Introducción a la analítica de datos

Pedro O. Pérez M., PhD.

Herramientas computacionales: el arte de la analítica
Tecnológico de Monterrey

pperezm@tec.mx

03-2023

¿Qué es el análisis de datos?

- El análisis de datos es el proceso de limpiar, analizar, interpretar y visualizar datos para descubrir información valiosa que impulsa decisiones comerciales más inteligentes y efectivas.
- Aunque, no sólo es el análisis de datos en sí, sino también la recopilación, la organización, el almacenamiento y las herramientas y técnicas utilizadas para profundizar en los datos, así como las que se utilizan para comunicar los resultados, por ejemplo, las herramientas de visualización de datos.

¿Por qué es importante el análisis de datos?

- Los datos están en todas partes: en hojas de cálculo, el área de ventas, plataformas de redes sociales, encuestas de satisfacción, tickets de atención al cliente y más. En nuestra era de la información moderna, se crean a velocidades deslumbrantes y, cuando los datos se analizan correctamente, pueden ser el activo muy valioso.
- El análisis de datos puede ayudar a mejorar aspectos específicos sobre productos y servicios, así como la imagen de marca general, la experiencia del cliente, análisis de experimentos, etc.

- Análisis de texto.
- **Análisis descriptivo**.
- Análisis inferencial.
- **Análisis de diagnóstico**.
- Análisis predictivo.
- Análisis prescriptivo.

- El análisis de datos descriptivos proporciona el "¿Qué sucedió?" al analizar datos cuantitativos. Es la forma más básica y común de análisis de datos que se ocupa de describir, resumir e identificar patrones a través de cálculos de datos existentes, como media, mediana, moda, porcentaje, frecuencia y rango.

Análisis de diagnóstico

- El análisis de diagnóstico, también conocido como análisis de causa raíz, tiene como objetivo responder "¿Por qué sucedió 'X'?". Utiliza conocimientos del análisis estadístico para intentar comprender la causa o la razón detrás de las estadísticas, identificando patrones o desviaciones dentro de los datos para responder por qué.
- El análisis de diagnóstico puede ayudar a calcular la correlación entre estas posibles causas y los puntos de datos existentes.

¿Cómo analizar los datos?

- Si bien puede ser complejo realizar análisis de datos, según el tipo de datos que esté analizando, existen algunas reglas estrictas y rápidas que puede seguir. Incluyen establecer objetivos, recopilar, limpiar y analizar datos, y luego visualizarlos en cuadros de mando llamativos para facilitar la detección de patrones y tendencias.

Algunos ejemplos...

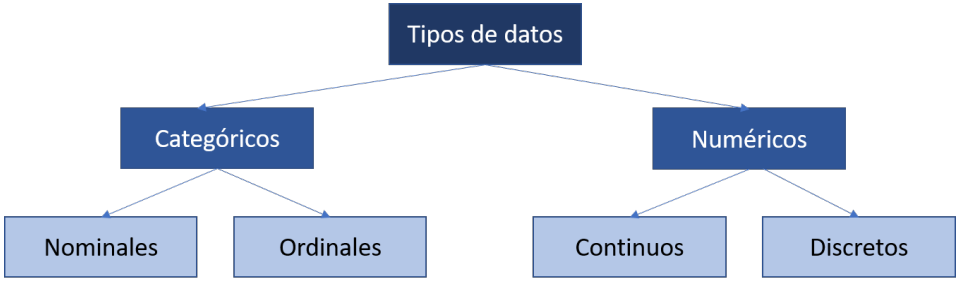
- Google Tracks Flu Trends
  https://www.youtube.com/watch?v=ytMzI3aphmo
- Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats
  https://www.youtube.com/watch?v=ahp7QhbB8G4&list=PLBE30C2B39FE4BD1C&index=4&t=14s
- Coca-Cola: Real-World Data Analytics Example
  https://www.youtube.com/watch?v=JIcXC_3Gfow

Tipos de datos

Los datos pueden provenir de muchas fuentes: sensores, texto, imágenes y vídeos. Gran parte de estos datos no están estructurados:

- Las imágenes son una colección de pixeles en donde cada píxel contiene información de color en formato RGB.
- Los textos son secuencias de palabras y caracteres, a menudo organizados por secciones, subsecciones, etc.
- Flujos de clic que representan las acciones de un usuario cuando interactúa con una aplicación o página web.

- Para aplicar las diversas herramientas de la ciencia de datos, los datos brutos deben procesarse y manipularse en forma ordenada. Para ello, debemos convertirlos en algún tipo de datos estructurado.
- Tener una buena comprensión de los diferentes tipos de datos es un requisito previo crucial para realizar un análisis exploratorio de datos, dado que algunas mediciones estadísticas solo se pueden usar con ciertos tipos de datos.
- También se necesita saber con qué tipo de datos estamos tratando para elegir el método de visualización correcto. Piensa en los tipos de datos como una forma de categorizar diferentes tipos de variables.

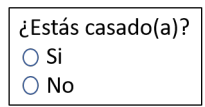


- Los datos categóricos toman un solo conjunto de valores fijos, como el tipo de computadora (laptop, computadora de escritorio, estación de trabajo, etc.) o los nombres de los días de la semana (lunes, martes, miércoles, etc.).
- Los datos numéricos son información que se puede medir y, por lo tanto, son datos representados como números y no como palabras o texto.

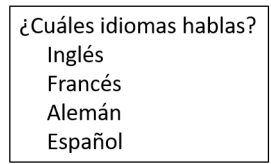
Datos nominales

- Los valores nominales representan unidades discretas y se utilizan para etiquetar variables que no tiene valor cuantitativo. Piensa en ellos como etiquetas. Ten en cuenta que los datos nominales no tienen orden. Por lo tanto, si se cambia el orden de los valores, el significado no cambiará.

- El ejemplo (a), que describe se una persona está casada, se llamaría dicotómica, que es un tipo de datos nominales que contiene solo dos categorías.
- Otro subtipo de datos nominales son "nominales con orden" (como "frío", "tibio", "caliente", "muy caliente") y nominales sin orden ("masculino", "femenino").



(a)

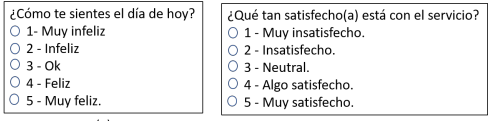


(b)

- Como puedes observar, las escalas son mutuamente excluyentes (no se sobreponen) y ninguna de ellas tiene ningún significado numérico. Una buena forma de recordar todo esto es que “nominal” suena mucho como “nombre” y las escalas nominales son como “nombre” o etiquetas.

Datos ordinales

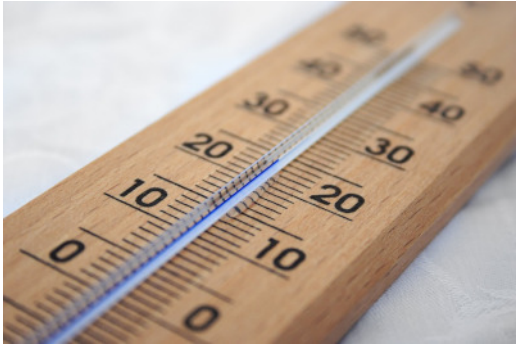
- Los datos ordinales representan unidades discretas y ordenadas. Por lo tanto, el orden de los valores es importante y significativo. En la siguiente imagen podemos ver un ejemplo:
- En cada caso, sabes que un 4 es menor que un 3 o un 2, pero no sabemos, y no podemos cuantificar, cuánto mejor es. Por ejemplo, ¿la diferencia entre "Ok" e "Infeliz" es la misma diferencia entre "Muy feliz" y "Feliz"? No podemos decirlo.



- Las escalas ordinales son típicamente medidas de conceptos no numéricos como satisfacción, felicidad, malestar, etc.
- "Ordinal" es fácil de recordar porque suena como "orden" y esa es la clave para recordar con "escalas ordinales"; es el orden lo que importa, pero eso es todo lo que realmente obtienes de estas.

Escalas de intervalo

- Las escalas de intervalo (o datos continuos) son escalas numéricas en las que conocemos tanto el orden como las diferencias exactas entre los valores. El ejemplo clásico es la temperatura Celsius porque la diferencia entre cada valor es la misma. Por ejemplo, la diferencia entre 60 y 50 grados es de 10 grados medibles, al igual que la diferencia entre 80 y 70 grados.
- Las escalas de intervalo son buenas porque nos permiten utilizar herramientas estadísticas para su análisis. Por ejemplo, la tendencia central puede ser medida por la moda, mediana o media. También se puede calcular la desviación estándar.



- Lo importante a recordar: las escalas de intervalo no solo nos dicen sobre el orden, sino también sobre el valor entre cada elemento.
- El problema con las escalas de intervalo es que no tienen un "cero verdadero". Por ejemplo, no existe "sin temperatura", al menos no con grados Celsius. En el caso de las escalas de intervalo, cero no significa la ausencia de valor, en realidad es otro número utilizado en la escala, como 0 grados centígrados.Sin un cero verdadero, es imposible calcular razones (o tasas). Con datos de intervalo, podemos sumar y restar, pero no multiplicar, ni dividir.

- ¿Confuso? Ok, considera esto: 10 grados C + 10 grados C = 20 grados C. No hay problema ahí. Sin embargo, 20 grados C no es dos veces más caliente que 10 grados C, porque no existe la "no temperatura" cuando se trata de la escala Celsius. Cuando se convierte a Fahrenheit, está claro: 10C = 50F y 20C = 68F, que claramente no es el doble de caliente. En pocas palabras, las escalas de intervalo son geniales, pero no podemos calcular las proporciones, lo que nos lleva a nuestra última escala de medición.

Proporción

- Las escalas de proporción (ratio en inglés) son el nirvana definitivo cuando se trata de escalas de medición de datos porque nos informan sobre el orden, nos dicen el valor exacto entre las unidades y también tienen un cero absoluto, lo que permite que una amplia gama de estadísticas tanto descriptivas como inferenciales puedan ser aplicadas.
- Las escalas de proporción brindan una gran cantidad de posibilidades cuando se trata de análisis estadístico. Estas variables se pueden sumar, restar, multiplicar, dividir (proporciones, de ahí su nombre) de manera significativa. La tendencia central se puede medir por moda, mediana o media. Las medidas de dispersión, como la desviación estándar y el coeficiente de variación, también se pueden calcular a partir de escalas de proporción.

En resumen...

	Nominal	Ordinal	Intervalo	Proporción
El “orden” de los valores es conocido		X	X	X
Frecuencia de distribución	X	X	X	X
Moda	X	X	X	X
Mediana		X	X	X
Media			X	X
Se puede cuantificar la diferencia entre cada valor			X	X
Se pueden sumar o restar los valores			X	X
Se pueden multiplicar o dividir los valores				X
Tiene un “cero” verdadero				X

¿Por qué los tipos de datos son importantes?

- Pues, resulta que para efectos del análisis de datos y el modelado predictivo, el tipo de datos es importante para determinar el tipo de visualización, análisis de datos o modelo estadístico. De hecho, herramientas para ciencia de datos, como Python, utilizan estos tipos de datos para mejorar el rendimiento computacional. Más importante aún, el tipo de datos para un variable define cómo el software manejará los cálculos para esa variable.

- Mucho nos puede desconcertar por qué incluso necesitamos la noción de datos categóricos y ordinales para el análisis. Sin embargo, la identificación explícita de datos como categóricos, a diferencia del texto, ofrece algunas ventajas:
  1. Saber que son categóricos, le permite saber al software cómo deben comportarse los procedimientos estadísticos, cómo producir un gráfico o ajustar un modelo.
  2. El almacenamiento e indexación se puede optimizar.
  3. Los valores que puede tomar una variable categórica dada se pueden representar como una enumeración.

- Un conjunto de datos puede ser representado como una matriz bidimensional con renglones que indican registros (casos) y columnas que indican características (variables). Es importante mencionar, que los datos no siempre comienzan de esta forma. Los datos no estructurados deben procesarse y manipularse para que se pueden representar como un conjunto de características en una matriz.

Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
Nashs.Field	3.6	11	Grassland	4.1	F	4
Silwood.Bottom	5.1	2	Arable	5.2	F	7
Nursery.Field	2.8	3	Grassland	4.3	F	2
Rush.Meadow	2.4	5	Meadow	4.9	T	5
Gunness.Thicket	3.8	0	Scrub	4.2	F	6
Oak.Mead	3.1	2	Grassland	3.9	F	2
Church.Field	3.5	3	Grassland	4.2	F	3
Ashurst	2.1	0	Arable	4.8	F	4
The.Orchard	1.9	0	Orchard	5.7	F	9
Rookery.Slope	1.5	4	Grassland	5	T	7
Garden.Wood	2.9	10	Scrub	5.2	F	8
North.Gravel	3.3	1	Grassland	4.1	F	1
South.Gravel	3.7	2	Grassland	4	F	2
Observatory.Ridge	1.8	6	Grassland	3.8	F	0
Pond.Field	4.1	0	Meadow	5	T	6
Water.Meadow	3.9	0	Meadow	4.9	T	8
Cheapside	2.2	8	Scrub	4.7	T	4
Pound.Hill	4.4	2	Arable	4.5	F	5
Gravel.Pit	2.9	1	Grassland	3.5	F	1
Farm.Wood	0.8	10	Scrub	5.1	T	3