

EST-46111: Fundamentos de Estadística con Remuestreo

Teresa Ortiz (001), Alfredo Garbuno (002), Felipe González

Contenido

Información del curso	v
Temario	v
Evaluación	VI
1. Principios de visualización	1
El cuarteto de Ascombe	1
Introducción	2
Visualización popular de datos	3
Teoría de visualización de datos	4
Principios generales del diseño analítico	4
Técnicas de visualización	4
Indicadores de calidad gráfica	4
Factor de engaño y Chartjunk	6
Pequeños múltiplos y densidad gráfica	7
Más pequeños múltiplos	8
Tinta de datos	10
Decoración	13
Percepción de escala	14
Ejemplo: gráfica de Minard	14
2. Análisis exploratorio	17
El papel de la exploración en el análisis de datos	17
Algunos conceptos básicos	17
Media y desviación estándar	25
Ejemplos	26
Precios de casas	26

Prueba Enlace	30
Estados y calificaciones en SAT	32
Tablas de conteos	36
Loess	40
Ajustando curvas <code>loess</code>	44
Series de tiempo	46
Caso de estudio: nacimientos en México	46
Datos de natalidad para México	48
Tendencia	49
Componente anual	50
Día de la semana	50
Residuales	51
Reestimación	52
Análisis de componentes	53
Residuales: antes y después de 2006	55
Otros días especiales: más de residuales	56
Semana santa	56

Información del curso

Notas del curso *Fundamentos de Estadística con Remuestreo* del programa de maestría en Ciencia de Datos del ITAM. En caso de encontrar errores o tener sugerencias del material se agradece la propuesta de correcciones mediante pull requests.

Ligas

- Notas: <https://fundamentos-est.netlify.app/>
- Correos: teresa.ortiz.mancera@gmail.com, alfredo.garbuno@itam.mx.
- GitHub: <https://github.com/tereom/fundamentos>
- Foros de discusion:
 - Grupo Teresa: slack
 - Grupo Alfredo: canvas

Este trabajo está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

Temario

Datos y análisis exploratorio

Referencias: (Cleveland, 1994), (Chihara and Hesterberg, 2018)

1. Visualización y análisis exploratorio
2. Tipos de datos o estudios
 - Muestras diseñadas y muestras naturales
 - Experimentos y datos observacionales

Introducción a Pruebas de Hipótesis

Referencias: (Chihara and Hesterberg, 2018)

3. Introducción a pruebas de hipótesis. Pruebas de permutaciones
4. Muestras pareadas y otros ejemplos

Estimación y distribución de muestreo

Referencias: (Chihara and Hesterberg, 2018), (Hesterberg, 2015)

5. Estimadores y su distribución de muestreo
6. Repaso de probabilidad y Teorema del límite central

Introducción a estimación por intervalos

Referencias: (Chihara and Hesterberg, 2018), (Efron and Tibshirani, 1993), (Hesterberg, 2015)

7. El método plugin y el bootstrap
8. Bootstrap e Intervalos de confianza. Ejemplos.

Estimación

Referencias: (Chihara and Hesterberg, 2018), (Wasserman, 2013)

9. Estimación por máxima verosimilitud
10. Ejemplos de estimación por máxima verosimilitud y Bootstrap paramétrico
11. Propiedades de estimadores de máxima verosimilitud

Más de pruebas de hipótesis

Referencias: (Chihara and Hesterberg, 2018), (Wasserman, 2013)

12. Pruebas de hipótesis para medias y proporciones: una y dos poblaciones.

Introducción a inferencia bayesiana

Referencias: (Kruschke, 2015)

13. Introducción a inferencia bayesiana
14. Ejemplos de distribuciones conjugadas
15. Introducción a métodos computacionales básicos: Muestreadores Metrópolis y Gibbs

Evaluación

- Tareas semanales 20 %
- Parcial teórico + parcial a casa 40 %
- Final a casa 40 %

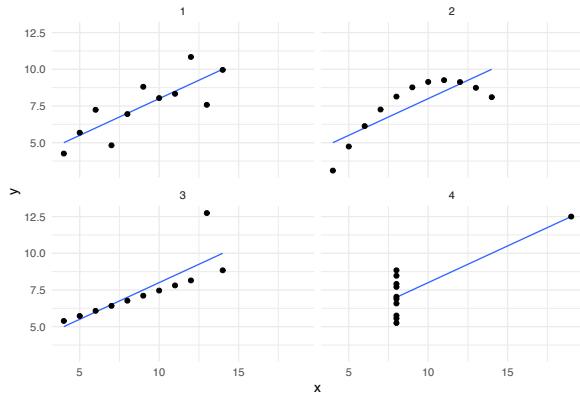
Sección 1

Principios de visualización

El cuarteto de Ascombe

En 1971 un estadístico llamado Frank Anscombe (fundador del departamento de Estadística de la Universidad de Yale) publicó cuatro conjuntos de dato. Cada uno consiste de 11 observaciones. La peculiaridad de estos conjuntos es que tienen las mismas propiedades estadísticas.

Sin embargo, cuando analizamos los datos de manera gráfica en un histograma encontramos rápidamente que los conjuntos de datos son muy distintos.



Media de x : 9
Varianza muestral de x : 11
Media de y : 7.50
Varianza muestral de y : 4.12
Correlación entre x y y : 0.816
Línea de regresión lineal: $y = 3,00 + 0,500x$

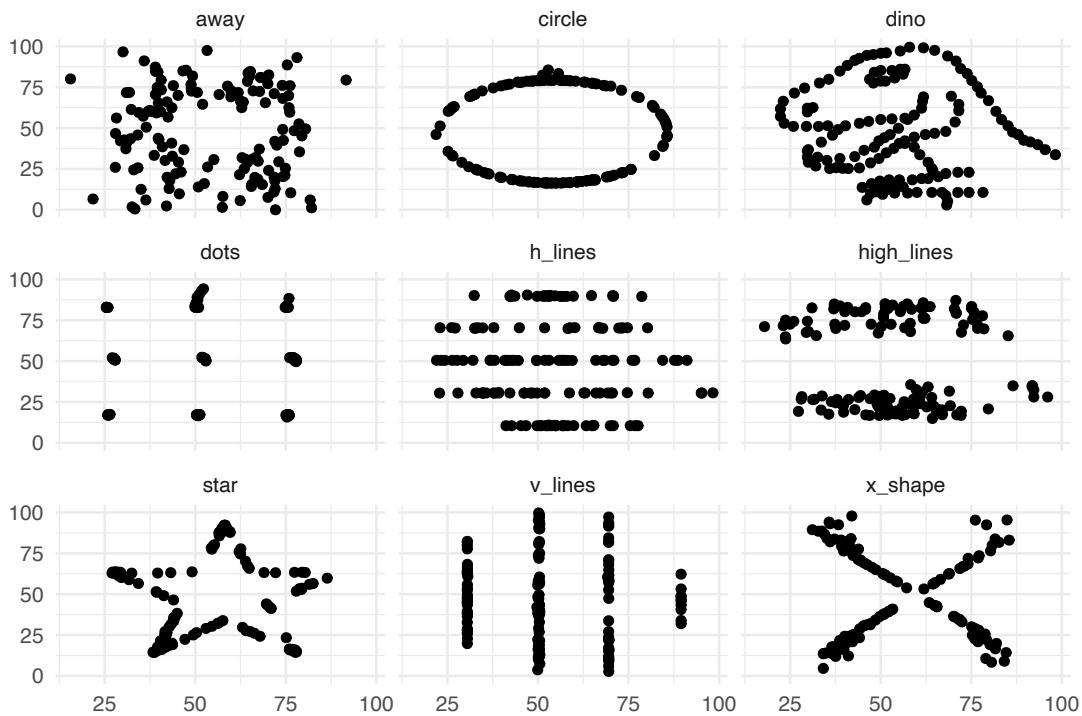
En la gráfica del primer conjunto de datos, se ve clara una relación lineal simple con un modelo que cumple los supuestos de normalidad. La segunda gráfica (arriba a la derecha) muestra unos datos que tienen una asociación pero definitivamente no es lineal. En la tercera gráfica (abajo a la izquierda) están puntos alineados perfectamente en una línea recta, excepto por uno de ellos. En la última gráfica podemos ver un ejemplo en el cual basta tener una observación atípica para que se produzca un coeficiente de correlación alto aún cuando en realidad no existe una asociación lineal entre las dos variables.

El cuarteto de Ascombe inspiró una técnica reciente para crear datos que comparten las mismas propiedades estadísticas al igual que en el cuarteto, pero que producen gráficas muy distintas (Matejka, Fitzmaurice).

```
## Warning: package 'datasauRus' was built under R version 4.0.2
```

```
## Warning: package 'ggridge' was built under R version 4.0.2
```

```
## Warning: package 'animation' was built under R version 4.0.2
```

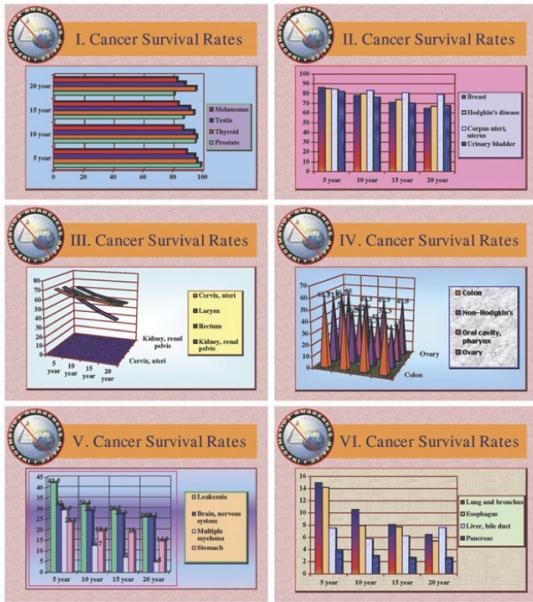


Introducción

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 (Tufte, 2006), 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 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).

The New York Times

Friday, September 9, 2016 | Today's Paper | Video | 60°F | S. & P. 500 +1.44%

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

North Korea Tests a Mightier Nuclear Bomb, Raising Tension

By CHOI SANG-HUN and JANE PERLEZ

South Korean officials said the underground test, North Korea's fifth, produced a more powerful explosive yield than the North's previous detonations.

Where murder rates rose significantly in 2015

Murder Rates Rose in a Quarter of the Nation's 100 Largest Cities

In our analysis of new data, half of the increase in homicides came from just seven cities.

By HAEYOUN PARK and JOSH KATZ

The Opinion Pages

A Debate Disaster Waiting to Happen

By THE EDITORIAL BOARD

Moderators at future presidential debates must do a better job of holding the candidates accountable.

- Brooks: Time for a Realignment
- Krugman: Donald Trump's 'Big Liar' Technique
- Egan: The Conscience of the Contrarian Voter
- Op-Ed: What Trump Doesn't Understand About the Military

Sunday Review

'Trump's Going to Get Us Killed,' Trump Voter Says

By ROGER COHEN

Appalachian voters know perfectly well the candidate is dangerous. But they're desperate for change.

THE CROSSWORD »
Play Today's Puzzle

Teoría de visualización de datos

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.

En resumen, hablaremos de las siguientes guías:

Principios generales del diseño analítico

Aplicables a una presentación o análisis completos, y como guía para construir nuevas visualizaciones (Tufte, 2006).



- 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.

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 (1977), Cleveland (1993), Cleveland (1994), Tufte (2006)).



- 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, suavizamiento y bandas de confianza.
- Pequeños múltiplos**

Indicadores de calidad gráfica

Aplicables a cualquier gráfica en particular. Estas son guías concretas y relativamente objetivas para evaluar la calidad de una gráfica (Tufte, 1986).



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 [@cleveland94].

Factor de engaño y Chartjunk

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.

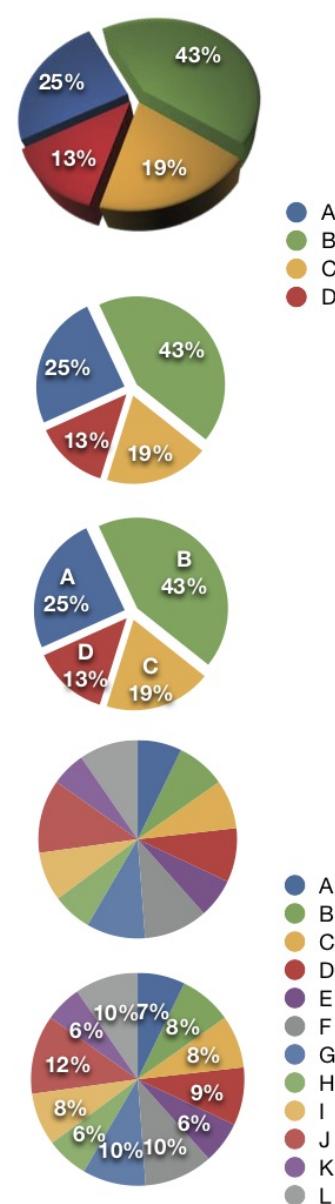
Estos son los indicadores de calidad más fáciles de entender y aplicar, y afortunadamente cada vez son menos comunes.

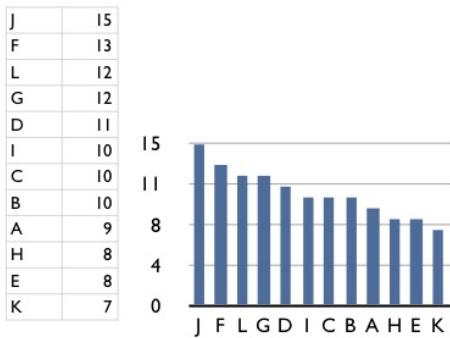
Un diseño popular que califica como chartjunk y además introduce factores de engaño es el *pie* de 3D. En la gráfica de la derecha, podemos ver como la rebanada C se ve más grande que la rebanada A, aunque claramente ese no es el caso (factor de engaño). La razón es la variación en la perspectiva que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

Crítica gráfica: Gráfica de *pie*

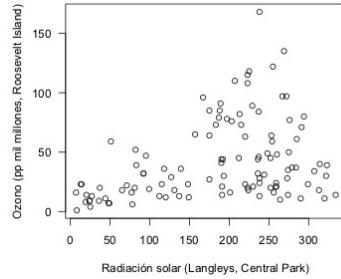
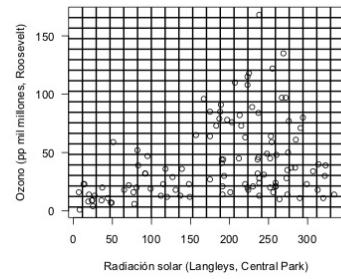
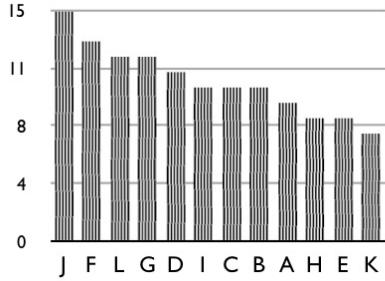
Todavía elementos que pueden mejorar la comprensión de nuestra gráfica de *pie*: se trata de la decodificación que hay que hacer categoría - color - cuantificación. Podemos agregar las etiquetas como se muestra en la serie de la derecha, pero entonces: ¿por qué no mostrar simplemente la tabla de datos? ¿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 izquierda 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.



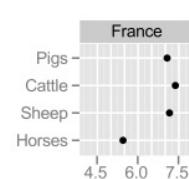
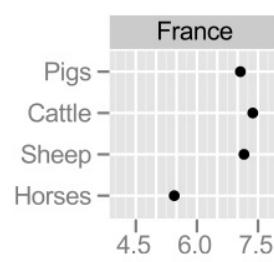
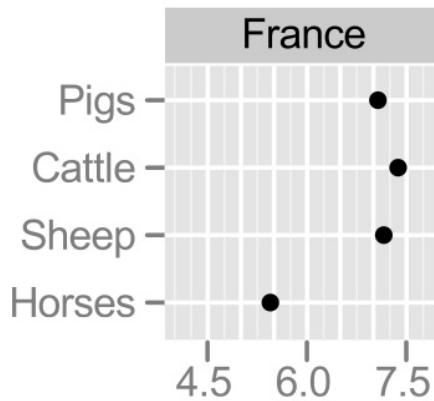


Hay otros tipos de **chartjunk** comunes: uno es la textura de barras, por ejemplo. El efecto es la producción de un efecto moiré que es desagradable y quita la atención de los datos, como en la gráfica de barras de abajo. Otro común son las rejillas, como mostramos en las gráficas de la izquierda. Nótese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

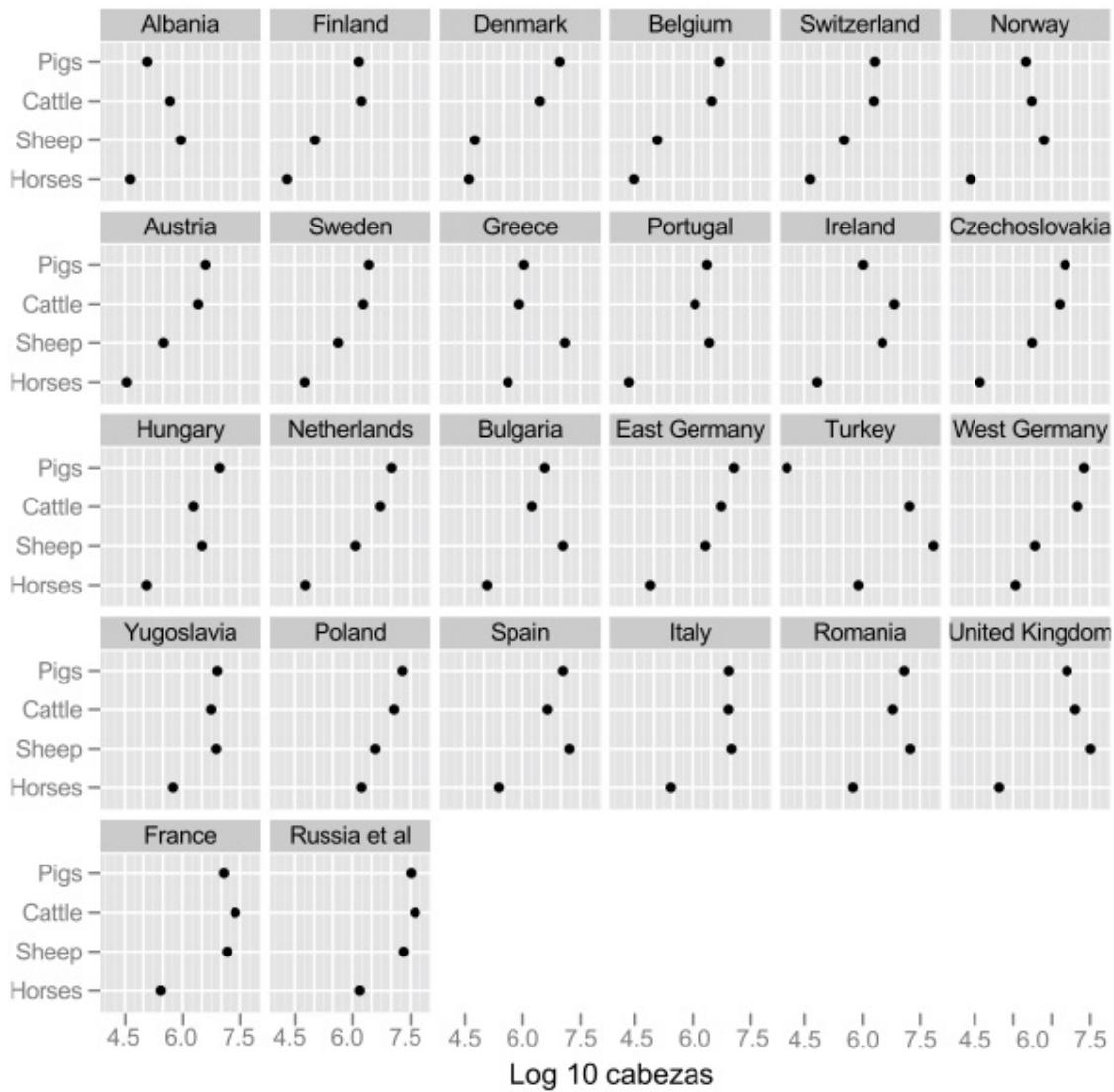


Pequeños múltiplos y densidad gráfica

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:

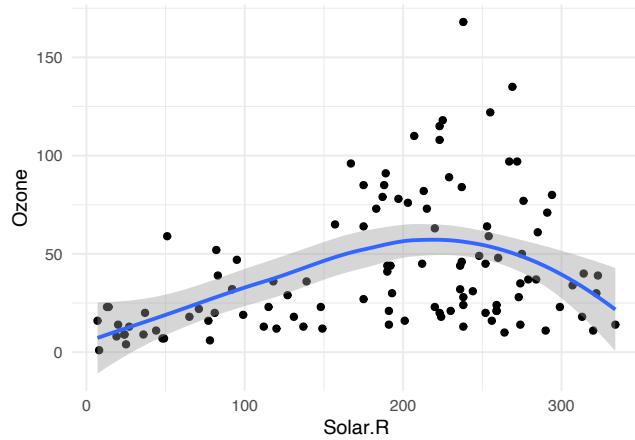


Esta es una gráfica de puntos. Es útil como sustituto de una gráfica de barras, y es superior en el sentido de que una mayor proporción de la tinta que se usa es tinta de datos. Otra vez, mayor proporción de tinta de datos representa más oportunidades que se pueden capitalizar, como muestra la gráfica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

Más pequeños múltiplos

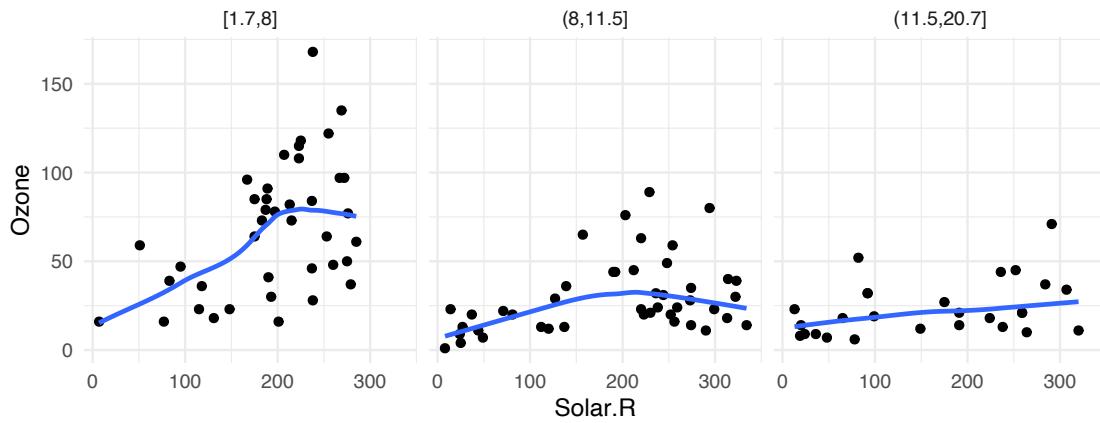
Los pequeños múltiplos presentan oportunidades para mostrar más acerca de nuestro problema de interés. Consideraremos por ejemplo la relación de radiación solar y niveles de ozono:

```
ggplot(airquality, aes(x=Solar.R, y=Ozone)) + geom_point() +
  geom_smooth(method = "loess", span = 1)
```



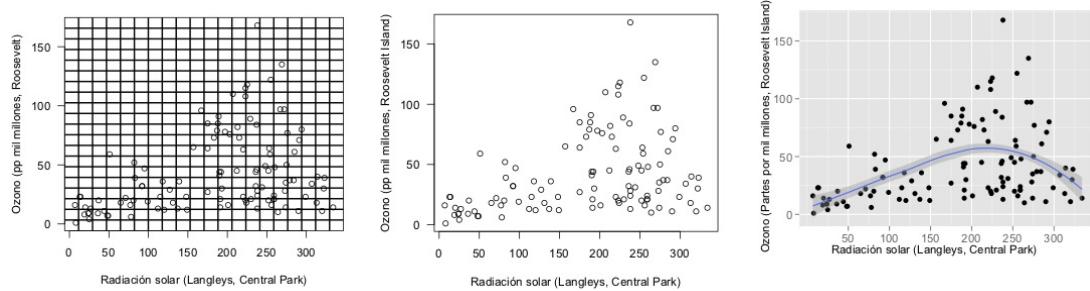
En el ejemplo anterior incluyendo una variable adicional (velocidad del viento) podemos entender más acerca de la relación de radiación solar y niveles de ozono:

```
airquality$Wind_cat <- cut(airquality$Wind,
  breaks = quantile(airquality$Wind, c(0, 1/3, 2/3, 1)),
  include.lowest = TRUE)
ggplot(airquality, aes(x=Solar.R, y=Ozone)) + geom_point() +
  facet_wrap(~Wind_cat) +
  geom_smooth(method = "loess", span = 0.8, se = FALSE,
  method.args = list(degree = 1, family="symmetric"))
```



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 Mieles	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	

Para un ejemplo del proceso de rediseño de una tabla, ver aquí. Finalmente, podemos ver un ejemplo que intenta incorporar los elementos del diseño analítico, incluyendo pequeños múltiplos:

Ejemplo: gráficas, tablas y texto

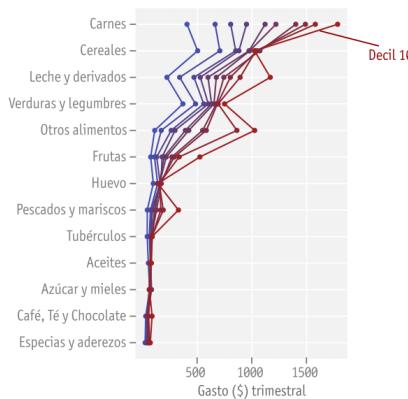
Gasto de hogares en Alimentos y Bebidas (miles de pesos), ENIGH 2006. Hogares agrupados en deciles según ingreso total monetario. Cada decil agrupa unos 2.65 millones de hogares.

	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	1395104	1582291	1783207	1966252	2123150	2360369	3091577	17025176
Huevo	255321	360471	421613	442603	405520	404737	451280	418855	398713	365472	3924585
Acetos Y Grasas	135823	190052	179945	183546	193544	197424	188956	180809	182252	208959	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 Mielas	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	1.9E+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



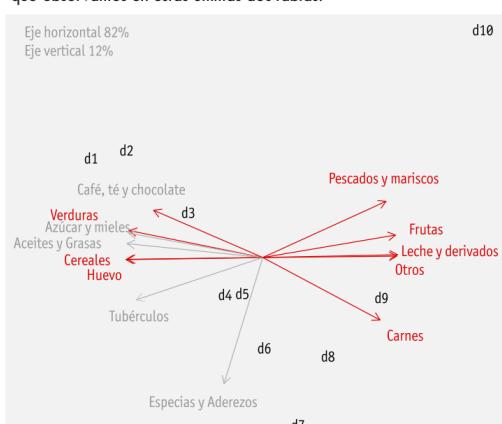
Porcentaje del gasto en cada categoría, por decil.

	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	14.8
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	1.9	1.9	1.7	1.5	1.5	1.4	1.2	1.6	1.6
Aceites Y Grasas	2.5	2.4	1.9	1.7	1.6	1.3	1.2	1.1	1.1	1.5	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	1.1	1.1	1.0	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	38	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 Mielas	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 Mielas	188	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	38	5	-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	

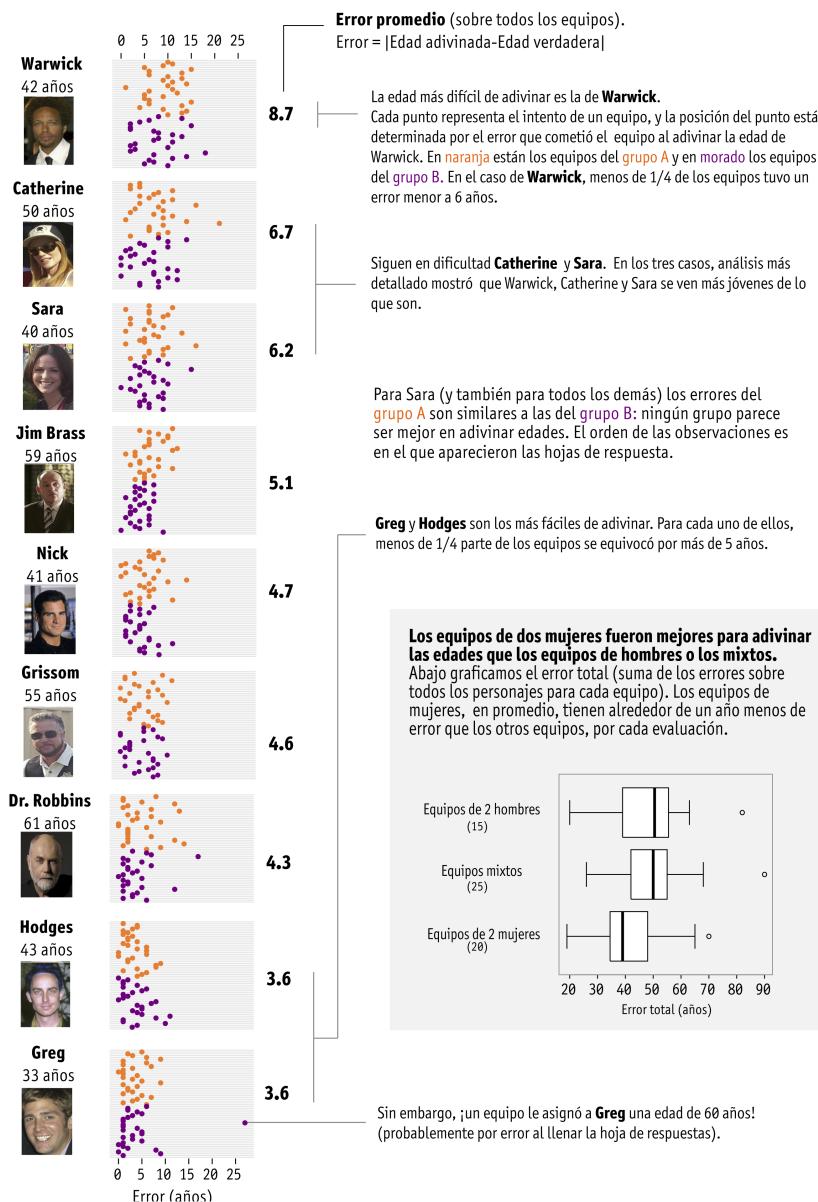
Podemos entender las diferencias de gasto entre los deciles calculando que tanto se aparta cada decil del patrón total de gasto, como en la tabla de la izquierda. Esta última tabla funciona mucho mejor si ordenamos según las diferencias del decil de ingreso más bajo. Finalmente, construimos un biplot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



Decoración

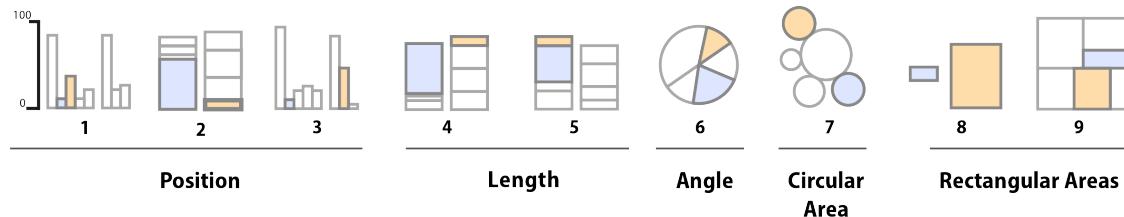
La decoración tiene su lugar

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.

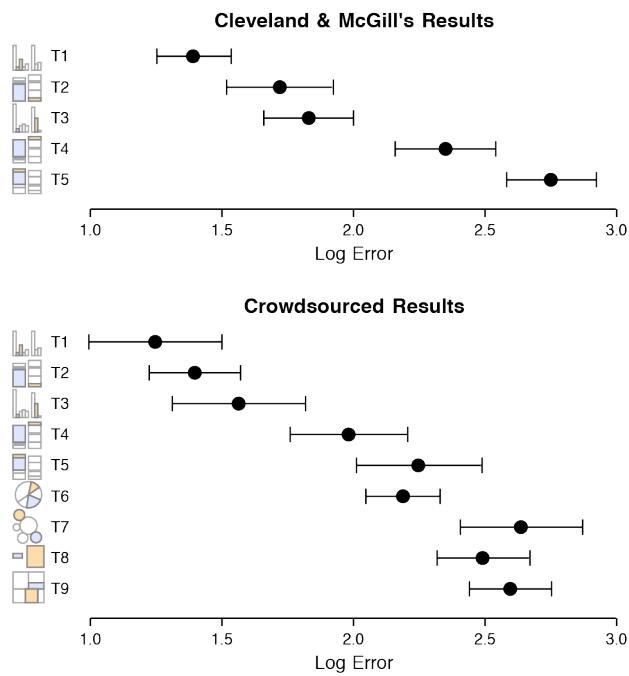


Percepción de escala

Entre la percepción visual y la interpretación de una gráfica están implícitas tareas visuales específicas que las personas debemos realizar para ver correctamente la gráfica. En la década de los ochenta, William S. Cleveland y Robert McGill realizaron algunos experimentos identificando y clasificando estas tareas para diferentes tipos de gráficos (Cleveland and McGill, 1984). En estos, se le pregunta a la persona que compare dos valores dentro de una gráfica, por ejemplo, en dos barras en una gráfica de barras, o dos rebanadas de una gráfica de pie.



Los resultados de Cleveland y McGill fueron replicados por Heer y Bostock en 2010 y los resultados se muestran en las gráficas de la derecha:

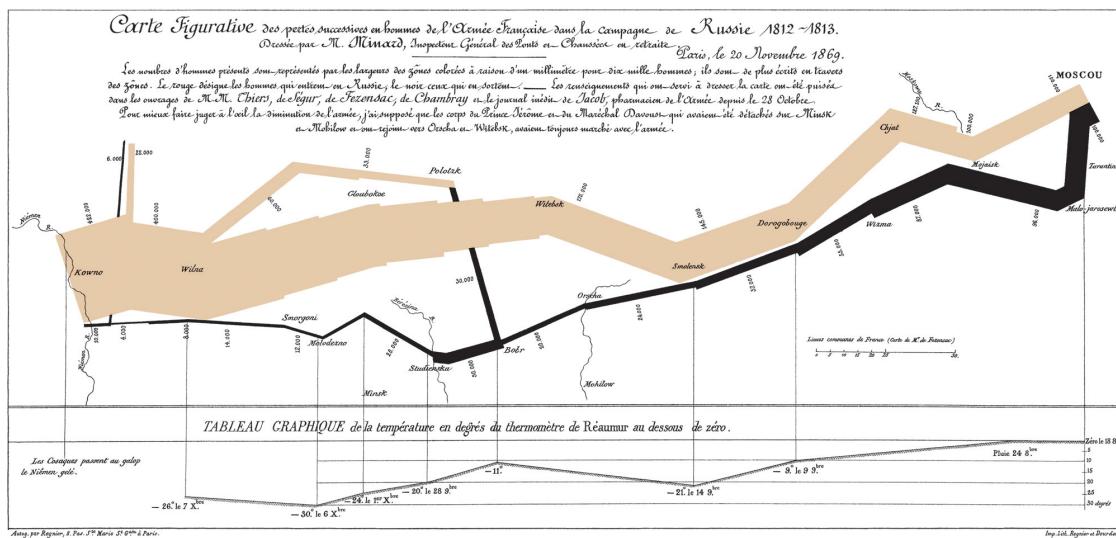


Ejemplo: gráfica de Minard

Concluimos esta sección con una gráfica que, aunque poco común, ejemplifica los principios de una buena gráfica, y es reconocida como una de las mejores visualizaciones de la historia.

Una gráfica excelente, presenta datos interesantes de forma bien diseñada: es una cuestión de fondo, de diseño, y estadística... [Se] compone de ideas complejas comunicadas con claridad, precisión y eficiencia. ... [Es] lo que da al espectador la mayor cantidad de ideas, en el menor tiempo, con la menor cantidad de tinta, y en el espacio más pequeño. ... Es casi siempre multivariado. ... Una excelente gráfica debe decir la verdad acerca de los datos. (Tufte, 1983)

La famosa visualización de Charles Joseph Minard de la marcha de Napoleón sobre Moscú, ilustra los principios de una buena gráfica. Tufte señala que esta imagen “bien podría ser el mejor gráfico estadístico jamás dibujado”, y sostiene que “cuenta una historia rica y coherente con sus datos multivariados, mucho más esclarecedora que un solo número que rebota en el tiempo”. Se representan seis variables: el tamaño del ejército, su ubicación en una superficie bidimensional, la dirección del movimiento del ejército y la temperatura en varias fechas durante la retirada de Moscú”.



Hoy en día Minard es reconocido como uno de los principales contribuyentes a la teoría de análisis de datos y creación de **infografías** con un fundamento estadístico.

Se grafican 6 variables: el número de tropas de Napoleón, la distancia, la temperatura, la latitud y la longitud, la dirección en que viajaban las tropas y la localización relativa a fechas específicas.

La gráfica de Minard, como la describe E.J. Marey, parece “desafiar la pluma del historiador con su brutal elocuencia”, la combinación de datos del mapa, y la serie de tiempo, dibujados en 1869, “retratan una secuencia de pérdidas devastadoras que sufrieron las tropas de Napoleón en 1812”. Comienza en la izquierda, en la frontera de Polonia y Rusia, cerca del río Niemen. La línea gruesa dorada muestra el tamaño de la Gran Armada (422,000) en el momento en que invadía Rusia en junio de 1812.

El ancho de esta banda indica el tamaño de la armada en cada punto del mapa. En septiembre, la armada llegó a Moscú, que ya había sido saqueada y dejada desértica, con sólo 100,000 hombres.

El camino del retiro de Napoleón desde Moscú está representado por la línea oscura (gris) que está en la parte inferior, que está relacionada a su vez con la temperatura y las fechas en el diagrama de abajo. Fue un invierno muy frío, y muchos se congelaron en su salida de Rusia. Como se muestra en el mapa, cruzar el río Berezina fue un desastre, y el ejército de Napoleón logró regresar a Polonia con tan sólo 10,000 hombres.

También se muestran los movimientos de las tropas auxiliares, que buscaban proteger por atrás y por la delantera mientras la armada avanzaba hacia Moscú. La gráfica de Minard cuenta una historia rica y cohesiva, coherente con datos multivariados y con los hechos históricos, y que puede ser más ilustrativa que tan sólo representar un número rebotando a lo largo del tiempo.

Sección 2

Analisis exploratorio

“Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone –as the first step.”

— John Tukey

“The simple graph has brought more information to the data analyst’s mind than any other device.”

— John Tukey

El papel de la exploración en el análisis de datos

El estándar científico para contestar preguntas o tomar decisiones es uno que se basa en el análisis de datos. Es decir, en primer lugar se deben reunir todos los datos disponibles que puedan contener o sugerir alguna guía para entender mejor la pregunta o la decisión a la que nos enfrentamos. Esta recopilación de datos —que pueden ser cualitativos, cuantitativos, o una mezcla de los dos— debe entonces ser analizada para extraer información relevante para nuestro problema.

En análisis de datos existen dos distintos tipos de trabajo:

- El trabajo **exploratorio** o de **detective**: ¿cuáles son los aspectos importantes de estos datos? ¿qué indicaciones generales muestran los datos? ¿qué tareas de análisis debemos empezar haciendo? ¿cuáles son los caminos generales para formular con precisión y contestar algunas preguntas que nos interesen?
- El trabajo **inferencial, confirmatorio**, o de **juez**: ¿cómo evaluar el peso de la evidencia de los descubrimientos del paso anterior? ¿qué tan bien soportadas están las respuestas y conclusiones por nuestro conjunto de datos?

Algunos conceptos básicos

Empezamos explicando algunas ideas que no serán útiles más adelante. Por ejemplo, los siguientes datos fueron registrados en un restaurante durante cuatro días consecutivos:

```

library(tidyverse)
library(patchwork)
source("R/funciones_auxiliares.R")
# usamos los datos tips del paquete reshape2
tips <- reshape2::tips
# renombramos variables y niveles
propinas <- tips %>%
  rename(cuenta_total = total_bill,
         propina = tip, sexo = sex,
         fumador = smoker,
         dia = day, momento = time,
         num_personas = size) %>%
  mutate(sexo = recode(sexo, Female = "Mujer", Male = "Hombre"),
         fumador = recode(fumador, No = "No", Si = "Si"),
         dia = recode(dia, Sun = "Dom", Sat = "Sab", Thur = "Jue", Fri = "Vie"),
         momento = recode(momento, Dinner = "Cena", Lunch = "Comida")) %>%
  select(-sexo) %>%
  mutate(dia = fct_relevel(dia, c("Jue", "Vie", "Sab", "Dom")))

```

Y vemos una muestra

```
sample_n(propinas, 10) %>% formatear_tabla()
```

cuenta_total	propina	fumador	dia	momento	num_personas
23.68	3.31	No	Dom	Cena	2
21.01	3.00	Yes	Vie	Cena	2
20.53	4.00	Yes	Jue	Comida	4
39.42	7.58	No	Sab	Cena	4
18.29	3.76	Yes	Sab	Cena	4
24.59	3.61	No	Dom	Cena	4
12.46	1.50	No	Vie	Cena	2
24.06	3.60	No	Sab	Cena	3
11.59	1.50	Yes	Sab	Cena	2
9.60	4.00	Yes	Dom	Cena	2

Aquí la unidad de observación es una cuenta particular. Tenemos tres mediciones numéricas de cada cuenta: cuánto fue la cuenta total, la propina, y el número de personas asociadas a la cuenta. Los datos están separados según se fumó o no en la mesa, y temporalmente en dos partes: el día (Jueves, Viernes, Sábado o Domingo), cada uno separado por Cena y Comida.



Denotamos por x el valor de medición de una *unidad de observación*. Usualmente utilizamos subíndices para identificar entre diferentes *puntos de datos* (observaciones), por ejemplo, x_n para la n -ésima observación. De tal forma que una colección de N observaciones la escribimos como

$$\{x_1, \dots, x_N\}. \quad (2.1)$$

El primer tipo de comparaciones que nos interesa hacer es para una medición: ¿Varían mucho o poco los datos de un tipo de medición? ¿Cuáles son valores típicos o centrales? ¿Existen valores atípicos?

Supongamos entonces que consideraremos simplemente la variable de `cuenta_total`. Podemos comenzar por **ordenar los datos**, y ver cuáles datos están en los extremos y cuáles están en los lugares centrales:



En general la colección de datos no está ordenada por sus valores. Esto es debido a que las observaciones en general se recopilan de manera *aleatoria*. Utilizamos la notación de $\sigma(n)$ para denotar un *reordenamiento* de los datos de tal forma

$$\{x_{\sigma(1)}, \dots, x_{\sigma(N)}\}, \quad (2.2)$$

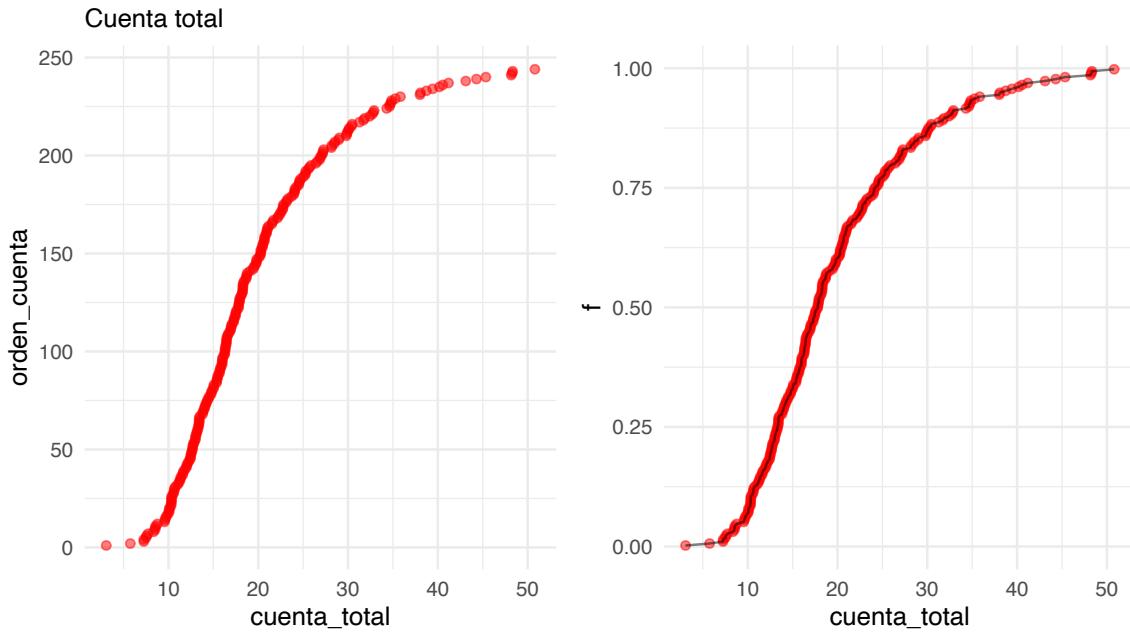
y que satisface la siguiente serie de desigualdades

$$x_{\sigma(1)} \leq \dots \leq x_{\sigma(N)}. \quad (2.3)$$

```
propinas <- propinas %>%
  mutate(orden_cuenta = rank(cuenta_total, ties.method = "first"),
        f = (orden_cuenta - 0.5) / n())
cuenta <- propinas %>% select(orden_cuenta, f, cuenta_total) %>% arrange(f)
bind_rows(head(cuenta), tail(cuenta)) %>% formatear_tabla()
```

orden_cuenta	f	cuenta_total
1	0.0020492	3.07
2	0.0061475	5.75
3	0.0102459	7.25
4	0.0143443	7.25
5	0.0184426	7.51
6	0.0225410	7.56
239	0.9774590	44.30
240	0.9815574	45.35
241	0.9856557	48.17
242	0.9897541	48.27
243	0.9938525	48.33
244	0.9979508	50.81

También podemos graficar los datos en orden, interpolando valores consecutivos.



A esta función le llamamos la **función de cuantiles** para la variable `cuenta_total`. Nos sirve para comparar directamente los distintos valores que observamos los datos según el orden que ocupan.



La función de cuantiles muestral esta definida por

$$\hat{F}(x) = \frac{1}{N} \sum_{n=1}^N 1\{x_n \leq x\}, \quad (2.4)$$

donde la función indicadora está definida por

$$1\{x \leq t\} = \begin{cases} 1, & \text{si } x \leq t \\ 0, & \text{en otro caso} \end{cases}. \quad (2.5)$$



Observación: la función de cuantiles definida arriba también es conocida como la *función de acumulación empírica*. Se puede encontrar la siguiente notación en la literatura

$$\hat{F}(x) = F_N(x) = \Pr_N(X \leq x), \quad (2.6)$$

así como

$$\Pr_N(X \geq x) = 1 - \hat{F}(x). \quad (2.7)$$



Observación: Para mediciones continuas y $0 \leq q \leq 1$, el cuantil q es el valor $x = x(q)$ —esta notación sirve para definir x como una función de q —tal que

$$\hat{F}(x) \geq q \quad \text{y} \quad 1 - \hat{F}(x) \geq 1 - q \quad (2.8)$$

Es decir, x acumula el $q\%$ de los casos.



Observación: Si $N \times q$ no es un número entero, entonces el cuantil q es único y es igual a la $\sigma(\lceil Nq \rceil)$ -ésima observación.

Si Nq es un número entero, entonces el cuantil q no es único y es igual a cualquier x tal que

$$x_{\sigma(Nq)} \leq x \leq x_{\sigma(Nq+1)}. \quad (2.9)$$



Para una medición de interés x con posibles valores en el intervalo $[a, b]$. Comprueba que $\hat{F}(a) = 0$ y $\hat{F}(b) = 1$ para cualquier colección de datos de tamaño N .

La gráfica anterior, también nos sirve para poder estudiar la **dispersión y valores centrales** de los datos observados. Por ejemplo, podemos notar que:

- El **rango** de datos va de unos 3 dólares hasta 50 dólares
- Los **valores centrales** —del cuantil 0.25 al 0.75, por decir un ejemplo— están entre unos 13 y 25 dólares
- El cuantil 0.5 (o también conocido como **mediana**) está alrededor de 18 dólares.



¿Cómo definirías la mediana en términos de la función de cuantiles? *Pista:* Considera los casos por separado para N impar o par.

Éste último puede ser utilizado para dar un valor *central* de la distribución de valores para `cuenta_total`. Asimismo podemos dar resúmenes más refinados si es necesario. Por ejemplo, podemos reportar que:

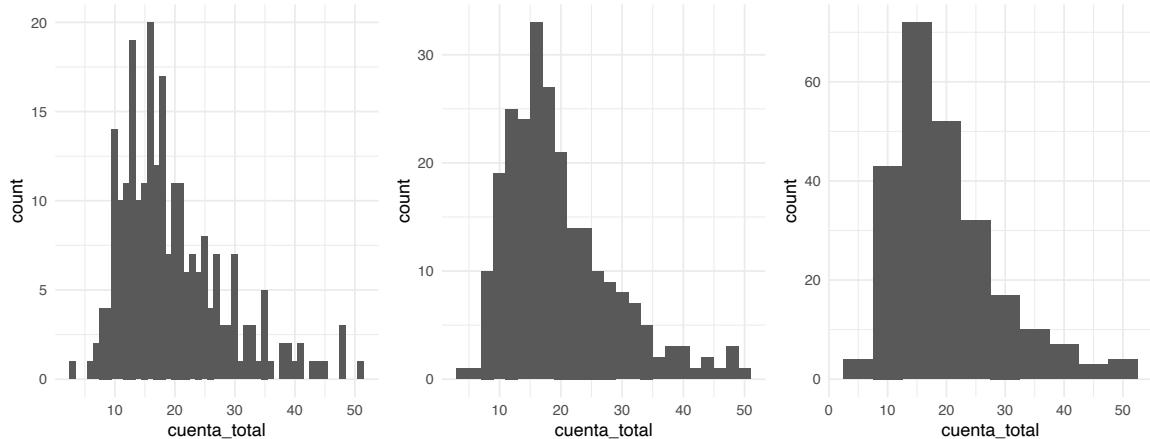
- El cuantil 0.95 es de unos 35 dólares — sólo 5% de las cuentas son de más de 35 dólares
- El cuantil 0.05 es de unos 8 dólares — sólo 5% de las cuentas son de 8 dólares o menos.

Finalmente, la forma de la gráfica se interpreta usando su pendiente (tasa de cambio) haciendo comparaciones en diferentes partes de la gráfica:

- La distribución de valores tiene asimetría: el 10% de las cuentas más altas tiene considerablemente más dispersión que el 10% de las cuentas más bajas.

- Entre los cuantiles 0.2 y 0.5 es donde existe *mayor* densidad de datos: la pendiente (tasa de cambio) es alta, lo que significa que al avanzar en los valores observados, los cuantiles (el porcentaje de casos) aumenta rápidamente.
- Cuando la pendiente es casi plana, quiere decir que los datos tienen más dispersión local o están más separados.

En algunos casos, es más natural hacer un **histograma**, donde dividimos el rango de la variable en cubetas o intervalos (en este caso de igual longitud), y graficamos por medio de barras cuántos datos caen en cada cubeta:

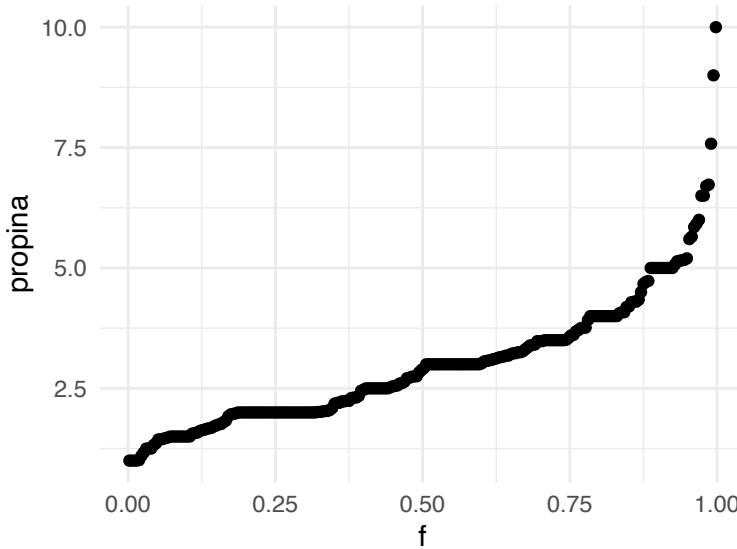


Es una gráfica más popular, pero perdemos cierto nivel de detalle, y distintas particiones resaltan distintos aspectos de los datos.



¿Cómo se ve la gráfica de cuantiles de las propinas? ¿Cómo crees que esta gráfica se compara con distintos histogramas?

```
g_1 <- ggplot(propinas, aes(sample = propina)) +
  geom_qq(distribution = stats::qunif) + xlab("f") + ylab("propina")
g_1
```



Observación. Cuando hay datos repetidos, los cuantiles tienen que interpretarse como sigue: el cuantil f con valor q satisface que existe una proporción aproximada f de los datos que están en el valor q o por debajo de éste, pero no necesariamente exactamente una proporción f de los datos estan en q o por debajo.

Finalmente, una gráfica más compacta que resume la gráfica de cuantiles o el histograma es el **diagrama de caja y brazos**. Mostramos dos versiones, la clásica de Tukey (T) y otra versión menos común de Spear/Tufte (ST):

```
library(ggthemes)
cuartiles <- quantile(cuenta$cuenta_total)
t(cuartiles) %>% formatear_tabla()
```

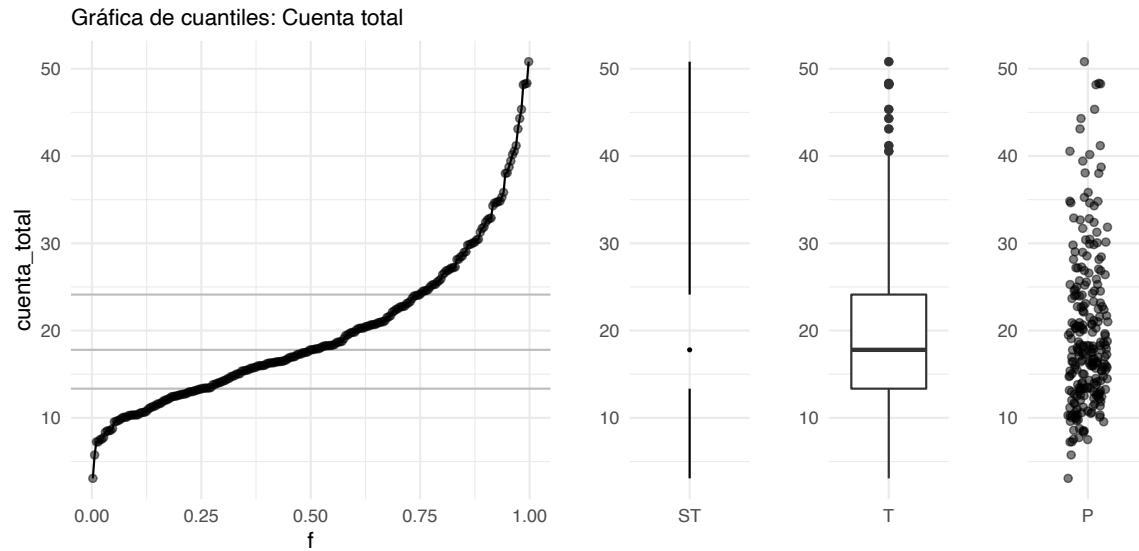
	0 %	25 %	50 %	75 %	100 %
	3.07	13.3475	17.795	24.1275	50.81

```
g_1 <- ggplot(cuenta, aes(x = f, y = cuenta_total)) +
  labs(subtitle = "Gráfica de cuantiles: Cuenta total") +
  geom_hline(yintercept = cuartiles[2], colour = "gray") +
  geom_hline(yintercept = cuartiles[3], colour = "gray") +
  geom_hline(yintercept = cuartiles[4], colour = "gray") +
  geom_point(alpha = 0.5) + geom_line()
g_2 <- ggplot(cuenta, aes(x = factor("ST", levels =c("ST")), y = cuenta_total)) +
  geom_tufteboxplot() +
  labs(subtitle = " ") + xlab("") + ylab("")
g_3 <- ggplot(cuenta, aes(x = factor("T"), y = cuenta_total)) +
  geom_boxplot() +
```

```

  labs(subtitle = " ") + xlab("") + ylab("")
g_4 <- ggplot(cuenta, aes(x = factor("P"), y = cuenta_total)) +
  geom_jitter(height = 0, width = 0.2, alpha = 0.5) +
  labs(subtitle = " ") + xlab("") + ylab("")
g_5 <- ggplot(cuenta, aes(x = factor("V"), y = cuenta_total)) +
  geom_violin() +
  labs(subtitle = " ") + xlab("") + ylab("")
g_1 + g_2 + g_3 + g_4 +
  plot_layout(widths = c(8, 2, 2, 2))

```



El diagrama de la derecha explica los elementos de la versión típica del diagrama de caja y brazos (*box-plot*). **RIC** se refiere al **Rango Intercuantílico**, definido por la diferencia entre los cuantiles 25 % y 75 %.

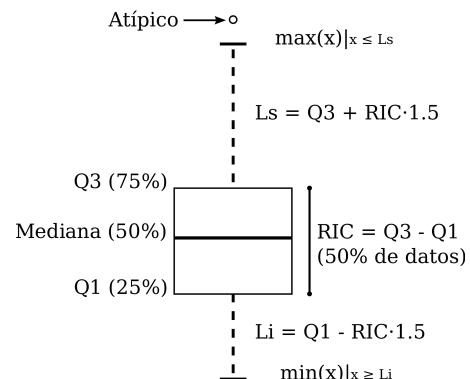


Figura: Jumanbar / CC BY-SA



Hasta ahora hemos utilizado la definición general de *cuantiles*. Donde consideramos el cuantil q , para buscar x tal que $\hat{F}(x) = q$. Hay valores típicos de interés que corresponden a q igual a 25 %, 50 % y 75 %. Éstos valores se denominan **cuartiles**.

Ventajas en el análisis inicial

En un principio del análisis, estos resúmenes (cuantiles) pueden ser más útiles que utilizar medias y varianzas, por ejemplo. La razón es que los cuantiles:

- Son cantidades más fácilmente interpretables
- Los cuantiles centrales son más resistentes a valores atípicos que medias o varianzas
- Sin embargo, permite identificar valores extremos
- Es fácil comparar cuantiles de distintos conjuntos de datos

Media y desviación estándar

Las medidas más comunes de localización y dispersión para un conjunto de datos son la media muestral y la desviación estándar muestral.

En general, no son muy apropiadas para iniciar el análisis exploratorio, pues:

- Son medidas más difíciles de interpretar y explicar que los cuantiles. En este sentido, son medidas especializadas. Por ejemplo, intenta explicar intuitivamente qué es la media.
- No son resistentes a valores atípicos o erróneos. Su falta de resistencia los vuelve poco útiles en las primeras etapas de limpieza y descripción.



La media, o promedio, se denota por \bar{x} y se define como

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (2.10)$$

La desviación estándar muestral se define como

$$\text{std}(x) = \sqrt{\frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2}. \quad (2.11)$$

Sin embargo,

- La media y desviación estándar son computacionalmente convenientes.
- Para el trabajo de modelado estas medidas de resumen tienen ventajas claras (bajo ciertos supuestos teóricos).
- En muchas ocasiones conviene usar estas medidas pues permite hacer comparaciones históricas o tradicionales —pues análisis anteriores pudieran estar basados en éstas.



1. Considera el caso de tener N observaciones y asume que ya tienes calculado el promedio para dichas observaciones. Este promedio lo denotaremos por \bar{x}_N . Ahora, considera que has obtenido M observaciones más. Escribe una fórmula recursiva para la media del conjunto total de datos \bar{x}_{N+M} en función de lo que ya tenías precalculado \bar{x}_N .
2. ¿En qué situaciones esta propiedad puede ser conveniente?

Ejemplos

Precios de casas

En este ejemplo consideremos los datos de precios de ventas de la ciudad de Ames, Iowa. En particular nos interesa entender la variación del precio de las casas.

Por este motivo calculamos los cuantiles que corresponden al 25 %, 50 % y 75 % (**cuartiles**), así como el mínimo y máximo de los precios de las casas:

```
quantile(casas %>% pull(precio_miles))
```

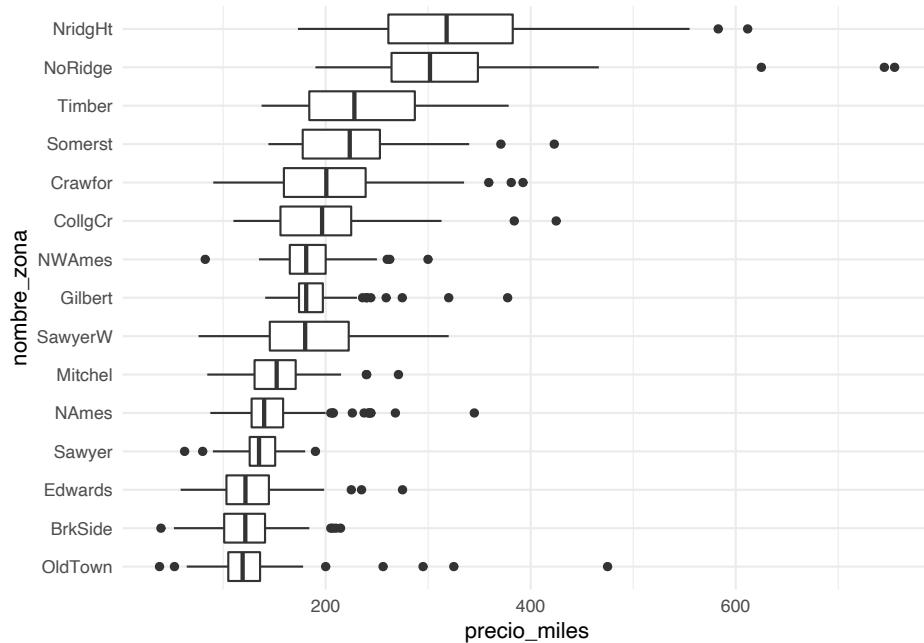
```
##      0%    25%    50%    75%   100%
## 37.9 132.0 165.0 215.0 755.0
```



Comprueba que el mínimo y máximo están asociados a los cuantiles 0 % y 100 %, respectivamente.

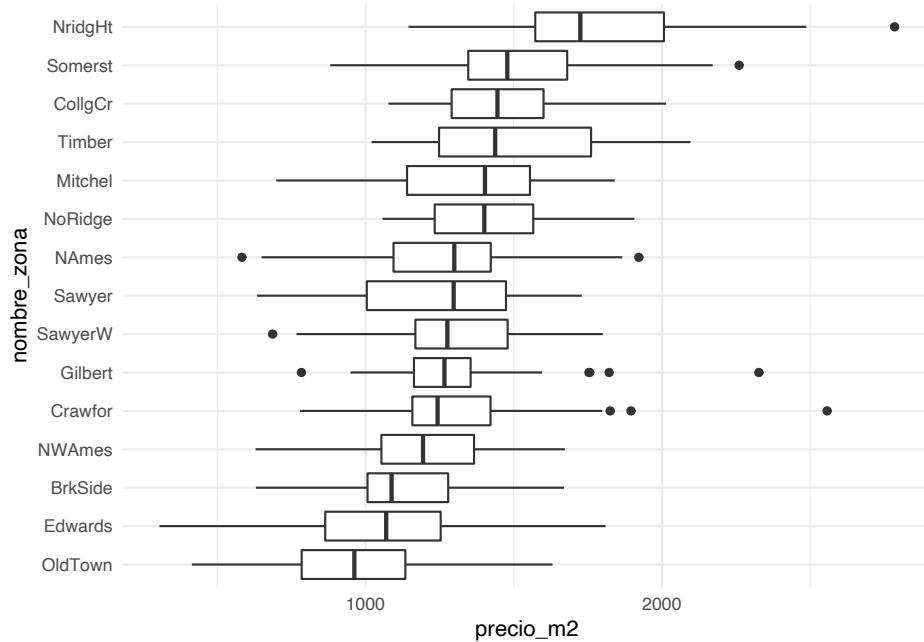
Una posible comparación es considerar los precios y sus variación en función de zona de la ciudad en que se encuentra una vivienda. Podemos usar diagramas de caja y brazos para hacer una **comparación burda** de los precios en distintas zonas de la ciudad:

```
ggplot(casas, aes(x = nombre_zona, y = precio_miles)) +
  geom_boxplot() +
  coord_flip()
```



La primera pregunta que nos hacemos es cómo pueden variar las características de las casas dentro de cada zona. Para esto, podemos considerar el área de las casas. En lugar de graficar el precio, graficamos el precio por metro cuadrado, por ejemplo:

```
ggplot(casas, aes(x = nombre_zona, y = precio_m2)) +
  geom_boxplot() +
  coord_flip()
```



Podemos cuantificar la variación que observamos de zona a zona y la variación que hay dentro de cada una de las zonas. Una primera aproximación es observar las variación del precio al calcular la mediana dentro de cada zona, y después cuantificar por medio de cuantiles cómo varía la mediana entre zonas:

```
casas %>%
  group_by(nombre_zona) %>%
  summarise(mediana_zona = median(precio_m2), .groups = "drop") %>%
  pull(mediana_zona) %>%
  quantile() %>%
  round()
```

```
##   0%  25%  50%  75% 100%
##  963 1219 1298 1420 1725
```



Tratar con datos por segmento es una situación común en aplicaciones. Usualmente denotamos por

$$x_{k,n} \quad (2.12)$$

a la n -ésima observación del k -ésimo grupo. Usualmente tenemos un universo de K posibles grupos y para cada grupo tenemos un total diferente de observaciones. Esto lo denotamos por N_k , el número total de observaciones del grupo k para cualquier $k = 1, \dots, K$. El número total de muestras lo denotamos por N , donde

$$N = \sum_{k=1}^K N_k. \quad (2.13)$$

Finalmente, nos puede interesar, como en el ejemplo, los promedios por grupo

$$\bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} x_{k,n}, \quad (2.14)$$

y contrastar contra el promedio total

$$\bar{x} = \frac{1}{K} \sum_{k=1}^K \bar{x}_k = \frac{1}{K} \sum_{k=1}^K \left(\frac{1}{N_k} \sum_{n=1}^{N_k} x_{k,n} \right). \quad (2.15)$$

Por otro lado, las variaciones con respecto a las medianas **dentro** de cada zona, por grupo, se resume como:

```
quantile(casas %>% group_by(nombre_zona) %>%
  mutate(residual = precio_m2 - median(precio_m2)) %>%
  pull(residual)) %>%
  round()

##    0%   25%   50%   75% 100%
## -765  -166     0  172 1314
```

Nótese que este último paso tiene sentido pues la variación dentro de las zonas, en términos de precio por metro cuadrado, es similar. Esto no lo podríamos haber hecho de manera efectiva si se hubiera utilizado el precio de las casas sin ajustar por su tamaño.

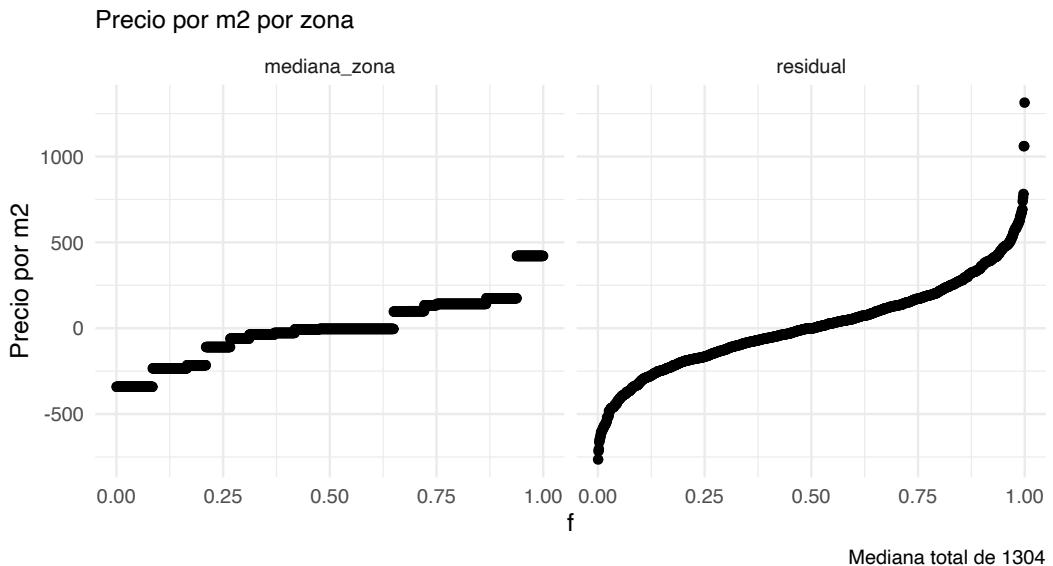
Podemos resumir este primer análisis con un par de gráficas de cuantiles (Cleveland (1993)):

```
mediana <- median(casas$precio_m2)
resumen <- casas %>%
  group_by(nombre_zona) %>%
  mutate(mediana_zona = median(precio_m2)) %>%
  mutate(residual = precio_m2 - mediana_zona) %>%
```

```

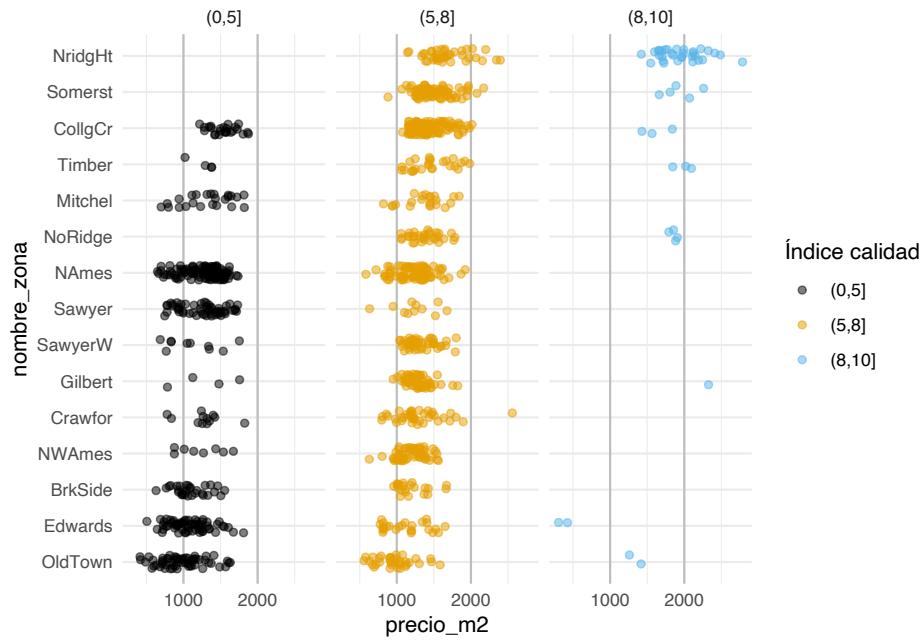
ungroup() %>%
  mutate(mediana_zona = mediana_zona - mediana) %>%
  select(nombre_zona, mediana_zona, residual) %>%
  pivot_longer(mediana_zona:residual, names_to = "tipo", values_to = "valor")
ggplot(resumen, aes(sample = valor)) +
  geom_qq(distribution = stats::qunif) +
  facet_wrap(~ tipo) +
  ylab("Precio por m2") + xlab("f") +
  labs(subtitle = "Precio por m2 por zona",
       caption = paste0("Mediana total de ", round(mediana)))

```



Vemos que la mayor parte de la variación del precio por metro cuadrado ocurre dentro de cada zona, una vez que controlamos por el tamaño de las casas. La variación dentro de cada zona es aproximadamente simétrica, aunque la cola derecha es ligeramente más larga con algunos valores extremos.

Podemos seguir con otro indicador importante: la calificación de calidad de los terminados de las casas. Como primer intento podríamos hacer:



Lo que indica que las calificaciones de calidad están distribuidas de manera muy distinta a lo largo de las zonas, y que probablemente no va ser simple desentrañar qué variación del precio se debe a la zona y cuál se debe a la calidad.

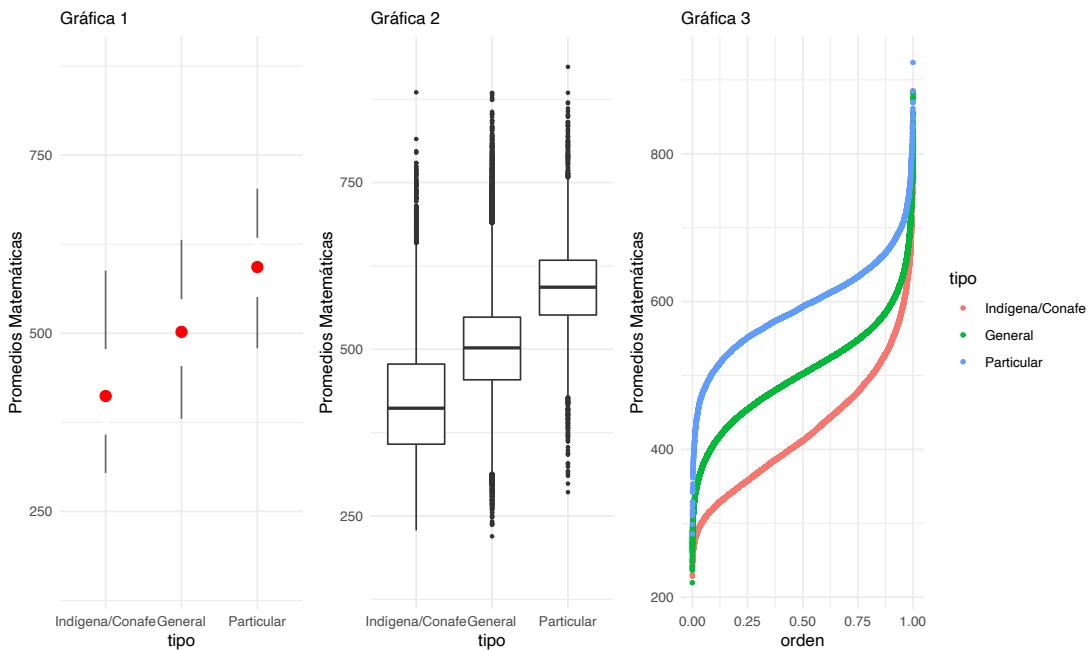
Prueba Enlace

Consideremos la prueba Enlace (2011) de matemáticas para primarias. Una primera pregunta que alguien podría hacerse es: ¿cuáles escuelas son mejores en este rubro, las privadas o las públicas?

```
enlace_tbl <- enlace %>% group_by(tipo) %>%
  summarise(n_escuelas = n(),
            cuantiles = list(cuantil(mate_6, c(0.05, 0.25, 0.5, 0.75, 0.95))) %>%
  unnest(cols = cuantiles) %>% mutate(valor = round(valor))
enlace_tbl %>%
  spread(cuantil, valor) %>%
  formatear_tabla()
```

tipo	n_escuelas	0.05	0.25	0.5	0.75	0.95
Indígena/Conafe	13599	304	358	412	478	588
General	60166	380	454	502	548	631
Particular	6816	479	551	593	634	703

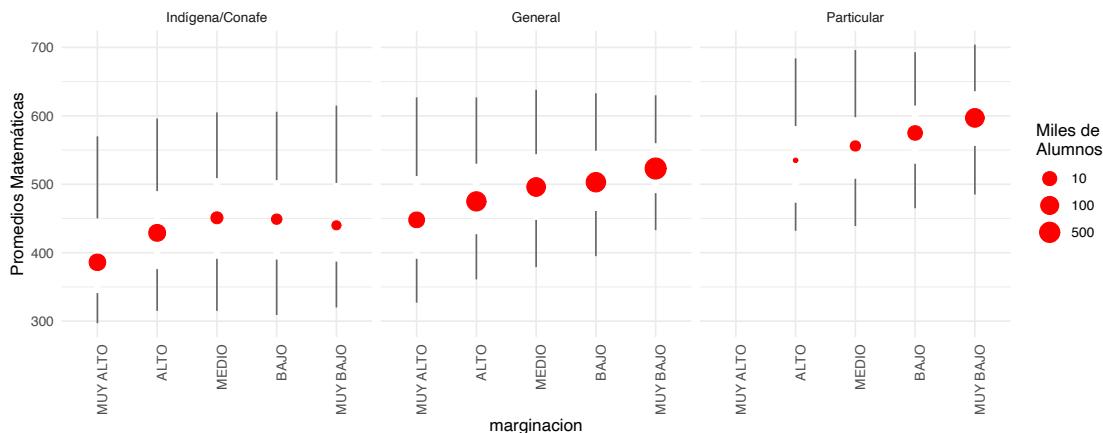
Para un análisis exploratorio podemos utilizar distintas gráficas. Por ejemplo, podemos utilizar nuevamente las gráficas de caja y brazos, así como graficar los percentiles. Nótese que en la gráfica 1 se utilizan los cuantiles 0.05, 0.25, 0.5, 0.75 y 0.95:



Se puede discutir qué tan apropiada es cada gráfica con el objetivo de realizar comparaciones. Sin duda, graficar más cuantiles es más útil para hacer comparaciones. Por ejemplo, en la Gráfica 1 podemos ver que la mediana de las escuelas generales está cercana al cuantil 5 % de las escuelas particulares. Por otro lado, el diagrama de caja y brazos muestra también valores “atípicos”. Es importante notar que una comparación más robusta se puede lograr por medio de **pruebas de hipótesis**, las cuales veremos mas adelante en el curso.

Regresando a nuestro análisis exploratorio, notemos que la diferencia es considerable entre tipos de escuela. Antes de contestar prematuramente la pregunta: ¿cuáles son las mejores escuelas? busquemos mejorar la interpretabilidad de nuestras comparaciones usando los principios 2 y 3. Podemos comenzar por agregar, por ejemplo, el nivel del marginación del municipio donde se encuentra la escuela.

Para este objetivo, podemos usar paneles (pequeños múltiplos útiles para hacer comparaciones) y graficar:



Esta gráfica pone en contexto la pregunta inicial, y permite evidenciar la dificultad de contestarla. En particular:

1. Señala que la pregunta no sólo debe concentrarse en el tipo de “sistema”: pública, privada, etc. Por ejemplo, las escuelas públicas en zonas de marginación baja no tienen una distribución de calificaciones muy distinta a las privadas en zonas de marginación alta.
2. El contexto de la escuela es importante.
3. Debemos de pensar qué factores –por ejemplo, el entorno familiar de los estudiantes– puede resultar en comparaciones que favorecen a las escuelas privadas. Un ejemplo de esto es considerar si los estudiantes tienen que trabajar o no. A su vez, esto puede o no ser reflejo de la calidad del sistema educativo.
4. Si esto es cierto, entonces la pregunta inicial es demasiado vaga y mal planteada. Quizá deberíamos intentar entender cuánto “aporta” cada escuela a cada estudiante, como medida de qué tan buena es cada escuela.

Estados y calificaciones en SAT

¿Cómo se relaciona el gasto por alumno, a nivel estatal, con sus resultados académicos? Hay trabajo considerable en definir estos términos, pero supongamos que tenemos el siguiente conjunto de datos (Guber, 1999), que son datos oficiales agregados por `estado` de Estados Unidos. Consideraremos el subconjunto de variables `sat`, que es la calificación promedio de los alumnos en cada estado (para 1997) y `expend`, que es el gasto en miles de dólares por estudiante en (1994-1995).

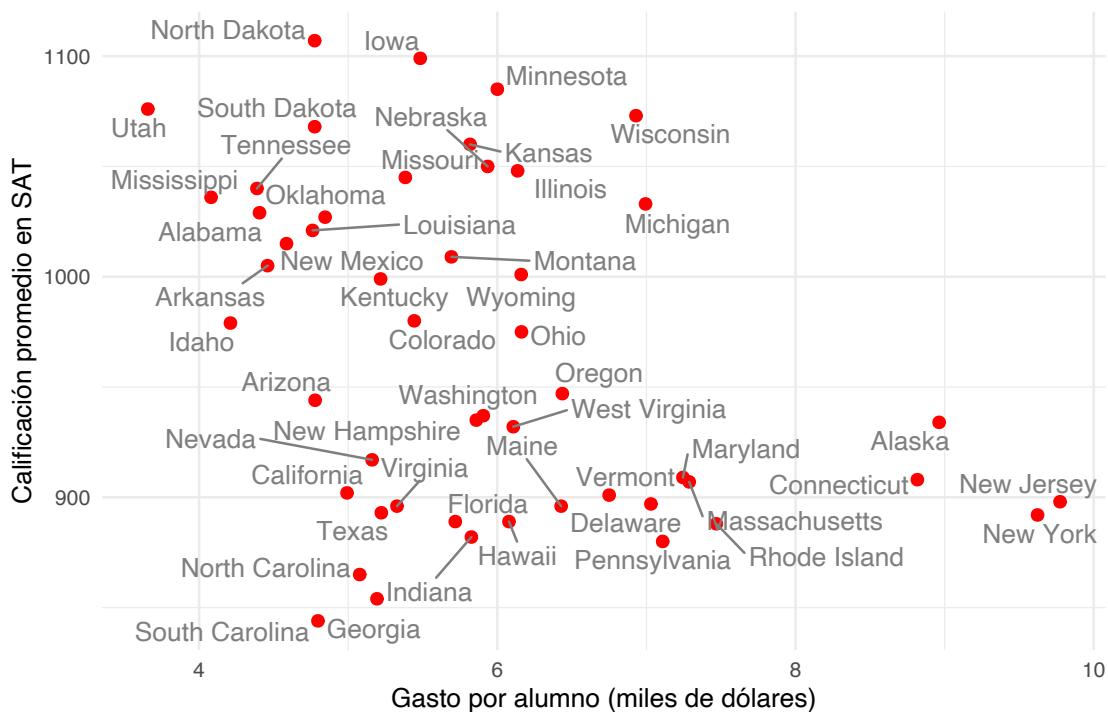
```
sat <- read_csv("data/sat.csv")
sat_tbl <- sat %>% select(state, expend, sat) %>%
  gather(variable, valor, expend:sat) %>%
  group_by(variable) %>%
  summarise(cuantiles = list(cuantil(valor))) %>%
  unnest(cols = c(cuantiles)) %>%
  mutate(valor = round(valor, 1)) %>%
  spread(cuantil, valor)
sat_tbl %>% formatear_tabla
```

variable	0	0.25	0.5	0.75	1
expend	3.7	4.9	5.8	6.4	9.8
sat	844.0	897.2	945.5	1032.0	1107.0

Esta variación es considerable para promedios del SAT: el percentil 75 es alrededor de 1050 puntos, mientras que el percentil 25 corresponde a alrededor de 800. Igualmente, hay diferencias considerables de gasto por alumno (miles de dólares) a lo largo de los estados.

Ahora hacemos nuestro primer ejercicio de comparación: ¿Cómo se ven las calificaciones para estados en distintos niveles de gasto? Podemos usar una gráfica de dispersión:

```
library(ggrepel)
ggplot(sat, aes(x = expend, y = sat, label = state)) +
  geom_point(colour = "red", size = 2) + geom_text_repel(colour = "gray50") +
  xlab("Gasto por alumno (miles de dólares)") +
  ylab("Calificación promedio en SAT")
```

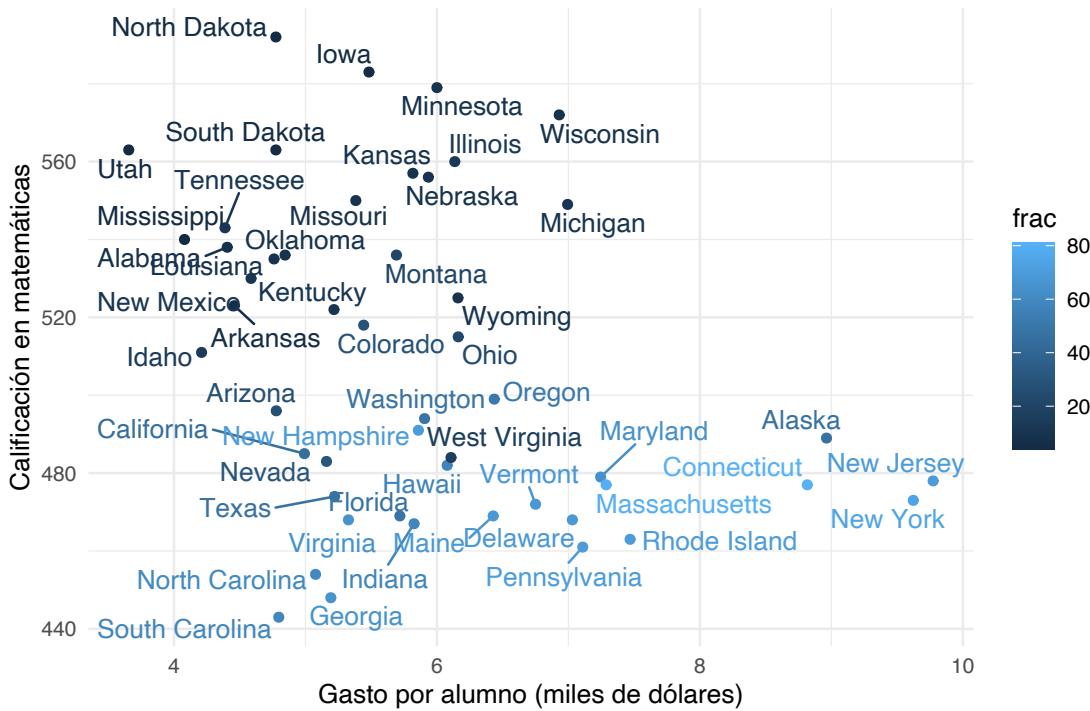


Estas comparaciones no son de alta calidad, solo estamos usando 2 variables —que son muy pocas— y no hay mucho que podamos decir en cuanto explicación. Sin duda nos hace falta una imagen más completa. Necesitaríamos entender la correlación que existe entre las demás características de nuestras unidades de estudio.

Las unidades que estamos comparando pueden diferir fuertemente en otras propiedades importantes (*aka*, dimensiones), lo cual no permite interpretar la gráfica de manera sencilla.

Sabemos que es posible que el IQ difiera en los estados. Pero no sabemos cómo producir diferencias de este tipo. Sin embargo, ¡descubrimos que existe una variable adicional! Ésta es el porcentaje de alumnos de cada estado que toma el SAT. Podemos agregar como sigue:

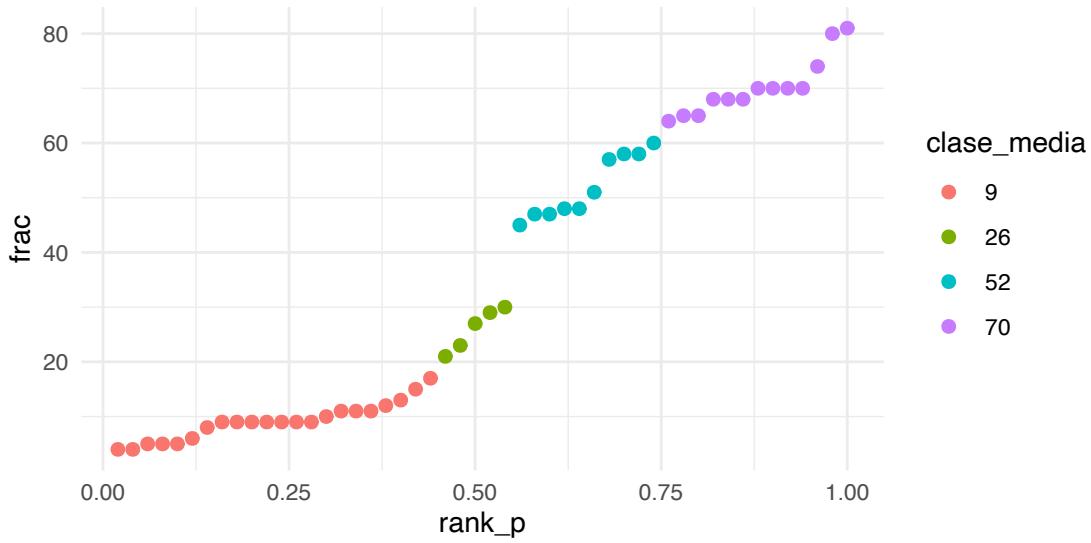
```
ggplot(sat, aes(x = expend, y = math, label=state, colour = frac)) +
  geom_point() + geom_text_repel() +
  xlab("Gasto por alumno (miles de dólares)") +
  ylab("Calificación en matemáticas")
```



Esto nos permite entender por qué nuestra comparación inicial es relativamente pobre. Los estados con mejores resultados promedio en el SAT son aquellos donde una fracción relativamente baja de los estudiantes toma el examen. La diferencia es considerable.

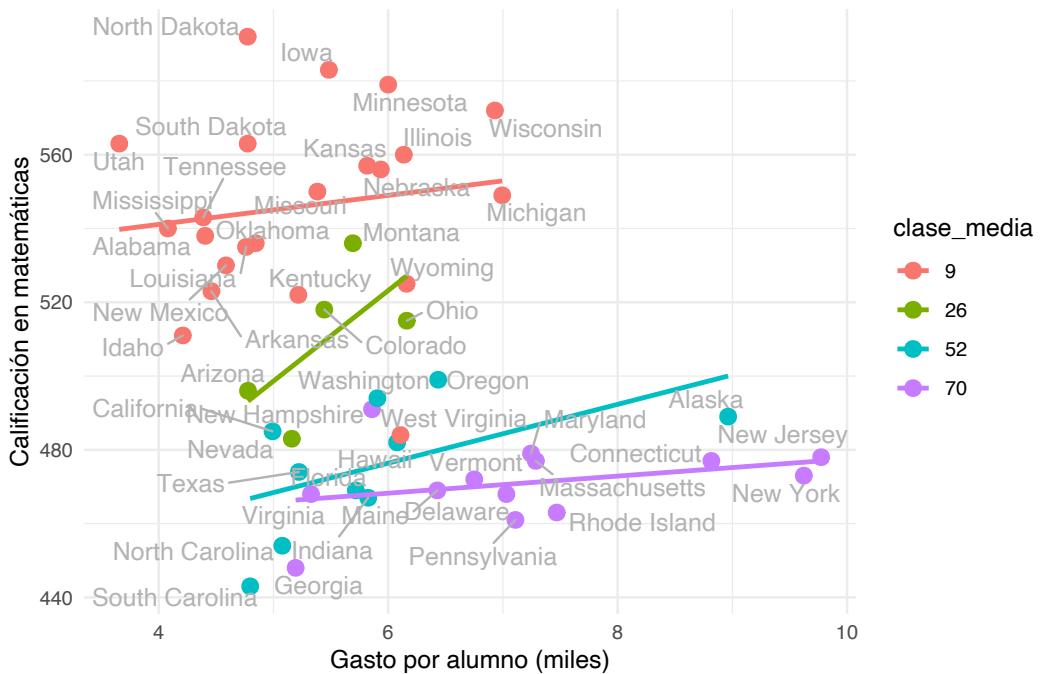
En este punto podemos hacer varias cosas. Una primera idea es intentar comparar estados más similares en cuanto a la población de alumnos que asiste. Podríamos hacer grupos como sigue:

```
set.seed(991)
k_medias_sat <- kmeans(sat %>% select(frac), centers = 4, nstart = 100, iter.max = 100)
sat$clase <- k_medias_sat$cluster
sat <- sat %>% group_by(clase) %>%
  mutate(clase_media = round(mean(frac))) %>%
  ungroup %>%
  mutate(clase_media = factor(clase_media))
sat <- sat %>%
  mutate(rank_p = rank(frac, ties= "first") / length(frac))
ggplot(sat, aes(x = rank_p, y = frac, label = state,
                 colour = clase_media)) +
  geom_point(size = 2)
```



Estos resultados indican que es más probable que buenos alumnos decidan hacer el SAT. Lo interesante es que esto ocurre de manera diferente en cada estado. Por ejemplo, en algunos estados era más común otro examen: el ACT.

Si hacemos *clusters* de estados según el % de alumnos, empezamos a ver otra historia. Para esto, ajustemos rectas de mínimos cuadrados como referencia:



Sin embargo, el resultado puede variar considerablemente si categorizamos de distintas maneras.

Tablas de conteos

Consideremos los siguientes datos de tomadores de té (del paquete FactoMineR (Lê et al., 2008)):

```
tea <- read_csv("data/tea.csv")
# nombres y códigos
te <- tea %>% select(how, price, sugar) %>%
  rename(presentacion = how, precio = price, azucar = sugar) %>%
  mutate(
    presentacion = fct_recode(presentacion,
      suelto = "unpackaged", bolsas = "tea bag", mixto = "tea bag+unpackaged"),
    precio = fct_recode(precio,
      marca = "p_branded", variable = "p_variable", barato = "p_cheap",
      marca_propia = "p_private label", desconocido = "p_unknown", fino = "p_upscale"),
    azucar = fct_recode(azucar,
      sin_azucar = "No.sugar", con_azucar = "sugar"))
```

```
sample_n(te, 10)
```

```
## # A tibble: 10 x 3
##   presentacion precio  azucar
##   <fct>        <fct>  <fct>
## 1 mixto         variable sin_azúcar
## 2 suelto        fino    con_azúcar
## 3 bolsas        fino    con_azúcar
## 4 mixto         variable sin_azúcar
## 5 bolsas        variable sin_azúcar
## 6 suelto        variable con_azúcar
## 7 bolsas        variable con_azúcar
## 8 mixto         fino    sin_azúcar
## 9 bolsas        marca   con_azúcar
## 10 mixto        marca   sin_azúcar
```

Nos interesa ver qué personas compran té suelto, y de qué tipo. Empezamos por ver las proporciones que compran té según su empaque (en bolsita o suelto):

```
precio <- te %>% group_by(precio) %>%
  tally() %>% mutate(prop = round(100 * n / sum(n))) %>%
  select(-n)
tipo <- te %>% group_by(presentacion) %>% tally() %>%
  mutate(pct = round(100 * n / sum(n)))
tipo %>% formatear_tabla
```

	presentacion	n	pct
bolsas	170	57	
mixto	94	31	
suelto	36	12	

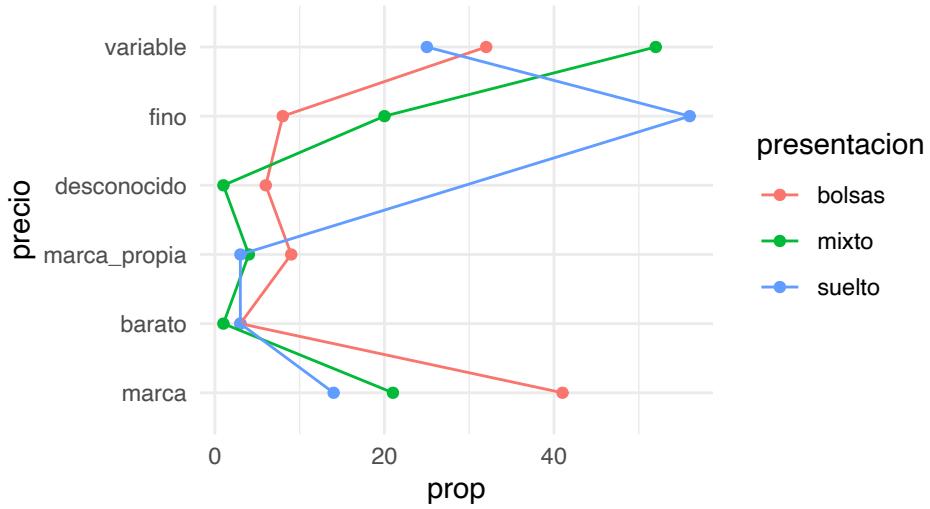
La mayor parte de las personas toma té en bolsas. Sin embargo, el tipo de té (en términos de precio o marca) que compran es muy distinto dependiendo de la presentación:

```
tipo <- tipo %>% select(presentacion, prop_presentacion = pct)
tabla_cruzada <- te %>%
  group_by(presentacion, precio) %>%
  tally() %>%
  # porcentajes por presentación
group_by(presentacion) %>%
  mutate(prop = round(100 * n / sum(n))) %>%
  select(-n)
tabla_cruzada %>%
  pivot_wider(names_from = presentacion, values_from = prop,
             values_fill = list(prop = 0)) %>%
  formatear_tabla()
```

precio	bolsas	mixto	suelto
marca	41	21	14
barato	3	1	3
marca_propia	9	4	3
desconocido	6	1	0
fino	8	20	56
variable	32	52	25

Estos datos podemos examinarlos un rato y llegar a conclusiones. Notemos que el uso de tablas no permite mostrar claramente patrones. Tampoco por medio de gráficas como la siguiente:

```
ggplot(tabla_cruzada %>% ungroup %>%
  mutate(price = fct_reorder(precio, prop)),
  aes(x = precio, y = prop, group = presentacion, colour = presentacion)) +
  geom_point() + coord_flip() + geom_line()
```



En lugar de eso, calcularemos *profiles columna*. Esto es, comparamos cada una de las columnas con la columna marginal (en la tabla de tipo de té):

```
num_grupos <- n_distinct(te %>% select(presentacion))
tabla <- te %>%
  group_by(presentacion, precio) %>%
  tally() %>%
  group_by(presentacion) %>%
  mutate(prop_precio = (100 * n / sum(n))) %>%
  group_by(precio) %>%
  mutate(prom_prop = sum(prop_precio)/num_grupos) %>%
  mutate(perfil = 100 * (prop_precio / prom_prop - 1))
tabla
```

```
## # A tibble: 17 x 6
## # Groups:   precio [6]
##   presentacion precio     n prop_precio prom_prop perfil
##   <fct>        <fct>   <int>      <dbl>      <dbl>    <dbl>
## 1 bolsas        marca    70       41.2      25.4     61.8
## 2 bolsas        barato     5        2.94      2.26     30.1
## 3 bolsas        marca_propia 16       9.41      5.48     71.7
## 4 bolsas        desconocido 11       6.47      2.51    158.
## 5 bolsas        fino      14       8.24      28.0    -70.6
## 6 bolsas        variable   54      31.8      36.3    -12.5
## 7 mixto         marca    20       21.3      25.4    -16.4
## 8 mixto         barato     1        1.06      2.26    -52.9
## 9 mixto         marca_propia 4        4.26      5.48    -22.4
## 10 mixto        desconocido 1        1.06      2.51    -57.6
## 11 mixto        fino      19       20.2      28.0    -27.8
## 12 mixto        variable   49      52.1      36.3     43.6
## 13 suelto       marca     5        13.9      25.4    -45.4
```

```

## 14 suelto      barato        1     2.78    2.26   22.9
## 15 suelto      marca_propia  1     2.78    5.48  -49.3
## 16 suelto      fino          20    55.6    28.0   98.4
## 17 suelto      variable      9     25     36.3  -31.1

```

```

tabla_perfil <- tabla %>%
  select(presentacion, precio, perfil, pct = prom_prop) %>%
  pivot_wider(names_from = presentacion, values_from = perfil,
              values_fill = list(perfil = -100.0))
if_profile <- function(x){
  any(x < 0) & any(x > 0)
}
marcar <- marcar_tabla_fun(25, "red", "black")
tab_out <- tabla_perfil %>%
  arrange(desc(bolsas)) %>%
  select(-pct, everything()) %>%
  mutate_if(if_profile, marcar) %>%
  knitr::kable(format_table_salida(), escape = FALSE, digits = 0, booktabs = T) %>%
  kableExtra::kable_styling(latex_options = c("striped", "scale_down"),
                            bootstrap_options = c("hover", "condensed"),
                            full_width = FALSE)

if (knitr::is_latex_output()) {
  gsub("marca_propia", "marca-propia", tab_out)
} else {
  tab_out
}

```

precio	bolsas	mixto	suelto	pct
desconocido	157.641196013289	-57.641196013289	-100	3
marca-propia	71.6967570081604	-22.3711470973743	-49.325609910786	5
marca	61.8106471150781	-16.389635229291	-45.4210118857871	25
barato	30.0871348026653	-52.9472065607381	22.8600717580728	2
variable	-12.4877880581576	43.6124360668927	-31.1246480087351	36
fino	-70.5895012167464	-27.8146572417104	98.4041584584568	28

Leemos esta tabla como sigue: por ejemplo, los compradores de té suelto compran té *fino* a una tasa casi el doble (98 %) que el promedio.

También podemos graficar como:

```

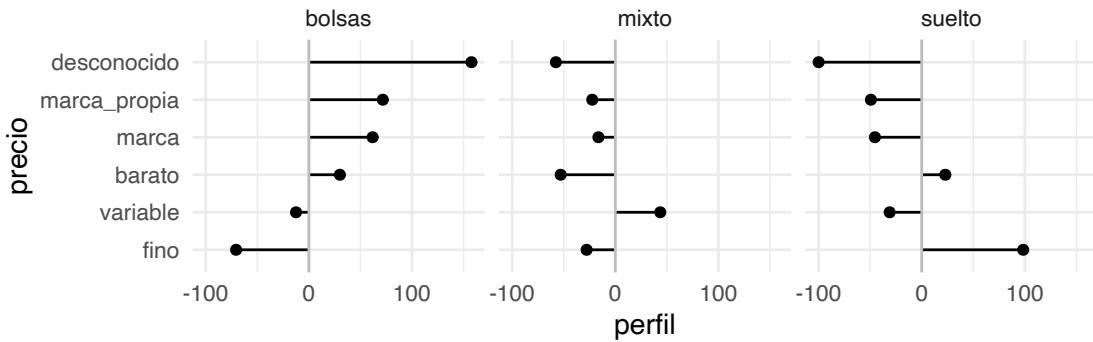
tabla_graf <- tabla_perfil %>%
  ungroup %>%

```

```

  mutate(precio = fct_reorder(precio, bolsas)) %>%
  select(-pct) %>%
  pivot_longer(cols = -precio, names_to = "presentacion", values_to = "perfil")
g_perfil <- ggplot(tabla_graf,
  aes(x = precio, xend = precio, y = perfil, yend = 0, group = presentacion)) +
  geom_point() + geom_segment() + facet_wrap(~presentacion) +
  geom_hline(yintercept = 0, colour = "gray") + coord_flip()
g_perfil

```



Observación: hay dos maneras de construir la columna promedio: tomando los porcentajes sobre todos los datos, o promediando los porcentajes de las columnas. Si los grupos de las columnas están desbalanceados, estos promedios son diferentes.

- Cuando usamos porcentajes sobre la población, perfiles columna y renglón dan el mismo resultado
- Sin embargo, cuando hay un grupo considerablemente más grande que otros, las comparaciones se vuelven vs este grupo particular. No siempre queremos hacer esto.

Interpretación

En el último ejemplo de tomadores de té utilizamos una muestra de personas, no toda la población de tomadores de té. Eso quiere decir que tenemos cierta incertidumbre de cómo se generalizan o no los resultados que obtuvimos en nuestro análisis a la población general.

Nuestra respuesta depende de cómo se extrajo la muestra que estamos considerando. Si el mecanismo de extracción incluye algún proceso probabilístico, entonces es posible en principio entender qué tan bien generalizan los resultados de nuestro análisis a la población general, y entender esto depende de entender qué tanta variación hay de muestra a muestra, de todas las posibles muestras que pudimos haber extraído.

En las siguientes secciones discutiremos estos aspectos, en los cuales pasamos del trabajo de “detective” al trabajo de “juez” en nuestro trabajo analítico.

Loess

Las gráficas de dispersión son la herramienta básica para describir la relación entre dos variables cuantitativas, y como vimos en ejemplos anteriores, muchas veces podemos apreciar mejor la relación

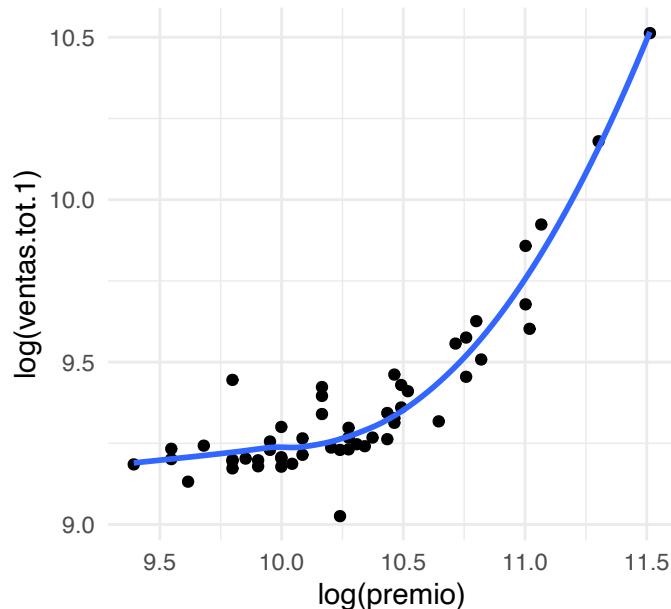
entre ellas si agregamos una curva `loess` que suavice.

Los siguientes datos muestran los premios ofrecidos y las ventas totales de una lotería a lo largo de 53 sorteos (las unidades son cantidades de dinero indexadas). Graficamos en escalas logarítmicas y agregamos una curva `loess`.

Los siguientes datos muestran los premios ofrecidos y las ventas totales de una lotería a lo largo de 53 sorteos (las unidades son cantidades de dinero indexadas). Graficamos en escalas logarítmicas y agregamos una curva `loess`.

```
# cargamos los datos
load(here::here("data", "ventas_sorteo.Rdata"))

ggplot(ventas.sorteo, aes(x = log(premio), y = log(ventas.tot.1))) +
  geom_point() +
  geom_smooth(method = "loess", alpha = 0.5, degree = 1, se = FALSE)
```



El patrón no era difícil de ver en los datos originales, sin embargo, la curva lo hace más claro, el logaritmo de las ventas tiene una relación no lineal con el logaritmo del premio: para premios no muy grandes no parece haber gran diferencia, pero cuando los premios empiezan a crecer por encima de $20,000$ (aproximadamente e^{10}), las ventas crecen más rápidamente que para premios menores. Este efecto se conoce como *bola de nieve*, y es frecuente en este tipo de loterías.

Antes de adentrarnos a los modelos `loess` comenzamos explicando cómo se ajustan familias paramétricas de curvas a conjuntos de datos dados.



Ajustando familias paramétricas El modelo de *regresion lineal* ajusta una recta a un conjunto de datos. Por ejemplo, consideremos la familia

$$f_{a,b} = ax + b, \quad (2.16)$$

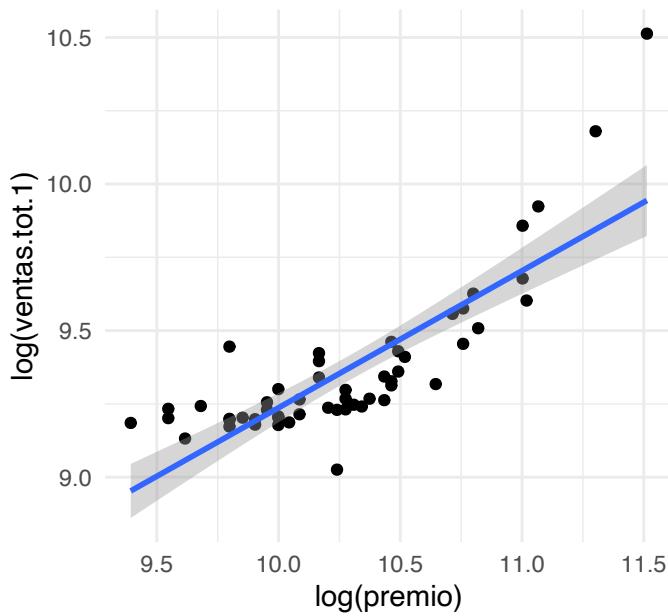
para un conjunto de datos bivariados $\{(x_1, y_1), \dots, (x_N, y_N)\}$. Buscamos encontrar a y b tales que $f_{a,b}$ sea un ajuste óptimo a los datos. Para esto, se minimiza la suma de errores cuadráticos

$$\frac{1}{N} \sum_{n=1}^N (y_n - ax_n - b)^2. \quad (2.17)$$

Observación: Los modelos de regresión lineal, cuando se pueden ajustar de manera razonable, son altamente deseables por su simplicidad: los datos se describen con pocos parámetros y tenemos incrementos marginales constantes en todo el rango de la variable que juega como factor, de modo que la interpretación es simple. Por esta razón, muchas veces vale la pena transformar los datos con el fin de enderezar la relación de dos variables y poder ajustar una función lineal (lo veremos más adelante).

En este caso, las constantes a y b se pueden encontrar diferenciando la función de mínimos cuadrados. Nótese que podemos repetir el argumento con otras familias de funciones (por ejemplo cuadráticas).

```
ggplot(ventas.sorteo, aes(x = log(premio), y = log(ventas.tot.1))) +
  geom_point() +
  geom_smooth(method = "lm")
```



Donde los parámetros a y b están dados por:

```
mod_lineal <- lm(log(ventas.tot.1) ~ log(premio), data = ventas.sorteo)
round(coef(mod_lineal), 2)
```

```
## (Intercept) log(premio)
##           4.56      0.47
```

De modo que la curva ajustada es $\log(V) = 4,6 + 0,47 \log(P)$, o en las unidades originales $V = 100P^{0,47}$, donde V son las ventas y P el premio. Si observamos la gráfica notamos que este modelo lineal (en los logaritmos) no es adecuado para estos datos. Podríamos experimentar con otras familias (por ejemplo, una cuadrática o cúbica, potencias, exponenciales, etc.); sin embargo, en la etapa exploratoria es mejor tomar una ruta de ajuste más flexibles (aún cuando esta no sea con funciones algebraicas), que al mismo tiempo sea robusto.



Curvas loess (regresión local): Una manera de mejorar la flexibilidad de los modelos lineales es considerar rectas de manera local. Esto lo logramos al minimizar una suma de errores cuadráticos ponderados por pesos $w_n(x)$. Para esto minimizamos las N funciones de error

$$\sum_{n=1}^N w_n(x_k)(y_n - ax_n - b)^2, \quad (2.18)$$

donde $x_k \in \{x_1, \dots, x_N\}$.



Observación: Notemos que el ajuste tradicional por medio de mínimos cuadrados considera un ajuste *global* con pesos uniformes

$$w_n(x) = \frac{1}{N}. \quad (2.19)$$



Interpretación: En la práctica, los pesos $w_n(\cdot)$ contienen a su vez un conjunto de parámetros que dictan el comportamiento del ajuste local. Si por ejemplo tomamos en cuenta pesos Gaussianos

$$w_n(x) = \exp\left(-\frac{(x - x_n)^2}{2\sigma^2}\right), \quad (2.20)$$

el parámetro σ dicta el efecto local de los puntos ancla x_n . Dicho de otro modo, dicta la vecindad de incidencia de los datos x_n .



Regresión polinomial: Una generalización sencilla es considerar polinomios para el ajuste

$$f_p(x) = b + a_1x + \cdots + a_px^p, \quad (2.21)$$

donde p es el máximo grado del polinomio. De esta forma, p controla el suavizado de la curva. El caso de recta lineal claramente es para $p = 1$.

Ajustando curvas loess

La idea es producir ajustes locales de rectas o funciones cuadráticas. Consideraremos especificar dos parámetros:

- Parámetro de suavizamiento α : cuando α es más grande, la curva ajustada es más suave.
- Grado de los polinomios locales que ajustamos λ : generalmente se toma $\lambda = 1, 2$.

Entonces, supongamos que los datos están dados por $(x_1, y_1), \dots, (x_N, y_N)$, y sean α un parámetro de suavizamiento fijo, y $\lambda = 1$. Denotamos como $\hat{y}(x)$ la curva `loess` ajustada, y como $w_n(x)$ a una función de peso (que depende de x) para la observación (x_n, y_n) .

Para poder calcular $w_n(x)$ debemos comenzar calculando $q = \lfloor N\alpha \rfloor$ que suponemos mayor que uno. Ahora definimos la función *tricubo*:

$$T(u) = \begin{cases} (1 - |u|^3)^3, & \text{para } |u| < 1. \\ 0, & \text{en otro caso.} \end{cases} \quad (2.22)$$

entonces, para el punto x definimos el peso correspondiente al dato (x_n, y_n) , denotado por $w_n(x)$ como:

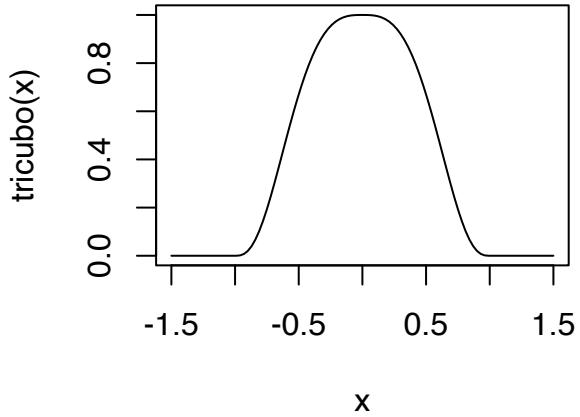
$$w_n(x) = T\left(\frac{|x - x_n|}{d_q(x)}\right)$$

donde $d_q(x)$ es el valor de la q -ésima distancia más chica (la más grande entre las q más chicas) entre los valores $|x - x_j|$, $j = 1, \dots, N$. De esta forma, las observaciones x_n reciben más peso cuanto más cerca estén de x .

En palabras, de x_1, \dots, x_N tomamos los q datos más cercanos a x , que denotamos $x_{i_1}(x) \leq x_{i_2}(x) \leq \dots \leq x_{i_q}(x) \leq$. Los re-escalamos a $[0, 1]$ haciendo corresponder x a 0 y el punto más alejado de x (que es x_{i_q}) a 1.

Aplicamos el tricubo (gráfica de abajo), para encontrar los pesos de cada punto. Los puntos que están a una distancia mayor a $d_q(x)$ reciben un peso de cero, y los más cercanos un peso que depende de que tan cercanos estén a x . Nótese que x es el punto ancla en dónde estamos ajustando la regresión local.

```
tricubo <- function(x) {
  ifelse(abs(x) < 1, (1 - abs(x) ^ 3) ^ 3, 0)
}
curve(tricubo, from = -1.5, to = 1.5)
```



Finalmente, para cada valor de x_k que está en el conjunto de datos $\{x_1, \dots, x_n\}$, ajustamos una recta de mínimos cuadrados ponderados por los pesos $w_n(x)$, es decir, minimizamos

$$\sum_{i=1}^n w_n(x_k)(y_i - ax_n - b)^2.$$

Observaciones:

1. Cualquier función (continua y quizás diferenciable) con la forma de flan del tricubo que se desvanece fuera de $(-1, 1)$, es creciente en $(-1, 0)$ y decreciente en $(0, 1)$ es un buen candidato para usarse en lugar del tricubo. La razón por la que escogemos precisamente esta forma algebráica no tiene que ver con el análisis exploratorio, sino con las ventajas teóricas adicionales que tiene en la inferencia.
2. El caso $\lambda = 2$ es similar. La única diferencia es en el paso de ajuste, donde usamos funciones cuadráticas, y obtendríamos entonces tres familias de parámetros $a(x_k), b_1(x_k), b_2(x_k)$, para cada $k \in \{1, \dots, N\}$.

Escogiendo de los parámetros. Los parámetros α y λ se encuentran por ensayo y error. La idea general es que debemos encontrar una curva que explique patrones importantes en los datos (que *ajuste* los datos) pero que no muestre variaciones a escalas más chicas difíciles de explicar (que pueden ser el resultado de influencias de otras variables, variación muestral, ruido o errores de redondeo, por ejemplo). En el proceso de prueba y error iteramos el ajuste y en cada paso hacemos análisis de residuales, con el fin de seleccionar un suavizamiento adecuado.

Ejemplo de distintas selecciones de λ , en este ejemplo consideraremos las ventas semanales de un producto a lo largo de 5 años.

Series de tiempo

Podemos usar el suavizamiento `loess` para entender y describir el comportamiento de series de tiempo, en las cuáles intentamos entender la dependencia de una serie de mediciones indexadas por el tiempo. Típicamente es necesario utilizar distintas componentes para describir exitosamente una serie de tiempo, y para esto usamos distintos tipos de suavizamientos. Veremos que distintas componentes varían en distintas escalas de tiempo (unas muy lentas, como la tendencia, otras más rápidamente, como variación quincenal, etc.).

Caso de estudio: nacimientos en México

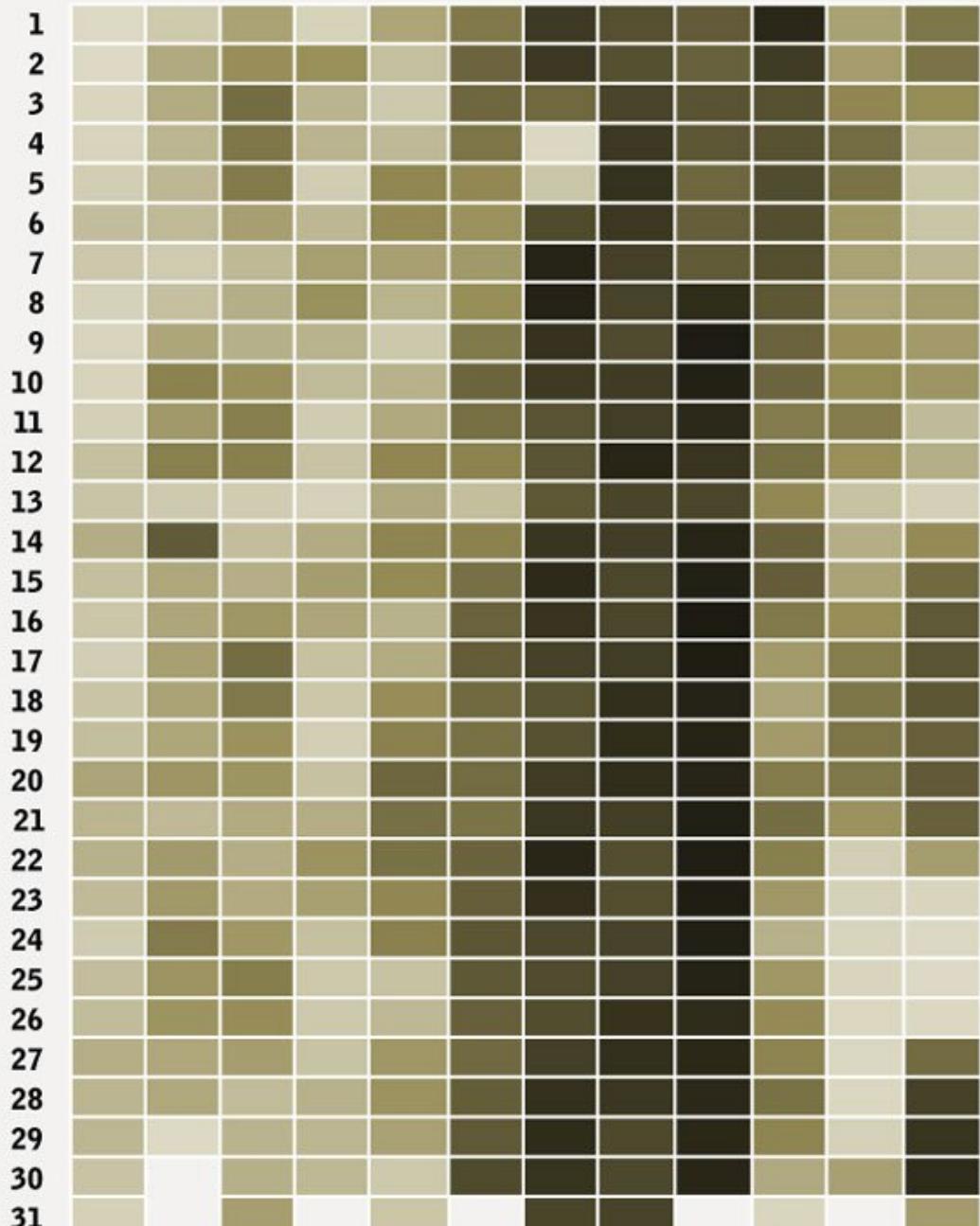
Este caso de estudio está basado en un análisis propuesto por A. Vehtari y A. Gelman, junto con un análisis de serie de tiempo de Cleveland (1993).

En nuestro caso, usaremos los datos de nacimientos registrados por día en México desde 1999. Los usaremos para contestar las preguntas: ¿cuáles son los cumpleaños más frecuentes? y ¿en qué mes del año hay más nacimientos?

Podríamos utilizar una gráfica popular (ver por ejemplo esta visualización) como:

Which Birth Dates Are Most Common?

DAY JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC



BIRTHDAY RANK

Less common

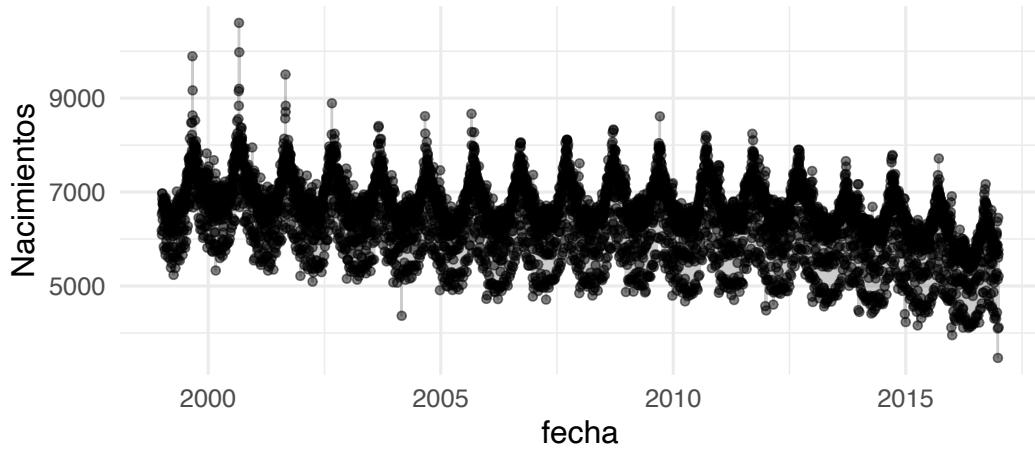
More common

Sin embargo, ¿cómo criticarías este análisis desde el punto de vista de los tres primeros principios del diseño analítico? ¿Las comparaciones son útiles? ¿Hay aspectos multivariados? ¿Qué tan bien explica o sugiere estructura, mecanismos o causalidad?

Datos de natalidad para México

```
library(lubridate)
library(ggthemes)
theme_set(theme_minimal(base_size = 14))
natalidad <- readRDS("./data/nacimientos/natalidad.rds") %>%
  mutate(dia_semana = weekdays(fecha)) %>%
  mutate(dia_año = yday(fecha)) %>%
  mutate(año = year(fecha)) %>%
  mutate(mes = month(fecha)) %>% ungroup %>%
  mutate(dia_semana = recode(dia_semana, Monday = "Lunes", Tuesday = "Martes", Wednesday = "Miércoles",
                             Thursday = "Jueves", Friday = "Viernes", Saturday = "Sábado", Sunday = "Domingo")) %>%
  mutate(dia_semana = fct_relevel(dia_semana, c("Lunes", "Martes", "Miércoles",
                                                "Jueves", "Viernes", "Sábado", "Domingo")))
```

Consideremos los *datos agregados* del número de nacimientos (registrados) por día desde 1999 hasta 2016. Un primer intento podría ser hacer una gráfica de la serie de tiempo. Sin embargo, vemos que no es muy útil:



Hay varias características que notamos. Primero, parece haber una tendencia ligeramente decreciente del número de nacimientos a lo largo de los años. Segundo, la gráfica sugiere un patrón anual. Y por último, encontramos que hay dispersión producida por los días de la semana.

Sólo estas características hacen que la comparación entre días sea difícil de realizar. Supongamos que comparamos el número de nacimientos de dos miércoles dados. Esa comparación será diferente dependiendo: del año donde ocurrieron, el mes donde ocurrieron, si semana santa ocurrió en algunos de los miércoles, y así sucesivamente.

Como en nuestros ejemplos anteriores, la idea del siguiente análisis es aislar las componentes que observamos en la serie de tiempo: extraemos componentes ajustadas, y luego examinamos los residuales.

En este caso particular, asumiremos una **descomposición aditiva** de la serie de tiempo (Cleveland, 1993).



En el estudio de **series de tiempo** una estructura común es considerar el efecto de diversos factores como tendencia, estacionalidad, ciclicidad e irregularidades de manera aditiva. Esto es, consideraremos la descomposición

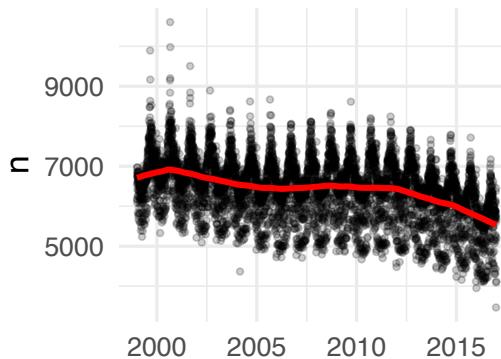
$$y(t) = f_t(t) + f_e(t) + f_c(t) + \varepsilon. \quad (2.23)$$

Una estrategia de ajuste, como veremos más adelante, es proceder de manera *modular*. Es decir, se ajustan los componentes de manera secuencial considerando los residuales de los anteriores.

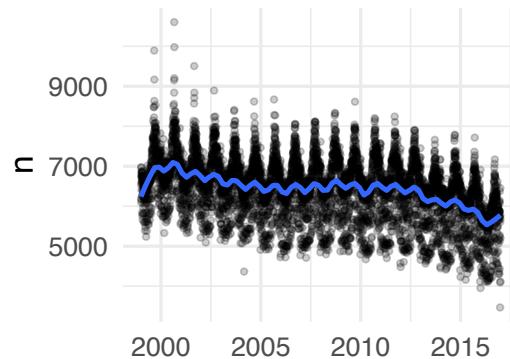
Tendencia

Comenzamos por extraer la tendencia, haciendo promedios **loess** (Cleveland, 1979) con vecindades relativamente grandes. Quizá preferiríamos suavizar menos para capturar más variación lenta, pero si hacemos esto en este punto empezamos a absorber parte de la componente anual:

```
mod_1 <- loess(n ~ as.numeric(fecha), data = natalidad, span = 0.2, degree = 1)
datos_dia <- natalidad %>% mutate(ajuste_1 = fitted(mod_1)) %>%
  mutate(res_1 = n - ajuste_1)
```



Suavizado apropiado



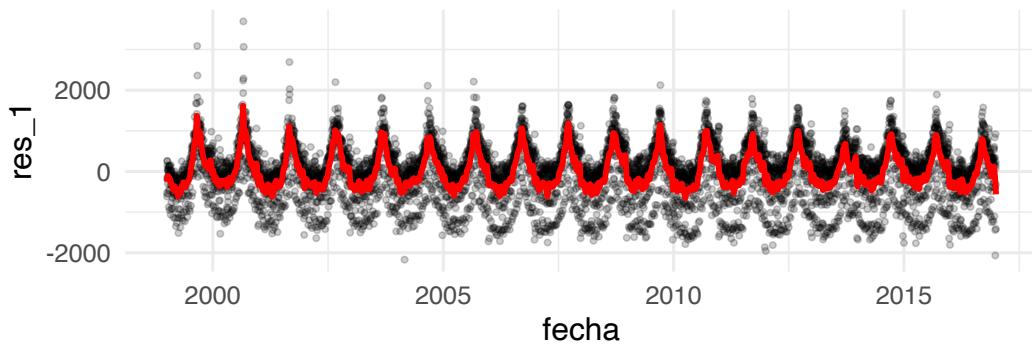
Requiere mayor suavizado

Notemos que a principios de 2000 el suavizador está en niveles de alrededor de 7000 nacimientos diarios, hacia 2015 ese número es más cercano a unos 6000.

Componente anual

Al obtener la tendencia podemos aislar el efecto a largo plazo y proceder a realizar mejores comparaciones (por ejemplo, comparar un día de 2000 y de 2015 tendría más sentido). Ahora, ajustamos **los residuales del suavizado anterior**, pero con menos suavizamiento. Así evitamos capturar tendencia:

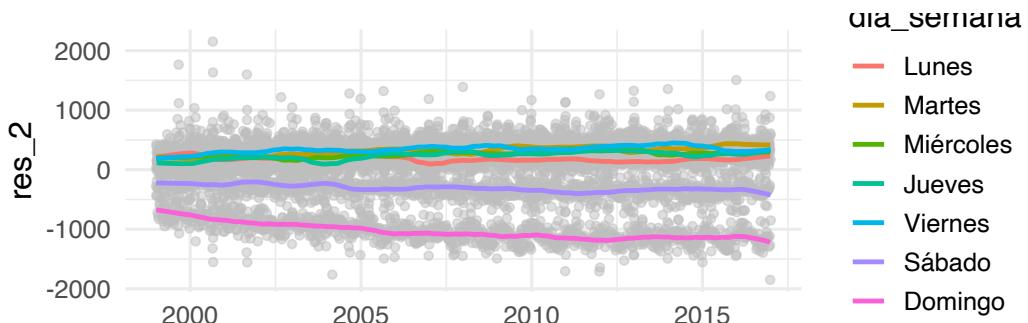
```
mod_anual <- loess(res_1 ~ as.numeric(fecha), data = datos_dia, degree = 2, span = 0.005)
datos_dia <- datos_dia %>% mutate(ajuste_2 = fitted(mod_anual)) %>%
  mutate(res_2 = res_1 - ajuste_2)
```



Día de la semana

Hasta ahora, hemos aislado los efectos por plazos largos de tiempo (tendencia) y hemos incorporado las variaciones estacionales (componente anual) de nuestra serie de tiempo. Ahora, veremos cómo capturar el efecto por día de la semana. En este caso, podemos hacer suavizamiento `loess` para cada serie de manera independiente

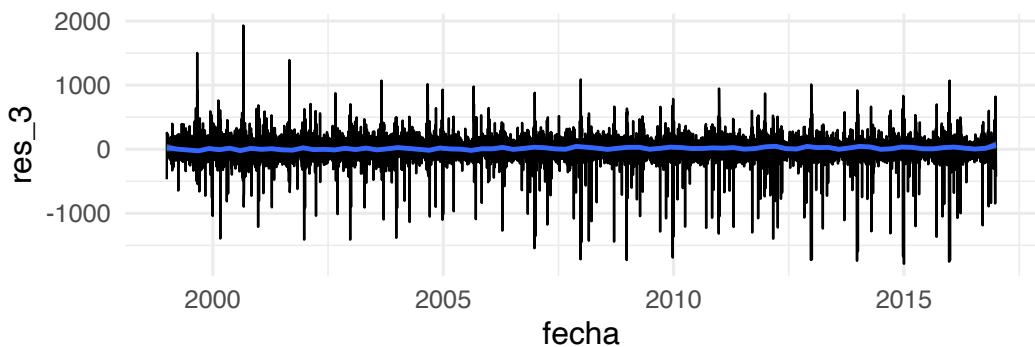
```
datos_dia <- datos_dia %>%
  group_by(dia_semana) %>%
  nest() %>%
  mutate(ajuste_mod =
    map(data, ~ loess(res_2 ~ as.numeric(fecha), data = .x, span = 0.1, degree = 1))) %>%
  mutate(ajuste_3 = map(ajuste_mod, fitted)) %>%
  select(-ajuste_mod) %>% unnest(cols = c(data, ajuste_3)) %>%
  mutate(res_3 = res_2 - ajuste_3) %>% ungroup
```



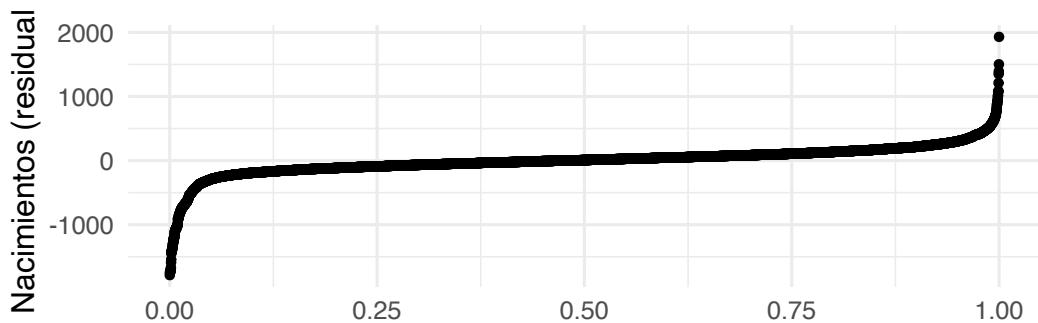
Residuales

Por último, examinamos los residuales finales quitando los efectos ajustados:

```
## `geom_smooth()` using formula 'y ~ x'
```



Observación: nótese que la distribución de estos residuales presenta irregularidades interesantes. La distribución es de *colas largas*, y no se debe a unos cuantos datos atípicos. Esto generalmente es indicación que hay factores importantes que hay que examinar mas a detalle en los residuales:

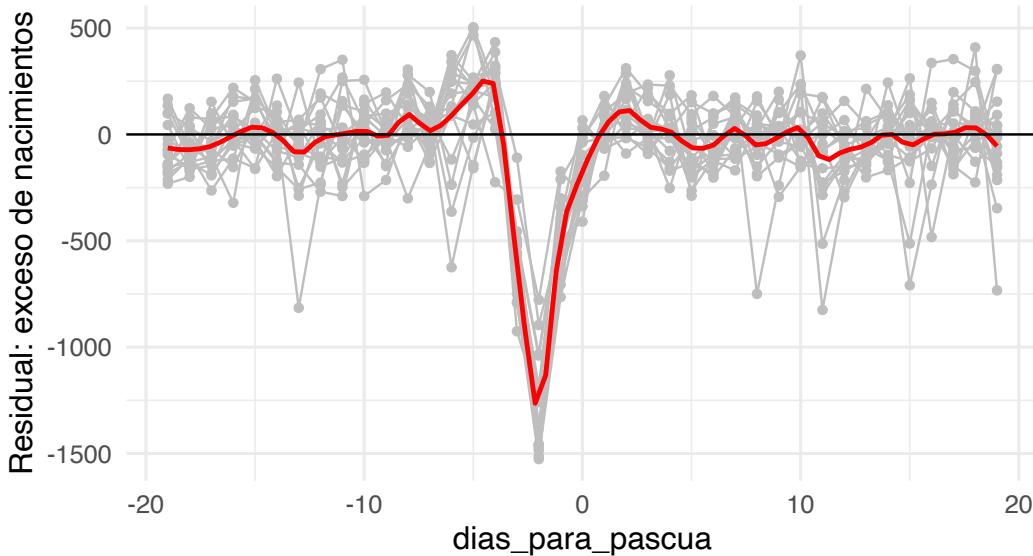
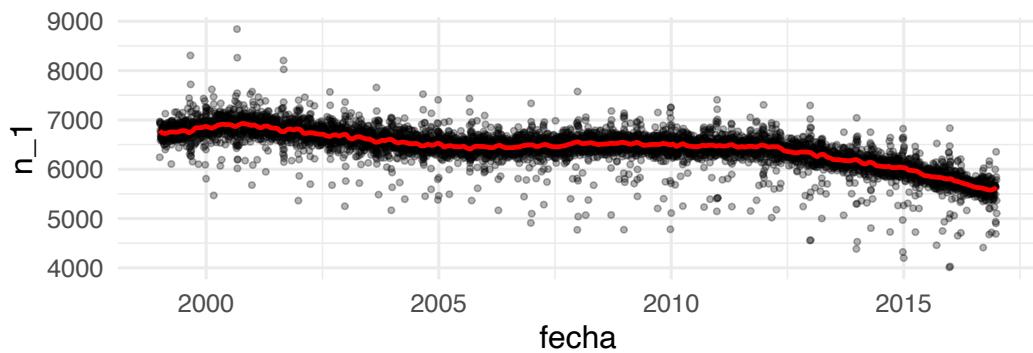


Reestimación

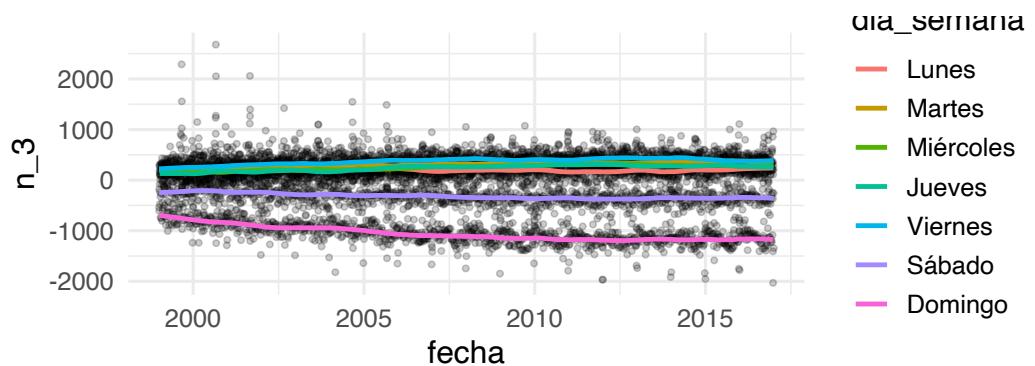
Cuando hacemos este proceso secuencial de llevar el ajuste a los residual, a veces conviene iterarlo. La razón es que en una segunda o tercera pasada podemos hacer mejores estimaciones de cada componente, y es posible suavizar menos sin capturar *componentes de más alta frecuencia*.

Así que podemos regresar a la serie original para hacer mejores estimaciones, más suavizadas:

```
# Quitamos componente anual y efecto de día de la semana
datos_dia <- datos_dia %>% mutate(n_1 = n - ajuste_2 - ajuste_3)
# Reajustamos
mod_1 <- loess(n_1 ~ as.numeric(fecha), data = datos_dia, span = 0.02, degree = 2,
                family = "symmetric")
```

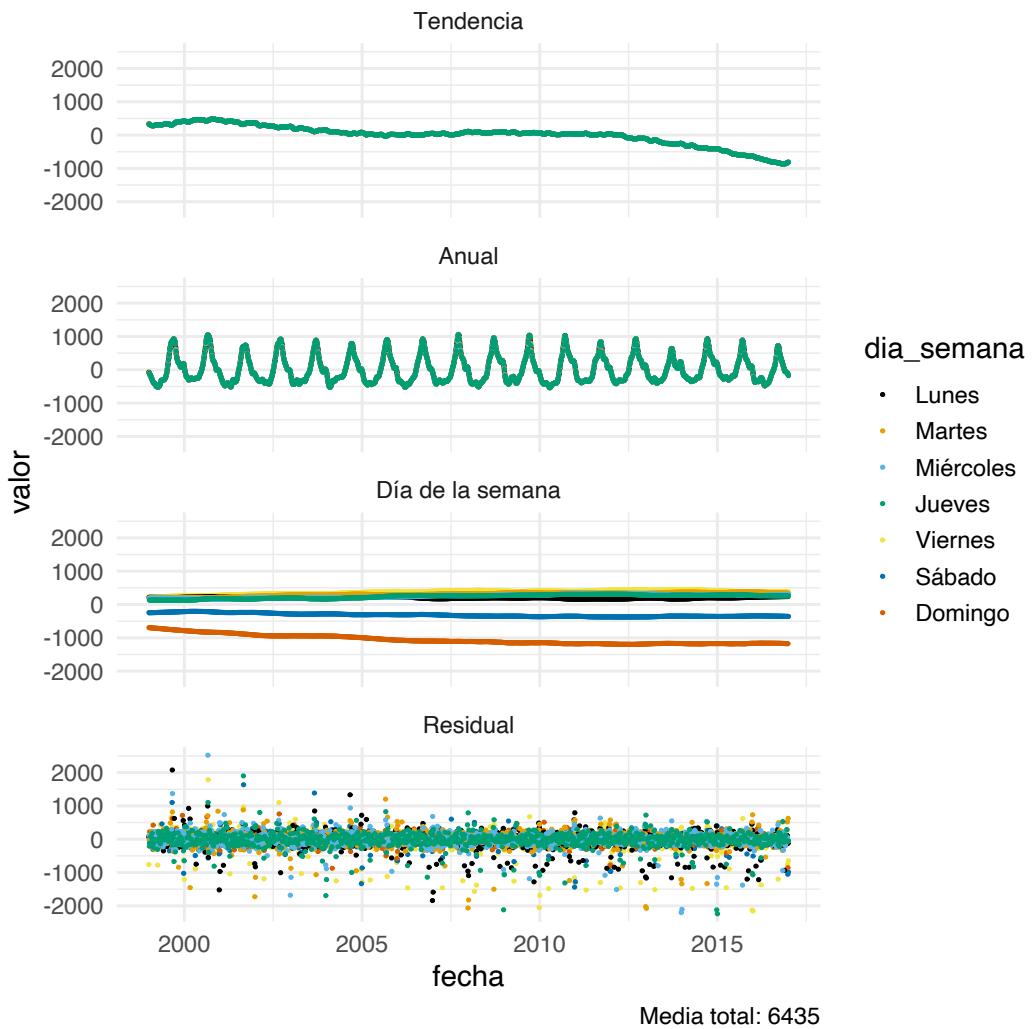


Y ahora repetimos con la componente de día de la semana:



Análisis de componentes

Ahora comparamos las componentes estimadas y los residuales en una misma gráfica. Por definición, la suma de todas estas componentes da los datos originales.



Este último paso nos permite diversas comparaciones que explican la variación que vimos en los datos. Una gran parte de los residuales está entre ± 250 nacimientos por día. Sin embargo, vemos que las colas tienen una dispersión mucho mayor:

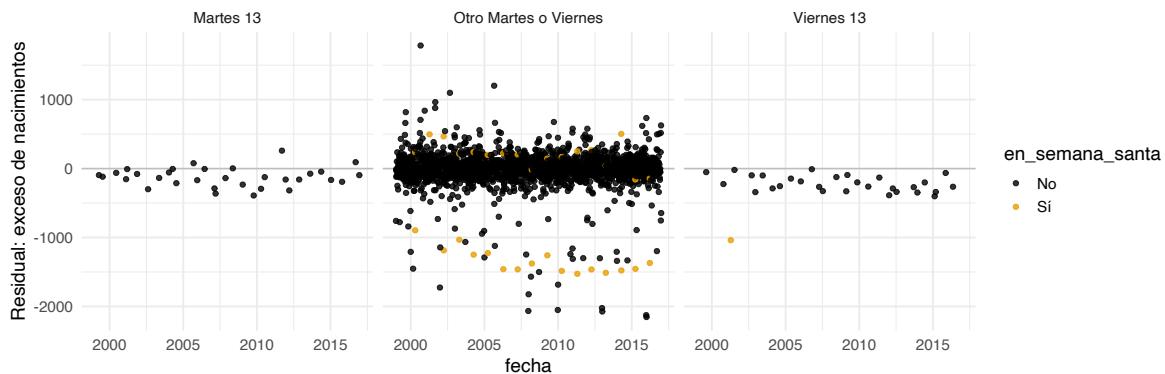
```
quantile(datos_dia$res_6, c(0.00, 0.01, 0.05, 0.10, 0.90, 0.95, 0.99, 1)) %>% round
```

```
##      0%     1%     5%    10%    90%    95%    99%   100%
## -2238 -1134  -315  -202   188   268   516   2521
```

¿A qué se deben estas colas tan largas?

Viernes 13?

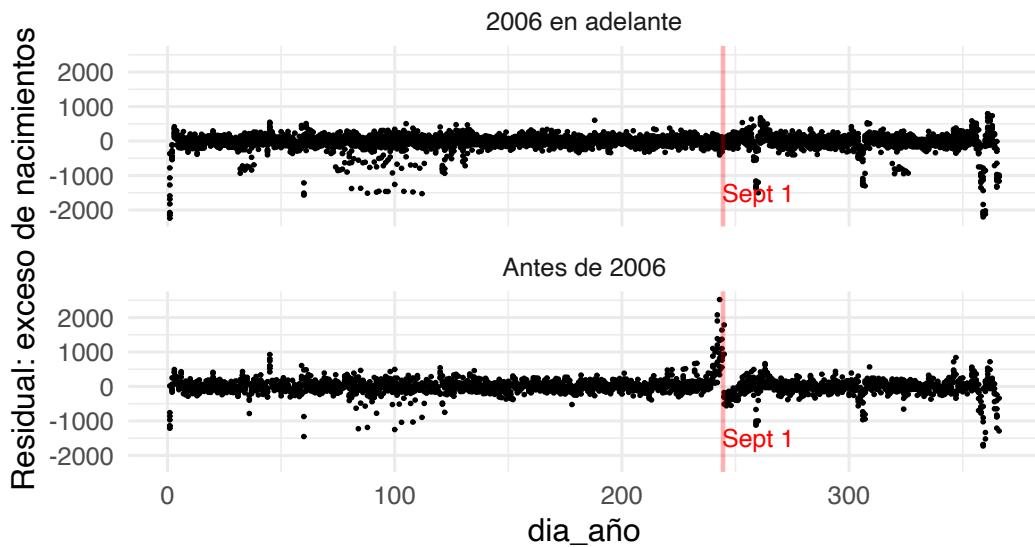
Podemos empezar con una curiosidad. Los días Viernes o Martes 13, ¿nacen menos niños?



Nótese que fue útil agregar el indicador de Semana santa por el Viernes 13 de Semana Santa que se ve como un atípico en el panel de los viernes 13.

Residuales: antes y después de 2006

Veamos primero una agregación sobre los años de los residuales. Lo primero es observar un cambio que sucedió repentinamente en 2006:



La razón es un cambio en la ley acerca de cuándo pueden entrar los niños a la primaria. Antes era por edad y había poco margen. Ese exceso de nacimientos son reportes falsos para que los niños no tuvieran que esperar un año completo por haber nacido unos cuantos días antes de la fecha límite.

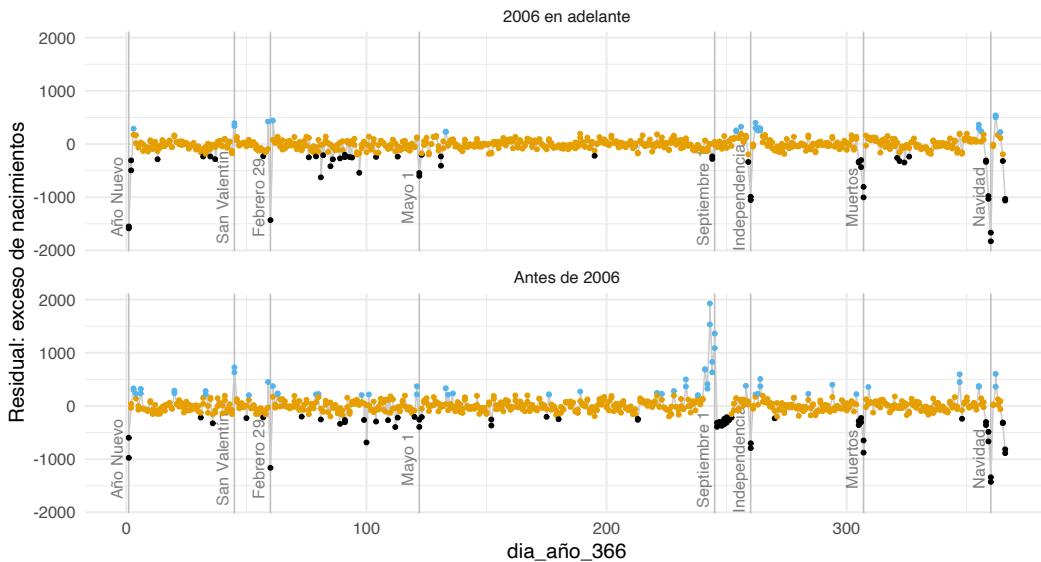
Otras características que debemos investigar:

- Efectos de Año Nuevo, Navidad, Septiembre 16 y otros días feriados como Febrero 14.
- Semana santa: como la fecha cambia, vemos que los residuales negativos tienden a ocurrir dispersos alrededor del día 100 del año.

Otros días especiales: más de residuales

Ahora promediamos residuales (es posible agregar barras para indicar dispersión a lo largo de los años) para cada día del año. Podemos identificar ahora los residuales más grandes: se deben, por ejemplo, a días feriados, con consecuencias adicionales que tienen en días adjuntos (excesos de nacimientos):

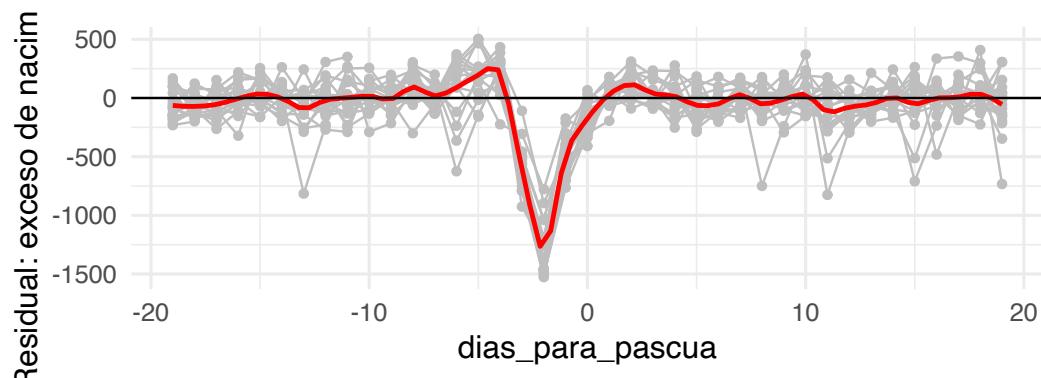
```
## `summarise()` regrouping output by `dia_año_366` , `antes_2006` (override with `groups` argument)
```



Semana santa

Para Semana Santa tenemos que hacer unos cálculos. Si alineamos los datos por días antes de Domingo de Pascua, obtenemos un patrón de caída fuerte de nacimientos el Viernes de Semana Santa, y la característica forma de “valle con hombros” en días anteriores y posteriores estos Viernes. ¿Por qué ocurre este patrón?

```
## `geom_smooth()` using formula `y ~ x`
```



Nótese un defecto de nuestro modelo: el patrón de “hombres” alrededor del Viernes Santo no es suficientemente fuerte para equilibrar los nacimientos faltantes. ¿Cómo podríamos mejorar nuestra descomposición?

Bibliografía

- Chihara, L. M. and Hesterberg, T. C. (2018). *Mathematical Statistics with Resampling and R*. John Wiley & Sons, Hoboken, NJ, 2 edition.
- Cleveland, W. (1994). *The Elements of Graphing Data*. AT&T Bell Laboratories.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.
- Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press.
- Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):pp. 531–554.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA.
- Guber, D. (1999). Getting what you pay for: The debate over equity in public school expenditures. *Journal of Statistics Education*, 7(2).
- Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386. PMID: 27019512.
- Kruschke, J. (2015). *Doing Bayesian Data Analysis (Second Edition)*. Academic Press.
- Lê, S., Josse, J., Husson, F., et al. (2008). Factominer: an r package for multivariate analysis. *Journal of statistical software*, 25(1):1–18.
- Tufte, E. R. (1986). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA.
- Tufte, E. R. (2006). *Beautiful Evidence*. Graphics Press, Cheshire, CT.
- Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company.
- Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

