

Empleos del Futuro





INTRODUCCIÓN A LA CIENCIA DE DATOS

Leonardo Ignacio Córdoba

IN GOD WE TRUST, ALL OTHERS
MUST BRING DATA

Deming

Agenda del curso

- Segundo encuentro:



Agenda del curso

- Segundo encuentro:
 - Visualización



Agenda del curso

- Segundo encuentro:
 - Visualización
 - Manipulación de datos



Introducción a la visualización de datos

¿Para qué hacer gráficos?

Parte de la tarea de los analistas consiste en entender qué nos dicen los datos, en este sentido la visualización de datos se emplea para:

- Generar un **conocimiento** en mayor profundidad, complementario, al que se hace empleando la estadística descriptiva.
- **Comunicar** hallazgos y características de la información de manera efectiva.
- **Influenciar** al oyente acerca de una interpretación posible de la información.

Introducción a la visualización de datos

El cuarteto de Anscombe

Consideremos el siguiente problema. Se encuentra un dataset con una variable X y una variable Y en el cual las observaciones se dividen en cuatro grupos. Para cada grupo y para cada variable calculamos los estadísticos descriptivos y obtenemos que:

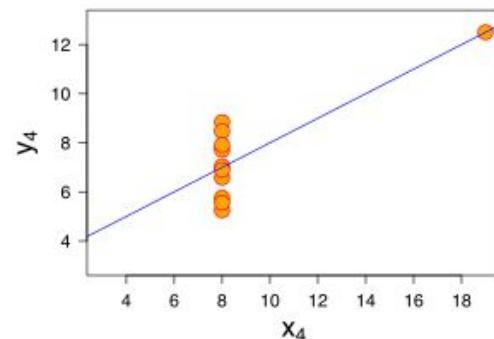
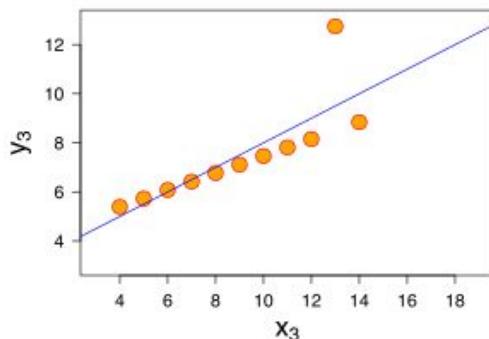
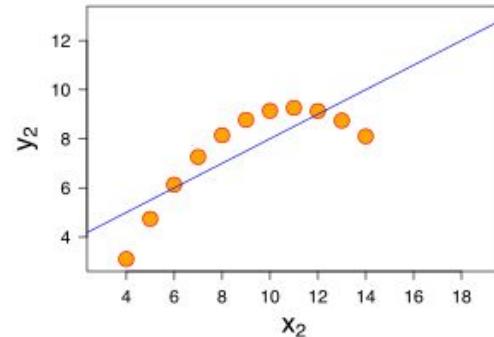
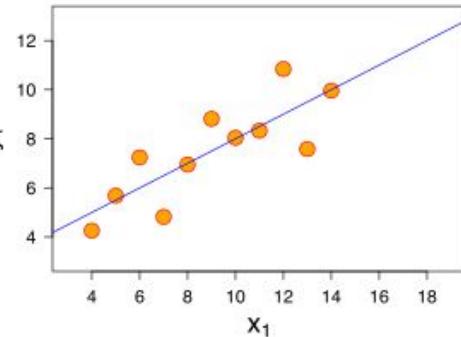
Plot	sum X	sum Y	avg X	avg Y	stdev X	stdev Y
I	99.0	82.5	9.00	7.50	3.32	2.03
II	99.0	82.5	9.00	7.50	3.32	2.03
III	99.0	82.5	9.00	7.50	3.32	2.03
IV	99.0	82.5	9.00	7.50	3.32	2.03

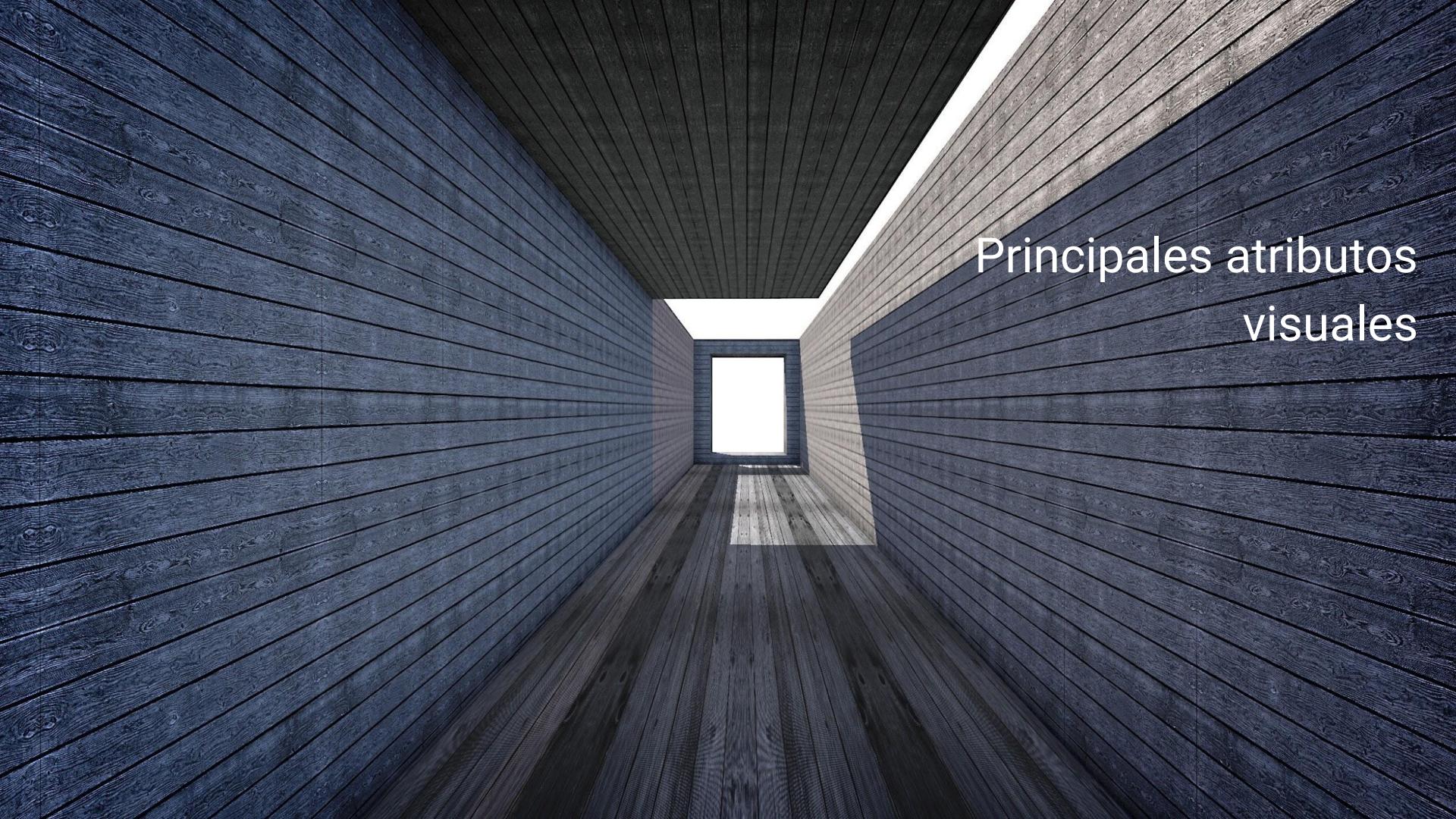
Además, la correlación en todos los casos entre X e Y es 0.816 y el R2 es 0.68, ¿podemos concluir que los 4 conjuntos de datos son iguales?

Introducción a la visualización de datos

El cuarteto de Anscombe

- Podemos ver que, a pesar de que la estadística descriptiva nos dio resultados iguales en todos los casos, los datos son realmente distintos.



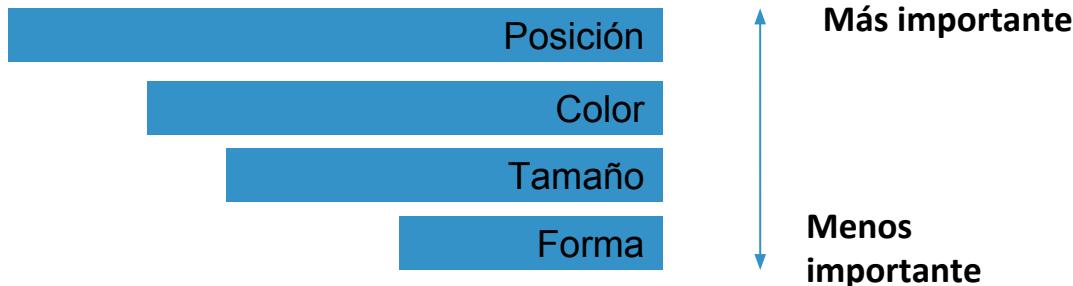
A perspective view of a dark wooden corridor. The walls and ceiling are made of vertical wooden planks. The floor is also made of wood. At the end of the corridor, there is a white rectangular opening. The lighting is dramatic, with the dark wood contrasting against the bright white opening.

Principales atributos visuales

Introducción a la visualización de datos

Percepción visual

- Algunos elementos tienen un impacto más grande en nuestra percepción:



Introducción a la visualización de datos

Percepción visual

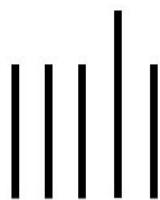
¿Cuántos cuadrados hay? ¿Cuántos círculos?

¿Cuál es la mejor forma de transmitir la información?

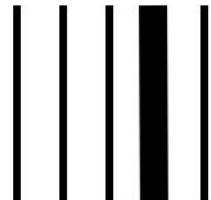


Introducción a la visualización de datos

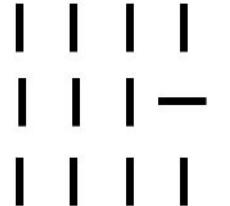
Recursos



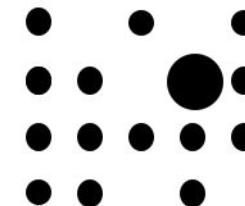
Length



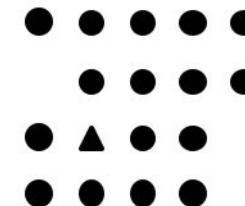
Width



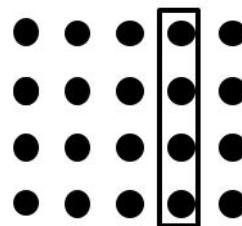
Orientation



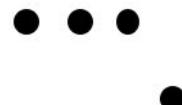
Size



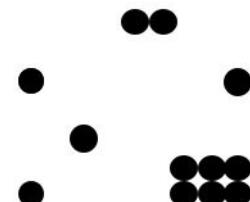
Shape



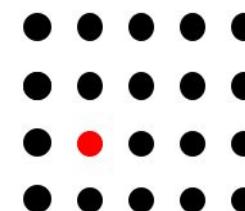
Enclosure



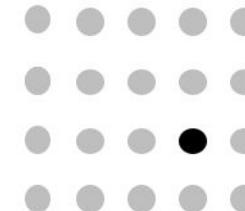
2D Position



Grouping

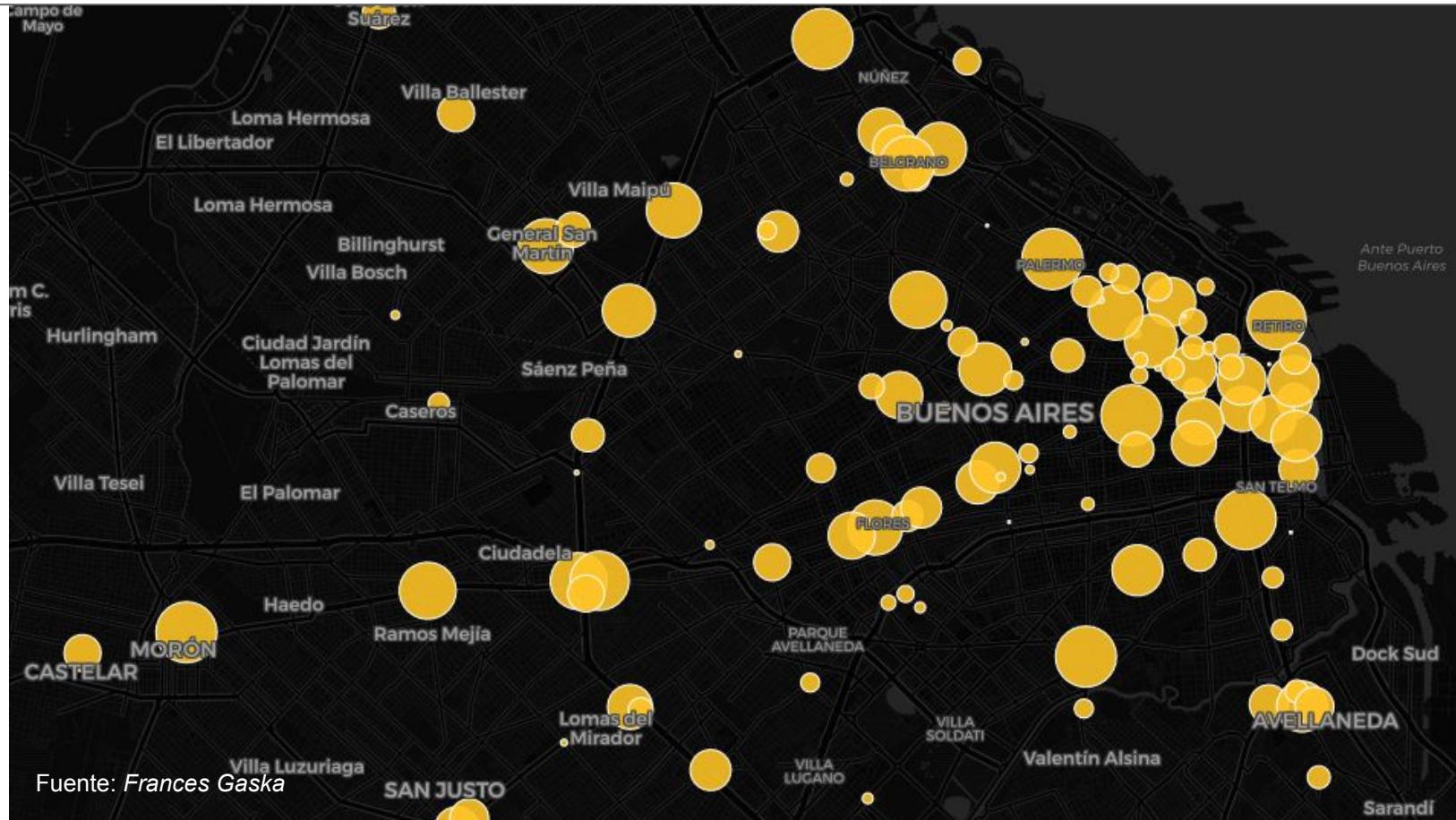


Color (Hue)



Color (Intensity)

Introducción a la visualización de datos



Introducción a la visualización de datos



Introducción a la visualización de datos

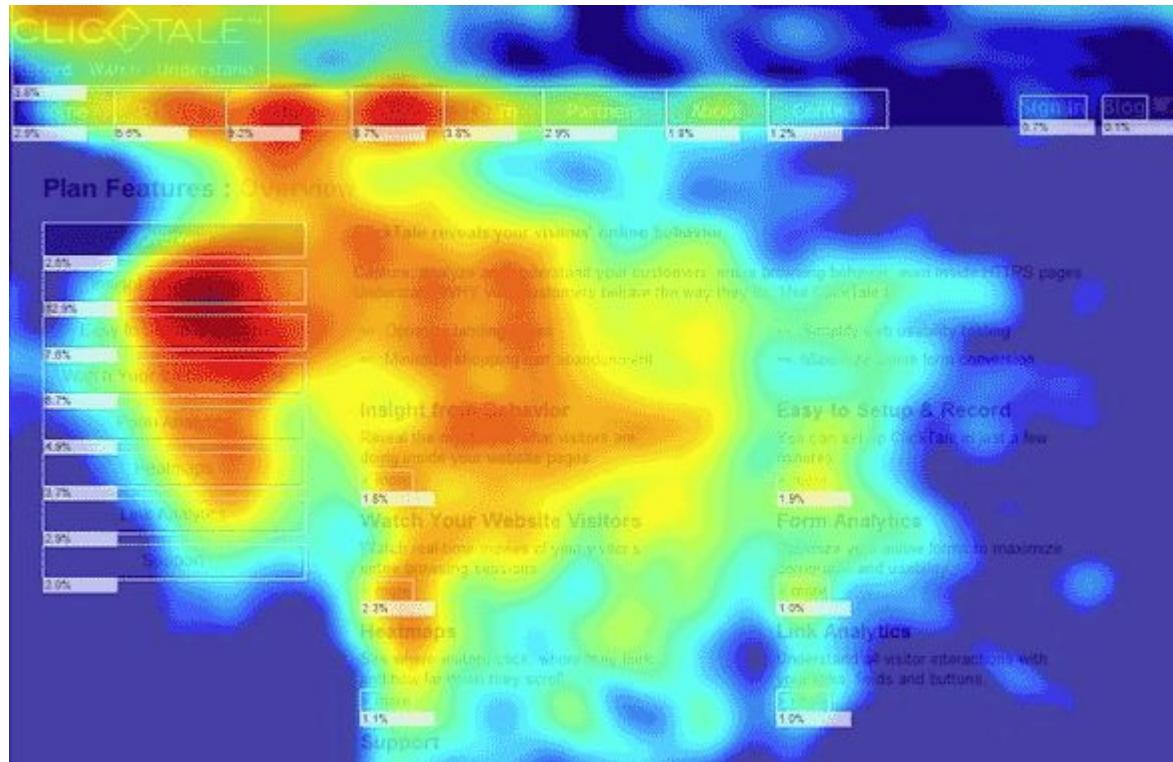
Color

Los colores se usan en visualización para facilitar la interpretación de:

- Secuencias
- Divergencias
- Categorías

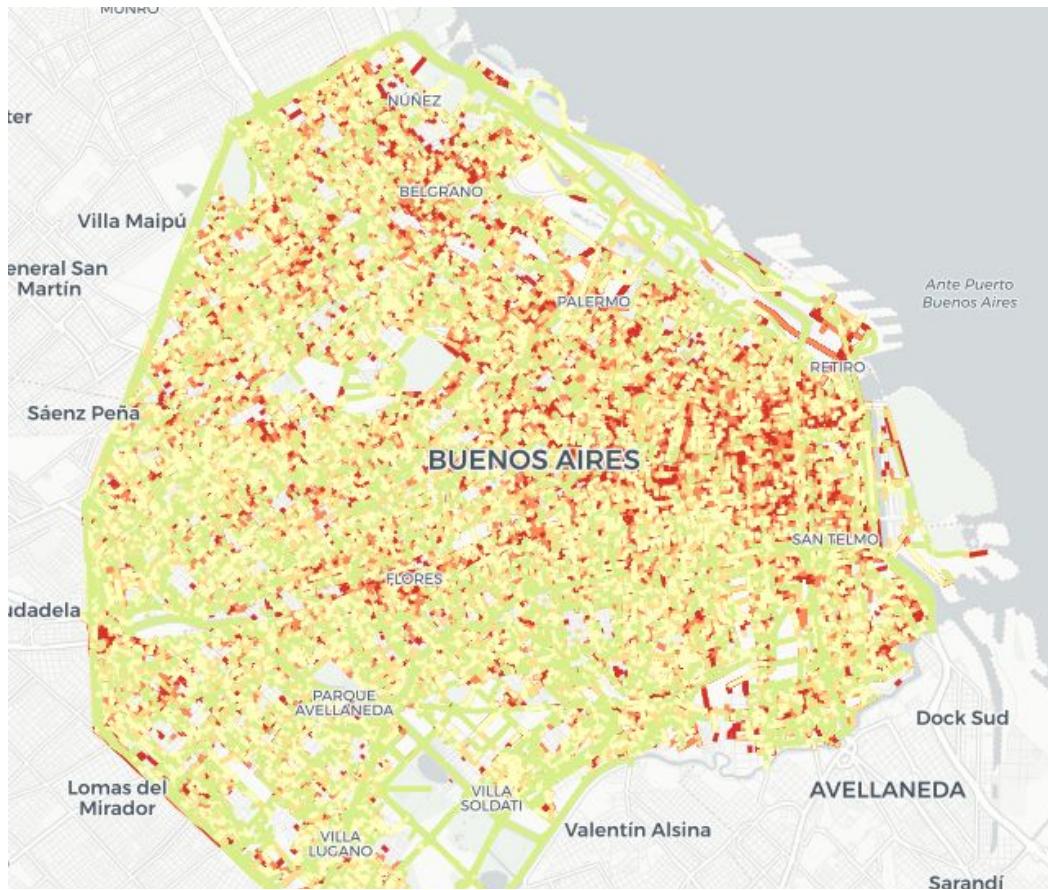
Introducción a la visualización de datos

Secuencias



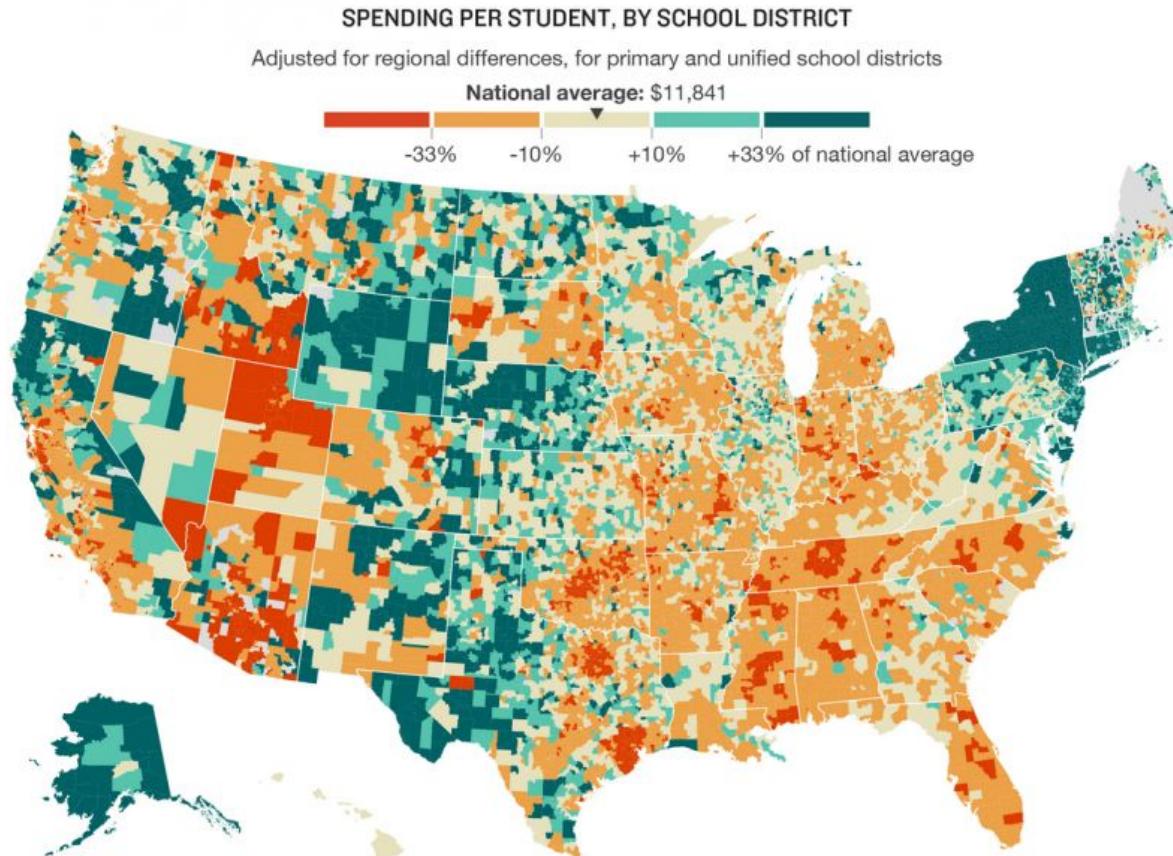
Introducción a la visualización de datos

Secuencias



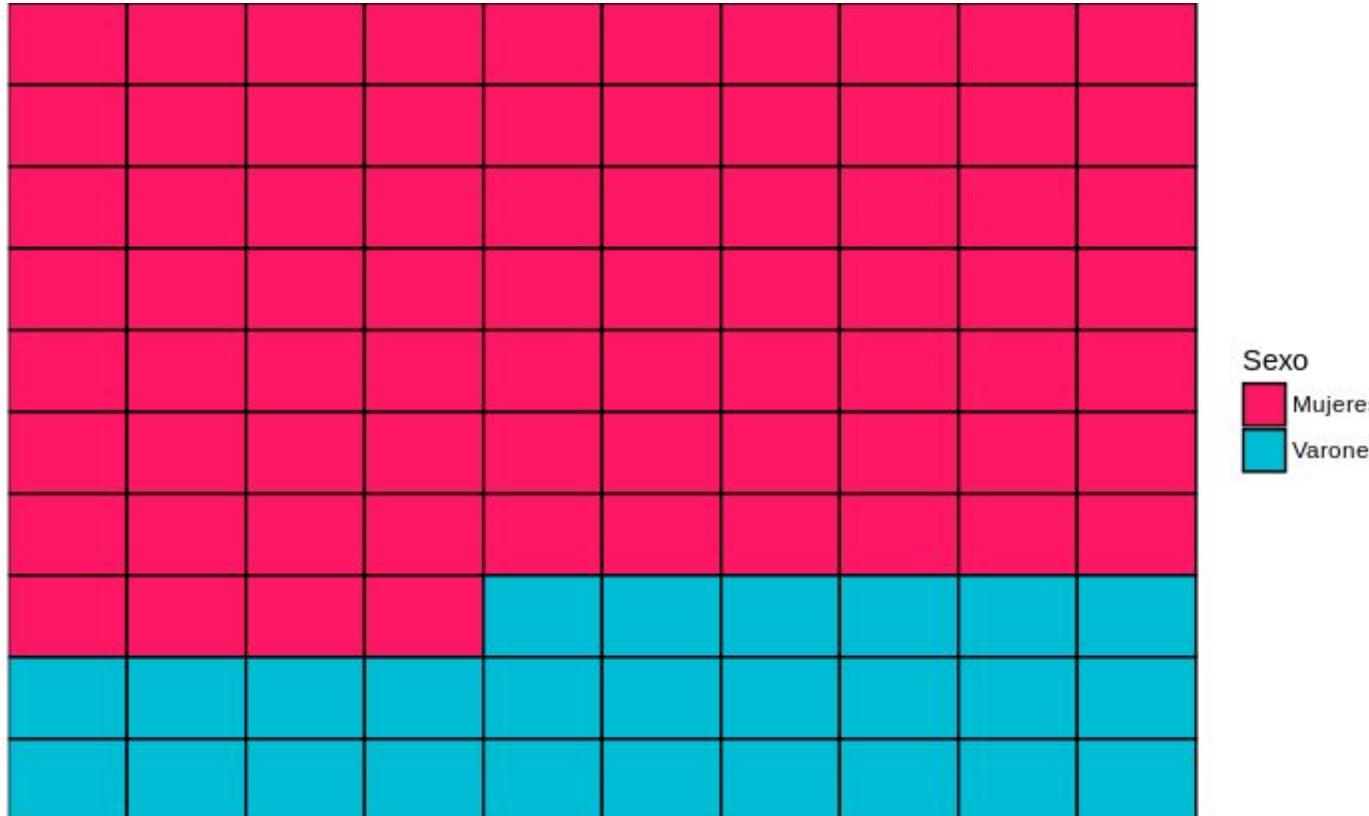
Introducción a la visualización de datos

Divergencia



Introducción a la visualización de datos

Categorías



Tipos de visualizaciones

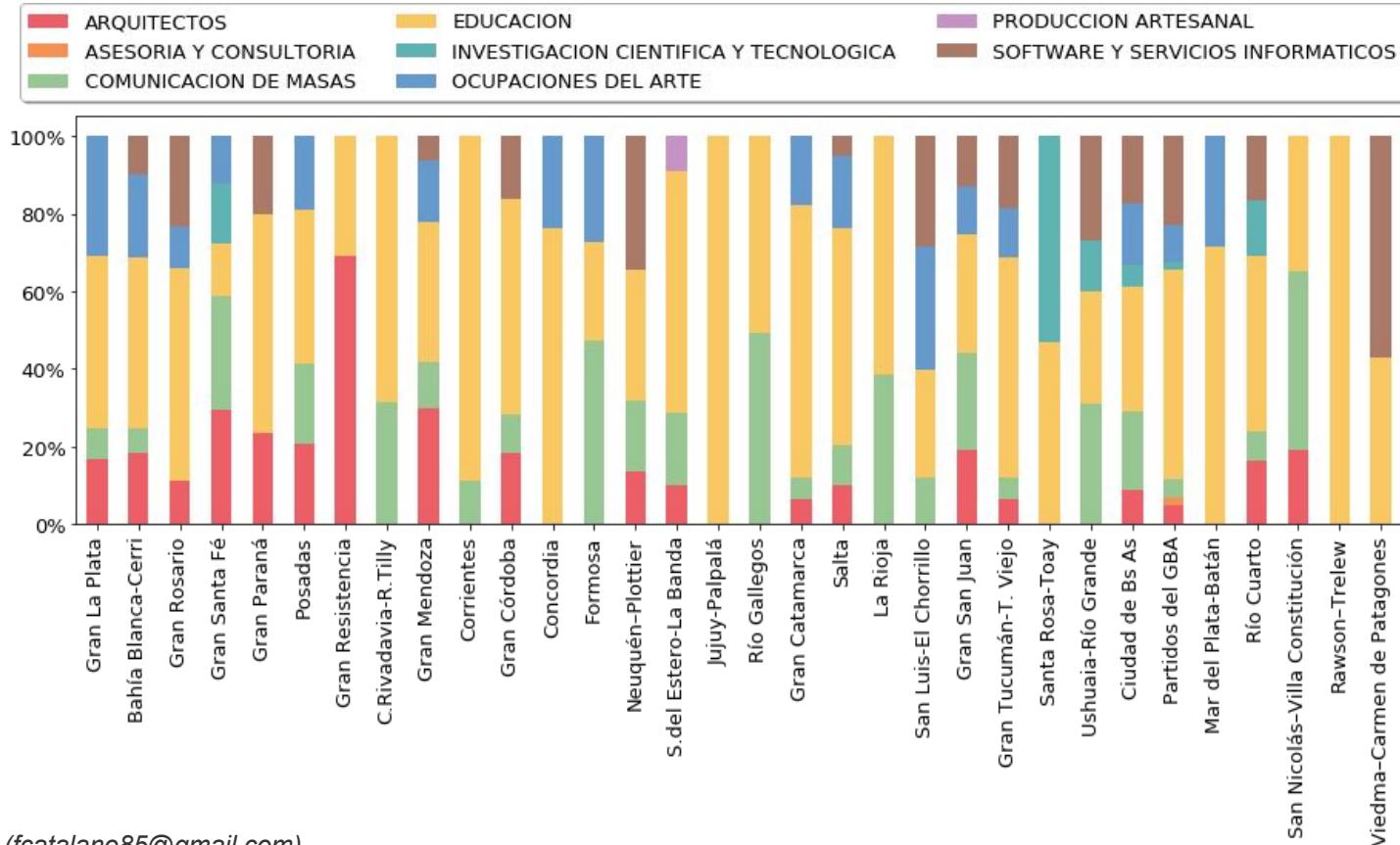
Tipos de visualizaciones

Gráfico de barras (verticales)



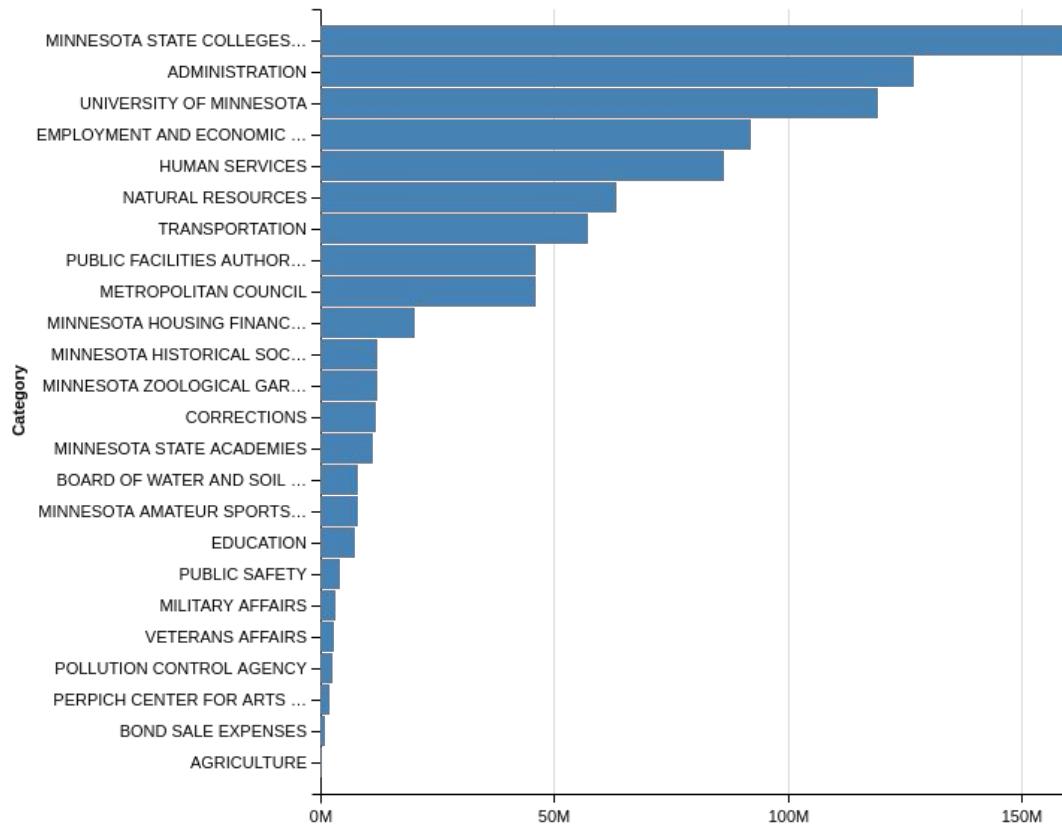
Tipos de visualizaciones

Gráfico
de barras
(apiladas)



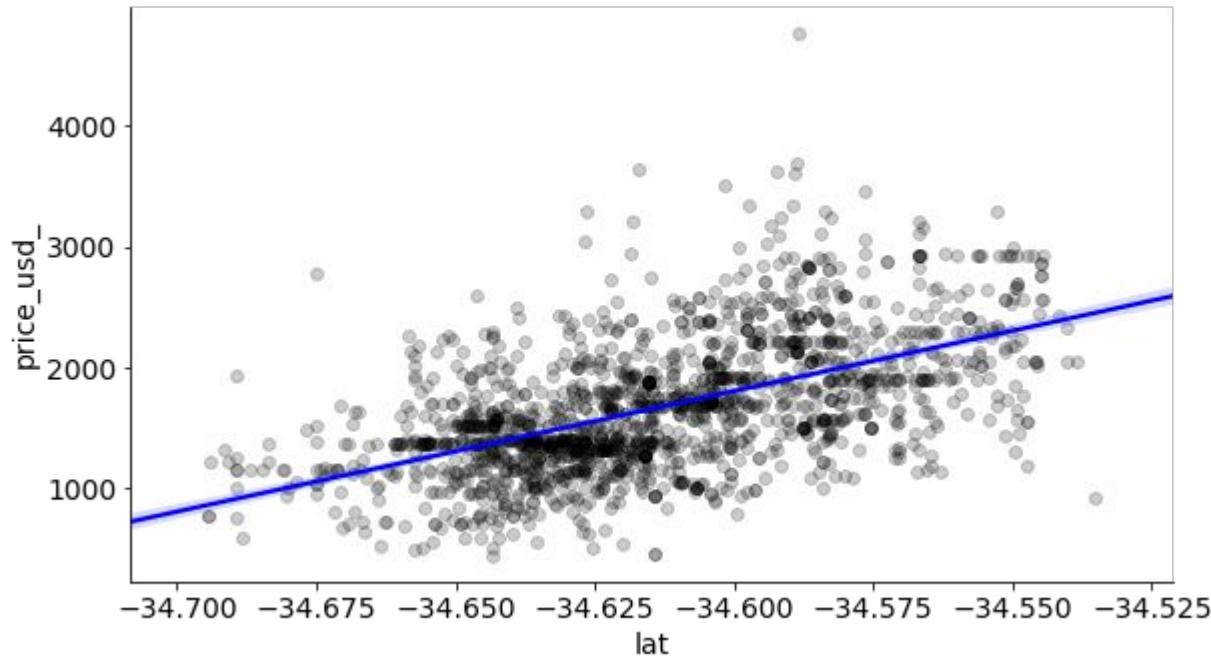
Tipos de visualizaciones

Gráfico de barras (horizontales)



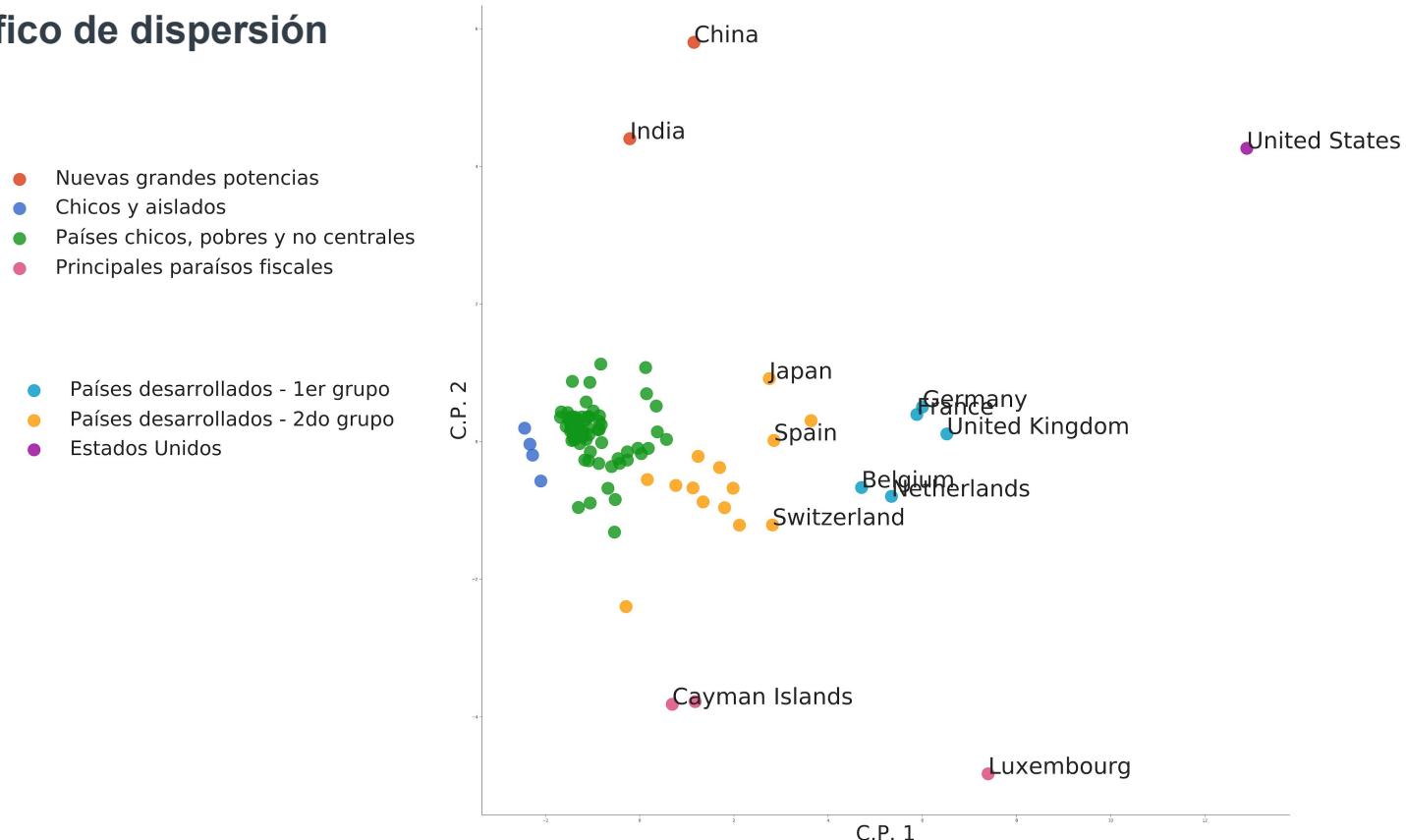
Tipos de visualizaciones

Gráfico de dispersión



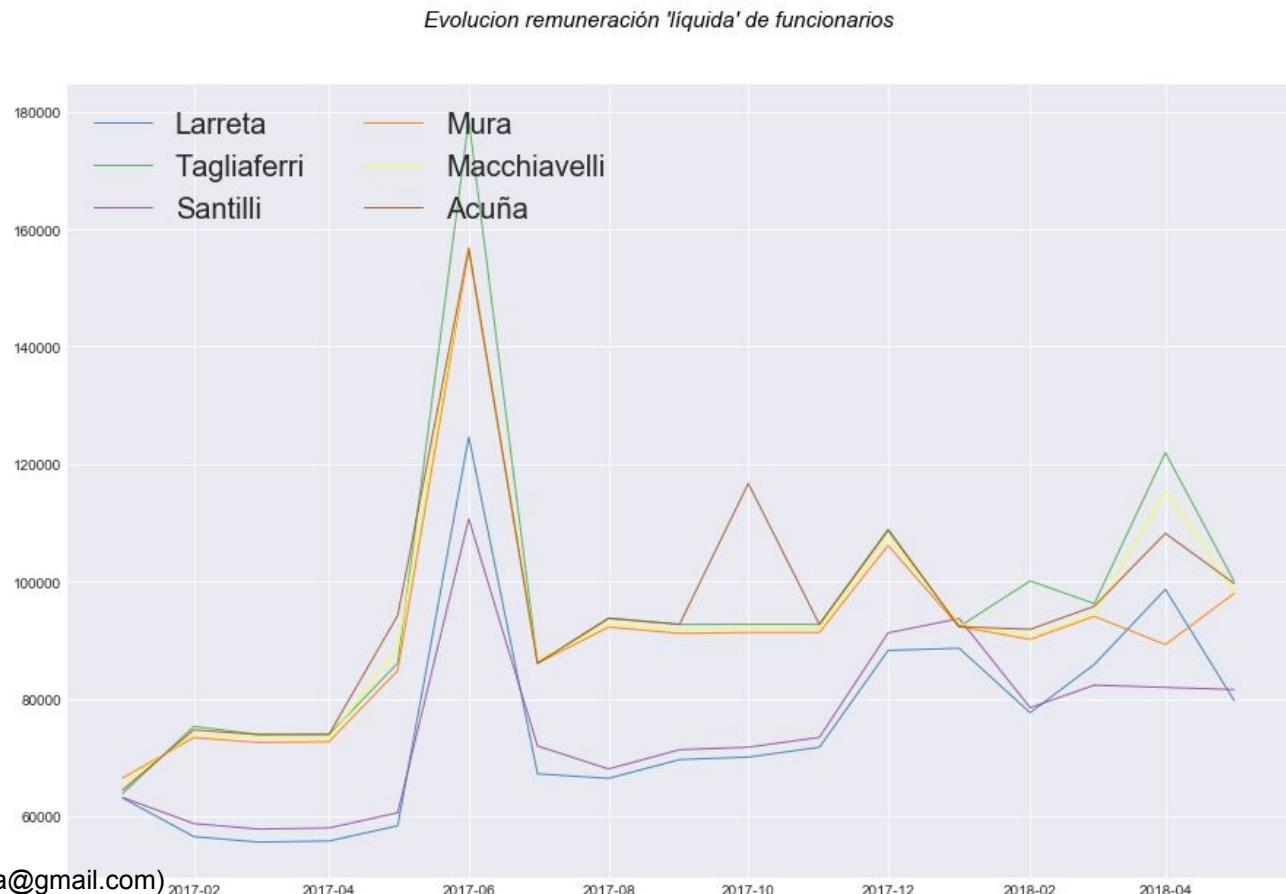
Tipos de visualizaciones

Gráfico de dispersión



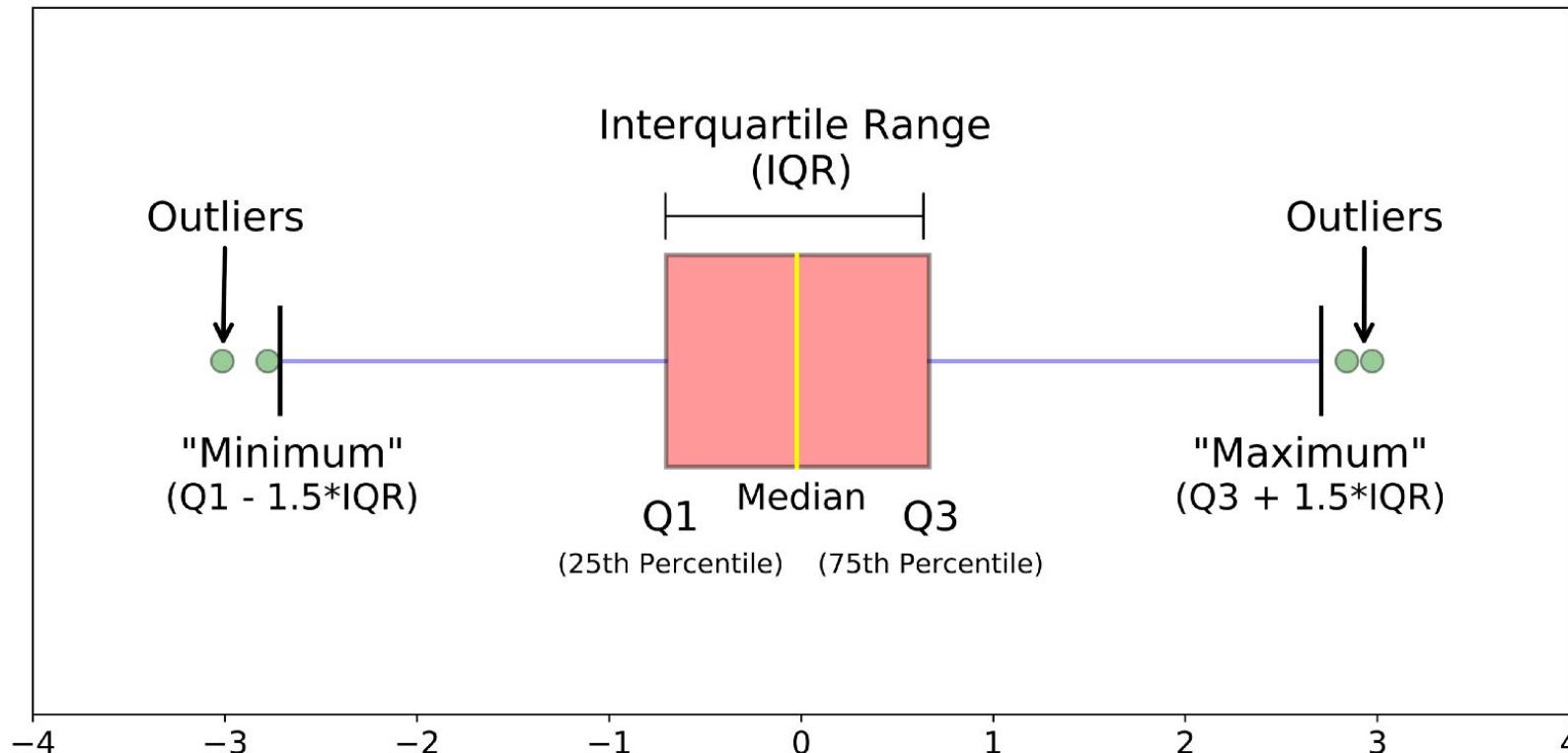
Tipos de visualizaciones

Gráfico de líneas



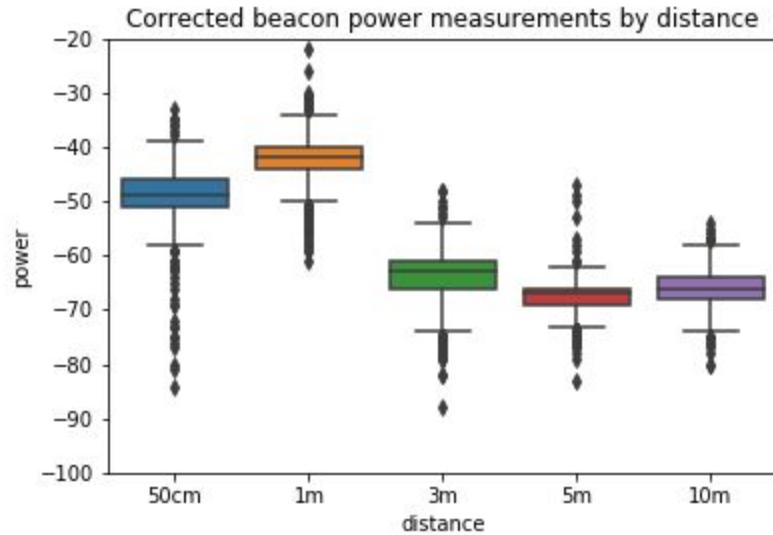
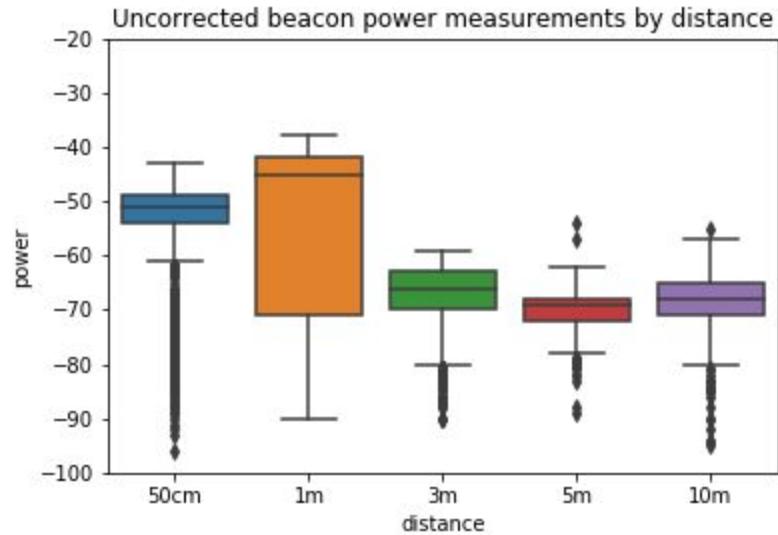
Tipos de visualizaciones

Boxplot



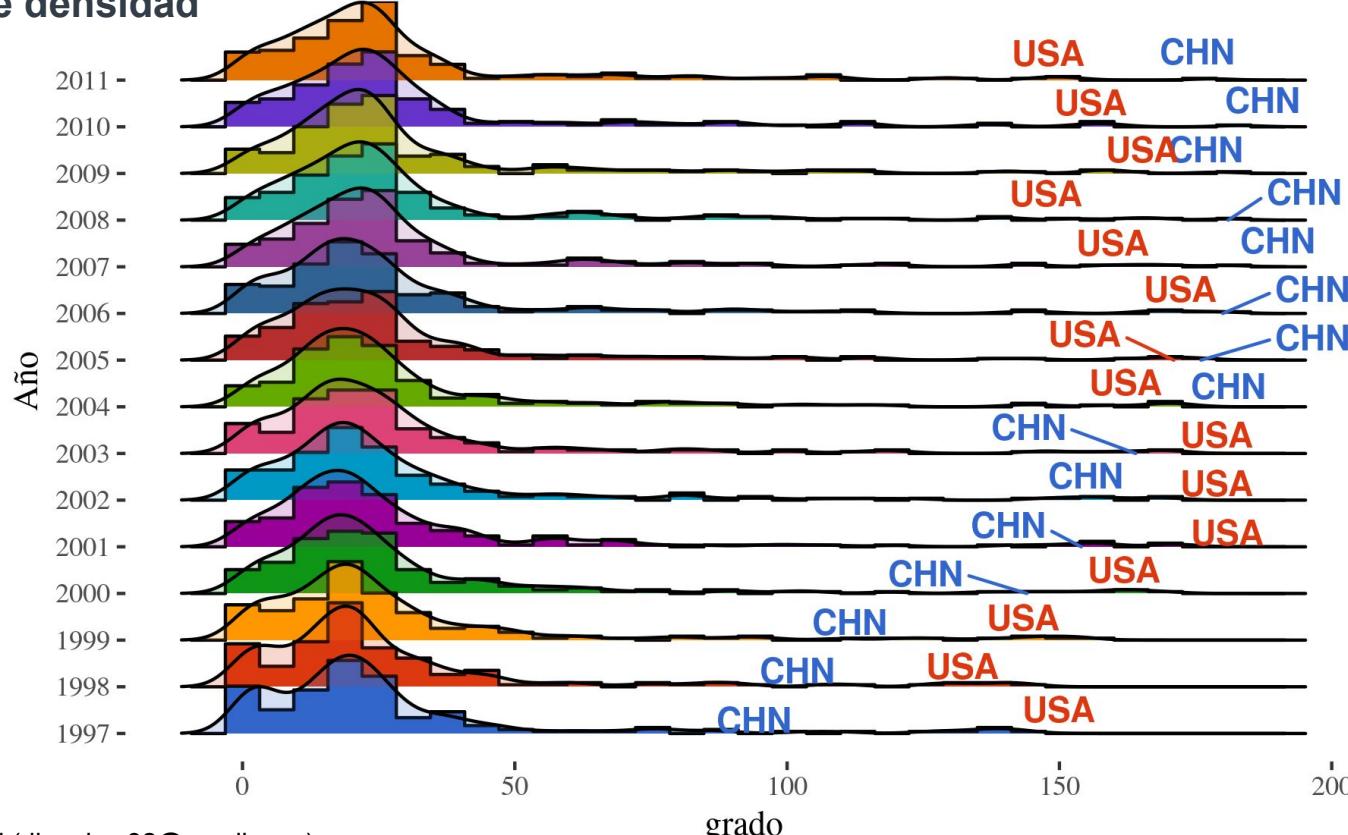
Tipos de visualizaciones

Boxplots



Tipos de visualizaciones

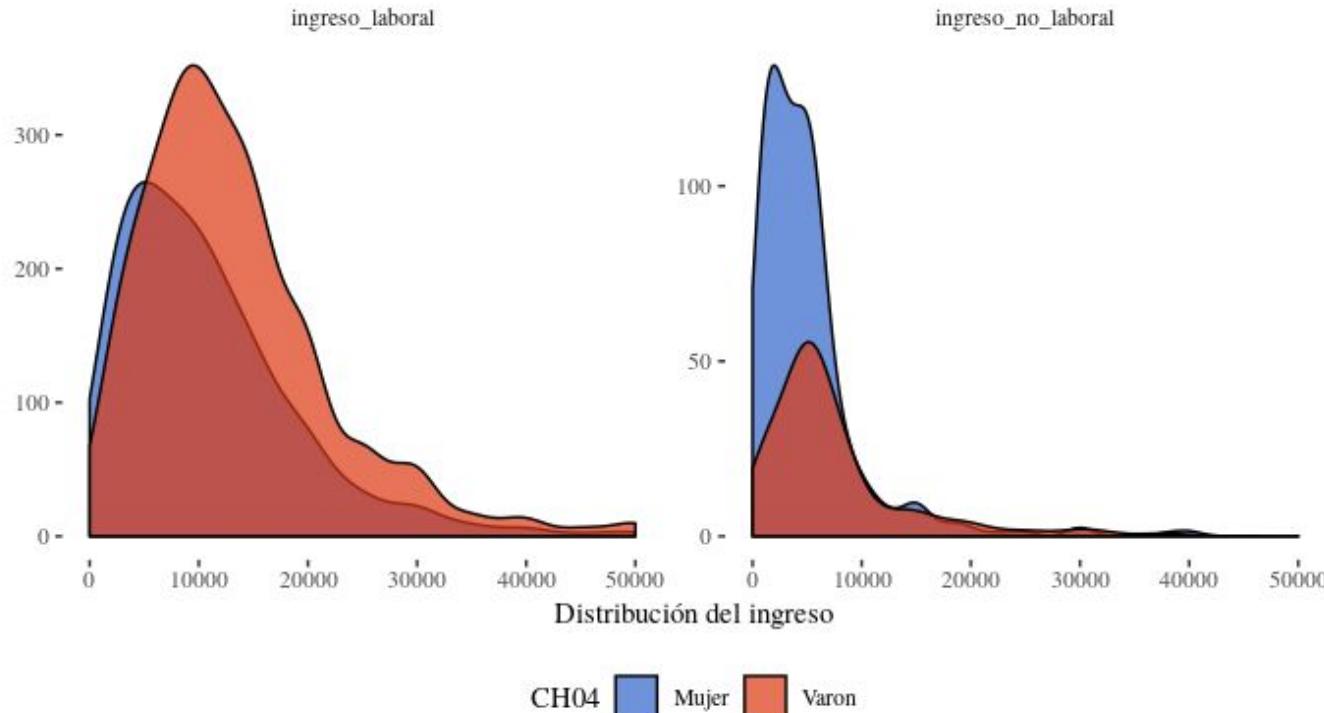
Gráfico de densidad



Tipos de visualizaciones

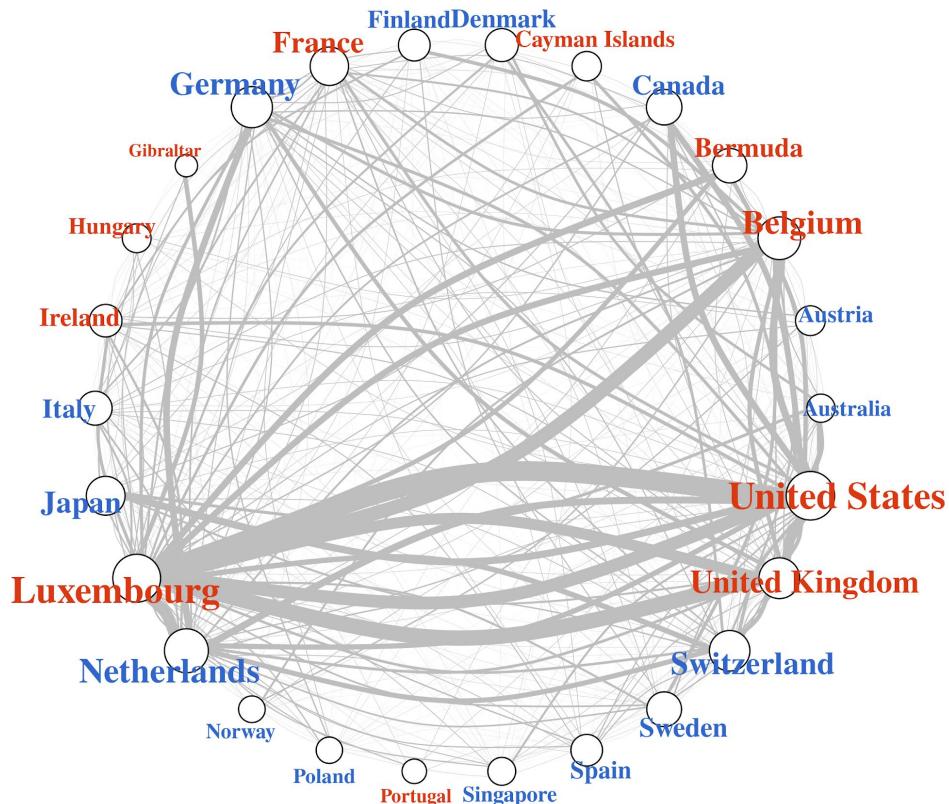
Gráfico de densidad

Total según tipo de ingreso y género



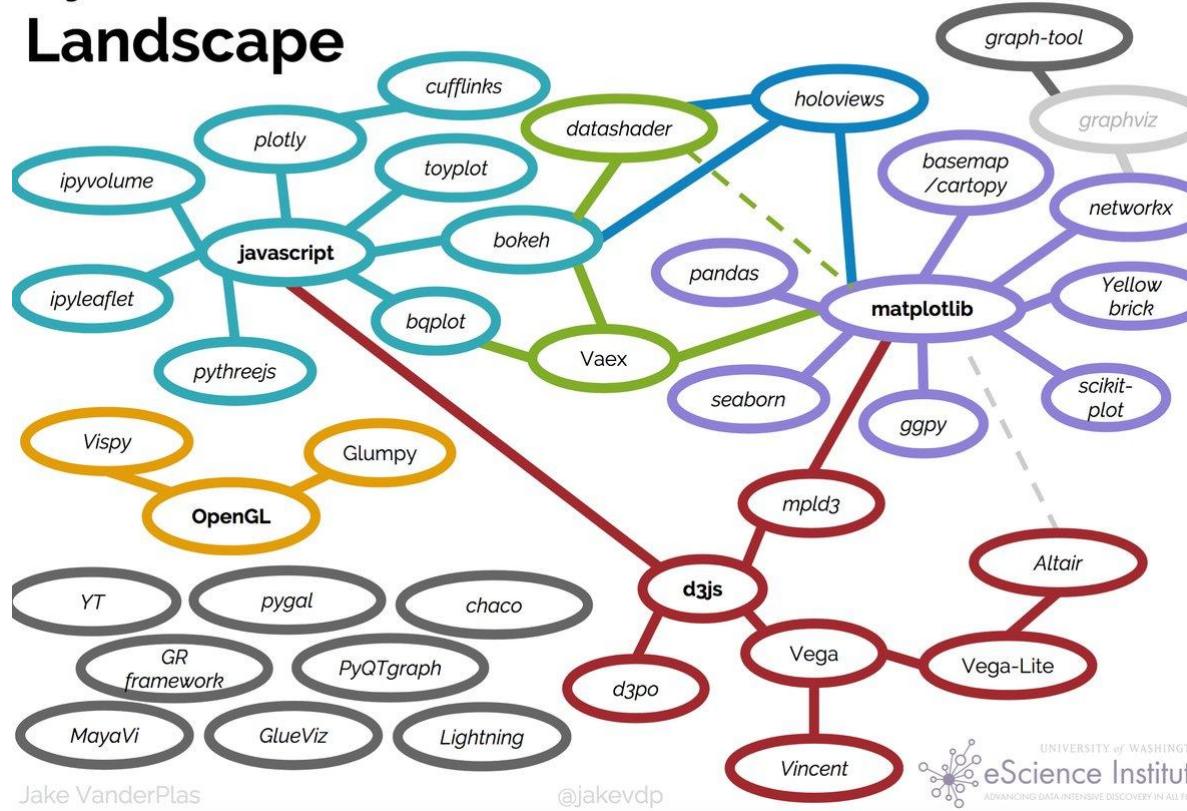
Tipos de visualizaciones

Gráfico de flujos



Tipos de visualizaciones

Python's Visualization Landscape



Tipos de visualizaciones

Links

Mapas:

<https://carto.com/gallery/>

Python:

<https://bokeh.pydata.org/en/latest/docs/gallery.html>

<https://plot.ly/python/>

<https://seaborn.pydata.org/examples/index.html>

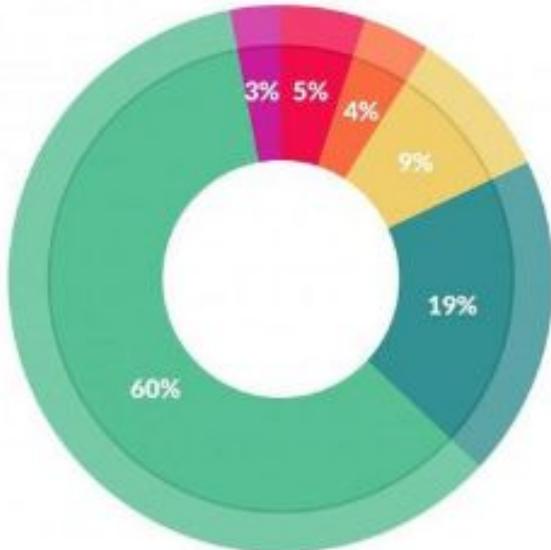
<https://altair-viz.github.io/gallery/index.html>



INTRODUCCIÓN A PANDAS

Introducción a PANDAS

Manipulación de datos

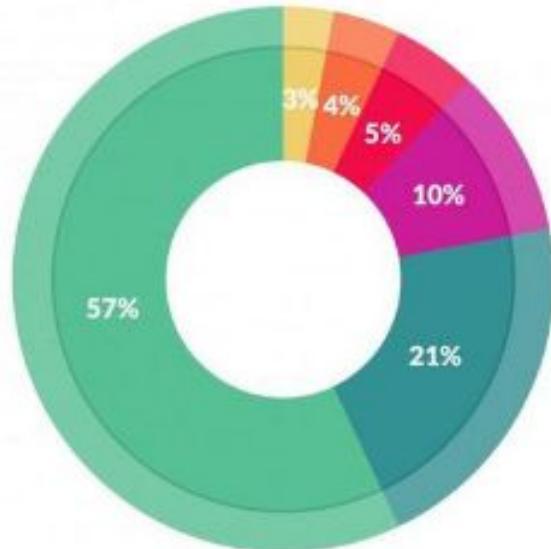


What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Introducción a PANDAS

Manipulación de datos



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Introducción a PANDAS

¿Por qué hacer análisis en Python?

- Reproducibilidad
- Facilidad en la lectura
- Comunidad enorme con gran variedad de librerías generales y específicas, recursos, foros, etc.
- Preferido (junto a R) en ambientes de ciencia de datos
- Gran capacidad para interoperar con otros sistemas (bases de datos, servicios cloud, incluso softwares propietarios)
- Escalabilidad
- Posibilidad de generar tests automáticos



Introducción a PANDAS

¿PANDAS?



Introducción a PANDAS

PANDAS: PANel DAta System

PANDAS es la librería por excelencia en Python para la manipulación de datos.

Esta librería está creada tomando por base los objeto Array de la librería Numpy, la cual es una librería que facilita las operaciones numéricas con un gran rendimiento de performance.

Está basada en el concepto de Data Frame del lenguaje R, que esencialmente imita una tabla u hoja de cálculo de Excel, pero extendiendo enormemente sus funcionalidades.

Permite realizar operaciones vectorizadas, joins y selección de columnas y filas como con SQL, creación de nuevas columnas, aplicación de operaciones sobre las mismas, visualizaciones rápidas, etc.

Introducción a PANDAS

PANDAS: PANel DAta System

PANDAS es, además, enormemente útil para:

- Emplear las diversas librerías de visualización
- Realizar tareas de análisis estadístico con StatsModel o de Machine Learning con Scikit-learn
- Hacer procesamientos geográficos con la extensión GeoPandas.
- Estructurar información proveniente de la web en formatos semiestructurados.

Introducción a PANDAS

PANDAS: PANel DAta System

PANDAS se basa en tres objetos fundamentales:

- DataFrame
- Serie
- Index

Estas estructuras tienen por objeto principal poder representar apropiadamente la información tabular, es decir, aquella en la que las filas representan **observaciones** y las columnas representan **características** de esas observaciones.

Introducción a PANDAS

Series

Las Series de Pandas constan de un Index (que sirve para identificar a cada fila), de los values (que son los valores) y el nombre de la misma.

El index es de gran utilidad para indicar fechas en series de tiempo.

Index	País	Values
1	Argentina	
2	Brasil	
3	Chile	
4	Colombia	

Introducción a PANDAS

DataFrame

Los DataFrame de PANDAS permiten representar información en formato de tabla, donde cada observación tiene un Index y cada atributo pertenece a una columna.

Los datos propiamente dicho son atributos llamados “values”, almacenados en Array de Numpy.

Index	País	Capital	Index	Values
1	Argentina	Buenos Aires	1	Argentina
2	Brasil	Brasilia	2	Brasil
3	Chile	Santiago de Chile	3	Chile
4	Colombia	Bogotá	4	Colombia

GRACIAS

Contacto

Docente: Leonardo Ignacio Córdoba

E-mail: cordoba.leonardoignacio@gmail.com

LinkedIn Leonardo Ignacio Córdoba



Buenos Aires Ciudad

