

Laboratorio de Datos



Calidad de Datos



Laboratorio de Datos

Calidad de Datos
... por Viviana Cotik (y modificaciones de P. Turjanski)

1º parte de
la materia

Recorrido de la materia (hasta ahora)

- ✓ Lenguaje de programación para trabajar en nuestros proyectos



- ✓ Etapas de un proyecto de Ciencias de Datos

- ✓ Modelado de Datos



- ✓ Representación de los Datos

Materia	
Código	Nombre
1	Laboratorio de Datos
2	Análisis II
3	Álgebra Lineal

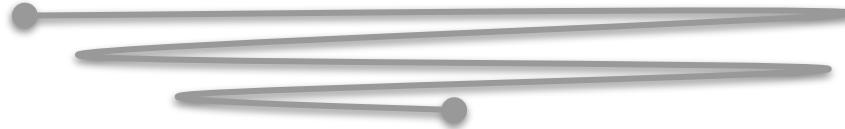
Unidad			
Código	Materia	Título	Descripción
1	Laboratorio de Datos	Administración de datos	Obtención y Manejo de los datos
1	Análisis II	Modelos Explicativos	Construcción de modelos explicativos
1	Álgebra Lineal	Modelos Predictivos	Construcción de modelos predictivos
2	Laboratorio de Datos	Integrales sobre curvas y superficies	Integrales en múltiples variables
2	Análisis II	Ecuaciones Diferenciales	Diseño y análisis de sistemas dinámicos

- ✓ Maneras de consultar los Datos AR/SQL

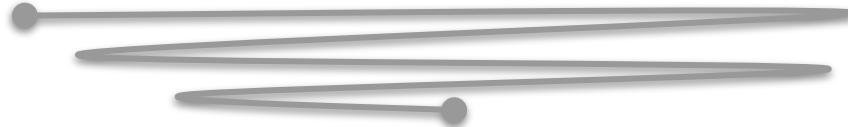
- ✓ Estrategias para mejorar la calidad de los datos desde el diseño de la BD

Introducción a Calidad de Datos

Trabajo individual



Actividad - Consigna



- Ingresar al siguiente *LINK* y completar la encuesta

<https://forms.gle/YEpBpHq29ZXdLXq7>

Importante: No se pueden hacer preguntas sobre el formulario

- Objetivo. Conocer cuál será el transporte más utilizado (por los alumnos de nuestro curso) el día de hoy para retirarse de Ciudad Universitaria. En caso de que piense que va a utilizar más de un transporte responder con respecto al primero que vaya utilizar.

Trabajo Grupal



- *Solicitar al docente que comparta las respuestas de la encuesta*
- *En grupos de 3 integrantes responder las siguientes preguntas*



1. ¿Cuántos estudiantes respondieron la encuestas?
2. ¿Cuántos estudiantes eran en total?
3. ¿Algún estudiante respondió más de una vez (respuestas repetidas)?
4. A partir de los datos, ¿pueden responder cuál será el transporte más utilizado por los alumnos para retirarse de Ciudad Universitaria? ¿Cuál será?
5. ¿Algunos se retiran caminando? Y en caso afirmativo, ¿existe alguna relación entre estos que se retiran caminando y el número de calzado que utilizan?
6. Las respuestas ¿tenían datos faltantes?
7. ¿Observaron respuestas con cierta inconsistencia?
8. ¿Se encontraron que todas las respuestas tenían el formato correcto?
Mencione al menos 3 casos en se encontraron con respuestas en un formato inesperado
9. ¿En cuántas respuestas obtuvieron “Ciudad Autónoma de Buenos Aires” como respuesta a Provincia?
(www.argentina.gob.ar/pais/provincias)

FIGURAR EN UN LISTADO DE INCUMPLIDORES ARRUINO UN NEGOCIO

Un banco debe pagar \$ 120.000 por incluir mal a un cliente en Veraz

Daniel Gutman

El Banco Río lo incluyó en las listas negras de deudores de la Organización Veraz y solicitó al Banco Central que lo inhabilitara. Pero todo era un error, porque no había existido ningún incumplimiento. El cliente hizo juicio y obtuvo una sentencia de Cámara a su favor. Hasta ahí, un caso igual a muchos otros que ha habido en los últimos años. Lo novedoso es que la Cámara en lo Comercial acaba de establecer la que seguramente sea la indemnización más alta en este tipo de casos: el Banco Río deberá pagarle 120.000 pesos a su ex cliente.

A esa cifra deberán sumársele los intereses a la tasa activa del Banco Nación desde la fecha de inhabilitación en los registros del Central, que es mayo de 1996, lo que llevaría la indemnización a más de medio millón de pesos, según los abogados del demandante.

La importancia de la indemnización —según se explicó en el fallo— tiene que ver con que el damnificado es un empresario que estaba en pleno proceso de ampliación de sus negocios.

El hombre, dueño de una confitería, estaba construyendo un edificio en la avenida Cruz en el cual pensaba instalar una concesionaria de autos, además de una confitería y salón de fiestas en la planta alta. Sin embargo, en mayo de 1996 quedó sin posibilidad de obtener crédito y operar con cheques, por lo que la obra y sus proyectos quedaron inconclusos.

Así, la Sala B de la Cámara —en un voto de la jueza María de Díaz Cordero, al que adhirió Enrique Butty— aplicó el concepto de "pérdida de chance". Es decir, el Río deberá indemnizar al empresario porque lo privó de una oportunidad de ganar dinero.

Las pruebas presentadas y la trayectoria de Eloy Domínguez Álvarez convencieron a la jueza de que él "tenía intención de culminar con la construcción del edificio y ampliar sus negocios" y de que lo hubiera hecho "de no haber existido la arbitraria y errónea decisión adoptada por la entidad bancaria demandada".

Caso

1. Leer el artículo
2. En grupos de 3 integrantes ...
 - a. Describir el problema
 - b. ¿Cuál es la causa del problema?
 - c. ¿Quiénes se benefician al contar con datos de calidad?



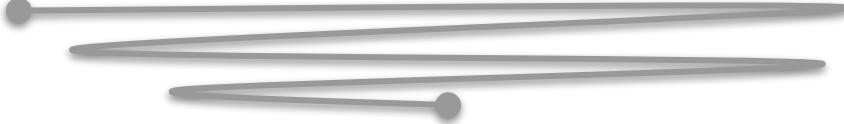
Calidad de Datos



¿Cuál es la definición de Calidad de Datos? (discutir 5 min.)



Definiciones



- “Un dato o conjunto de datos X tiene mayor calidad que un dato o conjunto de datos Y , si X satisface las necesidades del usuario mejor que Y ” [Redman, 1996]
- “Satisfacer de manera consistente las expectativas de los usuarios” [English, 1999]

... son definiciones subjetivas

Calidad de Datos



*¿Cuáles son las consecuencias de contar con datos de mala calidad?
(discutir 5 min.)*



Calidad de Datos

¿Cuáles son las consecuencias de contar con datos de mala calidad?

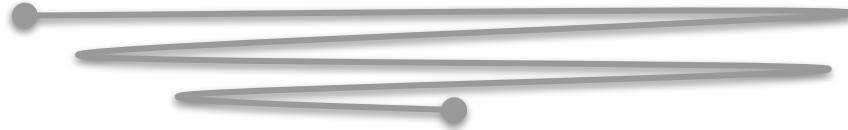
- Desconfianza
- Insatisfacción de los clientes
- Costos innecesarios
- Impacto en la toma de decisiones
- ...



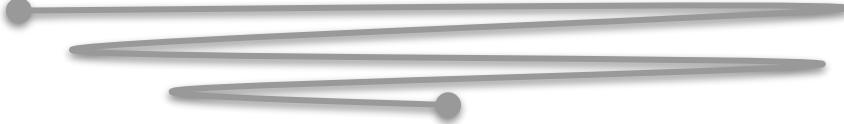
Gráfico tomado de AIT solutions

Problemas de Calidad de Datos

Trabajo en equipo

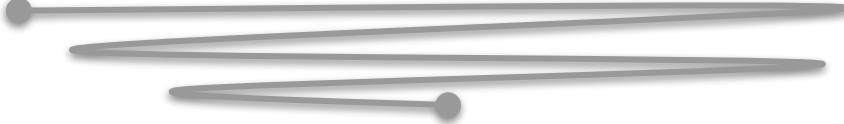


Ejercicio - Consigna



- ✓ Conformar grupos de 3 integrantes
- ✓ Descargar el registro de Datos de Dengue de 2020 correspondiente al Registro del Sistema Nacional de Vigilancia de la Salud 2.0 (<http://datos.salud.gob.ar/dataset/vigilancia-de-dengue-y-zika>)
-> Vigilancia de Dengue y Zika - 2020 (.xls)
- ✓ Responder las siguientes preguntas:
 1. ¿Detectan problemas de calidad de datos?
En caso afirmativo, mencionar cuáles son y caracterizarlos
 2. ¿Piensan que los datos provistos provienen de una única tabla de una BD relacional?

Ejercicio - Debate



Algunos problemas de Calidad de datos detectados ...

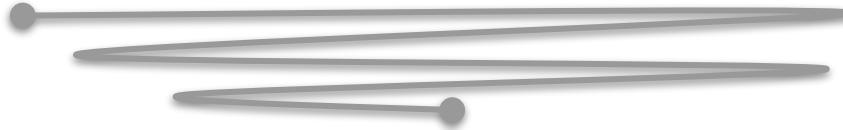
- ✓ Datos nulos: departamento_nombre, provincia_nombre, provincia_id
- ✓ Columnas intercambiadas:
 - grupo_edad_id vs grupo_edad_desc
 - evento_nombre vs semanas_epidemiologicas
- ✓ Códigos incorrectos (departamento_id)
- ✓ Sólo casos de dengue (no de zika)
- ✓ Rangos erróneos en grupo etario
- ✓ Algunos departamentos escritos con comillas (ej. Apostoles vs Apóstoles, etc.)

Trabajo Grupal



- Conformar grupo de 3 estudiantes
- Discutir qué acciones tomarian (en este caso) para mejorar la calidad de los datos

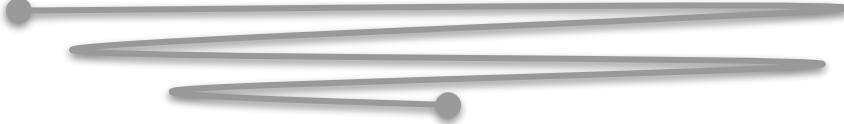
Ejercicio - Solución posible



Possible solución ... armar tablas normalizadas:

1. Chequear y corregir datos (aumentar la calidad de los datos)
2. Crear nuevas tablas con id y descripciones.
En la tabla original sólo deberían figurar los ids (y no las descripciones)
3. Tomar alguna decisión (y documentar) sobre qué hacer con los registros que aparecen dos veces
(pero con distinta cantidad de casos reportados)

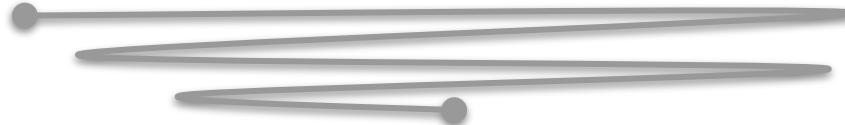
Algunos problemas habituales



- ✓ Valores no estandarizados
 - NETOFAGASTA
 - ANMTOFAGASTA
 - ANT0FAGASTA
 - ANTO9FAGASTA
 - ANTOAFAGASTA
 - ANTOFAAGASTA
- ✓ Valores imposibles o poco probables
 - Edad: 200 años
- ✓ Valores faltantes
 - Registros de personas con el campo e-mail vacío
- ✓ Valores desactualizados
 - Ocurrencias duplicadas
 - Falta de datos históricos
 - Inconsistencia entre aplicaciones o en una misma aplicación
 - Datos de pacientes en dos servicios distintos de un hospital
 - Datos de pozos petroleros en dos aplicaciones distintas (perforación, producción)
 - Información crítica que no es confiable
 - Hay personas habilitadas a votar que han fallecido

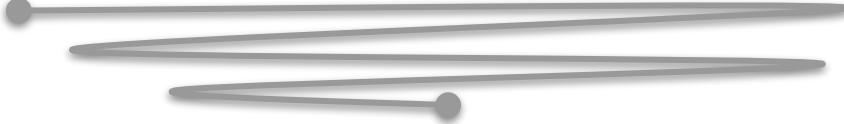
Causas de problemas en Calidad de Datos

Trabajo Grupal



- Conformar grupo de 3 estudiantes
- Discutir cuáles pueden ser las causas por las que los datos tienen problemas de calidad

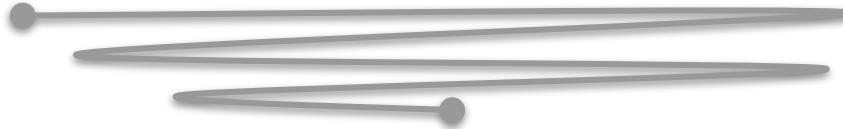
Posibles causas de problemas



Posibles causas de problemas ...

- ✓ Problemas masivos que reparan un dato, pero no reconstruyen información relacionada
Ej. Recuperar registros de pago (que por un problema de software fueron perdidos) y posteriormente no eliminar de la tabla de deudores a aquellos clientes que saldaron su deuda.
- ✓ Misma información cargada en distintos sistemas
Ej. Datos de Estudiante en SIU-Guarani y datos de estudiante en el Campus (puede diferir el email)
- ✓ Valores predeterminados
Ej. Fecha de Nacimiento 01-01-2021

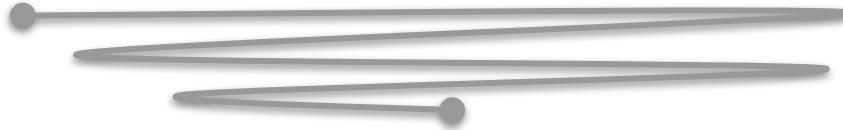
Posibles causas de problemas



Las causas pueden depender de ...

- *Calidad de software (usabilidad, interfaz [obligatoriedad de carga], seguridad)*
- *Definición de procesos asociados a los datos*
- *Diseño de la BD*
- *Falta de capacitación*

Posibles causas de problemas



Los problemas de calidad los podemos clasificar en ...

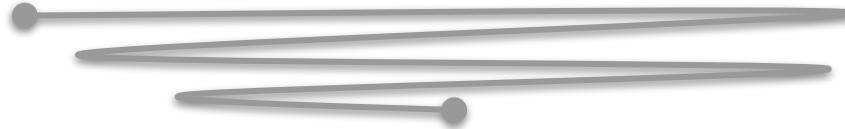
Posibles causas de problemas



1. Problemas asociados a la *INSTANCIA*

- ✓ Datos que han cambiado en el mundo real, y que no fueron actualizados
- ✓ Datos que provienen de distintas fuentes, deberían ser consistentes y sin embargo no lo son
- ✓ Datos que no han sido almacenados con la precisión necesaria (por ejemplo, Y2K)

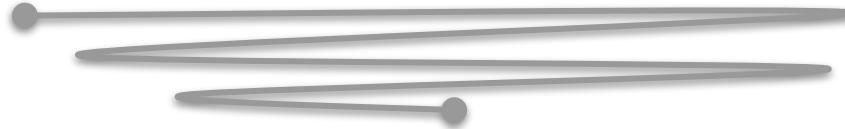
Posibles causas de problemas



2. Problemas asociados al **MODELO DE DATOS**

- ✓ Si se detecta que hay información que no está presente porque no hay forma de almacenarla
-> el modelo de datos físico está incompleto
- ✓ El mundo que se quiere representar evolucionó y no se tradujeron los cambios al modelo
-> pérdida de vigencia del modelo

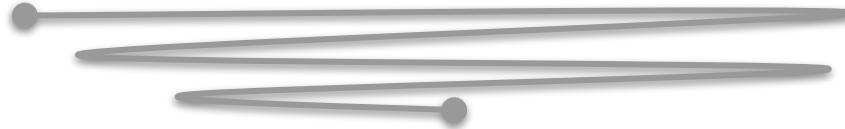
Posibles causas de problemas



3. Problemas asociados a los **PROCESOS**

- ✓ Distintas personas cargan la misma información haciendo distintas asunciones
- ✓ Se carga con una asunción y se usa con otras
- ✓ Modificaciones manuales por procesos
- ✓ Gente que hace modificaciones pero no debería estar autorizada para hacerlas

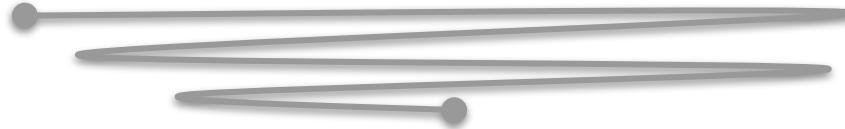
Posibles causas de problemas



4. Problemas asociados a **ERRORES DE SOFTWARE**

- ✓ Datos obligatorios que no se asumen como tales y por lo tanto no se cargan
- ✓ Interfaces poco amigables

Posibles causas de problemas



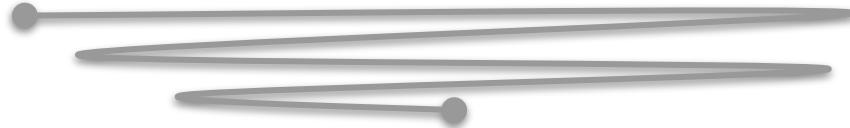
Importante: Sólo el **software** de buena calidad no garantiza la calidad de los datos

Se debe trabajar sobre:

- ✓ La instancia
- ✓ El modelo de datos
- ✓ Los procesos que intervienen en la generación y modificación del dato
- ✓ La consistencia entre las diferentes fuentes de datos

Atributos (o dimensión) de Calidad de Datos

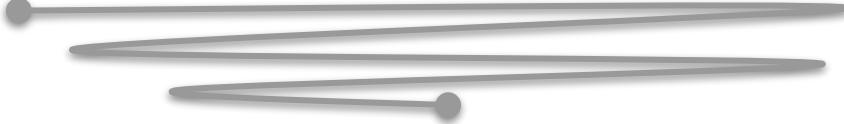
Atributos de Calidad



¿Qué características deberían cumplir los datos para ser de calidad? (5 min.)



Atributos de Calidad

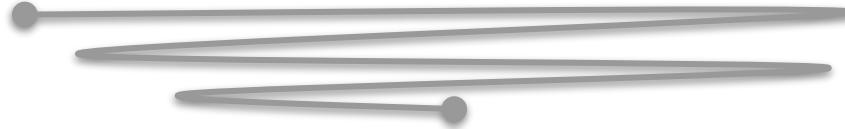


¿Qué características deberían cumplir los datos para ser de calidad?

Deberían ser ...

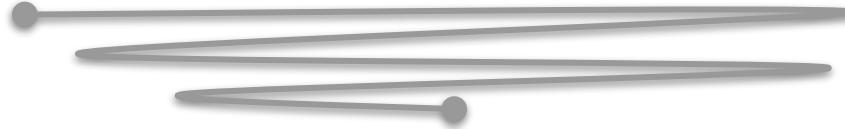
- ✓ completos
- ✓ oportunos (timeliness) y vigentes
- ✓ consistentes y correctos
- ✓ en cantidad adecuada
- ✓ disponibles/accesibles (ej. medicina), open data.
- ✓ seguros y privados (protección de datos personales)

Atributos de Calidad



Podemos definir los siguiente *Atributos de Calidad* ...

Atributos de Calidad



1. Completitud

- ✓ Están presentes todos los valores para representar la realidad
- ✓ Están presentes todas las instancias existentes en el mundo real

2. Relevancia

- ✓ Los datos son relevantes para representar la realidad

3. Vigencia

- ✓ Los datos se mantienen actualizados con la frecuencia adecuada

4. Disponibilidad

- ✓ Los datos están accesibles

5. Confiabilidad

- ✓ Se puede considerar que los datos representan información verídica

Atributos de Calidad



6. Consistencia

- ✓ No hay contradicciones entre distintos datos almacenados

7. Corrección

- ✓ Los datos representan la situación real

8. Seguridad/Privacidad

- ✓ Los datos cumplen con los requerimientos de privacidad adecuados de acuerdo a la reglamentación nacional-internacional / criterios éticos
- ✓ Los datos son sólo accesibles por los usuarios autorizados

Generalmente no nos vamos a encontrar con datos perfectos
Es necesario priorizar los atributos de calidad deseados

Cuán buenos son los datos (Calidad)

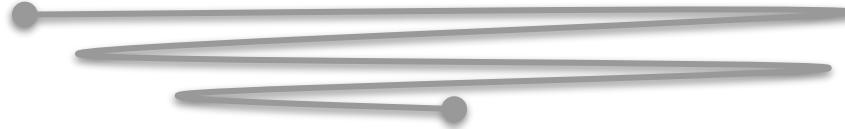
¿Son buenos los datos?



¿Cómo determinar cuán buena es la calidad de los datos? (5 min.)



¿Son buenos los datos?



¿Cómo determinar cuán buena es la calidad de los datos?

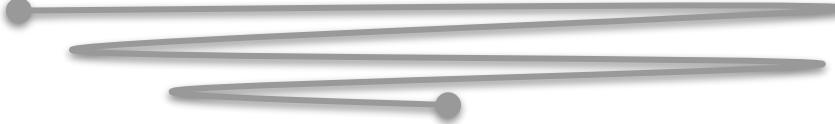
Tenemos que "entender" cuáles son los datos críticos y para ellos determinar los atributos de calidad de interés

Necesitamos seguir una metodología

1. Hacer relevamiento
2. Elaborar métricas de calidad
3. Recolectar valores de dichas métricas

Nos va a permitir cuantificar la calidad de los datos

1. Hacer relevamiento



Objetivo:

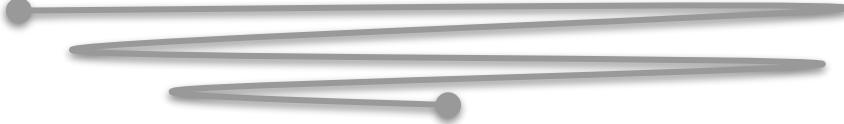
Determinar ...

- ✓ Datos críticos
- ✓ Ciclo de vida del dato
- ✓ Atributos de interés

Tareas a realizar:

- ✓ Identificar stakeholders (partes interesadas): CCC (creator, consumer, custodian)
- ✓ Conseguir el compromiso por parte del cliente - la organización
- ✓ Leer documentación sobre los sistemas, sobre el negocio, y estudiar modelos de datos
- ✓ Hacer cuestionarios tendientes a determinar cuáles son los datos críticos, cuál es el ciclo de vida del dato, cuáles son los atributos de interés y los problemas habituales
- ✓ Llevar adelante los cuestionarios con cada uno de los stakeholders identificados

2. Elaborar métricas de calidad



Para los datos críticos y los atributos de interés, armar métricas para cuantificar cuán grave es el problema

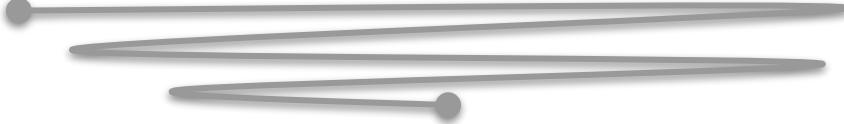
Possible Técnica: **Goal Question Metric** (Objetivo, Pregunta, Métrica). Conocida como **GQM**

Goal (Objetivo) Se define un objetivo

Question (Pregunta) Se plantea una pregunta (o más), cuya respuesta permitirá saber si se satisface el objetivo

Metric (Métrica) Se plantea una métricas (o más) -para cada una de las preguntas-, cuya ejecución permitirá responder las mismas

2. Elaborar métricas de calidad



Ejemplo (GQM. Goal Question Metric)

Queremos analizar la Completitud del dato Departamento asociado a los Empleados de la Compañía

Goal (Objetivo) El dato correspondiente al Departamento donde trabaja cada Empleado esté completo

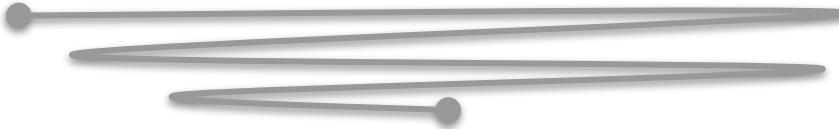
Question (Pregunta) ¿Cuál es la proporción de Empleados que tienen el dato correspondiente a Depto. vacío?

Metric (Métrica) M1: Proporción de registros con campo Departamento vacío en tabla Empleados, es decir,
Cantidad de registros de Empleado con campo id_Departamento vacío

Cantidad total de registros de Empleado

M2: Proporción de id de Departamento que tienen su nombre de departamento vacío
(en Tabla Departamento)

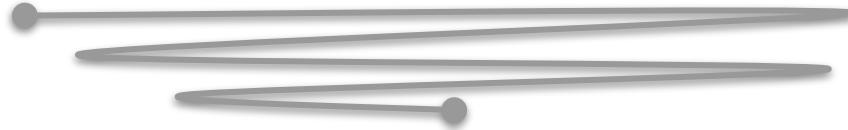
3. Recolectar valores de las métricas



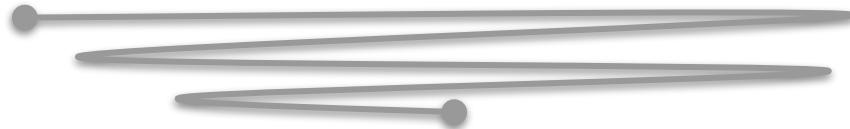
Los valores de las métricas se pueden obtener a través de consultas SQL

*Los objetivos en GQM también podrían establecerse sobre el modelo de datos
Las respuestas podrían ser si o no y como resultado podría concluirse la necesidad
de una revisión del modelo de datos*

Trabajo en equipo



Ejercicio - Consigna



- ✓ Conformar grupos de 3 integrante
- ✓ Descargar el registro de Datos de Dengue de 2020 correspondiente al Registro del Sistema Nacional de Vigilancia de la Salud 2.0 (<http://datos.salud.gob.ar/dataset/vigilancia-de-dengue-y-zika>)
-> Vigilancia de Dengue y Zika - 2020 (.xls)
- ✓ Aplicar la técnica GQM para comenzar a evaluar la calidad de datos de dicha fuente

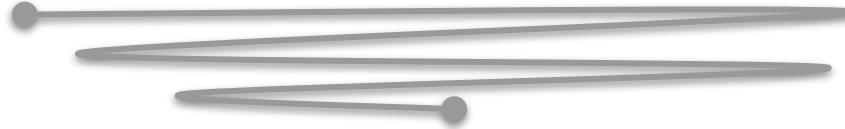
Diagnóstico y Mejoras

Diagnóstico



A partir de los resultados asociados a las ejecuciones de las métricas y del relevamiento, podemos determinar problemas existentes y sus causas.

Mejoras

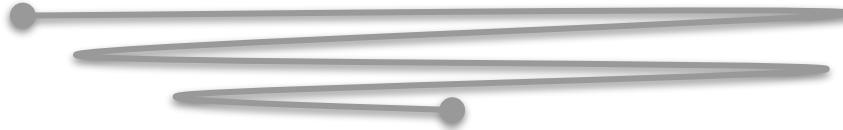


A partir del diagnóstico: Conclusiones y propuestas de mejora.

No sólo se trata de corregir, sino principalmente de prevenir. Posibles correcciones en:

- ✓ *Instancia*
- ✓ *Modelo de datos*
- ✓ *Procesos*
- ✓ *Capacitación*
- ✓ *Software*

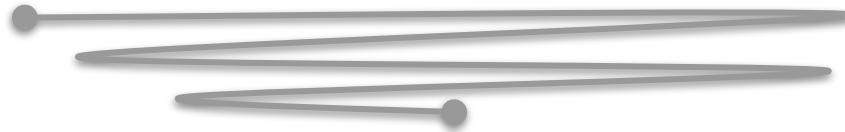
Herramientas de detección de problemas de Calidad de Datos



Existen muchas herramientas para automatizar la detección de:

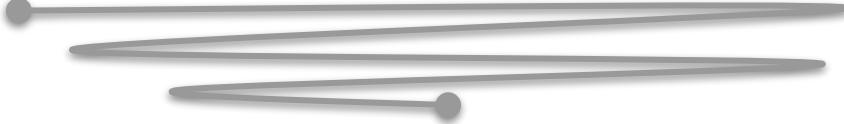
- ✓ Textos parecidos (soundex, keyboard distance, edit distance, ..., uso de diccionarios).
- ✓ Datos nulos
- ✓ Problemas de integridad referencial

Tareas para la próxima clase



- 1. Resolver la guía de ejercicios de “Calidad de Datos”*

Bibliografia



- ✓ English, 'Improving Data Warehouse and Business Information Quality', John Wiley & Sons (1999)
- ✓ Piattini, Calero, Genero (eds.): 'Information and Database Quality', Kluwer (2001)
Cap. 7: Bobrowski, Marré, Yankelevich, 'A NEAT Approach for Data Quality Assessment'
- ✓ Redman, 'Data Quality for the Information Age', Artech House (1996)
- ✓ Wang, Strong, Guarascio, 'Beyond Accuracy: What data quality means to data consumers', Total Data Quality Management Program (1996)



DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

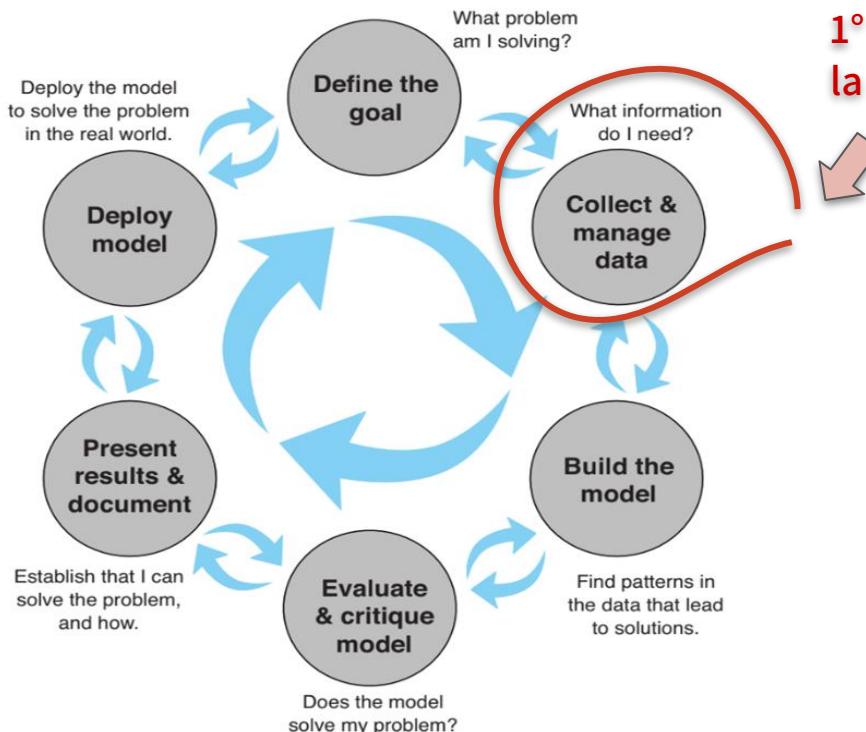
Laboratorio de datos

Visualización y Análisis

Exploratorio de Datos

Verano 2026

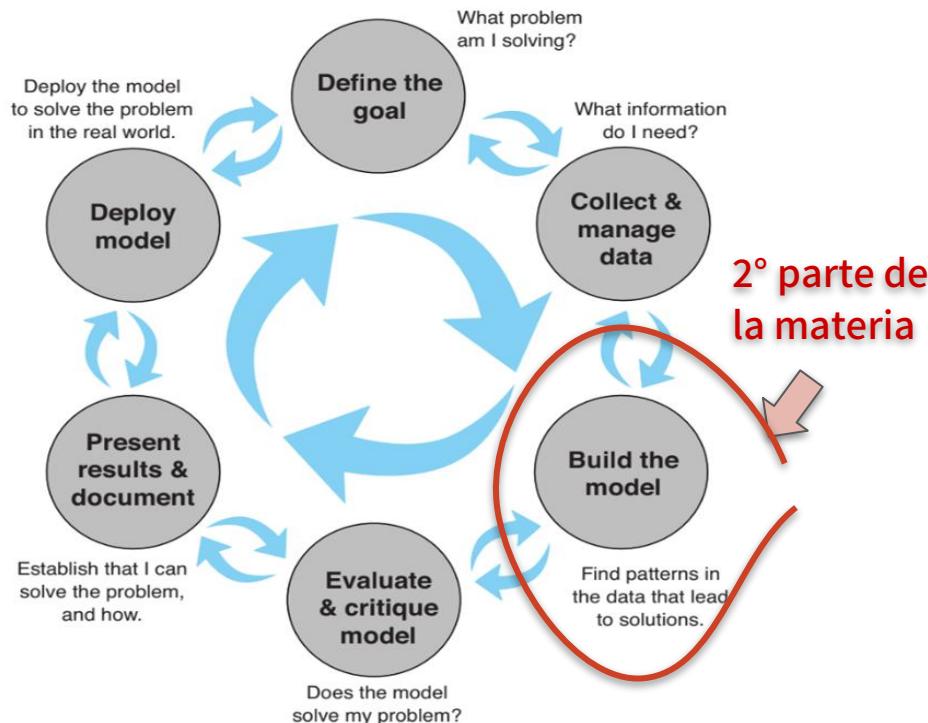
Hasta ahora vimos...



1º parte de
la materia

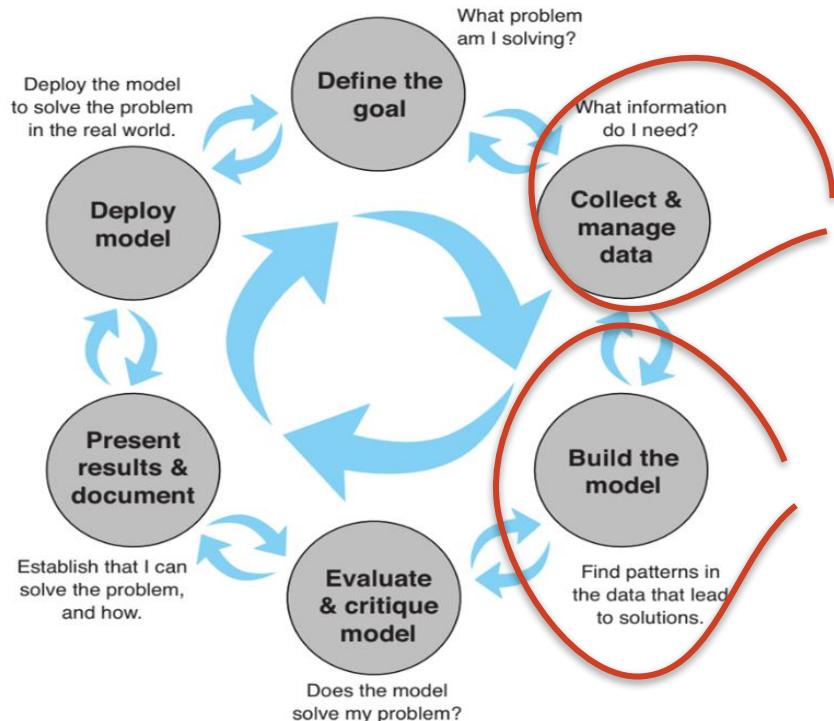
- Lenguaje de programación (Python)
- Modelado conceptual de los datos (DER)
- Representación de los datos (modelo relacional)
- Formas de consultar los datos (AR/SQL)
- Recomendaciones para el diseño (Normalización)
- Calidad de datos

Clase de hoy



- **Visualización y Exploración de los datos**

Clase de hoy



Administrar el **almacenamiento** de los datos:

- Qué datos me conviene guardar?
- Cuáles atributos definen a una entidad?
- Minimizar redundancias y anomalías
- Restricciones de integridad

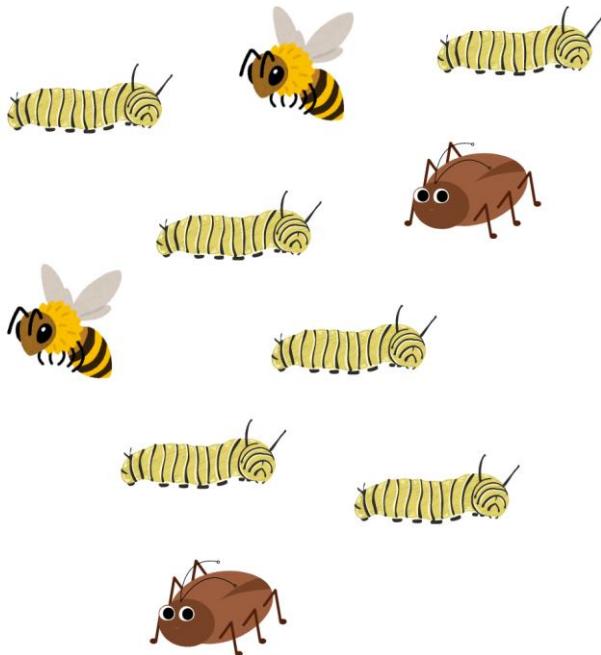
Encontrar **relaciones** entre los datos:

- Edad y altura crecen del mismo modo?
- Las causas de defunción tienen la misma frecuencia en todo el país?
- La frecuencia de las causas depende del rango etario?

¿Qué es el proceso de análisis de datos?

Proceso científico de **transformar** datos en información para tomar mejores decisiones

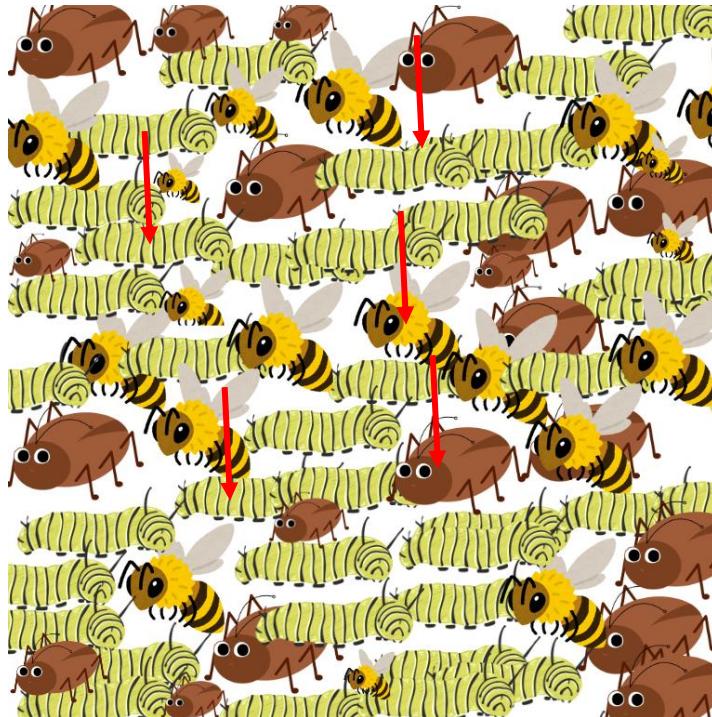
Alcance del análisis de datos



Queremos caracterizar la composición de esta **población** de insectos:

- 2 abejas
- 2 cucarachas
- 6 orugas

Alcance del análisis de datos



Recopilar datos de toda la población sería **costoso**

Vamos a recopilar datos de un **subconjunto** de la población → **Muestra (sample)**

Si asumimos que una **muestra** de datos es **representativa de la población**, podemos hacer **generalizaciones** sobre toda la población

Alcance del análisis de datos

Análisis descriptivo. Herramientas que describen lo que ha sucedido.

Ej. Queries, reportes, estadística descriptiva, visualización de datos. En general, resumen los datos existentes o los resultados de análisis predictivos o prescriptivos

Análisis predictivo. Técnicas que utilizan modelos matemáticos construidos a partir de datos pasados para predecir eventos futuros o comprender mejor las relaciones entre variables. Ej. Análisis de regresión, simulaciones computacionales

Análisis prescriptivo. Son modelos matemáticos o lógicos que sugieren una decisión o un curso de acción. Ej. modelos de optimización matemática, evaluación de escenarios, análisis de decisiones

Alcance del análisis de datos

- Análisis descriptivo
- Análisis predictivo
- Análisis prescriptivo



La **visualización** de datos es fundamental para el éxito de los tres tipos de análisis

Para qué visualizar

Al transformar los números y otras piezas de información en gráficos, el contenido se hace más fácil de **entender y usar**.

- resumir información
- encontrar tendencias o patrones
- detectar valores anómalos
- encontrar relaciones entre variables



This is news: <https://www.youtube.com/watch?v=SHb-3oIAFTs>

Para qué visualizar

Al transformar los números y otras piezas de información en gráficos, el contenido se hace más fácil de **entender y usar**.

- resumir información
- encontrar tendencias o patrones
- detectar valores anómalos
- encontrar relaciones entre variables

Para qué visualizar

Al transformar los números y otras piezas de información en gráficos, el contenido se hace más fácil de **entender y usar**.

- resumir información
- encontrar tendencias o patrones
- detectar valores anómalos
- encontrar relaciones entre variables
- hacerse preguntas, y en consecuencia elaborar hipótesis
- mostrar o reforzar una hipótesis
- mostrar resultados

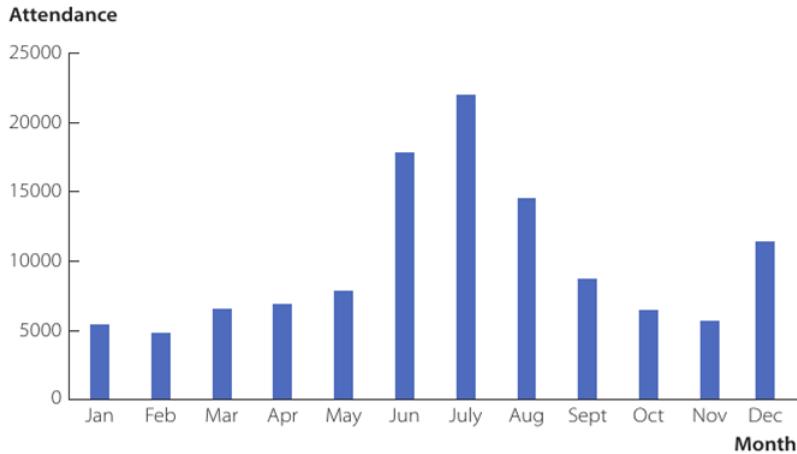
Ejemplo: encontrar tendencias o patrones

TABLE 1.1 Zoo Attendance Data						
Month	Jan	Feb	Mar	Apr	May	Jun
Attendance	5422	4878	6586	6943	7876	17843
Month	July	Aug	Sept	Oct	Nov	Dec
Attendance	21967	14542	8751	6454	5677	11422

¿Cómo varía la asistencia al zoológico a lo largo del tiempo?

FIGURE 1.1

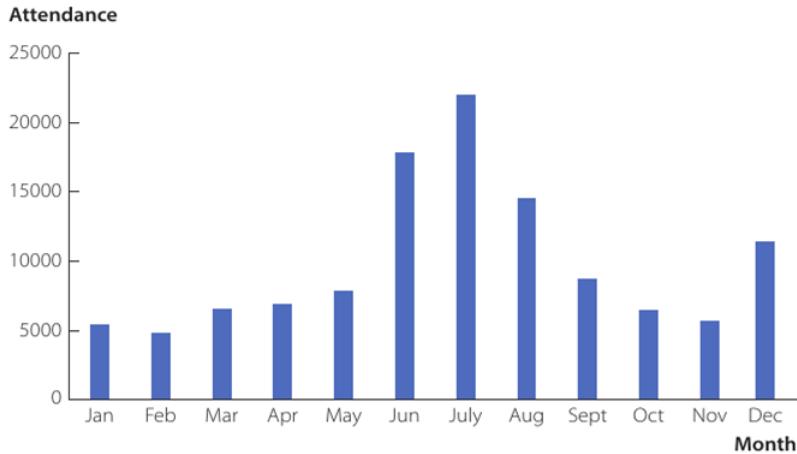
A Column Chart of Zoo Attendance by Month



- La asistencia aumenta en junio y julio, cuando los niños en edad escolar no asisten a la escuela (vacaciones de verano - hemisferio norte).
- Aumento gradual de la asistencia con la temperatura (desde febrero hasta mayo).

FIGURE 1.1

A Column Chart of Zoo Attendance by Month



- La asistencia en diciembre no sigue estos patrones → El zoo cuenta con el “Festival de las Luces” que se extiende desde finales de noviembre hasta principios de enero. También corresponde al periodo de vacaciones de invierno de los niños en edad escolar

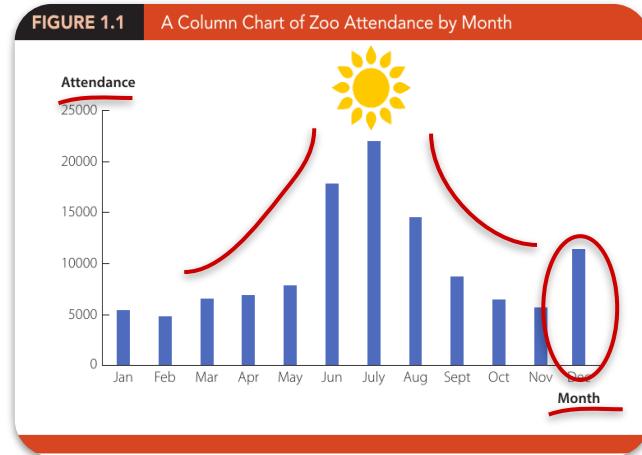
Visualización de datos según la finalidad

La **exploración** visual de datos es una parte crucial del análisis descriptivo.

La exploración de datos permite:

- Identificar patrones
- Reconocer anomalías e irregularidades
- Caracterizar la relación entre variables

La **visualización** nos da mayor capacidad para **detectar patrones, anomalías y relaciones** entre variables que al hacerlo mirando simplemente los datos crudos



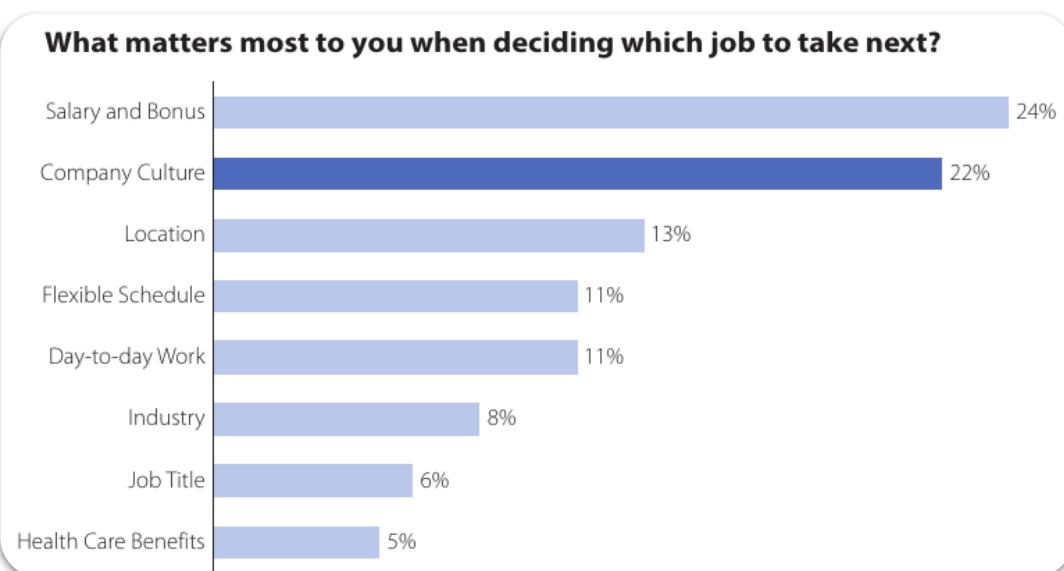
Mostrar o reforzar una hipótesis

Queremos mostrar que la “Cultura de la compañía” es uno de los factores más importantes a la hora de buscar trabajo

¿Resulta
obvio?

Factor	Porcentaje
Flexible Schedule	11,00%
Location	13,00%
Salary and Bonus	24,00%
Job Title	6,00%
Health Benefit Benefits	5,00%
Industry	8,00%
Company Culture	22,00%
Day-to-day Work	11,00%

Queremos mostrar que la “Cultura de la compañía” es uno de los factores más importantes a la hora de buscar trabajo

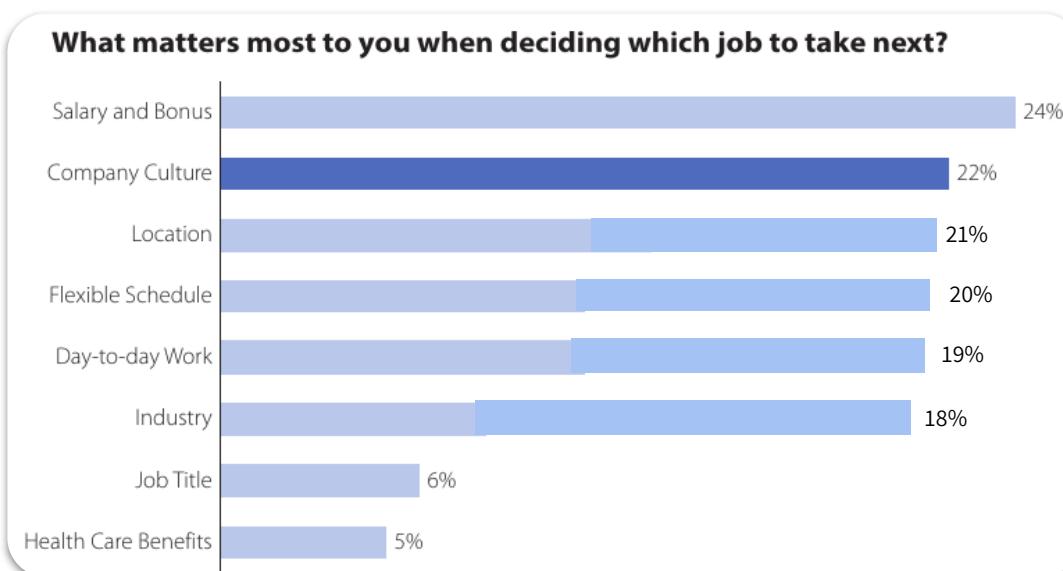


¿Resulta
obvio?

La **visualización** de datos es útil para **comunicar** a la audiencia y garantizar que **comprenda** y se centre en el mensaje deseado.

Visualización de datos según la finalidad: explicación

Queremos mostrar que la “Cultura de la compañía” es uno de los factores más importantes a la hora de buscar trabajo



¿Resulta
obvio?

En este caso, a pesar de que el ranking es el mismo y la conclusión sigue siendo estrictamente válida, estaríamos introduciendo un sesgo cuando en realidad tiene casi el mismo peso que otros 4 factores!

Tipos de datos

El tipo de gráfico a realizar depende de las características de los datos con los que se cuenta:

- Nominales - categorías
- Ordinales - escala discretas, rankings
- Cuantitativos - magnitudes
 - Proporciones - comparaciones de magnitudes
- Transversales
- Series Temporales

Tipos de datos: cuantitativos vs categóricos

¿Qué tipo de variable es *Length*?

Variable cuantitativa

Sex	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
F	16875	1275	375	3392017675	134660125	8618248	10942907	11
M	11	8875	3125	1353688625	3742134	231048425	5386405	9
F	15	1175	475	3216250775	13947954	735669525	956795625	10
M	13125	9625	25	1450076925	6973977	284912475	412485225	8
I	9875	675	25	846232575	409650275	17293195	2324659	5
I	13125	1	35	1971707725	681805475	453592	71724235	10
F	1325	1075	375	210069795	92135875	525883225	5556502	9
I	11375	9375	3125	151102835	66054335	3005047	52446575	8
I	1075	85	275	1033339275	45075705	242388225	29766975	7
M	1325	11	5125	236718325	907184	616601625	69456275	14
M	15375	13125	3875	3224755625	104042665	6690482	10489315	20
F	13125	10125	4	18653971	752679225	318931875	63786375	12

Variables cuantitativas:

- Permiten indicar una magnitud.
- Se les puede aplicar operaciones aritméticas (+, -, *, %, etc.).



Tipos de datos: cuantitativos vs categóricos

¿Qué tipo de variable es *Sex*?

Variable categórica

Sex	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
F	16875	1275	375	3392017675	134660125	8618248	10942907	11
M	11	8875	3125	1353688625	3742134	231048425	5386405	9
F	15	1175	475	3216250775	13947954	735669525	956795625	10
M	13125	9625	25	1450076925	6973977	284912475	412485225	8
I	9875	675	25	846232575	409650275	17293195	2324659	5
I	13125	1	35	1971707725	681805475	453592	71724235	10
F	1325	1075	375	210069795	92135875	525883225	5556502	9
I	11375	9375	3125	151102835	66054335	3005047	52446575	8
I	1075	85	275	1033339275	45075705	242388225	29766975	7
M	1325	11	5125	236718325	907184	616601625	69456275	14
M	15375	13125	3875	3224755625	104042665	6690482	10489315	20
F	13125	10125	4	18653971	752679225	318931875	63786375	12

Variables categóricas:

- Permiten identificar ítems similares mediante **etiquetas** o nombres.
- No se pueden realizar operaciones aritméticas con datos categóricos. Sin embargo, se pueden **sintetizar** los datos categóricos contando el número de observaciones o calculando las **proporciones** de las observaciones de cada categoría.

Tipos de datos: transversales vs series temporales

Los datos de esta tabla ¿Son transversales o corresponden a una serie temporal?

Sex	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
F	16875	1275	375	3392017675	134660125	8618248	10942907	11
M	11	8875	3125	1353688625	3742134	231048425	5386405	9
F	15	1175	475	3216250775	13947954	735669525	956795625	10
M	13125	9625	25	1450076925	6973977	284912475	412485225	8
I	9875	675	25	846232575	409650275	17293195	2324659	5
I	13125	1	35	1971707725	681805475	453592	71724235	10
F	1325	1075	375	210069795	92135875	525883225	5556502	9
I	11375	9375	3125	151102835	66054335	3005047	52446575	8
I	1075	85	275	1033339275	45075705	242388225	29766975	7
M	1325	11	5125	236718325	907184	616601625	69456275	14
M	15375	13125	3875	3224755625	104042665	6690482	10489315	20
F	13125	10125	4	18653971	752679225	318931875	63786375	12

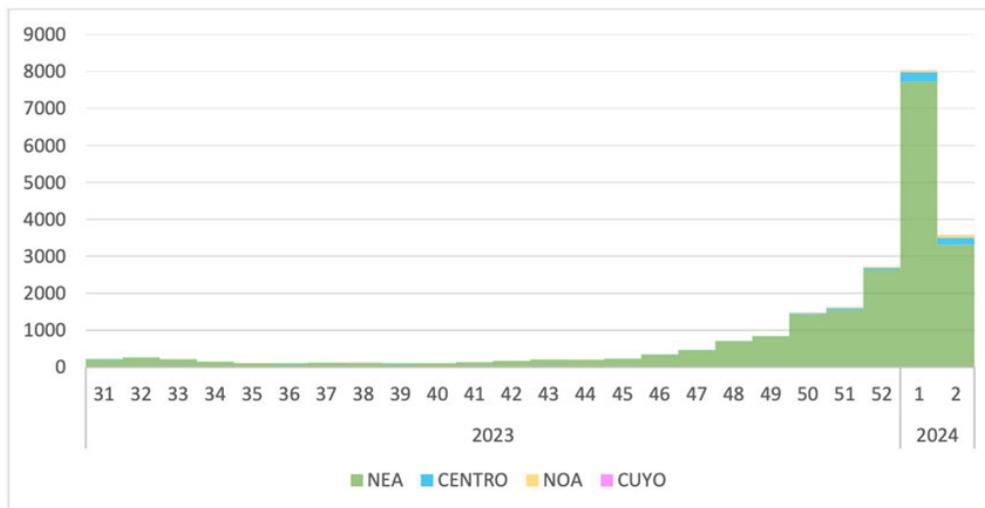
Datos transversales (Cross-Sectional Data)

Son datos que corresponden al mismo momento o aprox. del mismo tiempo.

Tipos de datos: transversales vs series temporales

Los datos de esta figura ¿Son transversales o corresponden a una serie temporal?

Gráfico 1. Casos de Dengue sin antecedentes de viaje por semana epidemiológica según región. SE 31/2023 a SE 2/2024 (n=22.466). Argentina.



Series de tiempo (Time Series Data)

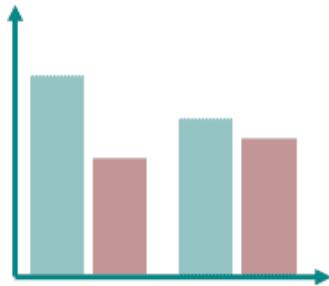
Son datos recopilados en varios puntos de tiempo (minutos, horas, días, meses, años, etc.).

Estos gráficos ayudan a los analistas a **comprender** lo que sucedió en **el pasado**, identificar **tendencias** a lo largo del tiempo y **proyectar** niveles **futuros** para la serie temporal.

Ej. Los gráficos de datos de series de tiempo se encuentran con frecuencia en publicaciones comerciales, económicas y científicas.

Visualización de datos: ejemplos comunes

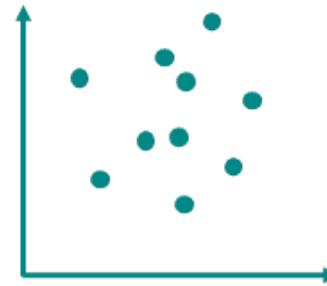
Bar chart



Histogram



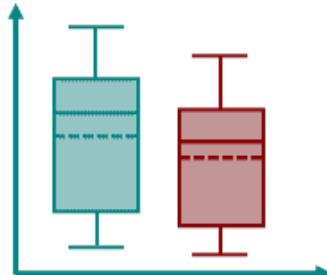
Scatter plot



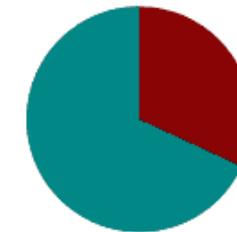
Line chart



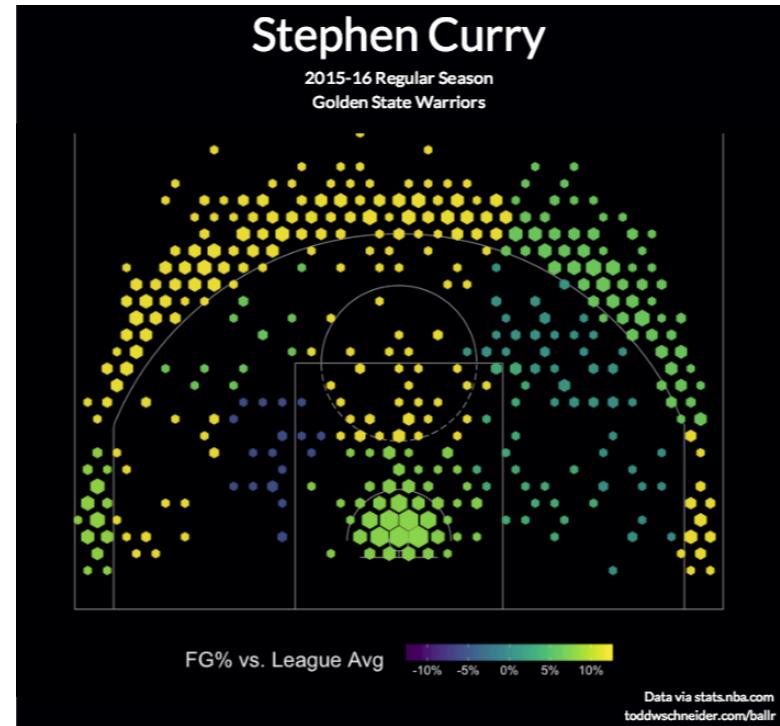
Boxplot



Pie chart



Visualización de datos: ejemplos menos comunes



Visualización de datos: ejemplos menos comunes

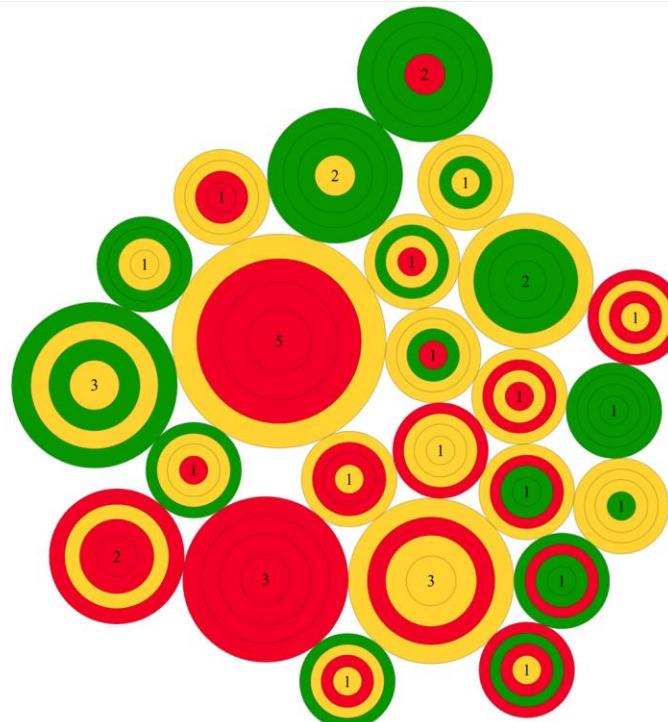
Medición de escuelas primarias

Desde adentro hacia afuera:

- Cognición y lenguaje
- Habilidades cotidianas
- Adaptación comportamental
- Soporte familiar

Cada diferente patrón de aros de colores es una combinación específica de cuatro valores

(López y Rosenfeld)



Selección del tipo de gráfico

1. Objetivo:
 - a. Explorar. Va a depender de la pregunta a responder y qué se espera de los datos
 - b. Explicar. Va a depender del mensaje que se quiere dar
2. Tipos de datos a graficar:
 - a. Variables **cuantitativas/cualitativas**
 - b. Datos **Transversales** vs. Series **Temporales**
3. Otros objetivos:
 - a. Ranking. Conocer el orden relativo de los elementos.
 - b. Correlación/Relación. Comprender cómo dos variables se relacionan entre sí. Ej. relación entre la temperatura mínima promedio y las nevadas anuales promedio en varias ciudades de Argentina.
 - c. Distribución. Saber cómo se dispersan los ítems. Ej. Cantidad de llamadas que recibe un call center a lo largo de un día.
 - d. Composición. Entender cómo se constituye una cierta entidad. Ej. Voto de las últimas elecciones.

Selección del tipo y formato de gráfico

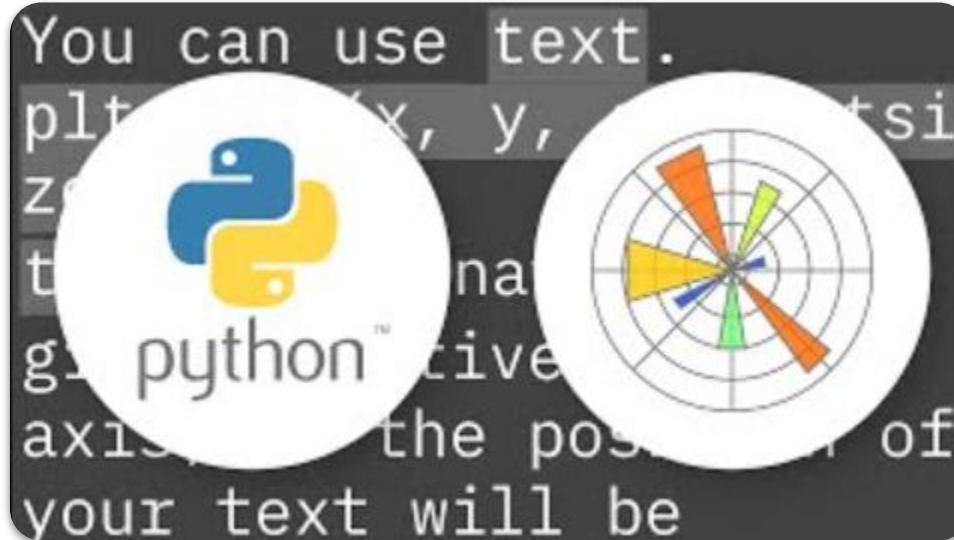
Objetivo

- Qué preguntas queremos responder
- Qué se espera de los datos
- Qué queremos mostrar
- Qué queremos enfatizar o resaltar
- Quién va a ver el gráfico
- Para qué se va a usar el gráfico
- Estilo que queremos utilizar

Tipos de datos a graficar

- Cuántas variables tenemos que representar
- Qué tipos variables tenemos que representar
- Tipos de interacción entre las variables
- Si los datos son transversales, temporales, u otros (espaciales...)

Nuestros primeros gráficos



Dataset de vinos



type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol	quality
white	8.4	0.17	0.31	6.7	0.038	29	132	3.1	0.32	10.6	7
white	6	0.18	0.31	1.4	0.036	14	75	3.34	0.58	11.1	8
white	8.6	0.36	0.26	11.1	0.03	43.5	171	3.03	0.49	12	5
white	6.9	0.4	0.17	12.9	0.033	59	186	3.08	0.49	9.4	5
red	6.8	0.785	0	2.4	0.104	14	30	3.52	0.55	10.7	6
red	10.8	0.29	0.42	1.6	0.084	19	27	3.28	0.73	11.9	6
white	7.1	0.21	0.32	2.2	0.037	28	141	3.2	0.57	10	7
white	6.1	0.17	0.21	1.9	0.09	44	130	3.07	0.41	9.7	5
white	9.2	0.28	0.46	3.2	0.058	39	133	3.14	0.58	9.5	5
red	11.5	0.59	0.59	2.6	0.087	13	49	3.18	0.65	11	6

Nuestros primeros gráficos

```
# Importar Bibliotecas
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt # Para graficar series multiples

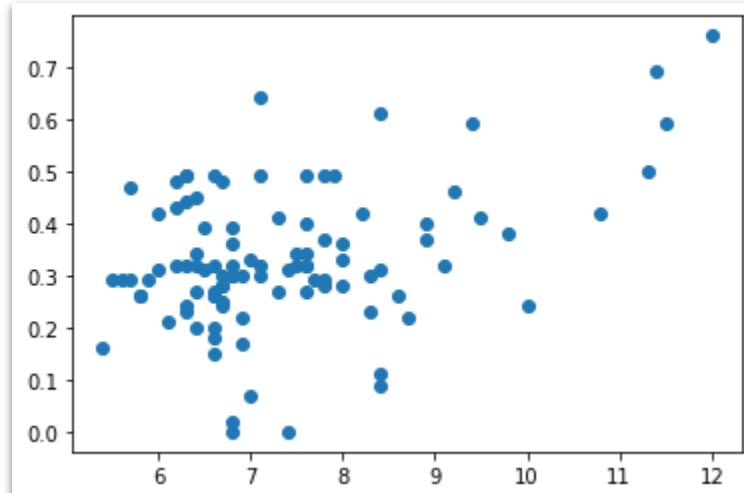
# Carpeta donde se encuentran los archivos a utilizar
carpeta = 'data/'
```

Leemos el *dataset*

```
wine = pd.read_csv(carpeta+"wine.csv", sep = ";")
# con sep indicamos que el separador es ;
```

Gráfico de dispersión/gráfico de puntos/scatter plot

```
# Genera el grafico que relaciona la acidez (no volatil) y el contenido de  
# acido citrico de cada vino  
plt.scatter(data = wine, x='fixed acidity', y='citric acid')  
# plt.scatter(wine['fixed acidity'], wine['citric acid']) #otra manera
```



Figuras de Matplotlib

```
fig, ax = plt.subplots() → devuelve una tupla!
```

In[7]: plt.subplots()

Out[7]: (<Figure size 864x576 with 1 Axes>, <Axes: >)

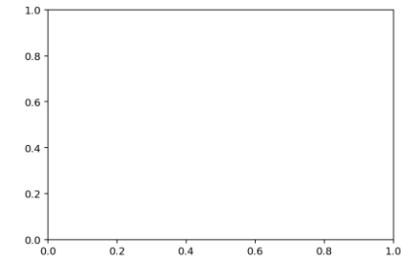


Fig: contenedor principal de todo el gráfico. Puede contener uno o varios ejes (en este caso contiene uno)

- Vamos a usar fig para hacer ajustes globales sobre toda la figura (tamaño, guardado de la imagen, etc)

Ax: Ejes, gráficos dentro de cada figura.

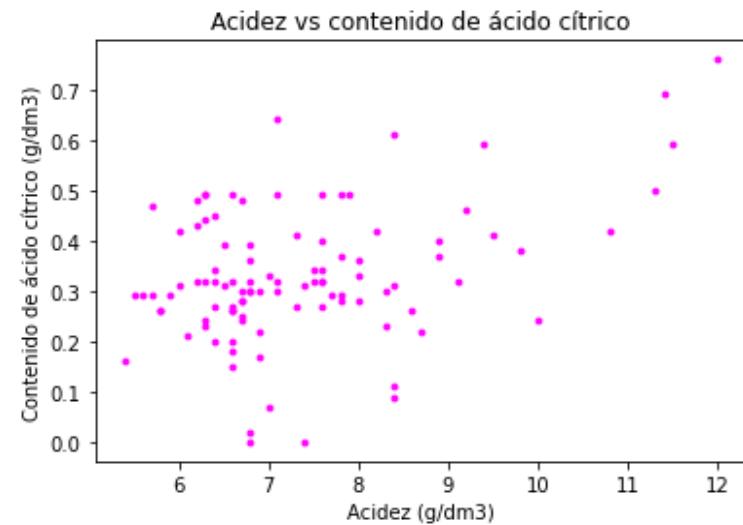
- Vamos a usar los atributos de ax para ajustar los elementos de cada gráfico, como etiquetas, líneas, colores, etc.

Gráfico de dispersión/gráfico de puntos/scatter plot

`fig, ax = plt.subplots()` → devuelve una tupla!

```
ax.scatter(data = wine,
           x='fixed acidity',
           y='citric acid',          # Tamaño de los puntos
           s=8,                      # Color de los puntos
           color='magenta')

ax.set_title('Acidez vs contenido de ácido cítrico') # Titulo del gráfico
ax.set_xlabel('Acidez (g/dm3)', fontsize='medium')    # Nombre eje X
ax.set_ylabel('Contenido de ácido cítrico (g/dm3)',   # Nombre eje Y
             fontsize='medium')
```



¿De qué tipo son las variables graficadas?

¿Cualitativas o cuantitativas?

Scatter Plot con color

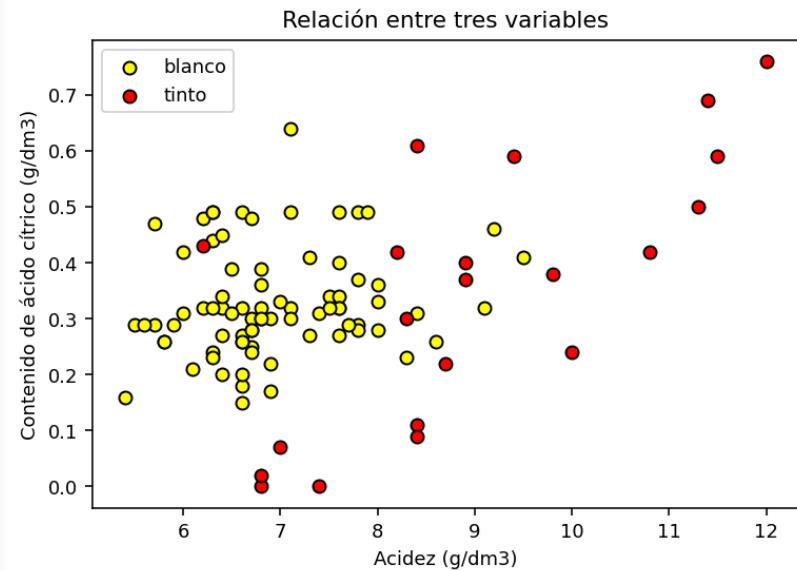
```
# Genera el grafico que relaciona tres variables en simultaneo
fig, ax = plt.subplots()

wine_blanco = wine[wine['type'] == 'white']
wine_tinto = wine[wine['type'] == 'red']

ax.scatter(data=wine_blanco, x='fixed acidity',
           y='citric acid', c='yellow', edgecolor='k', label='blanco')

ax.scatter(data=wine_tinto, x='fixed acidity',
           y='citric acid', c='red', edgecolor='k', label='tinto')

ax.set_title('Relación entre tres variables')
ax.set_xlabel('Acidez (g/dm3)', fontsize='medium')
ax.set_ylabel('Contenido de ácido cítrico (g/dm3)',
              fontsize='medium')
ax.legend()
del wine_blanco, wine_tinto
```



¿De qué tipo son las variables graficadas?
¿Cualitativas o cuantitativas?

Gráfico de globos/burbujas (*Bubble chart*)

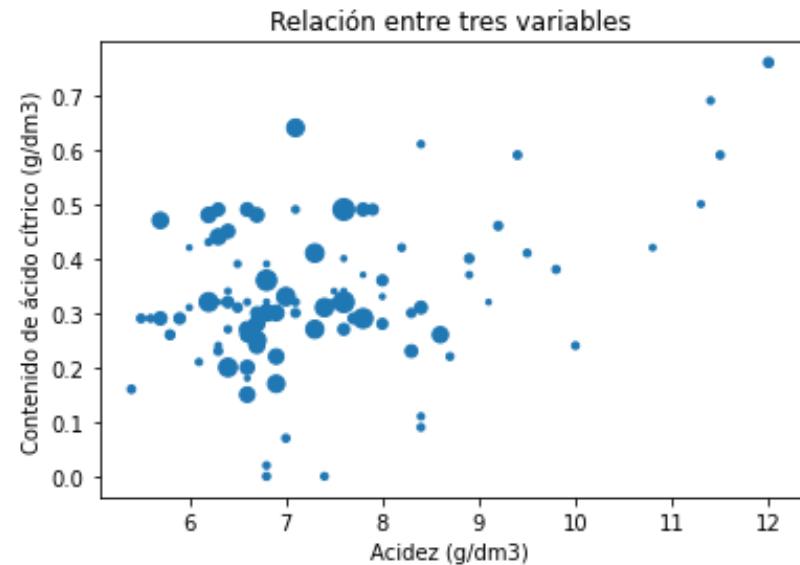
```
# Genera el grafico que relaciona tres variables en simultaneo
# (mejorando la informacion mostrada)
fig, ax = plt.subplots()

tamanoBurbuja = 5
# Cuanto queremos modificar el tamaño de cada burbuja

ax.scatter(data=wine, x='fixed acidity',
           y='citric acid', s=wine['residual sugar']*tamanoBurbuja)

ax.set_title('Relación entre tres variables')
ax.set_xlabel('Acidez (g/dm3)', fontsize='medium')
ax.set_ylabel('Contenido de ácido cítrico (g/dm3)',
              fontsize='medium')

# remueve la variable temporal tamanoBurbuja que ya no utilizaremos
del(tamanoBurbuja)
```



¿De qué tipo son las variables graficadas?
¿Cualitativas o cuantitativas?

Scatterplot - características

- Representa -al menos- dos variables, en los dos ejes.
 - Estas dos variables son de tipo numérico.
- Pueden sumarse otras variables mediante el uso del color, tipo o tamaño de los marcadores.
 - Estas otras variables pueden ser numéricas, categóricas, ordinales, etc.
- Puede ser útil para entender correlación entre variables.
- Puede ser útil para entender la distribución de valores para cada una de las variables.

Ejercicio

Con el dataset de vinos

¿Existe o no alguna **relación** entre el pH de los vinos (*pH*) y alguna de las otras variables? Mostrarlo gráficamente

Discutir con el resto de la clase:

- ¿Cuál fue su objetivo: Explorar, Explicar, Otro?
- ¿Qué tipos de variables estaban en juego?
- ¿Mejoró alguna característica del gráfico para cumplir con el objetivo?

Dataset CheetahRegion

Anio	regionEste	regionOeste	Ventas
1	59	28	87
2	57	33	90
3	68	42	110
4	91	54	145
5	109	61	170
6	96	58	154
7	110	67	177
8	72	103	175
9	63	120	183
10	65	130	195

Los datos de esta tabla ¿Son transversales o corresponden a una serie temporal?

Gráfico de líneas

```
# Genera el grafico de la serie temporal (grafico por defecto)
plt.scatter(data=cheetahRegion, x='Anio', y='Ventas')

# Genera el grafico de la serie temporal
#(mejorando la informacion mostrada)
fig, ax = plt.subplots()

ax.plot('Anio', 'Ventas', data=cheetahRegion, marker="o")

ax.set_title('Ventas de la compañía Cheetah Sports')
ax.set_xlabel('Año', fontsize='medium')
ax.set_ylabel('Ventas (millones de $)', fontsize='medium')
ax.set_xlim(0, 12)
ax.set_ylim(0, 250)
```

Tipo de gráfico muy utilizado para series temporales

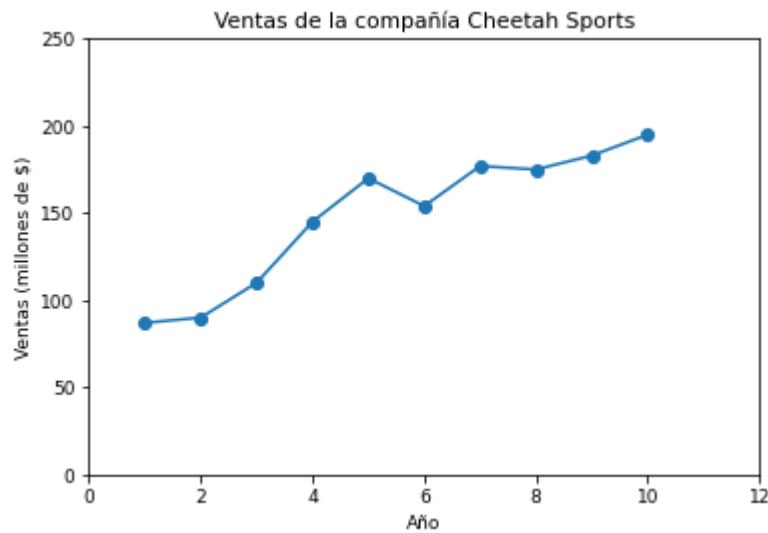


Gráfico de líneas

```
# Genera el grafico de ambas series temporales (mejorando la informacion mostrada)
fig, ax = plt.subplots()

# Grafica la serie regionEste
ax.plot('Anio', 'regionEste', data=cheetahRegion,
        marker='.',                      # Tipo de punto (redondo, triángulo, cuadrado, etc.)
        linestyle='-',                   # Tipo de linea (solida, punteada, etc.)
        linewidth=0.5,                  # Ancho de linea
        label='Región Este',            # Etiqueta que va a mostrar en la leyenda
        )

# Grafica la serie regionOeste
ax.plot('Anio', 'regionOeste', data=cheetahRegion,
        marker='.',                      # Tipo de punto (redondo, triángulo, cuadrado, etc.)
        linestyle='-',                   # Tipo de linea (solida, punteada, etc.)
        linewidth=0.5,                  # Ancho de linea
        label='Región Oeste'            # Etiqueta que va a mostrar en la leyenda
        )

# Agrega titulo, etiquetas a los ejes y limita el rango de valores de los ejes
ax.set_title('Ventas de la compañía Cheetah Sports según región')
ax.set_xlabel('Año')
ax.set_ylabel('Ventas (millones de $)')
ax.set_xlim(0,12)
ax.set_ylim(0,140)
# Muestra la leyenda
ax.legend()
```

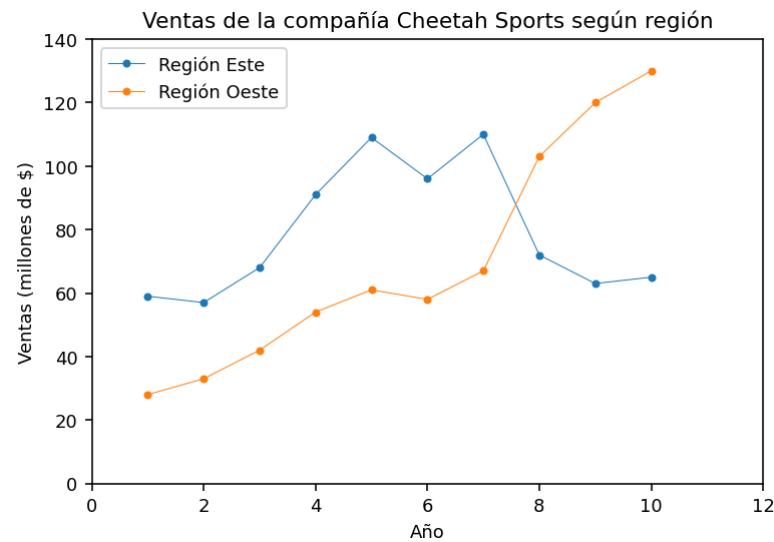


Gráfico de líneas - características

- Representa dos variables numéricas (x,y) donde para cada x hay como mucho un valor de y . La línea graficada interpola linealmente los valores de (x,y) existentes en la muestra.
 - La variable x a veces representa el paso del tiempo.
 - Idealmente debe haber bastantes valores distintos de x para que el gráfico cobre sentido.
- Se pueden representar más variables, mediante el uso de colores, o tipos de línea (llena/punteada/etc).
- Puede ser útil para entender la relación entre las variables.

Ejercicios

Considerar los siguientes datos correspondientes a los precios del biodiesel en distintos períodos en la Argentina
(se encuentran subidos en el campus)

Periodo	Precio
202312	686,986
202311	520
202310	434,006
202309	361,672
202308	346,000
202307	

1. Representarlos gráficamente
2. Analizar los resultados obtenidos
3. Discutir con el resto de la clase
 - a. ¿Cuál fue su objetivo: Explorar, Explicar, Otro?
 - b. ¿Qué tipos de variables estaban en juego?
 - c. ¿Qué tipo de gráfico decidió utilizar?
 - d. ¿Qué resultados obtuvo?
 - e. ¿Mejoró alguna característica del gráfico para cumplir con el objetivo?

Gráfico de barras

```
#### Genera el grafico de barras de las ventas mensuales
fig, ax = plt.subplots()

ax.bar(data=cheetahRegion, x='Anio', height='Ventas')

ax.set_title('Ventas de la compañía Cheetah Sports')
ax.set_xlabel('Año', fontsize='medium')
ax.set_ylabel('Ventas (millones de $)', fontsize='medium')
ax.set_xlim(0, 11)
ax.set_ylim(0, 250)

ax.set_xticks(range(1,11,1))           # Muestra todos los ticks del eje x
ax.set_yticks([])                     # Remueve los ticks del eje y
ax.bar_label(ax.containers[0], fontsize=8) # Agrega la etiqueta a cada barra
```

¿Por qué podemos usar un BarPlot para este dataset?

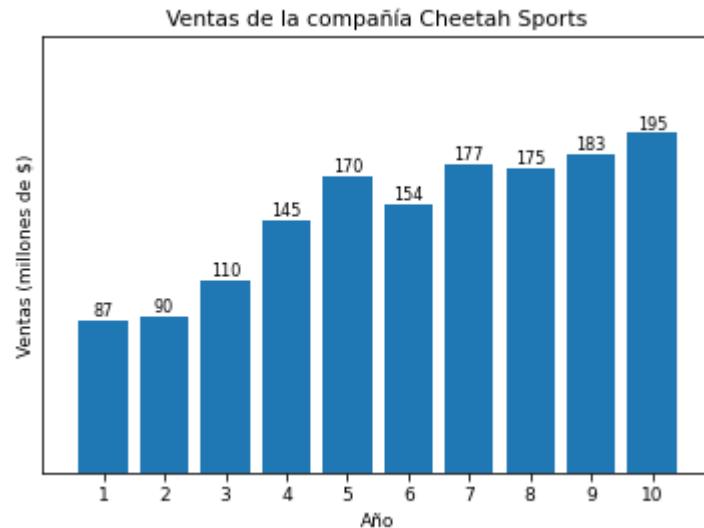


Gráfico de barras agrupadas

```
#### Genera el grafico de barras de ambas series temporales
fig, ax = plt.subplots()

x = cheetahRegion['Año']
east = cheetahRegion['regionEste']
west = cheetahRegion['regionOeste']

width = 0.4

ax.bar(x - width/2, east, width=width, label='Region Este')
ax.bar(x + width/2, west, width=width, label='Region Oeste')

ax.set_title('Ventas de la compañía Cheetah Sports según región')
ax.set_xlabel('Año')
ax.set_xticks([1,2,3,4,5,6,7,8,9,10])
ax.set_ylabel('Ventas (millones de $)')
ax.legend()

plt.show()
```

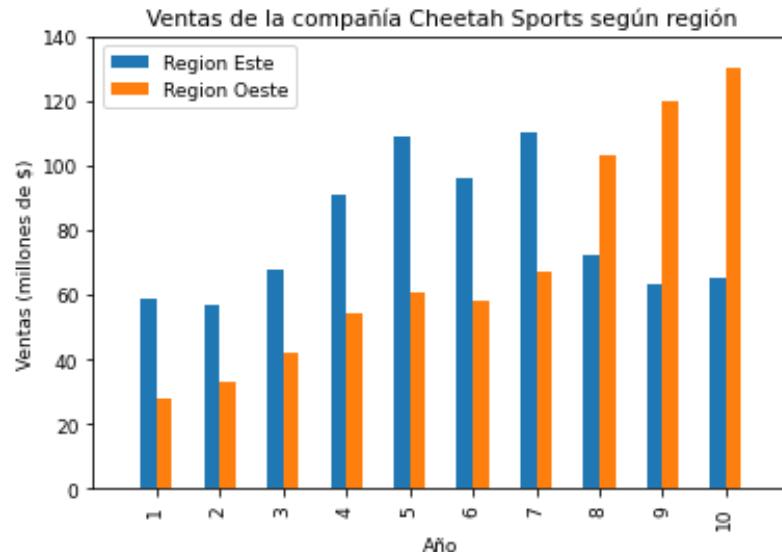


Gráfico de barras apiladas

```
# Genera el grafico de barras apiladas de ambas series temporales
fig, ax = plt.subplots()

# Grafica la serie regionEste
ax.bar(cheetahRegion['Anio'], cheetahRegion['regionEste'] ,
       label='Region Este', color = "#4A4063")
# Grafica la serie regionOeste
ax.bar(cheetahRegion['Anio'], cheetahRegion['regionOeste'],
       bottom=cheetahRegion['regionEste'], label='Region Oeste',
       color = '#BFACC8')

# Agrega titulo, etiquetas a los ejes y limita el rango de valores
# de los ejes
ax.set_title('Ventas de la compañía Cheetah Sports según región')
ax.set_xlabel('Año')
ax.set_ylabel('Ventas (millones de $)')
ax.set_xlim(0,10.9)
ax.set_ylim(0,250)
ax.set_xticks(range(1,11,1))      # Muestra todos los ticks del eje x
plt.legend()                      # Muestra la leyenda
```

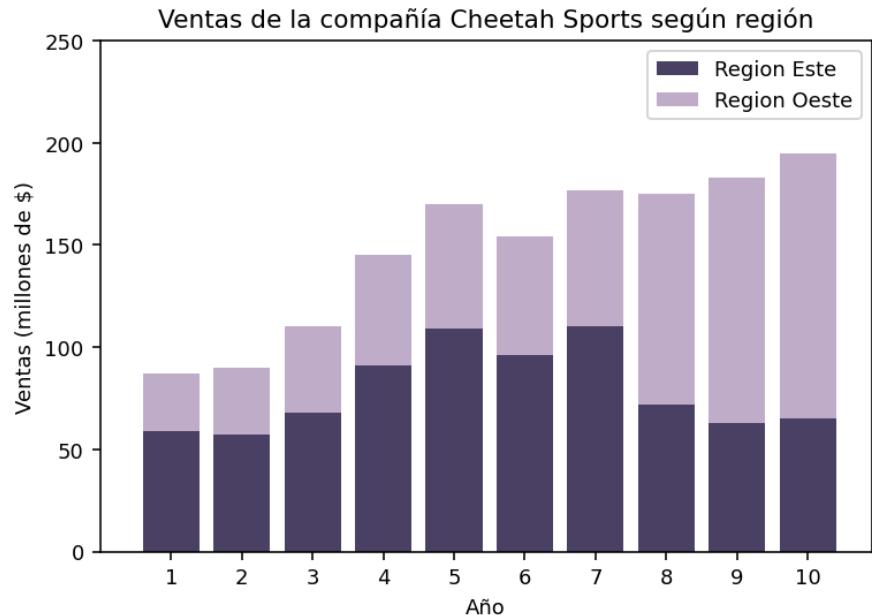


Gráfico de barras - características

- Representa dos variables. Una numérica (altura de las barras) y una categórica u ordinal (una barra para cada categoría).
 - A veces corresponden a una secuencia temporal
 - La altura de las barras suele representar una cantidad, (ej. No debería representar un año)
- Pueden representar además otra variable categórica, mediante el uso de colores.
 - Esta nueva variable debería tomar pocos valores (2 o 3) para que el gráfico no se vuelva incomprendible.
 - Sirve para comparar distribuciones.

Gráfico de torta/pie chart

Antes de graficar...

```
# Contamos cuantos vinos de cada tipo hay en el dataset  
wine['type'].value_counts()
```

Out[5]:

```
type  
white    79  
red      21  
Name: count, dtype: int64
```



type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol	quality
white	8.4	0.17	0.31	6.7	0.038	29	132	3.1	0.32	10.6	7
white	6	0.18	0.31	1.4	0.036	14	75	3.34	0.58	11.1	8
white	8.6	0.36	0.26	11.1	0.03	43.5	171	3.03	0.49	12	5
white	6.9	0.4	0.17	12.9	0.033	59	186	3.08	0.49	9.4	5
red	6.8	0.785	0	2.4	0.104	14	30	3.52	0.55	10.7	6
red	10.0	0.20	0.40	1.6	0.084	10	97	3.00	0.70	11.0	6

Gráfico de torta

```
# Transformamos la salida de value_counts en un dataframe
conteos = pd.DataFrame(wine['type'].value_counts()).reset_index()
conteos = conteos.rename(columns={'index': 'type', 0: 'count'})

# Genera el grafico de barras torta (mejorando la informacion mostrada)
fig, ax = plt.subplots()

ax.pie(data=conteos,
       x='count',
       labels='type',           # Etiquetas
       autopct='%1.2f%%',      # porcentajes
       colors=['gold',
               'purple'],
       shadow = True,
       explode = (0.1,0)        # separa las slices del pie plot
)
```

¿De qué tipo son las variables graficadas?
¿Cualitativas o cuantitativas?

Distribución de Tipos de Vino

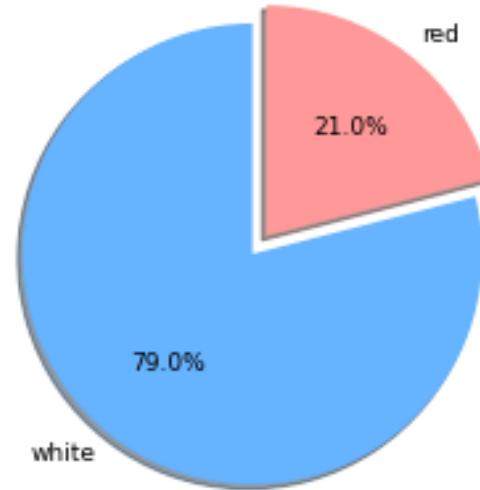


Gráfico de torta

- Un gráfico de torta (*pie plot/pie chart*) es un círculo dividido en porciones que representan partes de un conjunto
- Los humanos **no somos buenos para leer ángulos**

Hagamos un experimento para comprobarlo:

1. Tratemos de identificar al grupo más grande
2. Tratemos de ordenar a los grupos del mayoritario al minoritario

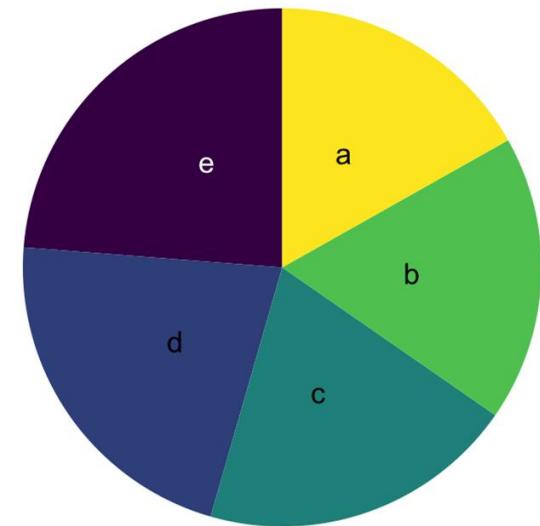


Gráfico de torta

Hagamos un experimento para comprobarlo:

1. Tratemos de identificar al grupo más grande
2. Tratemos de ordenar a los grupos del mayoritario al minoritario

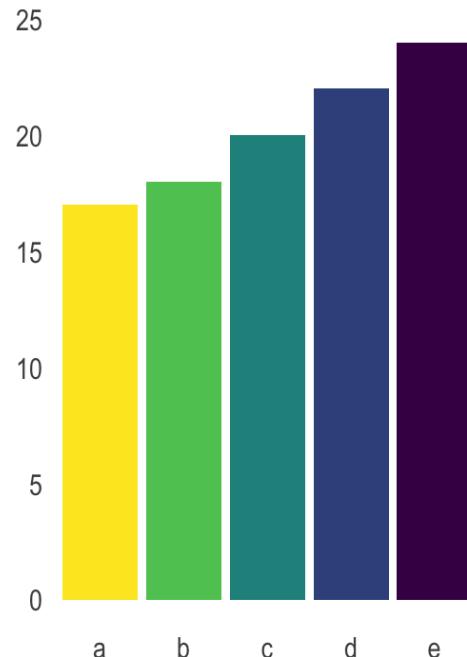


Gráfico de torta

Los dos gráficos fueron construidos usando **los mismos datos**, sin embargo encontrar al grupo mayoritario y ordenar los grupos resulta **mucho más sencillo** observando el **gráfico de barras**.

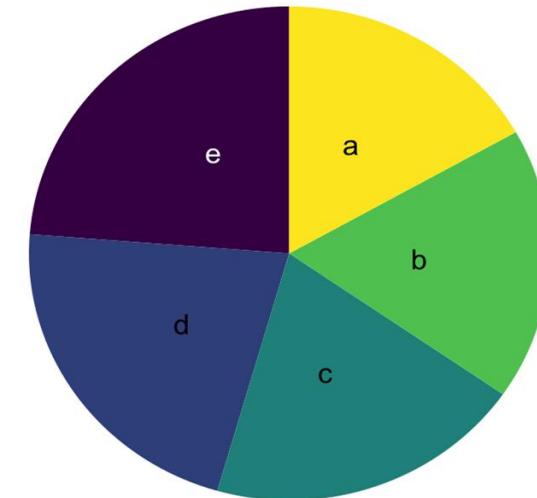
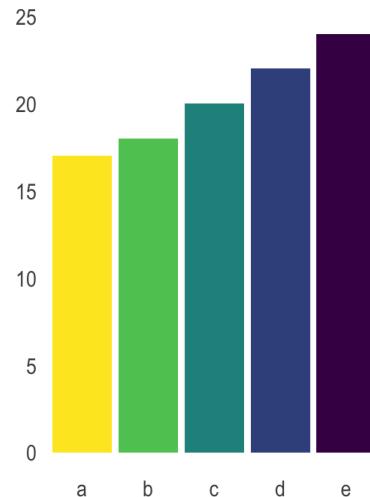


Gráfico de torta - características

- Representa proporciones o distribuciones porcentuales de una variable numérica (en el tamaño de las porciones) respecto de una variable categórica (color)
- Puede ser útil para entender la distribución entre las categorías.
- Puede no ser útil, si hay muchas categorías, o si hay proporciones similares.

Ejercicios

Considerar los siguientes datos a poseedores de teléfonos del archivo telefonosInteligentes.csv

RangoEtario	Telefono_Inteligente (%)	Telefono_NoInteligente (%)	SinTelefono (%)
18-24	49	46	5
25-34	58	35	7
35-44	44	45	11
45-54	28	58	14
55-64	22	59	19
65+	11	45	44

1. Generar un gráfico para representarlos gráficamente
2. Analizar los resultados obtenidos
3. Discutir con el resto de la clase
 - a. ¿Cuál fue su objetivo: Explorar, Explicar, Otro?
 - b. ¿Qué tipos de variables estaban en juego?
 - c. ¿Qué tipo de gráfico decidió utilizar?
 - d. ¿Qué resultados obtuvo?
 - e. ¿Mejoró alguna característica del gráfico para cumplir con el objetivo?
 - f. Responder Verdadero o Falso y justificar visualmente. **“Es más probable que las personas mayores posean un teléfono inteligente a que las personas más jóvenes posean uno inteligente.”**

Buenas Prácticas

- Elegir el tipo de gráfico adecuado
- Usar colores con sentido
- Usar pocos colores y diferenciables
- Hacer gráficos que aporten información útil
- No agregar mucha información en un solo gráfico
- Priorizar legibilidad frente a estética

Distribución de los Datos

Visualización - Distribución de los Datos

Edades de los atletas olímpicos de USA en los últimos 4 torneos de verano
(Bordes: izquierdo -> mínimo; derecho -> máximo)

FIGURE 5.1 Overlapping Range Bar Chart For Ages Of U.S. Olympic Gymnasts

Male Female Both

Age Range of U.S. Gymnasts in the Four Most Recent Summer Games

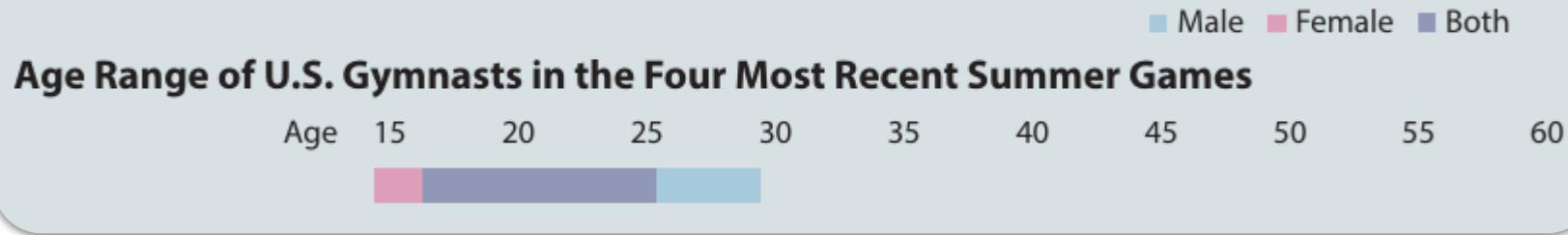


1. ¿En qué rango de edades hubo participación femenina? ¿Masculina? ¿Ambos?
2. ¿Cuántos individuos femeninos de 20 años de edad participaron? ¿Y masculinos?
3. ¿Para qué edades podemos afirmar que hubo participación?

Visualización - Distribución de los Datos

Edades de los atletas olímpicos de USA en los últimos 4 torneos de verano
(Bordes: izquierdo -> mínimo; derecho -> máximo)

FIGURE 5.1 Overlapping Range Bar Chart For Ages Of U.S. Olympic Gymnasts



¡No hay información sobre cómo se distribuyen lxs gimnastxs femeninos y masculinos en sus respectivos rangos!

Esta forma de visualizar datos:

- No es intuitiva
- Aumenta la carga cognitiva de la audiencia

Visualización - Distribución de los Datos

Edades de los atletas olímpicos de USA en los últimos 4 torneos de verano

FIGURE 5.2 Frequency Polygon for U.S. Olympic Gymnasts

Age Distribution of U.S. Olympic Gymnasts

Count of Gymnasts

40

35

30

25

20

15

10

5

0

< Female son más jóvenes

Males son más grandes >

Edades más comunes

Male — Female

Age (years)

1. ¿En qué rango de edad hubo participación femenina? ¿masculina? ¿ambos?
2. ¿Qué cantidad de individuos de sexo femenino, de 20 años, participó? ¿y de sexo masculino? ¿y de ambos?
3. ¿Para qué edades podemos afirmar que hubo participación?

¡Muestra información sobre la distribución por edades de gimnastxs masculinos y femeninos!

Visualización - Distribución de los Datos

El rol del **análisis descriptivo** es analizar y visualizar datos para comprender mejor la variación y su impacto

Distribución de frecuencia de una variable:

Describe qué valores se observaron y con qué frecuencia esos valores aparecen en dichos datos

Distribución de frecuencia



Variable categórica

(etiquetas que no pueden manipularse aritméticamente)

Variable cuantitativa

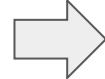
(valores numéricos que pueden manipularse aritméticamente)

Visualización - Distribución de los Datos - Categóricos

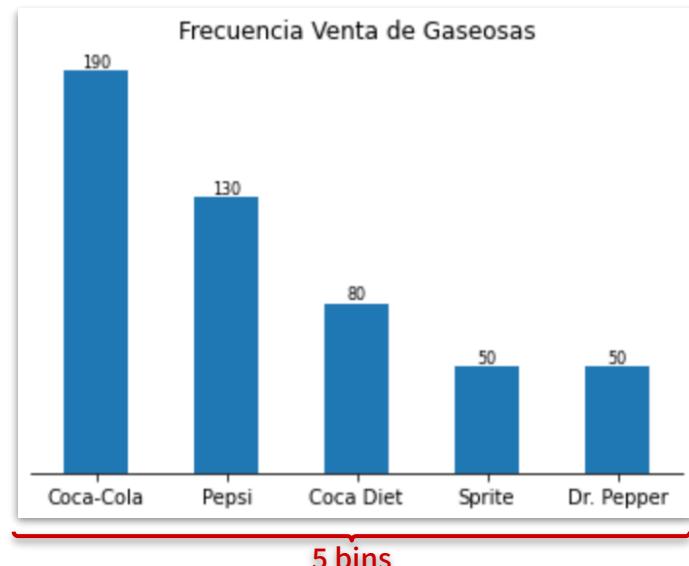
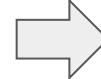
Una distribución de frecuencia (de variable categóricas) es un resumen de datos que muestra el **número** (frecuencia) de observaciones en **cada una de las clases** (no superpuestas), denominadas **bins**.

Compras_gaseosas
Coca-Cola
Sprite
Pepsi
Pepsi
Dr. Pepper
Coca Diet
Coca-Cola
C

500
compras



Compras_gaseosas	
Coca-Cola	190
Pepsi	130
Coca Diet	80
Dr. Pepper	50
Sprite	50



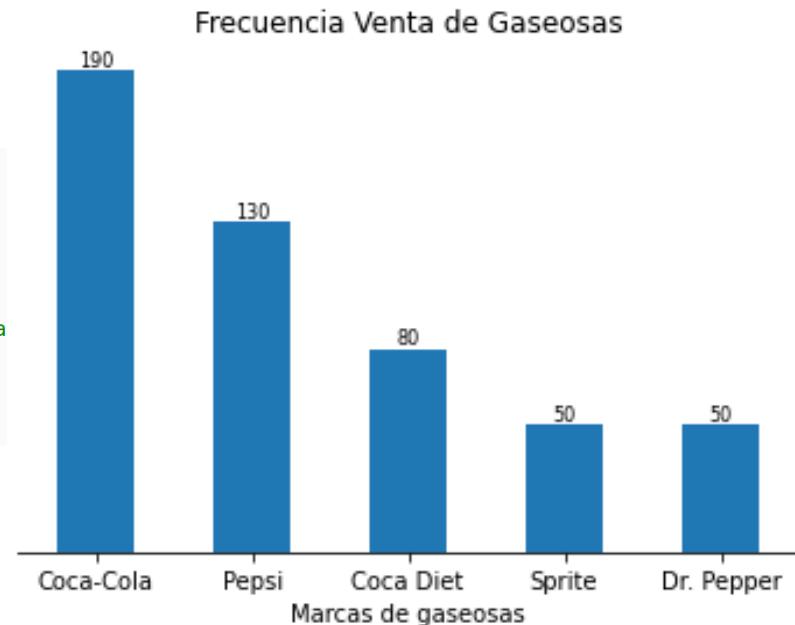
La distribución de frecuencia resume la información sobre la popularidad de las cinco gaseosas

Gráfico - Distribución de Datos categóricos

```
fig, ax = plt.subplots()
gaseosas['Compras_gaseosas'].value_counts().plot.bar(ax = ax)

ax.set_title('Frecuencia Venta de Gaseosas')
ax.set_xlabel('Marcas de gaseosas')
ax.set_yticks([])
ax.bar_label(ax.containers[0], fontsize=8)      # Remueve los ticks del eje y
                                                # Agrega la etiqueta a cada barra
ax.tick_params(axis='x', labelrotation=0)        # Rota las etiquetas del eje x

# Eliminar lineas del recuadro
ax.spines[['right', 'top', 'left']].set_visible(False)
```



Distribución de Datos categóricos

Distribución de **Frecuencia Absoluta** muestra la **cantidad** (recuento) de artículos en cada uno de los bins. A veces nos interesa la **proporción o porcentaje** de artículos en cada contenedor.

Frecuencia relativa de un bin. Fracción o proporción de ítems que pertenecen a ese bin (clase).

Frecuencia relativa de un bin =

$$\frac{\text{Frecuencia absoluta del bin}}{n}$$

donde n es la cantidad total de observaciones

Frecuencia porcentual de un bin. Frecuencia relativa multiplicada por 100.

Visualización - Distribución de los Datos - Categóricos

Compras_gaseosas
Coca-Cola
Sprite
Pepsi
Pepsi
Dr. Pepper
Coca Diet
Coca-Cola
C

Frecuencia relativa

Compras_gaseosas	
Coca-Cola	0.38
Pepsi	0.26
Coca Diet	0.16
Dr. Pepper	0.10
Sprite	0.10

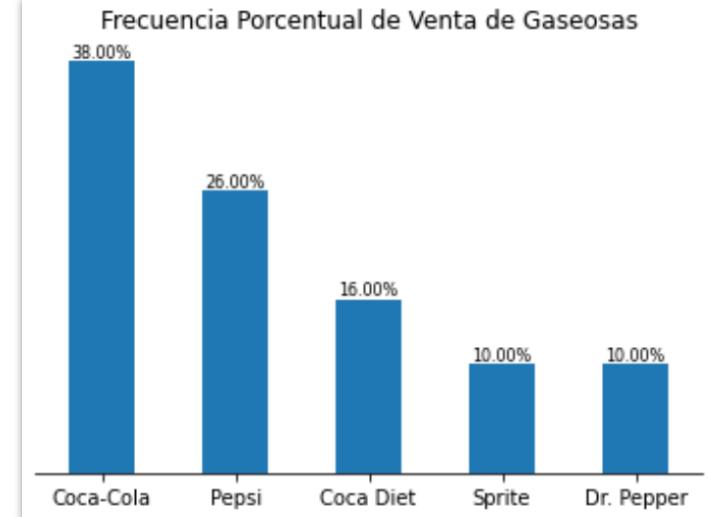


Gráfico - Distribución de Datos categóricos

```
# Tabla de frecuencias relativas
```

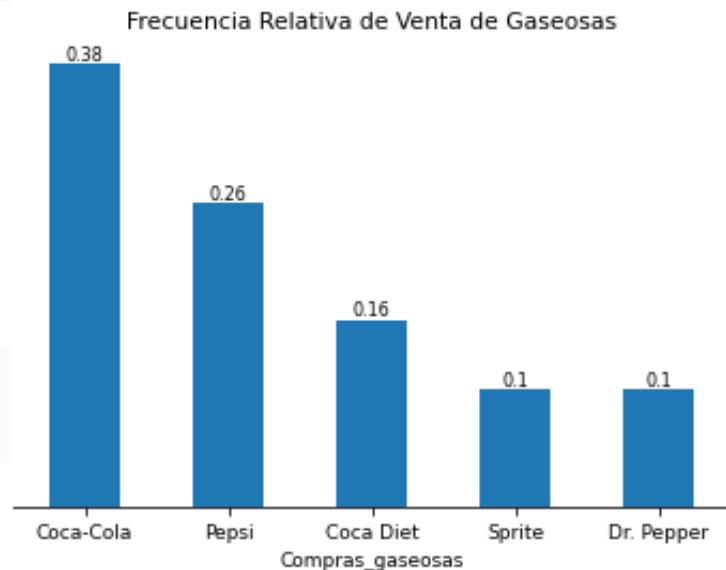
```
gaseosas['Compras_gaseosas'].value_counts(normalize=True)
```

```
Out[138]:
```

```
Compras_gaseosas
Coca-Cola      0.38
Pepsi          0.26
Coca Diet      0.16
Sprite         0.10
Dr. Pepper     0.10
Name: proportion, dtype: float64
```

```
fig, ax = plt.subplots()
```

```
ax = gaseosas['Compras_gaseosas'].value_counts(normalize=True).plot.bar()
```



Distribución de Datos categóricos

- Distribución de frecuencia relativa (o porcentual) se puede usar para estimar las probabilidades relativas de diferentes valores para una variable (aleatoria)
- Ej. Un puesto de comida ha determinado que adquirirá un total de 12.000 gaseosas para un próximo concierto. ¿Cómo dividirían este total entre los distintos tipos de gaseosas individuales?

Si los datos analizados (muestra) son representativos de la población de clientes del puesto de comida, se puede usar esta información para determinar los volúmenes apropiados de cada tipo de refresco.

Por ejemplo, los datos sugieren que se debería adquirir $12.000 * 0,38 = 4.560$ Coca-Colas.

Compras_gaseosas	
Coca-Cola	0.38
Pepsi	0.26
Coca Diet	0.16
Dr. Pepper	0.10
Sprite	0.10

Distribución de Datos continuos

¿Cuál es la dificultad para obtener la distribución de frecuencia en variables continuas?

Por ejemplo la distribución para la variable **Peso** (continua, asumir 3 decimales)

Peso (kg)
59,032
78,127
95,900

Distribución de Datos continuos

¿Cuál es la dificultad para obtener la distribución de frecuencia en variables continuas?

Por ejemplo la distribución para la variable **Peso** (continua, asumir 3 decimales)

Solución:

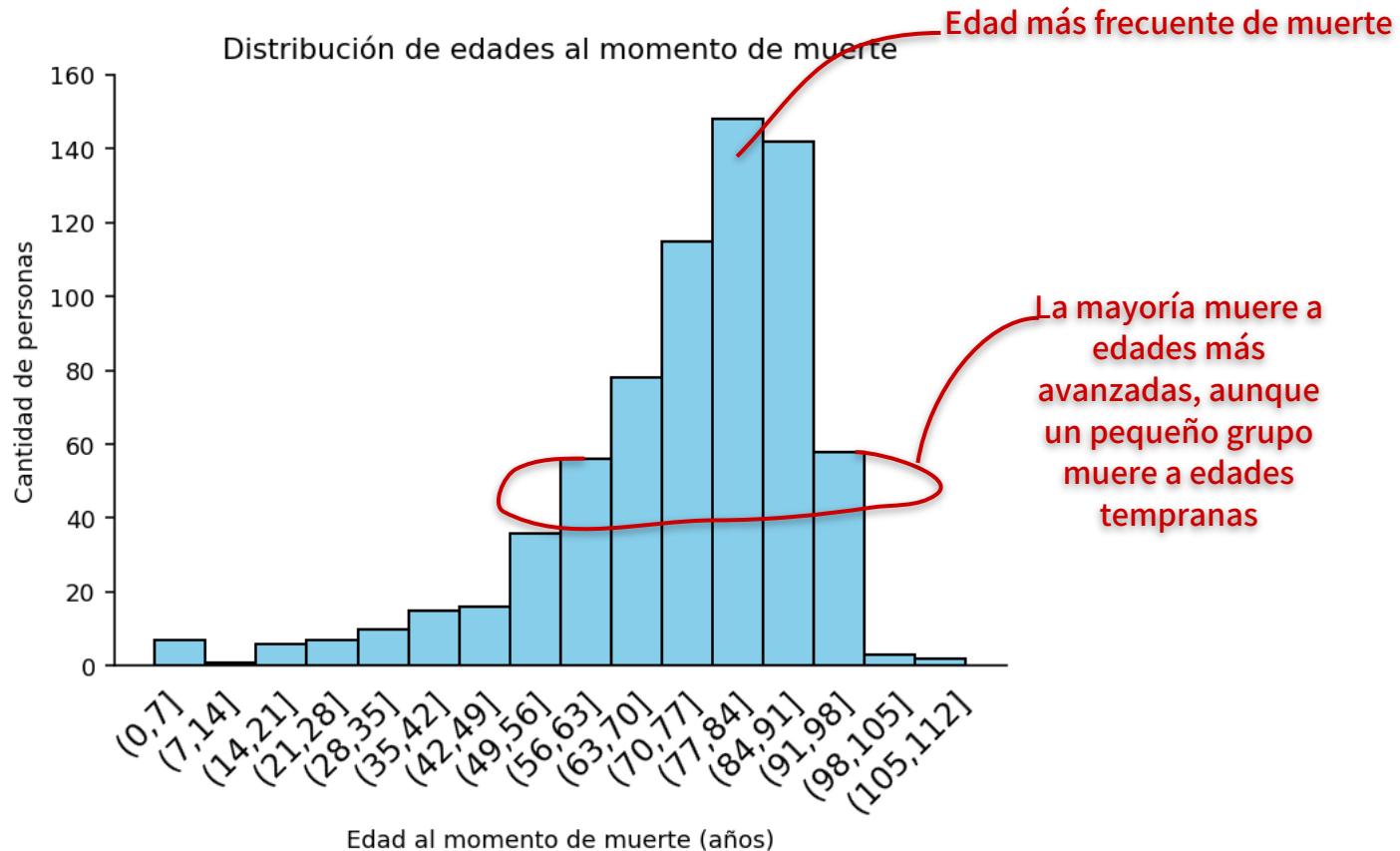
- Cada bin ahora contiene un **rango de valores** (en vez de 1 solo valor)
- Como antes, los bins **no se deben superponer**

Peso (kg)
59,032
78,127
95,900

Distribución de Datos continuos

AgeAtDeath
81
64
88
85
96
101
87
82

700
datos

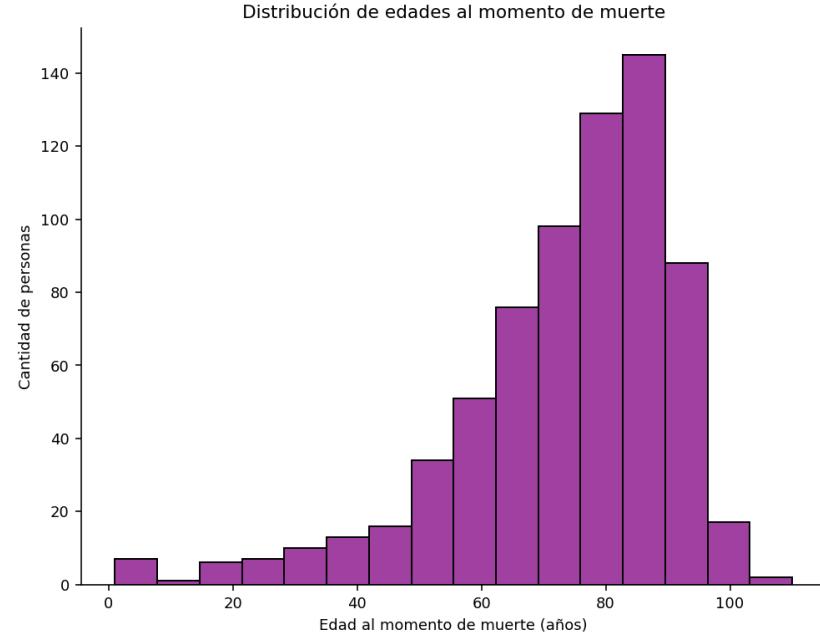


Distribución de Datos continuos

```
import seaborn as sns
#Graficar la distribucion usando seaborn
plt.figure(figsize=(8, 6))
sns.histplot(ageAtDeath['AgeAtDeath'], kde=False, bins=16,
color='purple') # kde=True agrega la curva de densidad

plt.title('Distribución de edades al momento de muerte')
plt.xlabel('Edad al momento de muerte (años)')
plt.ylabel('Cantidad de personas')

# Mostrar el grafico
plt.show()
```

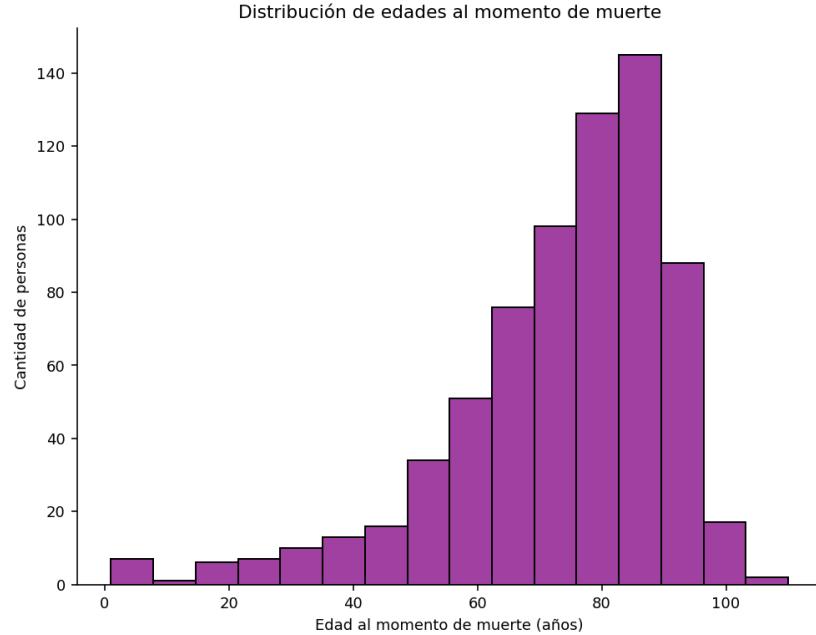


Distribución de Datos continuos

```
import seaborn as sns
#Graficar la distribucion usando seaborn
plt.figure(figsize=(8, 6))
sns.histplot(ageAtDeath['AgeAtDeath'], kde=False, bins=16,
color='purple') # kde=True agrega la curva de densidad

plt.title('Distribución de edades al momento de muerte')
plt.xlabel('Edad al momento de muerte (años)')
plt.ylabel('Cantidad de personas')

# Mostrar el grafico
plt.show()
```



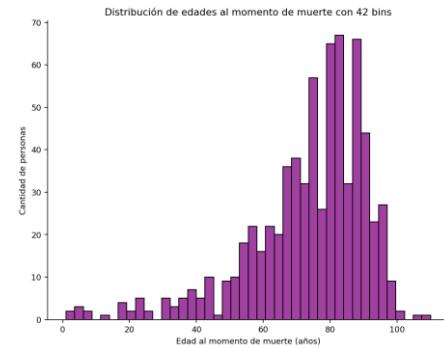
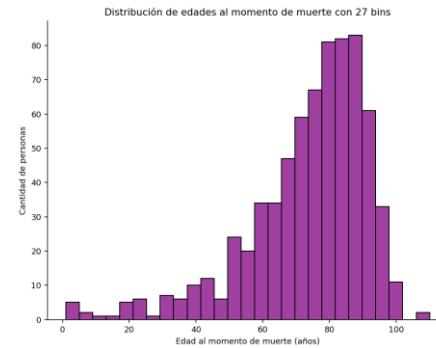
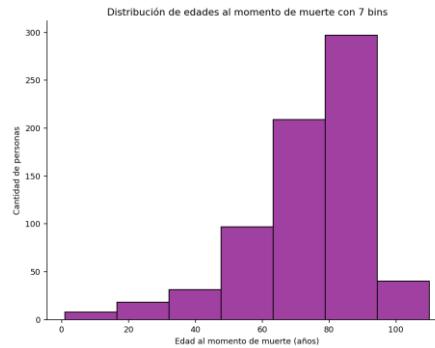
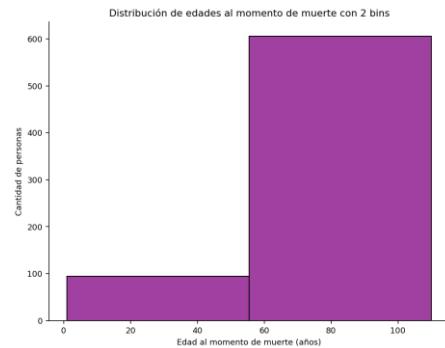
¿Qué pasa si cambiamos la cantidad de bins? Grafiquen usando los valores comentados en el código

Distribución de Datos continuos

La cantidad de bins y el ancho de los mismos puede afectar en gran medida la visualización de una distribución

Para graficar tenemos que definir:

- La cantidad de bins
- El ancho (rango numérico) de cada bin
- El rango total que abarca el conjunto de bins



Distribución de Datos continuos

1. Cantidad de bins

Muchos bins → contienen sólo unas pocas observaciones → no captura patrones generalizables (puede parecer irregular y "ruidoso")

Pocos bins → rango de valores muy amplio en mismo bin → no captura con precisión la variación en los datos y solo presenta patrones "borrosos" de alto nivel.

La elección de la cantidad de bins es **subjetiva**, depende del tema y el objetivo del análisis.

Recomendación: utilizar de 5 a 20 bins

Pocas observaciones → 5 o 6 bins

Muchas observaciones → Más bins.

Distribución de Datos continuos

2. Ancho de bins

Tomar anchos distintos para cada bin puede llevar a decisiones equivocadas

Recomendación: utilizar bins del mismo ancho

Distribución de Datos continuos

2. Rango de valores de bins

Todas las observaciones deberían caer dentro de un bin

Los bins **no deben superponerse**

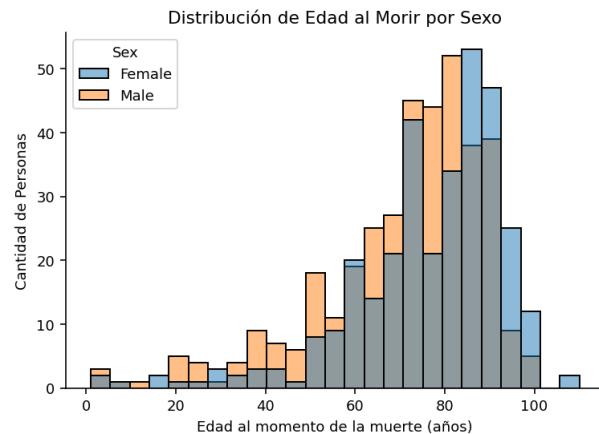
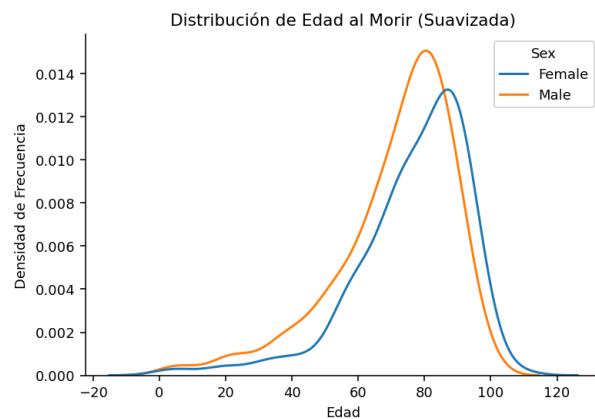
Tener cuidado con los extremos

Recomendación: utilizar rangos de bins que cumplan con lo anterior

Distribución de Datos continuos

¿Y si queremos analizar la variabilidad de dos variables?

Sex	AgeAtDeath
Female	81
Female	64
Male	88
Female	85
Female	96
Female	101
Female	87
Male	



Creemos nuestros gráficos



Ejercicio

Consigna.

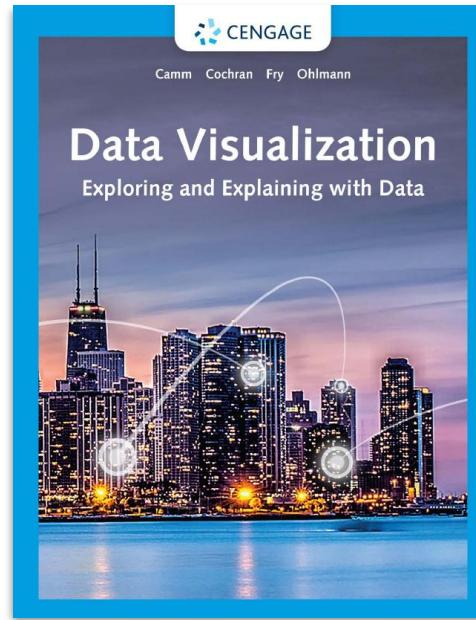
Sean los datos correspondientes a las propinas de un bar (están cargados en el campus en el archivo *tips.csv*)

1. Generar un gráfico para analizar la distribución de la propina en función del:
 - Sexo
 - Día de la semana
1. Comentar los resultados obtenidos

Tarea

- Resolver la guía de ejercicios de visualización (hasta ejercicio 12)

Bibliografía



Camm/Cochran/Fry/Ohlmann, Data Visualization: Exploring and Explaining with Data, 1st. Edition, Cengage Learning, 2022

Estadística descriptiva

Medidas de tendencia y dispersión

Repaso clase anterior

- 1. Exploración y Explicación**
- 2. Distintas maneras de visualizar y explorar datos**
- 3. Ejemplos (Barras, Puntos, Líneas, Torta, etc.)**
- 4. Distribución de Datos (Histogramas de variables categóricas y continuas)**

Análisis estadístico

¿Cómo caracterizarían/resumirían el precio de venta de estas propiedades en un único valor? (2 min.)

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

¿Cómo lo obtuvieron?

Medidas de tendencia

Medidas de tendencia

Una medida de tendencia (central) es un valor único asociado a una variable para caracterizar de alguna manera el conjunto completo de valores

- Existen distintas medidas
- Cada una posee ventajas y desventajas relativas respecto a las otras

Medidas de tendencia

Media (o valor promedio) es la sumatoria de todos los datos dividida la cantidad total de datos

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

$$\text{Media} = \frac{\$108.000 + \$138.000 + \dots + \$456.400}{12} = \$219.950$$

En general:

$$\text{media}(x) = \frac{\sum_{i=1}^N X_i}{N}$$

Medidas de tendencia

Mediana es el número del medio; se encuentra al ordenar todos los valores y elegir el que está en el medio (o si hay dos números en el medio, tomar el promedio de esos dos números)

Están
ordenado
s

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

$$\text{Mediana} = \frac{\$199.500 + \$208.000}{2} = \$203.750$$

Medidas de tendencia

Media (promedio) → Es influenciada por valores atípicos (valores extremadamente chicos/grandes)
Mediana → No influenciada por valores atípicos (su cálculo es robusto)

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

$$\underbrace{\$219.950}_{\text{Media}} > \underbrace{\$203.750}_{\text{Mediana}}$$

Aumenta la media
pero no la mediana

Valor
extremo

Si reemplazáramos los \$456.400 por \$1,5 millones
(media = \$306.916,67; mediana = \$203.750)
la mediana permanecería sin cambios



La mediana da una mejor idea del precio de venta

Siempre que un conjunto de datos contiene **valores extremos**, la **mediana** es la medida preferida de tendencia central (en particular para conjuntos de datos con pocas observaciones)

Medidas de tendencia

Moda es el número más frecuente, es decir, el número que se repite el mayor número de veces (en caso de empate, puede existir más de una moda; en caso de no existir repeticiones los datos no tienen moda)

		PrecioDeVenta
se repite 2 veces	{	\$108.000
		\$138.000
se repite 2 veces	{	\$138.000
		\$142.000
		\$186.000
		\$199.500
		\$208.000
		\$254.000
		\$254.000
		\$257.500
		\$298.000
		\$456.400

$$\text{Moda} = \{ \$138.000; \$254.000 \}$$

- Útil para variables que tienen pocos valores distintos
- Variables con muchos valores distintos (Ej. tiempos de maratonistas en una carrera)
 - Es posible que la moda no exista ¿por qué?
 - Alternativa: construir histograma y aplicar la noción de moda para referirse al bin con mayor cantidad de observaciones.

Medidas de Dispersión

Medidas de dispersión

Consigna. Sean:

Notas Estudiante A: 4; 5; 7; 7; 7 ; 9; 10

Notas Estudiante B: 7; 7; 7; 7; 7; 7; 7

Calcular la Media, Mediana y Moda para cada uno de los Estudiantes

Medidas de dispersión

Consigna. Sean:

Notas Estudiante A: 4; 5; 7; 7; 7 ; 9; 10

Notas Estudiante B: 7; 7; 7; 7; 7; 7; 7

Calcular la Media, Mediana y Moda para cada uno de los Estudiantes

Respuestas.

	Estudiante A	Estudiante B
Media	7	7
Mediana	7	7
Moda	7	7

- Medidas de tendencia central **no describen de qué manera varían los valores**
- Es necesario un valor que permita caracterizar a la “**dispersión**” de los valores ¿Cuál?

Medidas de dispersión

Rango es la diferencia numérica entre el valor **máximo** y el valor **mínimo**

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

Mínimo

$$\text{Rango} = \$456.400 - \$108.000 = \$348.400$$

Máximo

¿Problemas?

- Se basa sólo en 2 valores (máximo y mínimo)
- Influenciable por valores extremos

Medidas de dispersión

Desviación Estándar representa cuánto se apartan los valores del valor promedio

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

$$SD = \sqrt{\frac{(\$108.000 - \$219.950)^2 + \dots + (\$456.400 - \$219.950)^2}{12 - 1}} = \$95.100$$

Media

Medidas de dispersión

Desviación Estándar representa cuánto se apartan los valores del valor promedio

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

$$SD = \sqrt{\frac{(\$108.000 - \$219.950)^2 + \dots + (\$456.400 - \$219.950)^2}{12 - 1}} = \$95.100$$

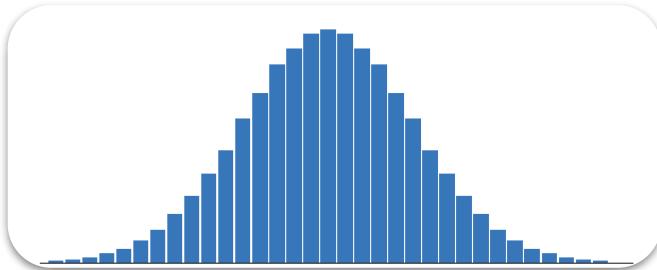
En general:

	Desviación Estándar
Población	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$
Muestra	$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$

Medidas de dispersión

Desviación Estándar representa cuánto se apartan los valores del valor promedio

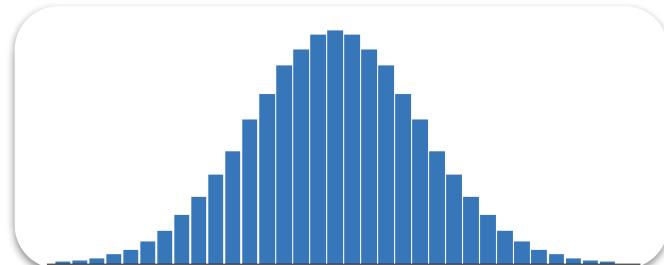
- Medida confiable cuando el histograma tiene forma de campana simétrica



Medidas de dispersión

Desviación Estándar representa cuánto se apartan los valores del valor promedio

- Medida confiable cuando el histograma tiene . forma de campana simétrica



En estos casos la variabilidad nos permite describir los datos usando intervalos :

- el 68% de los valores están en [media - 1 SD; media +1 SD]
- el 95% de los valores están en [media - 2 SD; media +2 SD]
- > 99% de los valores están en [media - 3 SD; media +3 SD]

¿Problemas?

- No es confiable para distribuciones asimétricas
- Influenciable por valores extremos

Medidas de dispersión

Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

Medidas de dispersión

Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

¿Cómo se calcula el valor del p-ésimo percentil?

1. Calcular su posición entre el conjunto de valores ordenados

Ejemplo. ¿percentil 25? ->

$$Lp = (n+1) \frac{P}{100}$$

Medidas de dispersión

Lp: Localización en la lista de datos ordenados

n: Tamaño muestral

P: Percentil a calcular

Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

¿Cómo se calcula el valor del p-ésimo percentil?

1. Calcular su posición entre el conjunto de valores ordenados

Ejemplo. ¿percentil 25? ->

¿Cuál es el Percentil 25 de altura?



$$Lp = (n+1) \frac{P}{100}$$

Lp: Localización en la lista de datos ordenados

n: Tamaño muestral

P: Percentil a calcular

Medidas de dispersión

Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

¿Cómo se calcula el valor del p-ésimo percentil?

1. Calcular su posición entre el conjunto de valores ordenados

Ejemplo. ¿percentil 25? -> $\frac{25 \times (12 + 1)}{100} = 3,25$

entre posiciones 3 y 4

2. Realizar la interpolación necesaria

$$\$138.000 + (3,25 - 3) \times (\$142.000 - \$138.000) = \$139.000$$

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

25% de los datos

$$Lp = (n+1) \frac{P}{100}$$

Lp: Localización en la lista de datos ordenados

n: Tamaño muestral

P: Percentil a calcular

Medidas de dispersión

Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

¿Cómo calcular el valor del p-ésimo percentil?

1. Calcular su posición entre el conjunto de valores ordenados

Ejemplo. ¿percentil 50? -> $\frac{50 \times (12 + 1)}{100} = 6,5$

entre posiciones 6 y 7

2. Realizar la interpolación necesaria

$$\$199.500 + (6,5 - 6) \times (\$208.000 - \$199.500) = \$203.750$$

¡ Coincide con la mediana !

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

50% de los datos

$$Lp = (n+1) \frac{P}{100}$$

Lp: Localización en la lista de datos ordenados

n: Tamaño muestral

P: Percentil a calcular

Medidas de dispersión

Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

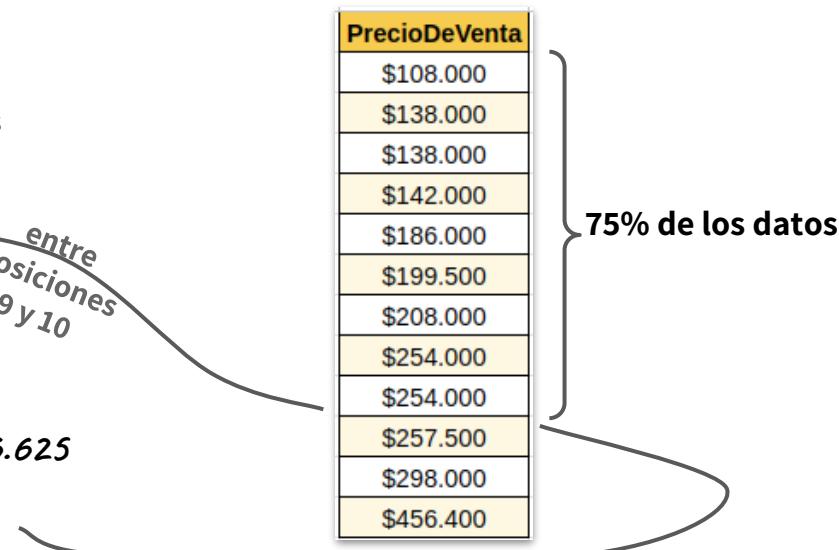
¿Cómo calcular el valor del p-ésimo percentil?

1. Calcular su posición entre el conjunto de valores ordenados

Ejemplo. ¿percentil 75? -> $\frac{75 \times (12 + 1)}{100} = 9,75$

2. Realizar la interpolación necesaria

$$\$254.000 + (9,75 - 9) \times (\$257.500 - \$254.000) = \$256.625$$



Medidas de dispersión

Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

En resumen, para calcular percentiles debemos:

- Ordena los datos
- Usa la fórmula para determinar la posición del percentil
 - Si es un número entero, toma el valor en esa posición
 - Si es decimal, interpola entre los valores adyacentes

Medidas de dispersión

Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

- Se puede calcular un percentil para cualquier valor entre 0% y 100%
- Percentiles más comunes: 25, 50 y 75 (primer cuartil, segundo cuartil y tercer cuartil)
- **Rango intercuartil (IQR)**. La diferencia entre el **tercer** y el **primer** cuartil (los percentiles 75 y 25)
- IQR **abarca el 50% medio de la distribución** de los valores y se utiliza como medida de variación
- Ventajas de percentiles y el rango intercuartil sobre el rango y la desviación estándar
 - **percentiles no requieren** que la distribución de la variable tenga **forma de campana**
 - **valores extremos no distorsionan** el valor de los percentiles

Medidas de dispersión

Consigna. Sean ...

Notas Estudiante A: 4; 5; 7; 7; 7 ; 9; 10

Notas Estudiante B: 7; 7; 7; 7; 7; 7; 7

Calcular la Media, Mediana y Moda para cada uno de los Estudiantes

Calcular el rango, desvío estándar, cuartiles e IQR

Respuestas

	Estudiante A	Estudiante B
Media	7	7
Mediana	7	7
Moda	7	7

Medidas de dispersión

Consigna. Sean ...

Notas Estudiante A: 4; 5; 7; 7; 7 ; 9; 10

Notas Estudiante B: 7; 7; 7; 7; 7; 7; 7

Calcular la Media, Mediana y Moda para cada uno de los Estudiantes

Calcular el rango, desvío estándar, cuartiles e IQR

Respuestas

	Estudiante A	Estudiante B
Media	7	7
Mediana	7	7
Moda	7	7

	Estudiante A	Estudiante B
Rango	6	0
STD	2,08	0,00
1Q	5	7
2Q	7	7
3Q	9	7
IQR	4	0

Estadística descriptiva con python

- Vamos a trabajar con el dataset 'tips'

Índice	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.5	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2

- Calculamos la **media** de la columna 'tip'

```
In [13]:  
...: tips['tip'].mean()  
Out[13]: 2.99827868852459
```

Estadística descriptiva con python

- Vamos a trabajar con el dataset '*tips*'

Índice	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.5	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2

- Calculamos la **mediana** de la columna 'tip'

```
In [15]:  
....: tips['tip'].median()  
Out[15]: 2.9
```

Estadística descriptiva con python

- Vamos a trabajar con el dataset '*tips*'

Índice	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.5	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2

- Calculamos la **moda** de la columna '*tip*'

```
In [16]:  
...: tips['tip'].mode()  
Out[16]:  
0    2.0  
Name: tip, dtype: float64
```

Estadística descriptiva con python

- Calculamos el **rango** de la columna 'tip'

```
In [18]:  
....: rango_tips = max(tips['tip']) - min(tips['tip'])  
....: print(rango_tips)  
9.0
```

- Calculamos la **desviación estándar** de la columna 'tip'

```
In [17]: tips['tip'].std()  
Out[17]: 1.3836381890011822
```

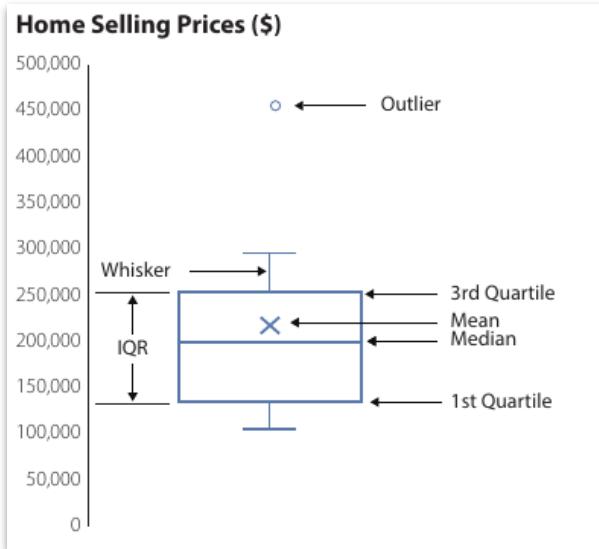
Estadística descriptiva con python

- Un comando **muy útil** es el método `.describe()`, el cual proporciona estadísticas descriptivas como la media, el percentil 25, el percentil 50 (mediana), el percentil 75, entre otros.

```
In [22]: tips['tip'].describe()
Out[22]:
count    244.000000
mean      2.998279
std       1.383638
min       1.000000
25%       2.000000
50%       2.900000
75%       3.562500
max      10.000000
Name: tip, dtype: float64
```

Boxplot

Boxplot es un resumen gráfico de la distribución de los datos. Se basa en los cuartiles.



Caja (Box)

$$\text{3er. cuartil} = \text{Percentil } 75 = \$256.625$$

$$\text{2do. cuartil} = \text{Percentil } 50 = \$203.750 \\ (\text{Mediana})$$

$$\text{1er. cuartil} = \text{Percentil } 25 = \$139.000$$

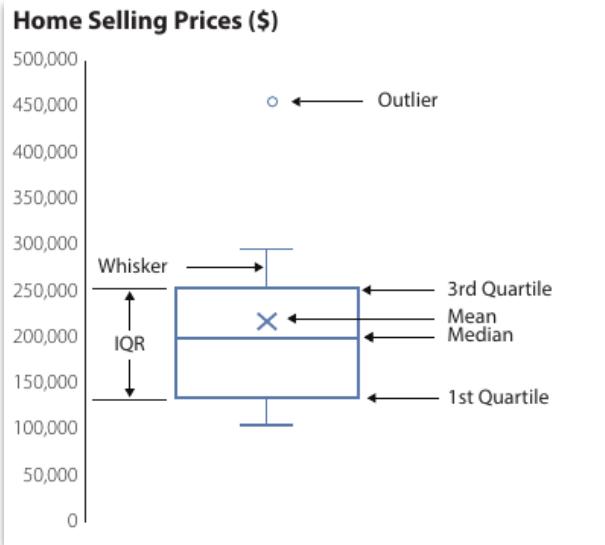
$$\text{IQR} = \text{3er. - 1er cuartil} = \$117.625$$

$$\text{Media} = \$219.950$$

PrecioDeVenta
\$456.400
\$298.000
\$257.500
\$254.000
\$254.000
\$208.000
\$199.500
\$186.000
\$142.000
\$138.000
\$138.000
\$108.000

Boxplot

Boxplot es un resumen gráfico de la distribución de los datos. Se basa en los cuartiles.



Bigotes (Whisker)

$$\text{Límite Sup.} = \text{3er. cuartil} + 1,5 * \text{IQR} = \$433,062.5$$

$$\text{Límite Inf.} = \text{1er. cuartil} - 1,5 * \text{IQR} = \$-37.437$$

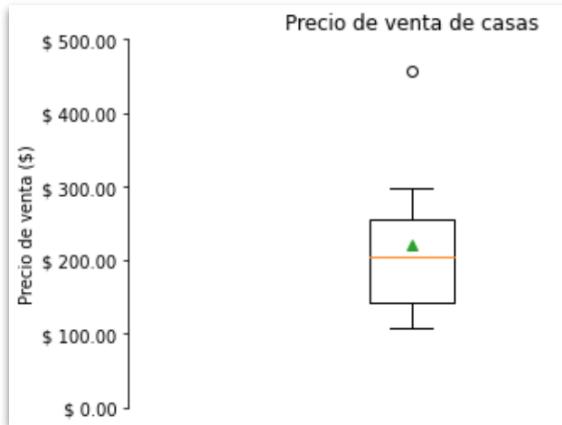
Los bigotes se extienden hasta el último dato que se encuentra dentro del límite

PrecioDeVenta
\$456.400
\$298.000
\$257.500
\$254.000
\$254.000
\$208.000
\$199.500
\$186.000
\$142.000
\$138.000
\$138.000
\$108.000

← Outlier
← Bigote Sup.
← Bigote inf.

Boxplot

Boxplot es un resumen gráfico de la distribución de los datos. Se basa en los cuartiles.



Bigotes (Whisker)

$$\text{Límite Sup.} = \text{3er. cuartil} + 1,5 * \text{IQR} = \$433,062.5$$

$$\text{Límite Inf.} = \text{1er. cuartil} - 1,5 * \text{IQR} = \$-37.437$$

Los bigotes se extienden hasta el último dato que se encuentra dentro del límite

PrecioDeVenta
\$456.400
\$298.000
\$257.500
\$254.000
\$254.000
\$208.000
\$199.500
\$186.000
\$142.000
\$138.000
\$138.000
\$108.000

← Outlier
← Bigote Sup.
Bigote inf.

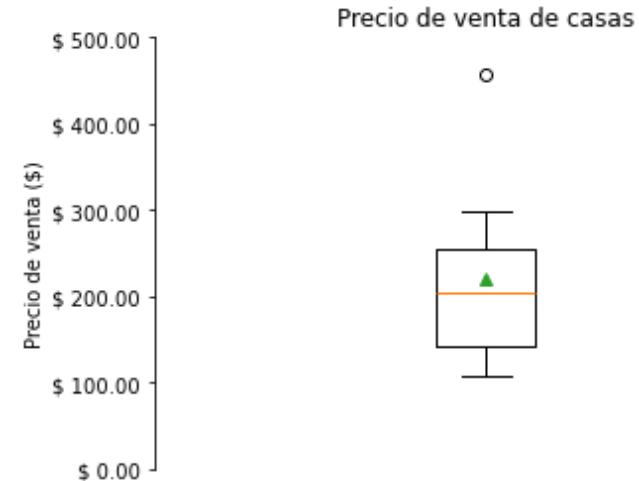
Boxplot

```
fig, ax = plt.subplots()

ax.boxplot(ventaCasas['PrecioDeVenta'], showmeans=True)

# Agrega titulo, etiquetas a los ejes
# y limita el rango de valores de los ejes
ax.set_title('Precio de venta de casas')
ax.set_xticks([])
ax.set_ylabel('Precio de venta ($)')
# Agrega separador de decimales y signo $
ax.yaxis.set_major_formatter(ticker.StrMethodFormatter("$ {x:,.2f}"));

ax.set_ylim(0,500)
```

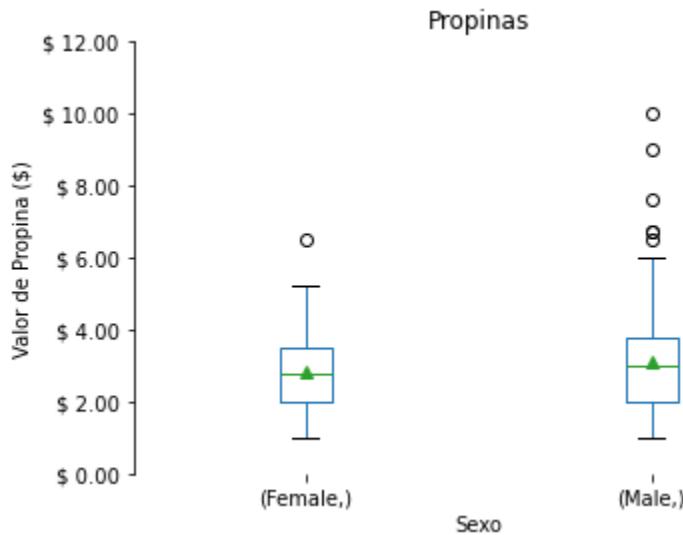


Boxplot: incorporando variables categóricas al análisis

```
fig, ax = plt.subplots()

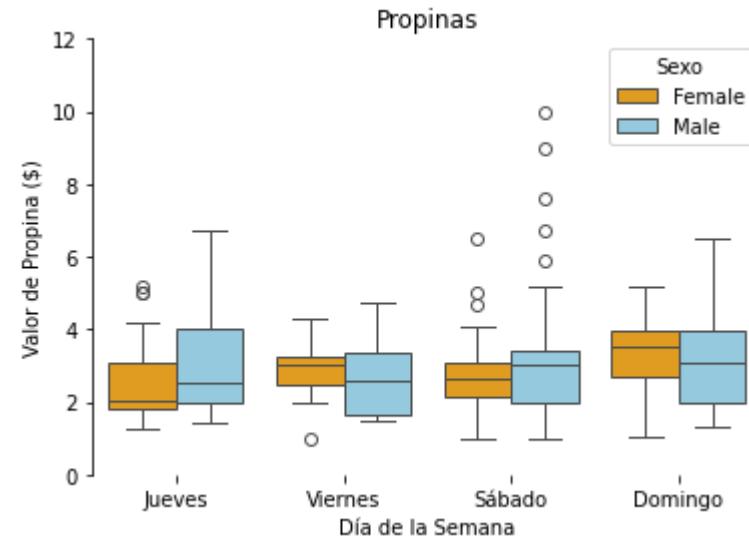
tips.boxplot(by=['sex'], column=['tip'],
             ax=ax, grid=False, showmeans=True)

# Agrega título, etiquetas a los ejes
fig.suptitle('')
ax.set_title('Propinas')
ax.set_xlabel('Sexo')
ax.set_ylabel('Valor de Propina ($)')
```



Boxplot: incorporando variables categóricas al análisis

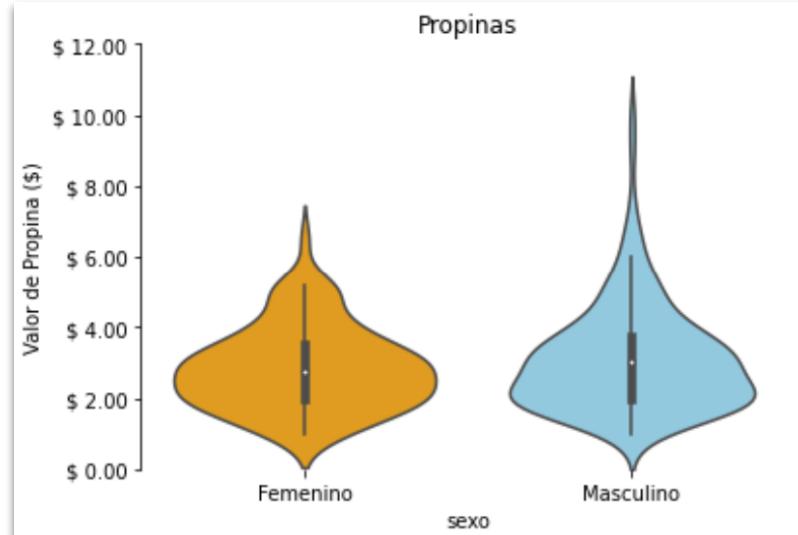
```
ax = sns.boxplot(x="day", ←  
                  y="tip", ←  
                  hue="sex", ←  
                  data=tips,  
                  order=['Thur', 'Fri', 'Sat', 'Sun'],  
                  palette={"Female": "orange", "Male": "skyblue" })  
  
ax.set_title('Propinas')  
ax.set_xlabel('Día de la Semana')  
ax.set_ylabel('Valor de Propina ($)')  
ax.set_ylim(0,12)  
ax.legend(title="Sexo")  
  
dias = ['Thur', 'Fri', 'Sat', 'Sun']  
labels = ['Jueves', 'Viernes', 'Sábado', 'Domingo']  
ax.set_xticks(range(len(dias)))  
ax.set_xticklabels(labels)
```



Violinplot

Violinplot es similar al boxplot, salvo que muestra también la densidad de probabilidad de los datos, generalmente suavizada por un estimador de densidad kernel.

```
ax = sns.violinplot(x ="sex", y ="tip", data = tips,  
                     palette={"Female": "orange", "Male": "skyblue" })  
  
ax.set_title('Propinas')  
ax.set_xlabel('sexo')  
ax.set_ylabel('Valor de Propina ($)')  
ax.yaxis.set_major_formatter(ticker.StrMethodFormatter("$ {x:,.2f}"));  
ax.set_ylim(0,12)  
ax.set_xticklabels(['Femenino','Masculino'])
```



Ejercicio

Consigna.

Utilizando el dataset de propinas (*tips*)

- Generar un gráfico de boxplot de la proporción de propina recibida (tip/total_bill)
- Analizar los resultados obtenidos
- Estudiar si hay diferencias en función del sexo y el sexo+día
- Discutir con el resto de la clase los resultados obtenidos

Cómo mejorar la visualización



Cómo mejorar la visualización de datos

Memoria sensorial	
<i>Definición</i>	Procesos cerebrales que interpretan estímulos por períodos mucho más breves que la memoria de corto plazo
<i>Sistemas derivados</i>	- Memoria icónica - Memoria ecoica - Memoria olfativa - Memoria gustativa - Memoria háptica
<i>Tiempo de permanencia de los datos</i>	Breve (milésimas de seg)
<i>Ejemplos</i>	Percibir un sonido en medio de una multitud

Fuente: <https://www.diferenciador.com/tipos-de-memoria/>

Cómo mejorar la visualización de datos

Memoria sensorial	
Definición	Procesos cerebrales que interpretan estímulos por períodos mucho más breves que la memoria de corto plazo
Sistemas derivados	<ul style="list-style-type: none">- Memoria icónica ←- Memoria ecoica- Memoria olfativa- Memoria gustativa- Memoria háptica
Tiempo de permanencia de los datos	Breve (milésimas de seg)
Ejemplos	Percibir un sonido en medio de una multitud

Visión
Audición
Olfato
Gusto
Tacto

Fuente: <https://www.diferenciador.com/tipos-de-memoria/>

Cómo mejorar la visualización de datos

	Memoria sensorial	Memoria a corto plazo
<i>Definición</i>	Procesos cerebrales que interpretan estímulos por períodos mucho más breves que la memoria de corto plazo	Procesos cerebrales encargados de interpretar los estímulos y conservar esa información durante un tiempo breve
<i>Sistemas derivados</i>	- Memoria icónica - Memoria ecoica - Memoria olfativa - Memoria gustativa - Memoria haptica	- Sistema ejecutivo - Almacén episódico - Bucle fonológico - Agenda visoespacial
<i>Tiempo de permanencia de los datos</i>	Breve (milésimas de seg)	Breve (7 a 40 seg)
<i>Ejemplos</i>	Percibir un sonido en medio de una multitud	Recordar la matrícula de un auto que acaba de pasar



Fuente: <https://www.diferenciador.com/tipos-de-memoria/>

Cómo mejorar la visualización de datos

	Memoria sensorial	Memoria a corto plazo	Memoria a largo plazo
Definición	Procesos cerebrales que interpretan estímulos por períodos mucho más breves que la memoria de corto plazo	Procesos cerebrales encargados de interpretar los estímulos y conservar esa información durante un tiempo breve	Procesos cerebrales encargados de conservar información durante períodos prolongados
Sistemas derivados	- Memoria icónica - Memoria ecoica - Memoria olfativa - Memoria gustativa - Memoria háptica	- Sistema ejecutivo - Almacén episódico - Bucle fonológico - Agenda visoespacial	- Memoria implícita - Memoria explícita
Tiempo de permanencia de los datos	Breve (milésimas de seg)	Breve (7 a 40 seg)	Prolongado (minutos a décadas)
Ejemplos	Percibir un sonido en medio de una multitud	Recordar la matrícula de un auto que acaba de pasar	Recordar cómo manejar bicicleta

Fuente: <https://www.diferenciador.com/tipos-de-memoria/>

Experimento - la memoria

Hagamos un experimento para ver los límites de la memoria a corto plazo:

Consigna:

- 1. Van a ver 5 series de letras**
- 2. Tienen que recordarlas en el orden correcto**

C X W

¿Las recuerdan?

C X W

M N K T Y

؟

M N K T Y

R P J H B Z S

؟

R P J H B Z S

G B M P V Q F J D

؟

G B M P V Q F J D

E G Q W J P B R H K A

؟

E G Q W J P B R H K A

- En general (a la mayoría, después de sólo 1 lectura) las siguientes series les resultaron:
 - 1 -> extremadamente fácil
 - 2 -> fácil
 - 3 -> un poco + difícil
 - 4 -> sumamente difícil
 - 5 -> casi imposible
- Muestra la limitada capacidad de la memoria de corto plazo
- Se estima que en memoria a corto plazo se puede retener alrededor de 4 fragmentos de información visual
- **Consecuencia.** Resulta difícil recordar qué color representa cada categoría si se utilizan más de 4 colores/categorías diferentes en un gráfico de barras o columns.

Atributos pre-atentivos

Atributos pre-atentivos representan aquellas características que pueden ser procesadas por la memoria icónica (memoria sensorial visual)

- Ver un gráfico o una tabla -> Ojos reciben estímulo y se transmite al cerebro
- Cerebro. Debe diferenciar y procesar

Percepción visual. Proceso mediante el cual nuestro cerebro interpreta esos reflejos producidos por la luz que entra por nuestros ojos. Relacionada con el funcionamiento de la memoria

- Memoria Icónica + Corto Plazo: las más importantes para el procesamiento visual
- Comprender qué aspectos se pueden procesar en estos tipos de memoria puede resultar útil para diseñar visualizaciones efectivas
- Ejemplo para ilustrar el poder de los atributos pre-atentivos

Experimento

Consigna ...

- Observar la siguiente figura
- Contar lo más rápido posible cuántas veces aparece el número de 7 en la figura

7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

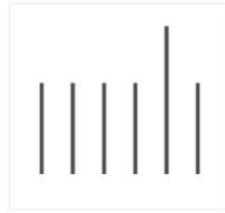
7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

Reflexiones ...

- Respuesta correcta. "Hay 14 sietes"
- Figura 1. Lleva tiempo y es probable haber cometido un error
- Figuras 2 y 3. Uso de atributos pre-atentivos hace el trabajo más fácil
- Figura 2. Hace uso del **tamaño** (grande vs pequeño)
- Figura 3. Hace uso del **color y forma**
- El color y el tamaño se procesan en la memoria icónica y por eso podemos diferenciarlos tan rápidamente
- El uso adecuado de atributos pre-atentivos reduce la cantidad de esfuerzo necesario para procesar de manera precisa y eficiente la información que se comunica mediante una visualización de datos

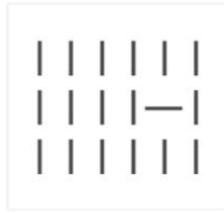
Atributos pre-atentivos



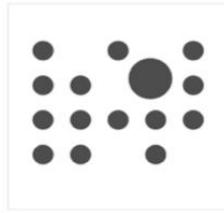
Length



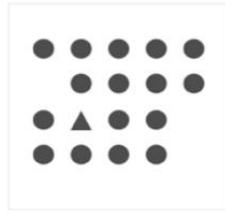
Width



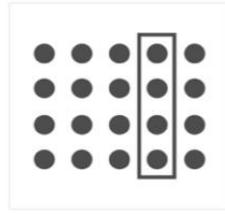
Orientation



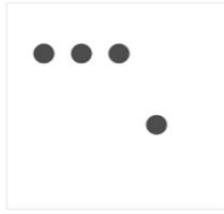
Size



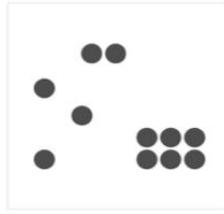
Shape



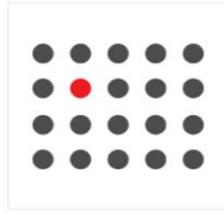
Enclosure



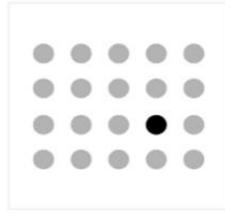
Position



Grouping



Color Hue

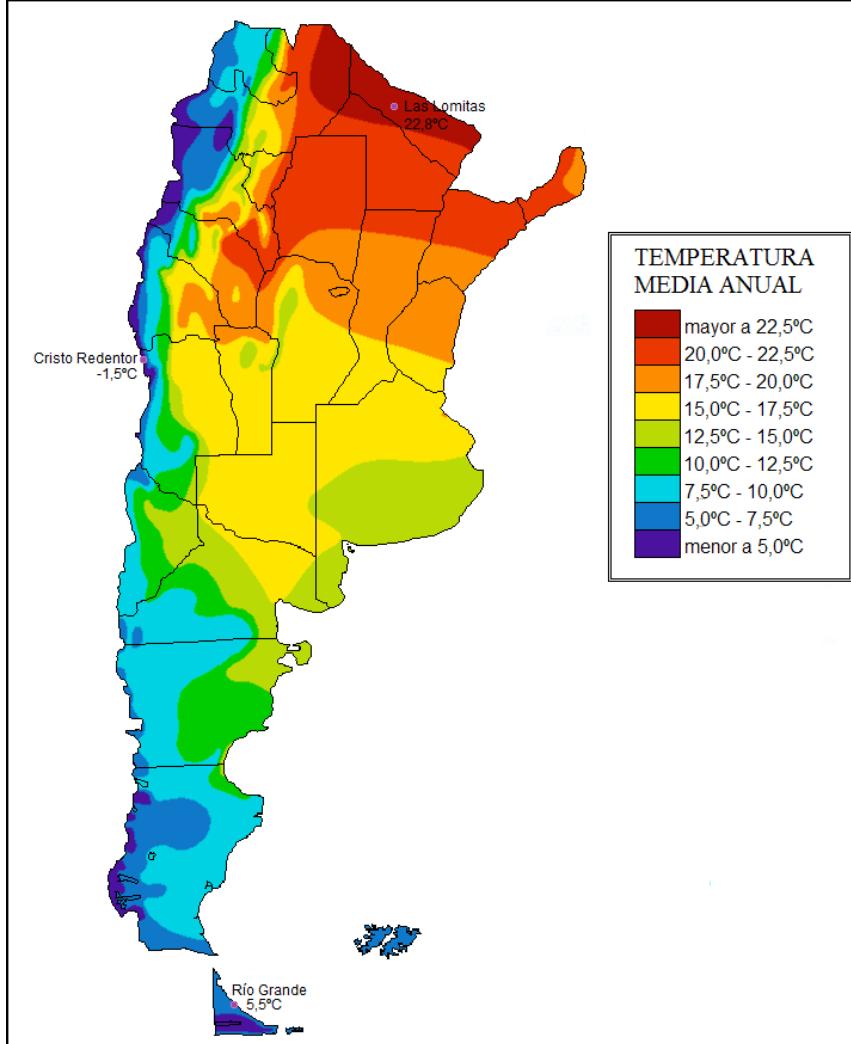
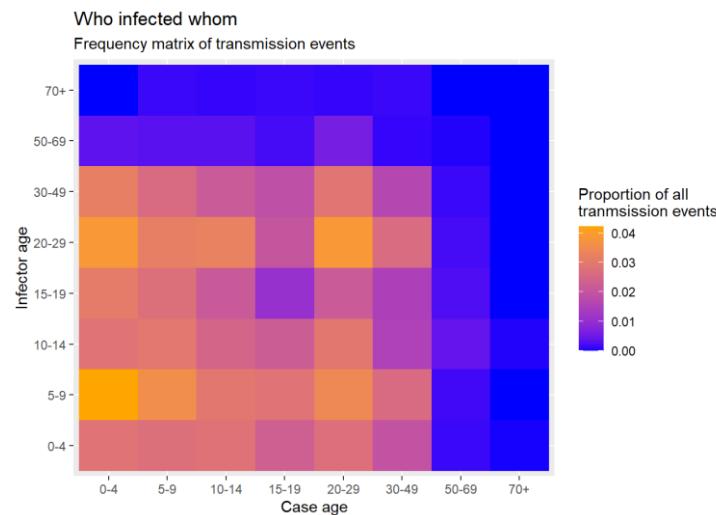


Color
Intensity

- Más info en el libro

Color

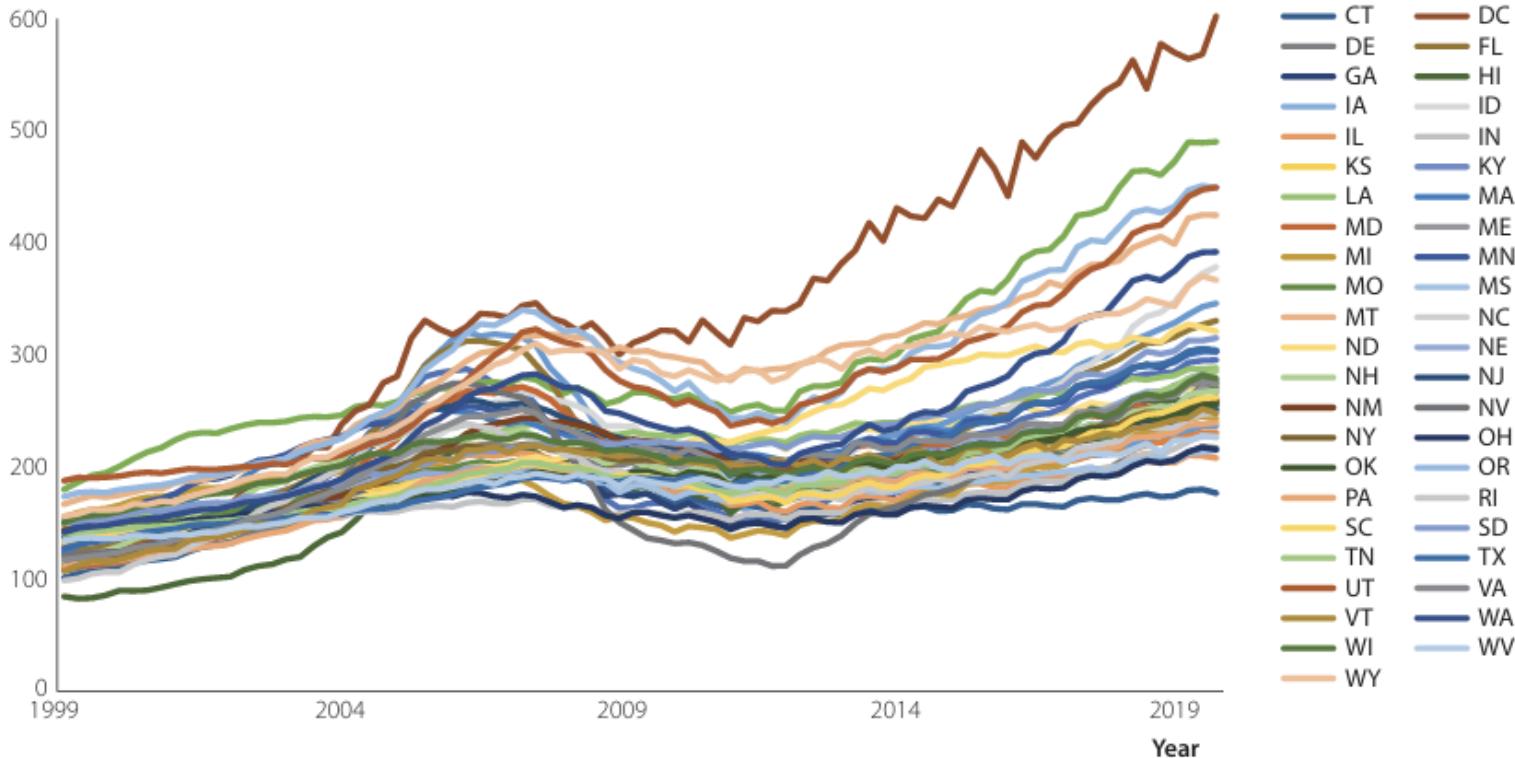
- Llama la atención
- Aumenta la comprensión
- Acelera la interpretabilidad



Color

Quarterly House Price Index by State (Base = 1991)

Home Price Index

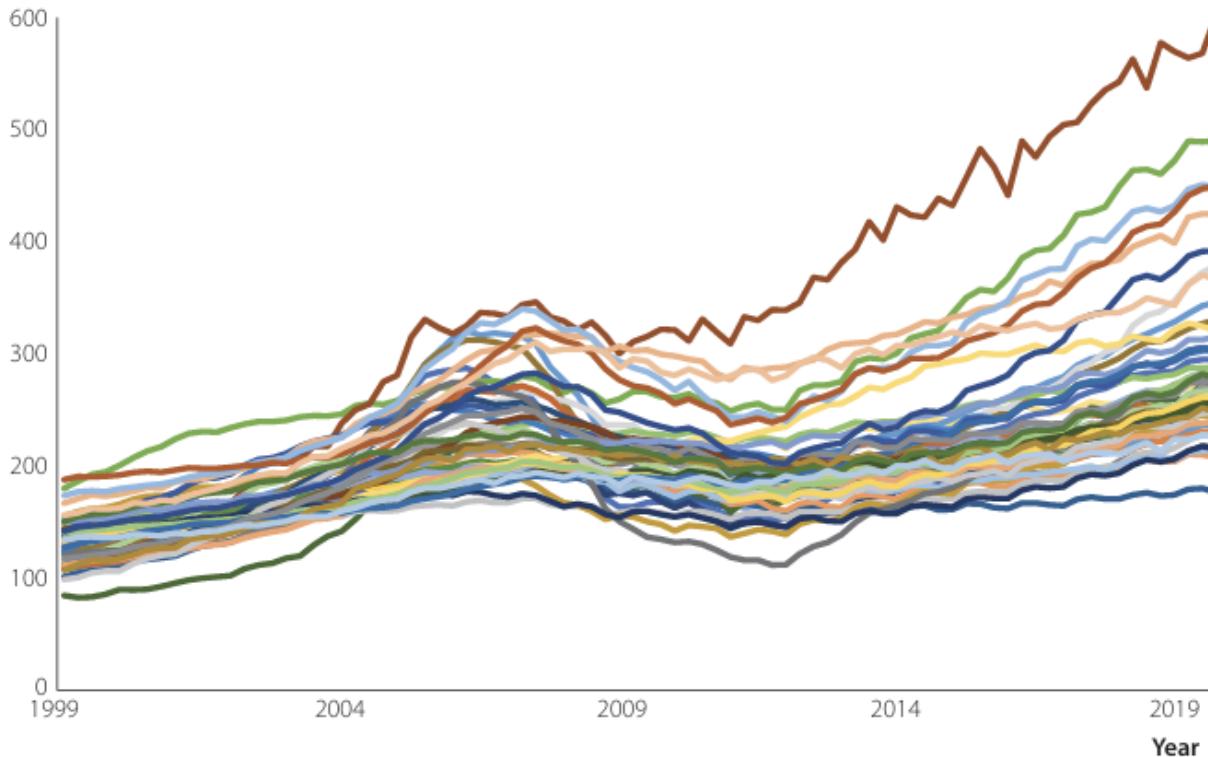


Color

Pero todo
en su justa
medida

Quarterly House Price Index by State (Base = 1991)

Home Price Index

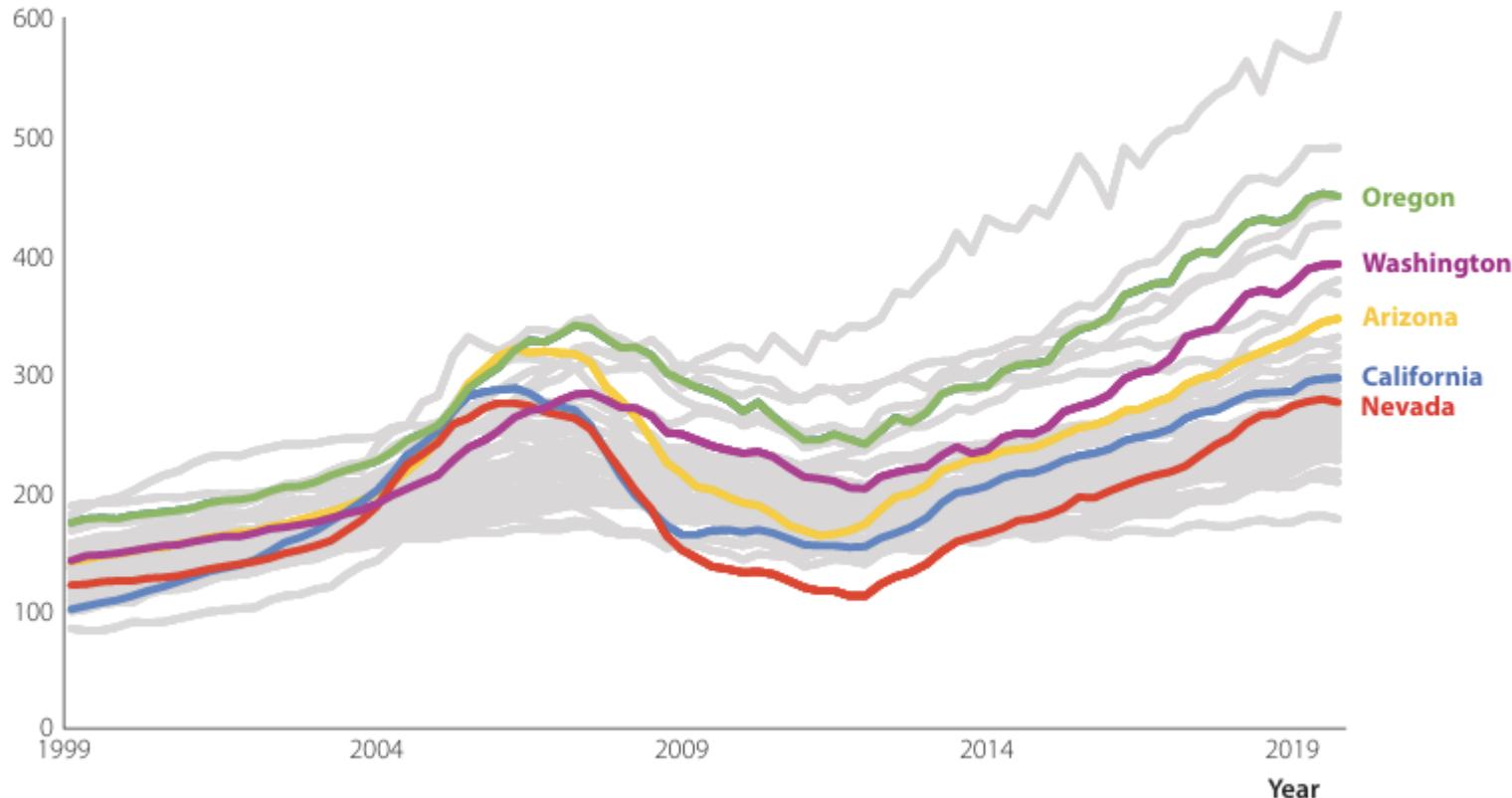


Color

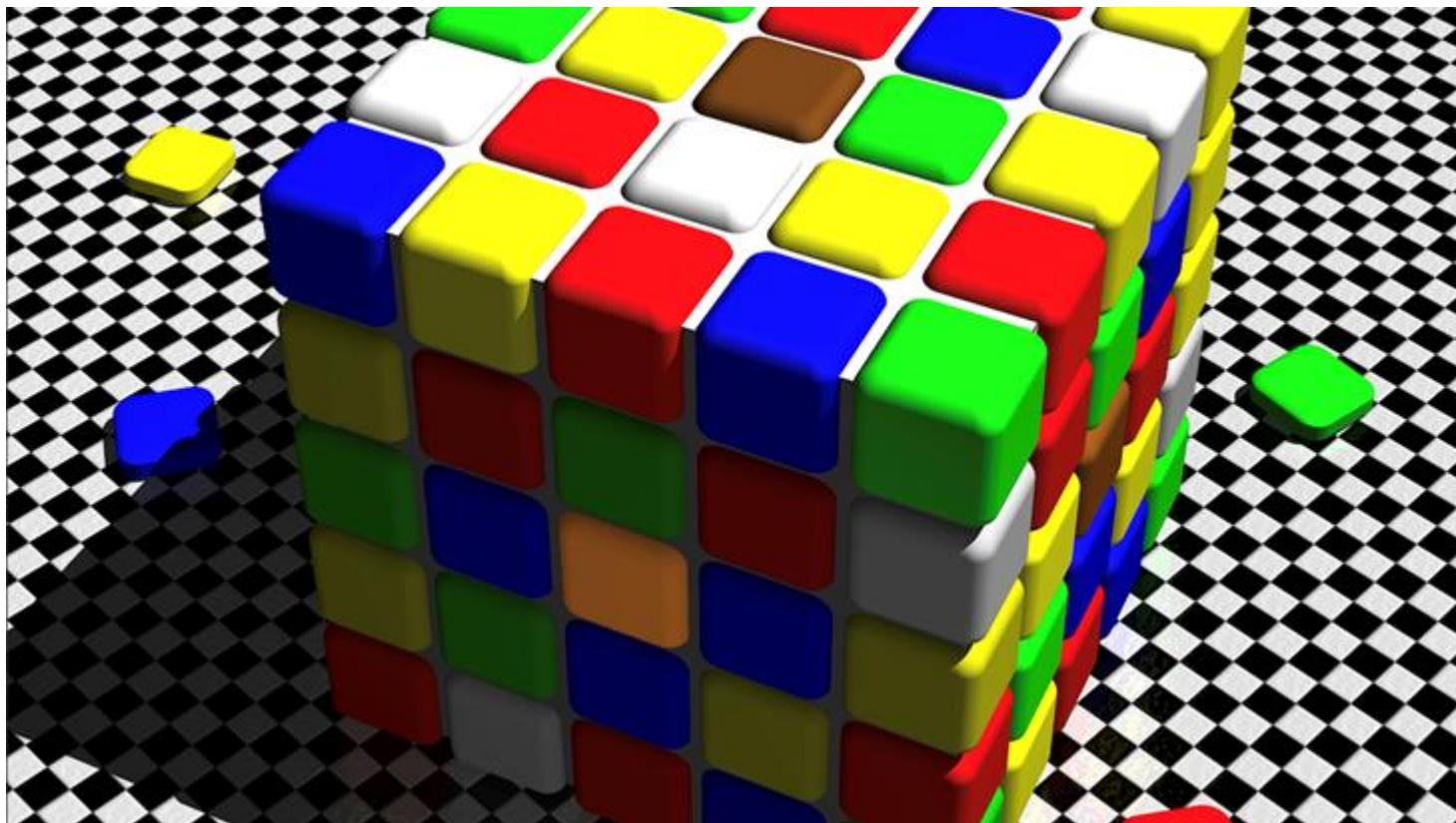
Pero todo
en su justa
medida

Quarterly House Price Index (Base = 1991) for Western States Demonstrates More Volatility

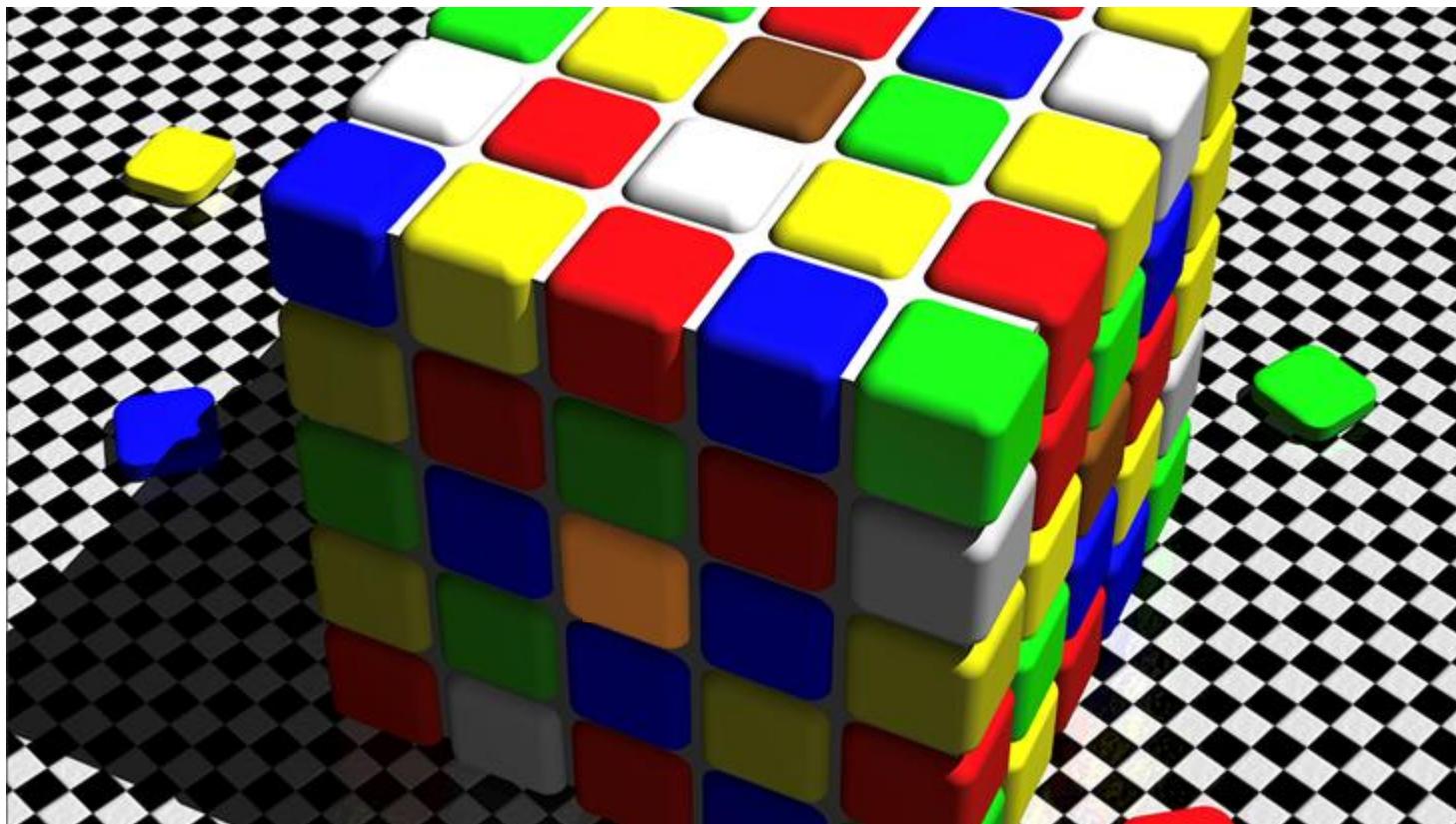
Home Price Index



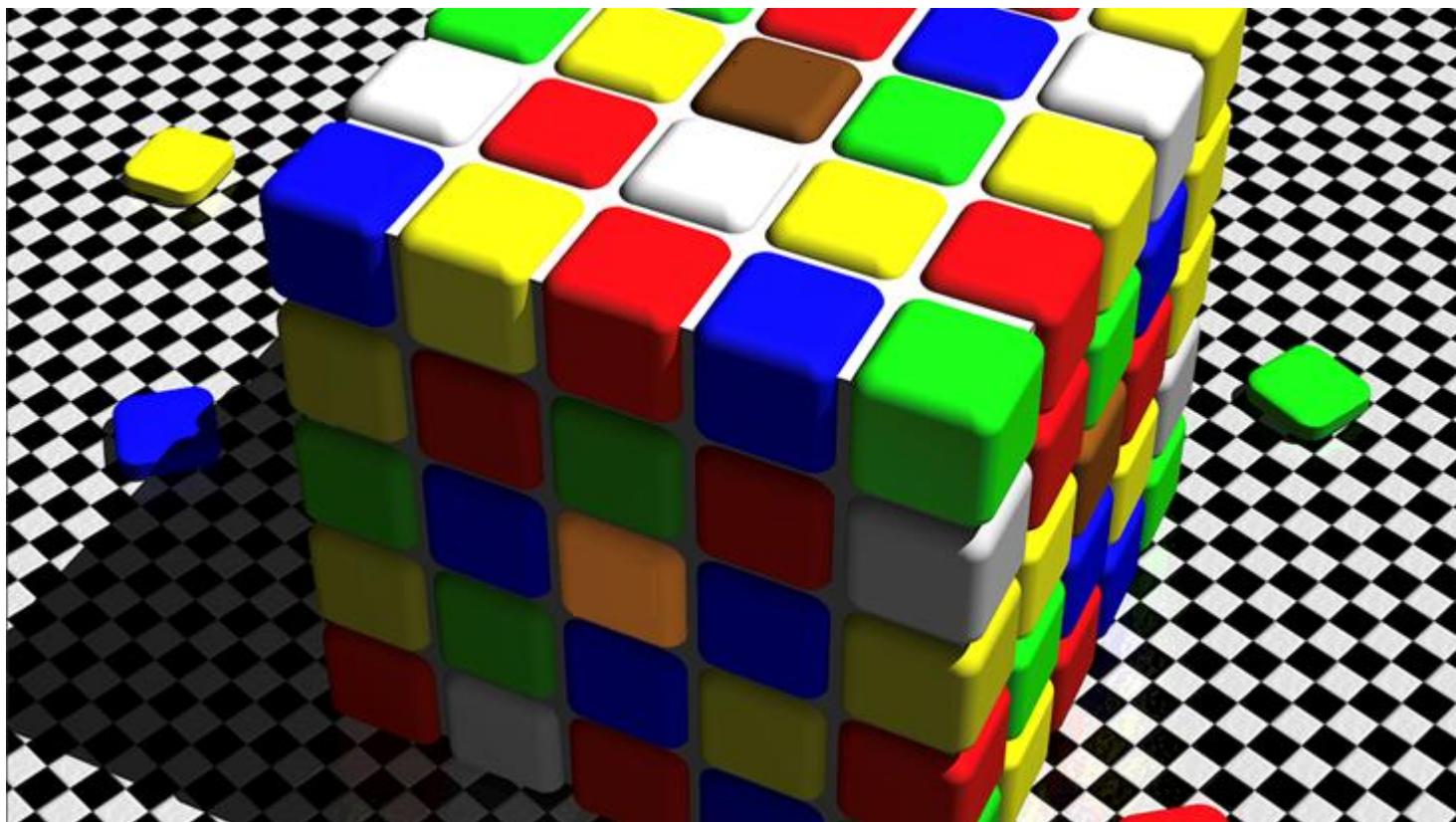
Color: Las apariencias engañan



Color: Las apariencias engañan



Color: Las apariencias engañan



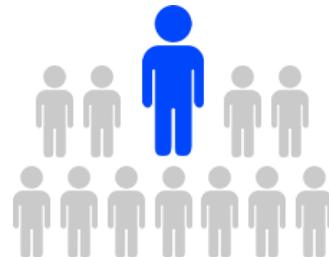
Conociendo al público

Color: Conociendo al público



Alrededor de 350.000.000 personas presentan algún tipo de ceguera al color en el mundo.
Aproximadamente la misma cantidad que toda la población de EEUU.

Color: Conociendo al público



Alrededor de 350.000.000 personas presentan algún tipo de ceguera al color en el mundo.

Aproximadamente la misma cantidad que toda la población de EEUU.

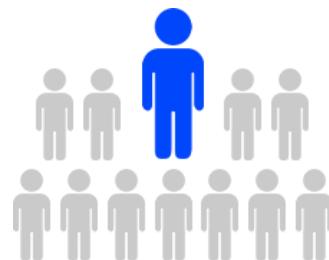
1 de cada 12 hombres presentan algún tipo de ceguera al color

Color: Conociendo al público

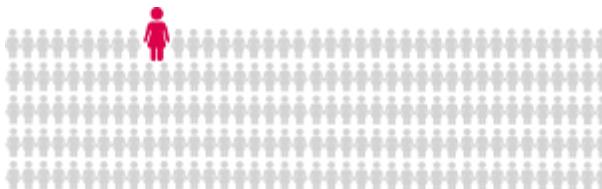


Alrededor de 350.000.000 personas presentan algún tipo de ceguera al color en el mundo.

Aproximadamente la misma cantidad que toda la población de EEUU.



1 de cada 12 hombres presentan algún tipo de ceguera al color



1 de cada 200 mujeres presentan algún tipo de ceguera al color

Color: Conociendo al público



Visión normal

Color: Conociendo al público



Visión normal



Deuteranomalia
(verde débil)



Protanomalia
(rojo débil)



Trianomalia
(azul débil)



Acromatopsia
(monocromacia)



Deuteranopia
(verde ciego)



Protanopia
(rojo ciego)

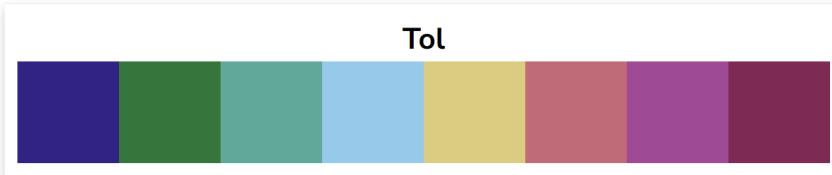
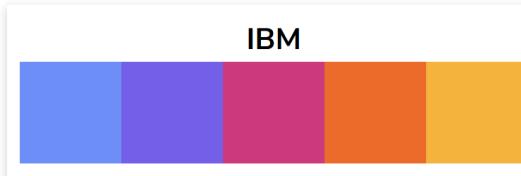
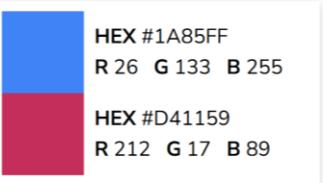
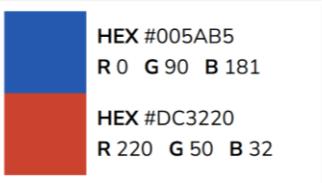
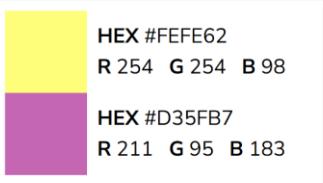
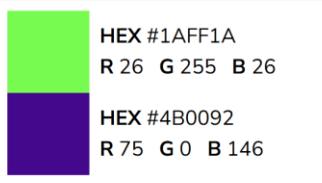
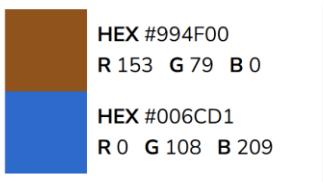
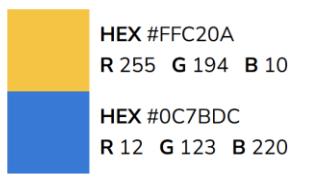


Tritanopia
(azul ciego)



Monocromacia
del cono azul

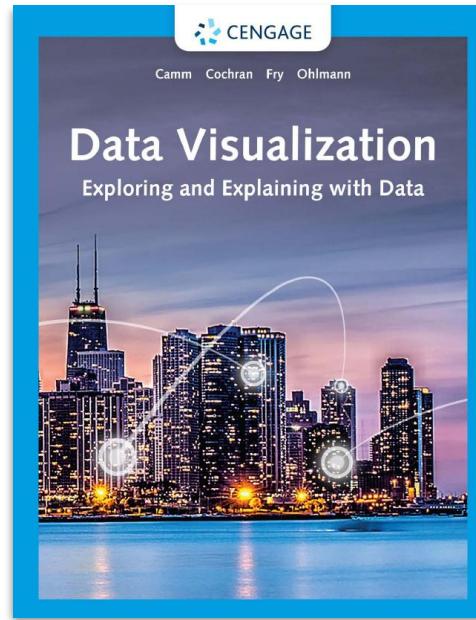
Paletas *colorblind* friendly



Tarea

- Resolver la guía de ejercicios de visualización

Bibliografía



Camm/Cochran/Fry/Ohlmann, Data Visualization: Exploring and Explaining with Data, 1st. Edition, Cengage Learning, 2022