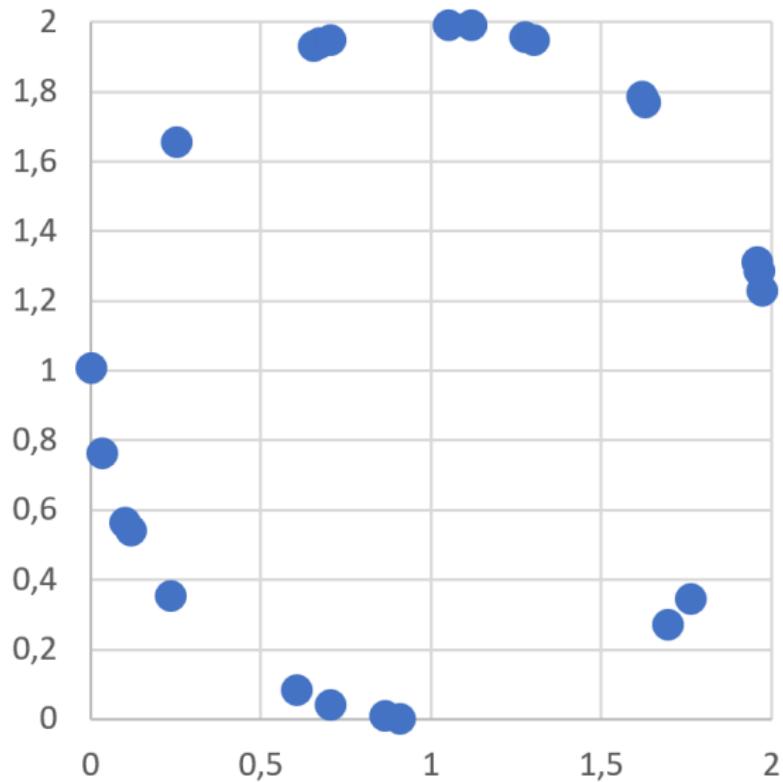
D



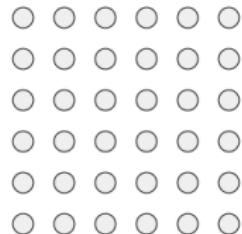
Source : https://en.wikipedia.org/wiki/Gestalt_psychology



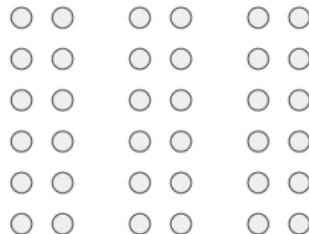
Source : https://en.wikipedia.org/wiki/Gestalt_psychology



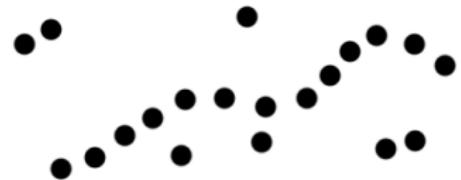
Prägnanz



Proximité

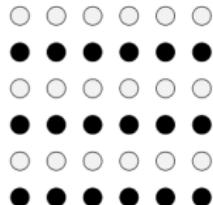


Fermeture

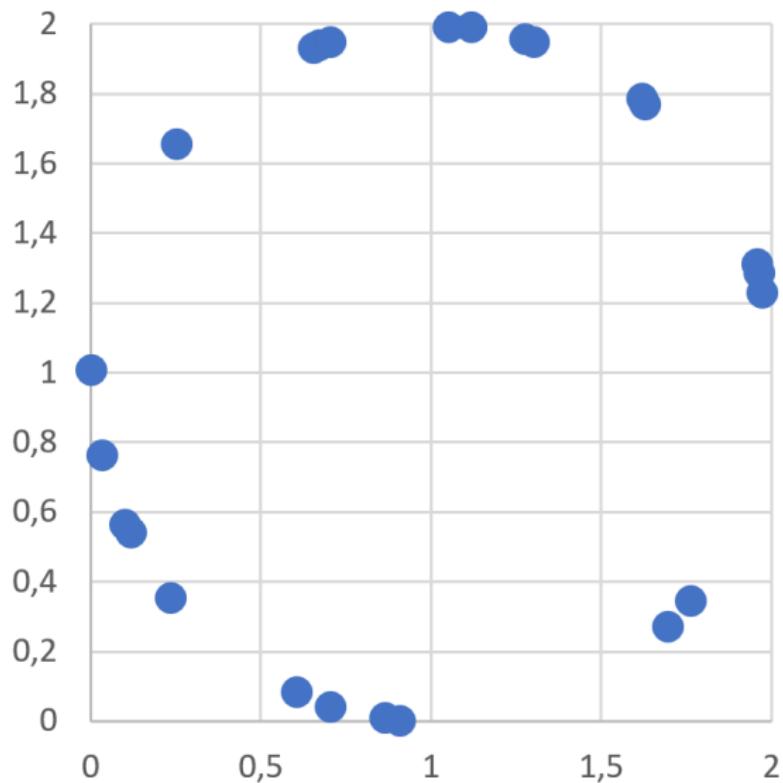


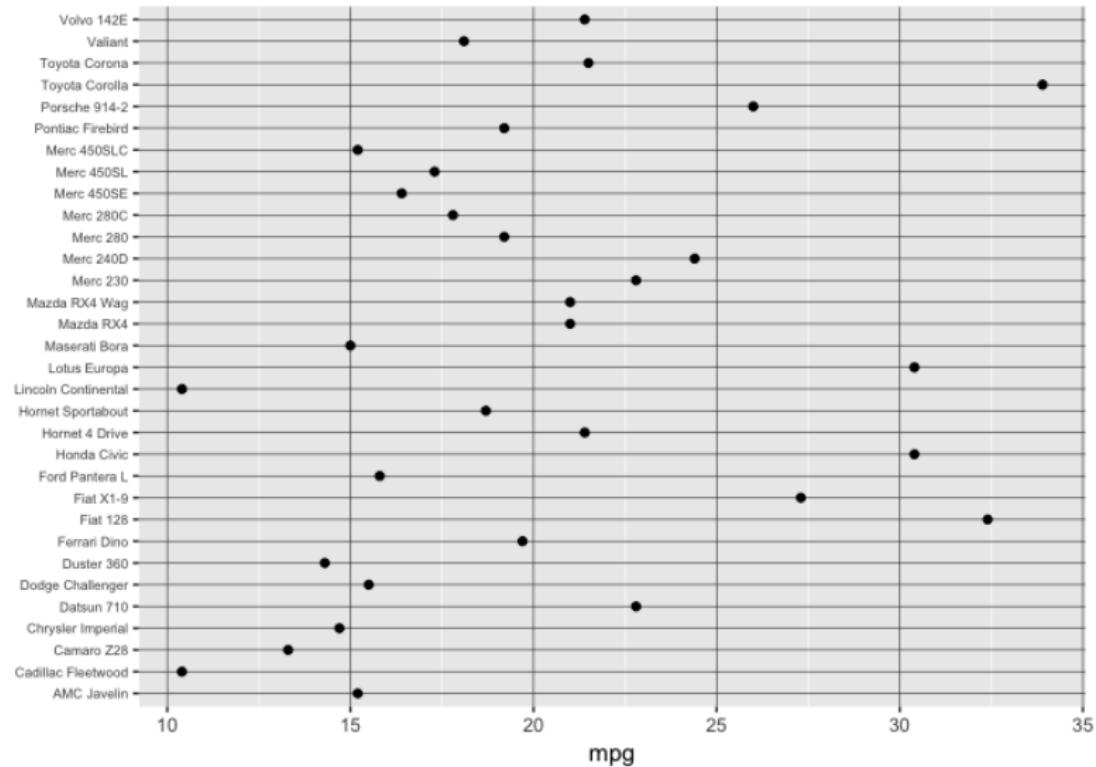
Continuité

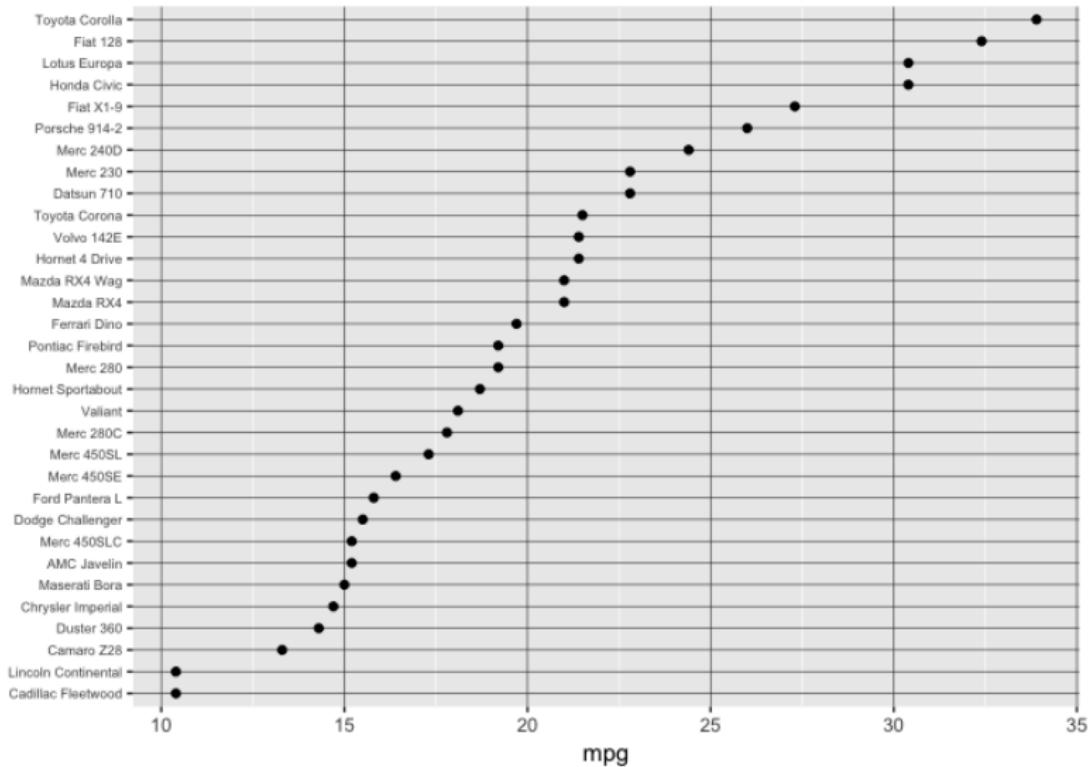
Similarité



Fermeture



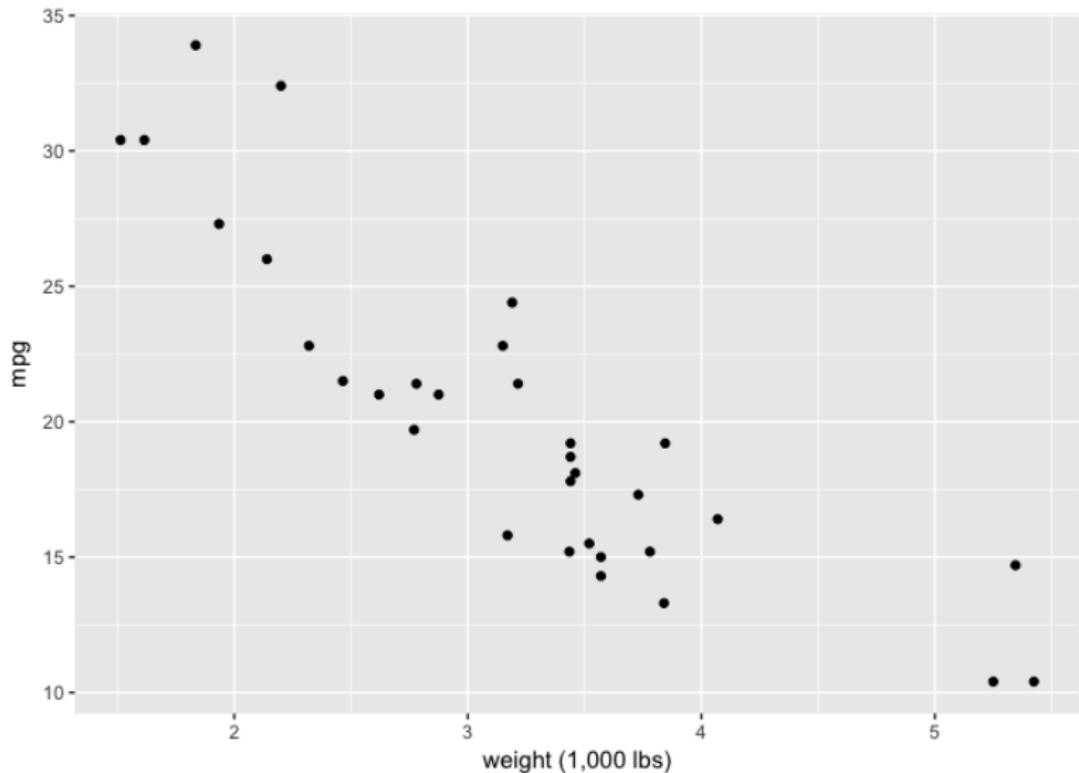




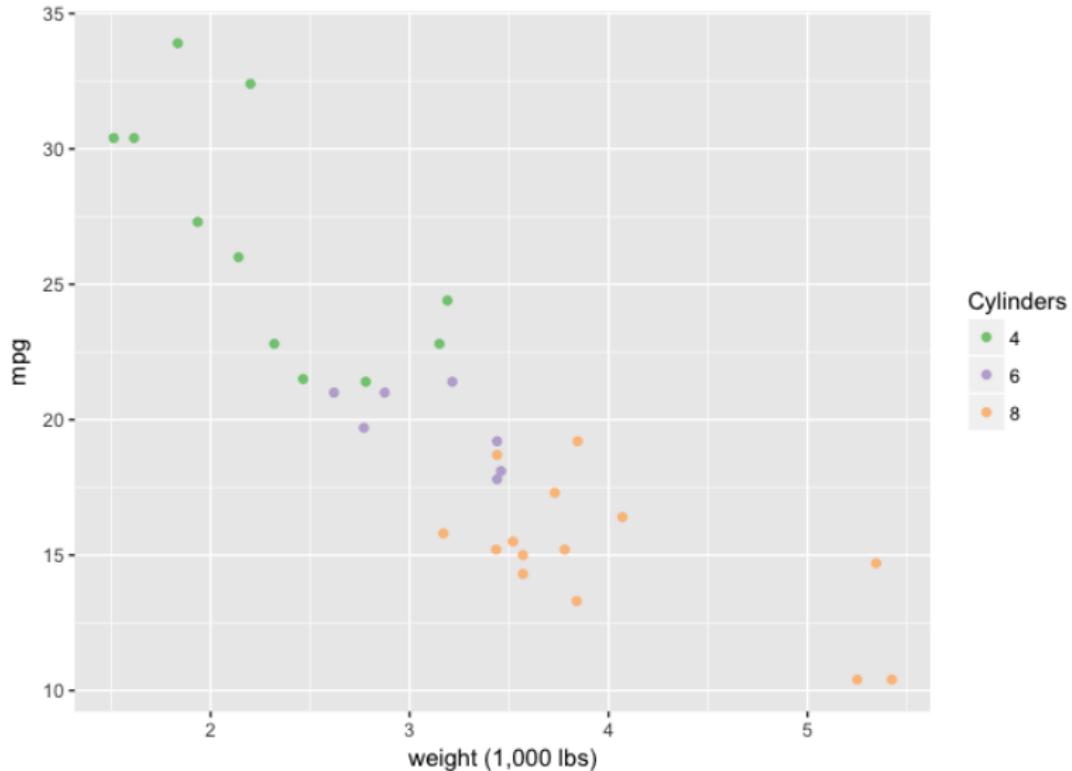
**Les bons graphiques
avantagent la continuité
pour améliorer la
construction.**

Loi de similarité

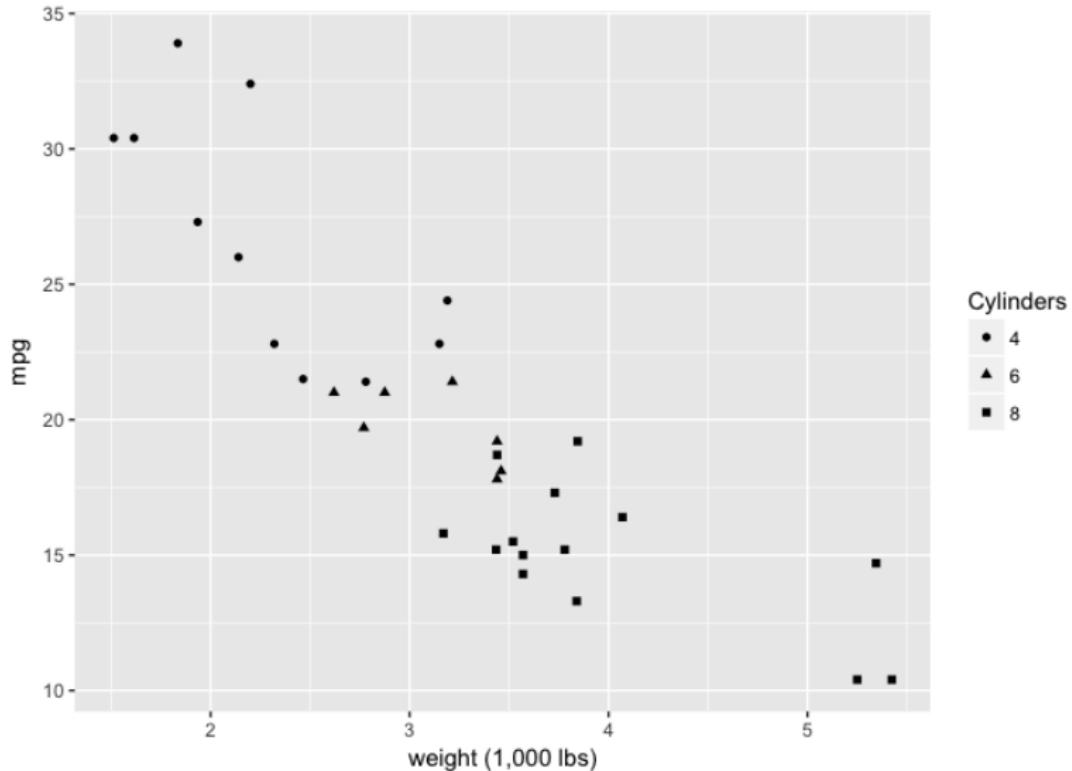
Similarité



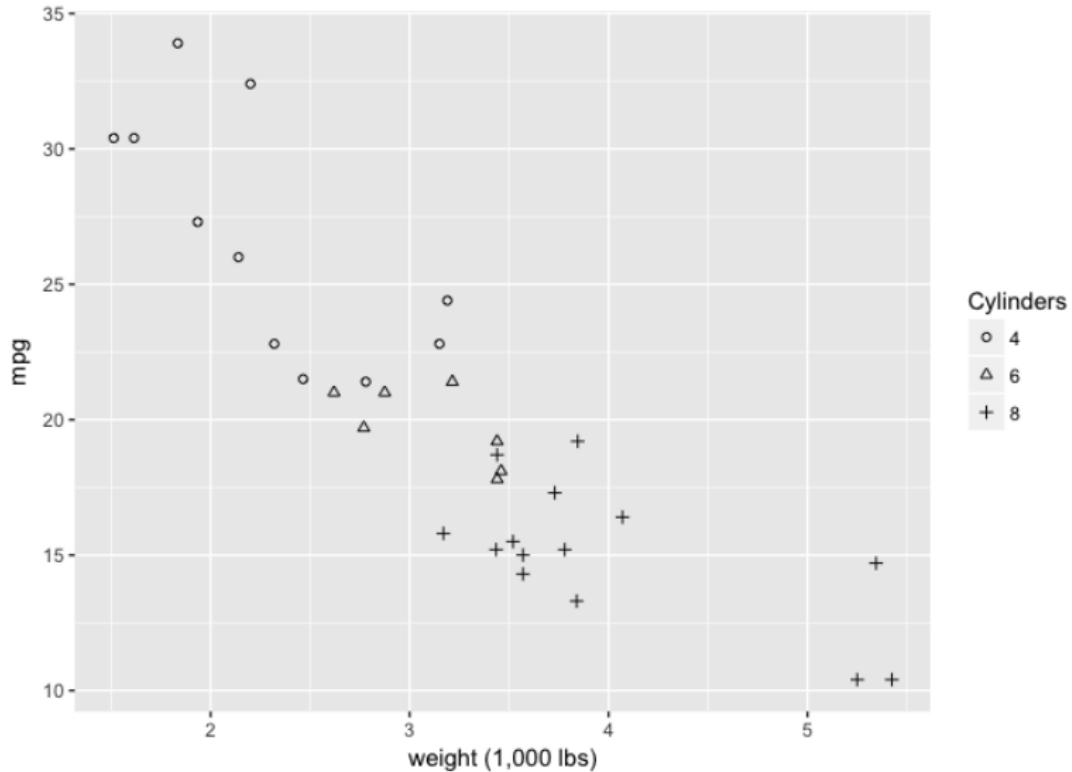
Similarité



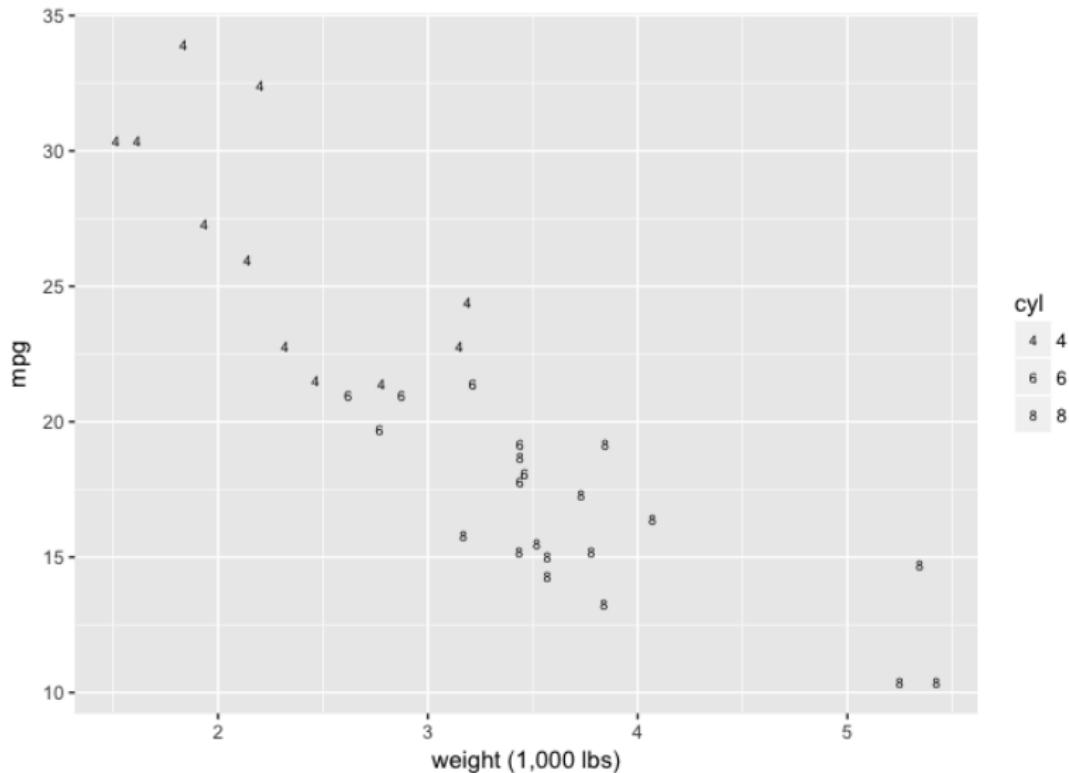
Similarité



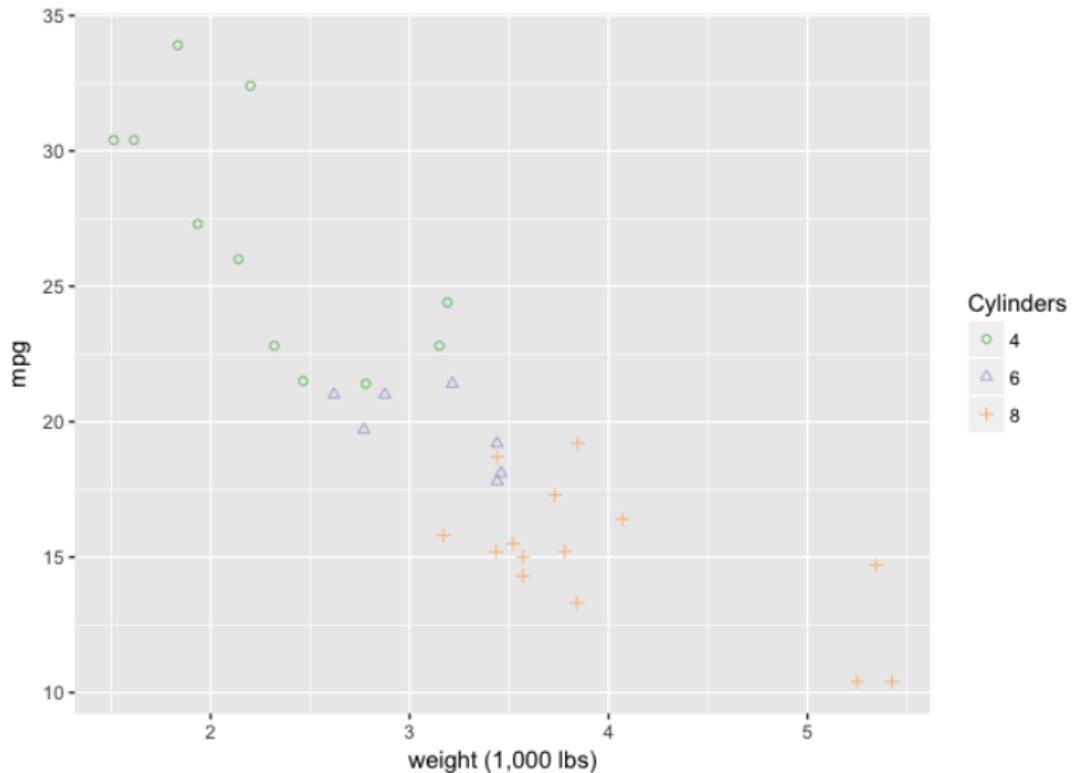
Similarité



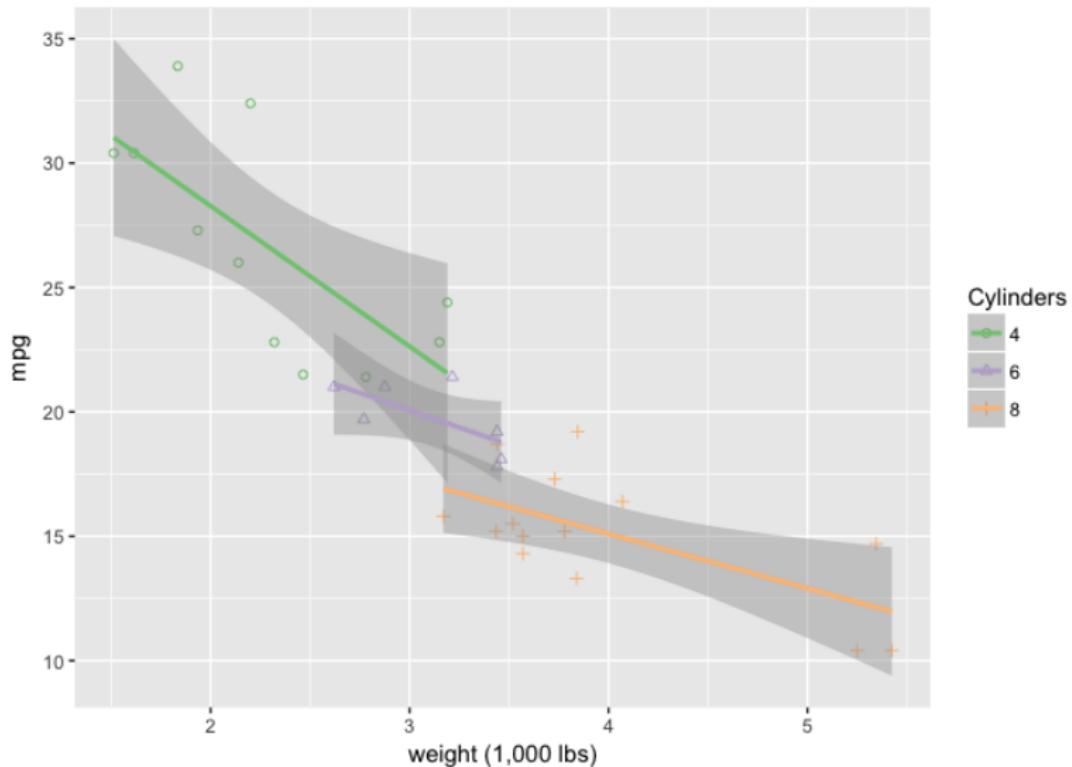
Similarité



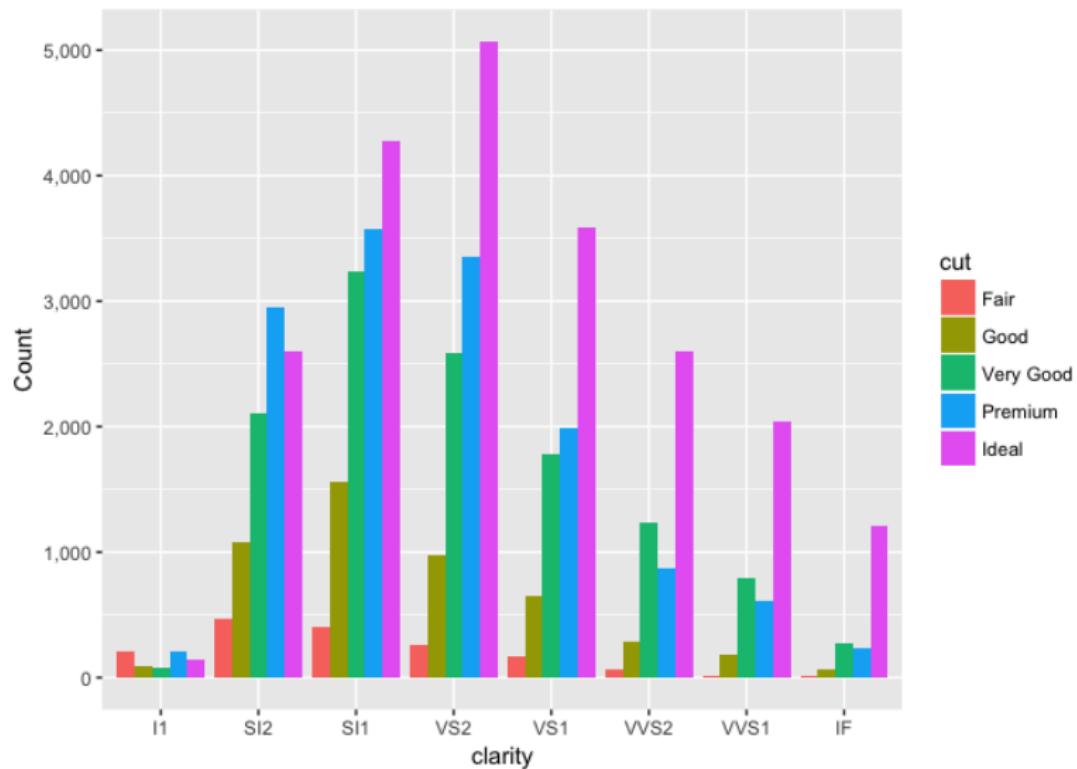
Similarité



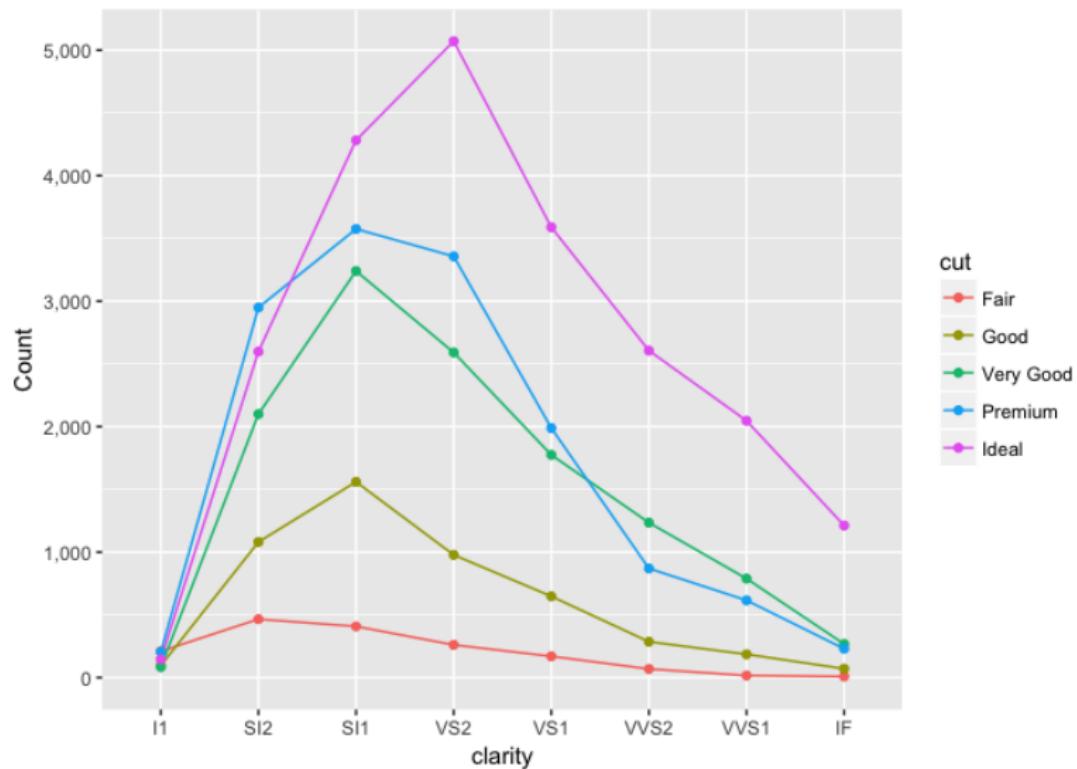
Similarité



Loi de proximité



Proximité

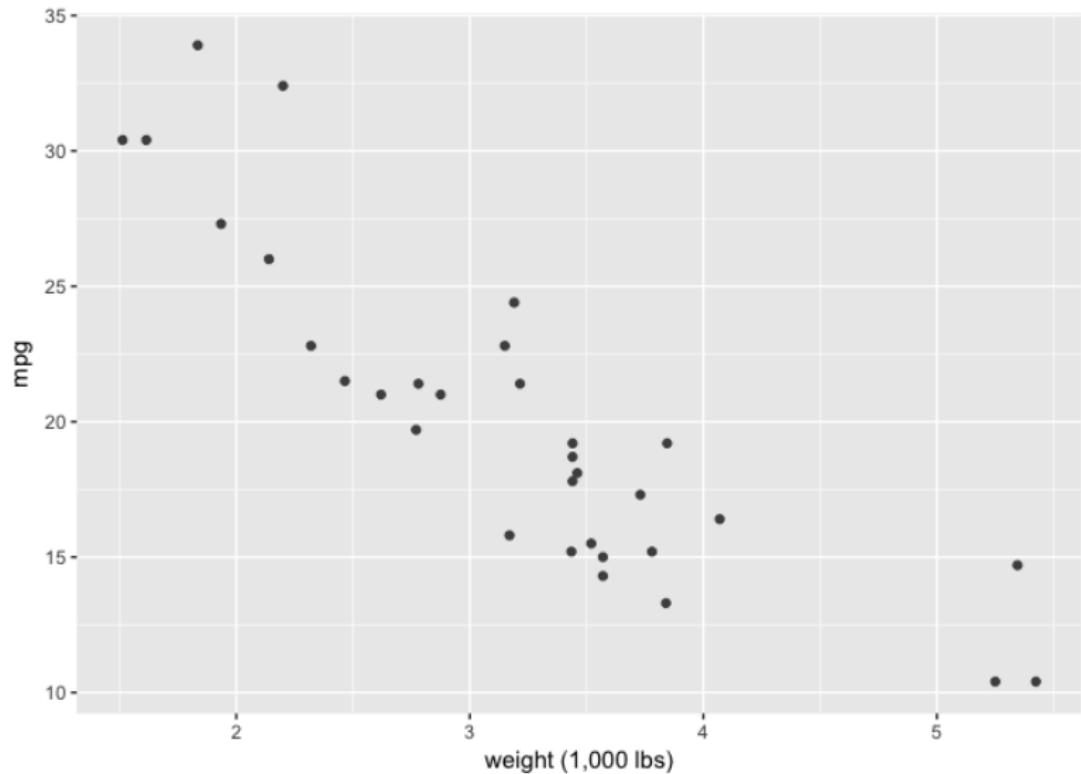


**Les diagrammes en bâtons
sont usuellement une
mauvaise idée.**

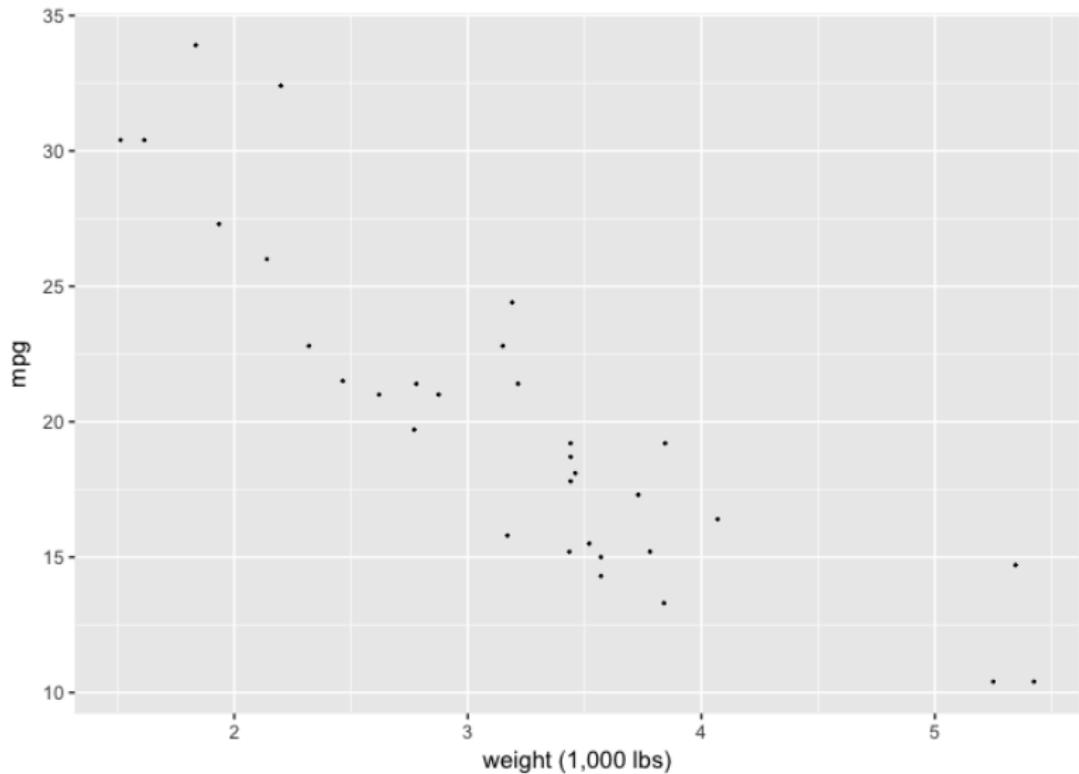
Trois opérations visuelles dans la perception de schémas :

- 1 Détection**
- 2 Construction**
- 3 Estimation**

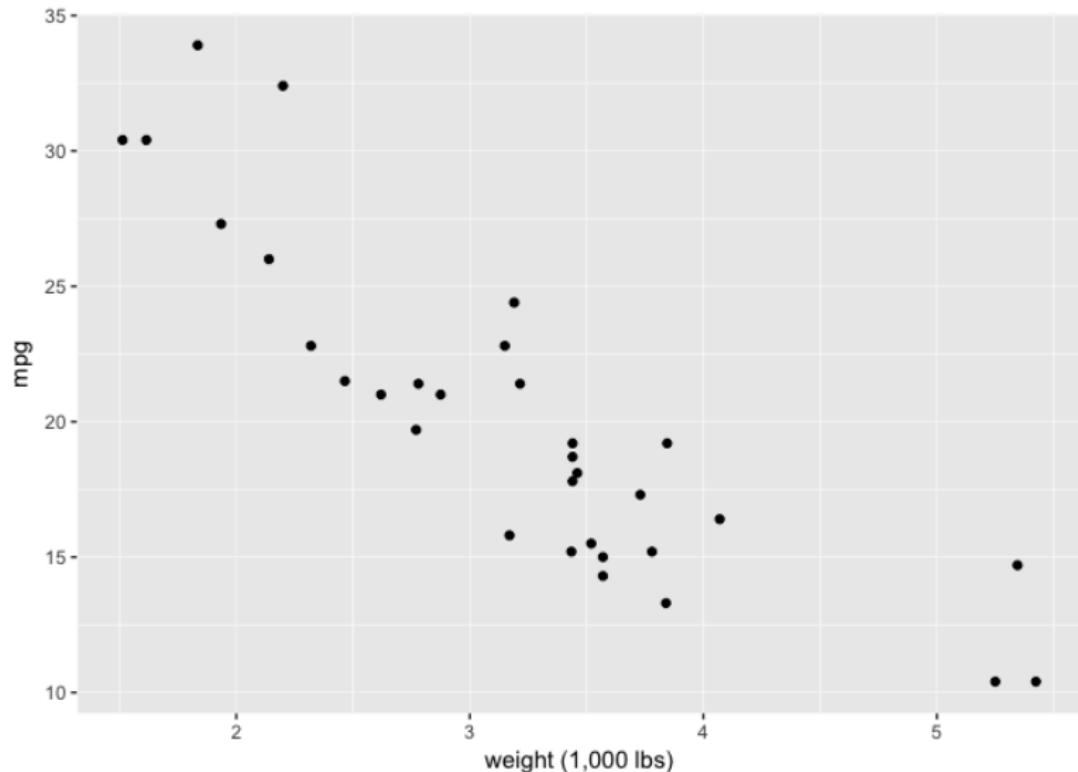
Détection



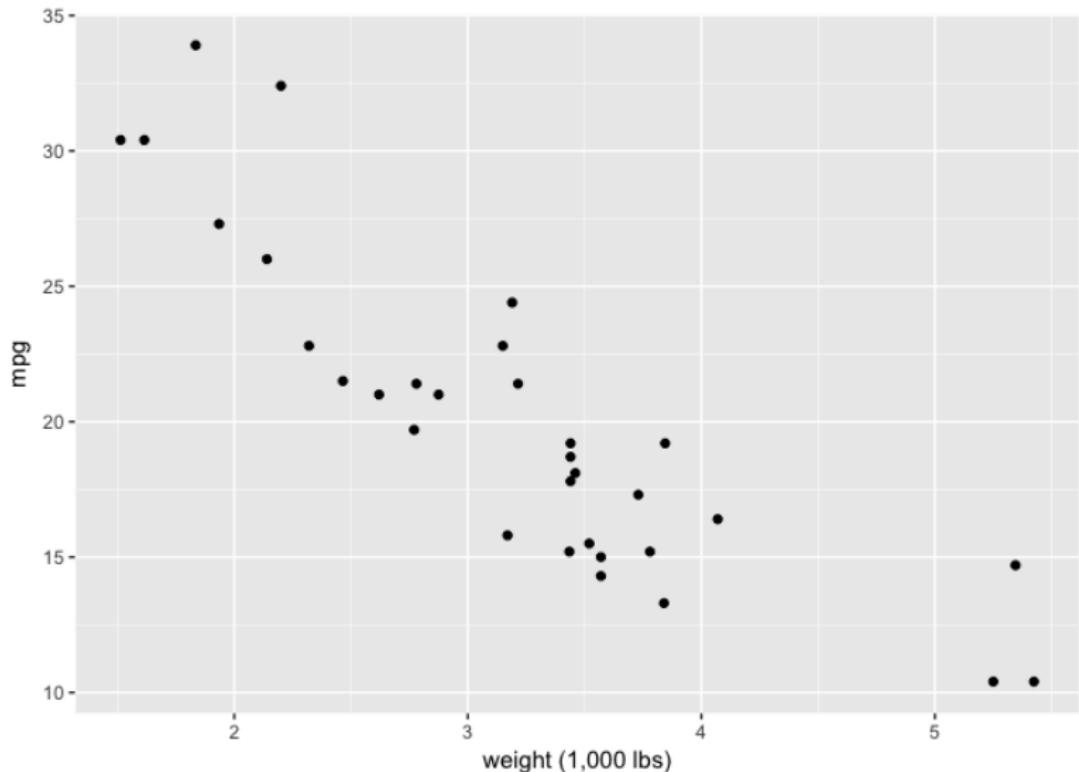
Détection



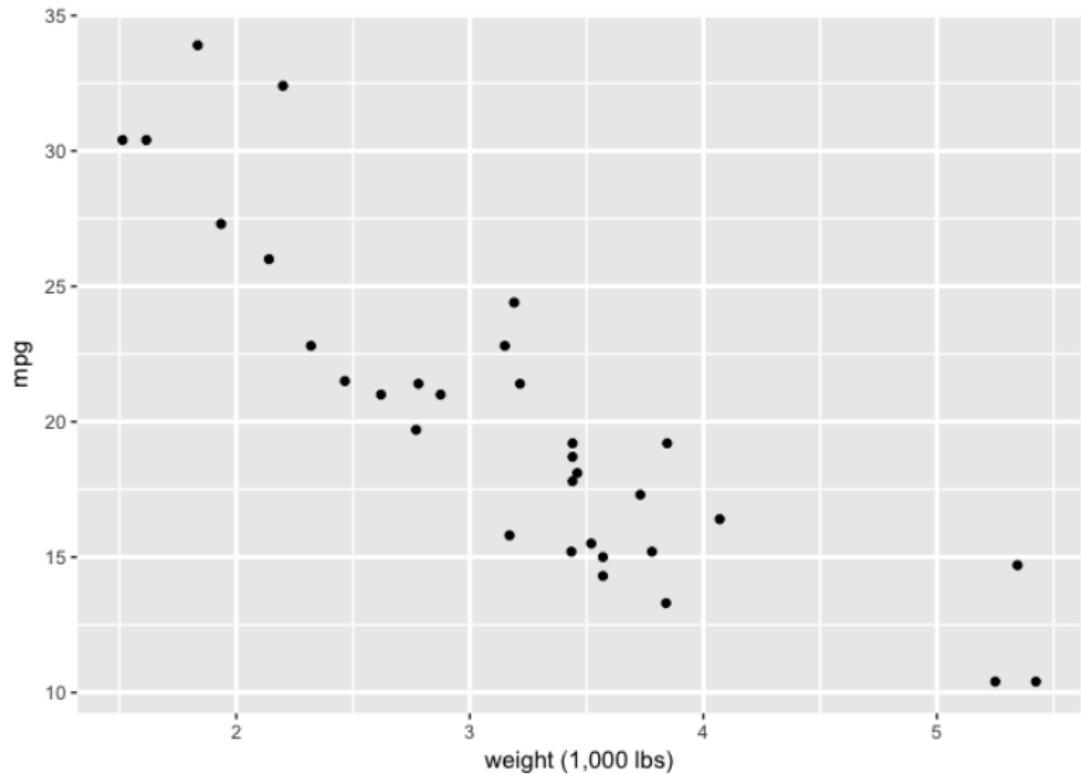
Détection



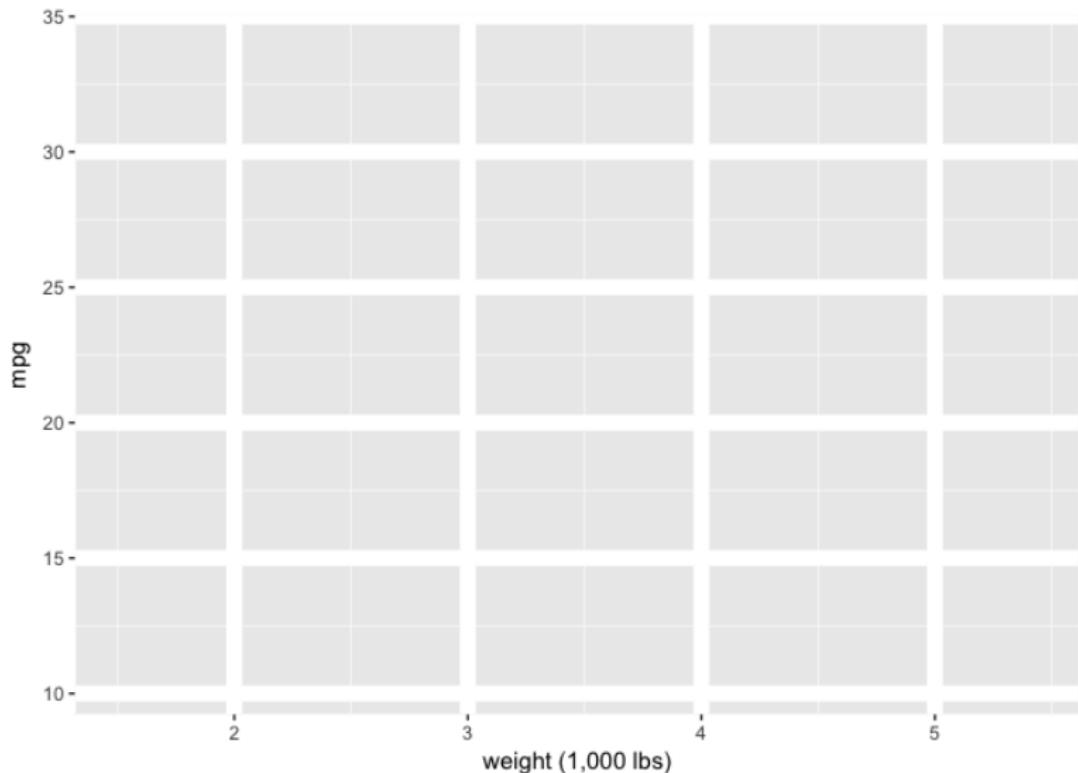
Détection



Détection

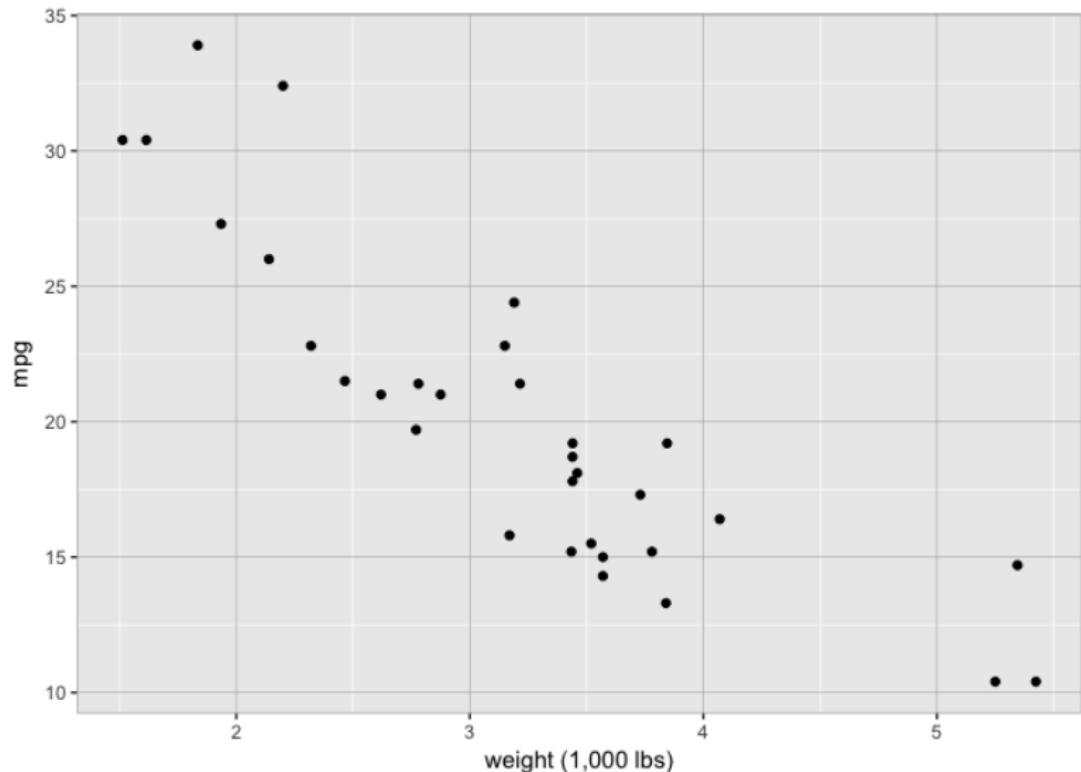


Détection



La détection n'est pas aussi triviale qu'elle ne paraît.

Détection



"Avant tout, montrez les données."

- *Tufte*

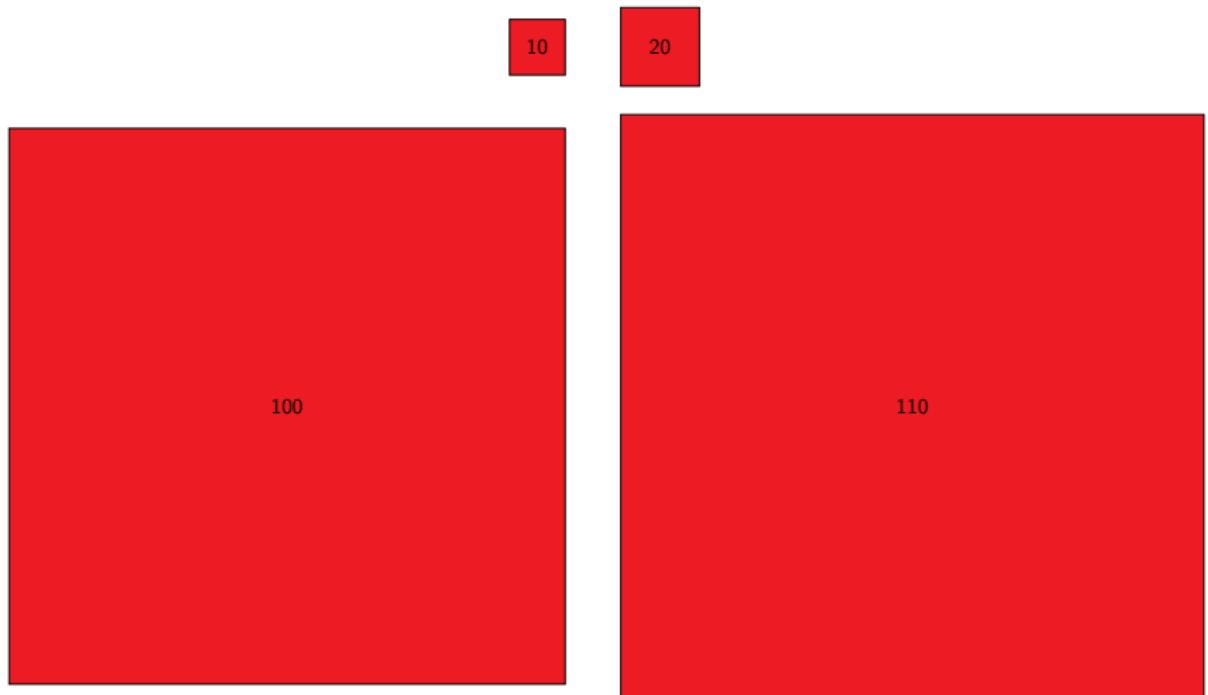
Autres résultats utiles

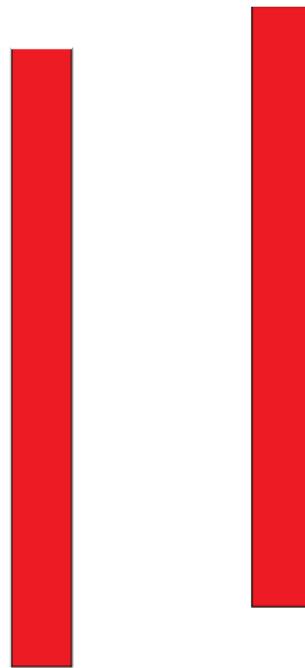
La "différence juste notable" est proportionnel à la taille du stimuli initial.

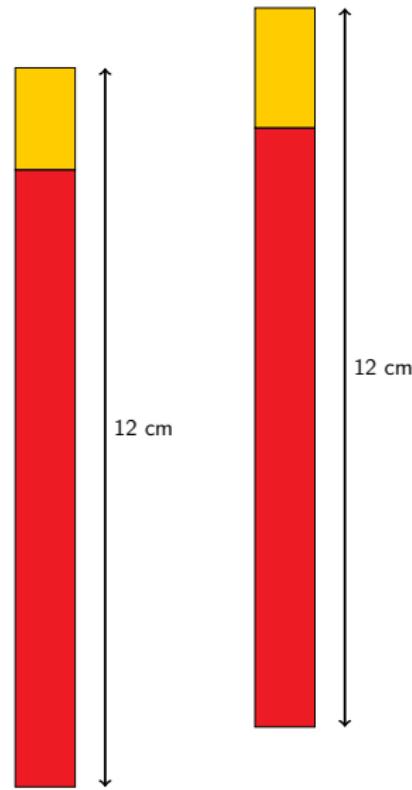
Loi de Weber



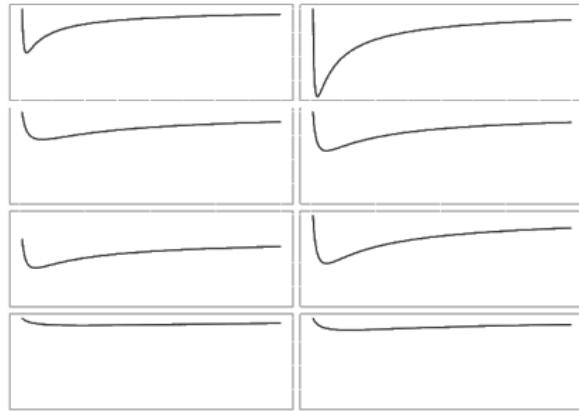
Loi de Weber

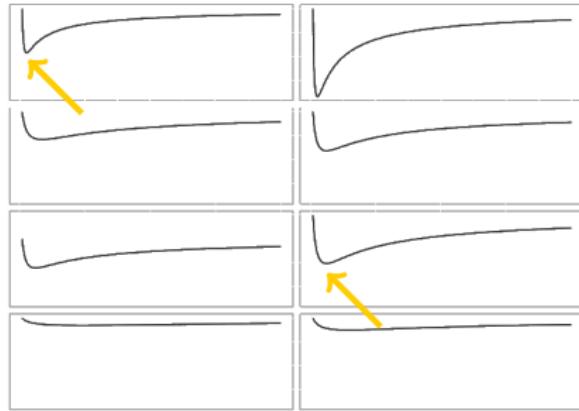




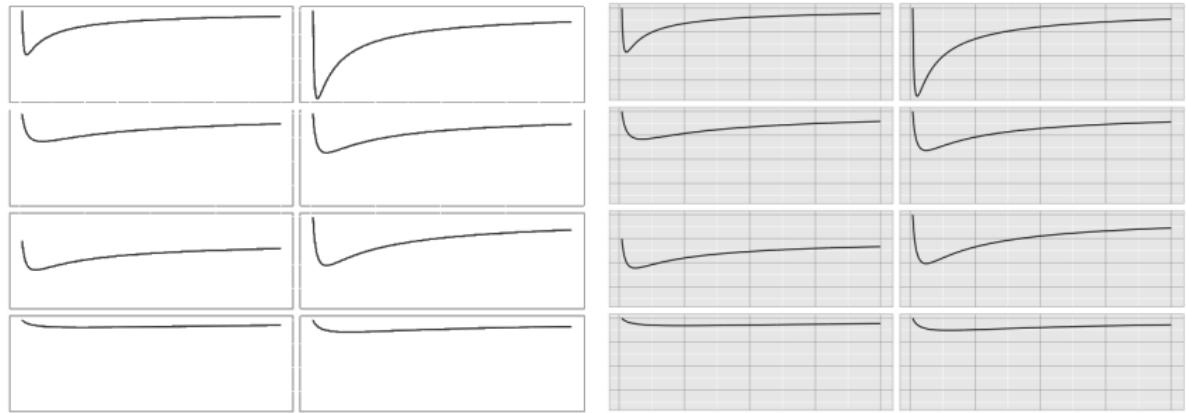


**La loi de Weber est la
raison de l'utilité des grilles
de fond.**





Loi de Weber



"Effacez tout ce qui n'est pas des données."

- *Tufte*

"Effacez tout ce qui n'est pas des données **avec une bonne raison.**"

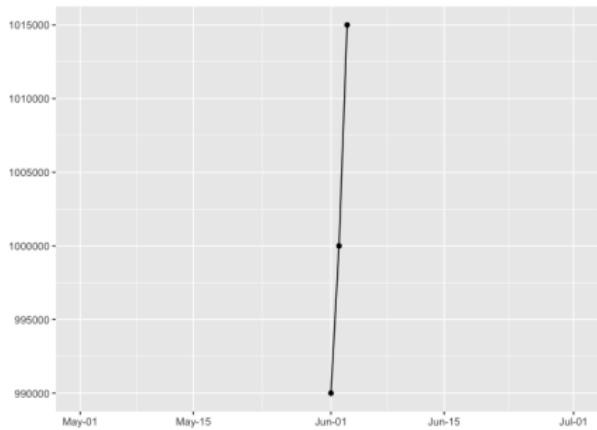
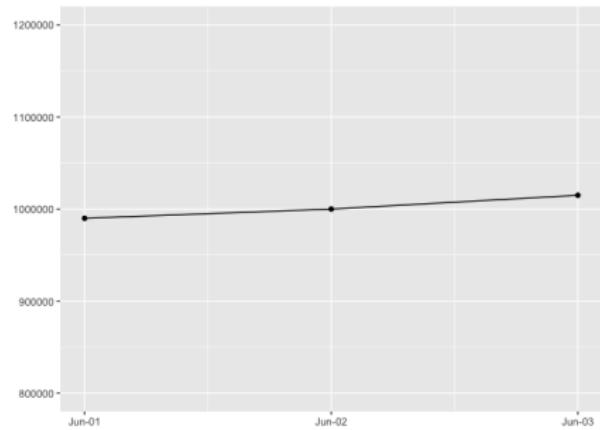
- *Tufte*

"Effacez tout ce qui n'est pas des données et qui interfère avec la détection, la construction ou l'estimation."

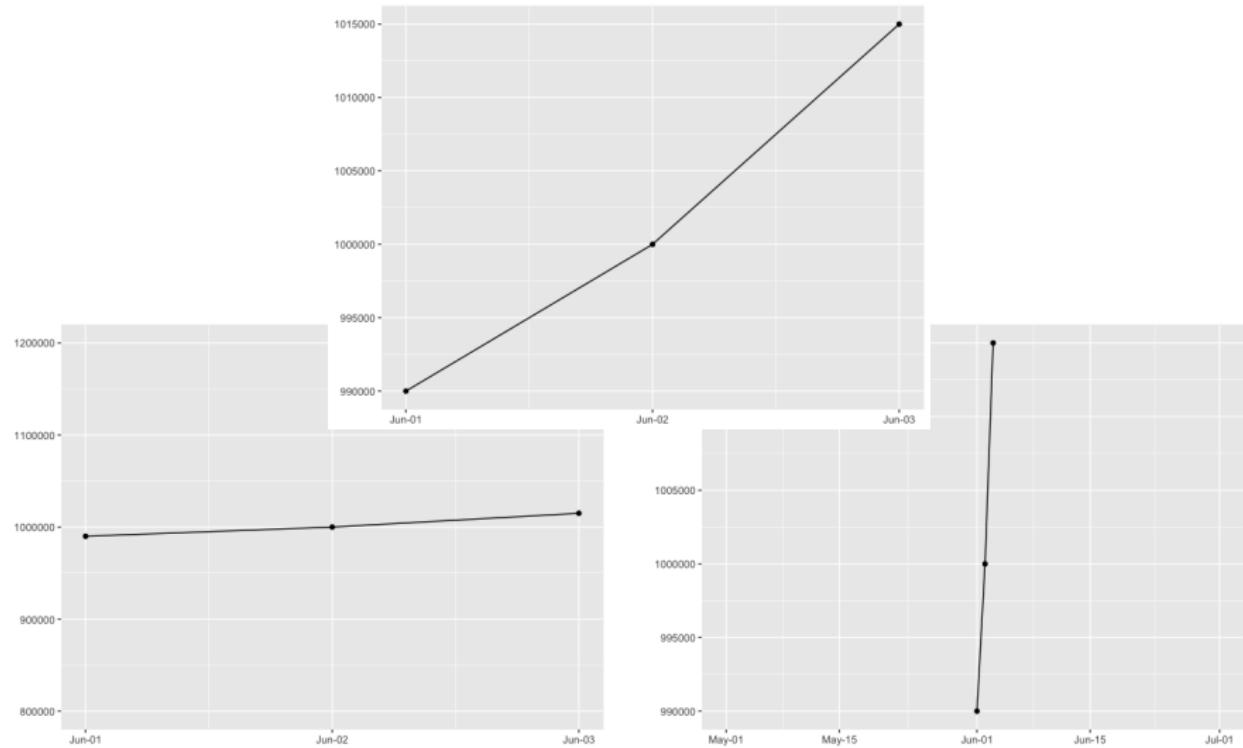
- *Rause (via Tufte)*

Observation

Vous êtes meilleur pour détecter les variations de pente autour de 45 degrés.



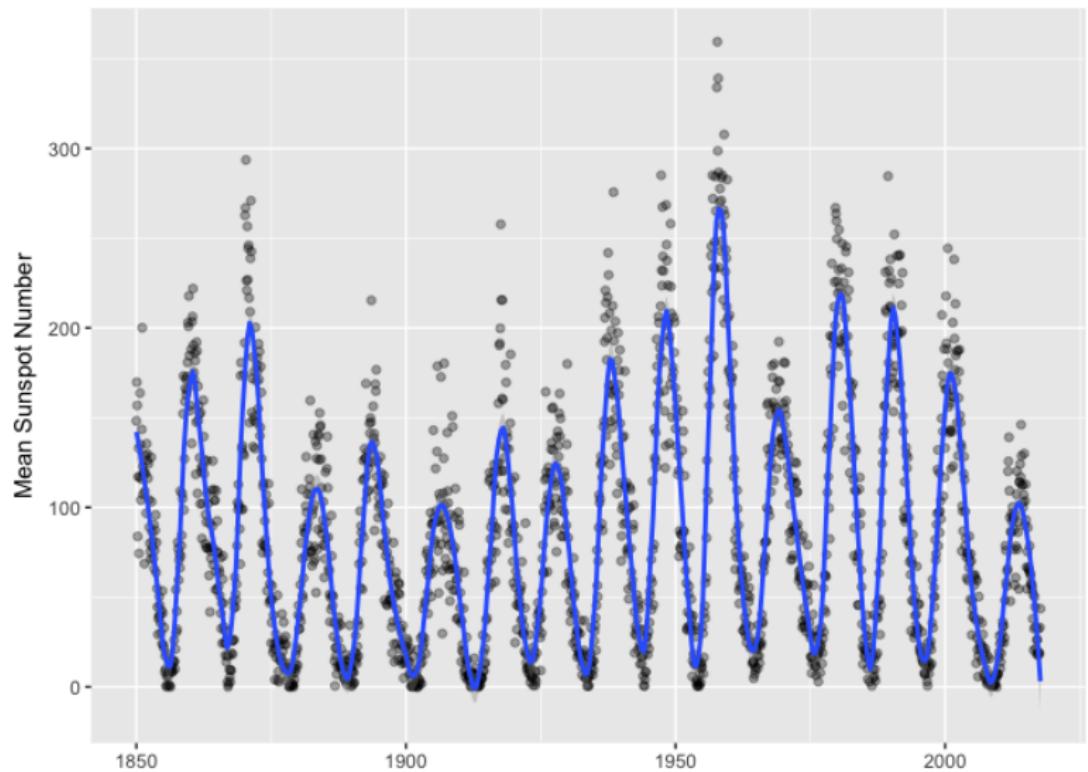
Pente



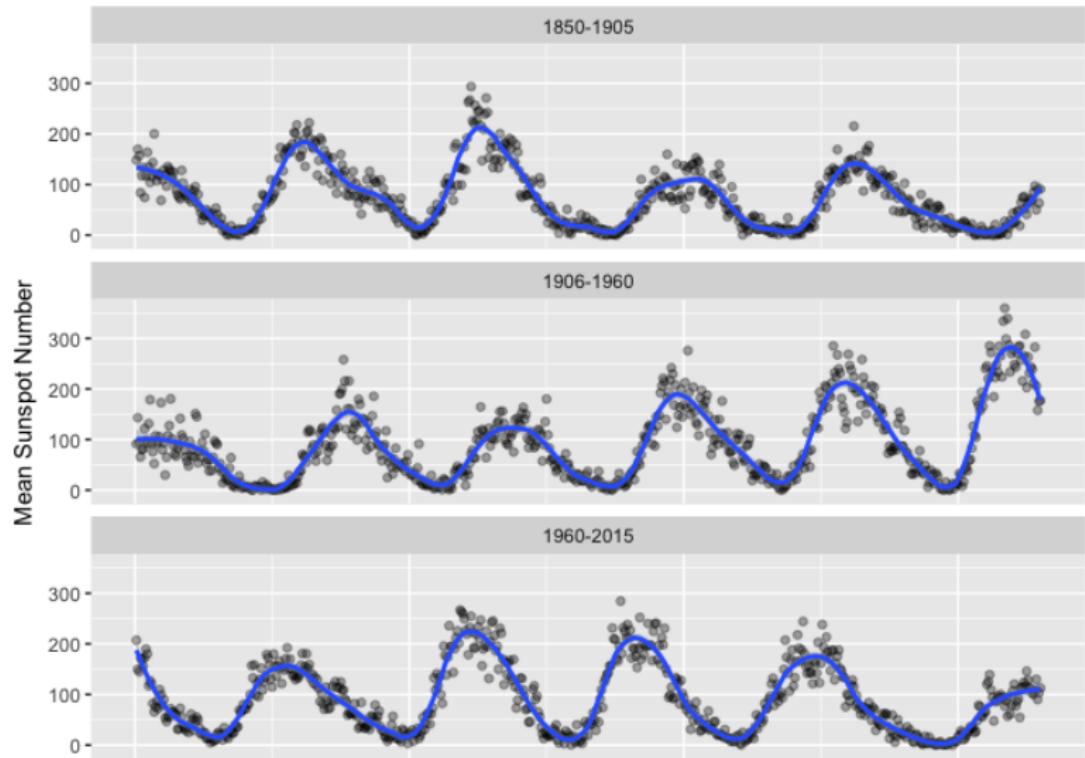
Observation

Juste autour de 45° montre les meilleures variations de pente.

Pente



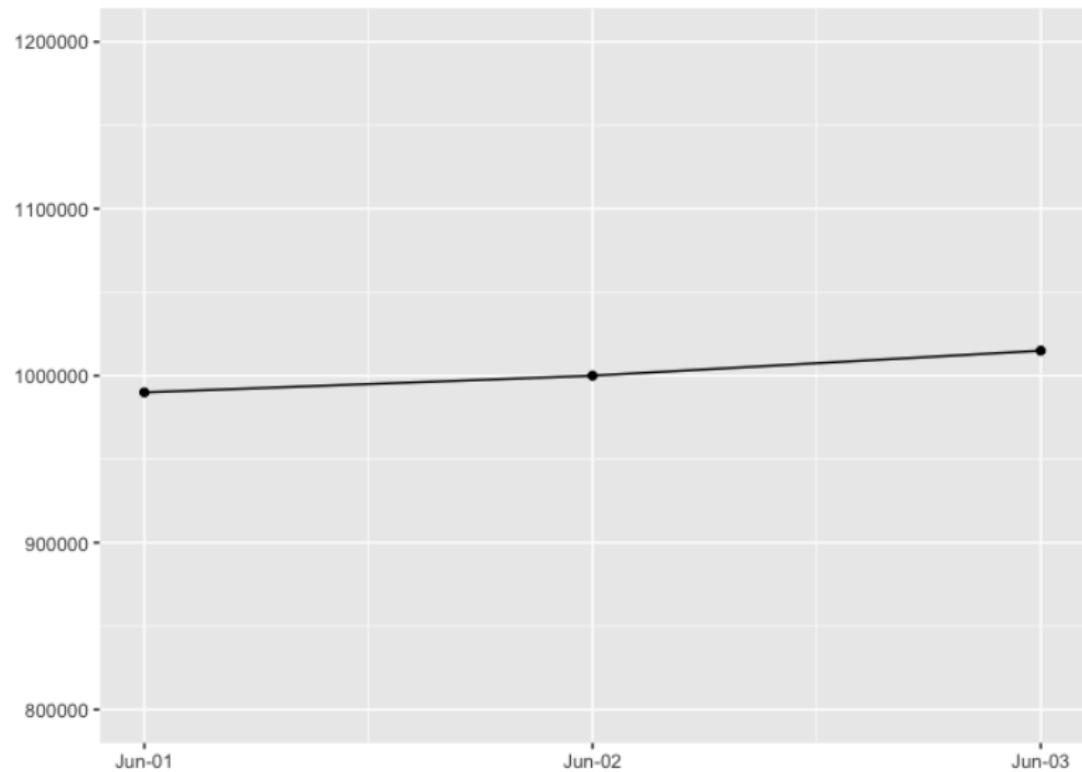
Pente



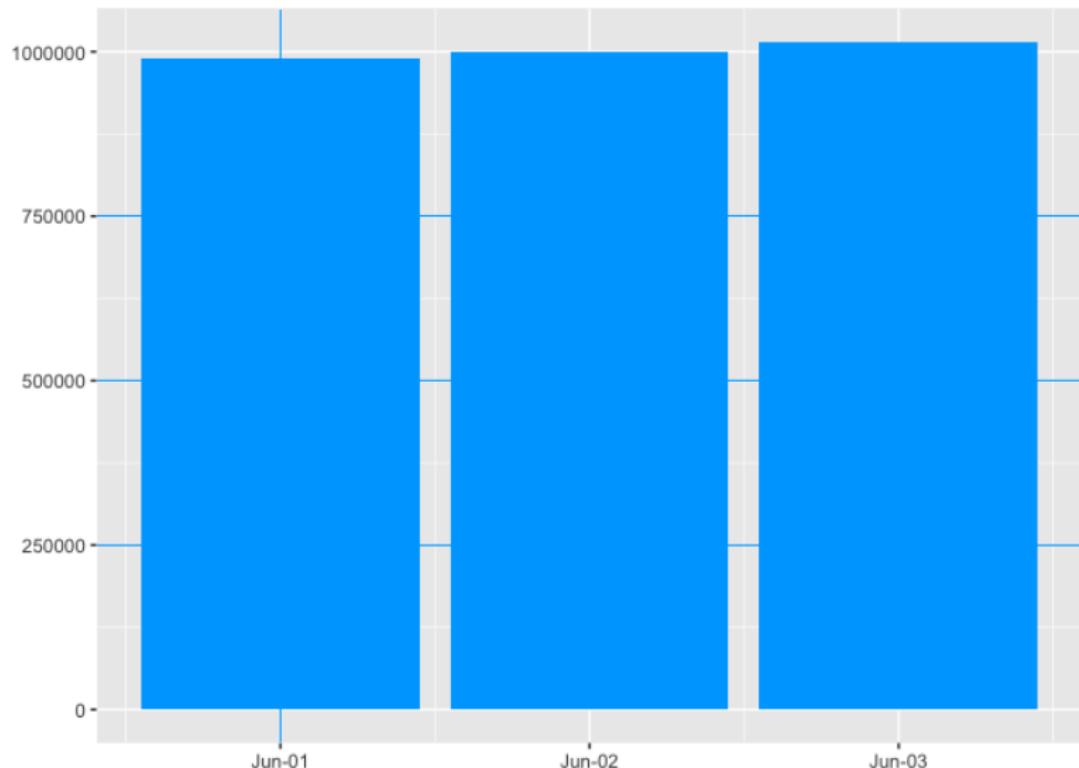
Question

Doit-on inclure 0 sur l'échelle ?

Zéro



Zéro



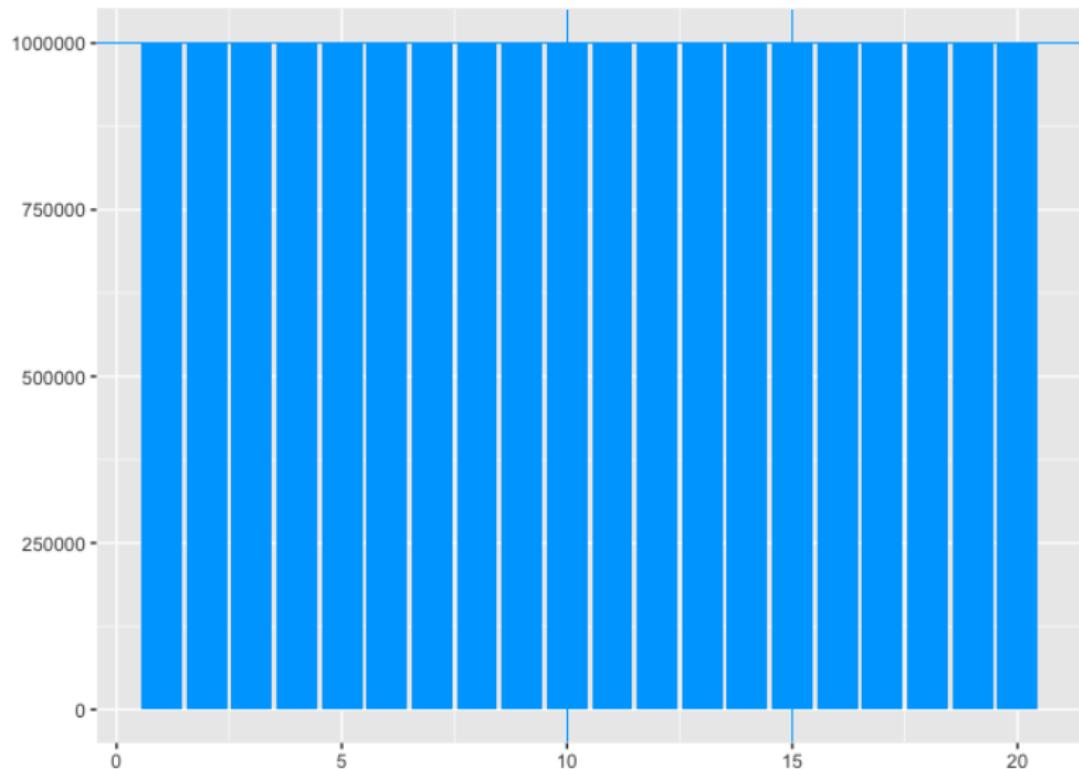
Question

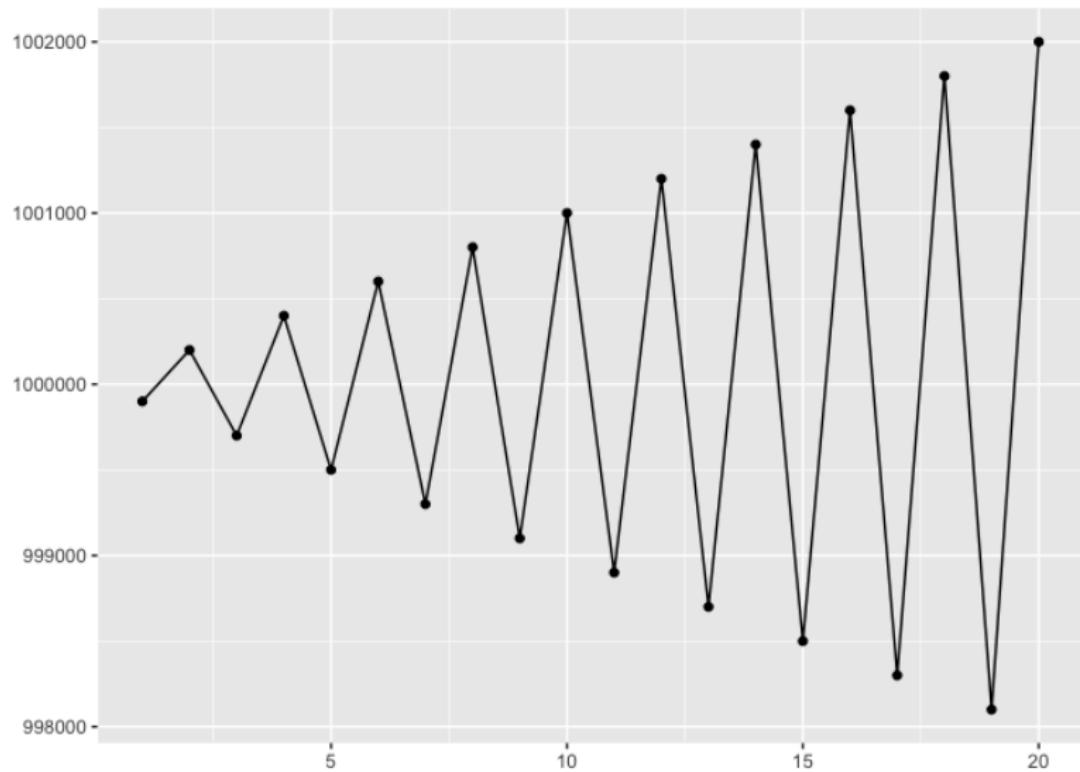
Doit-on inclure 0 sur l'échelle ? **Ca dépend.**

Doit-on inclure 0 sur l'échelle ?

- Se baser sur la perception pré-attentive de la taille ou de l'intensité ?
 - ▶ Oui, sinon vous biaiserez la perception.
- Et en utilisant la position ?
 - ▶ au choix ...

Zéro





"Avant tout, montrez les données."

- *Tufte*

"Avant tout, montrez **la variation** dans les données."

- *Tufte*

Addenda

La visualisation est de la communication

L'art
est de la
communication

La visualisation est de l'**art**







Pourquoi cela vous fait-il vous
sentir de cette manière ?

La **visualisation** a autant à apprendre de **l'art** que de la **science**.

Visualisation II : Exploration

Pourquoi ?

- "Connaître" les données
 - ▶ Métadonnées
 - ▶ Distributions
 - ▶ Domaine
- Préparer le nettoyage
- Préparer la définition des caractéristiques

Bases

Questions :

- Combien d'observations ai-je ?
 - ▶ Nombre d'exemples
- Combien de caractéristiques ?
 - ▶ Dimensionnalité
 - ★ Numériques?
 - ★ Catégoriques?
 - ▶ De quels types ?
- Ai-je une variable cible?

Exemples

Regarder des exemples d'observations :

	id_client	id_vehicle	id_policy	id_year	pol_bonus	pol_coverage	p
74906	A00068557	V01	A00068557-V01	Year 1	0.50		Maxi
90414	A00082734	V01	A00082734-V01	Year 1	0.50		Median1
26075	A00023900	V01	A00023900-V01	Year 1	0.50		Mini
40063	A00036703	V01	A00036703-V01	Year 1	0.50		Maxi
64546	A00059101	V01	A00059101-V01	Year 1	0.60		Median2
9522	A00008714	V02	A00008714-V02	Year 1	0.50		Maxi
41363	A00037884	V01	A00037884-V01	Year 1	0.50		Median1
55129	A00050502	V01	A00050502-V01	Year 0	0.50		Median2
96024	A00087861	V01	A00087861-V01	Year 0	0.57		Median2
50060	A00045842	V01	A00045842-V01	Year 1	0.50		Maxi

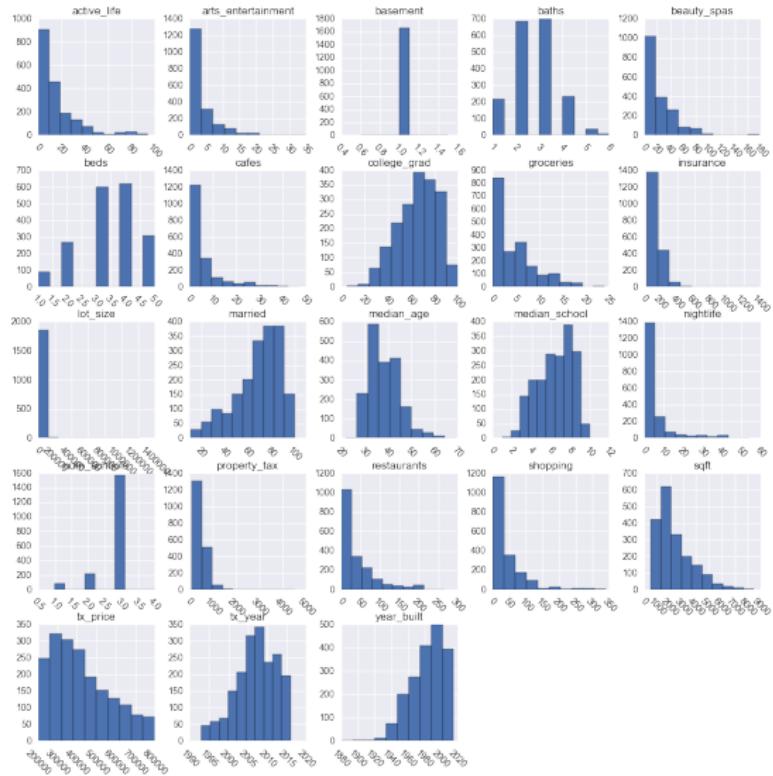
Exemples

Regarder des exemples d'observations pour avoir une "sensation" qualitative :

- Les colonnes ont-elles un sens?
- Les valeurs dans ces colonnes ont-elles un sens?
- Les valeurs sont-elles à la bonne échelle?
- Les données manquantes vont-elles poser un gros problème en se basant sur un test oculaire rapide?

Distributions univariées

Regarder les distributions des variables numériques :



Variables numériques :

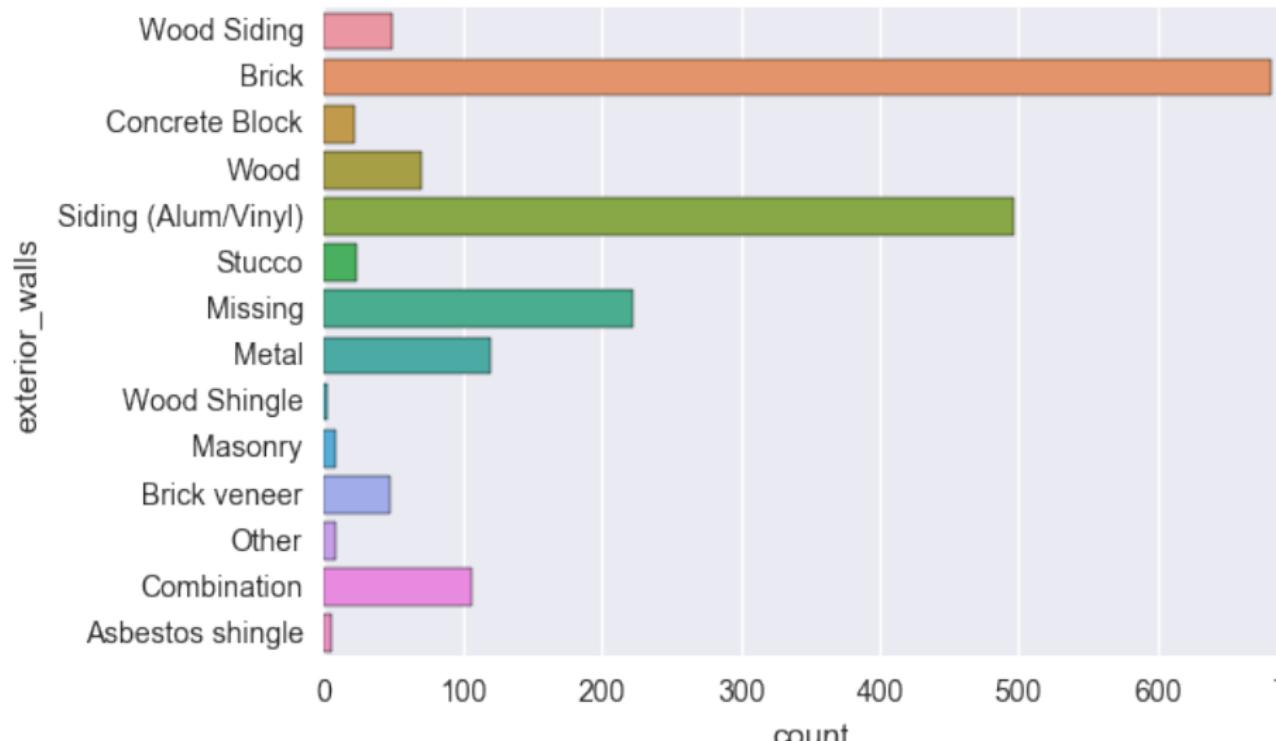
⇒ Histogrammes

Regarder les distributions :

- Des distributions inattendues
- Les valeurs aberrantes qui n'ont pas de sens
- Caractéristiques qui pourraient être binaires
- Des frontières qui n'ont pas de sens
- Erreurs de mesures potentielles

Distributions univariées

Regarder les distributions des variables catégorielles :



Variables catégorielles :

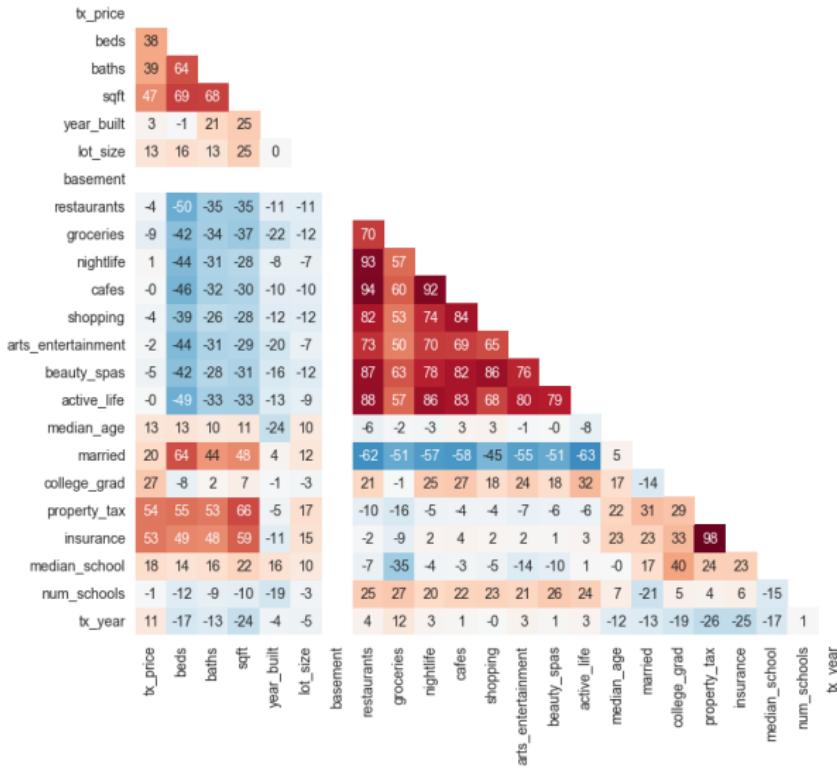
⇒ Diagramme en bâton (*barplot*)

Regarder les distributions :

- Voir le débancement de classes
- Encodage à adopter
- Regrouper les similaires

Distributions multivariées

Regarder les distributions numérique vs numérique :



Distributions multivariées

numérique vs numérique :

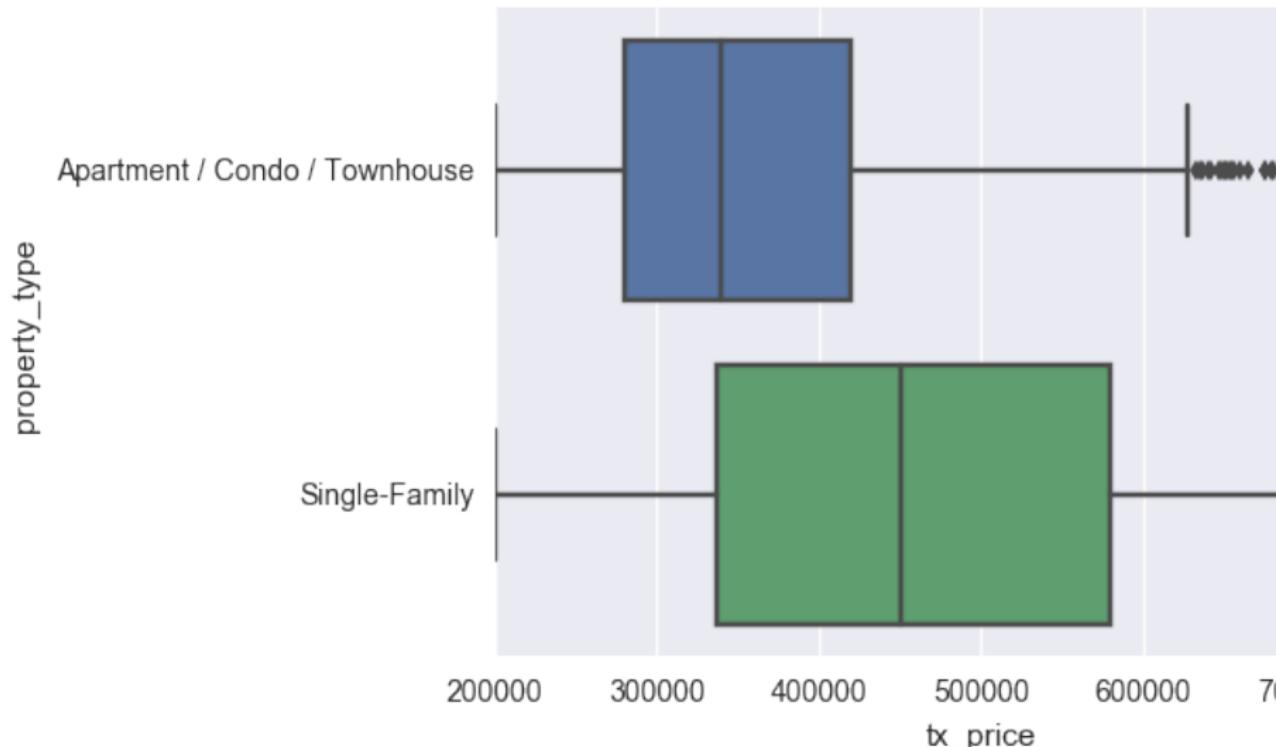
- ⇒ Carte de chaleur (*heatmap*)
- ⇒ Distributions de points (*pairplot*)

Regarder les corrélations :

- La corrélation positive signifie que lorsqu'une caractéristique augmente, l'autre augmente.
- La corrélation négative signifie que lorsque l'une des caractéristiques augmente, l'autre diminue.
- Des corrélations proches de -1 ou 1 indiquent une relation forte.
- Les plus proches de 0 indiquent une relation faible.
- 0 indique aucune relation.

Distributions multivariées

Regarder les distributions numérique vs catégorielle :



numérique vs catégorielle :

⇒ Boîte à moustaches (*boxplot*)

Comparer les distributions par catégorie :

- La médiane est la barre verticale du milieu
- Les deux principaux quartiles (du 25^e au 75^e percentile) sont dans la boîte (l'interquartile)
- Les moustaches couvrent $\pm 150\%$ de l'interquartile
- Les outliers sont à l'extérieur des moustaches

That's all folks !
Questions ?