Data Visualization

Dive Deep into Matplotlib and Seaborn

¡En español!

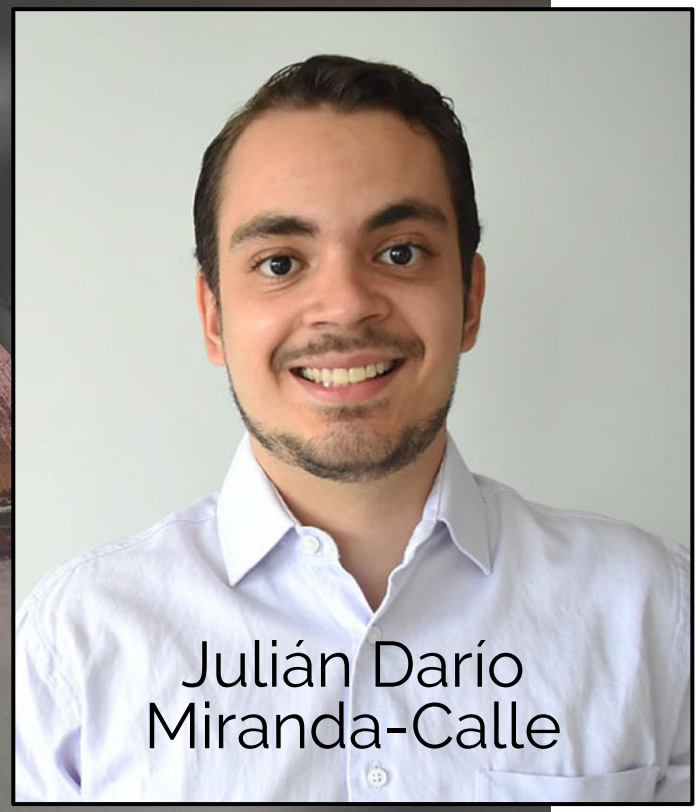
Julián Darío Miranda-Calle

Universidad Pontificia Bolivariana
SECCIONAL BUCARAMANGA

DPhi Tech
Data AnalysisBootcamp

DPhi Tech
Data Analysis Bootcamp

Presentación inicial



Auditor Interno ISO 27001:2013 (SGS Colombia), Especialista en Seguridad Informática, Ingeniero de Sistemas e Informática, Ingeniero Electrónico de la Universidad Pontificia Bolivariana de Bucaramanga (UPB), Colombia.

Docente de pregrado y posgrado de la Facultad de Ingeniería de Sistemas e Informática de la UPB. Scrum Master, desarrollador e investigador en criptografía, esteganografía, esteganálisis e ingeniería sísmica usando técnicas de Ciencia de Datos, Machine Learning y Deep Learning.

linkedin.com/in/juliandariomiranda/ 0000-0002-7580-2361
https://www.researchgate.net/profile/Julian_Miranda2

Agenda

01 Introducción
- Pipeline de Ciencia de Datos
- Metodología CRISP-DM
- Visualización de datos

02 Visualización de datos
- Definiciones
- Tipos de datos
- Relaciones entre datos

03 Tipos de visualización
Formatos y tipos de visualización según los datos y el objetivo

04 Caso de uso
Explorando eventos de ciberseguridad

05 Conclusiones
- Otros ejemplos de visualización
- Conclusiones

DPhi Tech
Data Analysis Bootcamp

Julián Darío Miranda-Calle
Visualization 101

Julián Darío Miranda-Calle
Visualization 101

Introducción

Pipeline de la Ciencia de Datos

Entorno del negocio

01 Adquisición

Conectividad con dispositivos y recepción de datos mediante protocolos definidos

02 Preparación de datos

Estructuración, limpieza y enriquecimiento de los datos.

03 Exploración de datos

Identificar patrones visibles en los datos y comprobar las hipótesis.

04 Modelado

Entrenar modelos de Machine Learning y Deep Learning para predecir variables de interés.

05 Validación

Validar que los modelos predictivos y prescriptivos ejecuten las tareas adecuadamente.

06 Visualización de datos

Comunicar los hallazgos con los stakeholders mediante visualización interactiva de datos.

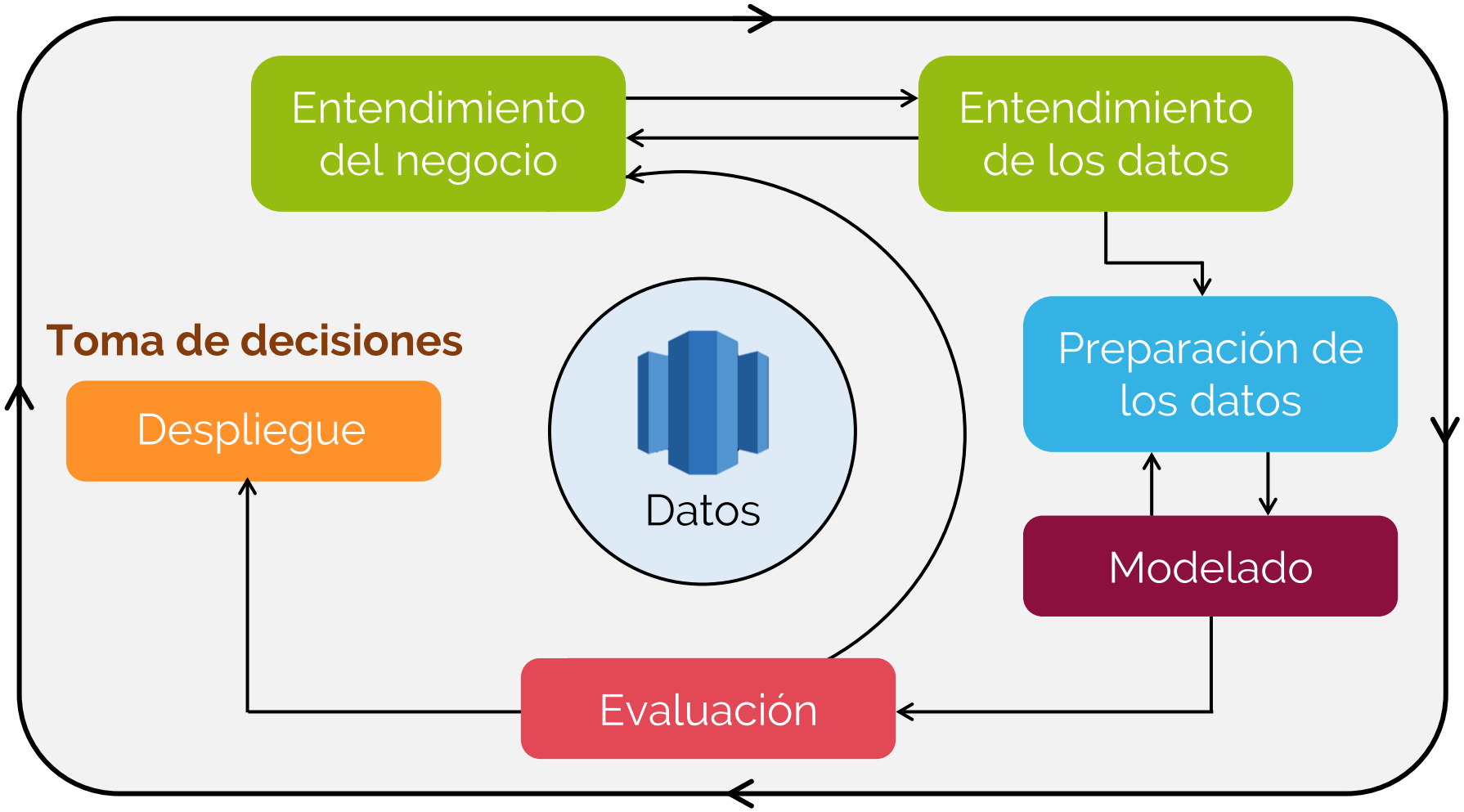
Toma de decisiones

Julián Darío Miranda-Calle
Visualization 101

Introducción

Metodología CRISP-DM

CRISP-DM es una metodología estándar para la minería de datos, que describe enfoques comunes utilizados para proceder con análisis de datos y toma de decisiones de negocio.



Julián Darío Miranda-Calle

Introducción

Visualización de datos

Enfocada al análisis ⟵ ⟶ **Enfocada a los resultados**

Entorno del negocio

Visualización de datos para el análisis y el descubrimiento de patrones.

Visualización de datos para la comunicación de hallazgos.

03 Exploración de datos

Identificar patrones visibles en los datos y comprobar las hipótesis.

06 Visualización de datos

Comunicar los hallazgos con los stakeholders mediante visualización interactiva de datos.

Toma de decisiones

Visualización Datos

Julián Darío Miranda-Calle
Visualization 101

Definición y pipeline

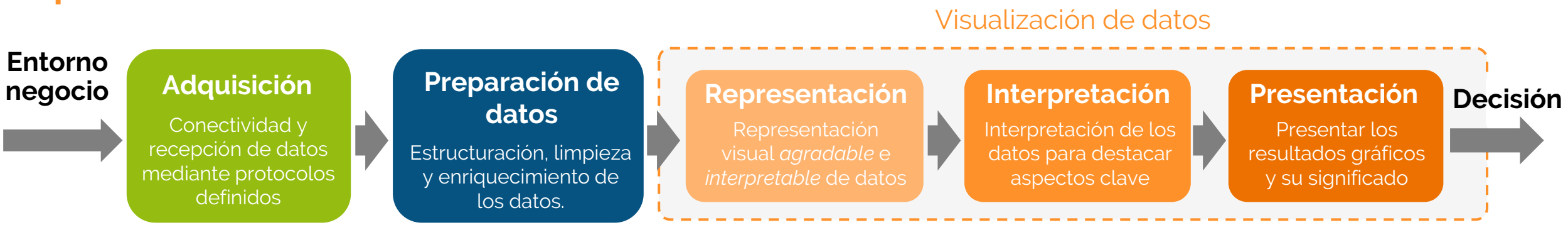
Visualización de datos

Comunicar los hallazgos con los stakeholders mediante visualización interactiva de datos.

Es una herramienta muy poderosa de generación de impacto en el contexto del negocio Netquest, 2020

La visualización de datos es esencial parta la **comunicación estratégica**, ayuda a la interpretación de datos, **detección de patrones**, tendencias y anomalías. Permite la **toma de decisiones** y el análisis inherente de procesos representados por los datos.

Pipeline de visualización:



Visualización Datos

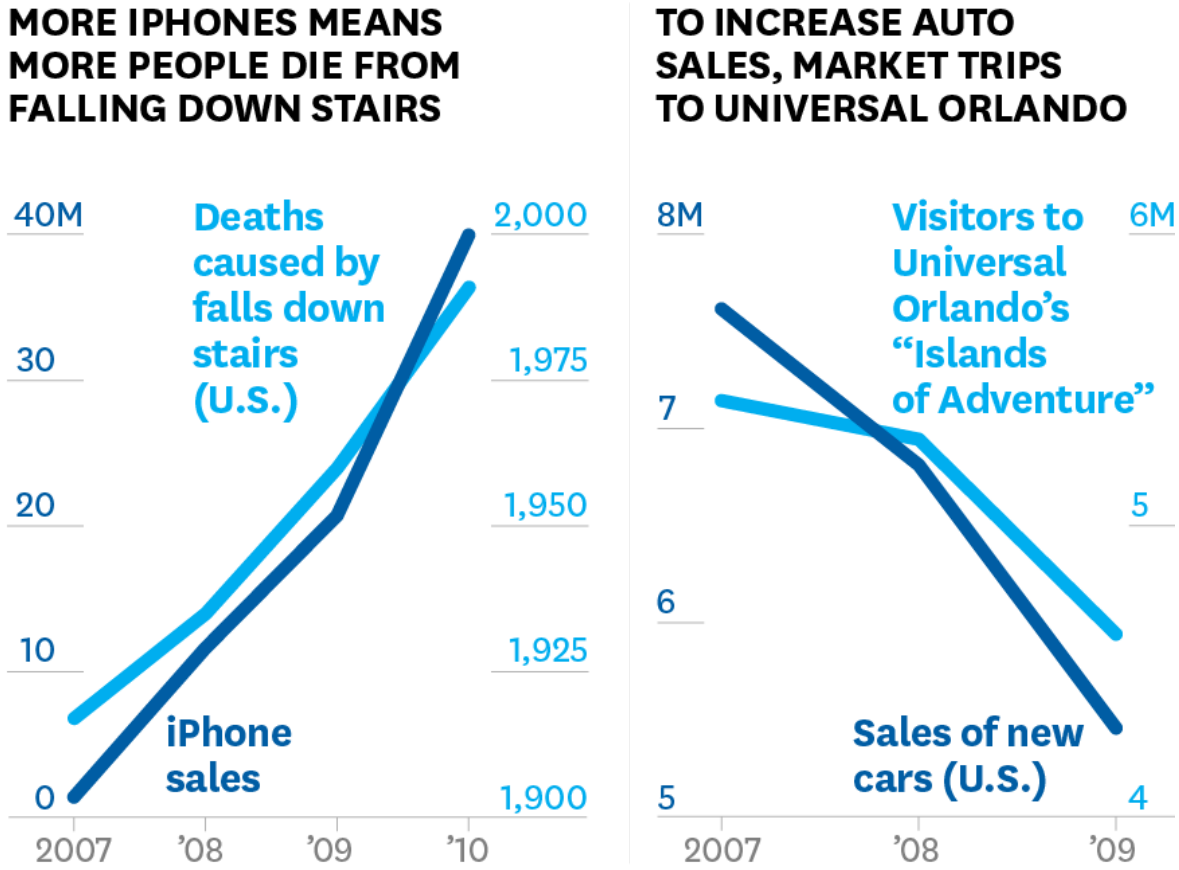
Julián Darío Miranda-Calle
Visualization 101

Importancia de la visualización de datos

Representa todo el proceso de análisis

Tomado de: Kimberly Herrington, LinkedIn, 2021.

Interpretación – Falsa correlación



Tomado de: tylervigen.com, 2015

Visualización Datos

Julián Darío Miranda-Calle

Tipos de datos

Cuantitativos (numéricos)

Datos que pueden ser cuantificados y medidos, representando una sucesión de mediciones que puede ser finita o infinita.

Discretos

Datos que consisten de agrupaciones finitas y cuantizables.

Número de estudiantes de una clase

Continuos

Datos que consisten, generalmente, intervalos con una cantidad infinita numerable de valores.

TRM (diaria) del dólar

Cualitativos (categóricos)

Datos que pueden ser divididos en categorías numerables de forma finita, con o sin orden aparente, que mide cualidades.

Ordinales

Datos que siguen un orden o secuencia que tiene un significado claro.

Posiciones del podio en carrera

Nominales

Datos que no siguen un orden o secuencia definidos.

Razas de perros (dog breeds)

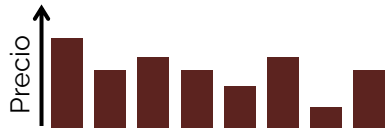
Visualización Datos

Julián Darío Miranda-Calle
Visualization 101

Relaciones entre los datos

Comparación

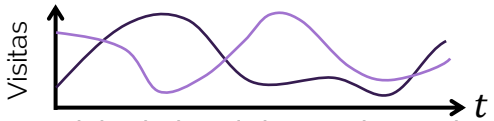
Comparar valores cuantitativos de diferentes categorías



Precio promedio de autos

Serie de tiempo

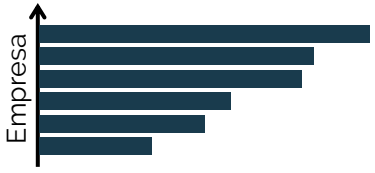
Seguimiento del comportamiento de variables en el tiempo



Cantidad de visitas a la web en un periodo definido

Ranking

Relaciona dos o más valores en la misma escala



Ventas en Noviembre de seis empresas

Distribución

Distribución espacial de datos sobre un punto central

El peso de los estudiantes de un salón

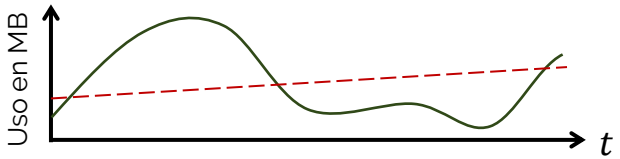
Porción/parcialidad

Muestra un subconjunto comparado con el total de mediciones

Porcentaje de clientes que compran productos puntuales

Desviación

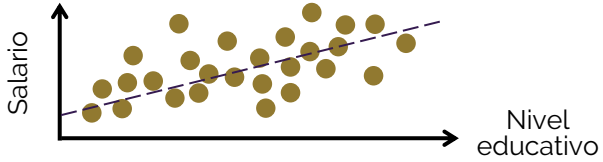
Examinar la relación entre mediciones y su divergencia



Baseline del uso de disco en PC

Correlación

Muestra la relación lineal positiva, negativa, fuerte o débil entre variables



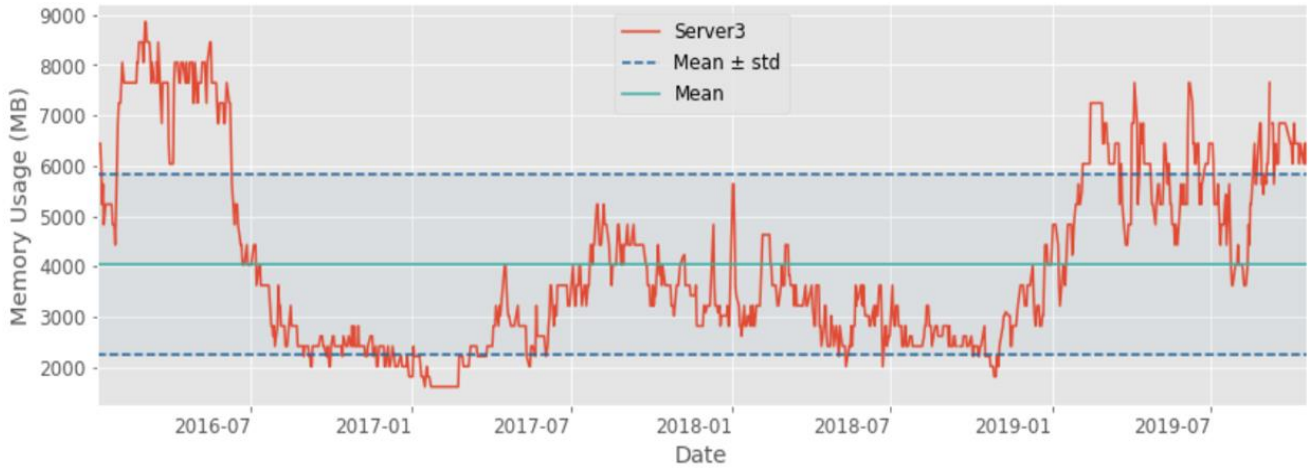
Relación entre salario y nivel educativo

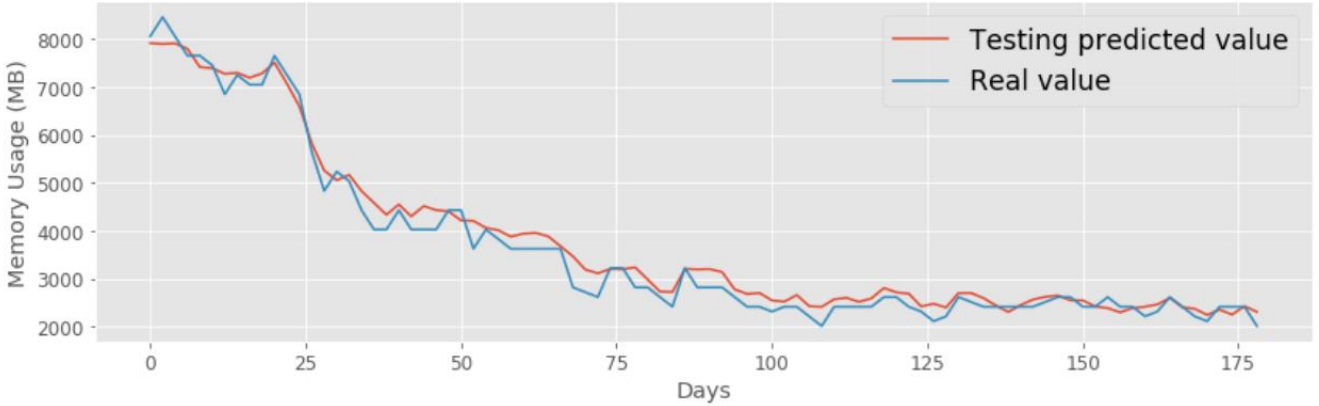
Visualización Datos

Julián Darío Miranda-Calle
Visualization 101

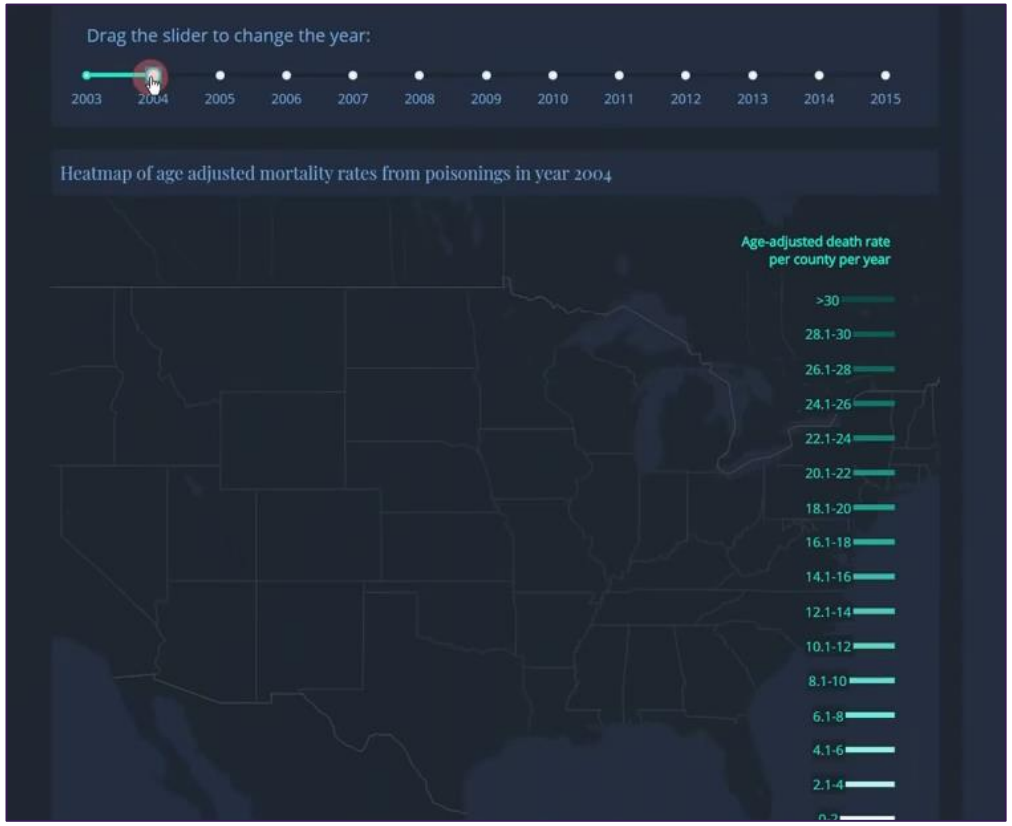
Tipos de formatos

Estáticos





Dinámicos



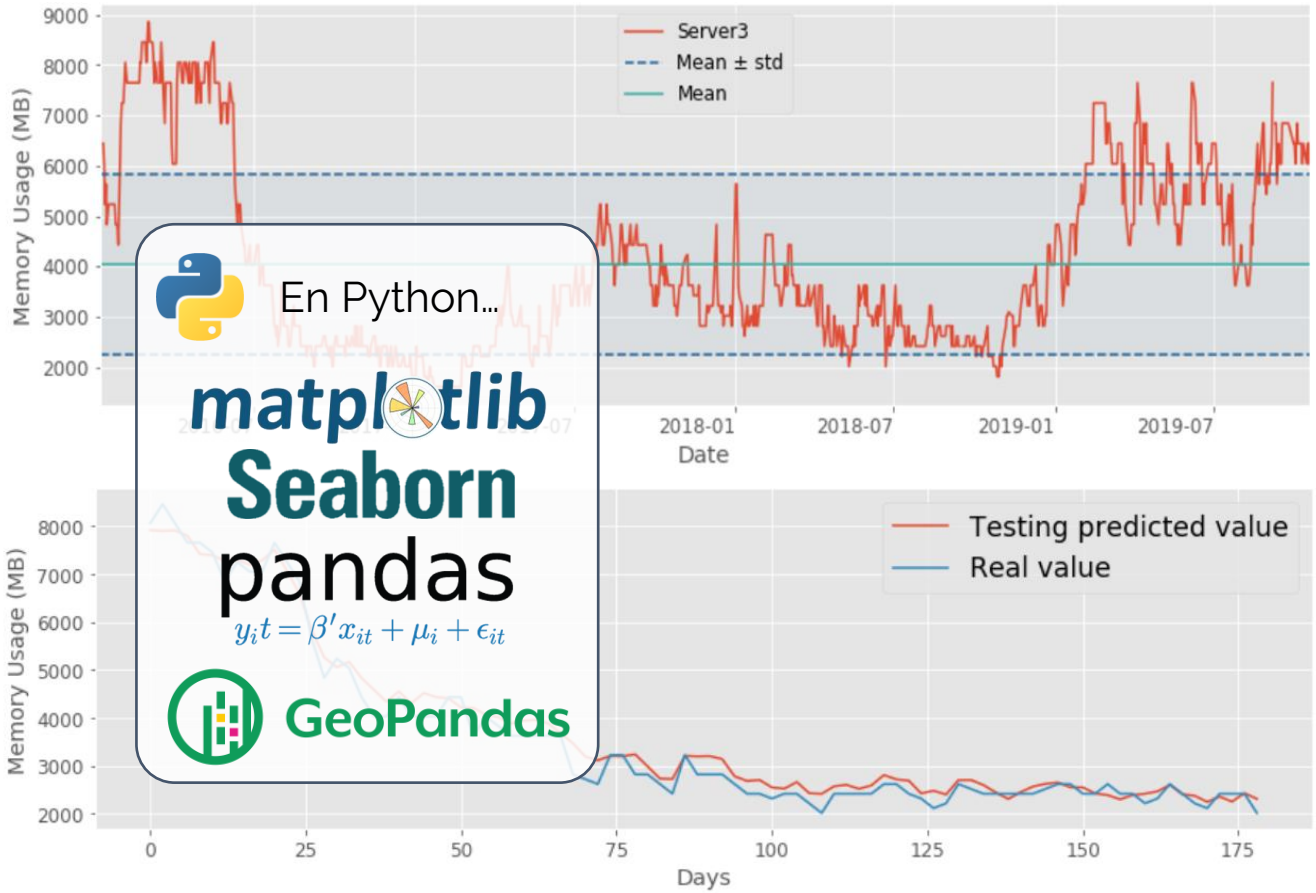
Tomado de: https://dash-gallery.plotly.host/Portal/

Visualización Datos

Julián Darío Miranda-Calle
Visualization 101

Tipos de formatos

Estáticos



Dinámicos

Tomado de: https://dash-gallery.plotly.host/Portal/

Julián Darío Miranda-Calle
Visualization 101

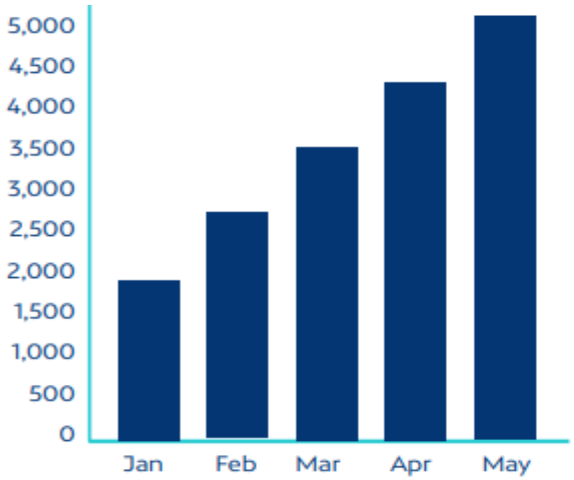
Tipos de gráficos

Gráfico de barras

Una de las formas más populares de presentar los datos, versátil y típicamente usadas para comparación de datos categóricos.

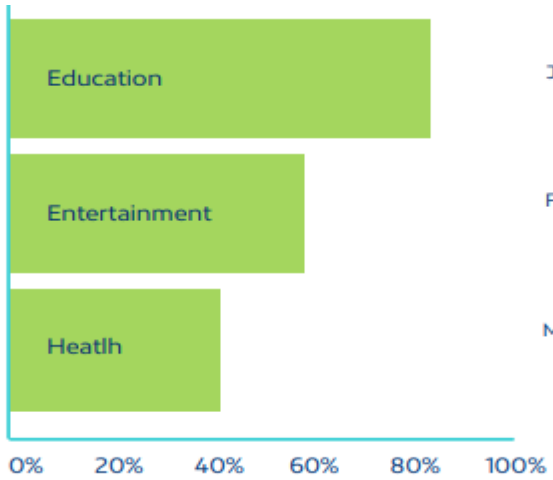
Barras verticales

Usadas generalmente para datos cronológicos, de izquierda a derecha



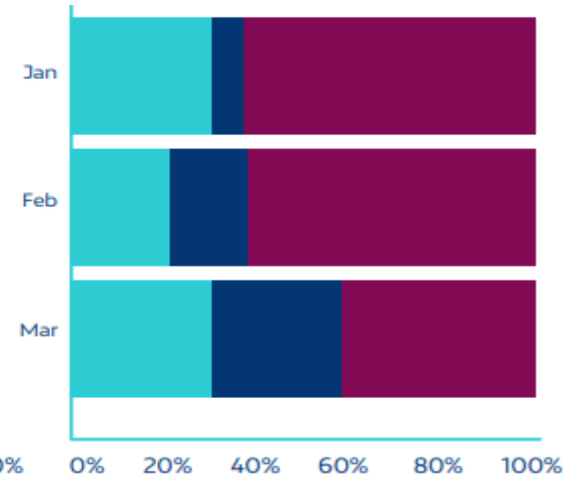
Barras horizontales

Usadas para la visualización de categorías



Barras apiladas

Usadas para visualizar categorías que hacen parte de un todo (100%)

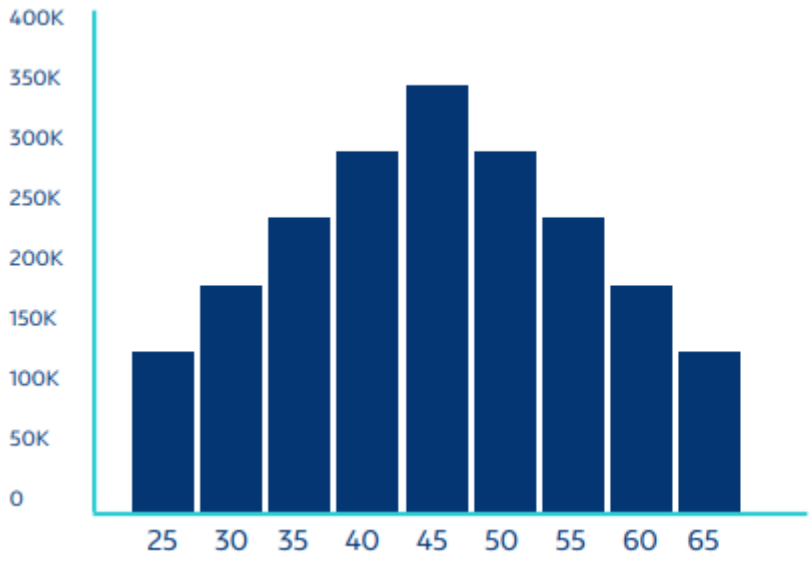


Tomado de: Netquest, 2020.

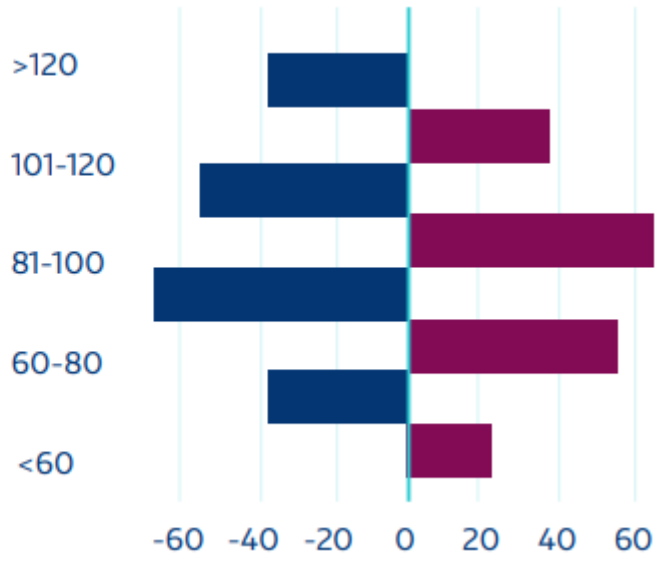
Tipos de gráficos

Histogramas

Los histogramas suelen representar la frecuencia de ocurrencia de la materialización de variables. Ofrecen la distribución de una población o muestra de datos.



Barras verticales



Barras horizontales

Tomado de: Netquest, 2020.

Julián Darío Miranda-Calle
Visualization 101

Tipos de gráficos

Gráfico de torta

Consisten en un círculo dividido en sectores que representan proporciones del total. Generalmente son divididos en no más de cinco categorías.

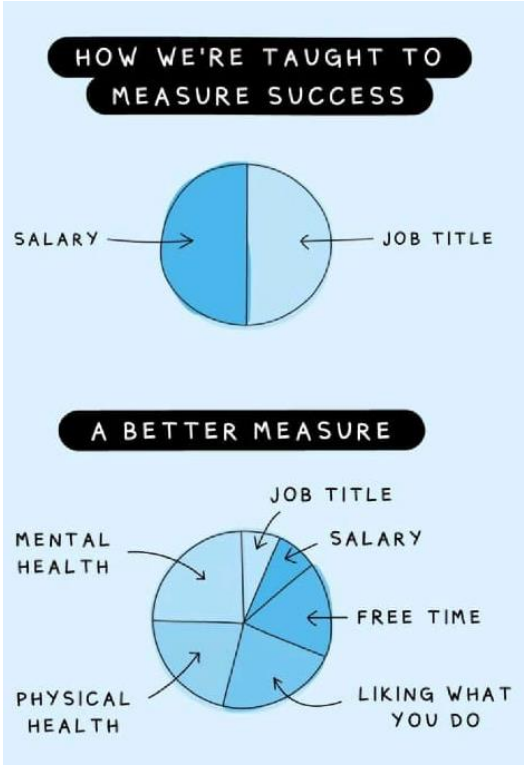
Torta estándar

Usados para exhibir la relación entre las partes de un todo

Dona

Una variación de estilo que facilita la inclusión de valores.

Tomado de: Netquest, 2020.



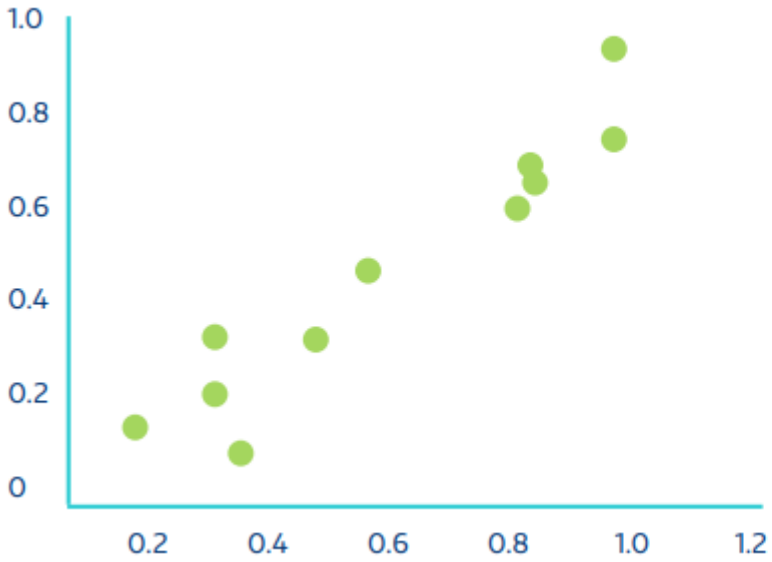
Tomado de: Lizand Mollie, 2021.

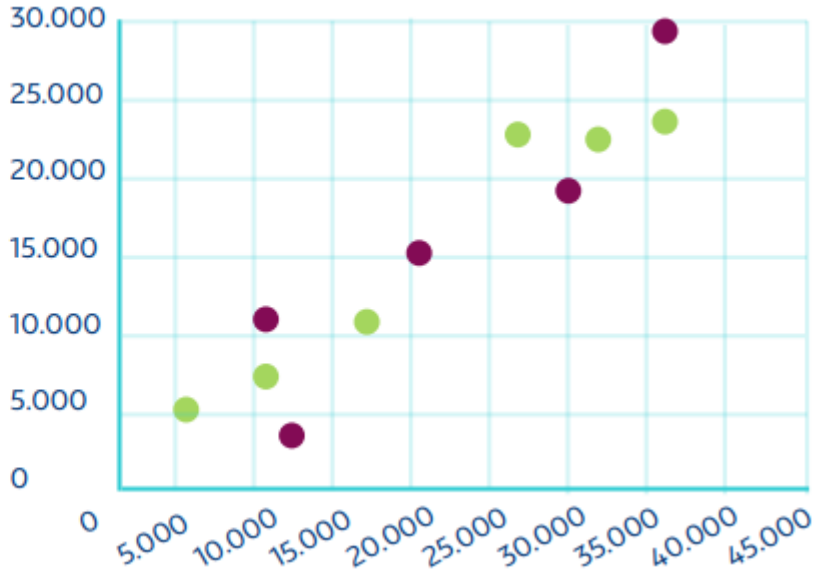
Tipos de gráficos

Gráfico de dispersión

Usan el esparcimiento de los datos para mostrar su relación. Suelen utilizarse para graficar pares de variables en un plano cartesiano.

Los gráficos de dispersión permiten identificar correlaciones altas, bajas positivas y/o negativas de forma rápida e interpretable.





Tomado de: Netquest, 2020.

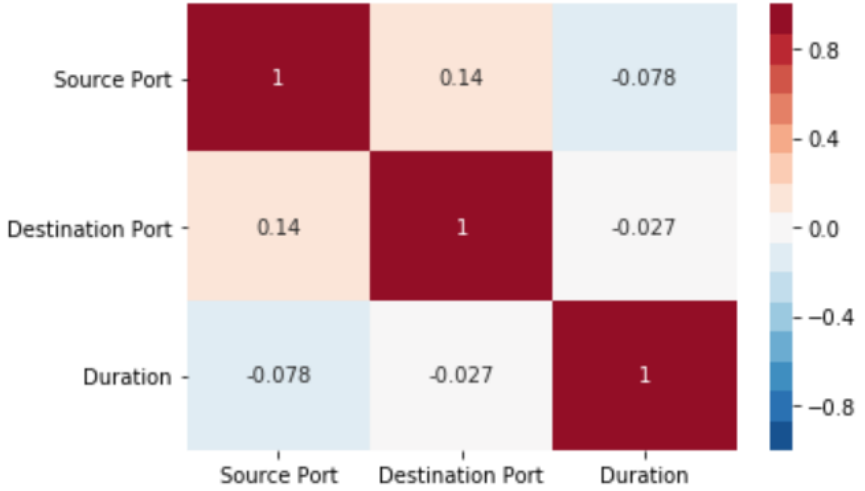
Tipos de gráficos

Mapas de calor

Representan valores individuales mediante la intensidad del color sobre un arreglo de datos generalmente bidimensional.

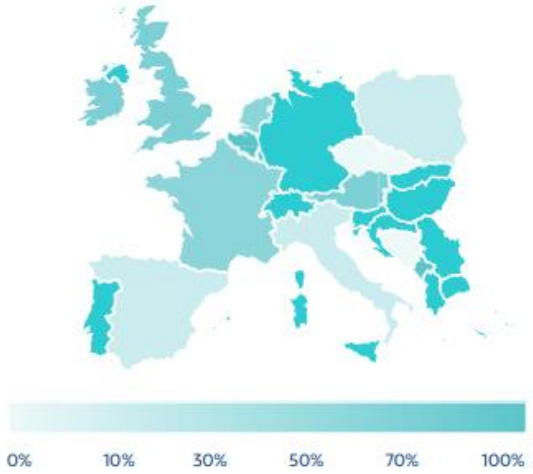
Diagrama de mosaico

Suele utilizarse para representar la correlación entre pares de variables, mediante su intensidad.



Mapa de color

Permite representar intensidad de una variable sobre un plano geográfico.

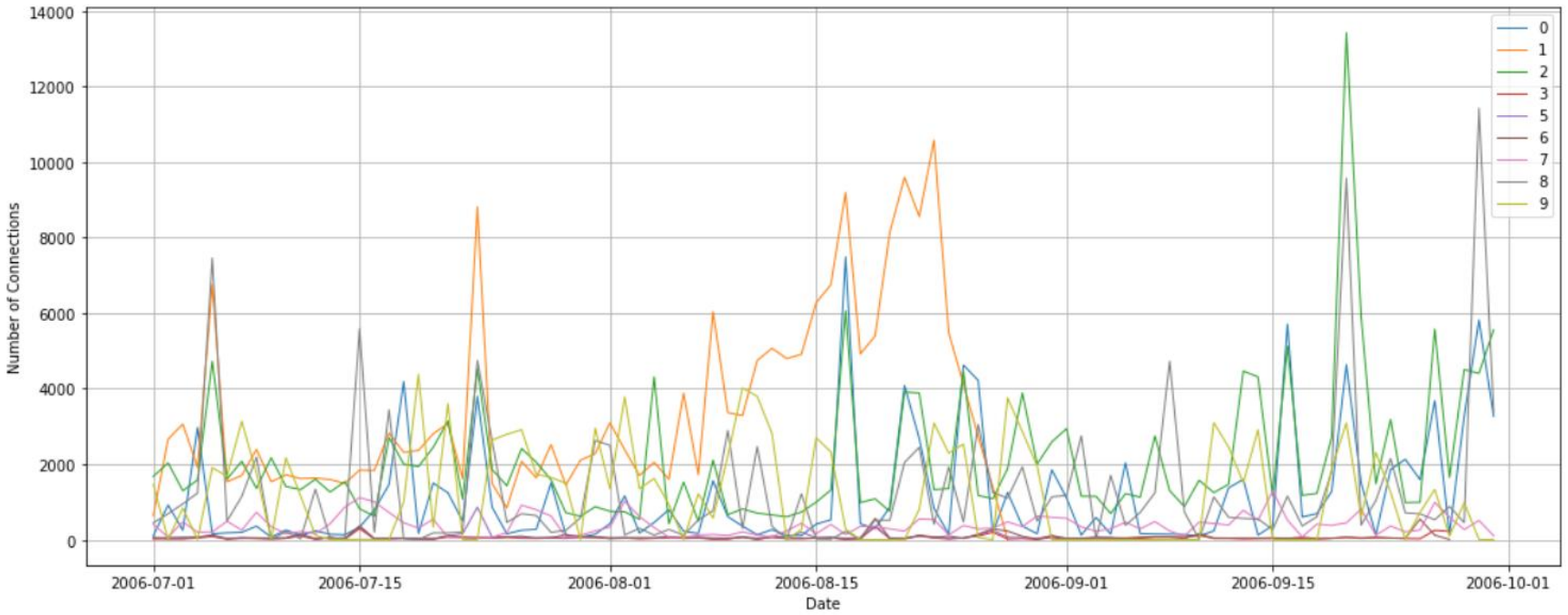


Tomado de: Netquest, 2020.

Tipos de gráficos

Gráficos de línea

Permiten representar cambios de una variable en el **tiempo**, **tendencias** y **repeticiones**, entre otros. Son especialmente útiles para mostrar relaciones, aceleración, desaceleración y volatilidad entre variables en el tiempo.



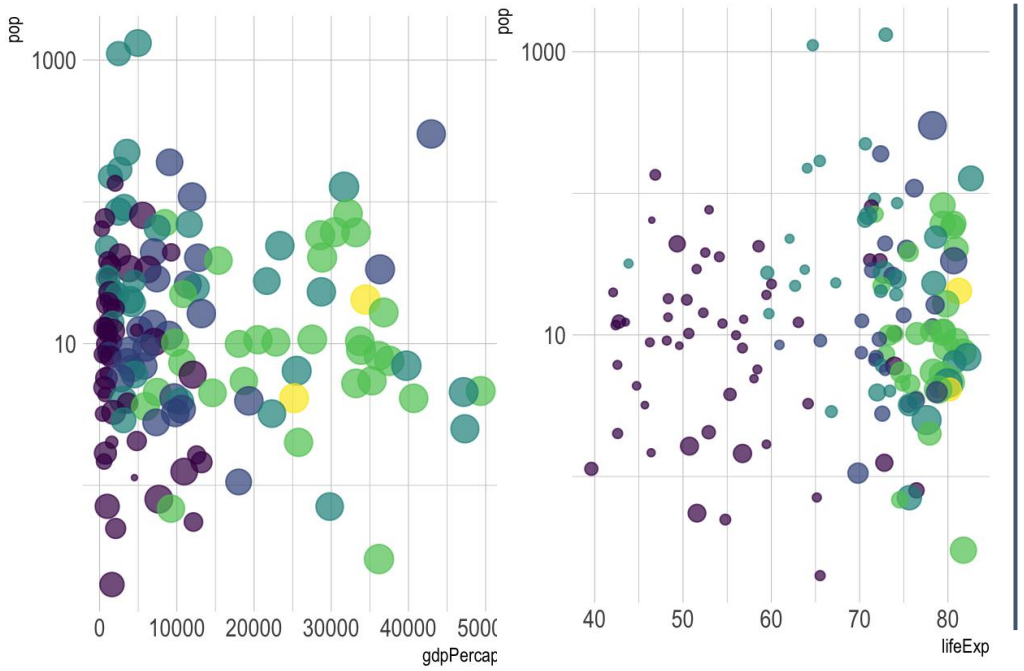
Julián Darío Miranda-Calle
Visualization 101

Tipos de gráficos

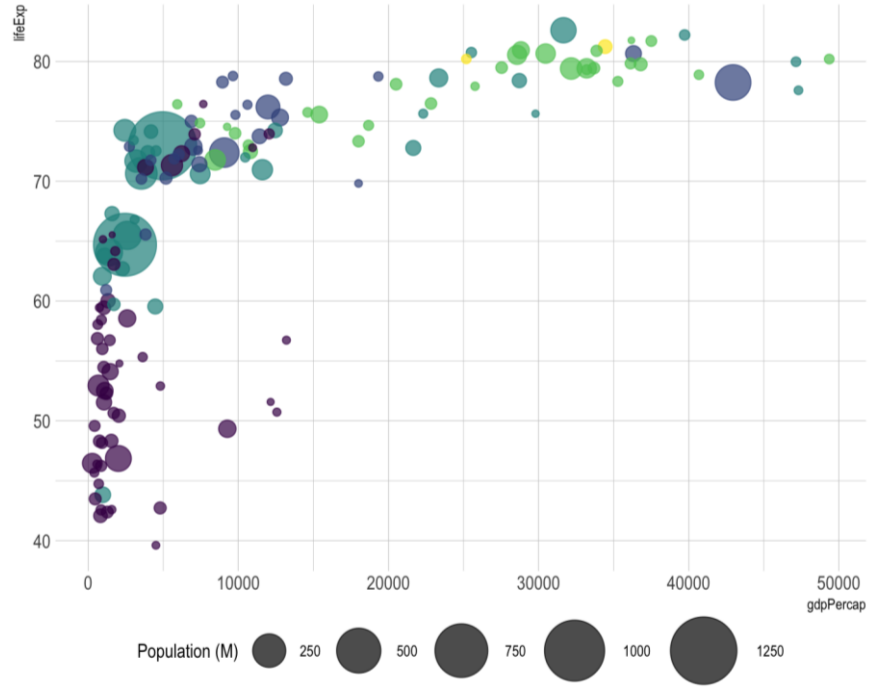
Gráfico de burbujas

Permiten mostrar la relación entre dos, tres o cuatro variables en un plano, acentuando la dispersión de los datos. Su propósito es destacar comparaciones nominales.

Gráficas incorrectas



Gráfica correcta



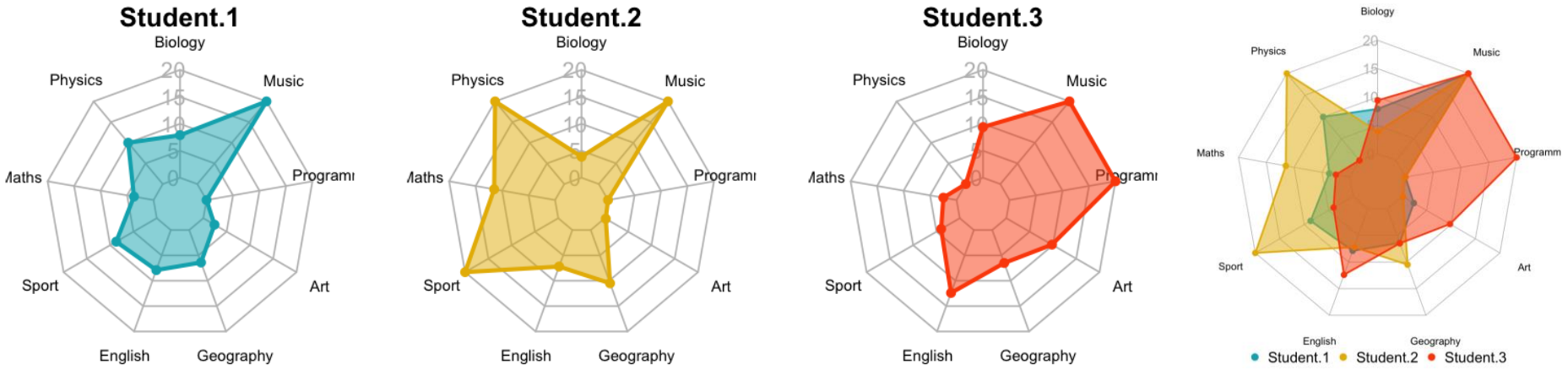
Tomado de: Data to Viz, 2021.

Tipos de gráficos

Gráficos de radar

Una forma de representación poligonal irregular contenida dentro de un polígono regular.

Los radios que guían los vértices de los polígonos irregulares, son los ejes sobre los cuales los valores están representados.



Tomado de: Datanovia, 2021.

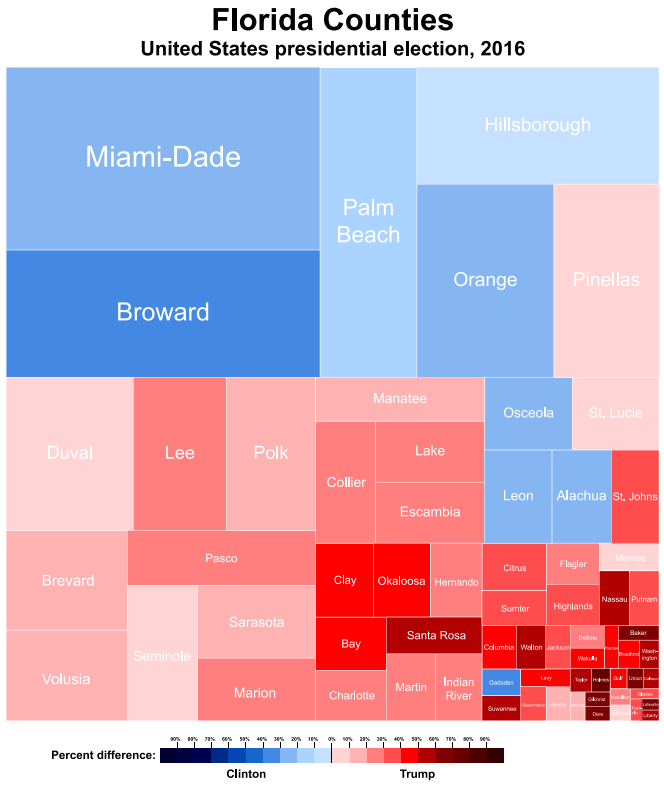
Julián Darío Miranda-Calle
Visualization 101

Julián Darío Miranda-Calle
Visualization 101

Tipos de gráficos

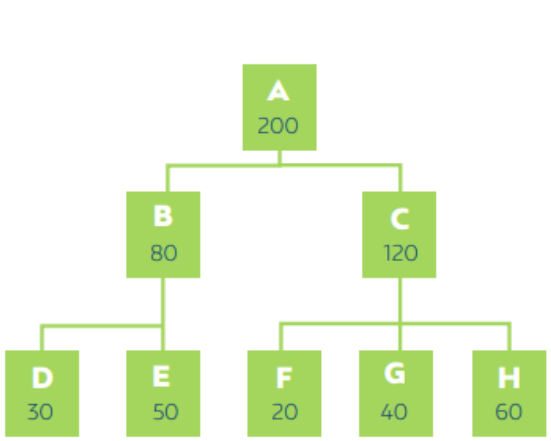
Mapas de árbol

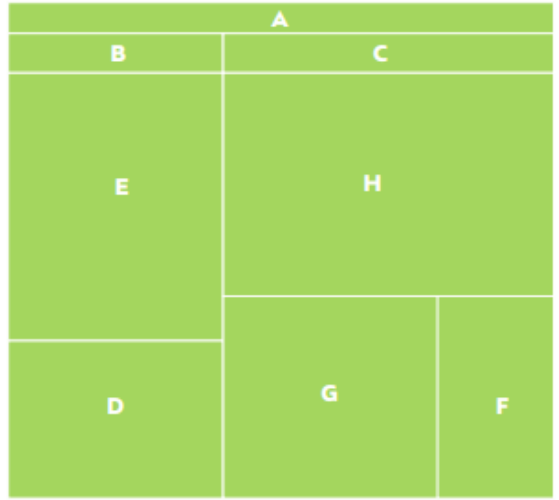
Muestra jerárquica de rectángulos anidados en el que su área es proporcional al valor de los datos que las variables representan.



Tomado de: Ali Zifan, 2017.

Cada rectángulo puede representarse como un nodo interno de un árbol, en donde el rectángulo inicial representa la raíz y los rectángulos más internos representan las hojas.





Tomado de: Netquest, 2020.

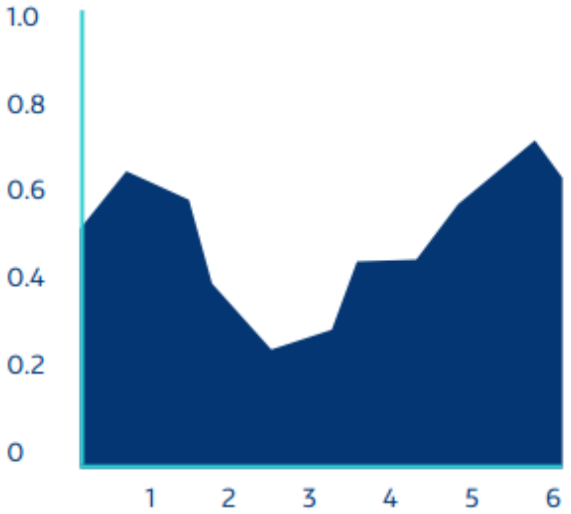
Tipos de gráficos

Gráficos de área

Los histogramas suelen representar la frecuencia de ocurrencia de la materialización de variables. Ofrecen la distribución de una población o muestra de datos.

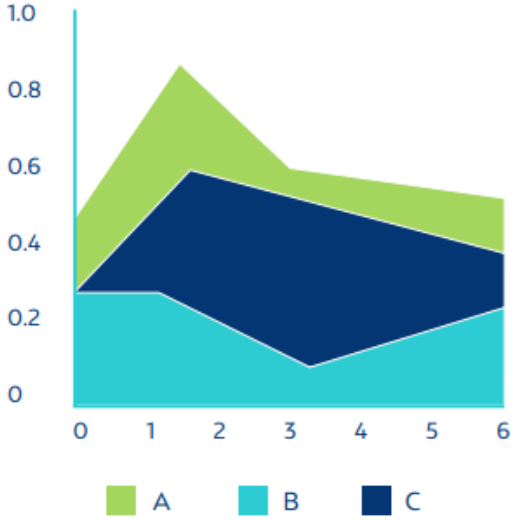
Área estándar

Comparar procesos en el tiempo



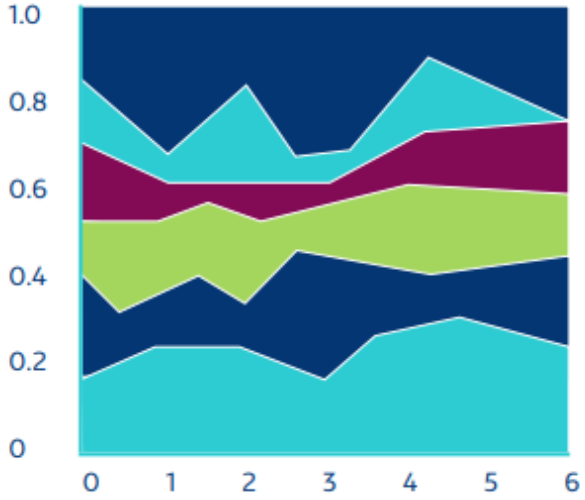
Área apilada

Visualizar relaciones e identificar contribuciones



Área apilada completa

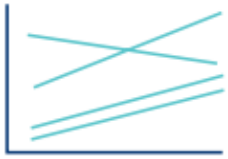
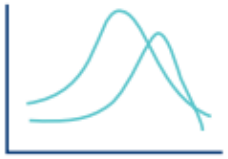
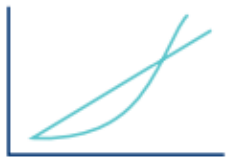
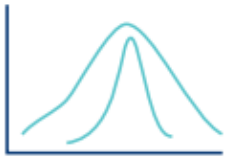
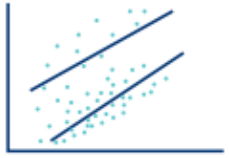
Ver distribución de categorías como parte de un todo



Tomado de: Netquest, 2020.

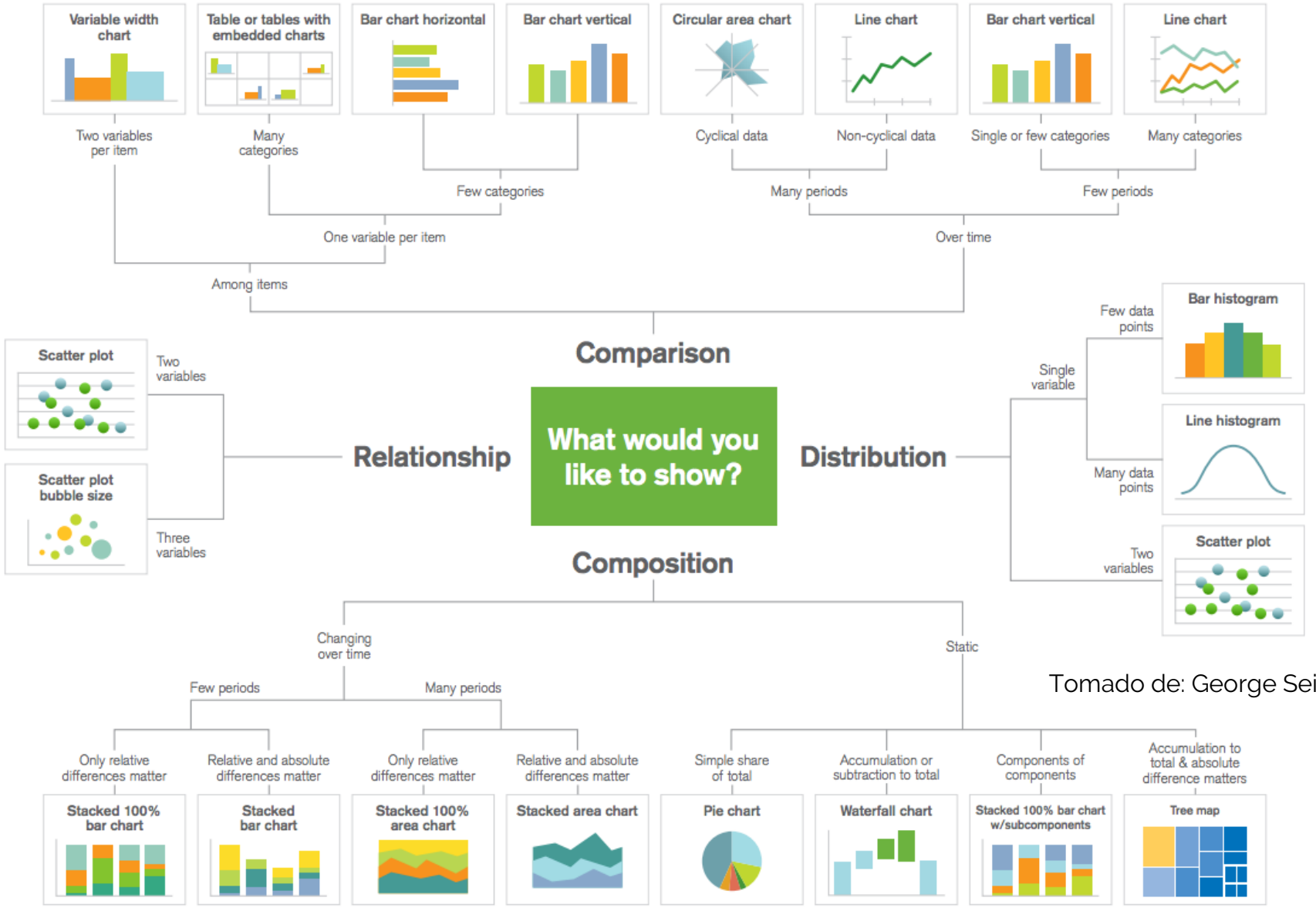
Julián Darío Miranda-Calle
Visualization 101

Tipos de gráficos

Pattern	Example	Pattern	Example	Pattern	Example
High, low and in between		Non-intersecting and intersecting		Random and repeating	
Going up, going down and remaining flat		Symmetrical and skewed		Straight and curved	
Steep and gradual		Wide and narrow		Tightly and loosely distributed	
Steady and fluctuating		Clusters and gaps		Normal and abnormal	

Tomado de: Netquest, 2020.

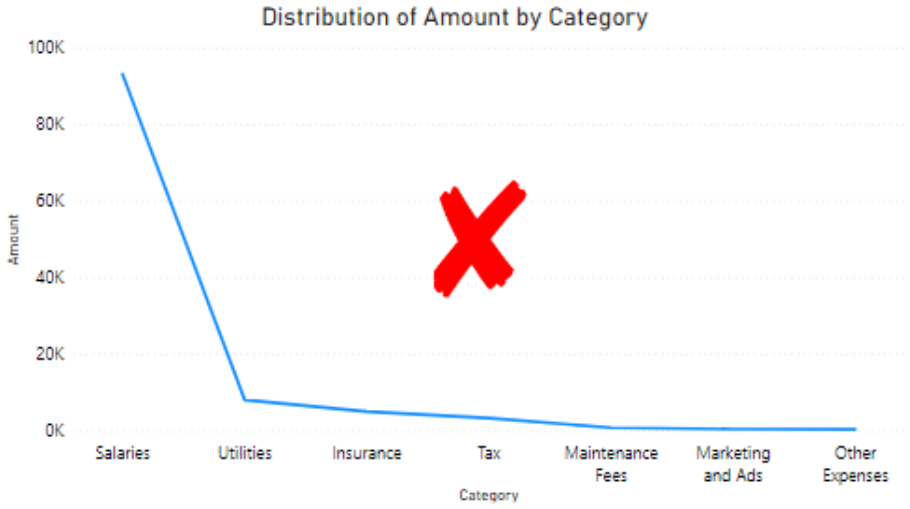
Tipos de gráficos

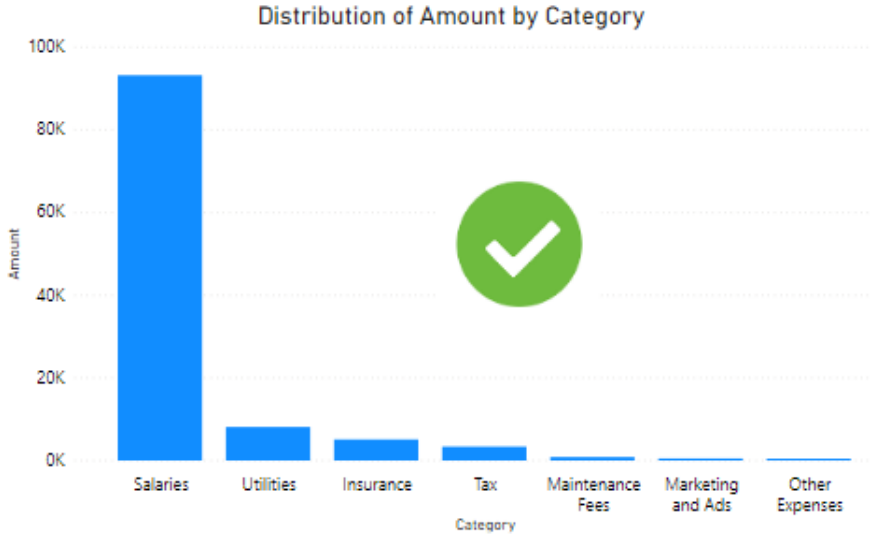


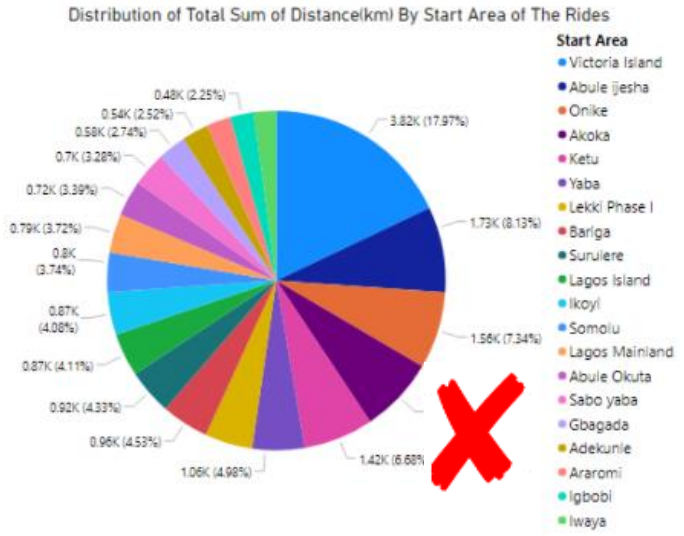
Tomado de: George Seif, 2018.

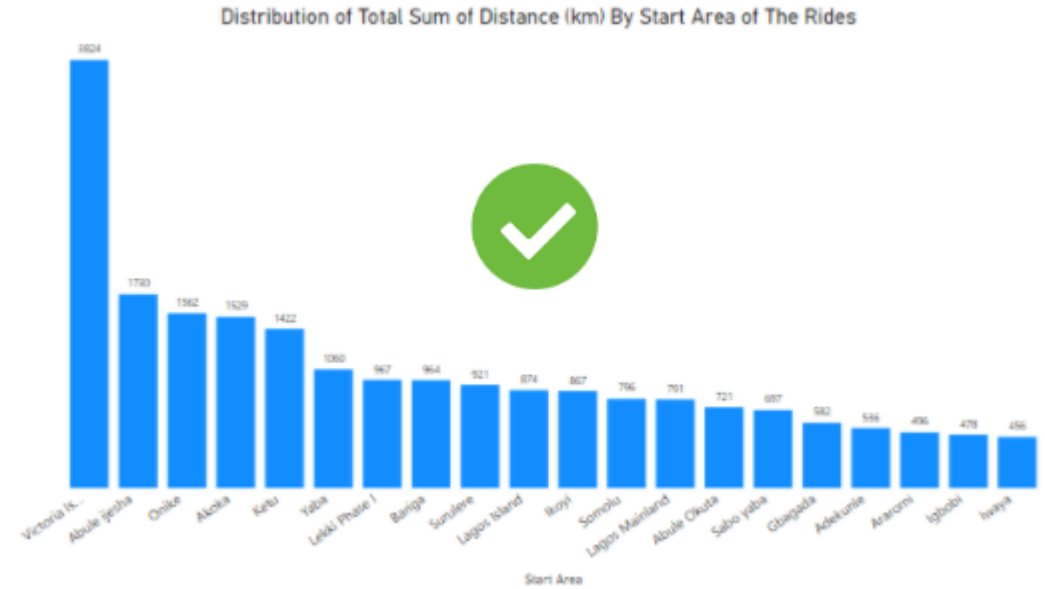
Julián Darío Miranda-Calle

Ejemplos de visualización









Tomado de: Resagratia, Research, Business Intelligence and Advanced Analytics, 2021.

Caso de uso

Julián Darío Miranda-Calle
Visualization 101

Vamos a identificar y visualizar datos sobre ataques de ciberseguridad adaptados del dataset UNSW-NB15 del Cyber Range Lab de Australian Centre for Cyber Security (ACCS)

Source code and datasets available **here**

Tipos de gráficas:

- Diagrama de barras
- Diagrama de torta
- Mapa de calor
- Diagrama de dispersión
- Diagrama de violín
- Gráfico de distribución
- Gráfico de KDE
- Gráfico categórico
- Gráficos de relación
- Gráfico de tiras

Attack category	Protocol	Source IP	Source Port	Destination IP	Destination Port	Start time
RECONNAISSANCE	TCP	175.45.176.0	13284	149.171.126.16	80	2015-01-22 11:50:14
EXPLOITS	UDP	175.45.176.3	21223	149.171.126.18	32780	2015-01-22 11:50:15
EXPLOITS	TCP	175.45.176.2	23357	149.171.126.16	80	2015-01-22 11:50:16
EXPLOITS	TCP	175.45.176.2	13792	149.171.126.16	5555	2015-01-22 11:50:17
EXPLOITS	TCP	175.45.176.2	26939	149.171.126.10	80	2015-01-22 11:50:18
...	...	...	...	...	...	...
DOS	TCP	175.45.176.0	33654	149.171.126.12	80	2015-02-18 12:21:06
FUZZERS	TCP	175.45.176.3	36468	149.171.126.15	445	2015-02-18 12:21:07
RECONNAISSANCE	TCP	175.45.176.2	64395	149.171.126.18	111	2015-02-18 12:21:07

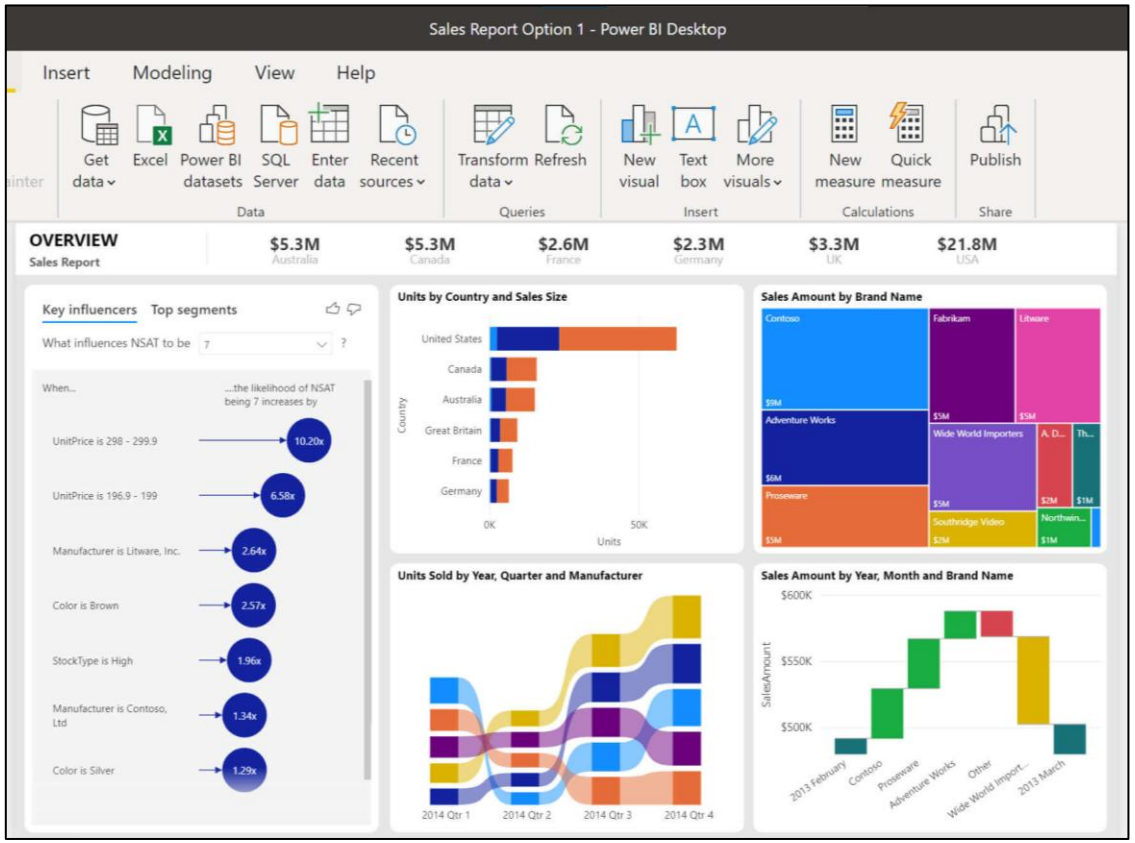
Miranda-Calle, J.D., Reddy C., V., Dhawan, P. and Churi, P. (2021), "Exploratory data analysis for cybersecurity", World Journal of Engineering.

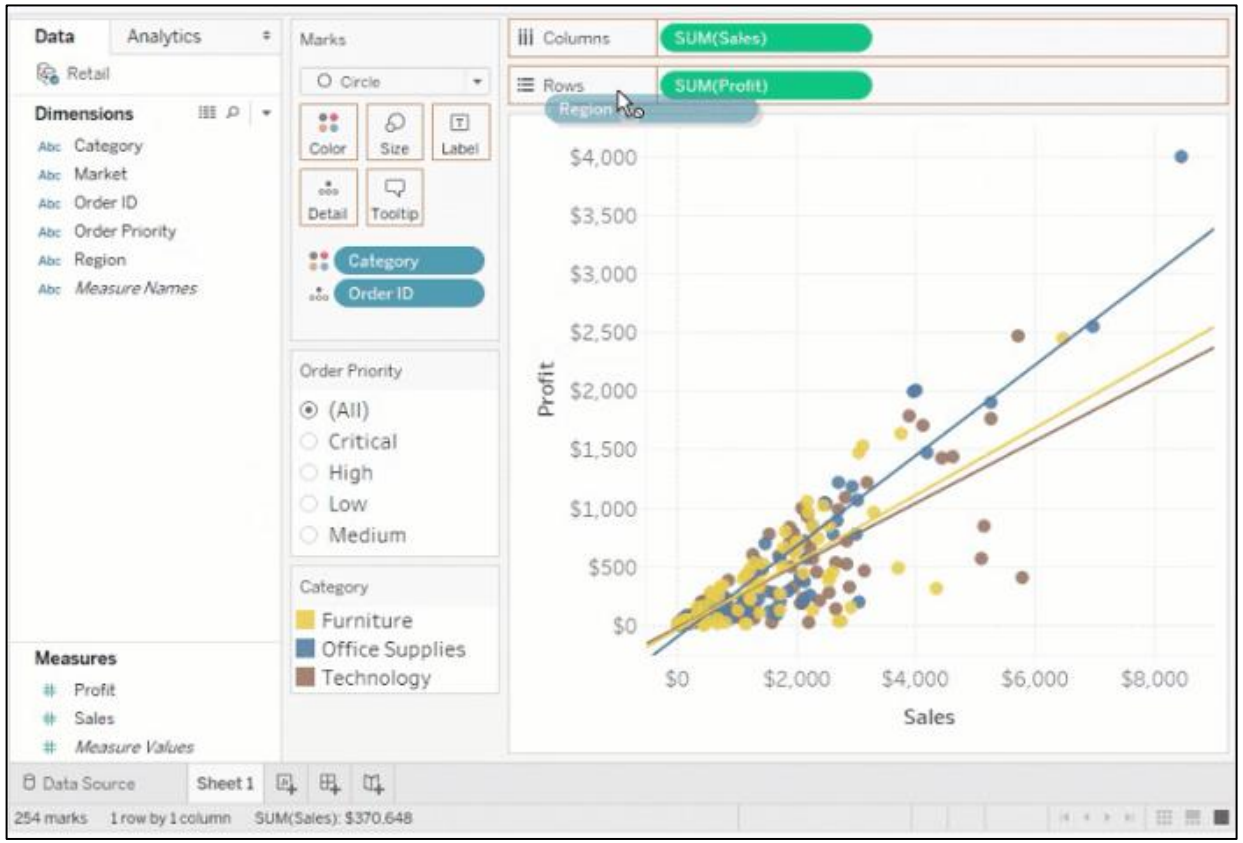
Entorno estratégico

Julián Darío Miranda-Calle
Visualization 101

Herramientas de visualización

Power BI





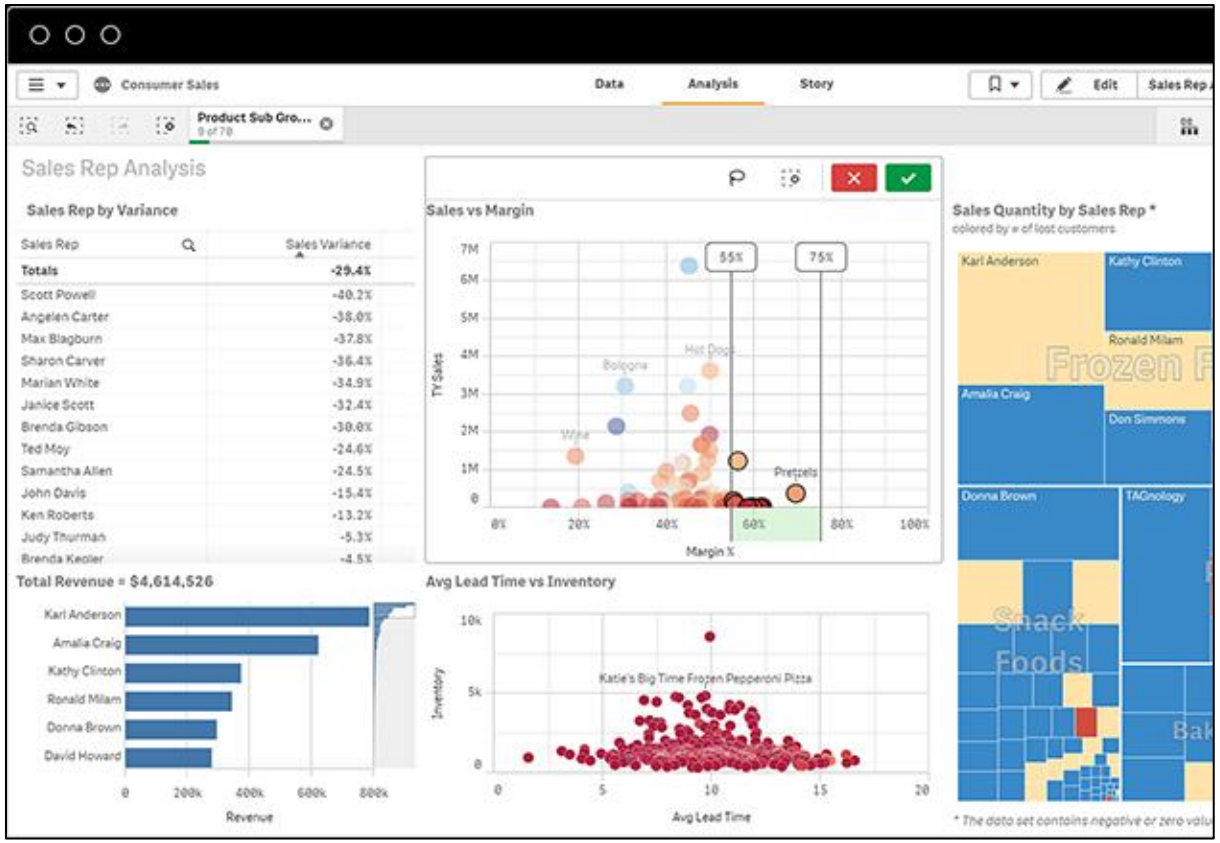
DPhi Tech
Data Analysis Bootcamp

Entorno estratégico

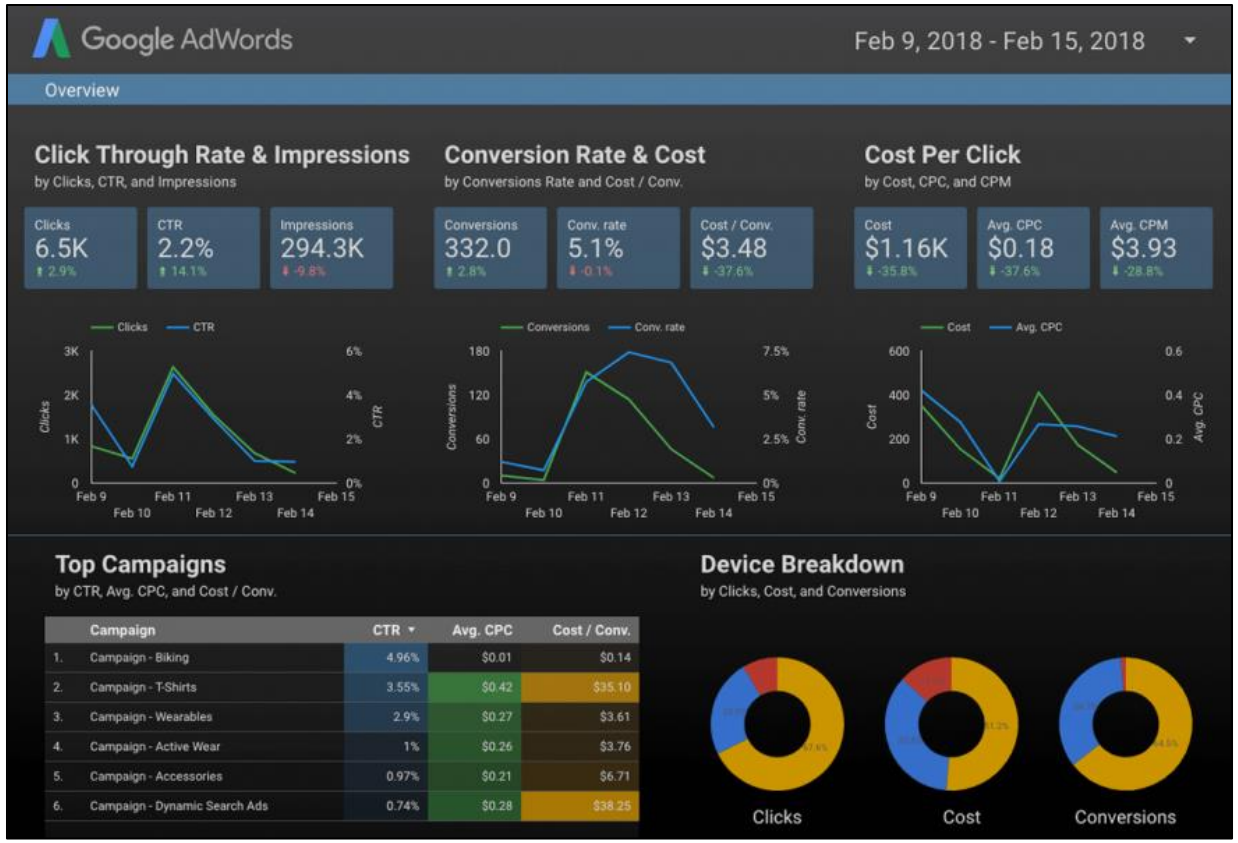
Julián Darío Miranda-Calle
Visualization 101

Herramientas de visualización

Qlik



Data Studio



Julián Darío Miranda

juliandariomiranda@gmail.com

/juliandariomiranda

0000-0002-7580-2361

www.researchgate.net/profile/Julian_Miranda2