Fundamentos de Ciencias de Datos

Semana 04 - Introducción a la Visualización de Datos

Visualización de datos

La visualización de datos es el componente vital en el análisis de datos, ya que con ella se pueden resumir grandes cantidades de información a través de un gráfico

Existen muchos tipos de gráficos, cada uno para casos de uso específicos

Uno de los procesos en el análisis de datos es escoger el gráfico adecuado para representar los datos y la información que desea entregar

Fuente: How to Choose the Right Data Visualization

Visualización de datos

- ¿Qué deseo representar?
- Mostrar los cambios en tiempo
- Mostrar una parte de todos los datos
- Representar flujos y procesos
- Ver como se distribuyen los datos
- Comparar valores entre grupos
- Observar valores entre variable
- Representar datos geográficos

Visualización de datos

- ¿Qué datos tengo?
- ¿Cómo son mis datos?
- ¿Cuántas variables voy a graficar?

El tipo de gráfico que se va a utilizar dependerá del tipo de datos que tengamos:

- Categórico
- Numerico
- Combinación de ambos

Essential Charts for Data Analysis

Raw Number

Single Value Chart

Show a raw singular value

Single Value w/ Indicator

Comparison of a single value against a previous value

Bullet Chart

Comparison of a single value against a benchmark value

Table

Show raw numbers for multiple data points on multiple variables

Change over Time

Line Chart

Change over time for a numeric variable or to compare 1-5 groups

Sparkline

Miniature line charts to compare many groups

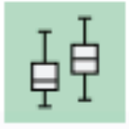
Distribution

Bar Chart

Comparison or distribution by a single categorical variable

Histogram

Distribution by a binned single numeric variable



Box Plot

Compare distribution summaries across a categorical variable

Part-to-Whole

Pie Chart

Part-to-whole breakdown by a single categorical variable

Stacked Bar Chart

Bar chart with additional part-to-whole breakdown

Stacked Area Chart

Line chart with additional part-to-whole breakdown

Relationship

Scatter Plot

Relationship between two numeric variables

Bubble Chart

Relationship between three numeric variables

Grouped Bar Chart

Comparison or distribution by two categorical variables

Heatmap

Distribution by two binned variables (categorical or numeric)

Geospatial

Bubble Map

Bubble chart built on top of a geographic map

Choropleth

Comparison between geopolitical regions by color

Fuente: How to Choose the Right Data Visualization

A continuación se presentarán algunos de los principales gráficos que estudiaremos en este curso

Tipos de Gráficos

Histograma

Nos da una vista general de la distribución de la población o de la muestra respecto a una característica cuantitativa y continua (ej: longitud, peso)

Se divide la variable por distintos rangos; luego cada barra representa el total de observaciones en cada rango

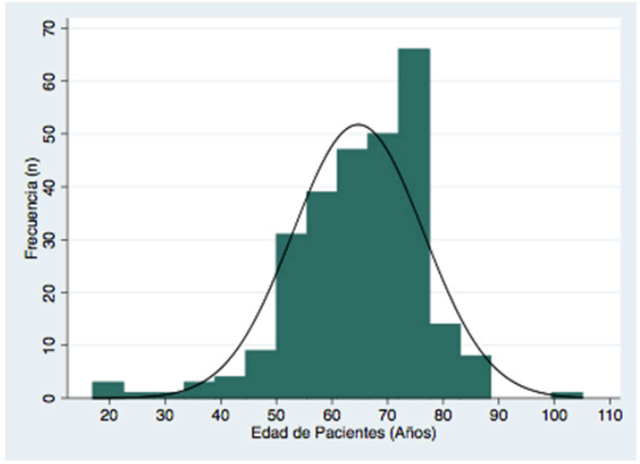


Figura 2. Histograma de edades entre pacientes ingresados a una Unidad de Cuidados Intensivos Cardiovasculares.

Tipos de Gráficos

Gráfico de Barras

Es una forma de resumir un conjunto de datos por categoría mediante barras

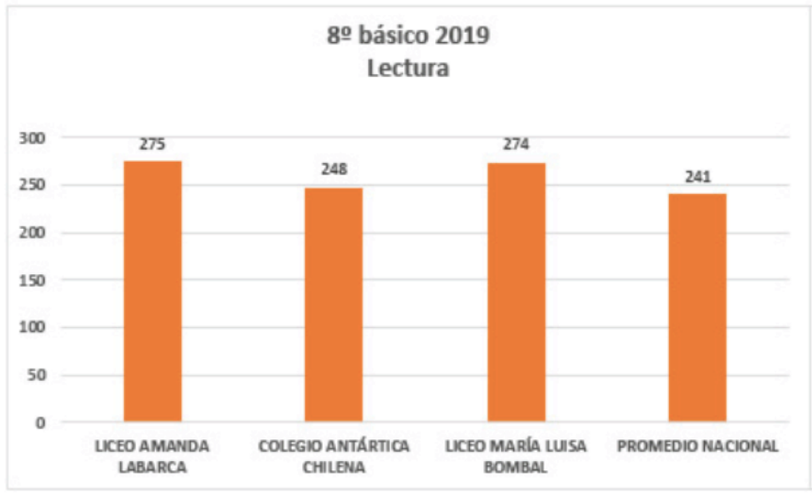
Muestra los datos utilizando varias barras de la misma anchura, en que la altura de la misma nos da información sobre una categoría concreta

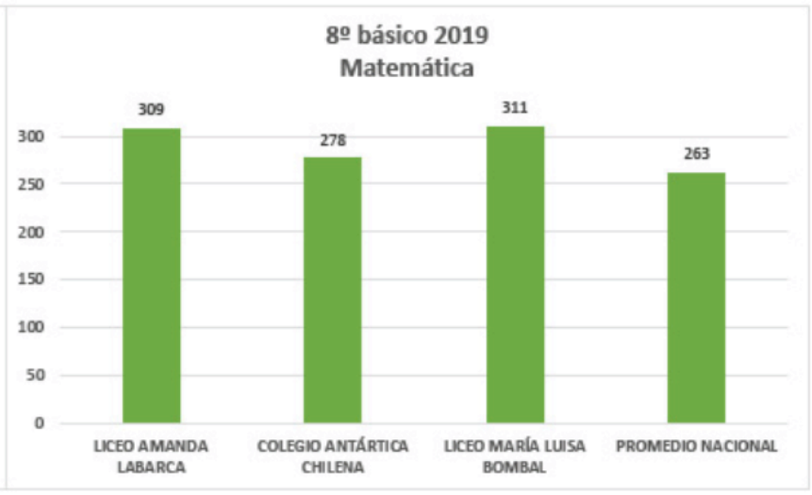
Gráfico de barras para los resultados del SIMCE 8º 2019

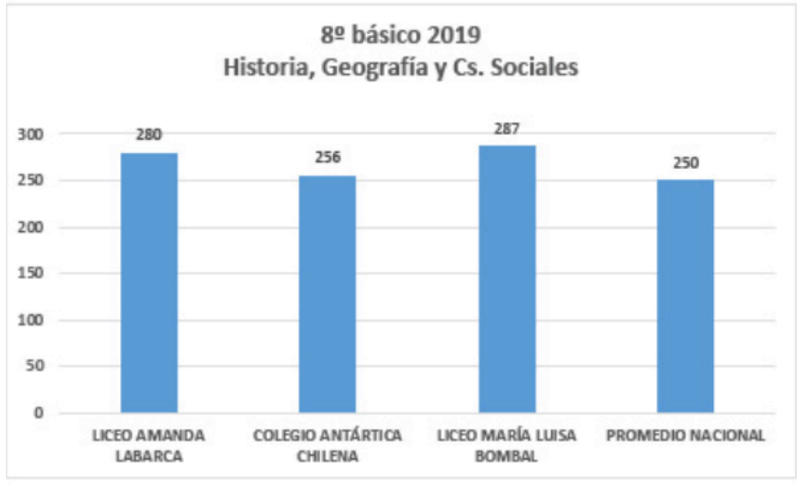
SIMCE 8° BÁSICO 2019

	Amanda Labarca	Antártica Chilena	María Luisa Bombal	Promedio Nacional
Lectura	275	248	274	241
Matemáticas	309	278	311	263
Historia, Geografía y Cs. Sociales	280	256	287	250

*** Agencia de la Calidad entregó resultados de prueba Simce rendida exclusivamente por 8° Básico el 2019.**





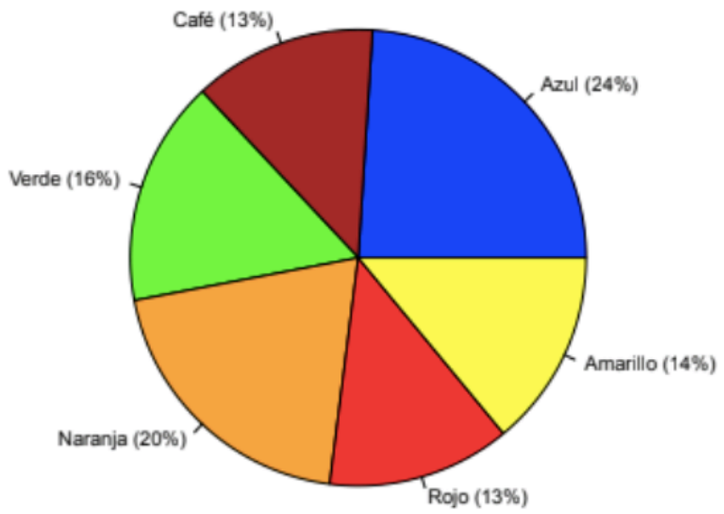


Tipos de Gráficos

Gráficos Circulares

También llamado gráfico de torta, es utilizado para representar magnitudes en frecuencias o porcentajes.

Un circulo se divide en sectores y cada sector representa el porcentaje de cada dato respecto al total de datos



Distribución de colores de bolsitas de M&M(chocolate de leche)

HOW PROFESSORS SPEND THEIR TIME

How they actually spend their time:

Source: Higher Education Research Institute Survey (1999)

How departments expect them to spend their time:

How Professors would *like* to spend their time:

WWW.PHDCOMICS.COM

Tipos de Gráficos

Boxplot

Es un método estandarizado para representar una serie de datos numéricos a través de sus cuartiles

Además, describe varias características importantes al mismo tiempo, tales como la dispersión y la simetría

Los **boxplot** pueden dividirse en dos secciones:

(1) la caja corresponde a la sección central y representa a la mayoría de los datos y

(2) las lineas muestran la variabilidad fuera de la caja, en un límite que corresponde a 1.5 veces el rango intercuartil

Al centro está expresada la mediana (o p50) con una línea horizontal, mientras que el límite superior de la caja es el p75 y el inferior el p25, lo que corresponde al rango intercuartil

Si existen valores más allá de esta frontera, éstos se dibujan como puntos externos y reciben el nombre de valores extremos o outliers

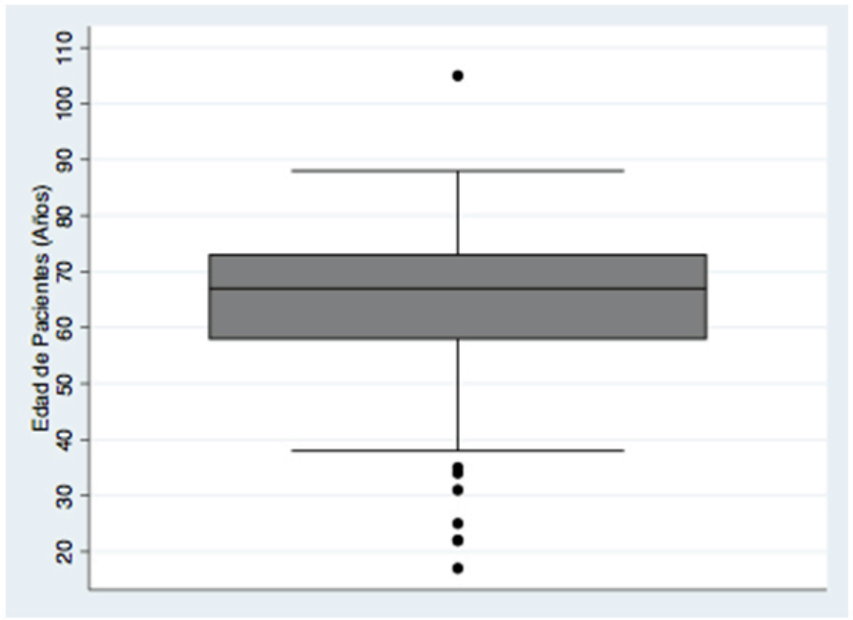


Figura 1. Gráfico de cajas: edades de pacientes ingresados a una Unidad de Cuidados Intensivos Cardiovasculares.

Tipos de Gráficos

Gráfico Poligonal

Representa la frecuencia de los datos a lo largo del tiempo; se usa un eje para el tiempo y otro para los datos

Se agregan puntos para representar cada observación cada dato y se unen los puntos para formar un polígono con el eje horizontal

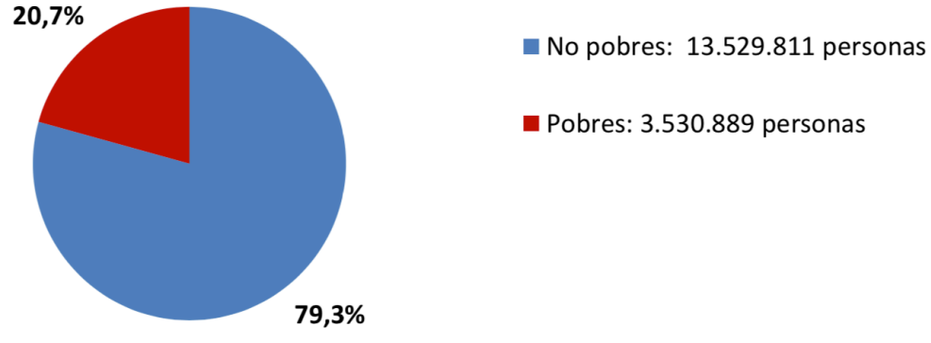
Principios básicos de la visualización

Principios básicos

Usar gráficos simples: mostrar lo que queremos lo más directo posible

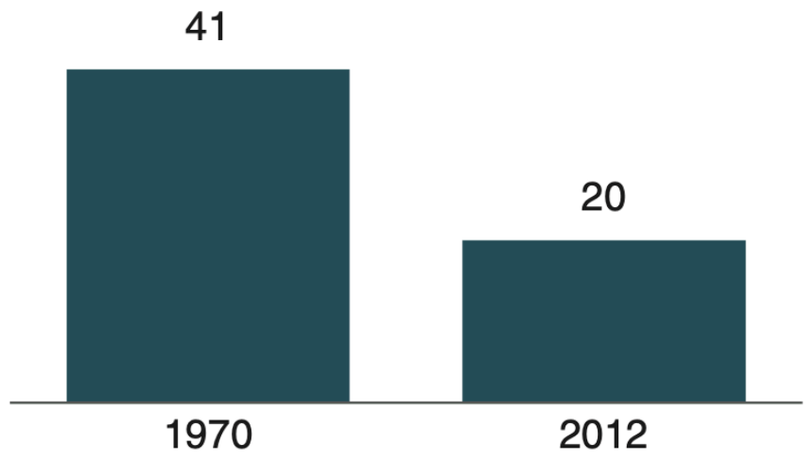
Si tiene pocos datos, quizás es mejor una tabla

Datos de pobreza	Número de personas	Porcentaje de personas
Pobres	3.530.889	20,7%
No pobres	13.529.811	79,3%



Children with a "Traditional" Stay-at-Home Mother

% of children with a married stay-at-home mother with a working husband



Note: Based on children younger than 18. Their mothers are categorized based on employment status in 1970 and 2012.

Source: Pew Research Center analysis of March Current Population Surveys Integrated Public Use Microdata Series (IPUMS-CPS), 1971 and 2013

Adapted from PEW RESEARCH CENTER

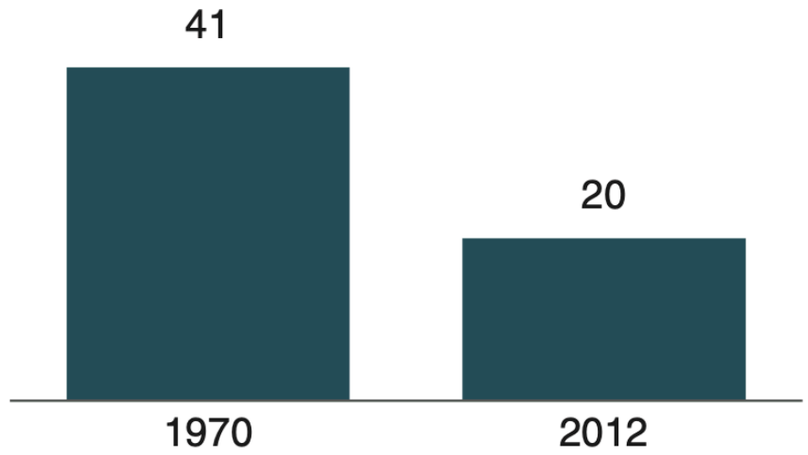
El texto también es efectivo

20%

of children had a
traditional stay-at-home mom
in 2012, compared to 41% in 1970

Children with a "Traditional" Stay-at-Home Mother

% of children with a married stay-at-home mother with a working husband



Note: Based on children younger than 18. Their mothers are categorized based on employment status in 1970 and 2012.

Source: Pew Research Center analysis of March Current Population Surveys Integrated Public Use Microdata Series (IPUMS-CPS), 1971 and 2013

Adapted from PEW RESEARCH CENTER

Principios básicos

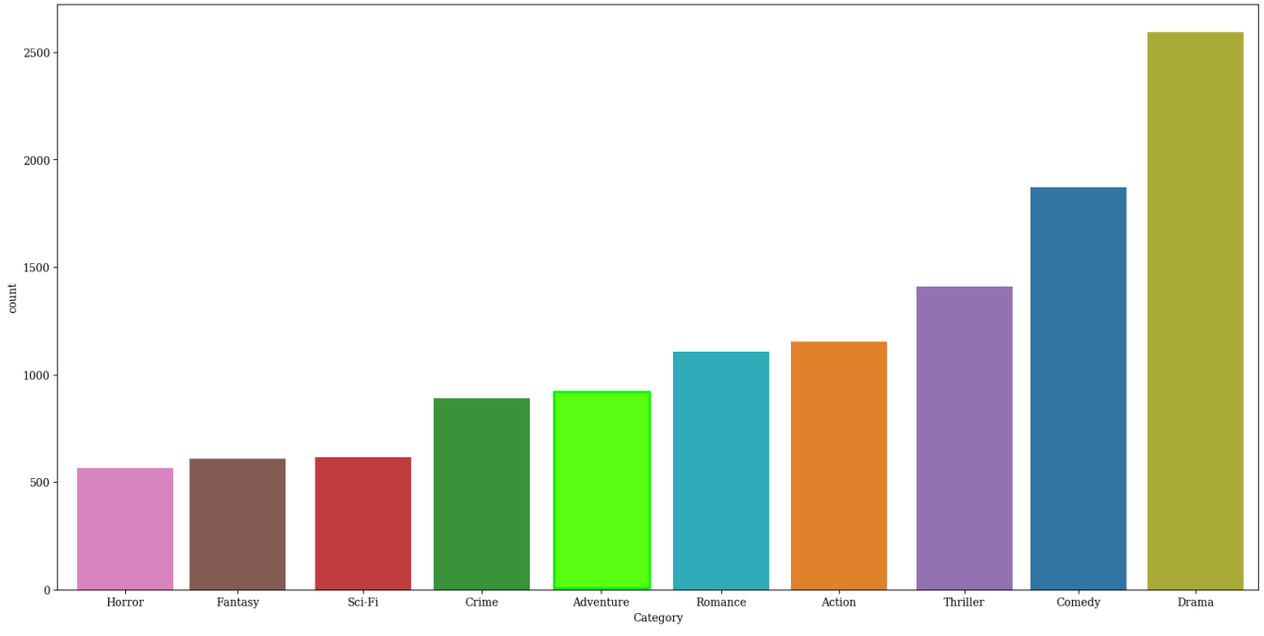
Facilitar la comparación de los datos ordenándolos en base a los valores de la variable, (y en general, nunca ordenar por nombre)

Los colores permiten resaltar lo que uno quiera, pero al mismo tiempo es un distractor, en caso de no ser necesario, es mejor no usar colores distintos

Principios básicos

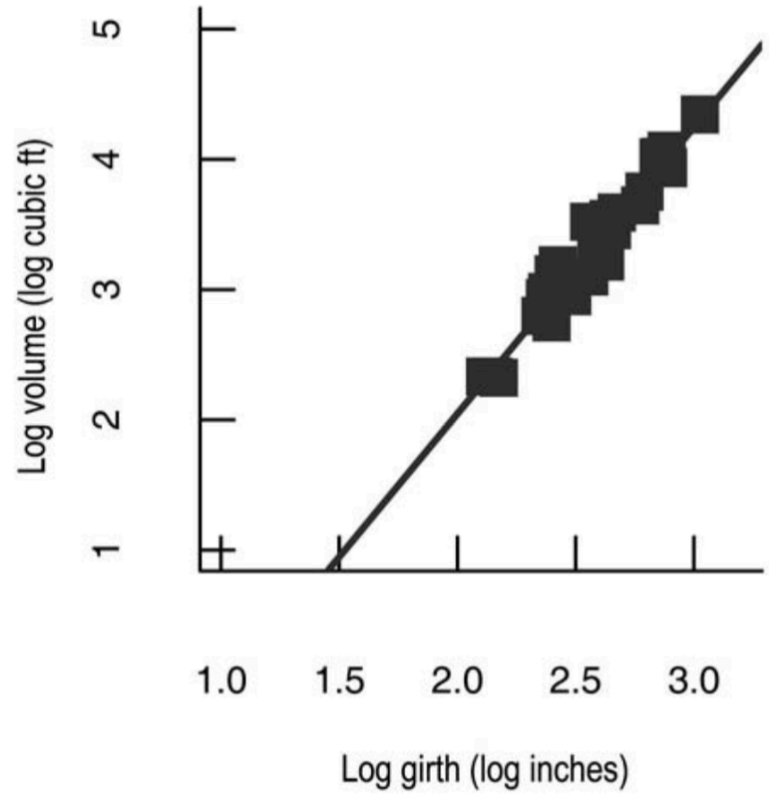
Facilitar la comparación de los datos ordenándolos en base a los valores de la variable, (y en general, nunca ordenar por nombre)

Los colores permiten resaltar lo que uno quiera, pero al mismo tiempo es un distractor, en caso de no ser necesario, es mejor no usar colores distintos



Principios básicos

Muestre toda la información posible: para poder detectar un patrón, toda la información debe estar visible.

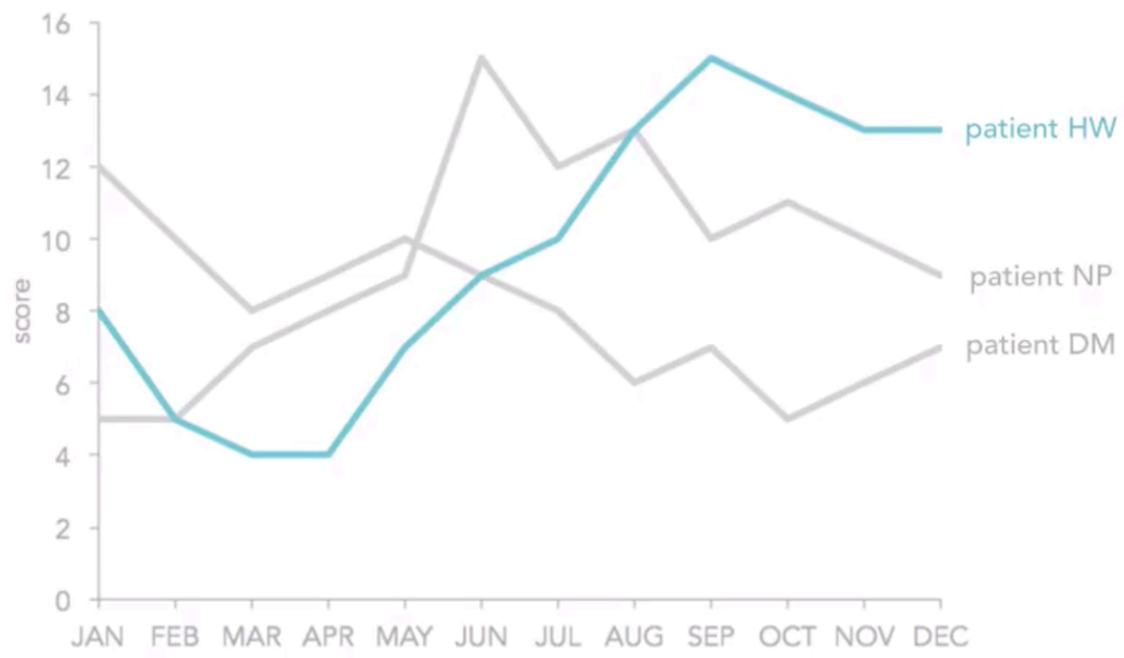


Principios básicos

Resaltar el objetivo: resalta lo que quieres mostrar en el gráfico, si quieres mostrar una serie en particular, solamente colorea solo esa serie

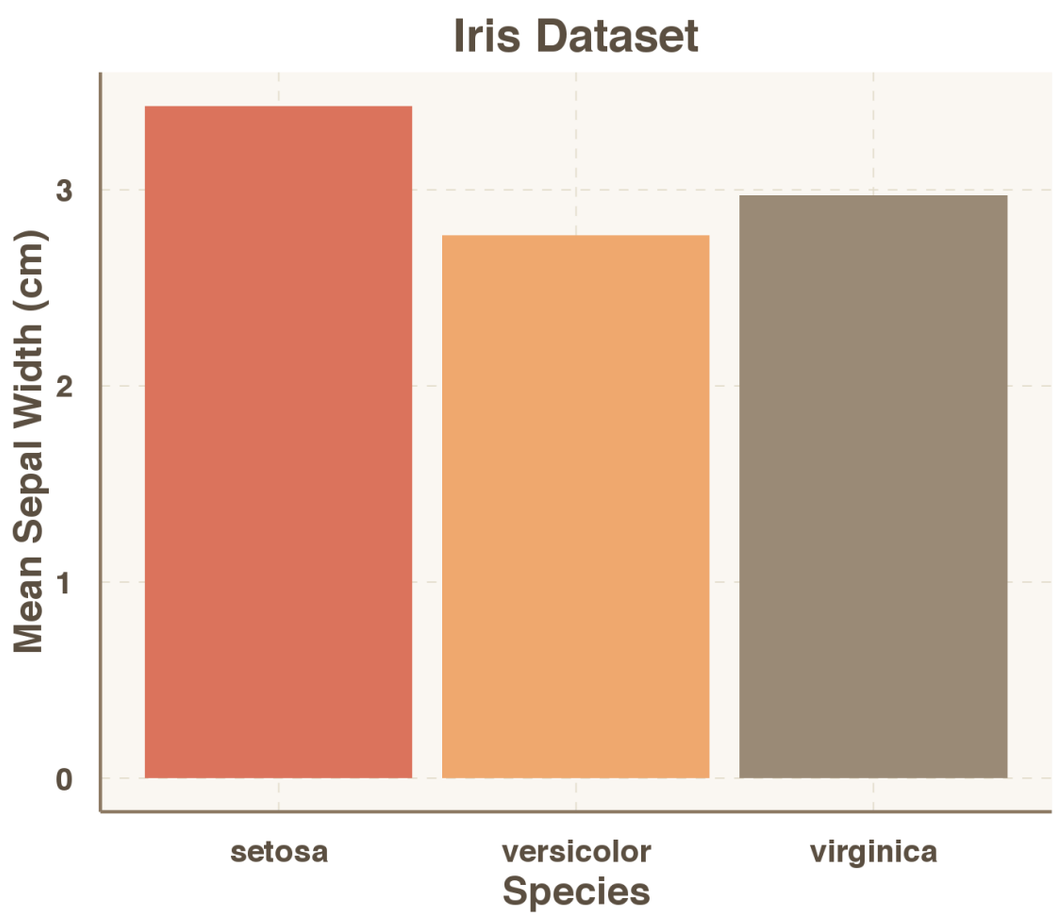
Principios básicos

Resaltar el objetivo: resalta lo que quieres mostrar en el gráfico, si quieres mostrar una serie en particular, solamente colorea solo esa serie



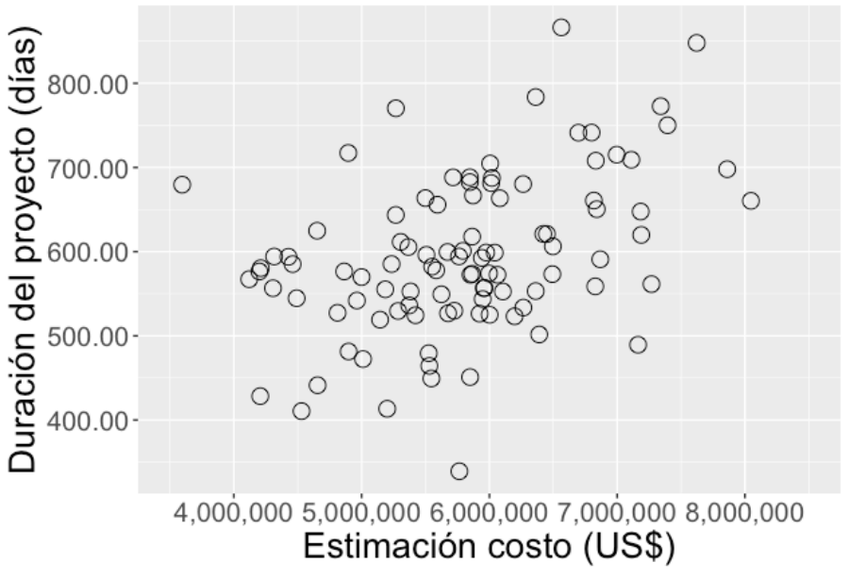
Principios básicos

Evitar leyendas: no usar leyendas en caso de no ser necesarias. Intenta usar etiquetas dentro de los gráficos, si es posible



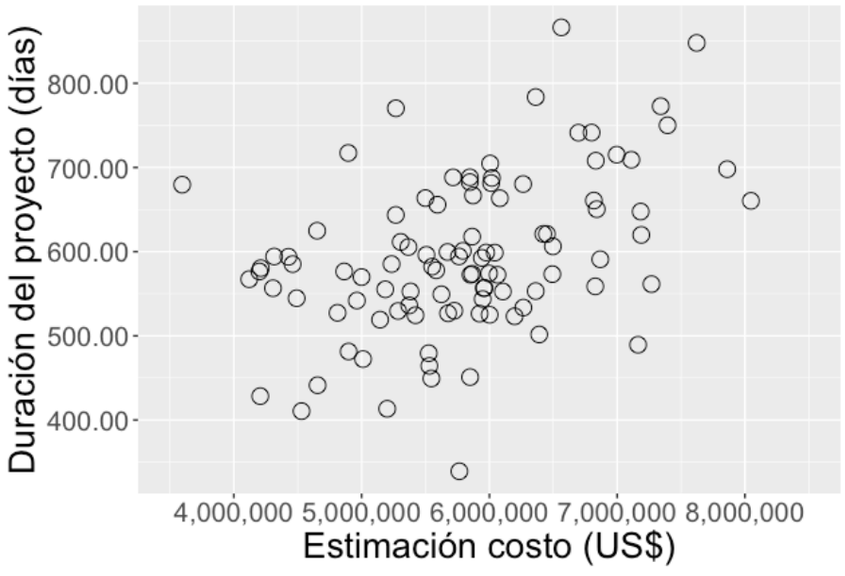
Principios básicos

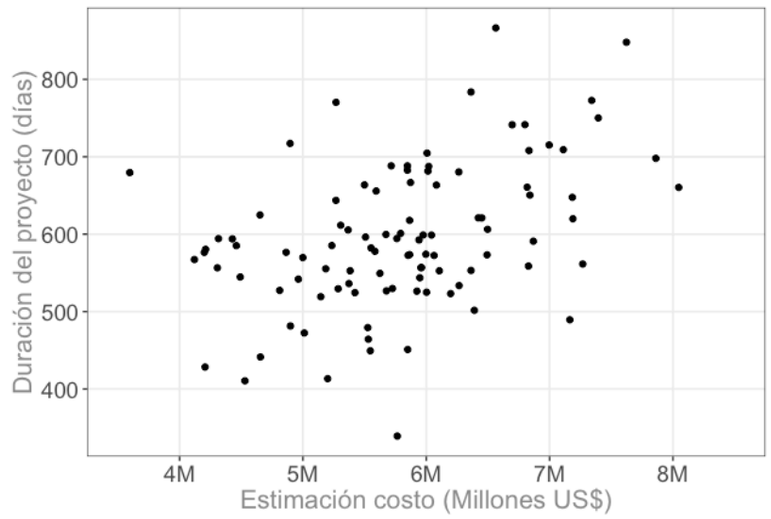
Evitar distracciones: disminuye toda la información que pueda distraer en forma significativa del gráfico que se observa



Principios básicos

Evitar distracciones: disminuye toda la información que pueda distraer en forma significativa del gráfico que se observa

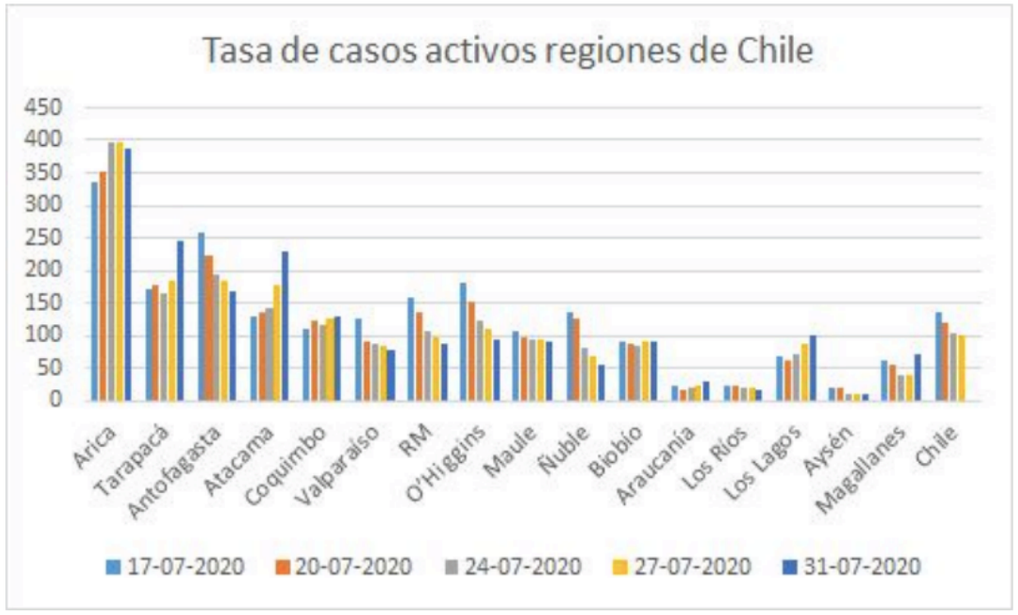




Principios básicos

Facilitar la comparación: los elementos más cercanos son más fáciles de comparar entre sí, evita que exista una gran distancia entre los elementos a comparar

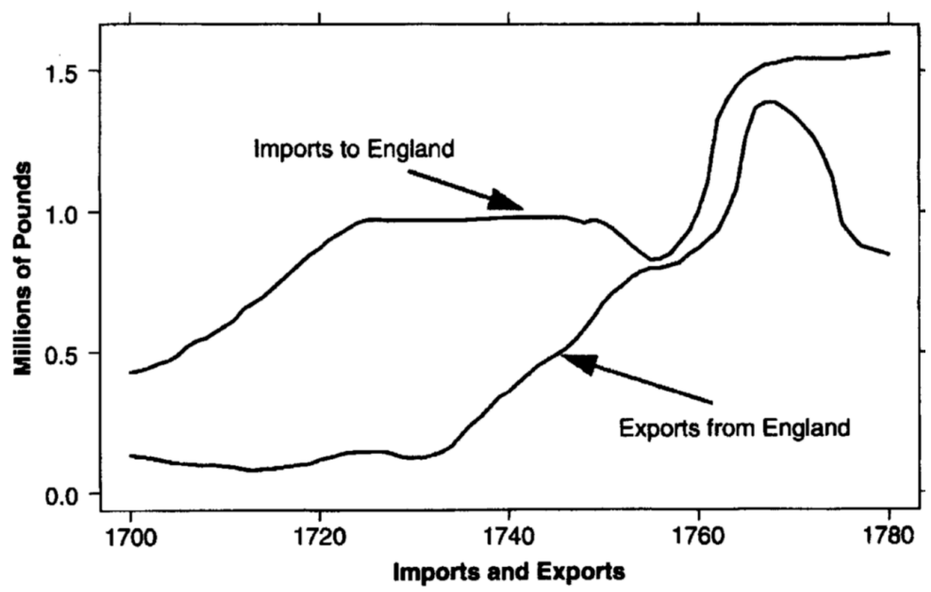
Gráfico 2. Evolución Tasa de casos activos durante entre el 17 y el 31 de julio.



Fuente: Elaboración propia en base a datos MINSAL.

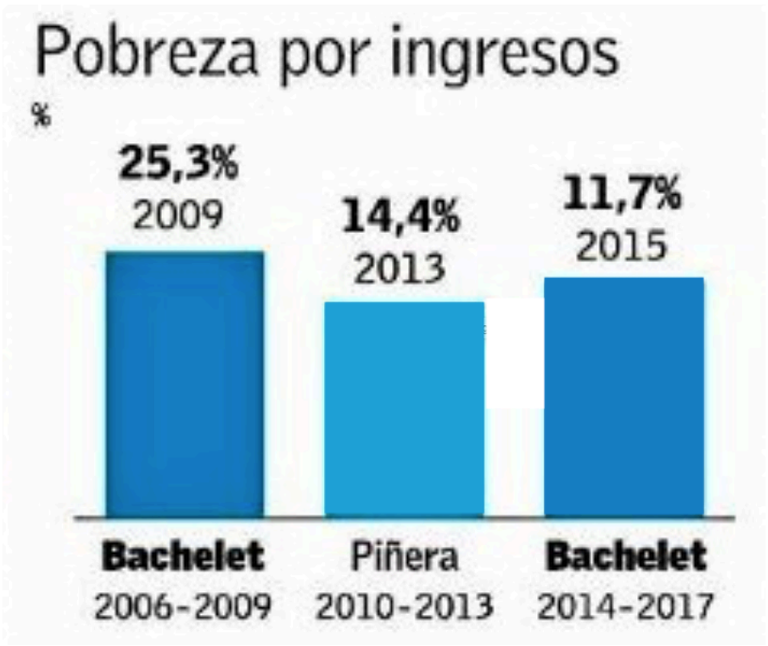
Principios básicos

Evitar cálculos: evita que la persona que observa tenga que realizar cálculos en el gráfico, muestra directamente lo que quieras resaltar



Principios básicos

Incluir el valor 0 en un gráfico de barras: un gráfico de barras tiene necesita tener el valor 0 de forma explícita, o podemos generar confusión



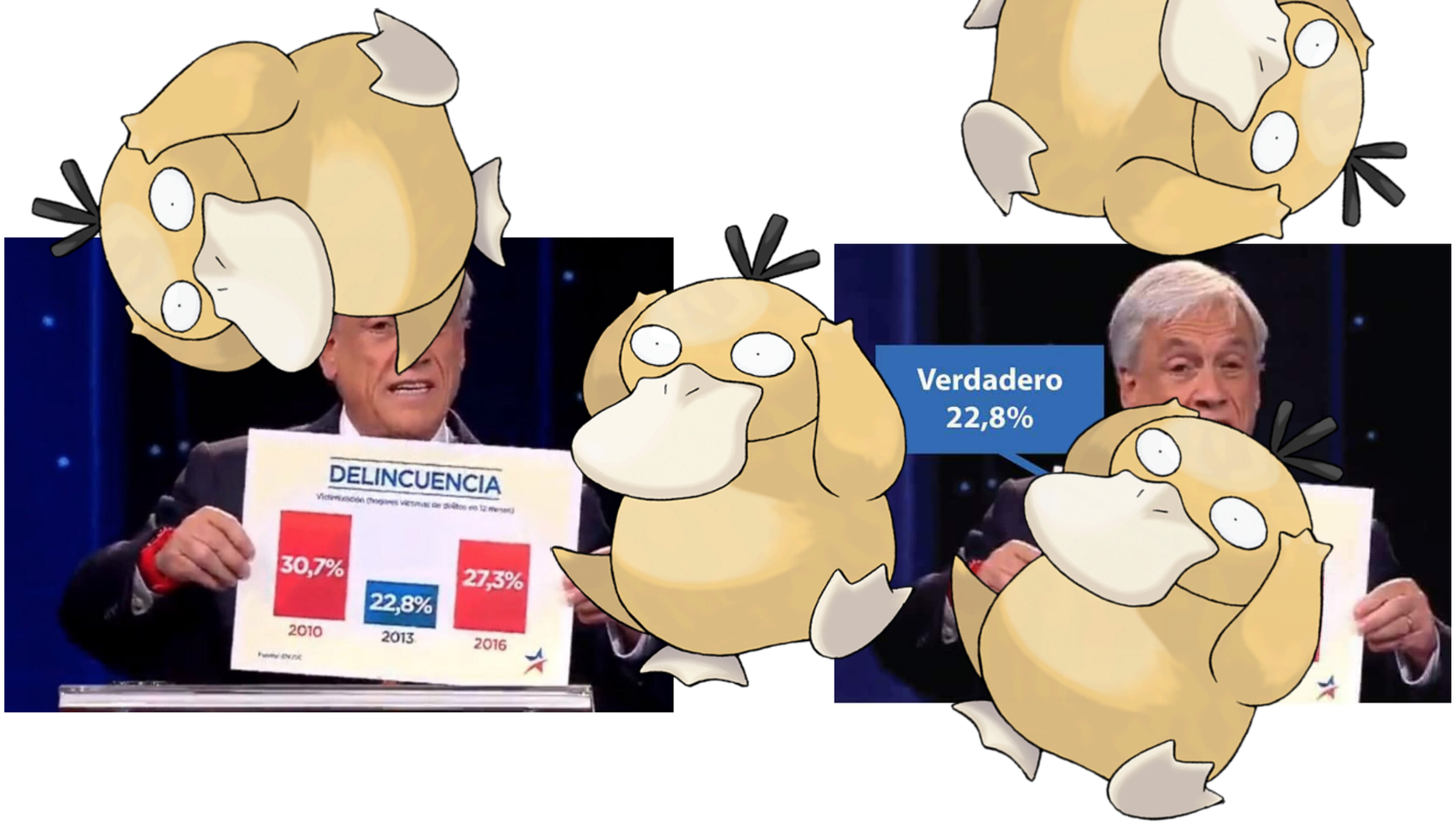
Principios básicos

Principios básicos

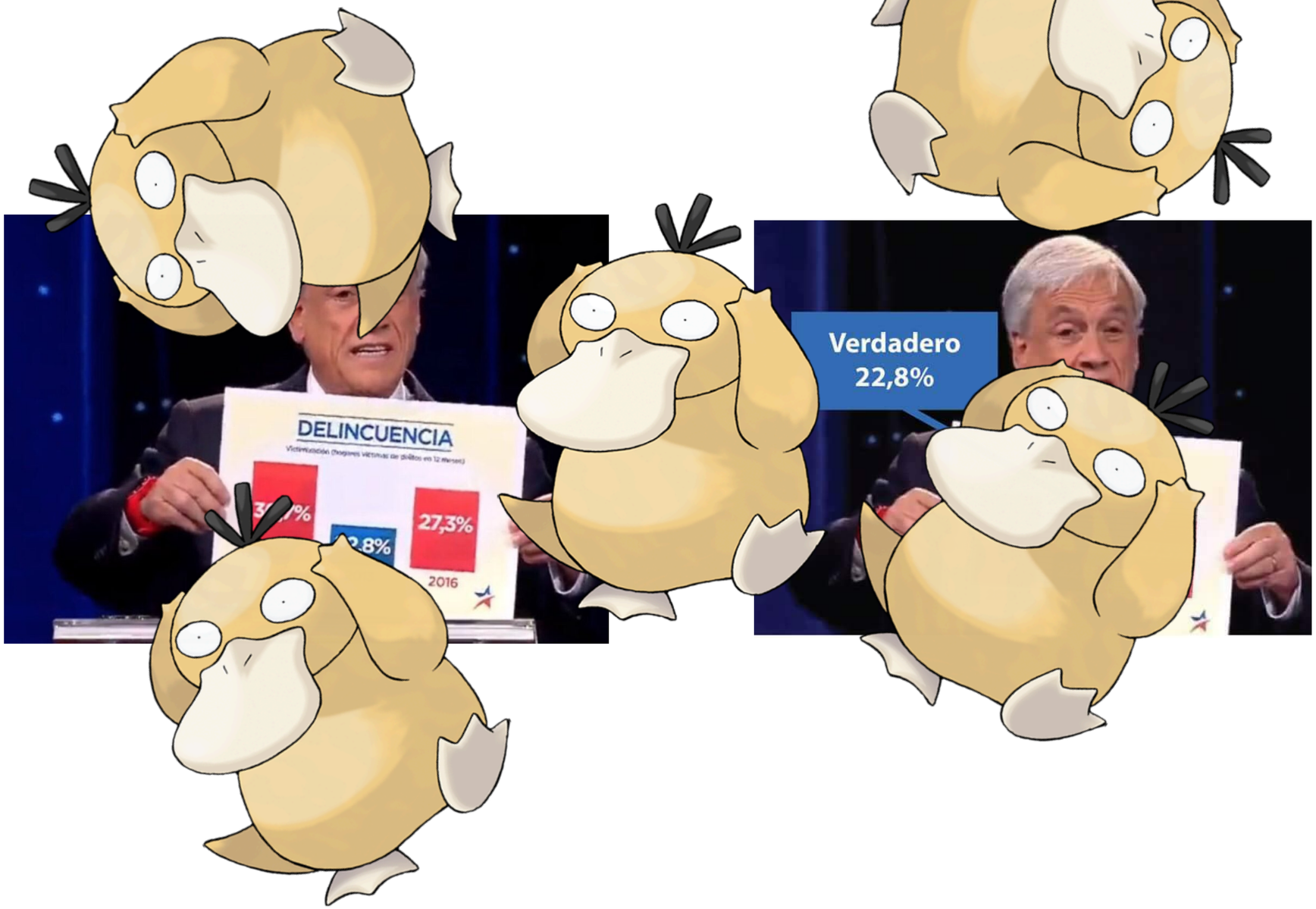
Principios básicos

Principios básicos

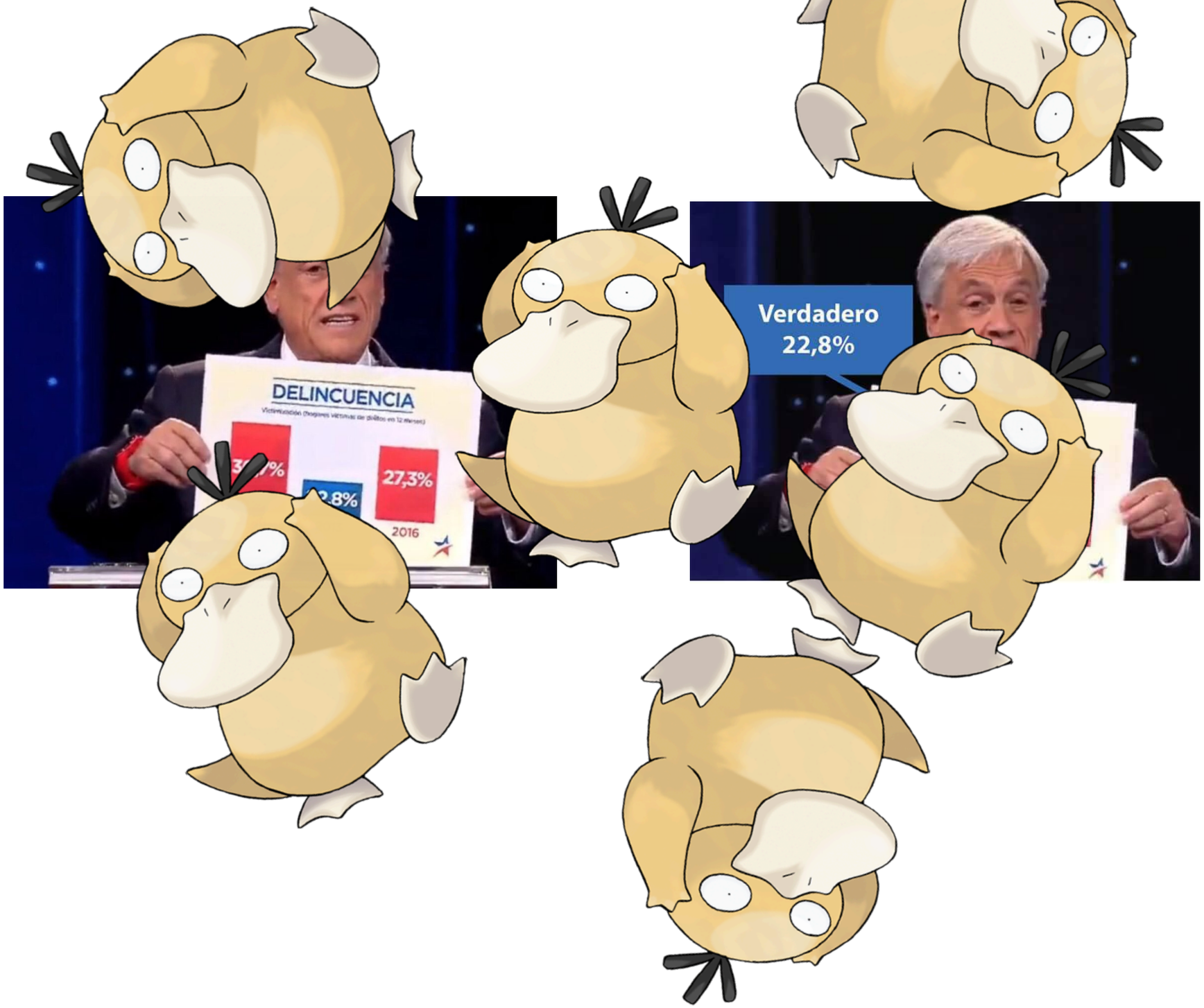
Principios básicos



Principios básicos



Principios básicos



Principios básicos

Para más malas visualizaciones: https://viz.wtf/

Visualización en Python

Existen varias herramientas para hacer visualización de datos en Python:

- Matplotlib
- Seaborn
- Plotnine
- Plotly
- Altair
- ...

En esta sección vamos a ver matplotlib y seaborn

Fundamentos de Ciencias de Datos

Semana 04 - Introducción a la Visualización de Datos