

Introducción a la visualización de datos

Felipe González

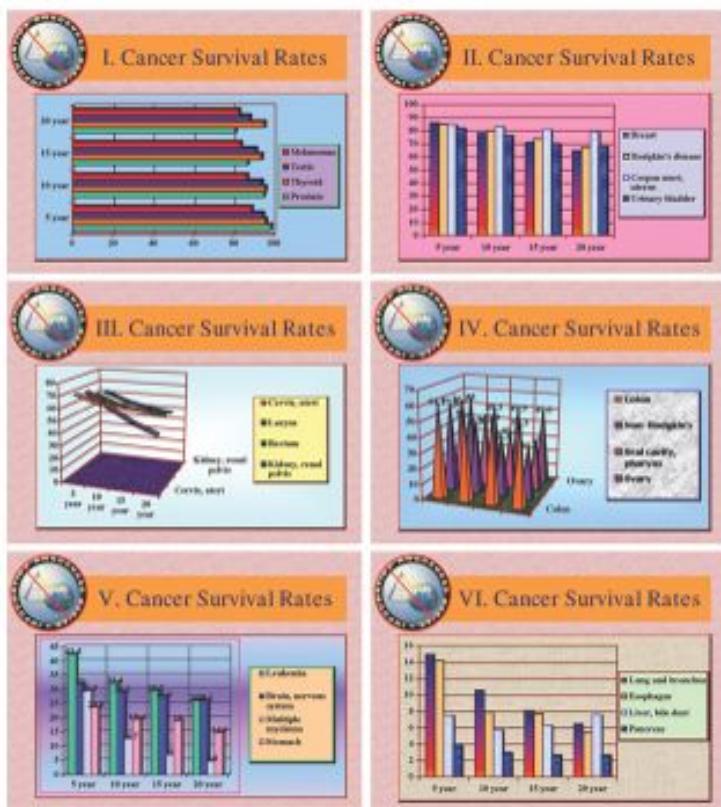
Introducción

La visualización de datos no trata de hacer gráficas "bonitas" o "divertidas", ni de simplificar lo complejo o ayudar a una persona "que no entiende mucho" a entender ideas complejas. Más bien, trata de aprovechar nuestra gran capacidad de procesamiento visual para exhibir de manera clara aspectos importantes de los datos.

El siguiente ejemplo de Edward Tufte (Beautiful Evidence, Graphics Press, 2008), ilustra claramente la diferencia entre estos dos enfoques. A la izquierda están gráficas (más o menos típicas de Powerpoint) basadas en la filosofía de simplificar, de intentar no "ahogar" al lector con datos. El resultado es una colección incoherente, de bajo contenido, que no tiene mucho qué decir y que es, "indefinible al contenido y la evidencia".

A la derecha está una variación del rediseño de Tufte en forma de tabla, que en este caso particular es una manera eficiente de mostrar claramente los patrones que hay en este conjunto simple de datos.

¿Qué principios son los que soportan la efectividad de esta tabla sobre la gráfica de la derecha? Veremos que hay dos conjuntos de principios importantes: unos relacionados con el diseño y otros con la naturaleza del análisis de datos, independientemente del método de visualización.



Estimates of relative survival rates and standard errors, by cancer site (% survival and standard error)

	5 year	10 year	15 year	20 year				
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

Rates derived from SEER 1973-98 databases (both sexes, all ethnic groups).

Visualización de datos en la estadística (las grafiqitas)

La estadística tradicionalmente se divide en dos partes: una parte de naturaleza exploratoria, donde jugamos el papel de detectives en búsqueda de los elementos de evidencia importante, y una parte de naturaleza inferencial, donde nos convertimos en jueces donde le damos pesos de credibilidad a la evidencia que presenta el detective. Estas dos partes tienen interacción fuerte en la práctica, pero por razones históricas se considera “superior” a la parte inferencial por encima de la exploratoria.

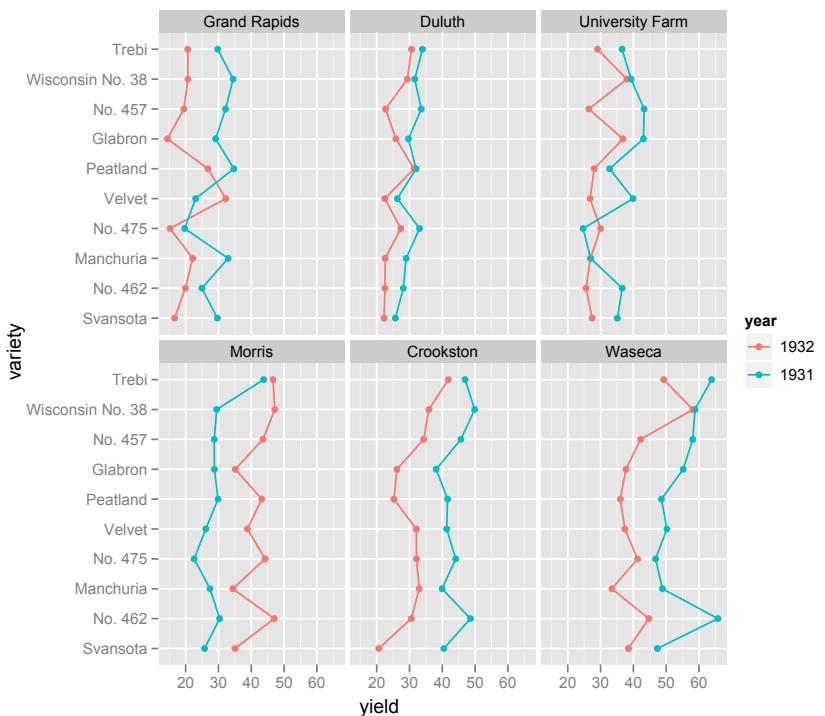
Aunque en el proceso de inferencia las gráficas cada vez son más importantes, la visualización entra más claramente dentro del análisis exploratorio de datos. Y como en un principio no es claro como la visualización aporta al proceso de la inferencia, se le consideró por mucho tiempo como un área de poca importancia para la estadística: una herramienta que en todo caso sirve para comunicar ideas simples, de manera deficiente, y a personas poco sofisticadas.

El peor lado de este punto de vista consiste en restringirse a el *análisis estadístico rutinario* (Visualizing Data, Cleveland): aplicar las recetas y negarse a ver los datos de distinta manera (¡incluso pensar que esto puede sesgar los resultados, o que nos podría engañar!). El siguiente ejemplo muestra un caso grave de este análisis estadístico rutinario (tomado de Visualizing Data, Cleveland).

A la derecha mostramos los resultados de un experimento de agricultura. Se cultivaron diez variedades de cebada en seis sitios de Minnesota, en 1921 y 1932. Este es uno de los primeros ejemplos en el que se aplicaron las ideas de Fisher en cuanto a diseño de experimentos.

Estos datos fueron reanalizados desde esa época por muchos agrónomos. Hasta muy recientemente se detectó la anomalía en el comportamiento de los años en el sitio Morris, el cual es evidente en la gráfica. Investigación posterior ha mostrado convincentemente que en algún momento alguien volteó las etiquetas de los años en este sitio.

Este ejemplo muestra, en primer lugar, que la visualización es crucial en el proceso de análisis de datos: sin ella estamos expuestos a no encontrar aspectos importantes de los datos (errores) que deben ser discutidos - aún cuando nuestra receta de análisis no considere estos aspectos. Ninguna receta puede aproximarse a describir todas las complejidades y detalles en un conjunto de datos de tamaño razonable (este ejemplo, en realidad, es chico). Sin embargo, la visualización de datos, por su enfoque menos estructurado, y el hecho de que se apoya en un medio con un “ancho de banda” mayor al que puede producir un cierto número de cantidades resumen, es ideal para investigar estos aspectos y detalles.



Visualización popular de datos

Publicaciones populares (periódicos, revistas, sitios internet) muchas veces incluyen visualización de datos como parte de sus artículos o reportajes. En general siguen el mismo patrón que en la visión tradicionalista de la estadística: sirven más para divertir que para explicar, tienden a explicar ideas simples y conjuntos chicos de datos, y se consideran como una “ayuda” para los “lectores menos sofisticados”. Casi siempre se trata de gráficas triviales (muchas veces con errores graves) que no aportan mucho a artículos que tienen un nivel de complejidad mucho mayor (es la filosofía: lo escrito para el adulto, lo graficado para el niño).

Ejemplo: gráficas, tablas y texto

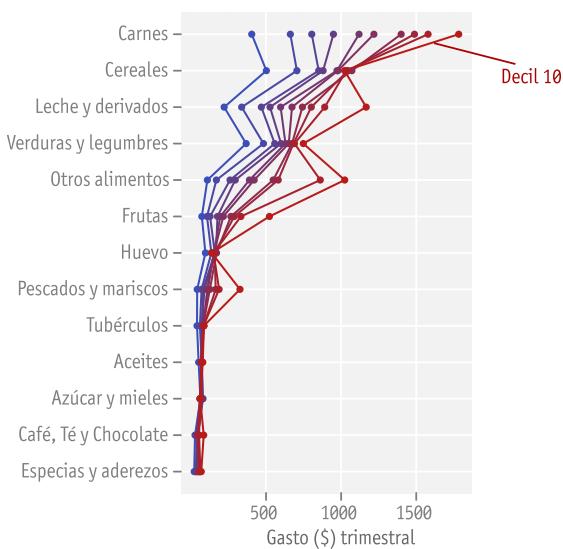
Gasto de hogares en Alimentos y Bebidas (miles de pesos), ENIGH 2006. Hogares agrupados en deciles según ingreso total monetario. Cada decil agrupa unos 2.65 millones de hogares.

	1	2	3	4	5	6	7	8	9	10	Total
Cereales	1330728	1869247	2254304	2331371	2576134	2593607	2839141	2770198	2740160	2710885	24015775
Carnes	1072718	1754012	2131706	2514365	2965671	3228132	3708675	3943535	4183472	4724145	30226431
Pescados Y Mariscos	110398	187546	213830	236001	286507	297299	333812	437266	496656	865432	3464747
Leche y Derivados	585910	895216	1242102	1395102	1582291	1783207	1966252	2123150	2360369	3091577	17025176
Huevo	255321	360471	421613	442603	405520	404737	451280	418855	398713	365472	3924585
Aceites Y Grasas	135823	190052	179945	183546	193544	197424	188956	180809	182252	208958	1841309
Tubérculos	107231	158078	190705	201664	229090	214818	214251	224368	221747	228002	1989954
Verduras, Legumbres	973984	1279986	1478179	1590063	1668224	1725576	1783611	1808792	1827177	1982693	16118285
Frutas	192462	283549	337608	468187	517938	571262	704867	765013	882037	1384251	6107174
Azúcar Y Mieles	167042	212941	200200	191048	202397	190093	157009	173545	164273	163299	1821847
Café, Té Y Chocolate	71945	120338	108609	97139	124502	128589	109801	126464	143134	225452	1255973
Especias Y Aderezos	57580	80636	91758	108561	116499	134123	155394	152145	167650	182256	1246602
Otros Alimentos	290038	448629	689605	781629	1031991	1115892	1451119	1540150	2282137	2713540	12344730
Total	5351180	7840701	9540164	1.1E+07	1.2E+07	1.3E+07	1.4E+07	1.5E+07	1.6E+07	1.7E+07	121382588

Esta tabla es difícil de leer, por varias razones: unidades, rejilla, renglones en desorden. Ningún intento de análisis acompaña a estas cifras.

Ordenar las categorías según gasto total (sobre todos los hogares) nos ayuda a entender estos datos con la gráfica de abajo. Adicionalmente, construimos una tabla con la proporción del gasto total por categoría según deciles de ingreso.

Gasto trimestral promedio por hogar, según deciles de ingreso



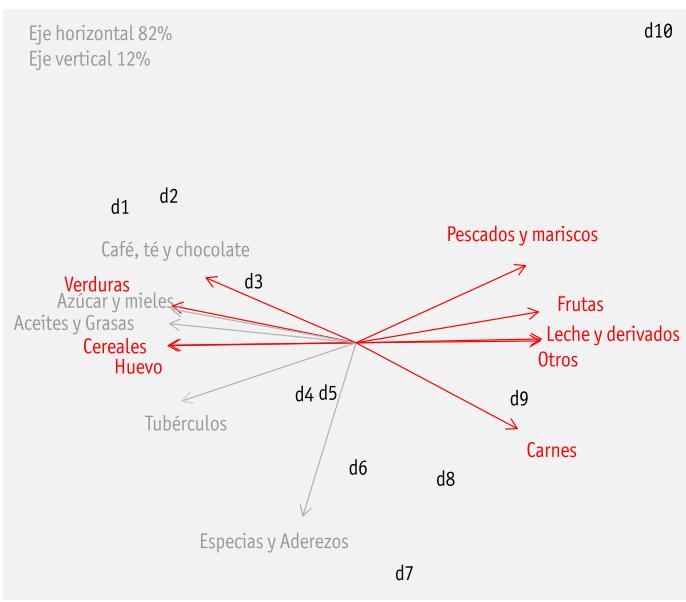
Porcentaje del gasto en cada categoría, por decil.

	1	2	3	4	5	6	7	8	9	10	Total %
Carnes	20.0	22.4	22.3	23.9	24.9	25.7	26.4	26.9	26.1	25.1	24.9
Cereales	24.9	23.8	23.6	22.1	21.6	20.6	20.2	18.9	17.1	14.4	19.8
Leche y Derivados	10.9	11.4	13.0	13.2	13.3	14.2	14.0	14.5	14.7	16.4	14.0
Verduras, Legumbres	18.2	16.3	15.5	15.1	14.0	13.7	12.7	12.3	11.4	10.5	13.3
Otros Alimentos	5.4	5.7	7.2	7.4	8.7	8.9	10.3	10.5	14.2	14.4	10.2
Frutas	3.6	3.6	3.5	4.4	4.4	4.5	5.0	5.2	5.5	7.3	5.0
Huevo	4.8	4.6	4.4	4.2	3.4	3.2	2.9	2.5	1.9	3.2	
Pescados Y Mariscos	2.1	2.4	2.2	2.2	2.4	2.4	3.0	3.1	4.6	2.9	
Tubérculos	2.0	2.0	2.0	1.9	1.9	1.7	1.5	1.5	1.4	1.2	
Aceites Y Grasas	2.5	2.4	1.9	1.7	1.6	1.6	1.3	1.2	1.1	1.1	
Azúcar Y Mieles	3.1	2.7	2.1	1.8	1.7	1.5	1.1	1.2	1.0	0.9	
Café, Té Y Chocolate	1.3	1.5	1.1	0.9	1.0	1.0	0.8	0.9	0.9	1.2	
Especias Y Aderezos	1.1	1.0	1.0	1.0	1.1	1.1	1.0	1.0	1.0	1.0	
Total (miles de millones)	5.4	7.8	9.5	10.5	11.9	12.6	14.1	14.7	16.0	18.8	

Diferencia relativa (%) con respecto al total

	1	2	3	4	5	6	7	8	9	10	Total %
Carnes	-19	-10	-10	-4	0	3	6	8	5	1	25
Cereales	26	20	19	12	9	4	2	-5	-14	-27	20
Leche y Derivados	-22	-19	-7	-6	-5	1	0	3	5	17	14
Verduras, Legumbres	37	23	17	14	6	3	-4	-7	-14	-21	13
Otros Alimentos	-47	-44	-29	-27	-15	-13	1	3	40	42	10
Frutas	-29	-28	-30	-12	-13	-10	0	4	9	46	5
Huevo	48	42	37	30	5	-1	-1	-12	-23	-40	3
Pescados Y Mariscos	-28	-16	-21	-22	-16	-17	-17	4	8	61	3
Tubérculos	22	23	22	17	17	4	-7	-7	-16	-26	2
Aceites Y Grasas	67	60	24	15	7	3	-11	-19	-25	-27	2
Azúcar Y Mieles	108	81	40	21	13	1	-26	-21	-32	-42	2
Café, Té Y Chocolate	30	48	10	-11	1	-1	-25	-17	-14	16	1
Especias Y Aderezos	5	0	-6	0	-5	4	8	1	2	-6	1
	1	2	3	4	5	6	7	8	9	10	Total %
Azúcar Y Mieles	108	81	40	21	13	1	-26	-21	-32	-42	2
Aceites Y Grasas	67	60	24	15	7	3	-11	-19	-25	-27	2
Huevo	48	42	37	30	5	-1	-1	-12	-23	-40	3
Verduras, Legumbres	37	23	17	14	6	3	-4	-7	-14	-21	13
Café, Té Y Chocolate	30	48	10	-11	1	-1	-25	-17	-14	16	1
Cereales	26	20	19	12	9	4	2	-5	-14	-27	20
Tubérculos	22	23	22	17	17	4	-7	-7	-16	-26	2
Especias Y Aderezos	5	0	-6	0	-5	4	8	1	2	-6	1
Carnes	-19	-10	-10	-4	0	3	6	8	5	1	25
Leche y Derivados	-22	-19	-7	-6	-5	1	0	3	5	17	14
Pescados Y Mariscos	-28	-16	-21	-22	-16	-17	-17	4	8	61	3
Frutas	-29	-28	-30	-12	-13	-10	0	4	9	46	5
Otros Alimentos	-47	-44	-29	-27	-15	-13	1	3	40	42	10
Total (miles de millones)	5.4	7.8	9.5	10.5	11.9	12.6	14.1	14.7	16.0	18.8	

Podemos entender las diferencias de gasto entre los deciles calculando que tanto se aparta cada decil del patrón total de gasto, como en la tabla de la izquierda. Esta última tabla funciona mucho mejor si ordenamos según las diferencias del decil de ingreso más bajo. Finalmente, construimos un biplot para reforzar el patrón más claro que observamos en estas últimas dos tablas.



Teoría de visualización de datos (Tufte, Cleveland, Tukey)

Existe teoría fundamentada acerca de la visualización. Después del trabajo pionero de Tukey, los principios e indicadores de Tufte se basan en un estudio de la historia de la graficación y ejercicios de muestreo de la práctica gráfica a lo largo de varias disciplinas (¿cuáles son las mejores gráficas? ¿por qué? El trabajo de Cleveland es orientado a la práctica del análisis de datos (¿cuáles gráficas nos han ayudado a mostrar claramente los resultados del análisis?), por una parte, y a algunos estudios de percepción visual.

Principios generales del diseño analítico. Aplicables a una presentación o análisis completos, y como guía para construir nuevas visualizaciones (E. Tufte, 06).

Principio 1 Muestra comparaciones, contrastes, diferencias.

Principio 2 Muestra causalidad, mecanismo, explicación, estructura sistemática

Principio 3 Muestra datos multivariados; es decir, más de una o dos variables

Principio 4 Integra palabras, números, imágenes y diagramas

Principio 5 Describe la totalidad de la evidencia. Muestra fuentes usadas y problemas relevantes

Principio 6 Las presentaciones analíticas, a fin de cuentas, se sostienen o caen dependiendo de la calidad, relevancia e integridad de su contenido.

Indicadores de calidad gráfica. Aplicables a cualquier gráfica en particular. Estas son guías concretas y relativamente objetivas para evaluar la calidad de una gráfica (E. Tufte, 86).

Integridad Gráfica. El factor de engaño, es decir, la distorsión gráfica de las cantidades representadas, debe ser mínimo.

Chartjunk. Minimizar el uso de decoración gráfica que interfiera con la interpretación de los datos: 3D, rejillas, rellenos con patrones

Tinta de datos. Maximizar la proporción de tinta de datos vs. tinta total de la gráfica. *For non-data-ink, less is more. For data-ink, less is a bore.*

Densidad de datos. Las mejores gráficas tienen mayor densidad de datos, que es la razón entre el tamaño del conjunto de datos y el área de la gráfica. Las gráficas se pueden encoger mucho.

Percepción visual. Algunas tareas son más fáciles para el ojo humano que otras (Cleveland, 94).

Técnicas de visualización. Esta categoría incluye técnicas específicas que dependen de la forma de nuestros datos y el tipo de pregunta que queremos investigar (Tukey, 77, Cleveland, 93-94, Tufte '06).

Tipos de gráficas: cuantiles, histogramas, caja y brazos, gráficas de dispersión, puntos/barras/líneas, series de tiempo.

Técnicas para mejorar gráficas: Transformación de datos, transparencia, vibración, banking 45, suavizado y bandas de confianza.

Pequeños múltiplos

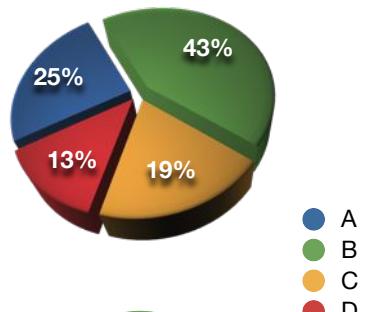
Tablas

Factor de engaño, chartjunk y pies

El **factor de engaño** es el cociente entre el efecto mostrado en una gráfica y el efecto correspondiente en los datos. Idealmente, el factor de engaño debe ser 1 (ninguna distorsión)

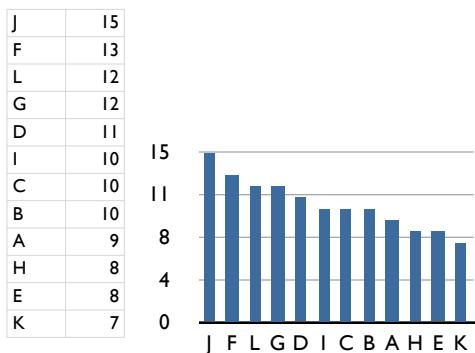
El **chartjunk** son aquellos elementos gráficos que no corresponden a variación de datos, o que entorpecen la interpretación de una gráfica.

Estos son los indicadores de calidad más fáciles de entender y aplicar, y afortunadamente cada vez son menos comunes.

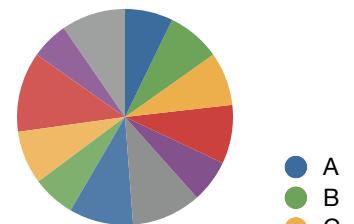
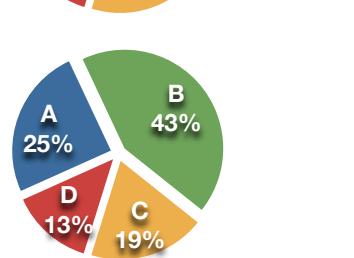


Un diseño popular que califica como chartjunk y además introduce factores de engaño es el pie de 3D. En la gráfica de la derecha, podemos ver como la rebanada C se ve más grande que la rebanada A, aunque claramente ese no es el caso (factor de engaño). La razón es la variación en la perspectiva que no corresponde a variación en los datos (chartjunk).

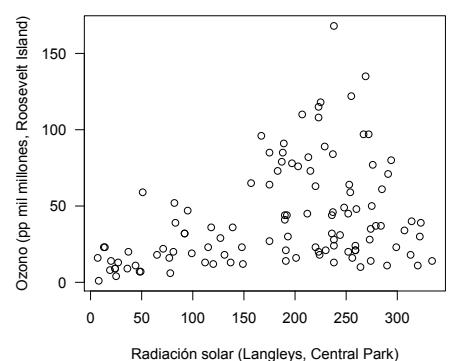
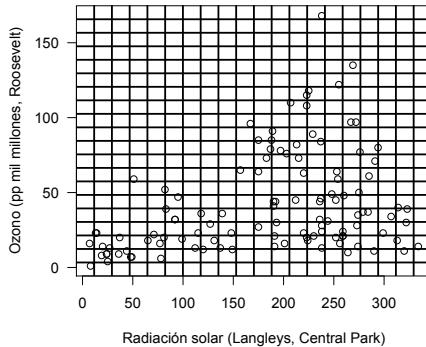
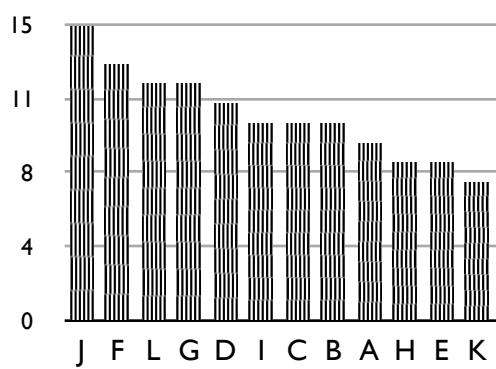
Corregimos quitando el efecto 3D. Esto reduce el factor de engaño pero hay todavía elementos que pueden mejorar la comprensión: se trata de la decodificación que hay que hacer categoría - color - cuantificación. Podemos agregar las etiquetas como se muestra en la serie de la derecha, pero entonces: ¿por qué no mostrar simplemente la tabla de datos? ¿qué agrega el pie a la interpretación?



La deficiencias en el pie se pueden ver claramente al intentar graficar más categorías (13). En el primer pie no podemos distinguir realmente cuáles son las categorías grandes y cuáles las chicas, y es muy difícil tener una imagen mental clara de estos datos. Agregar los porcentajes ayuda, pero entonces, otra vez, preguntamos cuál es el propósito del pie. La tabla de la derecha hace todo el trabajo (una vez que ordenamos las categorías de la más grande a la más chica). Es posible hacer una gráfica de barras como la de abajo a la izquierda que hace



Hay otros tipos de chartjunk comunes: uno es la textura de barras, por ejemplo. El efecto es la producción de un efecto moiré que es desagradable y quita la atención de los datos, como en la gráfica de barras de abajo. Otro común son las rejillas, como mostramos en las gráficas de la izquierda. Nótese como en estos casos hay efectos ópticos no planeados que degradan la percepción de los patrones en los datos.



Técnicas de visualización: descripción de variabilidad

Un concepto fundamental en la estadística es el de **bonche o conjunto de datos** (**conjunto no estructurado de datos**). Un bonche de datos numéricos se ve como sigue:
12.5, 32.32, 1.20, 13.4, 20.2, 4.20, 10.7

Mientras que un bonche de datos categóricos se ve como sigue:

rojo, azul, azul, verde, rojo, azul, amarillo

Estos datos no tienen estructura de orden, no están asociados a ninguna otra medición o etiqueta, ni tienen que ser resultado de un proceso de muestreo.

Los bonches de datos interesantes presentan variación: en el primer ejemplo, no todas las mediciones son iguales, en el segundo, no todos los colores son iguales. Y lo interesante es precisamente medir en qué consiste esa variabilidad, cómo es esa variabilidad. Entender y describir esa variabilidad es probablemente la tarea más fundamental de la estadística. Todas las preguntas y las respuestas estadísticas se expresan, a final de cuentas, en términos de variabilidad.

Los dos ejemplos de arriba fueron escogidos porque representan dos tareas claramente distintas: describir variabilidad para bonches de datos numéricos, y para bonches de datos categóricos.

La ejemplos de conjuntos de datos de arriba son fáciles de entender. Con conjuntos de datos más grandes, la inspección visual es insuficiente. Cantidades populares para resumir estos datos son la media y la desviación estándar:

Para los datos que se muestran a la derecha, podemos calcular:

	Plaza A	Plaza B
Media	11.05%	8.33%
Desviación Est	3.4%	2.5%

Donde vemos que la media de la plaza A y la media de la plaza B son considerablemente distintas, y también la Plaza A presenta dispersión mas grande que B. La media parece que se interpreta fácilmente. La desviación estándar es un poco más difícil.

Esto no es análisis exploratorio -es análisis estadístico condensado que depende de muchos supuestos, y que posiblemente cae en lo rutinario-. Nos quedamos cortos en nuestro propósito de describir la variación de los datos, y todavía estamos a oscuras en cuanto a nuestros datos. En vez de dar unos cuantos resúmenes numéricos de estos dos bonches de datos, intentamos primero algunas gráficas simples.

Proporción de Ingreso total (ventas nacionales) debidas al producto P en dos plazas A y B, respectivamente (mediciones en 199 semanas)

	Plaza A				Plaza B			
0.094	0.107	0.065	0.114	0.087	0.099	0.059	0.074	0.074
0.045	0.121	0.082	0.076	0.055	0.096	0.081	0.075	0.075
0.075	0.108	0.098	0.071	0.045	0.103	0.100	0.077	0.077
0.060	0.113	0.088	0.129	0.051	0.092	0.079	0.048	0.048
0.122	0.113	0.079	0.111	0.089	0.109	0.079	0.080	0.080
0.116	0.147	0.203	0.119	0.062	0.133	0.141	0.071	0.071
0.143	0.147	0.164	0.167	0.111	0.141	0.113	0.079	0.079
0.098	0.089	0.089	0.109	0.077	0.067	0.061	0.075	0.075
0.071	0.081	0.1	0.101	0.044	0.074	0.109	0.093	0.093
0.105	0.093	0.031	0.07	0.067	0.085	0.135	0.112	0.112
0.092	0.158	0.058	0.104	0.067	0.110	0.067	0.106	0.106
0.066	0.124	0.127	0.115	0.056	0.087	0.092	0.090	0.090
0.114	0.147	0.088	0.162	0.078	0.105	0.062	0.098	0.098
0.126	0.13	0.127	0.084	0.092	0.119	0.085	0.069	0.069
0.119	0.084	0.121	0.148	0.087	0.070	0.096	0.086	0.086
0.068	0.082	0.101	0.112	0.050	0.061	0.080	0.071	0.071
0.042	0.114	0.1	0.062	0.037	0.078	0.072	0.083	0.083
0.112	0.141	0.098	0.105	0.073	0.078	0.098	0.091	0.091
0.121	0.138	0.053	0.115	0.084	0.100	0.047	0.063	0.063
0.103	0.119	0.086	0.096	0.082	0.090	0.079	0.113	0.113
0.121	0.084	0.13	0.124	0.079	0.081	0.106	0.067	0.067
0.129	0.061	0.106	0.056	0.092	0.057	0.077	0.053	0.053
0.065	0.057	0.141	0.101	0.061	0.050	0.108	0.062	0.062
0.036	0.145	0.149	0.141	0.036	0.122	0.125	0.058	0.058
0.086	0.153	0.089	0.194	0.089	0.083	0.070	0.064	0.064
0.054	0.085	0.121	0.125	0.045	0.054	0.110	0.089	0.089
0.027	0.089	0.107	0.117	0.018	0.076	0.087	0.116	0.116
0.136	0.082	0.123	0.147	0.107	0.061	0.098	0.110	0.110
0.086	0.152	0.142	0.131	0.044	0.107	0.081	0.080	0.080
0.035	0.11	0.142	0.102	0.047	0.091	0.103	0.101	0.101
0.053	0.135	0.157	0.069	0.045	0.089	0.132	0.067	0.067
0.044	0.135	0.13	0.112	0.023	0.121	0.132	0.056	0.056
0.077	0.09	0.092	0.108	0.043	0.081	0.063	0.064	0.064
0.071	0.149	0.098	0.181	0.055	0.133	0.074	0.081	0.081
0.031	0.148	0.102	0.113	0.027	0.109	0.123	0.084	0.084
0.017	0.153	NA	0.157	0.014	0.122	0.086	0.100	0.100
0.084	0.134	0.093	0.094	0.075	0.086	0.082	0.080	0.080
0.113	0.155	0.115	0.134	0.096	0.105	0.079	0.068	0.068
0.155	0.141	0.135	0.16	0.111	0.113	0.119	0.067	0.067
0.088	0.163	0.105	0.125	0.086	0.104	0.101	0.134	0.134
0.057	0.122	0.185	0.126	0.046	0.072	0.132	0.109	0.109
0.123	0.07	0.135	0.128	0.094	0.052	0.120	0.112	0.112
0.066	0.077	0.1	0.105	0.062	0.049	0.096	0.099	0.099
0.137	0.087	0.118	0.106	0.100	0.061	0.099	0.095	0.095
0.132	0.083	0.09	0.126	0.107	0.046	0.083	0.093	0.093
0.083	NA	0.141	0.132	0.071	0.065	0.086	0.098	0.098
0.088	NA	0.139	0.107	0.080	0.083	0.093	0.065	0.065
0.129	NA	0.1	0.021	0.073	0.072	0.089	0.107	0.107
0.135	0.055	0.09		0.114	0.068	0.082		

Los cuantiles y su gráfica

Comenzamos con los datos de la plaza A. Los ordenamos del más chico al más grande y los graficamos según su orden. Obtenemos la gráfica de la derecha. Cada punto representa exactamente uno de los datos de la tabla de arriba. Esta es una versión de la gráfica de cuantiles de las participaciones para la plaza A. En esta gráfica vemos rápidamente, por ejemplo:

- 1) el máximo (alrededor de 0.20) y el mínimo (alrededor de 0.02)
- 2) la mediana (el valor intermedio), ligeramente arriba de 0.10
- 3) el grueso de los datos están entre 0.05 y 0.17
- 4) la mitad de los datos está entre 0.08 y 0.130, aproximadamente
- 5) existen cuatro datos (los más grandes) que están relativamente despegados del resto (igual que unos 8 datos por abajo de 0.05).
- 6) no existen valores repetidos, no parece haber huecos en los datos (que se verían como saltos en la gráfica).

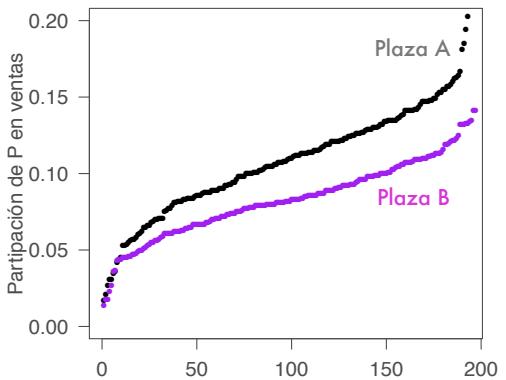
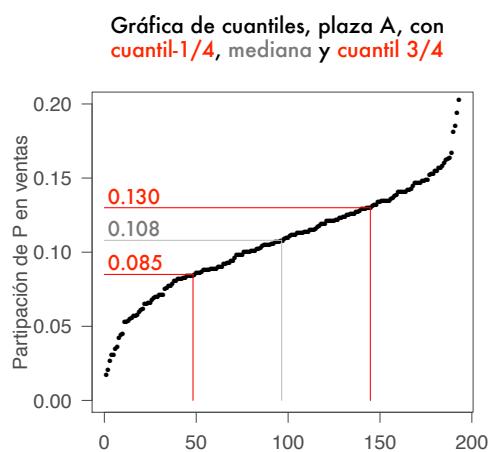
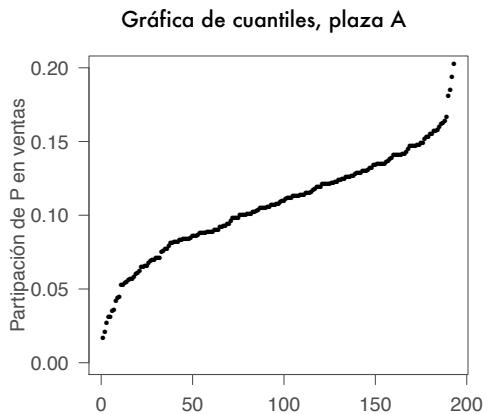
Los cuantiles se definen como sigue: Sea p una proporción. El cuantil- p de un conjunto de datos es el valor tal que aproximadamente una proporción p de los datos cae por debajo de x : por ejemplo, el cuantil 0.25 de las participaciones para la plaza A es 0.085, pues aproximadamente 25% de los datos están por debajo de 0.085.

Cuantiles útiles: cuartil inferior (0.25), mediana (0.5), cuartil superior (0.75).

Interpretación avanzada: ¿cómo se interpreta la distinta inclinación de la gráfica de cuantiles? La pendiente dice cuánto avanzamos en la escala (eje vertical) cada vez que nos movemos en el orden de los datos (eje horizontal). Cuando los datos están muy separados, avanzamos más hacia arriba cada vez que avanzamos un dato. Cuando los datos están muy pegados, avanzamos menos. La pendiente de la gráfica de cuantiles nos dice entonces qué tan densamente se acumulan los datos en cada parte de la escala, o también, qué tanta dispersión hay en cada parte de la escala.

En nuestro ejemplo, hay menos densidad de datos cerca de los extremos máximo y mínimo. Alrededor de la mediana los datos se acumulan más densamente.

Comparamos ahora los datos de la plaza A con los de la plaza B. Podemos ponerlos en la misma gráfica. Podríamos repetir el mismo análisis que hicimos arriba para la plaza B. Lo más interesante aquí es comparar las dos formas de las gráficas: notemos que los conjuntos de datos, en su rango inferior son similares. Sin embargo, conforme vamos avanzando en la escala, vemos que los datos de la plaza B tienden a dispersarse menos que los de la plaza A (la inclinación es menor): los datos de la plaza B tienen menos dispersión que los de la plaza A. Adicionalmente, no observamos datos grandes separados del resto como en la plaza A. La pendiente general de la gráfica indica la dispersión que hay en los datos: mayor pendiente=mayor dispersión.



¿Por qué cuantiles y gráficas en lugar de medias y desviación estándar?

Cuando usamos medias y desviación estándar debemos ser muy cuidadosos. En primer lugar, son sensibles a **valores atípicos** (¿cuál es la media de 1,2,3,4,5,1000? ¿en algún sentido esa media representa "el centro de los datos"? ¿qué significa la media aquí?), y esto las hace difíciles de interpretar cuando hay valores atípicos.

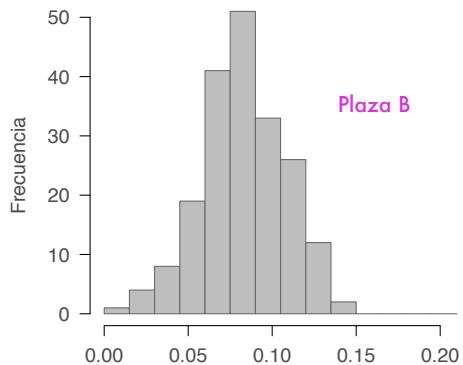
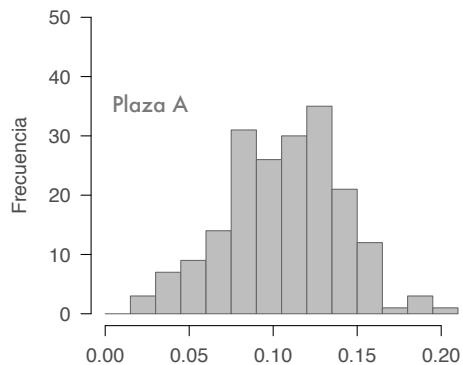
Pero incluso para datos bien portados son cantidades difíciles de interpretar, en particular la desviación estándar. Comparamos las siguientes dos descripciones simples (a la derecha) para los conjuntos de datos de arriba. ¿Cuál es más fácil de interpretar?

	Plaza A	Plaza B
Media	11.05%	8.33%
Desviación Est	3.4%	2.5%

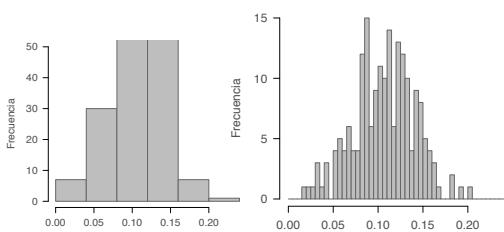
	Plaza A	Plaza B
Mín	1.7%	1.4%
Cuartil inferior	8.5%	6.7%
Mediana	10.8%	8.2%
Cuartil superior	13.0%	10.0%
Máx	20.3%	14.1%

Histogramas

Una alternativa popular a la gráfica de cuantiles es el histograma. Esta gráfica requiere más preparación: primero partimos el rango de los datos en intervalos de aproximadamente el mismo tamaño. En nuestro ejemplo, podemos partir el rango de 0.00 a 0.22 en unos 15 intervalos de longitud 0.015. Contamos cuántos datos caen en cada uno de los intervalos y hacemos una gráfica de barras donde la altura de cada barra cuenta cuántos datos cayeron en el intervalo correspondiente.

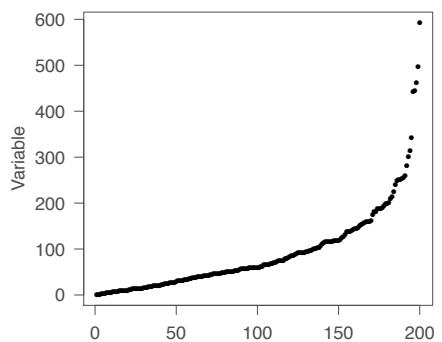


Los histogramas no deben hacerse con intervalos demasiado angostos ni demasiado anchos: hay que hacer un balance entre señal (patrones, forma de los datos) y ruido (variación a escalas muy chicas que no es interesante).



Asimetría y la gráfica de cuantiles

Los dos conjuntos de datos mostrados arriba son razonablemente simétricos alrededor de su valor central (por ejemplo, media o mediana). ¿Cómo se ven conjuntos de datos asimétricos? Ejercicio: ¿cómo se ve el histograma de este conjunto de datos



Gráficas cuantil-cuantil

Cuando queremos comparar directamente dos distribuciones de datos, conviene usar gráficas cuantil-cuantil. Si los dos conjuntos de datos son del mismo tamaño (digamos con n casos cada uno), podemos ordenar cada uno de más chico a más grande, y luego graficar los pares ordenados. Por ejemplo, si

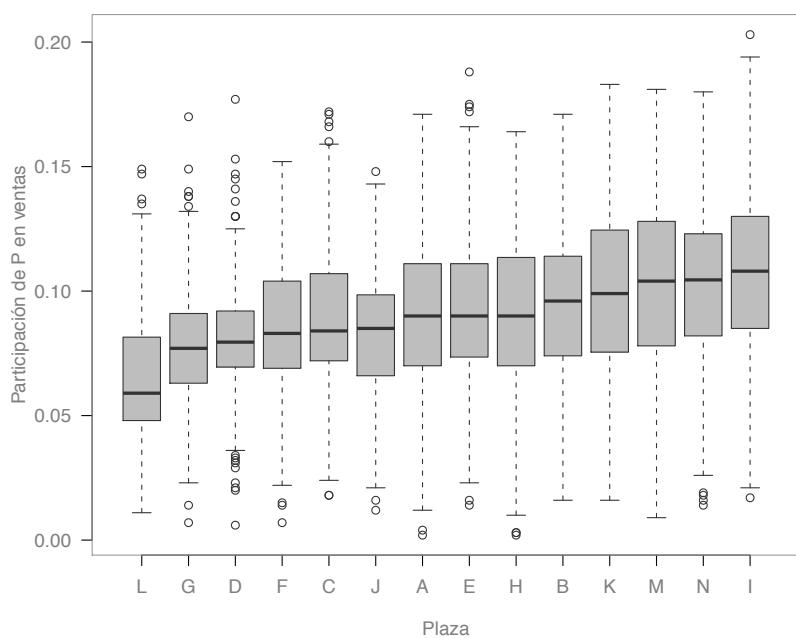
Grupo 1: 1,1,1,2,4,10

Grupo 2: 5,6,7,7,8,9

Entonces graficamos los puntos (1,5), (1,6), (1,7), (2,7), etcétera.

Técnicas de visualización: varios conjuntos de datos numéricos

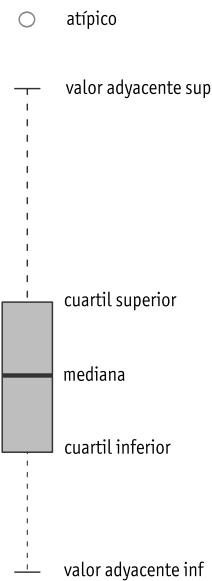
Las herramientas mostradas arriba son buenas para examinar con detalle la variabilidad de uno o dos conjuntos de datos. Cuando tenemos varios conjuntos de datos, conviene usar otro tipo de herramientas. Supongamos que tenemos datos similares a los del ejemplo anterior, pero ahora para 14 plazas distintas. Podemos graficar mediana y cuantiles en una misma gráfica para comparar todas las plazas fácilmente (ver explicación a la derecha), replicando el diseño de la derecha (**caja y brazos**) sobre cada una de las plazas.



El análisis de estos datos, usando la gráfica de caja y brazos, podría proceder como sigue:

- 1) La plaza con valores más chicos de participación de P es L, con una mediana alrededor de 0.6. Para la Plaza I tenemos una mediana de 0.10 (variación de más de 50% de un valor a otro).
- 2) La diferencia entre cuartiles superiores e inferiores va desde alrededor de 0.025 a 0.05 (dispersión).
- 3) Nótese también que hay traslape considerable en los valores que se toman en cada conjunto de datos: por ejemplo, el 25% de los datos más chicos para la plaza I son más chicos que el 25% de los datos más grandes para la plaza L.
- 4) La plaza D, en particular, muestra un comportamiento distinto a las demás: en general, varía poco (cuartiles cercanos, 50% de los datos), pero presenta, de manera más frecuente que las demás, variaciones no consistentes con esta variación pequeña del 50% de los datos.

La mediana es la línea negra. La caja está delimitada por el cuartil superior y el cuartil inferior. La línea punteada superior es 1.5 veces el tamaño de la caja. La línea punteada inferior llega hasta el valor mínimo de los datos. El criterio se escoge de manera que las líneas punteadas sean lo más cortas posible.



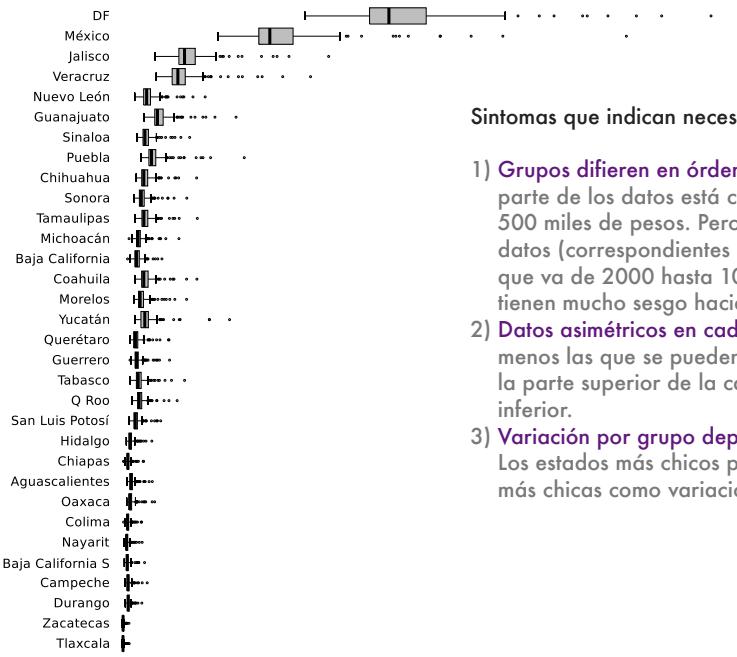
Esta es una aplicación del **principio de los pequeños múltiplos**: repetir una gráfica simple aplicada a distintos grupos de datos para comparar con efectividad. Funciona cuando cada pequeño múltiplo tiene el mismo diseño que cualquier otro, y diferencias en los múltiplos corresponden solamente a diferencias en los datos.

Ideas más avanzadas de variación: la plaza D presenta **platikurtosis** comparada con el resto de las plazas. Por ejemplo, ¿cómo es la variación para la plaza F comparada con la D?
- plaza F: variaciones moderadas regulares
- plaza D: regularmente variación chica, pero de vez en cuando variación fuerte.

Este es un concepto importante en análisis de riesgo: ¿cuál de los dos tipos de variación es preferible?

Técnicas de visualización: transformación de datos

Es común encontrar sesgos fuertes hacia la derecha cuando analizamos datos como número de personas, unidades vendidas, valor por unidad, etc. Generalmente, este sesgo fuerte tiene tres síntomas claros: los datos se concentran en valores chicos, pero hay variación considerable hacia arriba en un grupo considerable de casos.

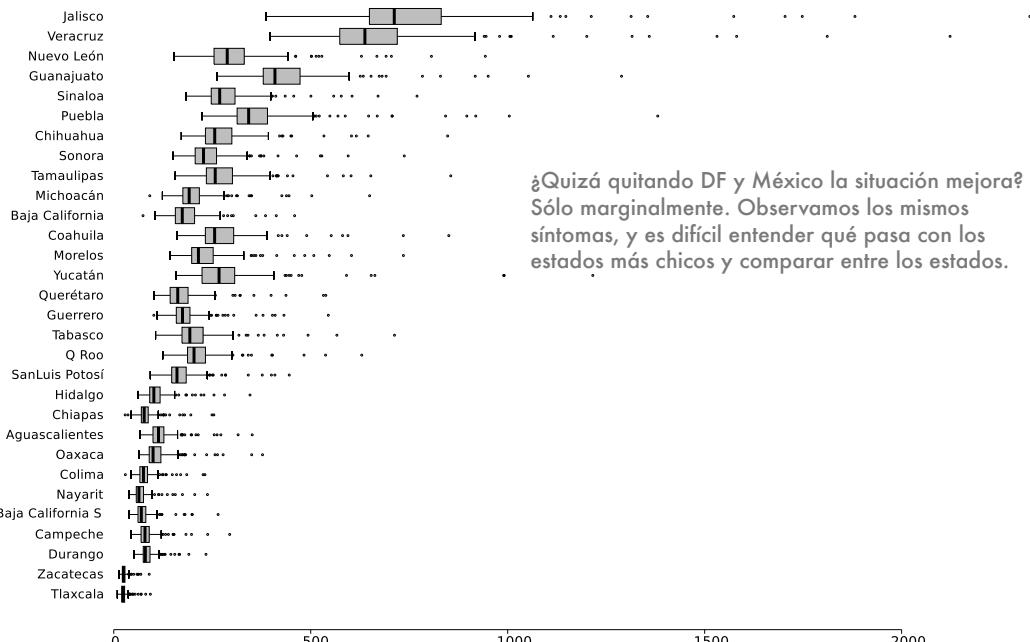


Síntomas que indican necesidad de transformar:

- 1) **Grupos difieren en órdenes de magnitud.** La mayor parte de los datos está concentrado de 0 a unos 500 miles de pesos. Pero al menos 1/32 de los datos (correspondientes al DF) varían en un rango que va de 2000 hasta 10000 miles. Estos datos tienen mucho sesgo hacia la derecha.
- 2) **Datos asimétricos en cada grupo.** Las cajas (por lo menos las que se pueden ver) son asimétricas, con la parte superior de la caja más larga que la parte inferior.
- 3) **Variación por grupo depende del nivel del grupo.** Los estados más chicos presentan tanto medianas más chicas como variación más chica (estados

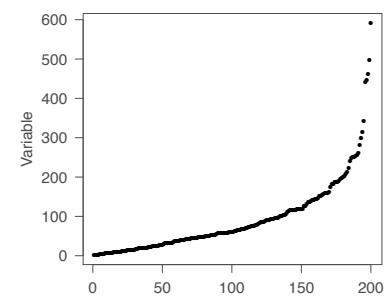
Problemas en la graficación:

- 1) Casi todo el espacio en la gráfica es utilizado para representar las unidades más grandes - quizás incluso algunos cuantos atípicos.
- 2) Difícil describir la variación en este tipo de datos (asimetría)

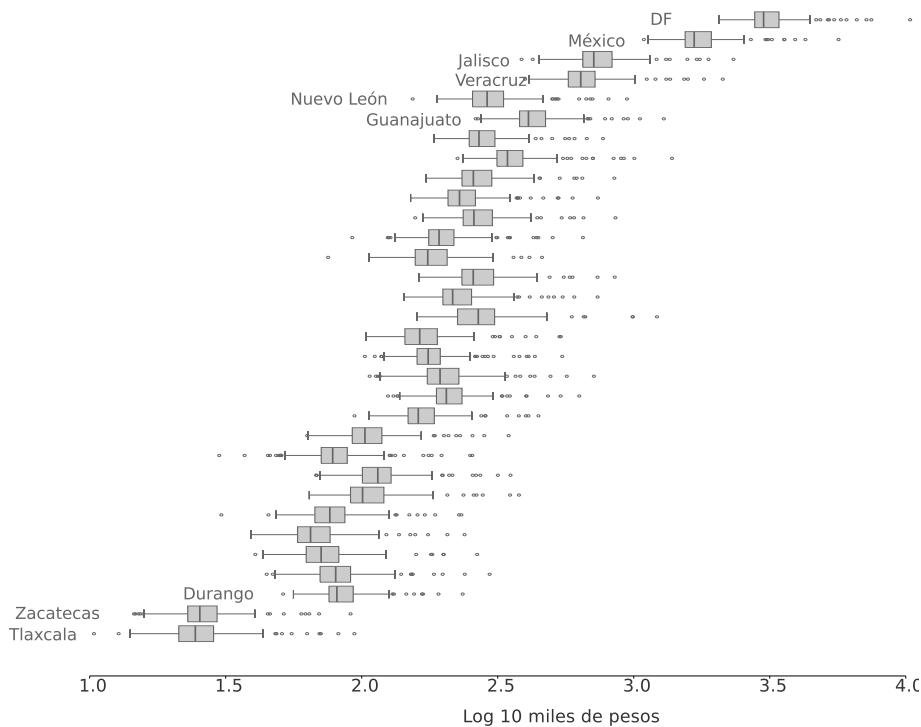


¿Quizá quitando DF y México la situación mejora?
Sólo marginalmente. Observamos los mismos síntomas, y es difícil entender qué pasa con los estados más chicos y comparar entre los estados.

Conjunto de datos sesgado hacia la derecha: gráfica de cuantiles



La solución es simple: transformar los datos usando el logaritmo (recordar la gráfica del logaritmo base 10, y algunas de sus propiedades). A la derecha mostramos la gráfica del logaritmo base 10: el logaritmo base 10 de 10 es 1, de 100 es 2, de 1000 es 3 y así sucesivamente. Obtenemos la siguiente gráfica de datos transformados:



Para analizar exitosamente una gráfica de este tipo, debemos cambiar un poco nuestro punto de vista: hay que considerar factores multiplicativos para entender la variación.

Por ejemplo:

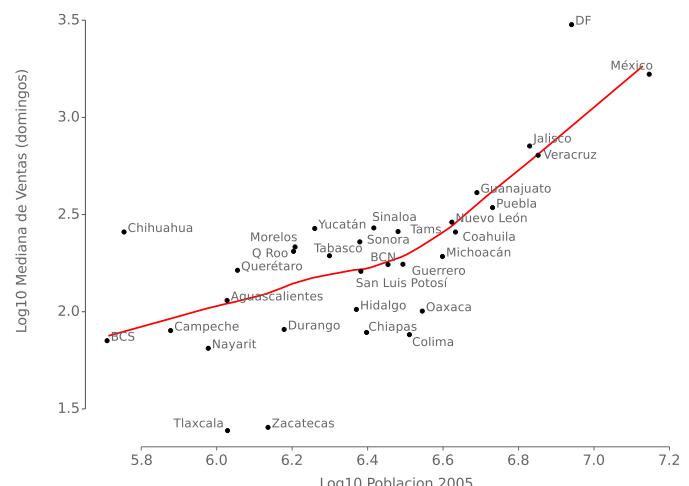
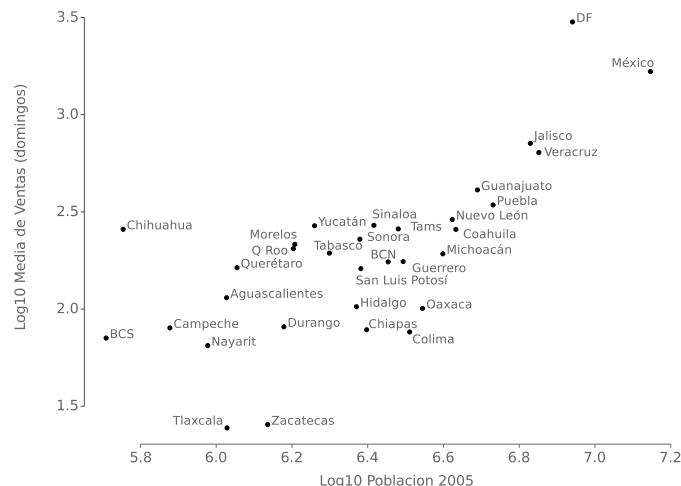
1) La diferencia entre los estados más grandes y los más chicos es alrededor de $3.5 - 1.3 = 2.2$. Como $10^{2.2} \approx 160$ aprox, eso quiere decir que el DF es, aproximadamente, 160 veces más grande que Tlaxcala o Zacatecas (considerando medianas).

2) El tamaño de las cajas es más o menos constante, igual a unas 0.2 unidades. Como $10^{0.2}$ es aproximadamente 1.6, esto quiere decir que en general el cuartil superior es 60% más grande que el cuartil inferior. Otra manera de decir esto, dada la simetría: el 75% de los datos varían un 30% o menos a partir del valor de la mediana.

3) Sin embargo, vemos que los datos atípicos, a lo largo de la mayoría de los estados, puede alejarse de la mediana hasta 0.5 unidades. Como $10^{0.5}$ es aprox 3.1, eso quiere decir que los ventas pueden ser más del doble que las medianas.

Gráficas de dispersión

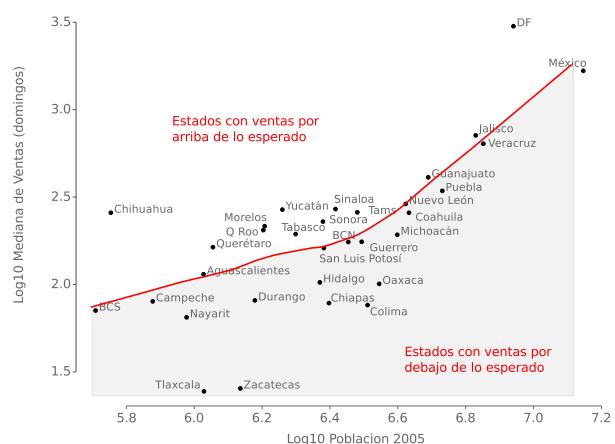
Ahora pensamos en datos que vienen por pares. En el siguiente ejemplo, cada caso (un estado) tiene dos mediciones: su población en 2005, y la media de ventas promedio un día dado de la semana, a lo largo de unos 3 años. La gráfica muestra lo obvio: en los estados más grandes se vende más. **Este es un patrón de asociación de dos variables numéricas.** ¿Cómo podemos seguir en el análisis de esta asociación? Primero tratamos de capturar la relación clara ajustando, por ejemplo una recta de regresión. En este caso, funciona mejor usar regresión local:



Buscamos datos que se salen del patrón "general" (donde el patrón general son los niveles locales de ventas). Por ejemplo, el DF tiene vende más de lo que se esperaría dada su población. Zacatecas, sin embargo, tiene una población similar a Morelos o Durango pero vende considerablemente menos. Podemos empezar marcando nuestra gráfica para resaltar estas observaciones, como a la derecha.

Una vez que hemos quitado el efecto obvio del tamaño de los estados, podemos seguir investigando estos datos: ¿Por qué Chihuahua o el DF están tan arriba de lo esperado? ¿Hay más tiendas? ¿Se podrían explicar estas diferencia con el PIB per cápita en cada estado, por ejemplo?

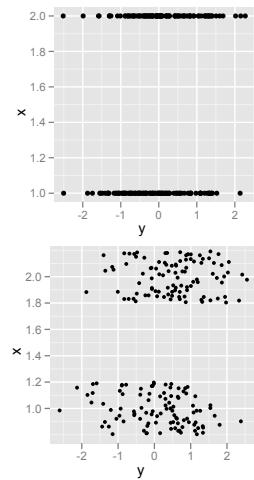
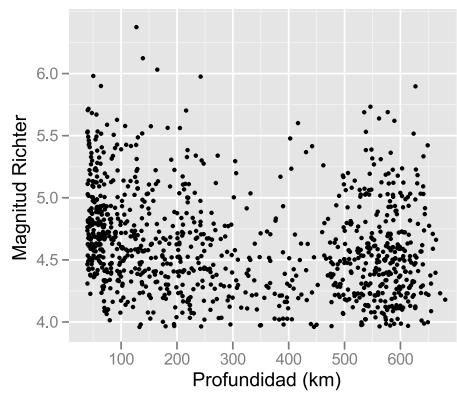
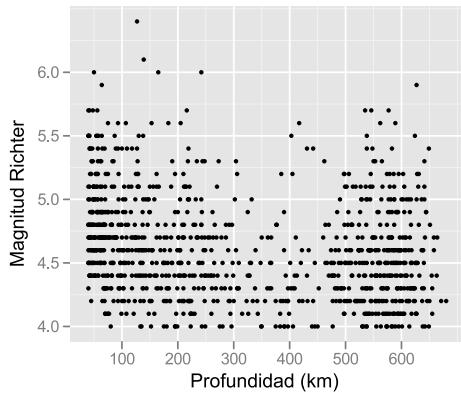
Esto se llama hacer el **análisis de los residuos** de los datos a partir de la curva ajustada.



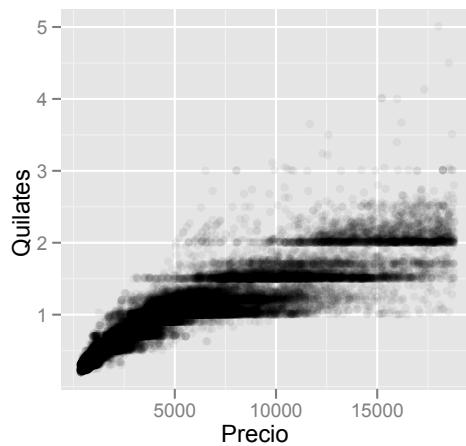
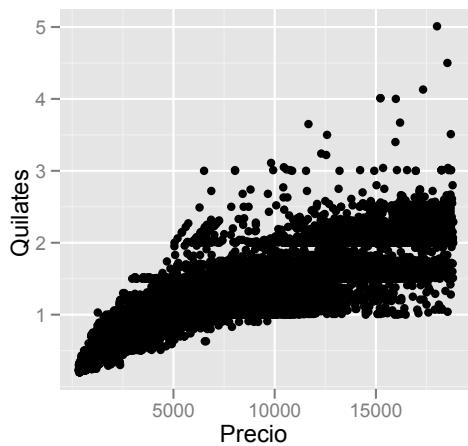
Este **análisis de residuos** es análogo al que vimos en el ejemplo del gasto en alimentos de los hogares mexicanos: primero describimos el patrón general, luego buscamos maneras eficientes de expresar cómo distintas unidades se apartan de dicho patrón. Son residuos porque es la variación que queda después de controlar por los patrones encontrados.

Gráficas de dispersión: vibración, transparencia

En esta parte discutimos cómo mejorar nuestras gráficas de dispersión para superar un problema típico: el traslape de datos y etiquetas. En el ejemplo de la derecha, hemos graficado contra una variable que sólo toma dos valores. El resultado no es bueno, pues es difícil entender cómo varía la otra variable dentro de cada grupo. La solución es **vibrar** los valores de la variable y , que consiste en agregar una cantidad aleatoria pequeña que evite el traslape. Abajo graficamos datos recogidos acerca de temblores ocurridos cerca de Fiji, y nos interesa la profundidad a la que ocurrió cada temblor y la magnitud correspondiente. El problema es que los datos de magnitud están redondeados a dos cifras decimales, como muestra la primera gráfica. La vibración da una idea más clara de la asociación entre estas dos variables. El patrón de bandas de la gráfica de la izquierda no es lo que estamos buscando en estos datos.

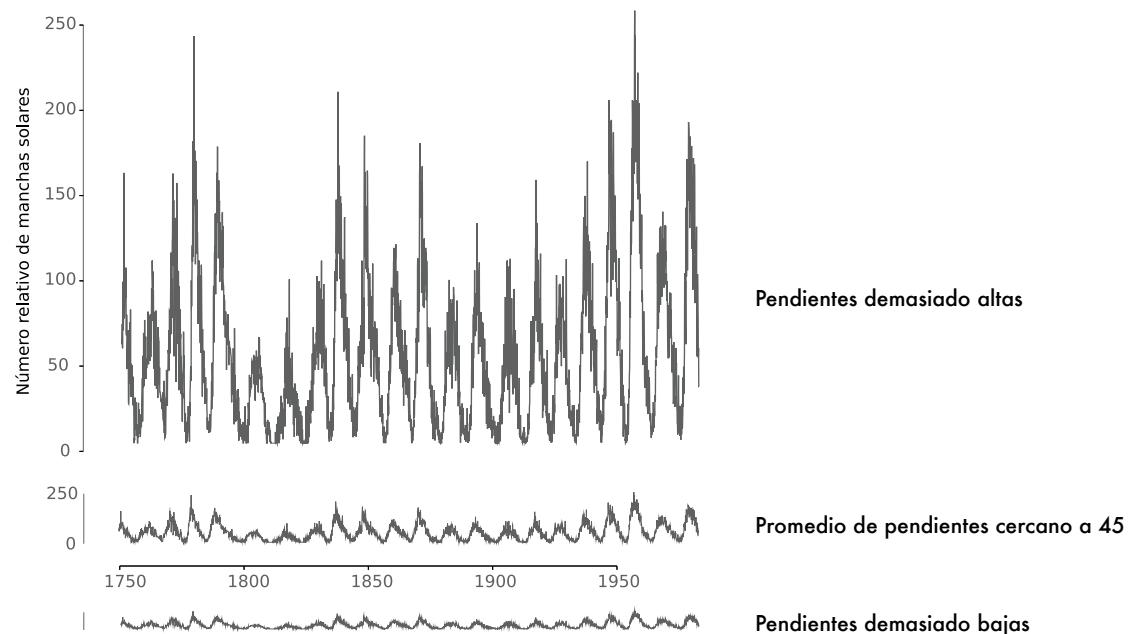


A veces vibrar los datos no es la solución. En particular, cuando graficamos grandes cantidades de datos, es prácticamente imposible evitar el traslape. La solución es la **transparencia**: permitimos que datos encimados contribuyan al valor de la opacidad de cada región de la gráfica de dispersión. El resultado se puede ver en el ejemplo de abajo, donde graficamos datos de precio y quilate para más de 50 mil diamantes.

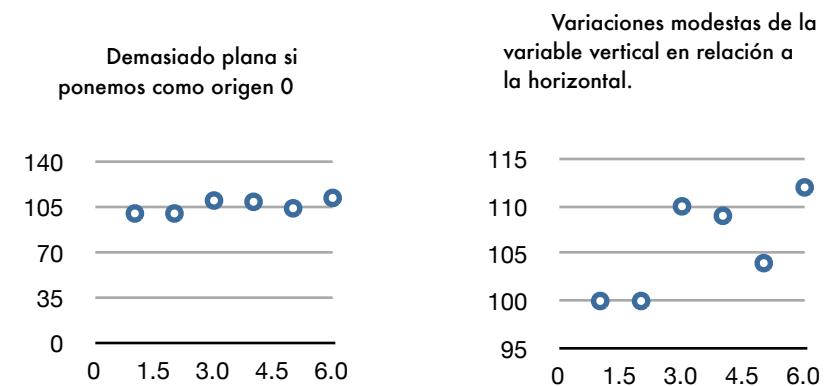


Series de tiempo y promedio de 45

Las series de tiempo son una especie particular de las gráficas de dispersión, en donde la dimensión horizontal es el tiempo. Buscamos entender cómo varía una medición dada en el tiempo. Estas gráficas son más útiles cuando se construyen usando el principio del **promedio de 45 grados**: los patrones de variación en el tiempo se distinguen mejor (aproximadamente) cuando el promedio de pendiente (en valor absoluto) en las gráficas está cercano a 45 grados. El siguiente ejemplo, que muestra la actividad de manchas solares del sol, muestra claramente este principio:



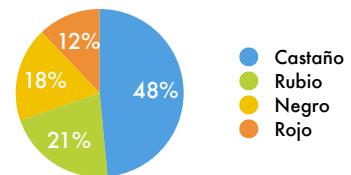
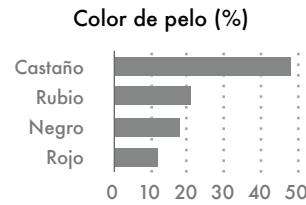
Este es también un principio para decidir la **razón de aspecto** de cualquier gráfica de dispersión (y también gráficas de barras, por ejemplo). Esta regla supera el principio de que "las escalas deben comenzar en cero". En realidad, este principio cuida contra dos errores en la graficación: no poner atención a la escala e intentar comparar gráficas que no están dibujadas en la misma escala. Poniendo atención a estos dos aspectos (incluso llamado a veces la atención a estos puntos. Stephen Few) no hay necesidad de seguir la regla del 0.



Técnicas simples de visualización para datos categóricos

La herramienta básica para explorar datos categóricos es la tabla, junto con algunas gráficas simples (gráficas de barras, por ejemplo) que la representan. En general, es buena idea evitar las gráficas de pie.

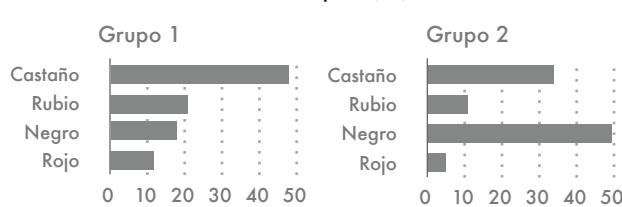
	No. casos	% del Total
Castaño	286	48
Rubio	127	21
Negro	108	18
Rojo	71	12
Total	592	



Tablas y gráficas de barras son útiles para comparar conjuntos de datos con las mismas categorías

Si ponemos los porcentajes, ¿por qué no usar mejor una tabla? ¿Qué rotación es apropiada? Porcentajes similares son difíciles de distinguir, y ¿cómo etiquetar? Usar leyendas en general dificulta la interpretación.

	Grupo 1 (%)	Grupo 2 (%)
Castaño	48	34
Rubio	21	11
Negro	18	50
Rojo	12	5



Usando el principio de los pequeños múltiplos.

Índices para dos conjuntos de datos categóricos

Una técnica usual, que veremos más adelante aplicada más en general, es la de producir índices. En este caso, podemos calcular el cociente del Grupo 2 vs. Grupo 1 para cada color de pelo. Marcamos índices por encima de 130 (diferencia +30%) o por debajo de 70 (diferencia -30%). Esta es una **tabla marcada**. Una tabla marcada apropiadamente puede ser mucho más clara que la misma tabla no marcada.

	Grupo 1 (%)	Grupo 2 (%)	Índice
Castaño	48	34	70
Rubio	21	11	51
Negro	18	50	274
Rojo	12	5	42

Cuando queremos entender y comparar varios conjuntos de datos categóricos, todos sobre las mismas categorías, usamos principios similares:

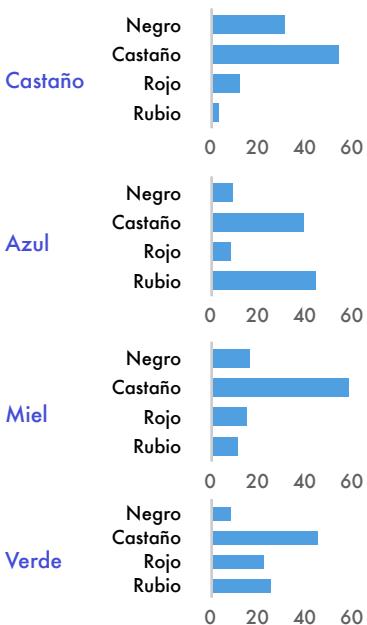
- 1) Tablas marcadas, aprovechar índices.
- 2) Gráficas simples que fácilmente se puedan replicar en pequeños múltiplos.

Entender cómo están asociadas dos variables categóricas es similar a nuestro ejemplo anterior de datos del ENIGH. En este nuevo ejemplo, consideraremos la relación entre color de pelo y color de ojos.

No. casos		Castaña	Azul	Miel	Verde	Total
	Negro	68	20	15	5	108
	Castaña	119	84	54	29	286
	Rojo	26	17	14	14	71
	Rubio	7	94	10	16	127
Total		220	215	93	64	592

%		Castaña	Azul	Miel	Verde	Total
	Negro	31	9	16	8	18
	Castaña	54	39	58	45	48
	Rojo	12	8	15	22	12
	Rubio	3	44	11	25	21
	Total	100	100	100	100	100

Índices		Castaña	Azul	Miel	Verde
	Negro	169	51	88	43
	Castaña	112	81	120	94
	Rojo	99	66	126	182
	Rubio	15	204	50	117



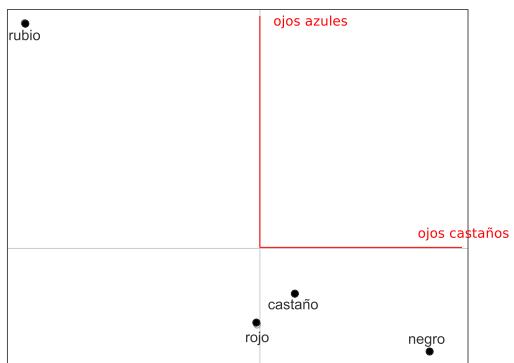
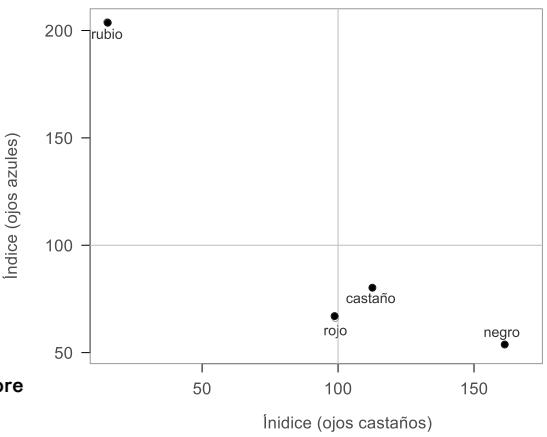
Nótese que, igual que en el ejemplo anterior de regresión, en este caso entendemos los niveles esperados como los dados por la columna Total (%). Buscamos entender justamente como varían los grupos con respecto a estos niveles observados. Esto es lo que significa entender la relación entre dos variables categóricas. Los índices juegan ahora el papel de los residuales: indican desviaciones con respecto a lo esperado. ¿Cómo podemos graficar estos datos?

Podemos en primer lugar producir pequeños múltiples de gráficas de barras para los porcentajes originales. Nótese sin embargo que, como pelo castaño es común para todos los grupos, y pelo rojo es poco común en todas las categorías, es más fácil usar los índices para entender las asociaciones entre estas dos variables. Podemos graficar los índices directamente, pero hay otras técnicas útiles en este punto.

Mapas de asociación

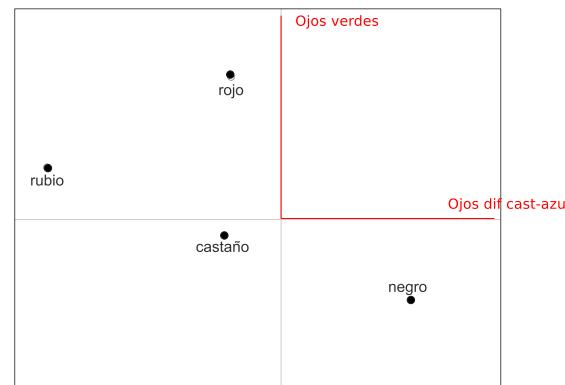
Si sólo nos interesa entender cómo se comporta ojos azules contra ojos castaños, sobre todos los posibles colores de pelo, podemos comenzar con la gráfica de la derecha. Quitamos los ejes para entender el patrón de asociación: como negro está abajo a la derecha, está positivamente asociado con ojos castaños, pero negativamente asociada con ojos azules. Como rubio está arriba a la izquierda, está positivamente asociado con ojos azules pero negativamente asociado con ojos castaños. Pelo rojo está abajo y en el centro: eso quiere decir que está negativamente asociado a ojos azules, pero no hay asociación para un lado o para el otro con ojos castaños.

Nótese que en la gráfica los colores de pelo están aproximadamente en una recta. Esto quiere decir que estamos poniendo información redundante en este mapa: si una color de pelo es alto en ojos azules, es bajo en ojos castaños, y a la inversa. Podemos resumir esta observación creando una nueva dimensión donde calculamos la diferencia Índice Ojos castaños-Índice ojos Azules.



Calculamos entonces la siguiente tabla, y graficamos en el eje horizontal Dif C-A (diferencia de perfil ojos castaños-ojos azules). Escogemos otra columna para el eje vertical, como ojos verdes. El resultado está a la derecha

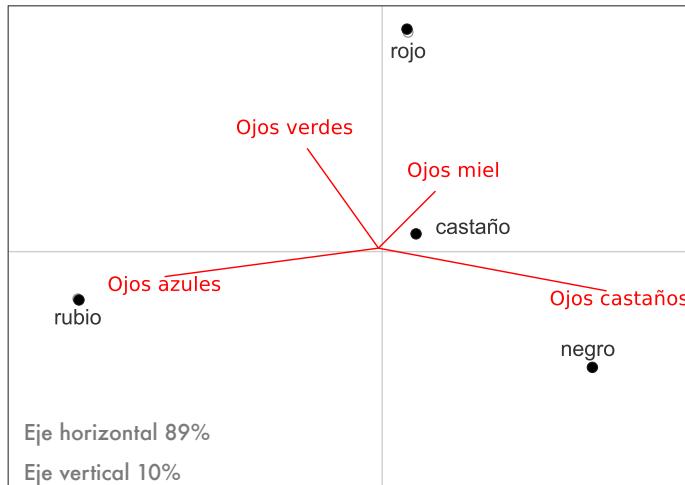
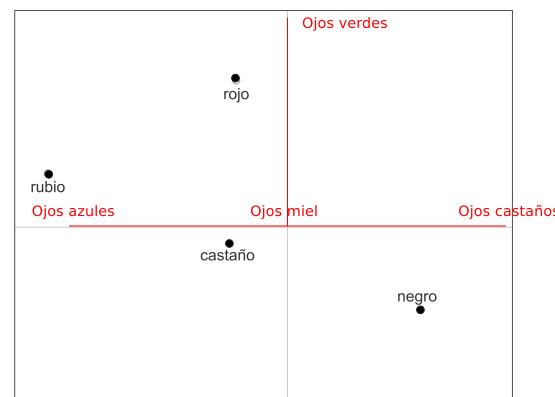
Índices	Castaño	Azul	Dif C-A	Verde
Negro	169	51	118	43
Castaño	112	81	31	94
Rojo	99	66	33	182
Rubio	15	204	-189	117



Colores de pelo a la izquierda cargan más a ojos azules, y colores de pelo a la derecha cargan más a ojos negros. Este eje horizontal resume bien lo que vimos en los dos ejes iniciales. Pero ahora tenemos también ojos verdes graficado (que está asociado con pelo rojo). Finalmente, ponemos ojos miel en el centro, pues por el momento no contribuye a nuestro mapa.

Que ojos azules y ojos castaños vayan en direcciones contrarias se interpreta como que están negativamente correlacionados (cuando uno es alto en perfil, el otro es bajo, a lo largo de los posibles colores de pelo)

Este proceso que hicimos manualmente se puede optimizar usando métodos matemáticos. La idea general es buscar combinaciones de atributos que discriminen mucho entre las categorías. Encontramos primero el eje horizontal. Después, buscamos una segunda combinación que discrimine entre las marcas, pero que no contenga la información del primer eje horizontal. El resultado es el mapa que está abajo:



Índices	Castaño	Azul	Miel	Verde
Negro	169	51	88	43
Castaño	112	81	120	94
Rojo	99	66	126	182
Rubio	15	204	50	117

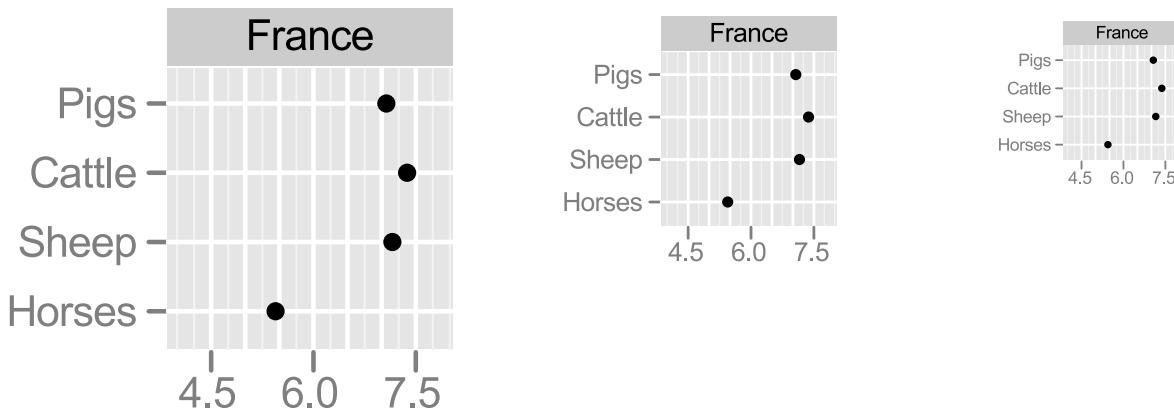
Interpretación

Principios generales:

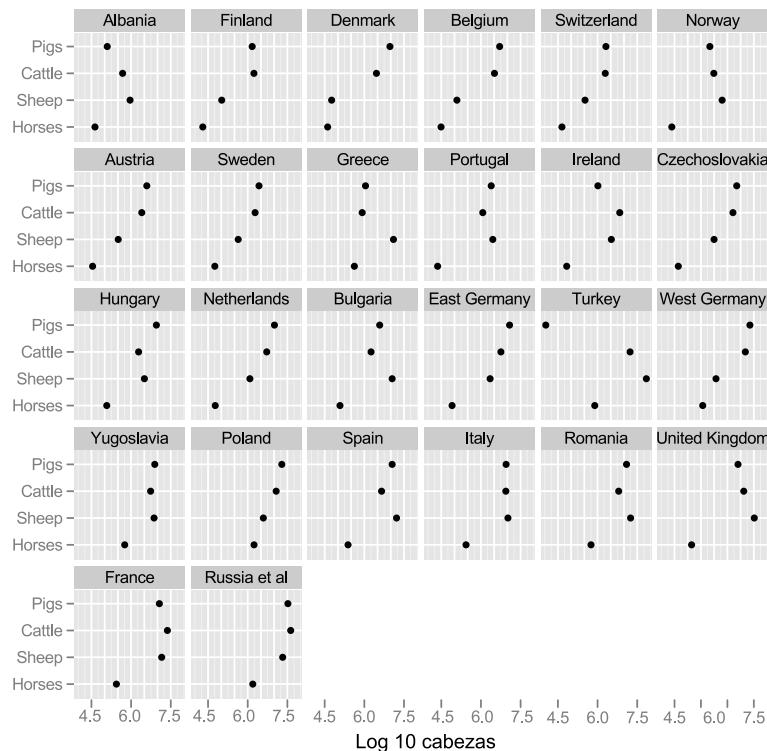
- 1) El mapa sirve para encontrar y explicar patrones que están en la tabla de datos original. La interpretación de un mapa siempre es más sólida cuando se hace junto con la tabla de datos. Siempre que sea posible se debe hacer el análisis paralelo (mapa/tabla).
- 2) Los mapas no se interpretan por cercanía de las direcciones con los puntos. Es decir, que Ojos miel esté cercano a pelo castaño no quiere decir que estén asociados. Los mapas se interpretan en direcciones.
- 3) Sin embargo, que dos puntos negros (renglones) estén cercanos quiere decir que tienen perfiles similares en las variables de las columnas, y viceversa.

Pequeños múltiplos y densidad gráfica

La densidad de una gráfica es el tamaño del conjunto de datos que se grafica comparado con el área total de la gráfica. En el siguiente ejemplo, graficamos en logaritmo-10 de cabezas de ganado en Francia (cerdos, res, ovejas y caballos). La gráfica de la izquierda es pobre en densidad pues sólo representa 4 datos. La manera más fácil de mejorar la densidad es hacer más chica la gráfica:

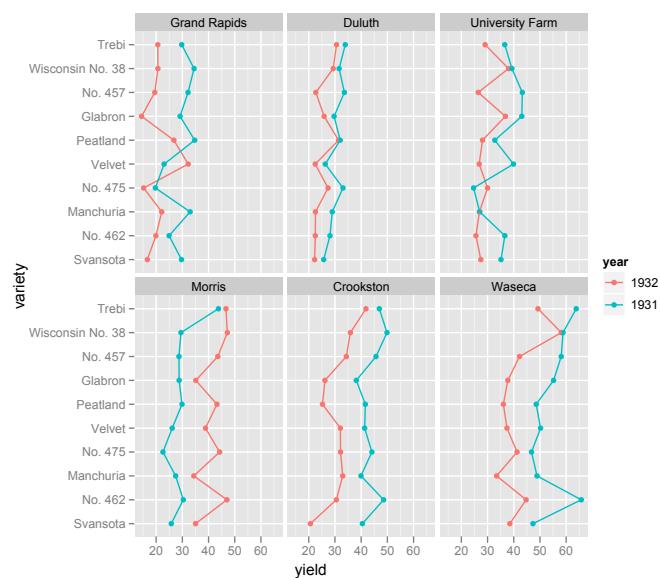


La razón de este encogimiento es una que tiene qué ver con las **oportunidades perdidas** de una gráfica grande. Si repetimos este mismo patrón (misma escala, mismos tipos de ganado) para distintos países obtenemos la siguiente gráfica:



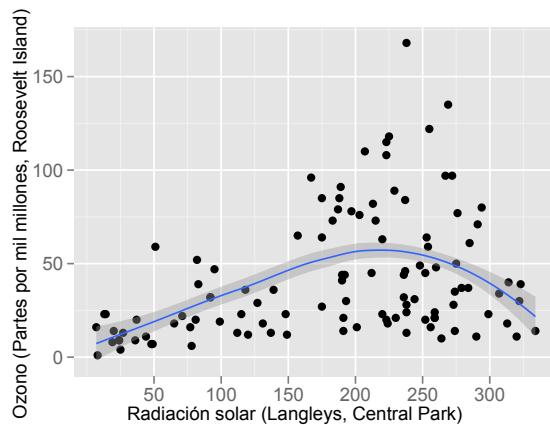
Esta gráfica da un ejemplo de pequeños múltiplos: graficar distintas partes de nuestro conjunto de datos repitiendo una gráfica simple.

Esta es una **gráfica de puntos**. Es útil como sustituto de una gráfica de barras, y es superior en el sentido de que una mayor proporción de la tinta que se usa es tinta de datos. Otra vez, mayor proporción de tinta de datos representa más oportunidades que se pueden capitalizar, como muestra la gráfica de punto y líneas que mostramos al principio:

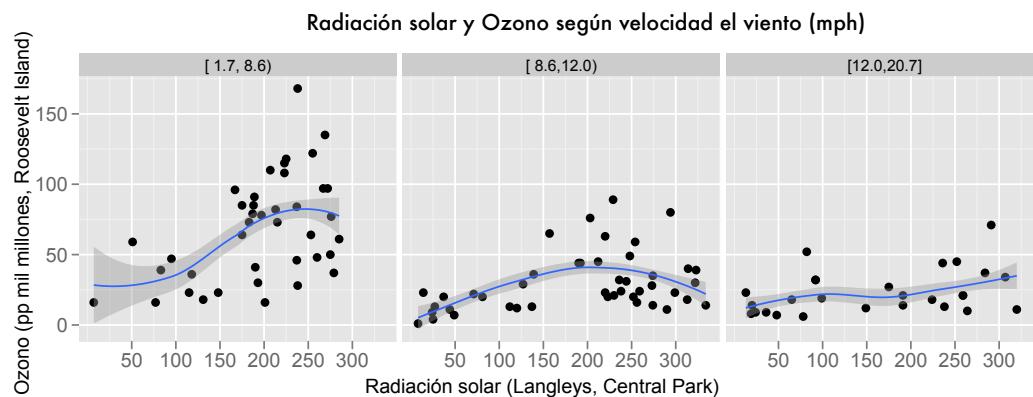


Más pequeños múltiplos

Los pequeños múltiplos son poderosos en el análisis de datos, y es un patrón que podemos aplicar a cualquier gráfica base. En el siguiente ejemplo, graficamos radiación solar contra niveles de ozono para ciertas estaciones de medición de Nueva York. La asociación entre estas dos variables no es tan simple:

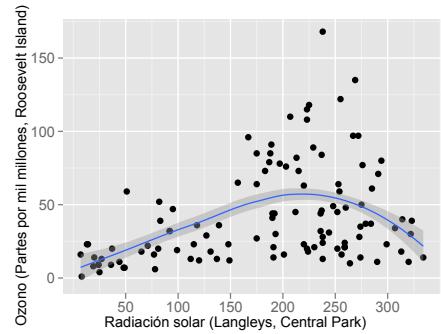
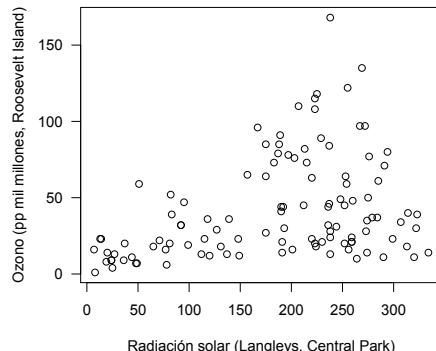
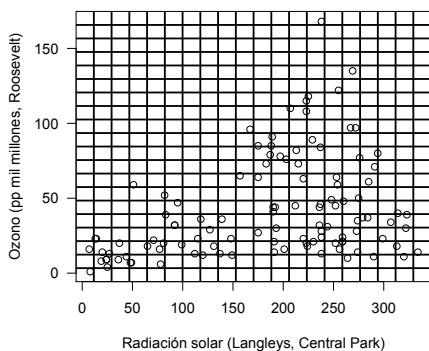


Mediante la idea de pequeños múltiplos podemos incluir otra variable en nuestro análisis: la velocidad del viento, que aclara la relación complicada que vimos en la gráfica anterior. Para hacer la gráfica de abajo, dividimos en categorías velocidad del viento. Dependiendo del tamaño de muestra, es posible hacer categorías más finas o tenemos que usar otras más gruesas (incluso es posible usar categorías traslapadas).



Tinta de datos

Maximizar la proporción de tinta de datos en nuestras gráficas tiene beneficios inmediatos. La regla es: si hay tinta que no representa variación en los datos, o la eliminación de esa tinta no representa pérdidas de significado, esa tinta debe ser eliminada. El ejemplo más claro es el de las rejillas en gráficas y tablas:



	1	2	3	4	5	6	7	8	9	10	Total %
Carnes	20	22	22	24	25	26	26	27	26	25	24.9
Cereales	25	24	24	22	22	21	20	19	17	14	19.8
Leche y Derivados	11	11	13	13	13	14	14	15	15	16	14.0
Verduras, Legumbres	18	16	16	15	14	14	13	12	11	11	13.3
Otros Alimentos	5.4	5.7	7.2	7.4	8.7	8.9	10	11	14	14	10.2
Frutas	3.6	3.6	3.5	4.4	4.4	4.5	5.0	5.2	5.5	7.3	5.0
Huevo	4.8	4.6	4.4	4.2	3.4	3.2	3.2	2.9	2.5	1.9	3.2
Pescados Y Mariscos	2.1	2.4	2.2	2.2	2.4	2.4	2.4	3.0	3.1	4.6	2.9
Tubérculos	2.0	2.0	2.0	1.9	1.9	1.7	1.5	1.5	1.4	1.2	1.6
Aceites Y Grasas	2.5	2.4	1.9	1.7	1.6	1.6	1.3	1.2	1.1	1.1	1.5
Azúcar Y Mielas	3.1	2.7	2.1	1.8	1.7	1.5	1.1	1.2	1.0	0.9	1.5
Café, Té Y Chocolate	1.3	1.5	1.1	0.9	1.0	1.0	0.8	0.9	0.9	1.2	1.0
Especias Y Aderezos	1.1	1.0	1.0	1.0	1.0	1.1	1.1	1.0	1.0	1.0	1.0
Total (miles de millones)	5.4	7.8	9.5	11	12	13	14	15	16	19	

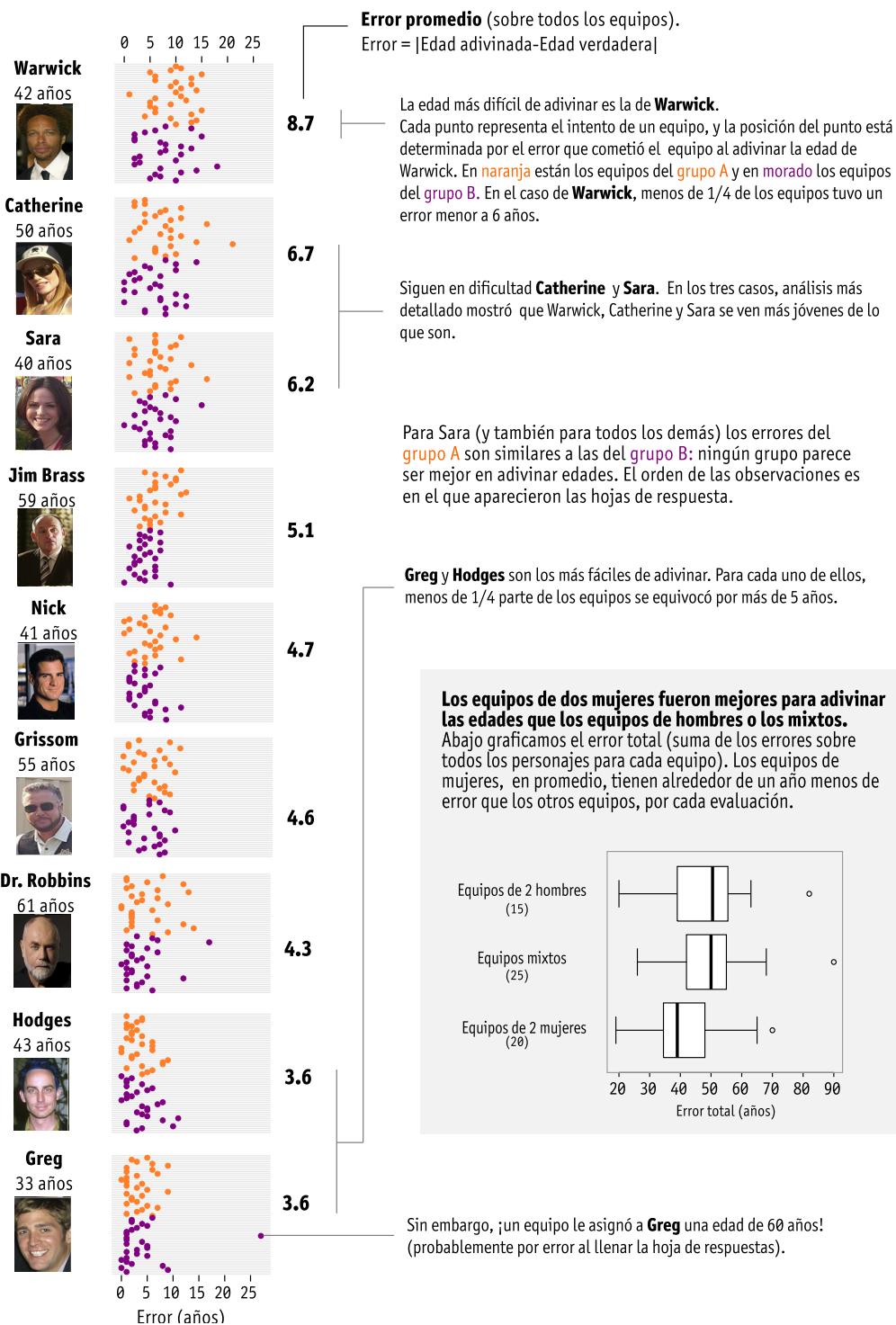
¿Por qué usar grises en lugar de negros? La respuesta tiene qué ver con el principio de tinta de datos: si marcamos las diferencias sutil pero claramente, tenemos más oportunidades abiertas para hacer énfasis en lo que nos interesa: a una gráfica o tabla saturada no se le puede hacer más - es difícil agregar elementos adicionales que ayuden a la comprensión. Si comenzamos marcando con sutileza, entonces se puede hacer más. Los mapas geográficos son un buen ejemplo de este principio.

El espacio en blanco es suficientemente bueno para indicar las fronteras en una tabla, y facilita la lectura:

	1	2	3	4	5	6	7	8	9	10	Total %
Carnes	20.0	22.4	22.3	23.9	24.9	25.7	26.4	26.9	26.1	25.1	24.9
Cereales	24.9	23.8	23.6	22.1	21.6	20.6	20.2	18.9	17.1	14.4	19.8
Leche y Derivados	10.9	11.4	13.0	13.2	13.3	14.2	14.0	14.5	14.7	16.4	14.0
Verduras, Legumbres	18.2	16.3	15.5	15.1	14.0	13.7	12.7	12.3	11.4	10.5	13.3
Otros Alimentos	5.4	5.7	7.2	7.4	8.7	8.9	10.3	10.5	14.2	14.4	10.2
Frutas	3.6	3.6	3.5	4.4	4.4	4.5	5.0	5.2	5.5	7.3	5.0
Huevo	4.8	4.6	4.4	4.2	3.4	3.2	3.2	2.9	2.5	1.9	3.2
Pescados Y Mariscos	2.1	2.4	2.2	2.2	2.4	2.4	2.4	3.0	3.1	4.6	2.9
Tubérculos	2.0	2.0	2.0	1.9	1.9	1.7	1.5	1.5	1.4	1.2	1.6
Aceites Y Grasas	2.5	2.4	1.9	1.7	1.6	1.6	1.3	1.2	1.1	1.1	1.5
Azúcar Y Mielas	3.1	2.7	2.1	1.8	1.7	1.5	1.1	1.2	1.0	0.9	1.5
Café, Té Y Chocolate	1.3	1.5	1.1	0.9	1.0	1.0	0.8	0.9	0.9	1.2	1.0
Especias Y Aderezos	1.1	1.0	1.0	1.0	1.0	1.1	1.1	1.0	1.0	1.0	1.0
Total (miles de millones)	5.4	7.8	9.5	10.5	11.9	12.6	14.1	14.7	16.0	18.8	

La decoración tiene su lugar

Adivinar la edad de una persona. ¿Hay personas cuya edad es más difícil que adivinar que otras? ¿Las mujeres son mejores que los hombres para adivinar edades? En dos grupos con unos 60 alumnos cada uno, formamos equipos de dos personas (algunos de dos hombres, otros de dos mujeres, y otros mixtos). Les preguntamos adivinar la edad de los nueve personajes principales del programa de televisión CSI, y calculamos el error en cada intento contando cuántos años se desvía cada intento de la edad verdadera que se pretendía adivinar.



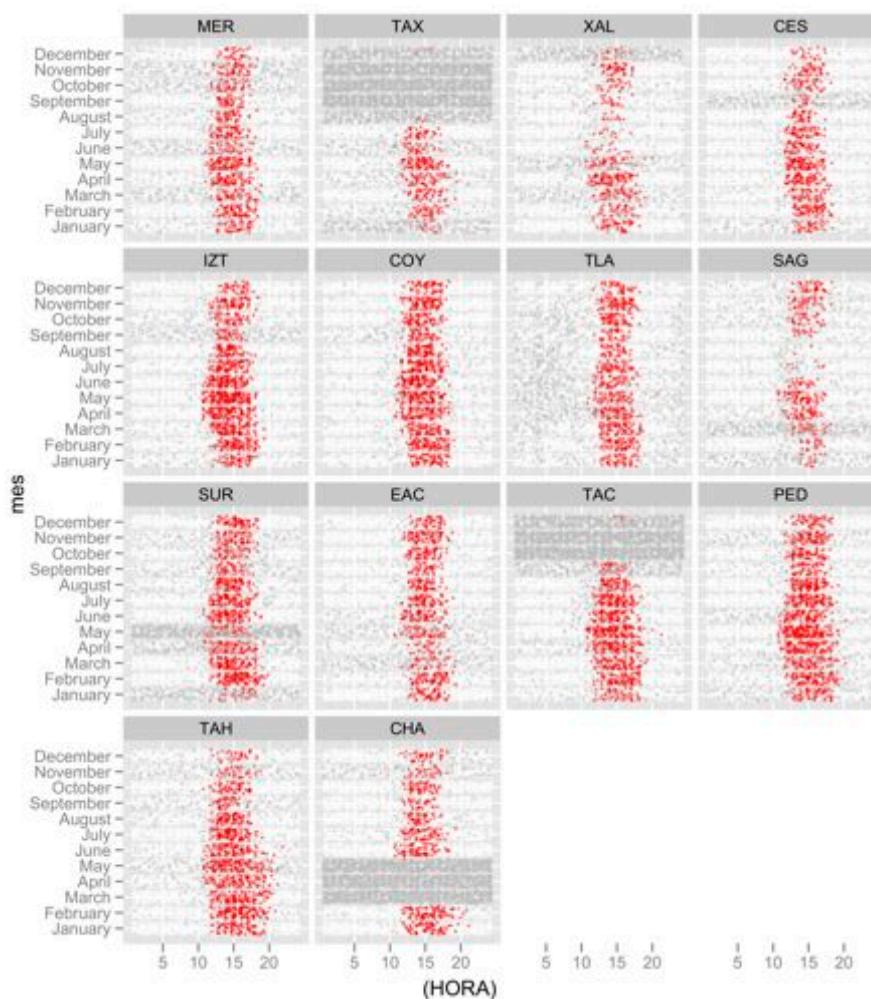
Ejemplo: niveles de ozono en el Valle de México

Buscamos entender, mediante una visualización:

- 1) ¿Qué tan altos/peligrosos son los niveles de ozono en el Valle de México?
- 2) ¿Hay lugares de mucha más alta contaminación de ozono que otros? ¿O los niveles de contaminación son similares a lo largo de las estaciones de medición?
- 3) ¿Qué horarios son los más peligrosos?

Usaremos los datos de la RAMA (red automática de medición atmosférica) de la SMA. Consideraremos varias estaciones de medición distribuidas en el Valle de México, para el año 2009. Adicionalmente, consideraremos la recomendación de la OMS acerca de la exposición a ozono, que indica que niveles de 0.07 pp miles de millones de ozono causan dificultades pulmonares bajo exposiciones de 8 horas o más.

Los datos tienen algunas dificultades: en primer lugar, hay gran cantidad de datos faltantes. Nuestra preparación consiste en marcar aquellos horas del día (en 2009) donde hubo concentraciones mayores a 0.07 pp miles de millones, y aquellas horas en donde no se registró ninguna medición. Bajo este criterio, marcamos como rojas las horas peligrosas, con blanco las horas seguras, y con gris oscuro las horas sin medición. El resultado es el siguiente:



En cada combinación de hora, mes y estación tenemos unas 30 mediciones. Vibrámos y usamos transparencia para que el color de cada una de estas combinaciones refleje la proporción de rojos/blancos/grises. De aquí vemos que bajo nuestro criterio, existen algunas regiones (Pedregal o Tlalhuac, por ejemplo), donde en ciertos meses nos acercamos al nivel marcado por la OMS.

La gráfica representa unos 120,00 datos distintos.

¿Qué otros aspectos de los datos se pueden entender con estas gráficas? ¿Qué mejorarías lo harás diferente?

Referencias

Edward Tufte

Beautiful Evidence. Cheshire, CT: Graphics Press, 2006
The Visual Display of Quantitative Information. Graphics Press., 1983
Envisioning Information. Graphics Press, 1990.

William Cleveland

Visualizing Data. Hobart Press, 1993
Elements of Graphing Data, 2a. ed., Hobart Press, 1994

Howard Wainer

Picturing the Uncertain World, Princeton University Press, 2009
Graphic Discovery: A Trout in the Milk and Other Visual Adventures,
Princeton University Press, 2004

Stephen Few

Now you see it, Analytics press, 2009
Show Me the Numbers, 2004

Herramientas

Este documento fue creado con:

R (el paquete de estadística libre),
<http://www.r-project.org/>

Inkscape (software para edición de imágenes vectoriales) y
<http://inkscape.org/>

Además de Excel (tablas) y Pages (layout del documento)