Introducción a Machine Learning

Delta Analytics construye capacidad técnica alrededor del mundo.

El contenido de este curso está siendo desarrollado activamente por Delta Analytics, una organización sin fines de lucro 501(c)3 del Área de la Bahía que apunta a capacitar a las comunidades para aprovechar sus datos.

Por favor comuníquese con cualquier pregunta o comentario a inquiry@deltanalytics.org.

Descubre más sobre nuestra misión aquí.

Resumen del Curso:

- ✓ Módulo 1: Introducción a Machine Learning
- ❑ Módulo 2: Machine Learning en Profundidad
- ❑ Módulo 3: Selección y Evaluación del Modelo
- ❑ Módulo 4: Regresión Lineal
- ❑ Módulo 5: Árboles de Decisión
- ❑ Módulo 6: Algoritmos de Conjunto
- ❑ Módulo 7: Algoritmos de Aprendizaje no Supervisados
- ❑ Módulo 8: Procesamiento del Lenguaje Natural Parte 1
- ❑ Módulo 9: Procesamiento del Lenguaje Natural Parte 2

Ahora pasemos a los datos que usaremos ...

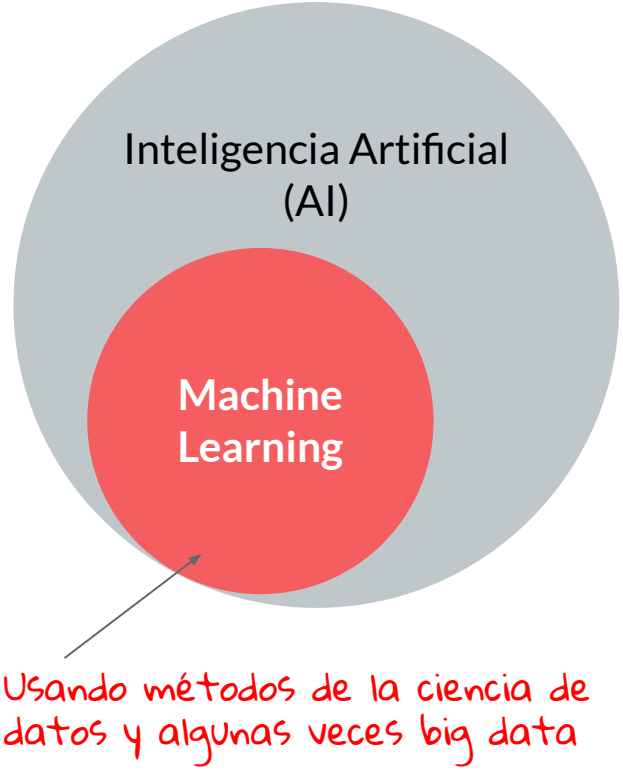
Módulo 1: Introducción a Machine Learning

Checklist del Módulo:

- ✓ ¿Qué es machine learning?
- ❑ ¿Cómo se define una pregunta de investigación?
- ❑ ¿Qué son las observaciones?
- ❑ ¿Qué son las características?
- ❑ ¿Qué son las variables de salida?
- ❑ Introducción a los datos de KIVA

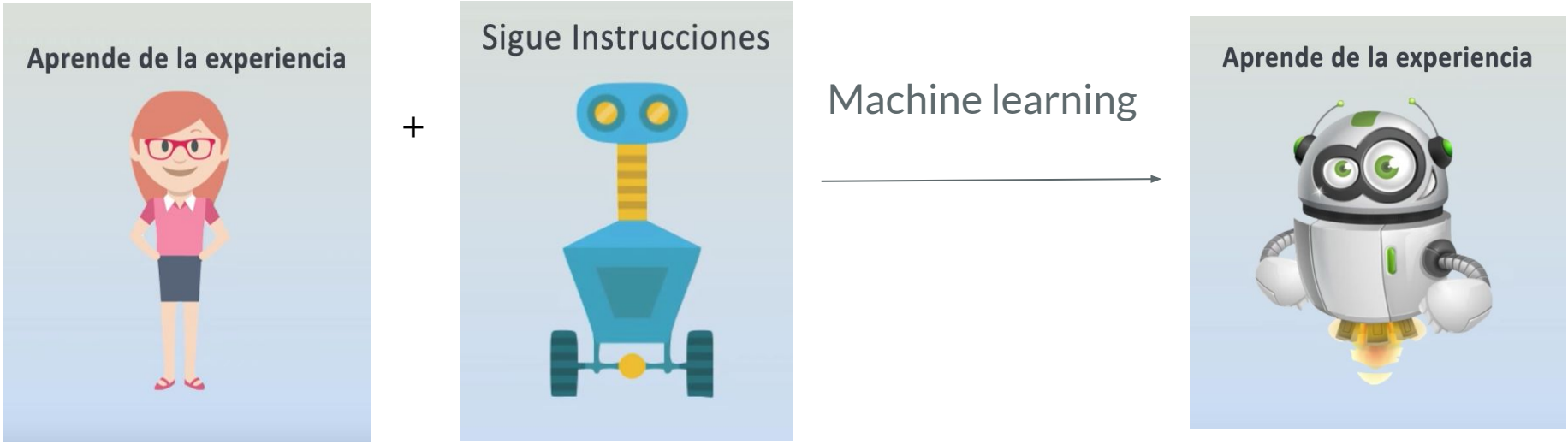
¿Qué es machine learning?

¿Qué es Machine Learning?



Llamamos a algo **machine learning** o aprendizaje automático cuando, en lugar de decirle a una computadora que haga algo, permitimos que una computadora cree su propia solución basada en los datos que se le proporcionan.

Machine learning es un subconjunto de la IA que permite a las máquinas **aprender de datos sin procesar.**



Los humanos aprenden de la experiencia. La programación de software tradicional implica dar a las máquinas instrucciones para realizar. **Machine learning implica permitir que las máquinas aprendan de datos en bruto para que el programa pueda cambiar cuando se expone a nuevos datos** (aprendiendo de la experiencia).

Fuente: https://www.youtube.com/watch?v=IpGxLWOIZy4

Machine learning es interdisciplinario

Machine learning es…

- Ciencias computacionales + estadísticas + matemáticas
- El uso de datos para **responder preguntas**

Pensamiento crítico combinado con herramientas técnicas

Hay una necesidad creciente por machine learning

- Hay una enorme cantidad de datos generados todos los días.
- Los problemas anteriormente imposibles ahora son solucionables.
- Las empresas exigen cada vez más soluciones cuantitativas.

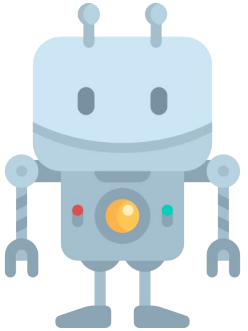
*"Todos los días, creamos **2.5 quintillones de bytes de datos** — tanto que el **90% de los datos en el mundo de hoy se han creado sólo en los últimos dos años.**"* [1]

Fuente:
[1] "What is Big Data," IBM,

Ejemplo de machine learning: Predicción de la malaria

La Dra. Delta trabaja para diagnosticar pacientes con malaria. Sin embargo, le lleva mucho tiempo ver a todos.

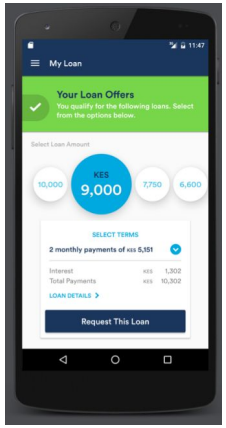
Afortunadamente, la Dra. Delta tiene **datos históricos de pacientes sobre qué factores predicen la malaria**, como la temperatura corporal, historial de viaje, edad, e historial médico.



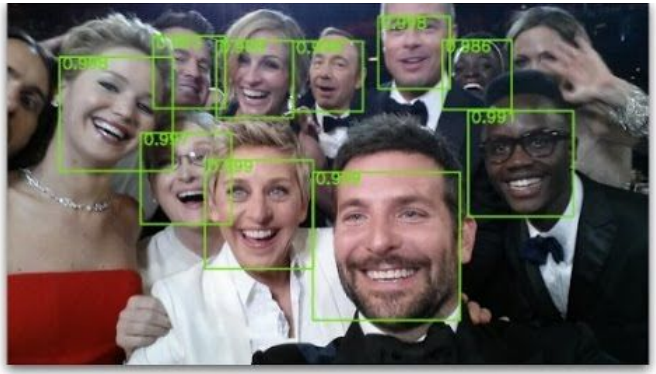
La Dra. Delta puede usar datos históricos como una entrada en un **algoritmo de machine learning** para ayudarle a predecir si un nuevo paciente tendrá malaria.

El algoritmo (la máquina) aprende de datos pasados, como lo haría un ser humano, y por lo tanto es capaz de hacer predicciones sobre el futuro.

Machine learning es una herramienta poderosa; puede...



Determina tu calificación crediticia basada en el uso del teléfono celular.

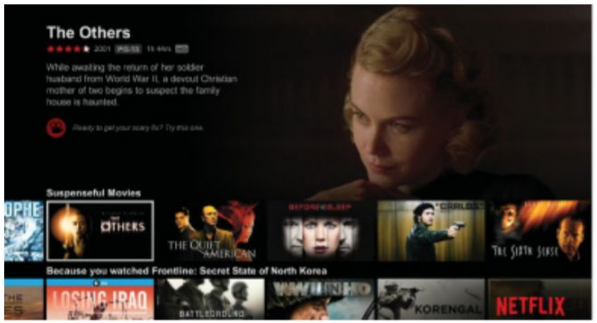


Reconoce tu rostro en una foto.

Emmanuel Macron Is Inaugurated as French President

Ceremony comes a week after victory over Marine Le Pen in presidential election

Determina el tema en un fragmento de texto.



Recomienda películas que gustarán.

Machine learning nos ayuda a responder preguntas. **¿Cómo definimos la pregunta?**

Antes de que lleguemos a los modelos/algoritmos, tenemos que aprender sobre nuestros datos y definir nuestra pregunta de investigación.

El científico de datos pasa ~ 80% de su tiempo aquí: preparando los datos para el análisis

Machine learning tiene lugar durante la fase de modelado.

Pregunta de Investigación

Una pregunta de investigación es la pregunta que queremos que nuestro modelo responda.

Ejemplos de preguntas de investigación:

- ¿Este paciente tiene malaria?
- ¿Podemos controlar la deforestación ilegal al detectar ruidos de motosierra en el audio transmitido desde las selvas tropicales?

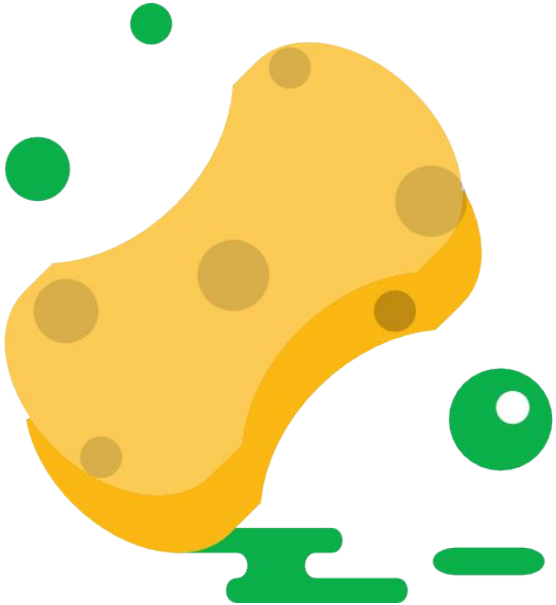
Es posible que ya tengamos una pregunta en mente antes de ver los datos, pero a menudo usaremos nuestra exploración de los datos para desarrollar o refinar nuestra pregunta de investigación.

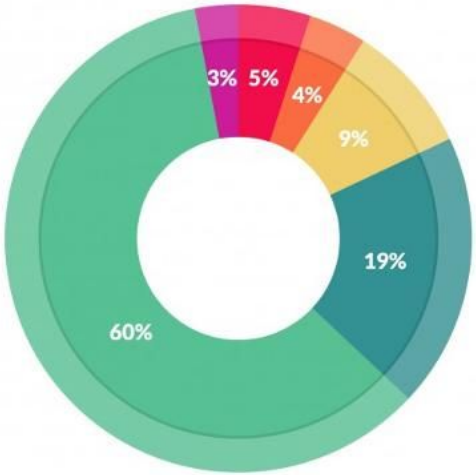
¿Qué fue primero, el huevo o la gallina?

Validación y Limpieza de los Datos

Limpieza de los Datos

"La preparación de datos representa aproximadamente el **80%** del trabajo de los científicos de datos."





¿Qué es lo que los científicos de datos pasan más tiempo haciendo?

- *Construyendo conjuntos de entrenamiento: 3%*
- *Limpiando y organizando datos: 60%*
- *Recolectando conjuntos de datos: 19%*
- *Minería de datos para patrones: 9%*
- *Refinando algoritmos: 4%*
- *Otro: 5%*

Fuente: Survey of 80 data scientists. Artículo de Forbes, 23 de marzo, 2016.

Limpieza de los Datos

¿Por qué necesitamos validar y limpiar nuestros datos?

Los datos a menudo provienen de múltiples fuentes

- ¿Se alinean los datos a través de diferentes fuentes?

Los datos son creados por humanos

- ¿Es necesario transformar los datos?
- ¿Están libre de sesgos y errores humanos?

Ir más lejos con estas lecturas: aquí y aquí

Limpieza de los Datos

La limpieza de datos implica identificar cualquier problema con nuestros datos y confirmar nuestra comprensión cualitativa de ellos.

Datos Faltantes
¿Hay datos faltantes?
¿Faltan sistemáticamente?

Tipos de Datos
¿Son todas las variables del tipo correcto?¿Una fecha es tratada como una fecha?

Validación de series de tiempo
¿Son los datos para el rango de tiempo correcto?
¿Hay máximos inusuales en el volumen de *préstamos* a lo largo del tiempo?

Rango de Datos
¿Están todos los valores en el rango esperado?
¿Todos los *préstamos* son mayores que 0?

Limpieza de los Datos

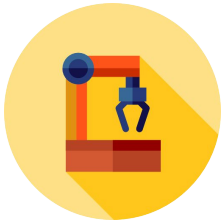
Veamos algunos ejemplos:

Limpieza de los Datos

- Datos faltantes

- Series de tiempo

- Tipos de datos
- Transformación de variables



Después de obtener una comprensión inicial de tus datos, es posible que debas **transformarlos** para utilizarlos en el análisis.

¿Hay datos faltantes? ¿Los datos faltan al azar o sistemáticamente?

Muy pocos conjuntos de datos tienen datos completos; La mayoría de las veces tendrás que lidiar con los datos faltantes.

La primera pregunta que debes hacer es qué tipo de datos faltantes tienes.

Faltan completamente al azar: No hay patrón en los datos faltantes. Este es el mejor tipo de falta que puedes esperar.

Faltan al azar: Hay un patrón en tus datos faltantes pero no en tus variables de interés.

No faltan al azar: hay un patrón en los datos faltantes que afecta sistemáticamente a tus variables primarias.

¿Hay datos faltantes? ¿Los datos faltan al azar o sistemáticamente?

Ejemplo: tienes datos de una encuesta de una muestra aleatoria de estudiantes de secundaria en los EE. UU. Algunos estudiantes no participaron:

Algunos alumnos se enfermaron el día de la encuesta.

Si faltan datos al azar, podemos usar el resto de los datos que no faltan sin preocuparnos por el sesgo

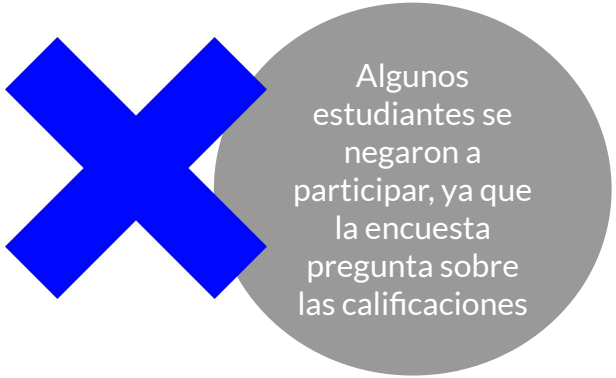
Algunos estudiantes se negaron a participar, ya que la encuesta pregunta sobre las calificaciones

Si faltan datos de forma no aleatoria o sistemática, tus datos no faltantes pueden estar sesgados

¿Hay datos faltantes? ¿Los datos faltan al azar o sistemáticamente?

Ejemplo: tienes datos de una encuesta de una muestra aleatoria de estudiantes de secundaria en los EE. UU. Algunos estudiantes no participaron:

Si faltan datos al azar, podemos usar el resto de los datos que no faltan sin preocuparnos por el sesgo



Si faltan datos de forma no aleatoria o sistemática, tus datos no faltantes pueden estar sesgados

A veces, puedes **reemplazar** los datos faltantes.

- Elimina observaciones faltantes.
- Rellena los valores faltantes con el promedio de los datos disponibles.
- Imputa datos

Lo que debes hacer depende en gran medida de lo que tenga sentido para la pregunta de investigación y los datos.

Técnicas comunes de imputación

Usa el promedio de los valores no faltantes	Toma el promedio de observaciones que tienes para completar las observaciones faltantes, es decir, puedes asumir que esta observación también está representada por el promedio de la población.
Usa una estimación fundamentada	Suena arbitrario y, a menudo, no se prefiere, pero puedes inferir un valor faltante. Para preguntas relacionadas, por ejemplo, como las que se presentan a menudo en una matriz, si el participante responde todo con la opción "4", asume que el valor que falta es un 4.
Utiliza la imputación de puntos comunes	Para una escala de calificación, utiliza el punto medio o el valor más comúnmente elegido. Por ejemplo, en una escala de cinco puntos, sustituye con un 3, el punto medio o un 4, el valor más común (en muchos casos). Esto es un poco más estructurado que adivinar, pero aún se encuentra entre las opciones más riesgosas. Se debe tener cuidado, a menos que se tenga una buena razón y datos para respaldar el uso del valor sustituto.

Fuente: Handling missing data

Si tenemos observaciones a lo largo del tiempo, debemos hacer validación de series de tiempo.

Pregúntate:

a. ¿Es el rango de tiempo correcto para los datos?
b. ¿Hay máximos inusuales en los datos a lo largo del tiempo?

¿Qué deberíamos hacer si hay máximos inusuales en los datos a lo largo del tiempo?

¿Cómo abordamos los máximos inesperados en nuestros datos?

Anomalía de datos

Máximo sistemático

Para ciertos conjuntos de datos (como datos de ventas) se esperan máximos estacionales sistemáticos. Por ejemplo, alrededor de Navidad veríamos un aumento en el lugar de ventas. Esto es normal, y no necesariamente debe ser eliminado.

Máximo aleatorio

Si el máximos es aislado, es probable que sea inesperado, es posible que queramos eliminar los datos corruptos. Por ejemplo, si durante un mes las ventas se registran en pesos chilenos en lugar de dólares estadounidenses, se inflarían las cifras de ventas. Deberíamos realizar una limpieza de datos convirtiendo a $ o tal vez eliminar este mes.

Ten en cuenta que, a veces, existen anomalías naturales en los datos que deben investigarse primero.

¿Son todas las variables del tipo correcto?

Muchas funciones en Python son de un tipo específico, lo que significa que debemos asegurarnos de que todos nuestros campos se traten como el tipo correcto:

entero float string fecha

	loan_amount	partner_id	sector	posted_date
1957	50	156.0	Personal Use	2017-04-11
78437	350	133.0	Clothing	2013-08-07
116723	575	156.0	Agriculture	2011-01-04

- **entero:** Un número sin decimales
- **Float**: Un número con decimales
- **String**: Campo de texto, o más formalmente, una secuencia de caracteres unicode
- **Booleano**: Solo puede ser Verdadero o Falso (también llamado indicador o variable ficticia)
- **Fecha y hora**: Valores destinados a contener datos de tiempo

Prueba sobre limpieza de datos!

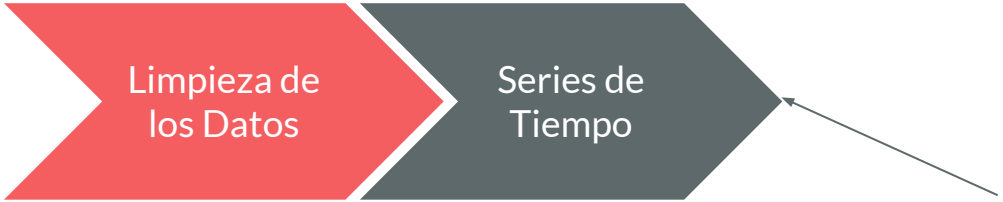
A medida que exploras los datos, surgen algunas preguntas…

Pregunta #1

Pregunta	Respuesta
Hay una observación del conjunto de datos de préstamos de KIVA que dice que un préstamo fue financiado en su totalidad en el año 1804, pero Kiva ni siquiera fue fundada en ese momento. ¿Qué debo hacer?	

Pregunta #1

Pregunta	Respuesta
Hay una observación del conjunto de datos de préstamos de KIVA que dice que un préstamo fue financiado en su totalidad en el año 1804, pero Kiva ni siquiera fue fundada en ese momento. ¿Qué debo hacer?	Consulte la documentación de datos. Si no existe una explicación, elimine esta observación.



Esta pregunta ilustra que siempre debes hacer la validación del rango de tiempo. Comprueba cuáles son las observaciones mínimas y máximas en tu conjunto de datos.

Pregunta #2

Pregunta	Respuesta
Hay una observación que indica que el cumpleaños de una persona es el 12/1/80 pero que falta la variable "edad". ¿Qué debo hacer?	

Pregunta #2

Pregunta	Respuesta
Hay una observación que indica que el cumpleaños de una persona es el 12/1/80 pero que falta la variable "edad". ¿Qué debo hacer?	Podemos calcular la edad: (ej: 2017 - 1980 = 37)

Esta pregunta ilustra cómo podemos ser capaces de aprovechar otros campos para hacer una conjetura acerca de la edad que falta.

Pregunta #3

Pregunta	Respuesta
La variable "amount_funded" tiene valores de "N / A" y "0". ¿Qué debo hacer?	

A medida que exploras los datos, surgen algunas preguntas ...

Pregunta #3

Pregunta	Respuesta
La variable "amount_funded" tiene valores de "N / A" y "0". ¿Qué debo hacer?	Verifica la documentación si hay una diferencia importante entre NA y 0.

Pregunta #4

Pregunta	Respuesta
No estoy seguro en qué moneda se reporta la variable "amount_funded". ¿Qué debo hacer?	

Pregunta #4

Pregunta	Respuesta
No estoy seguro en qué moneda se reporta la variable "amount_funded". ¿Qué debo hacer?	Consulta la documentación y otras variables, convierte a la moneda apropiada.

Una nota final...

Ten en cuenta que nuestros ejemplos fueron todos muy específicos: puedes o no encontrar estos ejemplos exactos en la naturaleza. Esto se debe a que la limpieza de datos a menudo es particular a esos datos y **no se puede completar adecuadamente siguiendo un conjunto predeterminado de pasos: ¡debes usar el sentido común!**

A continuación, pasamos al análisis exploratorio, para el cual a menudo tenemos que **transformar nuestros datos**.

Análisis Exploratorio

Análisis Exploratorio

El objetivo del análisis exploratorio es comprender mejor tus datos.

El análisis exploratorio puede revelar limitaciones en los datos, qué características son importantes e informar qué métodos usar para responder la pregunta de investigación.

¡Este es un primer paso indispensable en cualquier análisis de datos!

Let's explore our data!

Una vez que hemos realizado una validación inicial, exploramos los datos para ver qué modelos son adecuados y qué patrones podemos identificar.

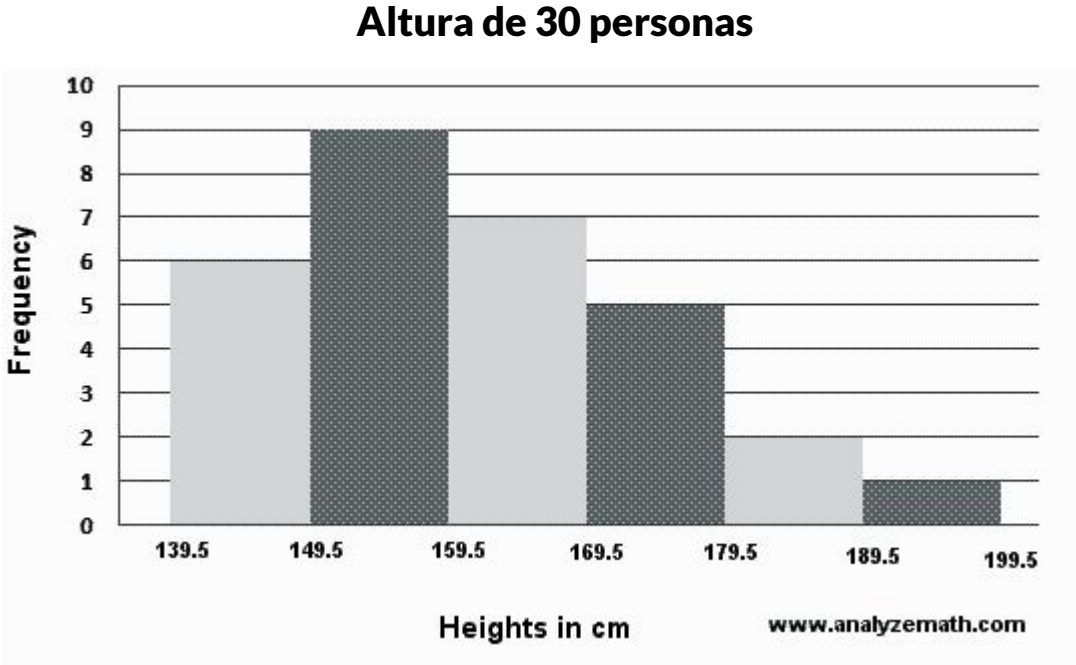
El proceso varía en función de los datos, tu estilo y las restricciones de tiempo, pero normalmente la exploración incluye:

- Histograma
- Gráficos de dispersión
- Tablas de correlación
- Diagramas de caja
- Resumen estadístico
  - Promedio, mediana, frecuencia

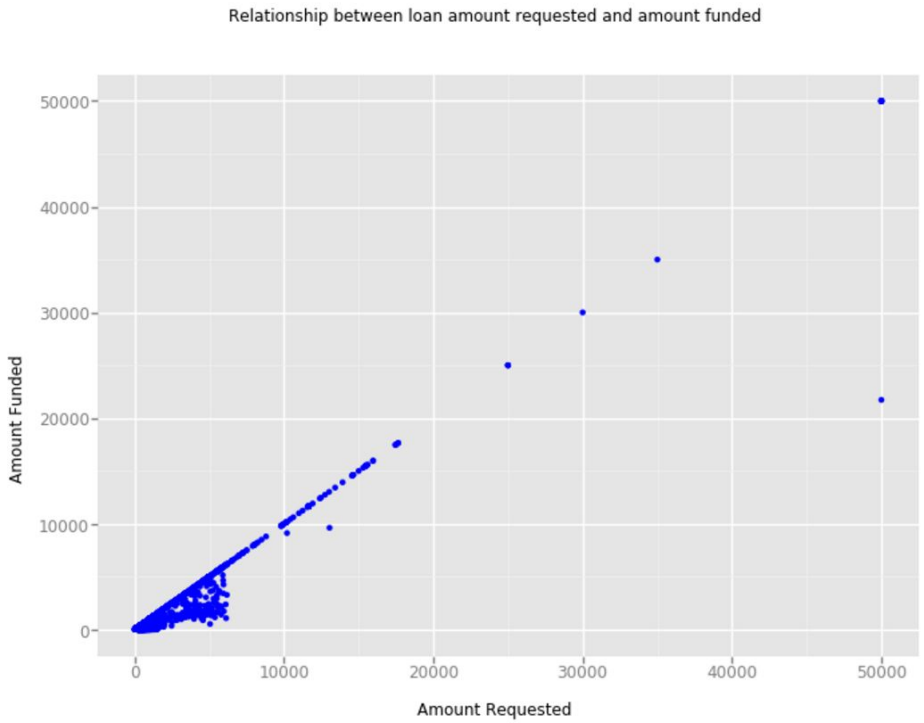
Los histogramas nos informan sobre la distribución de la característica.

Un histograma muestra la **distribución de frecuencia** de una característica continua.

Aquí, tenemos datos de altura de un grupo de personas. Vemos que la mayoría de las personas en el grupo tienen entre 149 y 159 cm de altura.



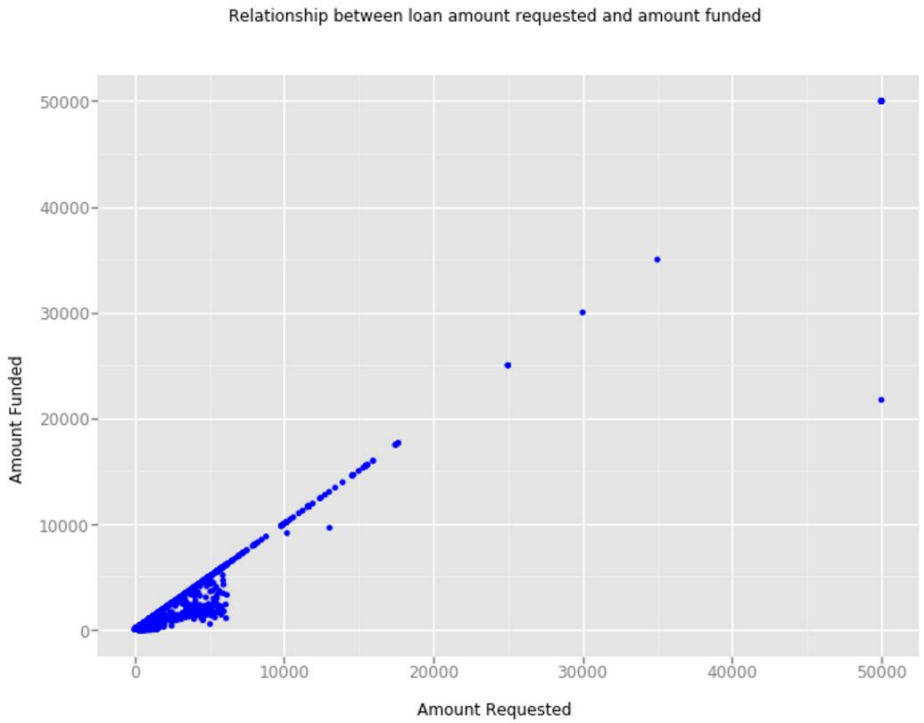
Los gráficos de dispersión proporcionan información sobre la relación entre dos características.



Los diagramas de dispersión visualizan las relaciones entre dos características como puntos en un gráfico. Son un primer paso útil para explorar una pregunta de investigación.

Aquí, ya podemos ver una relación positiva entre la cantidad financiada y la cantidad solicitada. ***¿Qué podemos concluir?***

Los gráficos de dispersión proporcionan información sobre la relación entre dos características.

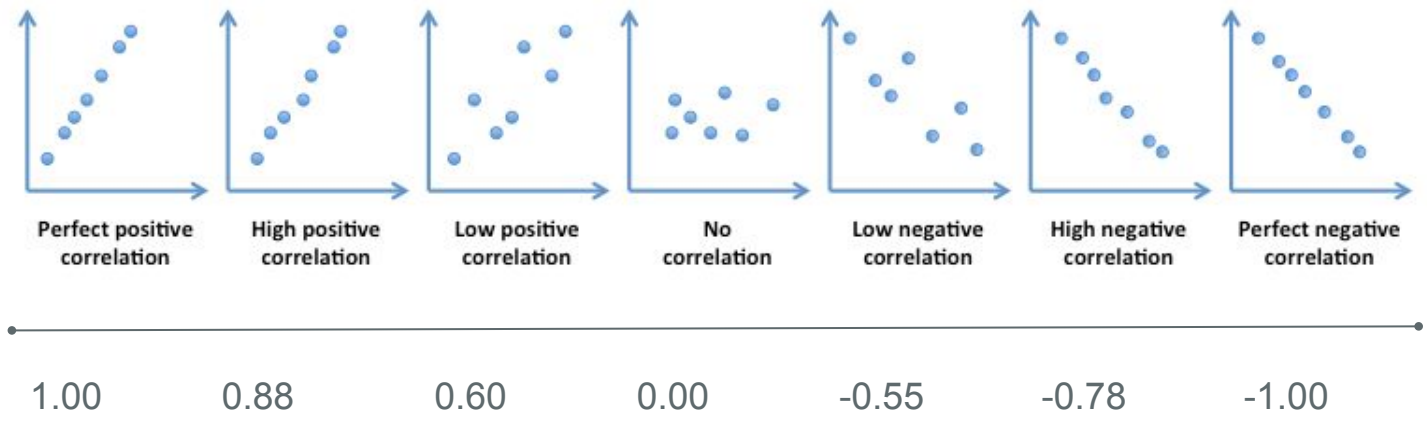


Los diagramas de dispersión visualizan las relaciones entre dos características como puntos en un gráfico. Son un primer paso útil para explorar una pregunta de investigación.

Aquí, ya podemos ver una relación positiva entre la cantidad financiada y la cantidad solicitada. ***¿Qué podemos concluir?***

Parece que existe una relación sólida entre el monto del préstamo que se solicita y lo que se financia.

La correlación es una medida útil de la fuerza de una relación entre dos variables. Va desde -1.00 a 1.00



Ve más lejos con este entretenido juego.

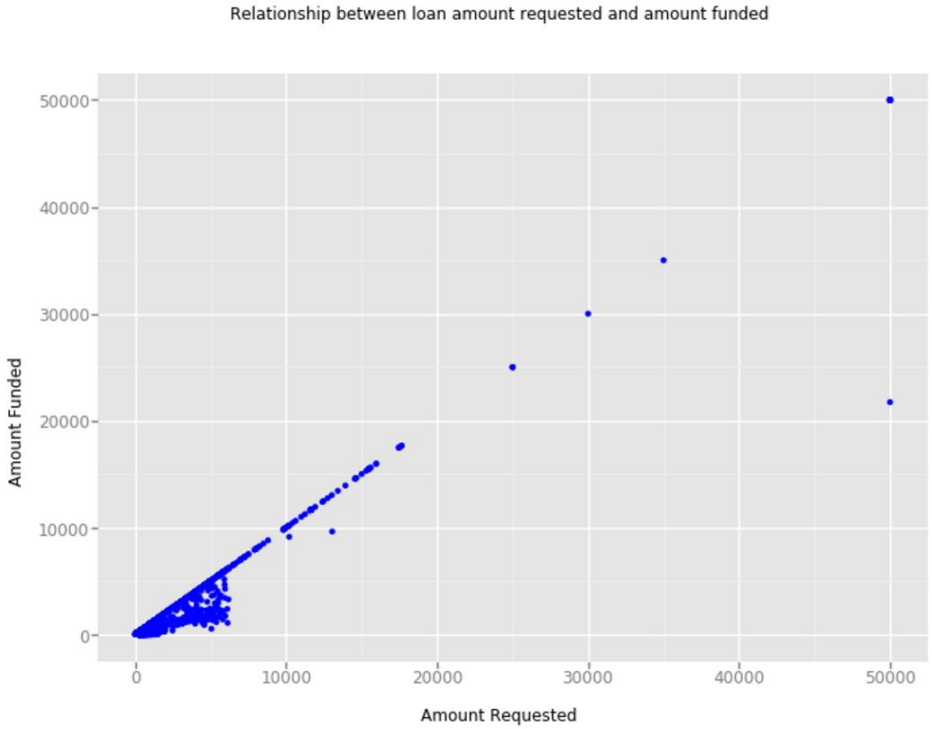
La correlación no es igual a la causalidad

Digamos que eres un ejecutivo de una empresa. Has recopilado los siguientes datos:

X = $ gastado en publicidad
Y= Ventas

Según el gráfico y la correlación positiva, estaría tentado de decir que los $ gastados en publicidad causaron un aumento en las ventas. **Pero espera**, ¡también es posible que un aumento en las ventas (y por lo tanto, las ganancias) conduzca a un aumento en $ gastado en publicidad! ***La correlación entre x e y no significa que x causa y; podría significar que y causa x!***

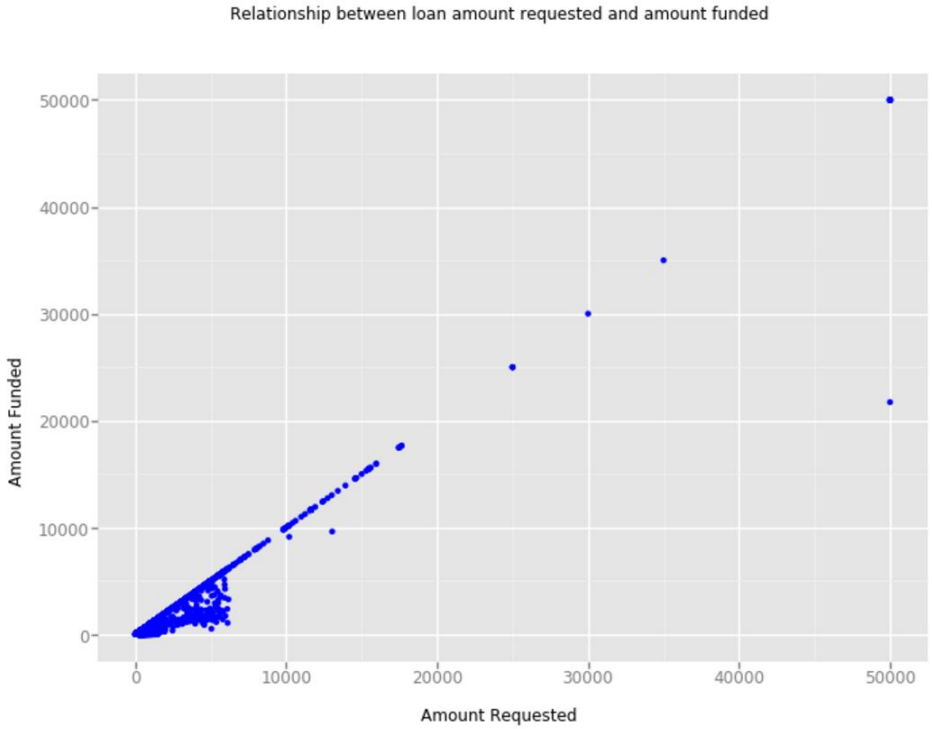
Ejemplo de Kiva: la correlación no es igual a la causación



Correlación: 0.96

Si deseas solicitar un préstamo a través de Kiva, y solo se te mostró este gráfico, puedes concluir que es una buena idea solicitar $ 1 millón de dólares.

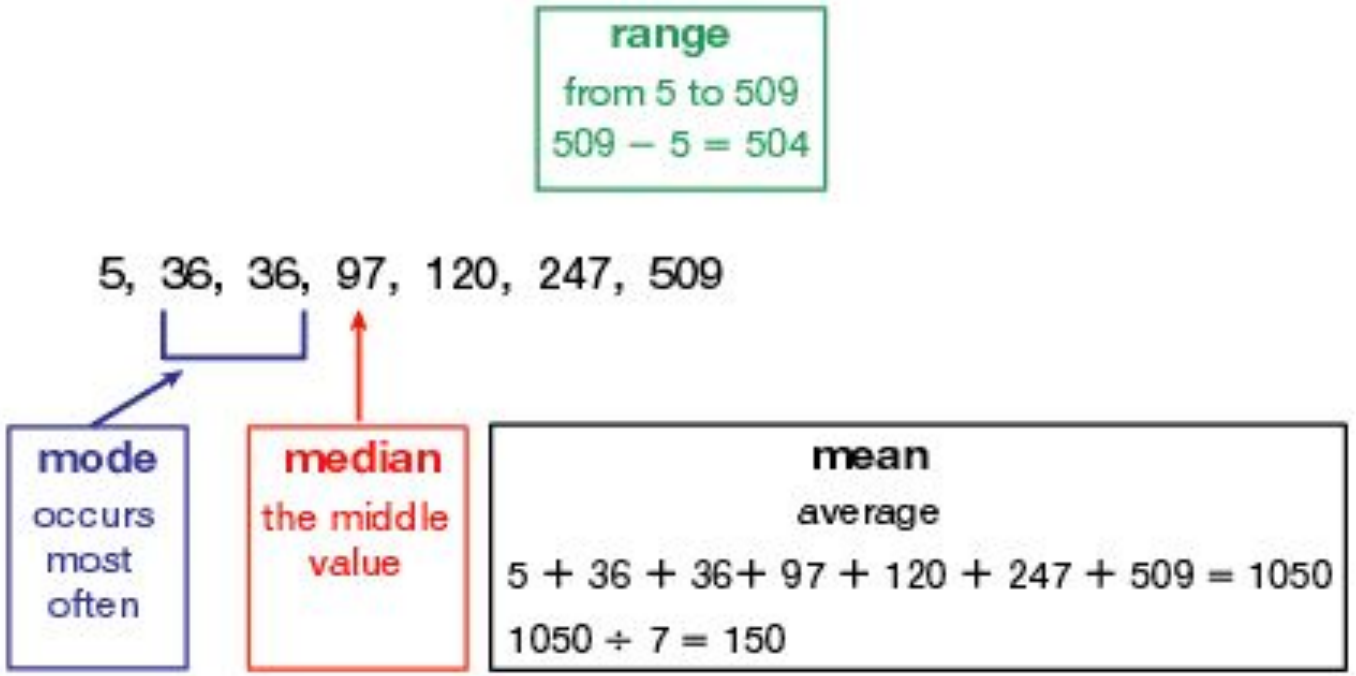
Ejemplo de Kiva: la correlación no es igual a la causación



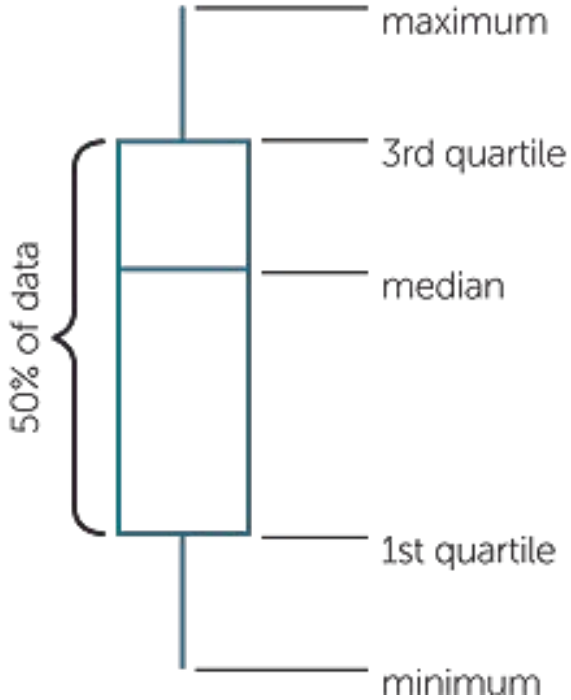
Pero el sentido común nos dice que esta conclusión no tiene mucho sentido. ¡Solo porque solicites mucho no significa que recibirás muchos fondos!

¡Las conclusiones pueden ser inválidas incluso cuando los datos son válidos!

El promedio, la mediana y la frecuencia son estadísticas de resumen útiles que te permiten saber qué hay en tus datos.



Los diagramas de caja son una representación visual útil de ciertas estadísticas de resumen.



Fuente de la imagen: Universidad de Florida, Quantitative introduction to the boxplot

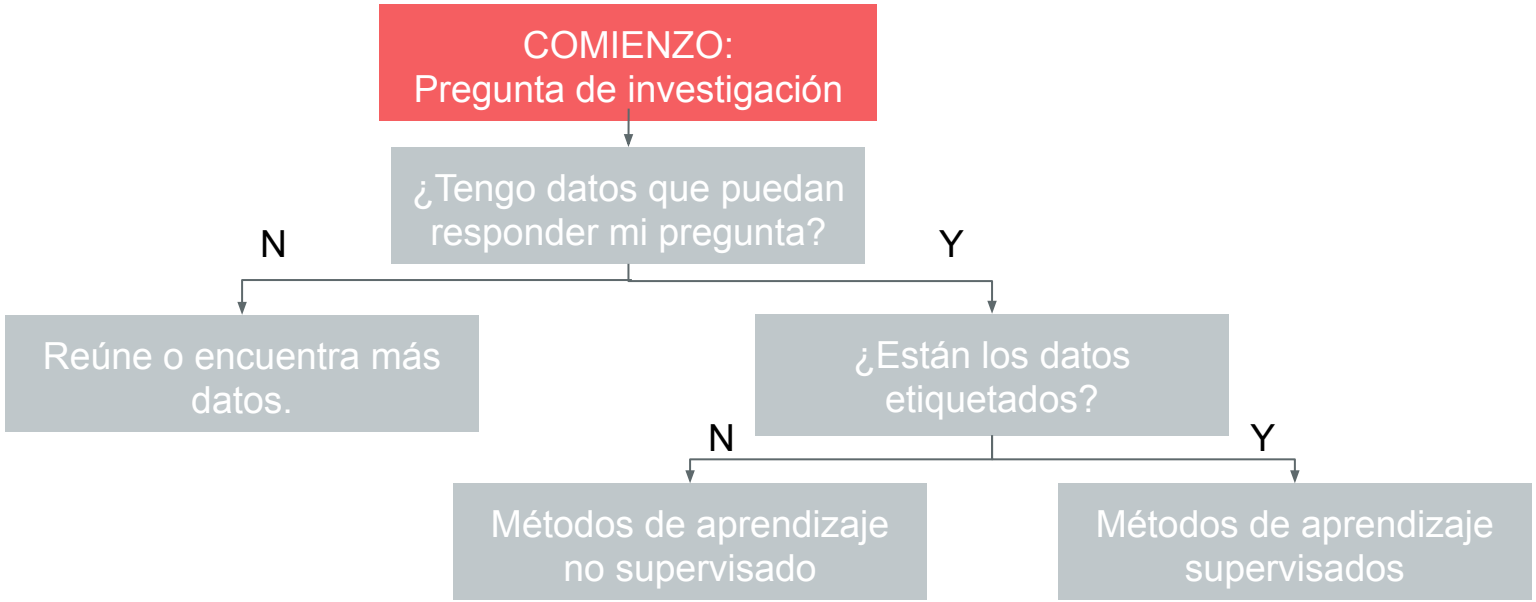
Elaborando una pregunta de investigación

Recuerda: es posible que tengamos una pregunta en mente antes de ver los datos, pero nuestra exploración de los datos a menudo desarrolla o refina nuestra pregunta de investigación.

¿Qué viene primero, el huevo o la gallina?

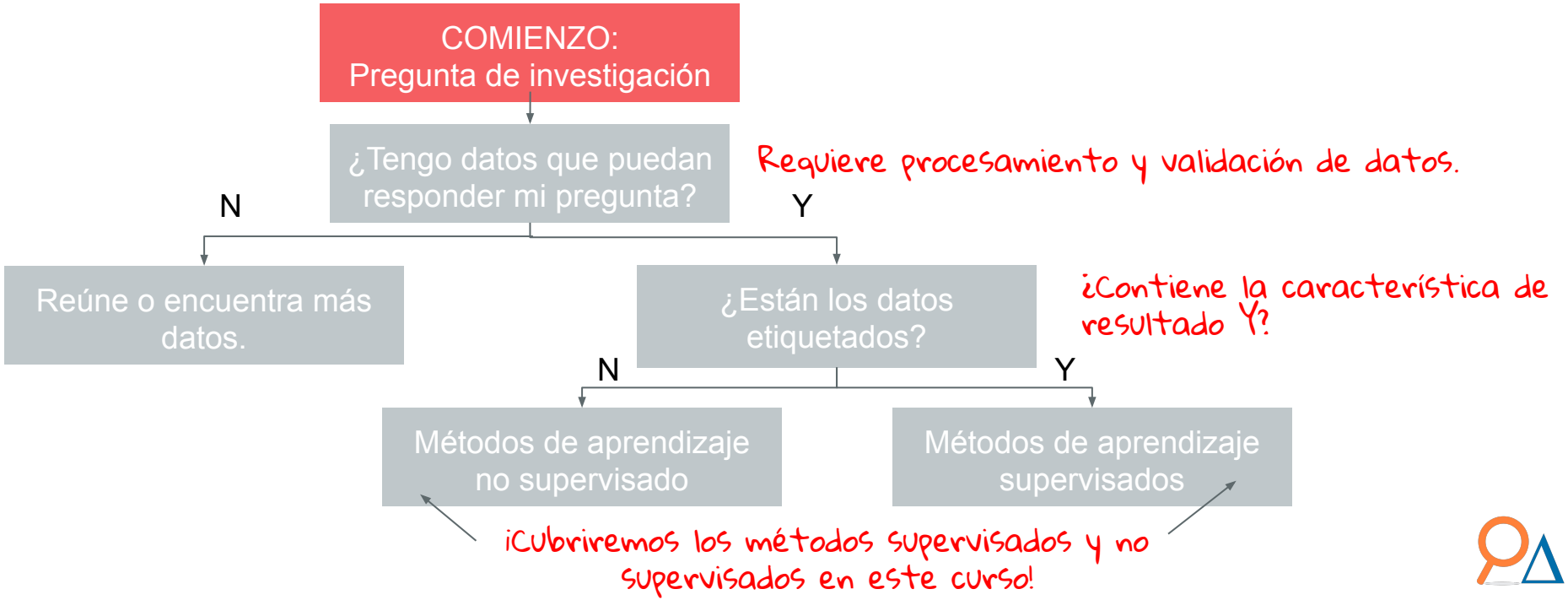
¿Cómo se define la pregunta de investigación?

- Hacemos una pregunta que esperamos que los datos respondan. *¿Qué viene primero, los datos o la pregunta?*



¿Cómo se define la pregunta de investigación?

- Hacemos una pregunta que esperamos que los datos respondan. *¿Qué viene primero, los datos o la pregunta?*



Pregunta de Investigación

Dados los datos de KIVA a continuación, podemos encontrar algunas preguntas interesantes.

Monto del préstamo solicitado por un prestatario de Kiva en Kenia

Ciudad en la que el solicitante del préstamo reside.

s	lender_count	loan_amount	location.country	location.country_code	location.geo.level	location.geo.pairs	location.geo.type	location.town	
	7	225	Kenya	KE	town	-1.166667 36.833333	point	Kiambu	J
	14	350	Kenya	KE	town	0.516667 35.283333	point	Eldoret	J
	33	1075	Kenya	KE	town	1 38	point	Kakamega North	T C

Una posible pregunta de investigación que podríamos estar interesados en explorar es: ¿El monto del préstamo solicitado varía según la ciudad?

Pregunta de Investigación

¿Cómo varía el monto del préstamo solicitado por ciudad?

Esta es una pregunta de investigación razonable, porque esperamos que la cantidad varíe porque el costo de los materiales y servicios varía de una región a otra.

Por ejemplo, esperaríamos que el costo de vida en un área rural sea más barato que en una ciudad urbana.

Mirando hacia el futuro: Modelado

Ahora tenemos nuestra pregunta de investigación, ¡podemos comenzar a modelar!

Fase de Modelado

Todos los modelos tienen 3 componentes clave: una tarea, una medida de desempeño y una metodología de aprendizaje.

Tarea	¿Cuál es el problema que queremos que resuelva nuestro modelo?
Metodología de Aprendizaje	Los algoritmos de ML pueden ser supervisados o no supervisados. Esto determina la metodología de aprendizaje.
Medida de Desempeño	Medida cuantitativa que utilizamos para evaluar el desempeño del modelo.

Fuente: Libro Deep Learning - Chapter 5: Introduction to Machine Learning

Revisaremos la tarea de aprendizaje automático y la metodología de aprendizaje en la próxima lección.

Hoy cubrimos esto:

- ✓ ¿Qué es Machine Learning?
- ✓ ¿Cómo se define una pregunta de investigación?
- ✓ ¿Qué son las observaciones?
- ✓ ¿Qué son las características?
- ✓ ¿Qué son las variables de salida?
- ✓ Introducción a los datos de KIVA