

Visualización de Información

Tablas y Modelos de Color

Daniela Opitz

dopitz@udd.cl

Instituto Data Science, Universidad del Desarrollo
Edición 2024

1. Visualización de Tablas

Scatterplot

Datos: 2 atributos cuantitativos

Expresa atributos cuantitativos. Usa solamente valores.

Marca: puntos.

Canales: posición horizontal y vertical.

Tareas: encontrar patrones, outliers, distribuciones, correlaciones, clusters.

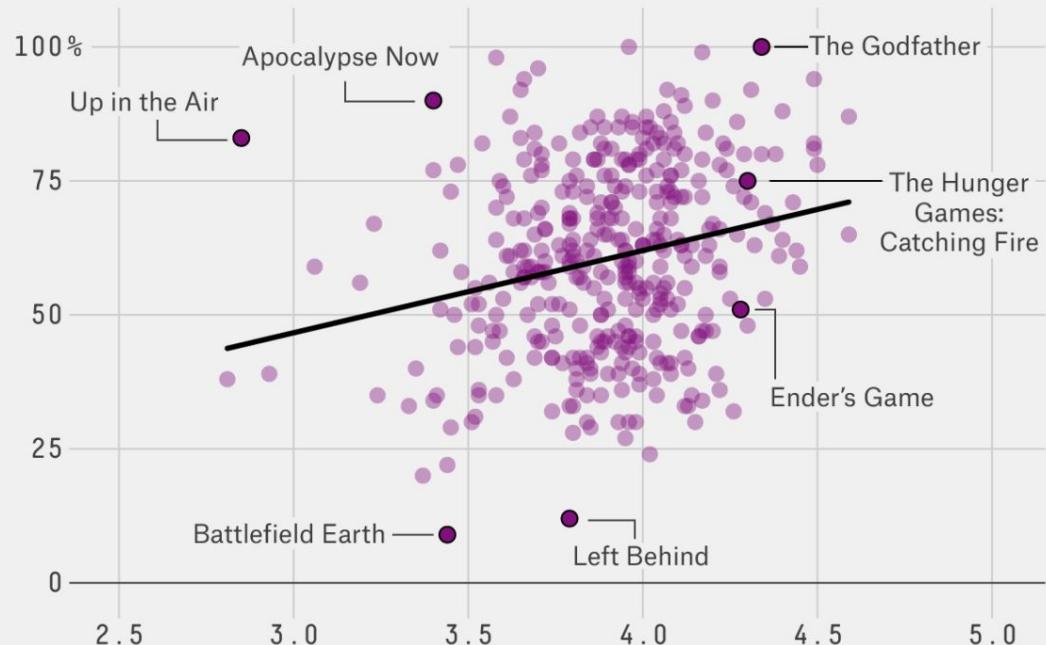
Escalabilidad: cientos de ítems (filas).

Películas

Libros

When Books Become Movies

Metacritic score of films vs. Goodreads score of source novel



Line Chart

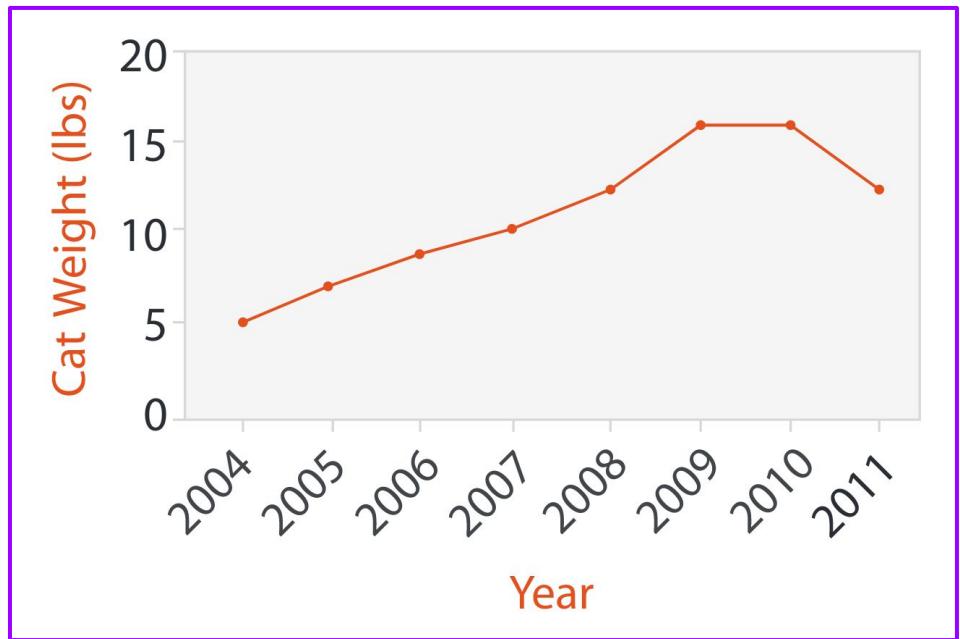
Datos: una llave (en este caso el año) y un valor: 2 atributos cuantitativos.

Marca: puntos, que se conectan a través de líneas

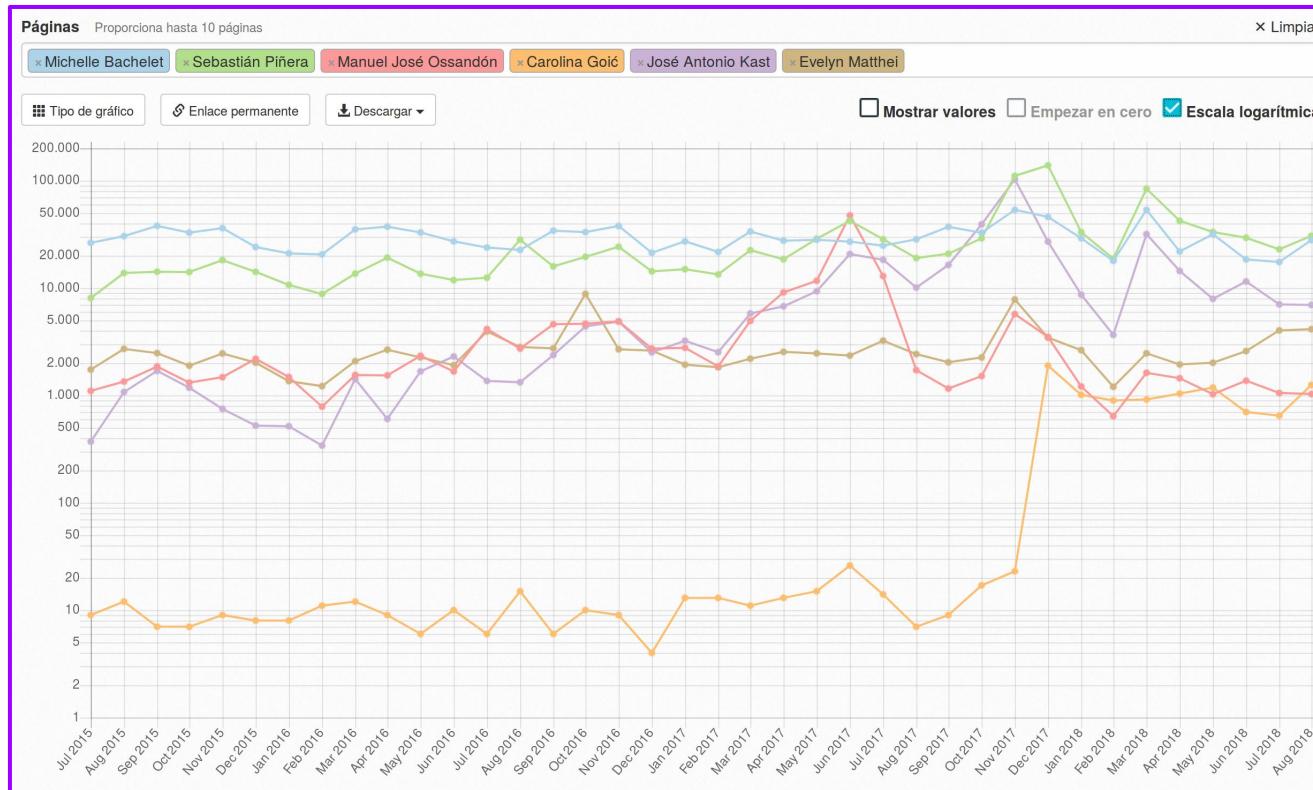
Canales: la posición vertical de los puntos expresa un valor cuantitativo, la posición horizontal es ordenada por el atributo llave.

Tareas: encontrar tendencias/patrones.

Las líneas que conectan cada par consecutivo de puntos enfatizan la relación que hay entre un ítem y el siguiente.



Line Chart



Fuente: [Wikipedia Pageviews](#)

Area Chart

Una línea puede convertirse en un área si se rellena el área debajo (o sobre ella) respecto a un eje de referencia.

Cuando trabajamos con áreas podemos hablar de **area chart**.



Fuente: Ministerio de Ciencia

Infartos, accidentes cardiovasculares y arritmias graves son algunos de los casos que se han disparado:

Demanda por camas UCI no asociada a covid-19 llega a niveles nunca vistos en el sistema de salud

MAX CHÁVEZ

Nunca el sistema de salud había tenido una necesidad tan alta de cupos en unidades de cuidados intensivos (UCI) como durante la pandemia. Sin embargo, con la mejoría de los indicadores tras la segunda ola de contagios, la demanda de camas críticas por el covid-19 ha bajado de manera considerable, habiendo hoy solo cerca de 370 pacientes con dicha patología internados en el país, casi el 10% de lo que hubo durante el *peak*. No obstante, así, la red sanitaria está funcionando a un ritmo mayor que antes de la llegada del coronavirus.

¿La razón? Jamás la demanda UCI por otras enfermedades ha-

bi sido tan elevada como ahora, con casi 1.600 personas ingresadas. Incluso, dicha cantidad supera con creces la capacidad total previa a la emergencia (cerca de 1.100).

Tal como se observa en los números de hospitalizaciones en UCI, las atenciones de urgencia por males crónicos durante los últimos meses han crecido muy por sobre los niveles normales.

Por ejemplo, durante las cinco semanas recientes, en el sistema público se han atendido 639 pacientes de urgencia por arritmias graves al corazón, muy por

encima de los 520 en el mismo período del año pasado o los 567 recibidos en 2019.

Esta situación se repite en distintas patologías. En el caso de los infartos, se registra un 15% más de ingresos este año que durante 2019, previo a la llegada del covid, y 18% más si se compara con el año pasado. Algo similar sucede con los accidentes cerebrovasculares, con un alza de cerca de un 8%.

Juan Pablo González, urgenciólogo y docente de Medicina de la U. de O'Higgins, señala que "en un comienzo se evidenció una acumulación de pacien-

tes que estaban en listas de espera para resolver varios problemas, como cirugías programadas o garantías pendientes. Pero si revisamos los ingresos que están ocurriendo ahora, hay un incremento de pacientes con patologías agudas o crónicas descompensadas por consulta de urgencia". Agrega que "lo que puede estar ocurriendo detrás de ese número es que son pacientes que posteriguran sus controles de salud, sus tratamientos, y este retraso en la atención ha terminado repercutiendo meses después en una mayor tasa de

enfermedades vasculares, cardíacas, etcétera".

Héctor Sánchez, director del Instituto de Salud Pública de la U. Andrés Bello, advierte que este fenómeno "se va a mantener de forma importante, por lo menos, en los próximos dos a tres años, independientemente de que el sistema de salud haga todos los esfuerzos posibles con el objeto de reducir las listas de espera, porque en la medida que vayamos reduciendo las listas, esta cantidad de pacientes va a ir apareciendo".

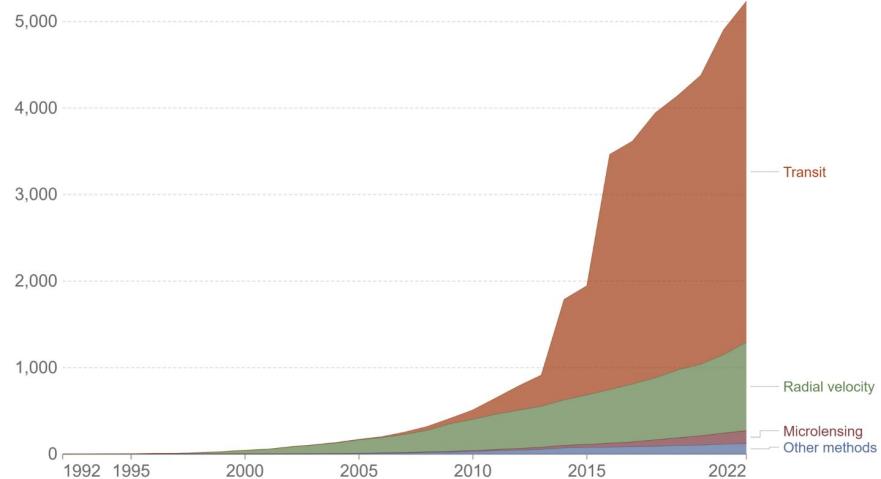
SPECIALISTAS ADVIERTEN POR AUMENTO DE MUERTES POR CÁNCER GÁSTRICO EL PRÓXIMO AÑO A 6

Stacked Area Charts

- Se usa para mostrar cómo se divide **un todo en sus partes componentes**.
- Muestra como el total de todas las categorías **cambia a lo largo del tiempo**. El área total bajo la curva en cualquier punto del eje horizontal representa la suma total de todas las categorías en ese punto.
- Aunque el enfoque principal está en el total acumulado, estos gráficos también permiten visualizar el tamaño relativo de las contribuciones de las categorías individuales a ese total.
- **Tareas: comparar tendencias y contribuciones por categoría.**

Cumulative number of exoplanets discovered, by method
Cumulative number of planets discovered outside the Solar System, broken down by their method of first identification.

Our World
in Data



Source: NASA Exoplanet Archive (2023)

OurWorldInData.org/space-exploration-satellites • CC BY

Bar Chart

Datos: 1 atributo categórico, 1 cuantitativo.

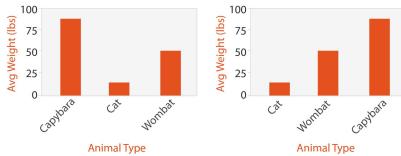
Marca: líneas.

Canales:

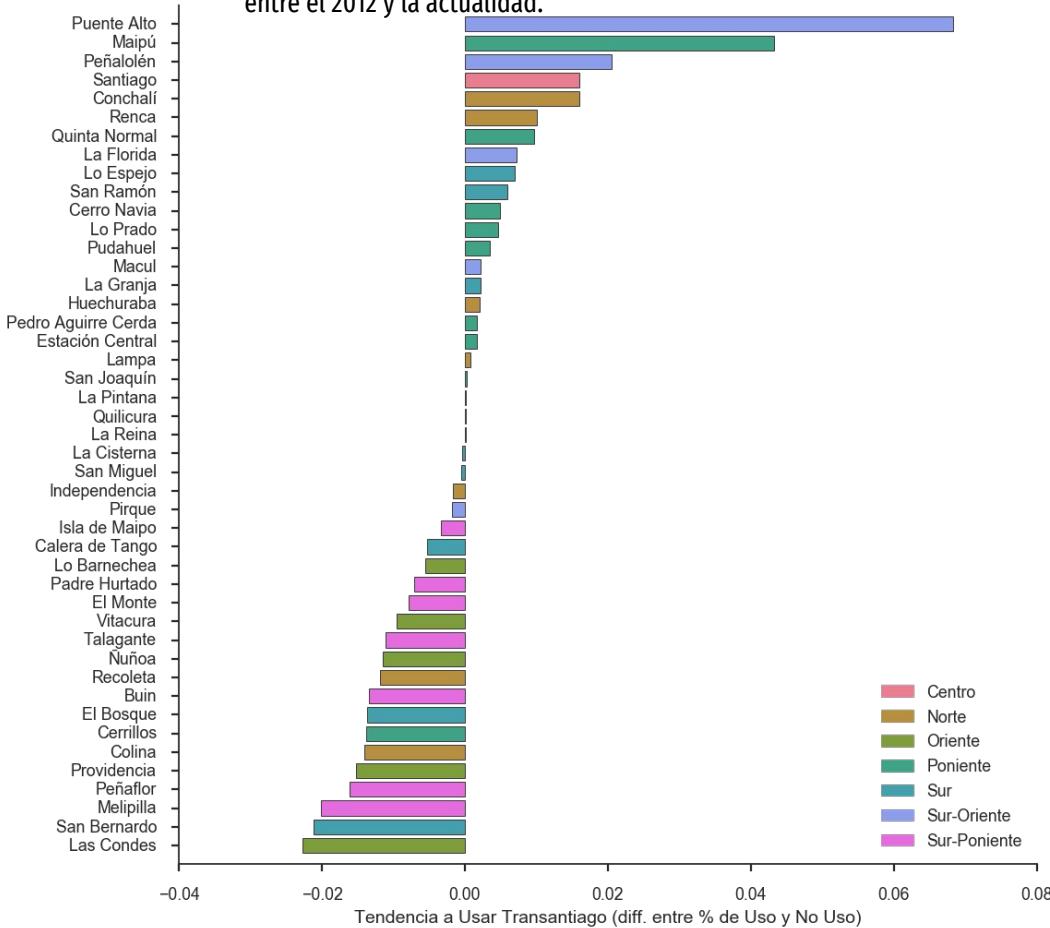
- 1) largo de la barra para expresar valores cuantitativos.
- 2) posición en el espacio: una por marca. Separadas horizontalmente, alineadas verticalmente. Las líneas pueden ser ordenadas (por ej., orden alfabetico).

Tareas: comparar, encontrar valores, encontrar extremos.

Escalabilidad: decenas de atributos.



Comparación entre el uso de transporte público en Santiago entre el 2012 y la actualidad.



Stacked Bar Chart

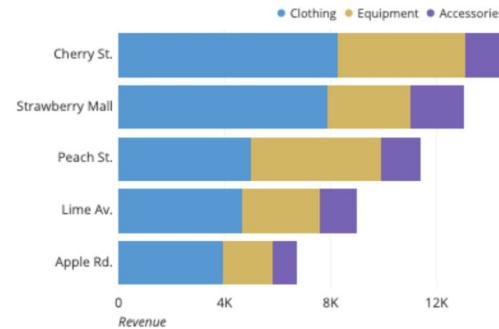
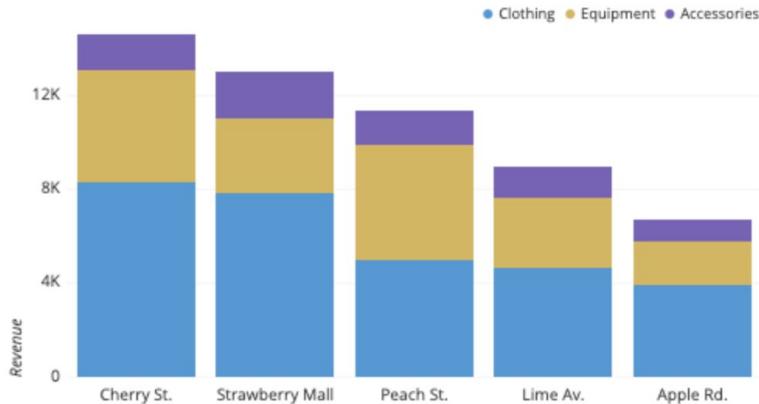
Datos: una llave adicional: 2 atributos categóricos, 1 atributo cuantitativo.

Marca: pila vertical de múltiples líneas, un **glifo** (objeto compuesto por múltiples marcas).

Canales: largo, tono de color (hue), regiones espaciales (una por glifo).

Tarea adicional: ver [relaciones de parte con el todo](#).

Escalabilidad: hasta una docena de niveles para el atributo utilizado en la pila.



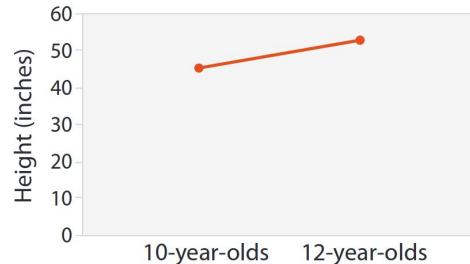
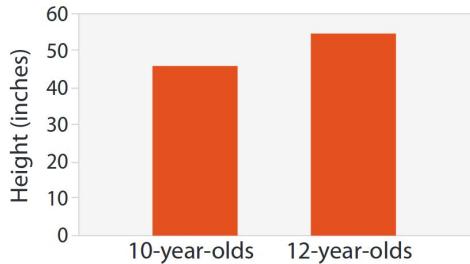
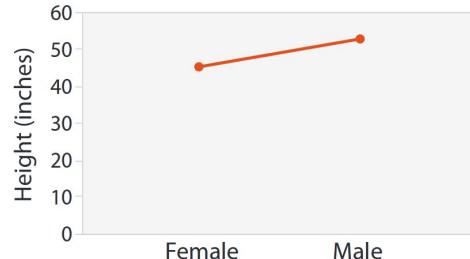
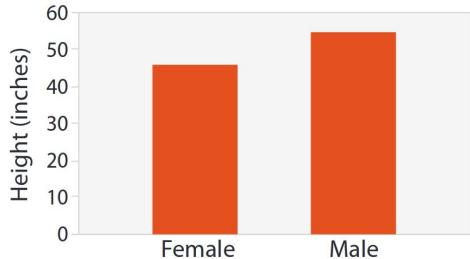
¿Barras o Líneas?

Depende del tipo de llave

- barras si es categórico
- líneas si es ordinal o cuantitativo

NO usar líneas con llaves categóricas no ordinadas.
Hacerlo viola el principio de expresividad. Como
consecuencia tendemos a ver tendencias o patrones
muy fuertes que no respetan la semántica

**“Mientras más hombre sea el sexo de una persona,
más alta es”**



Basado en [Bars and Lines: A Study of Graphic Communication](#), Zacks and Tversky.
Memory and Cognition 27:6 (1999), 1073–1079.

Heatmap

Datos: dos llaves, un valor: 2 atributos categóricos, 1 atributo cuantitativo.

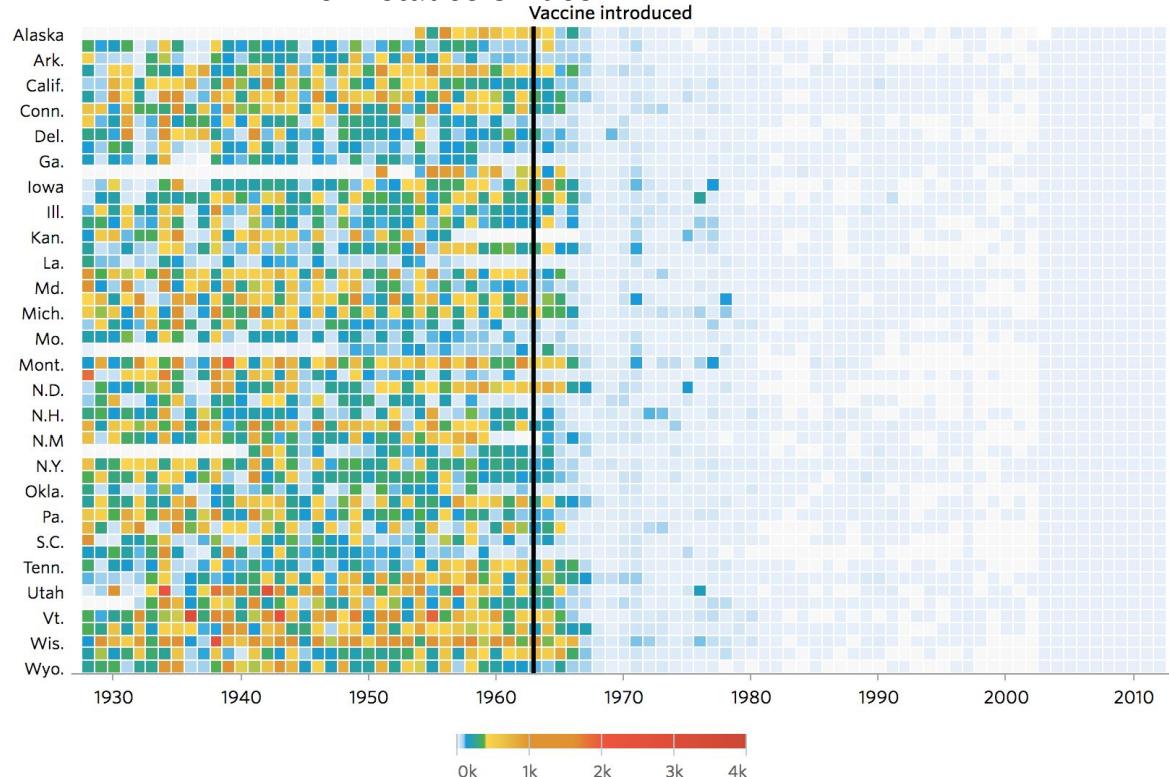
Marcas: área, separadas y alineadas en una matriz 2D.

Canales: color determinado por el atributo cuantitativo, posición en los ejes x e y.

Tareas: encontrar clusters, outliers.

Escalabilidad: Su escalabilidad es flexible, pudiendo mostrar hasta 1 millón de áreas, cientos de niveles categóricos, y aproximadamente 10 niveles en el atributo cuantitativo

Introducción de la Vacuna del Sarampión en Estados Unidos

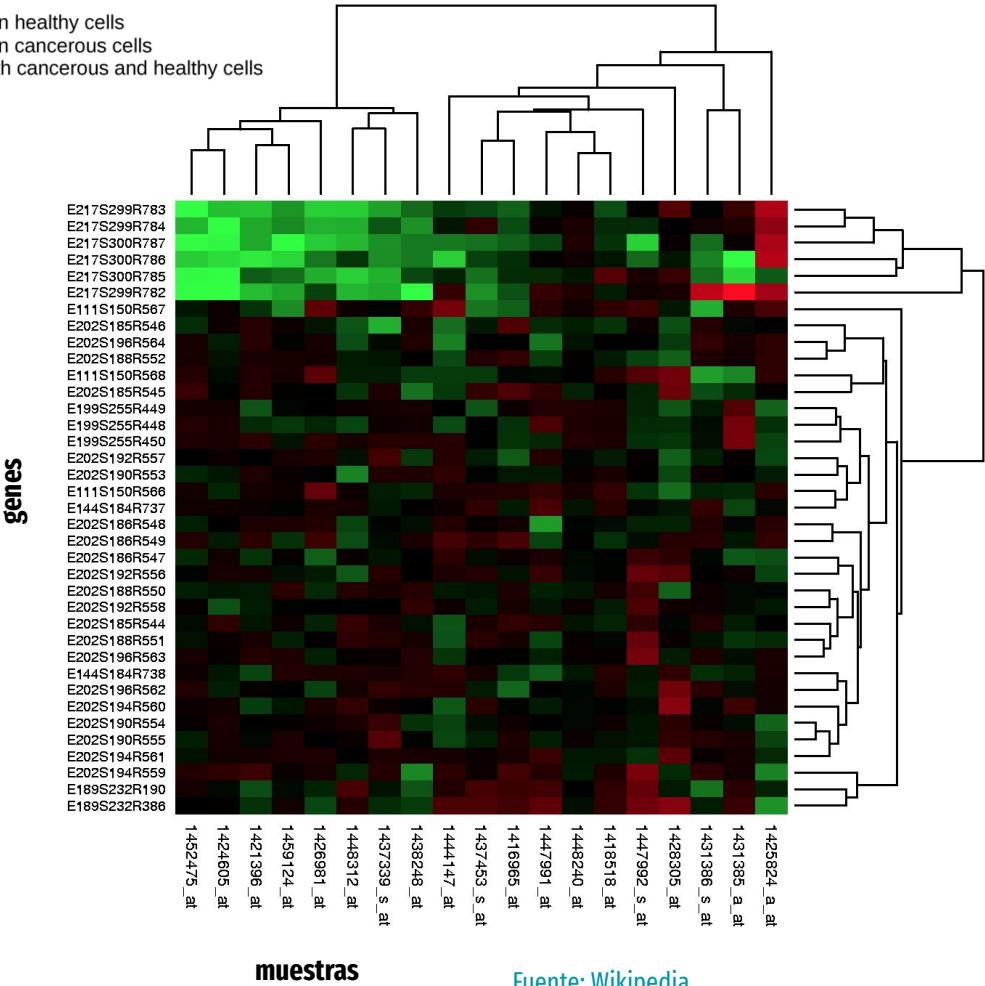


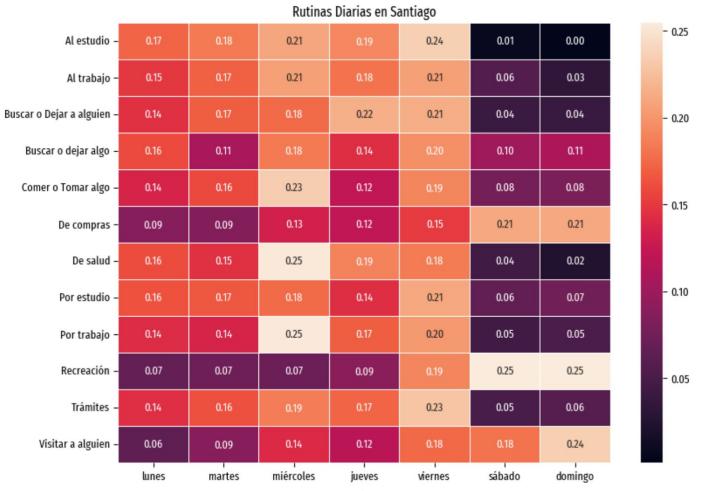
Note: CDC data from 2003-2012 comes from its Summary of Notifiable Diseases, which publishes yearly rather than weekly and counts confirmed cases as opposed to provisional ones.

Cluster Heatmap

- Heatmap ordenado. Se utiliza para representar la intensidad de los datos en forma de colores y para agrupar (o "clusterizar") datos similares utilizando métodos de agrupación jerárquica.
- Es común en estadística, biología computacional y otras disciplinas que tratan con grandes conjuntos de datos, como la expresión genética o el análisis de datos multidimensionales.
- Dendrograma: Representaciones gráficas de la estructura de agrupamiento que muestran cómo se agrupan los elementos individuales y qué tan similares son entre sí basados en la distancia o similitud.

- only expressed in healthy cells
- only expressed in cancerous cells
- expressed in both cancerous and healthy cells





¡Usamos un clustermap para organizar los resultados!



¿Cómo caracterizar las rutinas de Santiago en función de los viajes que realizan las personas? Lo podemos calcular a partir de la Encuesta Origen Destino.

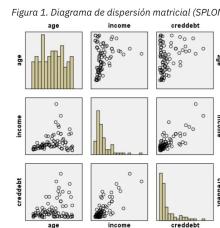
Scatterplot Matrix (splom)

La visualización explora todos los pares posibles de atributos con un scatterplot, heredando su codificación visual.

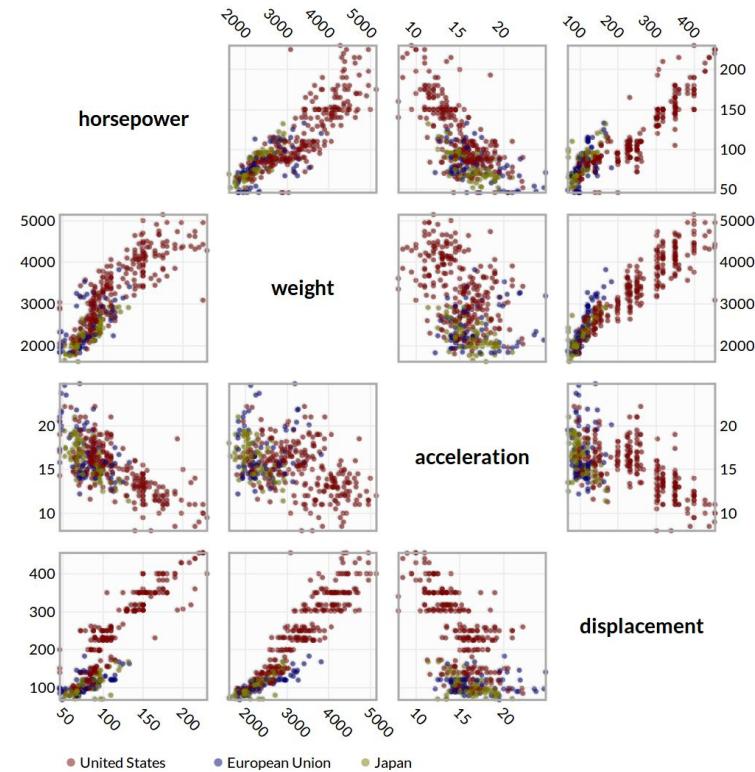
Datos: múltiples atributos cuantitativos, agrupados por pares.

Escalabilidad: una docena de atributos, decenas a cientos de ítems.

Tareas: identificar variables relevantes de acuerdo a correlaciones.



Scatter Plot Matrix of Automobile Data



Four dimensions of a database of cars plotted in a scatter plot matrix, with different colors to indicate the country of origin. Each pair of variables is represented in two (transposed) plots. Dragging a rectangle on any of the graphs highlights the selected points in all the graphs, a technique called *brushing and linking*.

Source: [GGobi](#)

Pie chart, Polar area Chart

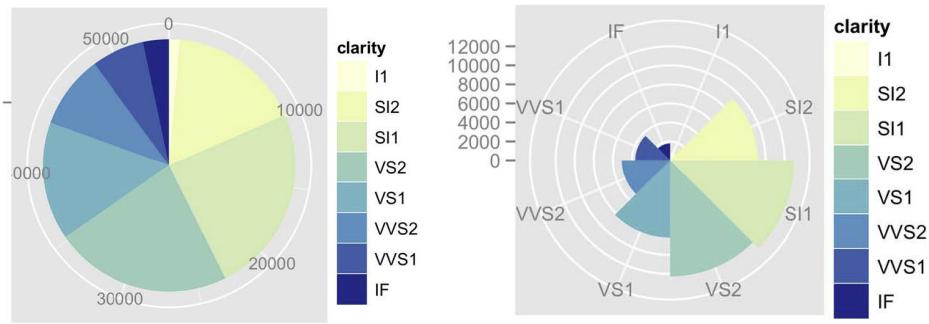
Pie chart: marcas son áreas con un canal angular. **Cuidado: ángulo y área son menos precisos que el largo de una línea o barra.**

Polar área chart: áreas con canal angular y de largo.

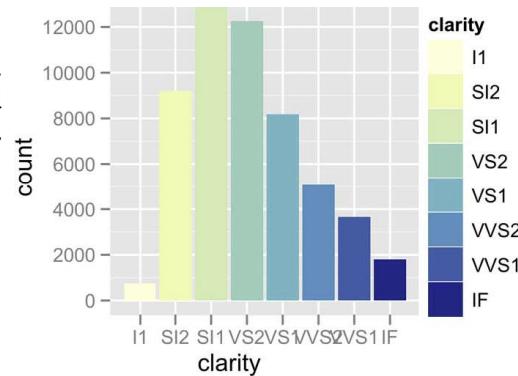
Datos: 1 atributo llave categórico, 1 atributo cuantitativo

Tareas: entendimiento de parte-de-un-todo.

Densidad de información: para poder entender las relaciones se necesita un círculo de gran tamaño.



[A layered grammar of graphics](#). Wickham.
Journ. Computational and Graphical Statistics 19:1 (2010), 3-28.



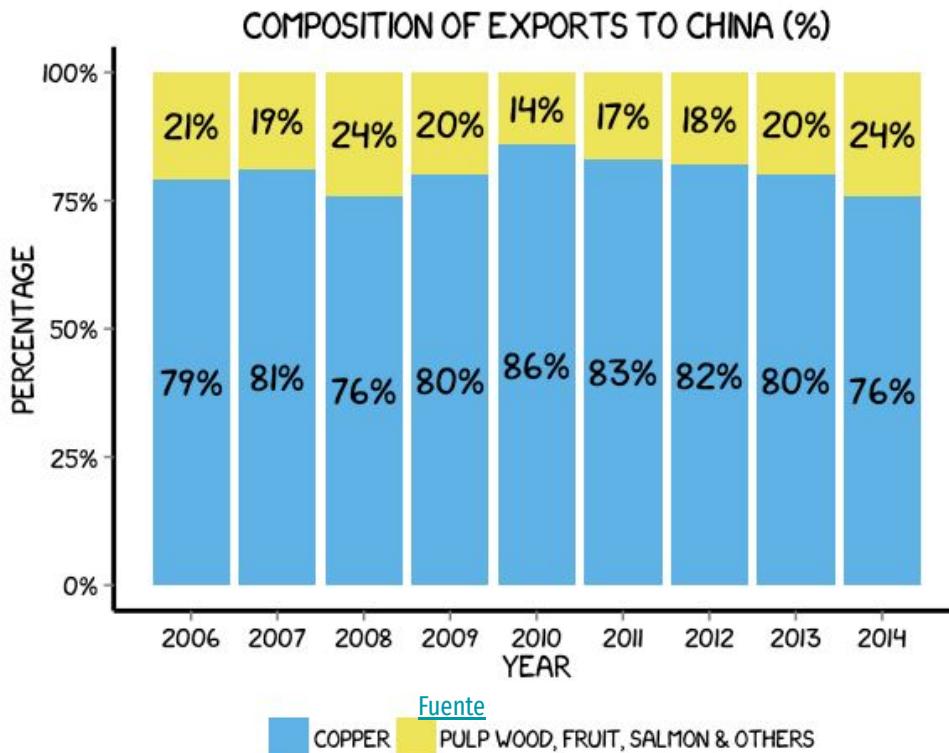
Normalized stacked bar chart

Tarea: entender relaciones parte-de-un-todo.

Normalized stacked bar chart: es un stacked bar chart, normalizado para que utilice todo el espacio vertical.

Una barra es el equivalente a un gráfico de torta (pie chart).

En contraste, tiene alta densidad de información: un rectángulo delgado codifica la distribución.



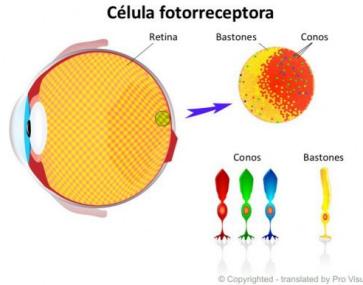
Resumen

- La mayoría de las visualizaciones que vimos hoy ya eran conocidas y las hemos utilizado en Python.
- La manera de describirlas utilizando marcas y canales es nueva, y permite indagar en cómo se define una visualización en función del qué (datos) y el cómo (codificación).
- Así, podemos comparar entre distintas visualizaciones para elegir la mejor opción en función de la tarea a realizar (¿por qué?).

2. Modelos de Color

¿Cómo Distinguimos Colores?

Nuestros ojos tienen dos tipos de células fotorreceptoras: los bastones y los conos. Estos últimos son los encargados de aportar la información del color, y hay de tres tipos: rojo (R), verde (G) y azul (B).



Color en una Visualización

En el caso de visualización de datos, la decisión de usar un color o una paleta de colores, no solo pasa por su aspecto estético. **Sino por consideraciones respecto de cómo representar los datos de forma efectiva.**

Modelos de Color

Existen muchos modelos de color. Los más comunes son **RGB** (usado en pantallas) y **CMYK** (usado para impresión).

- **Modelos Aditivos:**

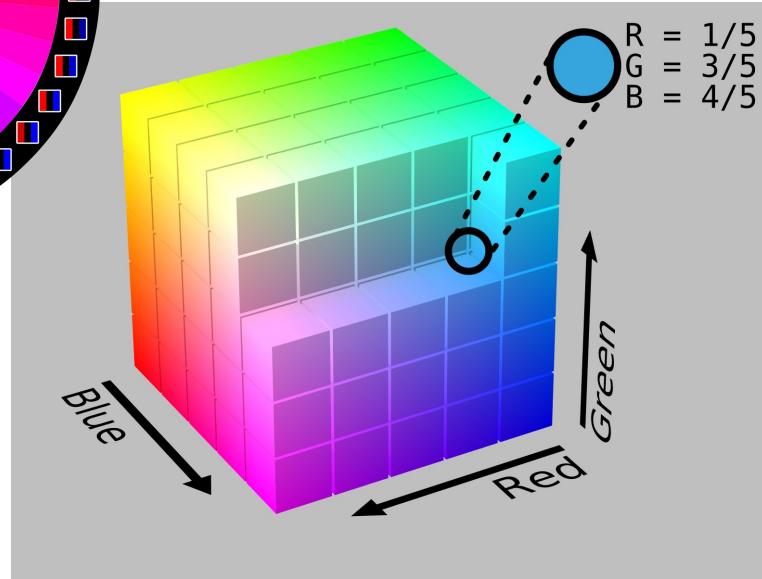
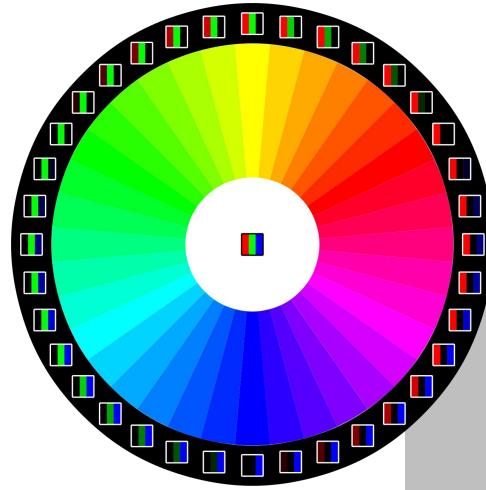
En los modelos de color aditivos, la luz se utiliza para mostrar colores. Y al agregar todos los colores del modelo, se llega al color blanco. Este tipo de modelo se usa para desplegar colores en pantallas de televisión, de computadores o de celulares.

- **Modelos Sustractivos:**

Por otro lado, en los modelos de color sustractivos, al agregar varios colores vamos llegando a matices más oscuros. Como un café muy oscuro o negro, este tipo de modelo se usa en impresoras.

Modelo RGB

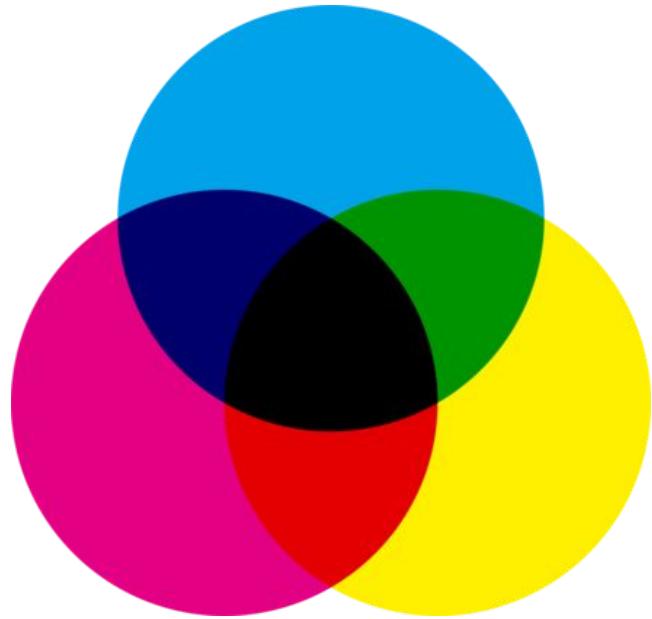
El **modelo de color RGB** define un espacio cúbico en el cual cada color es la suma ponderada de los tres colores primarios.



Fuente: [Wikipedia](#)

Modelo CMYK

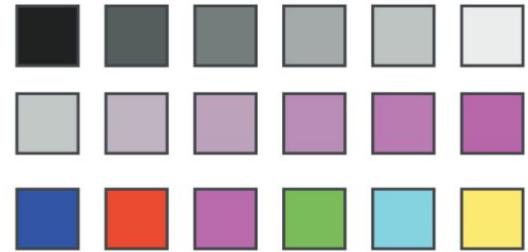
- Es un modelo sustractivo cuya sigla CMYK hace referencia a cuatro colores: cian (cyan), magenta, amarillo (yellow) y negro (black)
- Como regla general, cuanto más pigmento haya en el papel blanco, menos luz blanca refleja este; en otras palabras, la tinta restará luz blanca en cantidades crecientes.
- Si ciertos colores se combinan y se superponen, aparecerán tonos más oscuros de rojo, verde y azul, hasta llegar al negro.



Modelo HSL

Modelo HSL. Este modelo se identifica por sus siglas, el matiz o tono, en inglés hue. La saturación, en inglés saturation. Y la luminosidad, lightness o luminosity en inglés. El valor H o hue, permite describir el matiz o tono del color que se va a generar. Esto sería un rojo, verde, azul, amarillo, etc.

Luminance
Saturation
Hue



La saturación es el grado de pureza de un color, es la propiedad que define la intensidad del mismo.

La luminosidad es la proporción o los niveles de blanco o de negro que contiene un color. La luminosidad es la encargada de calificar a un color como claro u oscuro.

Interpretación del Color para Representar Datos

El modelo HSL es muy usado para la visualización de datos. Porque al separar la representación del color en tres aspectos, es posible controlar su uso para diferentes escalas de datos.

- Es posible codificar visualmente **datos categóricos** con el matiz o tono, el H.
- La saturación y luminosidad permiten codificar visualmente **datos ordinales o continuos, magnitudes**.

Paletas de Colores

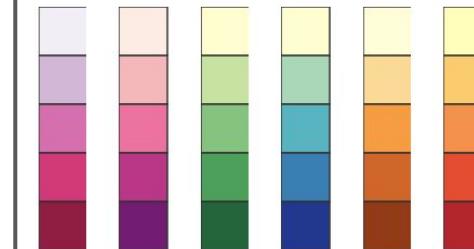
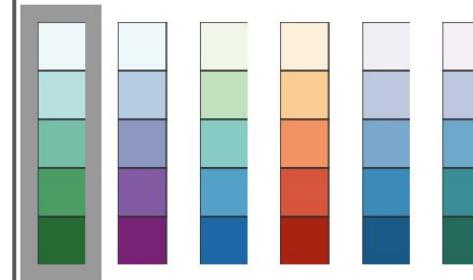
Paleta de colores secuencial:

Cuando tengo datos secuenciales, hay un orden o hay gradación.

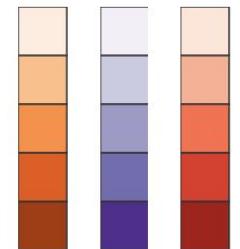
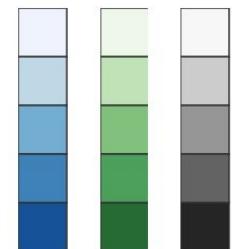


Pick a color scheme:

Multi-hue:



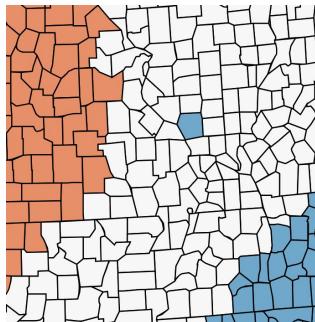
Single hue:



Paletas de Colores

Paleta de colores divergente:

Se usan en escenarios en los que hay un orden entre los valores pero hay también un "centro" definido. Por ejemplo, alturas con respecto al nivel del mar (que pueden ser positivas, negativas o cero).



Number of data classes: 3 ▲ ▼

Nature of your data: sequential diverging qualitative

Pick a color scheme:

1

2

3

4

5

6

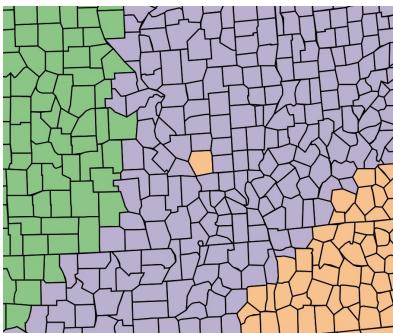
7

8

9

Paletas de Colores

Paleta de Colores Cualitativa: Cuando trabajamos con variables categóricas con cuyos valores no es posible establecer una relación de orden, debemos utilizar una paleta de colores que comuniquen independencia entre valores.



Number of data classes: 3 i

Nature of your data:

sequential diverging qualitative i

Pick a color scheme:

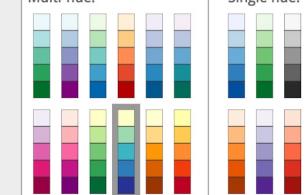
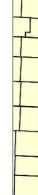


Color Brewer

Hay herramientas que nos facilitan elegir y configurar paletas de colores, incluyendo compatibilidad con impresión, fotocopias, y ceguera de colores.

Number of data classes: 8 i

Nature of your data:
 sequential diverging qualitative

Pick a color scheme:
Multi-hue:  Single hue: 

Only show:
 colorblind safe i
 print friendly
 photocopy safe

8-class YIGnBu i

RGB 

255,255,217
237,248,177
199,233,180
127,205,187
65,182,196
29,145,192
34,94,168
12,44,132

Context:
 roads
 cities
 borders

Background:
 solid color 
 terrain 
 color transparency

Export your selected color scheme:
Permalink
Share a direct link to this color scheme.
<http://colorbrewer2.org/?type=sequential&sch>

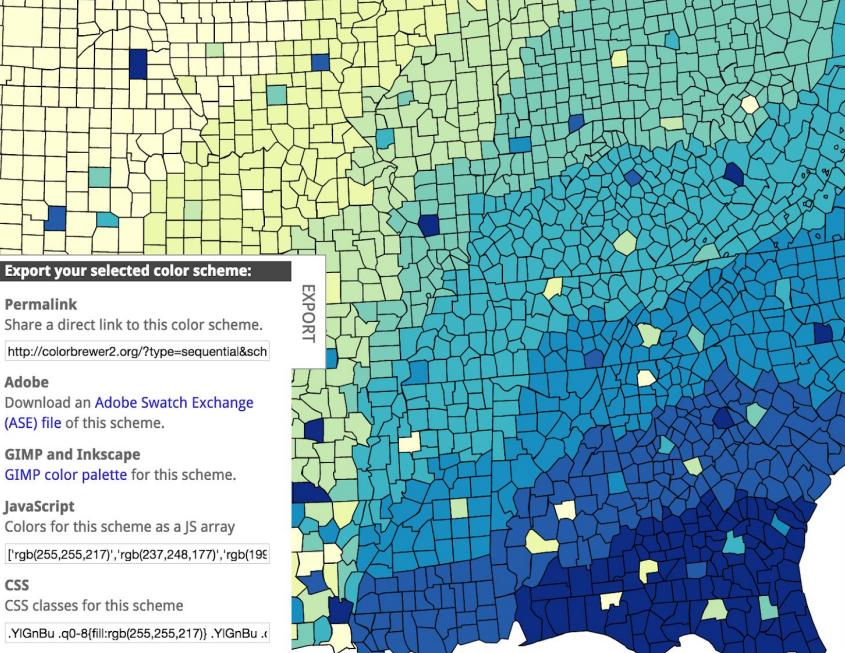
Adobe
Download an [Adobe Swatch Exchange \(ASE\) file](#) of this scheme.

GIMP and Inkscape
[GIMP color palette](#) for this scheme.

JavaScript
Colors for this scheme as a JS array
`[rgb(255,255,217),rgb(237,248,177),rgb(199,233,180),rgb(127,205,187),rgb(65,182,196),rgb(29,145,192),rgb(34,94,168),rgb(12,44,132)]`

CSS
CSS classes for this scheme
`.YIGnBu .q0-8{fill:rgb(255,255,217)} .YIGnBu .q1{fill:rgb(237,248,177)} .YIGnBu .q2{fill:rgb(199,233,180)} .YIGnBu .q3{fill:rgb(127,205,187)} .YIGnBu .q4{fill:rgb(65,182,196)} .YIGnBu .q5{fill:rgb(29,145,192)} .YIGnBu .q6{fill:rgb(34,94,168)} .YIGnBu .q7{fill:rgb(12,44,132)}`

EXPORT 



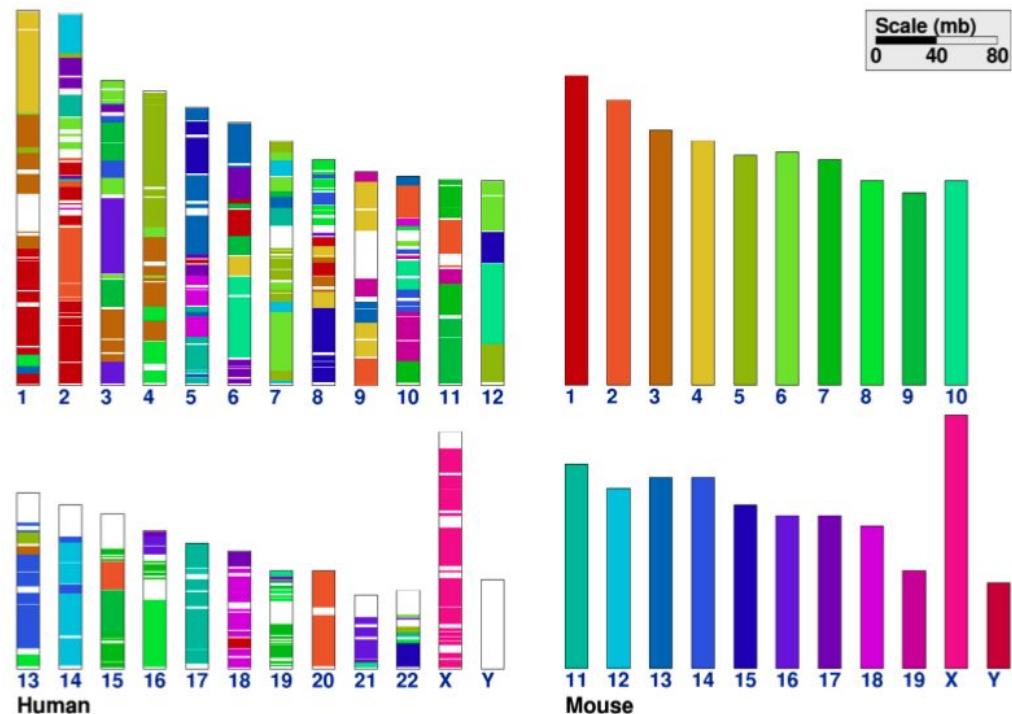
<http://colorbrewer2.org/?type=sequential&scheme=BuGn&n=3>

Consideraciones: Discriminabilidad

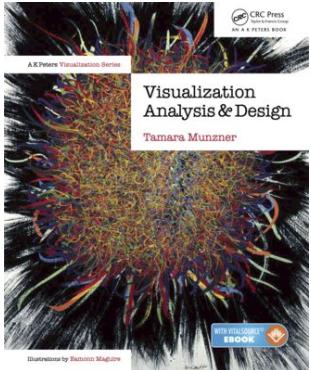
Al visualizar categorías nos cuesta mucho diferenciar colores distintos pero similares.

Por eso debemos usar colores discontiguos.

Entre 6 y 12 categorías está bien. Más se vuelve imposible de percibir.



¿Preguntas?



Esta clase incluye material del libro
Visualization Analysis & Design de Tamara
Munzner.
<http://www.cs.ubc.ca/~tmm/vadbook/>