FEB
19

SPRINT 1 - MEETING 6

Matplotlib

Presentación

El equipo docente subió contenido extra de la meeting

Lista de recursos

- Profundiza
- Challenge
- Rincón Matemático
- Notebook: DS_Bitácora_06_Matplotli

Contenido extra

- Clase 19 feb reemplazo clase martes 16

Un detector de fallas

> *"At their best, graphics are instruments for reasoning." —Edward Tufte. Artista, estadístico, y profesor emérito de la Universidad de Yale.*

No todo es lo que parece

En las últimas bitácoras fuimos compartiendo conceptos y herramientas que nos permiten interactuar con los datasets para empezar a explorarlos. Hasta ahora, esas herramientas fueron estadísticas, y lo que obtenemos al usarlas son algunos números representativos de las variables de nuestro dataset. Por ejemplo, valor medio, desviación estándar, percentiles, correlaciones, etc.

Si bien estas herramientas son muy útiles para trabajar, informarse y comunicar las características de un dataset, en general, tienen gusto a poco. El problema con esos cálculos es que restringen el acceso y eventual comprensión de la información a aquellas personas —nosotros/as como data scientists incluidos/as — con los conocimientos necesarios para interpretarlos. Pero, ¿cómo podemos hacer para involucrar a más personas en nuestras observaciones?

Las **visualizaciones** son una herramienta central para este objetivo. No sólo son el complemento perfecto para transmitir de manera clara y precisa lo que estamos observando en el dataset, sino que también nos sirven para detectar errores y terminar de entender realmente con qué datos estamos trabajando, y qué historia queremos contar con ellos.

Un ejemplo de este caso se da con el cuarteto de Anscombe. En 1973, este estadístico generó cuatro conjuntos de datos que tenían tres métricas en común: su media, su varianza y correlación. Sin embargo, al momento de graficarlos, los resultados eran distintos:

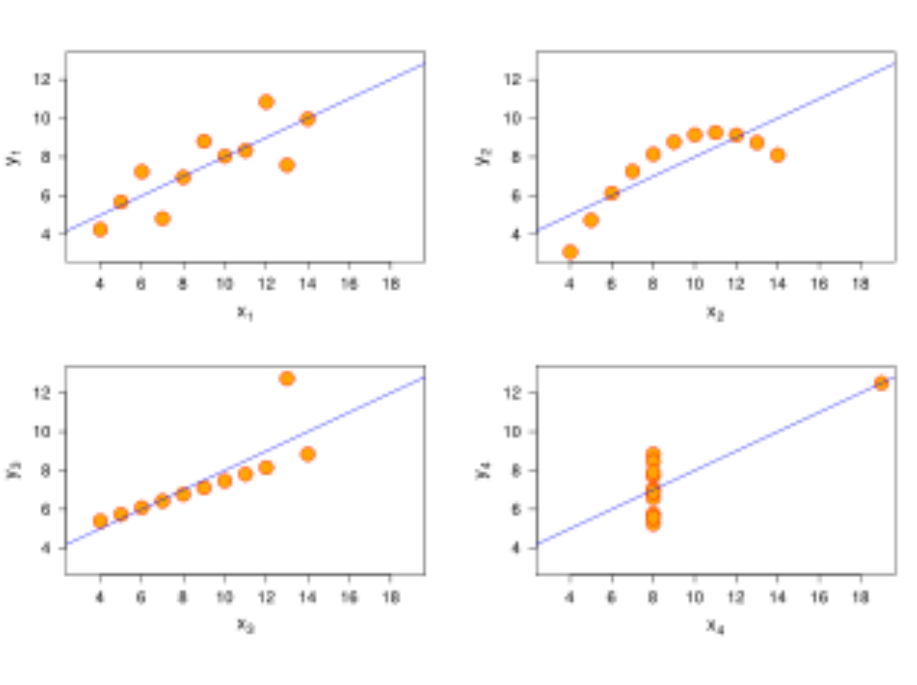
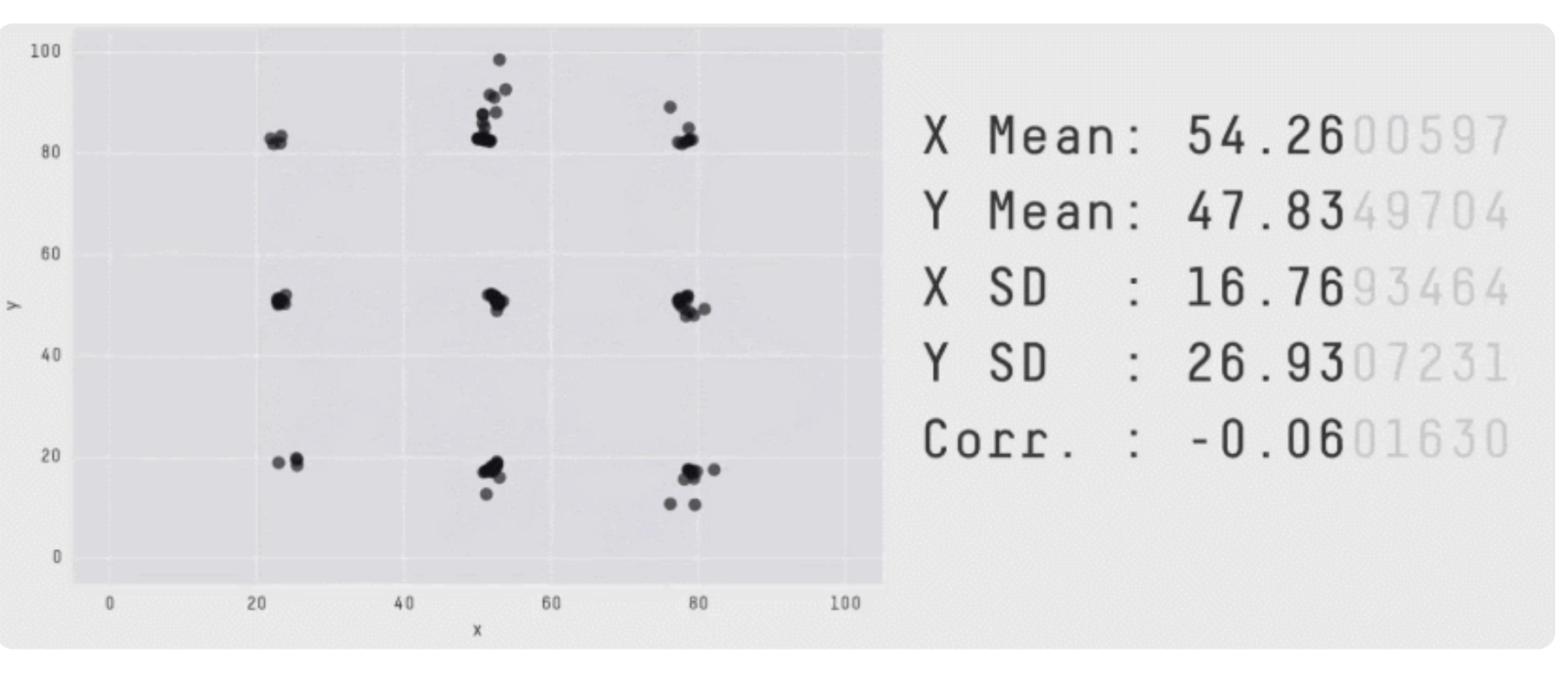


Figura 1. Fuente: Wikipedia

Un dato de color: basándose en esta idea, Justin Matejka y George Fitzmaurice fueron un paso más allá y generaron una herramienta que permite, manteniendo las mismas métricas fijas, generar figuras de lo más diversas, ¡hasta la de un dinosaurio! De ahí el nombre, The *Datasaurus Dozen.*



Data Science se diferencia del paradigma tradicional del método científico. ¿En qué sentido?

Mientras el modelo tradicional trabaja generando una hipótesis que puede ser avalada o refutada a partir de los datos que se generan con Data Science, primero generamos el conjunto necesario de datos para comenzar a identificar patrones que serán las hipótesis para futuros análisis. A esto último, lo conocemos como **Análisis Exploratorio de Datos.**

En resumen, encontrarás distintas formas de visualizar tu dataset, resaltando ciertos rasgos por sobre otros. La decisión de hacerlo de una u otra manera dependerá de los objetivos que guíen tu exploración y los objetivos de comunicación que tengas por delante. ¡Cada visualización es una forma de contar una historia!

Con estos ejemplos, esperamos haberte introducido la importancia de visualizar tu dataset. Ahora, es mejor hacer una aclaración: no todos los datos se visualizan de la misma manera.

Algunas formas: gráficos de puntos y de dispersión

Existen varios tipos de gráficos que puedes usar, dependiendo de lo que busques destacar. Es importante que prestes atención al tipo de variable/s que quieras graficar. No es lo mismo:

- graficar la distribución de una variable numérica
- graficar la relación entre dos variables numéricas
- graficar la relación entre una variable numérica y una categórica.

A continuación, vamos a describir los tipos de gráficos más comunes con los que trabajaremos. Es importante que prestes atención a sobre qué tipo de variable operan y cuándo debes utilizarlos.

Los **gráficos de puntos y líneas** son los más comunes y representan una visualización de la relación entre un conjunto de puntos *(x, y)*, donde *x* e *y* son variables numéricas. Los gráficos de puntos solo muestran los puntos de datos, mientras que los gráficos de líneas los conectan o interpolan para definir una función continua f (x). Por ejemplo:

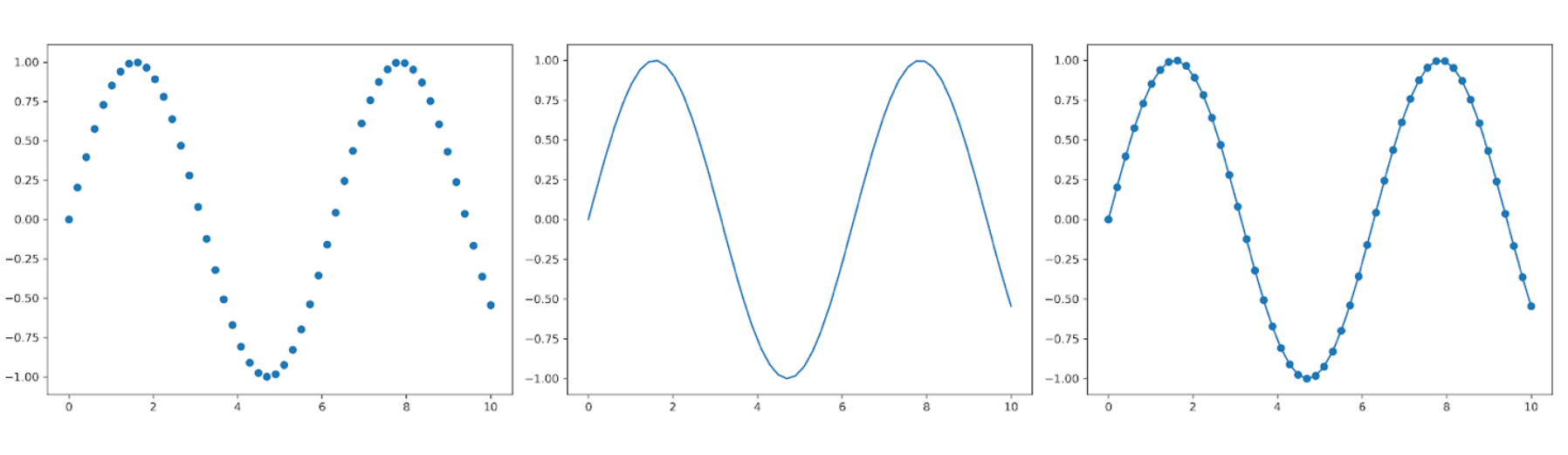


Figura 2.

Tres gráficos generados con el mismo conjunto de datos: uno de puntos (izquierda), uno de línea (centro) y superpuestos (derecha).

En estos ejemplos podemos observar estilos diferentes de gráficos de línea y de puntos, que varían en el grado de énfasis que le dan a los puntos contra la curva interpolada. Ten en cuenta que unir los puntos por líneas —es decir, interpolar en los lugares faltantes del nuestro dominio— proporciona una predicción para la función f(x) sobre el rango completo de x posible. Esto nos permite hacer referencia a valores que no fueron medidos.

Tip: Sin embargo, ¡mantente atento/a! La unión de puntos por líneas puede ser confuso, engañoso o directamente desprolijo en algunas situaciones. Por ejemplo, no tiene mucho sentido unir puntos cuando la función solamente está definida sobre números enteros. Si, además, los puntos no están correctamente ordenados o tienen mucho ruido, podemos terminar con líneas que van y vuelven de una forma muy desprolija, como ocurre en el siguiente ejemplo:

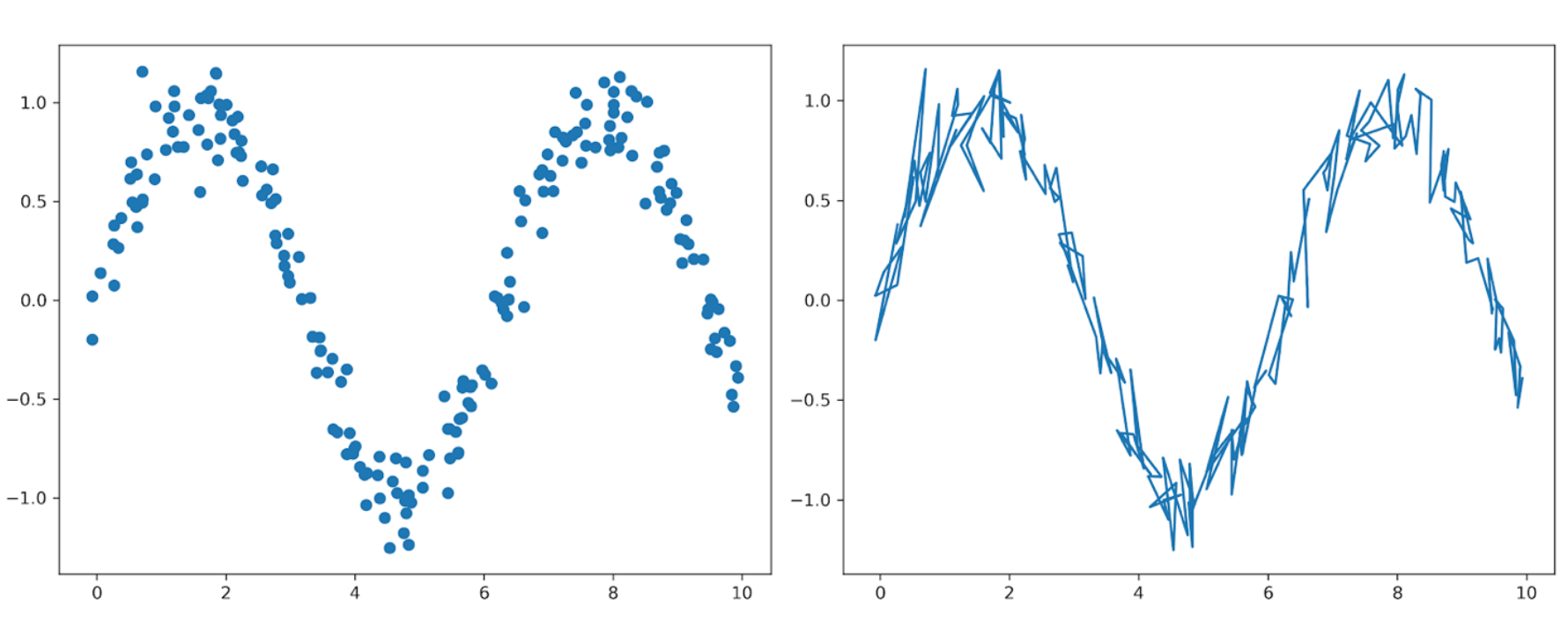


Figura 3.

Dos gráficos generados con el mismo conjunto de datos: uno de puntos (izquierda) y uno de líneas (derecho). ¿Cuál te parece mejor y por qué?

Otra forma de mostrar los datos —en particular para grandes conjuntos de datos — son los **gráficos de dispersión.** Un gran número de puntos abruman fácilmente las representaciones gráficas, lo que resulta en una imagen difícil de interpretar. Pero cuando se dibujan correctamente, los diagramas de dispersión son capaces de mostrar miles de puntos bivariados (bidimensionales) de una manera clara y comprensible. Los gráficos de dispersión muestran los valores de cada punto (x, y) de un conjunto de datos dado. Por ejemplo, este caso donde se muestran relacionados la edad y el peso de un conjunto de personas:

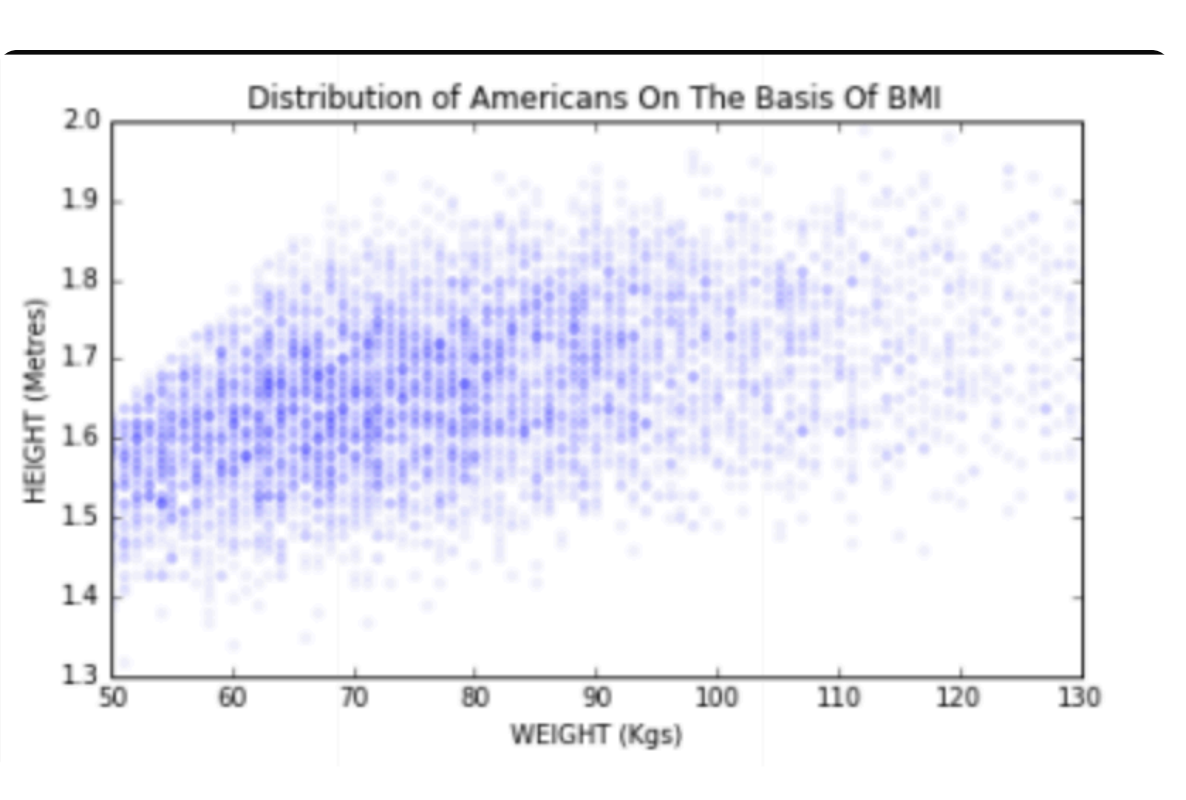


Figura 4. Fuente: Steve S. Skiena (2017): *The Data Science Design Manual*, Springer International Publishing.

¿Cuál es la diferencia entre un gráfico de puntos y un gráfico de dispersión? Si bien estos gráficos son, en el fondo, similares, en general se habla de un gráfico de puntos cuando existe una relación matemática entre las dos variables, mientras que se habla de un gráfico de dispersión cuando no está clara la relación entre esas variables y queremos empezar a explorarla.

Ejercicio: completa la siguiente ficha.

Los gráficos de puntos y líneas

operan sobre variables ....

y son útiles para ...

Una buena práctica es ...

Los histogramas

Cuando queremos explorar cómo están distribuidos los valores de una variable numérica, la mejor opción suelen ser los **histogramas.** Un histograma, en general, consiste en dividir el rango de posibles valores de una variable en rangos más pequeños, en general de la misma longitud, llamados *bines.*

Por ejemplo, si una variable puede tomar valores del 0 al 100, si dividimos en 10 bines estos serían: 0 al 10, 10 al 20, 20 al 30 y así hasta el último bin, 90 a 100. Luego, contamos cuántas instancias de nuestro conjunto de datos caen dentro de cada bin. Con esa información, construimos un gráfico de barras, como el de la siguiente figura:

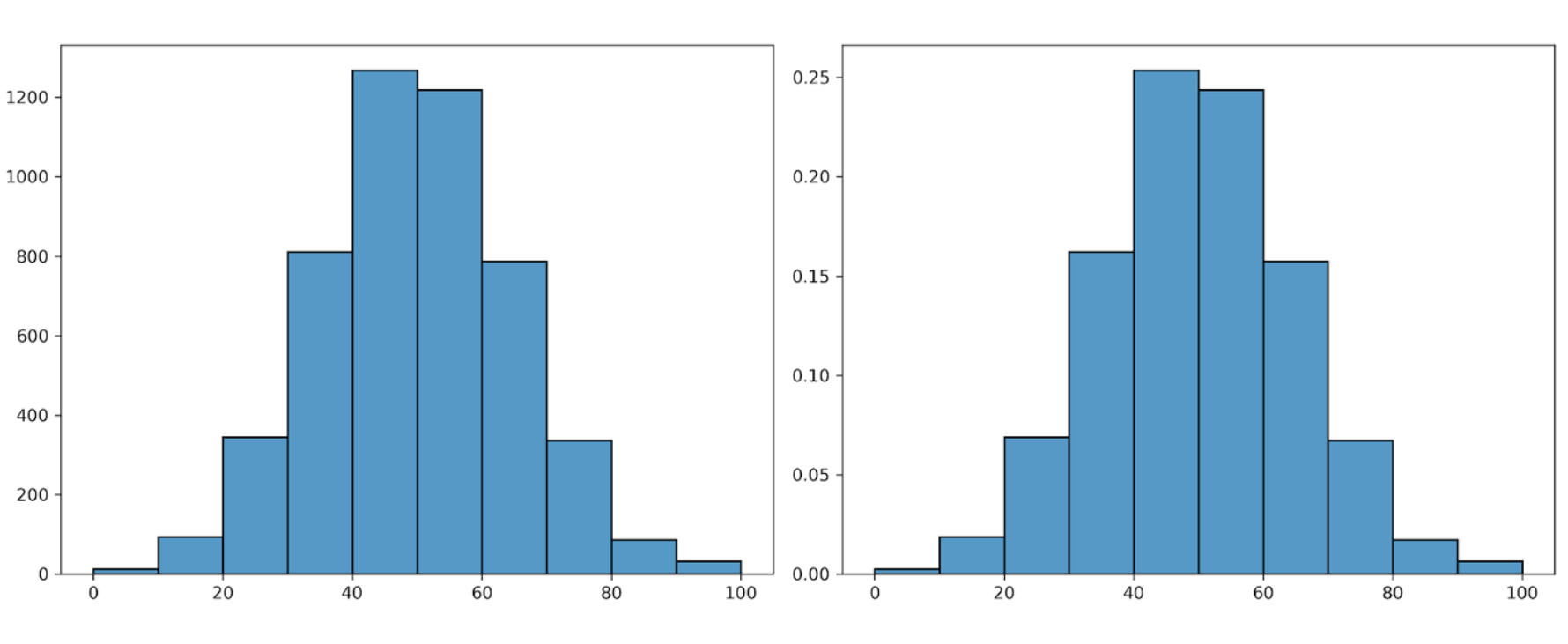


Figura 5.

¡Los histogramas y las distribuciones/densidades de probabilidad están fuertemente relacionados! Los histogramas son herramientas estadísticas que sirven para representar mediciones. Las distribuciones son herramientas probabilísticas que sirven para modelar lo que esperamos —idealmente— de una variable aleatoria. Entonces, **podemos pensar los histogramas como realizaciones de esas distribuciones de probabilidad.**

Existen histogramas de varios tipos: el más conocido es aquel en el que contamos la cantidad de instancias que caen en cada bin, y en el que hacemos un gráfico de frecuencias, es decir, dividimos por la cantidad de muestras totales. De esta forma, la altura de cada histograma podría ser interpretada como la probabilidad de obtener una muestra de ese bin.

Algunas preguntas interesantes sobre una variable son fáciles de responder con su histograma: ¿qué forma tiene la distribución? ¿es simétrica o asimétrica? ¿existen valores muy diferentes al resto (valores atípicos)? ¿cuál es la moda o qué valor ronda? ¿hay más de una moda?

TIP: a la hora de construir un histograma es importante prestar atención y decidir el número correcto de bins para usar. Si usas demasiados, solo habrá unos pocos puntos en cada bin. Si usas pocos, no verás con suficiente detalle para comprender la forma de la distribución. Aquí va un ejemplo:

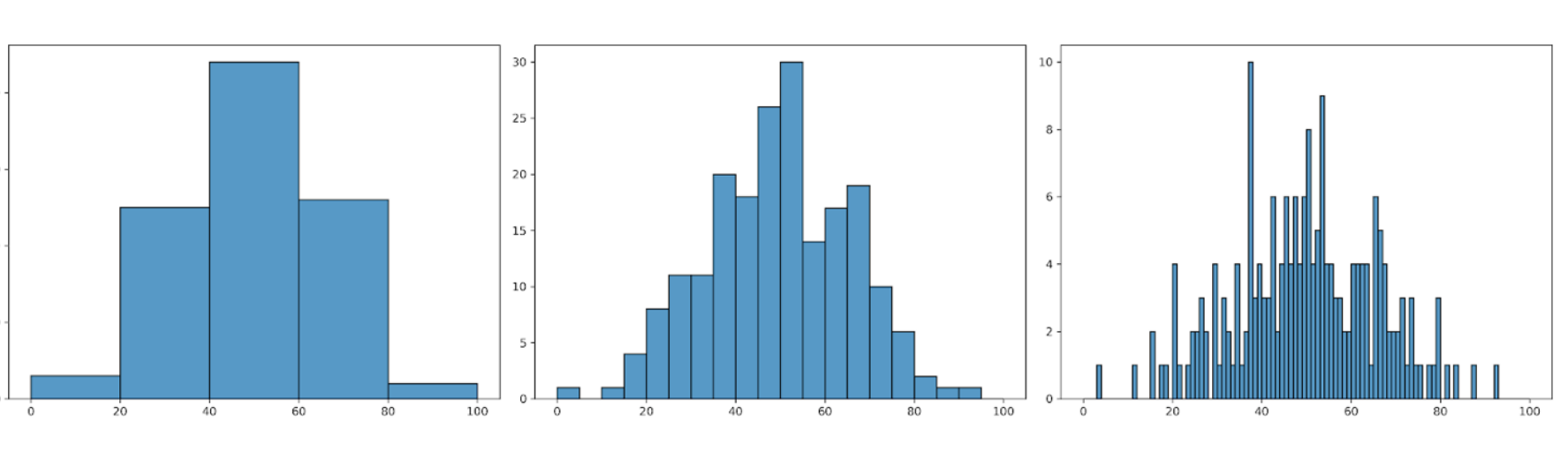


Figura 6.

Histogramas de las mismas 200 muestras, en 5 (izquierda), 20 (centro) y 100 (derecha) bines.

En general, nunca vamos a poder discriminar entre más de 100 bins a simple vista, lo que proporciona un límite superior lógico. Te recomendamos experimentar con diferentes valores y elegir el que mejor te resulte.

Ejercicio: completa la siguiente ficha.

Los histogramas

operan sobre variables ...,

y son útiles para ....

Una buena práctica es ...

Estos son sólo algunos tipos de visualizaciones, ¿en qué otras formas podríamos pensar?

Herramientas para todo: Matplotlib

Como ya te contamos en las bitácoras anteriores, todos los conceptos matemáticos y/o herramientas que vamos explorando para mejorar el análisis tienen una contraparte dentro de una librería de Python. Durante la cursada, trabajaremos con Matplotlib, la librería por defecto para graficar en Python. Como la mayoría de las librerías que venimos usando, cuenta con una excelente documentación, que encontrarás disponible para consultar en la sección "Herramientas". Además, Matplotlib es la base de otras librerías para graficar, como Seaborn, sobre la que te contaremos en la próxima bitácora.

Empieza a trabajar en el notebook

Trabaja en el siguiente notebook, ya sea en Colab o en tu computadora. Para el encuentro, lleva hecha las secciones 1 y 2, el resto del notebook lo trabajaremos en clase.

Notebook: DS_Bitácora_06_Matplotlib

Puede descargar el dataset aquí. Si deseas trabajar en Colab, deberás subir el dataset. ¡Inténtalo, no es difícil!

¡Prepárate para el próximo encuentro!

Profundiza
Te invitamos a conocer más sobre el tema de esta bitácora.

Rincón Matemático
¡¡¿¿Matemática??!! No te asustes... aquí tienes herramientas para revisar el contenido necesario para sobrevivir.

Challenge
Te proponemos el siguiente desafío, ¿te animas?

Contenido extra de la meeting

Clase 19 feb reemplazo clase martes 16
Tema: mentores92 Acámica's Zoom Meeting Hora de inicio de la reunión : 19 feb. 2021 07:07 Grabación de la reunión: https://acamica.zoom.us/rec/share/5l493tekRtIsSg4hUEqs_2npm0wo0afwPhajP0UoB2j61ORIC5ZQKXyNdyqe9Bz-.bQc0mFyABLoiYLqA