Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

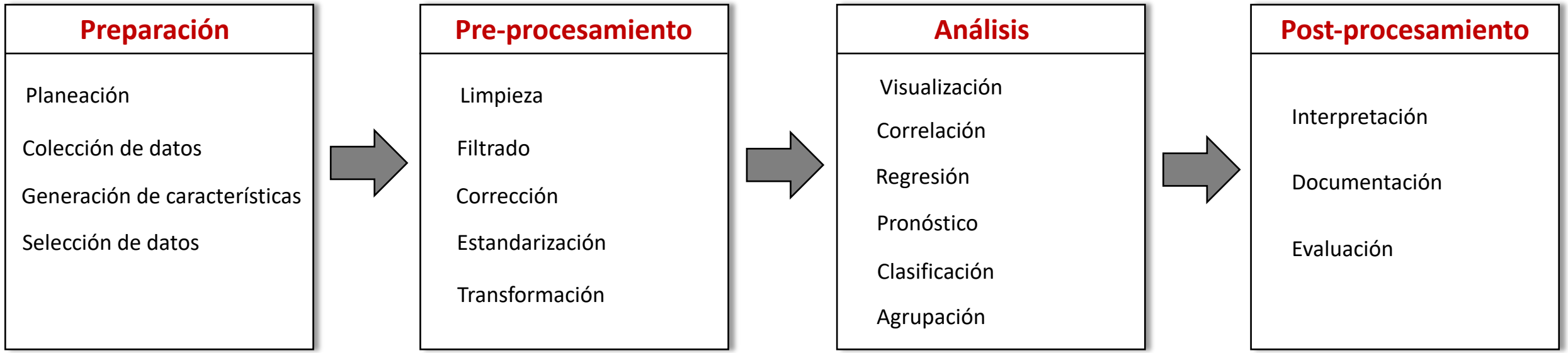
Agosto, 2022

Analítica de Datos, Minería de Datos y Descubrimiento de Conocimiento (1)

- **Minería de Datos**
  - Su objetivo es extraer conocimiento de los datos. En este contexto el *conocimiento* se define como patrones interesantes que generalmente son válidos, novedosos, útiles y entendibles para el ser humano.
  - El hecho de que los patrones extraídos sean interesantes o no, depende de la aplicación en particular y deben ser verificados por expertos. Con base en la retroalimentación, el proceso de extracción de conocimiento se refina de forma interactiva frecuentemente.
- **Analítica de Datos**
  - Se define como la aplicación de sistemas de cómputo para analizar grandes conjuntos de datos que brinden un soporte a las decisiones.
  - Es un campo multidisciplinario que ha adoptado aspectos de otras disciplinas tales como la estadística, el aprendizaje automático, el reconocimiento de patrones, la teoría de sistemas, investigación de operaciones o la Inteligencia Artificial.
  - Los proyectos típicos de análisis de datos se pueden dividir en varias fases. Los datos se evalúan y seleccionan, limpian y filtran, visualizan y analizan; y los resultados del análisis finalmente se interpretan y evalúan.

Analítica de Datos, Minería de Datos y Descubrimiento de Conocimiento (2)

- **Descubrimiento de Conocimiento en Bases de Datos**
  - El proceso de descubrimiento de conocimiento en bases de datos (KDD) está compuesto por 6 fases: selección, preprocesamiento, transformación, extracción de datos, interpretación y evaluación.
  - Para simplificar el proceso, únicamente se consideran 4 fases: preparación, **pre-procesamiento**, **análisis** y post-procesamiento.



Datos (1)

- **Conjunto de Datos de Iris**
  - Para presentar los conceptos básicos del análisis de datos, consideraremos uno de los conjuntos de datos de referencia históricos más populares:
    - El Conjunto de Datos Iris
      - *E. Anderson. The Irises of the Gaspe Peninsula. Bull. of the American Iris Society, 59:2–5, 1935.*
    - Fue creado originalmente en 1935 por el botánico Edgar Anderson, quien examinó la distribución geográfica de las flores de Iris en la península de Gaspé en Quebec.
    - En 1936, Sir Ronald Aylmer Fisher usó el conjunto de datos Iris como ejemplo para el análisis discriminante multivariable.
      - *R. A. Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179–188, 1936*.
    - Posteriormente, el conjunto de datos de Iris se convirtió en uno de los conjuntos de datos de referencia más utilizados en estadística y análisis de datos.

Datos (2)

- **Conjunto de Datos de Iris**
  - El conjunto de datos de Iris comprende mediciones de 150 muestras de flores de Iris:
    - 50 de cada una de las tres especies Iris Setosa, Iris Virginica e Iris Versicolor.
    - Para cada una de las 150 flores, se midieron los valores de cuatro características numéricas elegidas por Anderson:
      - la longitud y la anchura de las hojas sépalo y pétalo en centímetros.
    - Se puede descargar de: https://archive.ics.uci.edu/ml/datasets/Iris
    - El conjunto de datos contiene 3 clases de 50 instancias cada una, donde cada clase se refiere a un tipo de planta de Iris.
  - A continuación se presenta un fragmento del conjunto de datos Iris…

Datos (3)

- **Conjunto de Datos de Iris**

Setosa				Versicolor				Virginica			
Sepal		Petal		Sepal		Petal		Sepal		Petal	
Length	Width	Length	Width	Length	Width	Length	Width	Length	Width	Length	Width
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

- En el análisis de datos llamamos a cada una de las 150 flores de iris como **objetos**, cada una de las 3 especies como **clases** y cada una de las 4 dimensiones como **características**.
- Algunas preguntas típicas para responder mediante el análisis de datos:
  - ¿Cuál de los datos podría contener errores o asignaciones de clases falsas?
  - ¿Cuál es el error causado por redondear los datos a un decimal?
  - ¿Cuál es la correlación entre la longitud y el ancho de los pétalos?
  - ¿Qué par de dimensiones están más correlacionadas?
  - Ninguna de las flores en el *dataset* tiene un ancho de sépalo de 1.8 cm. ¿Qué longitud de sépalo esperaríamos para una flor que tuviera 1.8 cm como ancho de sépalo?
  - ¿A qué especie pertenecería un Iris con un ancho de sépalo de 1.8 cm?
  - Las 3 especies contienen subespecies que pueden identificarse a partir de los datos?

Tipos de datos estadísticos (1)

- **Datos categóricos**
  - También conocidos como datos cualitativos representan características como el género, idioma de las personas. También pueden tomar valores numéricos, por ejemplo: 1 – Mujeres, 0 – Hombres; sin que estos números tengan algún significado matemático. Los tipos de datos categóricos se clasifican en:
    - **Datos Nominales.** Que refieren a valores nominales que representan unidades discretas y se usan para etiquetar variables que no tienen un valor cuantitativo. Estos datos no tienen un orden; es decir, que aunque se cambiara el orden de sus valores, no cambia su significado.
    - **Datos Ordinales.** Éstos representan unidades discretas y ordenadas. Por tanto, es casi lo mismo que los datos nominales, excepto por su orden que es importante. Las escalas ordinales se usan para medir características no numéricas como: estado anímico, nivel de calidad, entre otros.
- **Datos numéricos**
  - También se conocen como datos cuantitativos y se refieren a una medida. Se clasifican de la siguiente manera:

Tipos de datos estadísticos (2)

- **Datos numéricos**
  - **Datos discretos.** Son discretos cuando sus valores son distintos y separados; es decir, cuando los datos sólo pueden tomar ciertos valores. Este tipo de datos no puede medirse, pero se pueden contar. Básicamente representan información que se puede clasificar.
  - **Datos continuos.** Representan mediciones y por lo tanto, sus valores no se pueden contar pero se pueden medir. A su vez, éstos se clasifican de la siguiente manera:
    - **Datos de intervalo.** Representan unidades ordenadas que tienen la misma diferencia. Por lo tanto, hablamos de datos de intervalo cuando tenemos una variable que contiene valores numéricos que están ordenados y donde conocemos las diferencias exactas entre los valores.
      - El problema con los datos de valores de intervalo es que podemos sumar y restar, pero no podemos multiplicar, dividir o calcular razones. Debido a que no existe un cero verdadero, no se pueden aplicar muchas estadísticas descriptivas e inferenciales.
    - **Datos de relación.** También son unidades ordenadas que tienen la misma diferencia. Los datos de relación son los mismos que los valores de intervalo, con la diferencia de que tienen un cero absoluto.

Datos Nominales (1)

- Son datos "etiquetados" o "nombrados" que pueden dividirse en varios grupos que no se traslapan. En este caso, los datos no se miden ni se evalúan, sino que se asignan a varios grupos.
- Estos grupos son únicos y no tienen elementos comunes. El orden de los datos recopilados no puede establecerse utilizando datos nominales y, por lo tanto, si cambia el orden de los datos, su significado no se verá alterado.
- Presentan una similitud entre los diversos elementos, pero es posible que no se revelen los detalles relativos a esta similitud. Se trata simplemente de facilitar el proceso de recolección y análisis de datos.
- Como ejemplo: Si los datos binarios representan datos de "dos valores", los datos nominales representan datos de "múltiples valores" y no pueden ser cuantitativos. Los datos nominales se consideran discretos.

Datos Nominales (2)

- Los datos nominales nunca pueden ser cuantificados. Los datos nominales siempre estarán en forma de nomenclatura. Por ejemplo, una *encuesta para países asiáticos* puede incluir una pregunta como: *¿cuál es su origen étnico?* – con las siguientes opciones de respuesta.
  - Centroasiático
  - Indonesio
  - Asiático occidental
  - Japonés
- El análisis estadístico, lógico o numérico sobre estos datos no es posible, es decir, no se puede sumar, restar o multiplicar los datos recolectados o concluir que la variable 1 es mayor que la variable 2.

Datos Nominales (3)

- Entre las principales características de los datos nominales se tienen:
  - **Ausencia de orden.** A diferencia de los datos ordinales, los datos nominales tampoco pueden ser asignados a un orden definido. En el ejemplo anterior, el orden de las opciones de respuesta es irrelevante para las respuestas que proporciona el encuestado.
  - **Propiedad cualitativa.** Los datos recopilados siempre tendrán una propiedad cualitativa, es muy probable que las opciones de respuesta sean de esa naturaleza.
  - **No es posible calcular la media.** La media de los datos nominales no se puede establecer, incluso si los datos están ordenados alfabéticamente. En el ejemplo anterior, es imposible calcular el promedio de respuestas presentadas para las etnias debido a la naturaleza cualitativa de las opciones.
  - **Los datos son principalmente alfabéticos.** En la mayoría de los casos, los datos nominales son alfabéticos y no numéricos. Por ejemplo, en el caso mencionado los datos no numéricos también se pueden categorizar en varios grupos.

Datos Nominales (4)

- Como ejemplos de datos nominales, se tienen los siguientes :
  - ¿Cuál es la raza de perros más querida?
    - Dálmata – 1
    - Doberman – 2
    - Labrador – 3
    - Pastor Alemán – 4
  - ¿A quién le gusta más viajar?
    - Hombres – 1
    - Mujeres – 2
  - ¿Qué tipo de casas prefieren los residentes de la ciudad de Nueva York?
    - Apartamentos – 1
    - Bungalows -2
    - Chalets – 3

Datos Ordinales (1)

- Los datos ordinales son un tipo estadístico de datos cuantitativos en los que existen variables en categorías ordenadas que se producen de forma natural. La distancia entre dos categorías no se establece utilizando datos ordinales.
- En estadística, un grupo de números ordinales se representan usando una escala ordinal. La principal diferencia entre los datos nominales y ordinales es que los ordinales tienen un orden de categorías mientras que los nominales no.
- La **Escala Likert** es un ejemplo popular de datos ordinales.
  - Por ejemplo: para una pregunta como: *"Exprese la importancia que tiene el precio al comprar un producto"*, una escala Likert tendrá las siguientes opciones que están codificadas a 1, 2, 3, 4, 5 (números), en donde 1 es menor que 2, que es menor que 3, que es menor que 4, y que a su vez es menor que 5.

Muy importante	Importante	Neutral	Irrelevante	Muy irrelevante
1	2	3	4	5

Datos Ordinales (2)

- Por lo tanto, los datos ordinales son un conjunto de variables ordinales, es decir, variables con un orden particular – "**bajo, medio, alto**", pueden representarse como datos ordinales. Hay dos factores importantes a considerar para los datos ordinales.
  - Hay múltiples términos que representan el "orden" como "Alto, Superior, Máximo" o "Satisfecho, Insatisfecho, Extremadamente Insatisfecho".
  - Pero la diferencia entre las variables no es uniforme.

Niveles de medición (1)

- El primer paso en el análisis de datos es simplemente entender lo que éstos significan. Esto se facilita clasificando cada variable según su nivel de medición. El nivel de medición se refiere a la relación entre los valores que se asignan a los atributos de una variable.
- Una variable es cualquier cantidad que puede ser medida y cuyo valor varía a través de la población o muestra.
  - Por ejemplo, si consideramos una población de estudiantes, la nacionalidad del estudiante, género, calificaciones, etc., son todas las variables definidas, y su valor correspondiente diferirá para cada estudiante.
  - Si se desea calcular el salario promedio de los ciudadanos de un país, podemos salir y registrar el salario de todas y cada una de las personas para calcular el promedio o elegir un muestreo aleatorio de toda la población y calcular el salario promedio para esa muestra, y luego usar las pruebas estadísticas para obtener conclusiones para una población más amplia.
- Entonces, el tipo de prueba estadística que puede utilizarse para llegar a una conclusión sobre la población en general depende del nivel de medición de la variable considerada.
- El nivel de medición de una variable no es otra cosa que la naturaleza matemática de una variable o cómo se mide una variable.

Niveles de medición (2)

- Tipos de niveles de medición
  - Los números se pueden agrupar en 4 tipos o niveles: nominal, ordinal, por intervalos y de razón.
  - **Nivel de medición nominal.** El nivel nominal es apenas una medida. Se refiere a la cualidad más que a la cantidad. Un nivel nominal de medición es simplemente una cuestión de diferenciar por nombre, por ejemplo, 0 = mujer, 1 = hombre. Aunque se usen los números 1 y 2, éstos no indican cantidad. En el mismo sentido, por ejemplo la categoría binaria de 0 y 1 utilizada para las computadoras es un nivel nominal de medición. Algunos ejemplos:
    - PREFERENCIA DE COMIDA: desayuno, comida, cena
    - PREFERENCIA RELIGIOSA: 1= Budista, 2= Musulmana, 3= Cristiana, 4= Judía, 5= Otra
    - ORIENTACIÓN POLÍTICA: Izquierda, Derecha, Independiente
    - Otros valores nominales son números de seguro social, códigos postales y números de teléfono.

Niveles de medición (3)

- Tipos de niveles de medición
  - **Nivel de medición ordinal.** Este nivel se refiere al orden en la medición. Una escala ordinal indica la dirección, además de proporcionar información nominal. Bajo/Medio/Alto o Más Rápido/Más Lento son ejemplos de niveles ordinales de medición. Calificar una experiencia con un "9" en una escala de 1 a 10 nos indica que fue mejor que una experiencia calificada con un "6". Muchas escalas o pruebas psicológicas utilizan la escala ordinal de medición. Algunos ejemplos:
    - CLASIFICACIÓN: 1er lugar, 2do lugar… último lugar
    - NIVEL DE ACUERDO: No, Tal vez, Sí
    - CALIFICACIÓN CURSO: 1, 2,…, 10

Niveles de medición (4)

- Tipos de niveles de medición
  - **Nivel de medición de intervalo.** Proporcionan información sobre el orden y también poseen intervalos iguales. Del ejemplo anterior, si conociéramos que la distancia entre 1 y 2 es la misma que entre 7 y 8 en nuestra escala de calificación de 10 puntos, entonces tendríamos una escala de intervalo.
  - Un ejemplo de una escala de intervalo es la temperatura, medida en una escala Fahrenheit o Celsius. Un grado representa la misma cantidad subyacente de calor, independientemente de dónde ocurra en la escala.
  - Si lo medimos en unidades Fahrenheit, la diferencia entre una temperatura de 46 y 42 es la misma que la diferencia entre 72 y 68. Las escalas de medición de intervalos iguales pueden ser utilizadas para medir opiniones y actitudes. Algunos ejemplos:
    - HORA DEL DÍA en un reloj de 12 horas
      - Intervalo de tiempo de día – intervalos iguales; reloj analógico (12 horas), la diferencia entre la 1 y 2 pm es la misma que la diferencia entre las 11 y 12 am.

Niveles de medición (5)

- Tipos de niveles de medición
  - **Nivel de medición de razón.** Además de poseer las cualidades de las escalas nominal, ordinal y de intervalo, una escala de razón tiene un cero absoluto (un punto donde no existe ninguna de las cualidades que se están midiendo).
  - Utilizar una escala de razón permite hacer comparaciones como: ser el doble de alto, o la mitad de alto de una persona. El tiempo de reacción (cuánto tiempo tarda en responder a una señal de algún tipo) utiliza una escala de medición de razón, el tiempo.
  - Aunque el tiempo de reacción de un individuo siempre es mayor que cero, conceptualizamos un punto cero en el tiempo y podemos afirmar que una respuesta de 24 milisegundos es dos veces más rápida que un tiempo de respuesta de 48 milisegundos. Algunos ejemplos:
    - REGLA: pulgadas o centímetros
    - INGRESOS: dinero ganado el año pasado
    - AÑOS: de experiencia laboral
    - De RAZÓN: el tiempo de 24 horas tiene un 0 absoluto (medianoche); 14 en punto está dos veces más lejos de la medianoche que las 7 en punto.

Escala Nominal (1)

- Una escala nominal es una escala de medición en la cual los números sirven como "etiquetas", solamente para identificar o clasificar un objeto. Una escala de medición nominal normalmente trata sólo con variables no numéricas (cualitativas).
- Por ejemplo, supongamos que se realiza esta pregunta: "¿Podrías seleccionar el grado de incomodidad de tu enfermedad?" 1-Leve; 2-Moderado; 3-Severo.
- Aquí los números simplemente son utilizados como etiquetas y no tienen ni un solo valor.
- La escala nominal posee solo la característica de descripción, y esto significa que posee etiquetas únicas que sirven para identificar o delegar valores de los objetos.
- Cuando la escala nominal se utiliza con fines de identificación, existe una correlación uno a uno entre un objeto y el valor asignado a él.
  - Por ejemplo, los números que están escritos en los autos de carrera simplemente están ahí para identificar al conductor asociado con el automóvil, la realidad es que estos números no tienen nada que ver con las características del automóvil.

Escala Nominal (2)

- Pero cuando se utiliza la escala nominal para fines de clasificación, los números asignados al objeto sirven como etiquetas para categorizar y organizar objetos por clase.
  - Por ejemplo, en el caso de una escala de género, un individuo puede clasificarse como masculino o femenino. En este caso, todos los objetos de la categoría tienen el mismo número, por ejemplo, todos los hombres pueden ser número 1 y todas las mujeres pueden ser número 2. Tomar en cuenta que ese valor es puramente utilizado para fines de conteo.
  - Desde el punto de vista estadístico, la escala nominal es una de las escalas de medición más fáciles de comprender. Como se mencionó anteriormente, la escala nominal se asigna a objetos o elementos que no son cuantitativos, ni están orientados a un número.
    - Por ejemplo, supongamos que tenemos 5 colores, naranja, azul, rojo, negro y amarillo. Podríamos enumerar éstos en cualquier orden que nos guste, ya sea del 1 al 5 o del 5 al 1 en orden ascendente o descendente.
      - Aquí los números se asignan a los colores sólo con el propósito de identificación.

Escala Nominal (3)

- Características de la escala nominal:
  - En una escala nominal, una variable se divide en dos o más categorías, por ejemplo, de acuerdo / en desacuerdo, si / no, etc. Es un mecanismo de medición en el que la respuesta a una pregunta en particular puede caer en cualquier categoría.
  - La escala nominal es de naturaleza cualitativa, lo que significa que los números se usan únicamente para categorizar o identificar objetos. Por ejemplo, en el fútbol, ¿has notado que los jugadores tienen un número en su camiseta? (cada uno tiene un número diferente). La realidad es que estos números no tienen nada que ver con la capacidad de los jugadores, sin embargo, pueden ayudar a identificar al jugador.
  - En una escala nominal, los números no definen las características relacionadas con el objeto, lo que significa que cada número se asigna a un objeto aleatorio o por decisión propia. El único aspecto permitido relacionado con los números en una escala nominal es que sirven para "contar".
  - En una escala nominal es fácil generar respuestas utilizando preguntas cerradas, es por eso que se pueden recopilar muchas respuestas en un corto periodo de tiempo, lo que a su vez aumenta la confiabilidad de las respuestas.
    - Si volvemos al ejemplo de la clasificación de hombres y mujeres, 1 siendo hombres y 2 siendo mujeres, los números nos servirán para saber cuántos hombres (1) hay y cuántas mujeres (2) hay.
- Ejemplos:
  - ¿Cómo describirías tu comportamiento?
    - E – Extrovertido; I – Introvertido; A – Ambas
  - Podrías seleccionar una opción que describa tu color de cabello
    - 0 – Negro; 1 – Café; 2 – Rojo; 3 – Amarillo; 4 – Otro
  - ¿Cuál es tu género? (Subtipo de escala nominal con solo dos categorías, conocido también por: escala nominal dicotómica)
    - H – Hombre; M – Mujer

Escala Ordinal (1)

- La escala ordinal es uno de los niveles de medición que nos otorga la clasificación y el orden de los datos sin que realmente se establezca el grado de variación entre ellos.
- Los datos ordinales son básicamente datos estadísticos que tienen la misma naturalidad pero existe una diferencia entre ellos que es desconocida. Estos datos pueden ser agrupados o clasificados.
- Por lo tanto, se utiliza una escala ordinal como parámetro para comprender si las variables son mayores o menores. La tendencia central de la escala ordinal es la mediana.
- La escala de Likert es un ejemplo de porque la diferencia de intervalo entre las variables ordinales no se puede concluir. En esta escala de hecho, las opciones de respuesta suelen ser polares, como por ejemplo, algo como "totalmente satisfecho" o "totalmente insatisfecho".
- La intensidad de la diferencia entre estas dos opciones no puede ser relacionada a valores específicos, ya que el valor de la diferencia entre totalmente satisfecho y totalmente insatisfecho es mucho mayor que la distancia entre satisfecho y neutral.
    - Supongamos que a una persona le encantan los automóviles Mercedes Benz, y se le aplica una encuesta que consta de una pregunta
    - ¿qué tan probable es que le recomiendes los automóviles de Mercedes Benz a tus amigos y familiares?
    - Supongamos que será muy fácil que este elija "Extremadamente probable" en lugar de "probable". Pero qué pasa si fuera una persona "neutral", a esta persona si le costaría tal vez un poco de trabajo elegir.
    - Es por eso que se utiliza una escala ordinal cuando se debe deducir el orden de las opciones, y no cuando se debe establecer una diferencia de intervalo.

Escala Ordinal (2)

- Propiedades de la escala ordinal:
  - Además de identificar y describir la magnitud, la escala ordinal suele mostrar el rango relativo de variables.
  - Las propiedades del intervalo no se conocen.
  - Se miden atributos no numéricos como frecuencia, satisfacción, felicidad, etc.
  - Además de la información proporcionada por la escala nominal, la escala ordinal identifica el rango de las variables.
  - Utilizando esta escala, los encuestadores pueden analizar el grado de acuerdo o desacuerdo de los encuestados con respecto a una pregunta realizada.
  - Facilidad de comparación entre variables, ya que son extremadamente convenientes para agrupar variables después de que son ordenadas.
- Ejemplos:
  - Ranking de estudiantes de secundaria: 1ero, 3ero, 4to, 5to , etc. Un estudiante con un puntaje de 99/100 sería el primer rango, otro estudiante con puntaje de 98/100 sería el segundo, y así sucesivamente.
  - Encuestas de calificación en restaurantes: cuando se recibe una encuesta con una pregunta como: "¿Qué tan satisfecho está con la experiencia gastronómica?" En ésta las opciones de respuesta pueden ser algo como calificar del 0 al 10, siendo 10 extremadamente satisfecho y 0 extremadamente insatisfecho.
  - Escala de Likert: la escala de Likert es una variante de la escala ordinal que se utiliza para calcular niveles de satisfacción.

Escala de Intervalo (1)

- La escala de intervalo se define como una escala de medición cuantitativa en la que se mide la diferencia entre dos variables. En otras palabras, las variables se miden en valores reales y no de forma relativa, donde la presencia de cero es arbitraria. Esto significa que la diferencia entre dos variables en una escala es una distancia real o igual.
  - Por ejemplo, la diferencia entre 40 grados centígrados y 50 grados centígrados es exactamente la misma que la diferencia entre 50 grados centígrados y 60 grados centígrados.
- El "Intervalo" equivale a la distancia entre dos variables. Otra manera fácil de recordar lo que es una escala de intervalo es considerando que ésta es la resta que se define entre dos variables. Esto es diferente a la escala de razón, donde la división se define entre dos variables.
- Los datos de la escala pueden ser:
  - Tipo discretos. Ejemplo: números tipo 8 grados, 4 años, 2 meses, etc.
  - Tipo continuos. Ejemplo: con números fraccionarios como 12.2 grados, 3.5 semanas o 4.2 kilómetros.

Escala de Intervalo (2)

- Las características de esta escala son las siguientes:
  - La escala de intervalo es preferible a la escala nominal o la escala ordinal porque las dos últimas son escalas cualitativas. La escala es cuantitativa en el sentido de que se pueden cuantificar la diferencia entre dos valores.
  - Se pueden restar valores entre dos variables y esto te ayuda a comprender la diferencia entre dos variables.
  - Esta escala permite calcular la media de las variables.
  - Esta es una escala preferida en estadística porque permite asignar un valor numérico a cualquier evaluación arbitraria.

Escala de Intervalo (3)

- Ejemplos:
- La escala de intervalo es el tipo de pregunta que se utiliza con mayor frecuencia en un estudio o investigación.
  - Para obtener cualquier tipo de respuesta, es indispensable que la pregunta solicitada requiera que los encuestados respondan en una escala numérica donde la diferencia entre los dos números sea la misma.
  - En la escala de intervalos, se requiere que la encuesta se diseñe de tal forma que la dimensión que se mida también se escale adecuadamente. Esto se puede lograr de manera numérica o verbal. Los tipos de preguntas para escalas de intervalo:
    - **Net Promoter Score (NPS).** En esta pregunta de intervalo, la pregunta se realiza usando una escala del 1 al 10 para responder. La pregunta NPS se basa en saber qué tan probable es que un cliente le recomiende tu negocio, producto o servicio a sus amigos, colegas y familiares.

Escala de Intervalo (4)

- Ejemplos:
  - **Matriz de Escala Bipolar.** Otra pregunta de intervalo es cuando un objeto es evaluado por el encuestado en una matriz de escala bipolar (*i.e.*, utilizando una escala de calificación de 5 puntos).
  - La escala de intervalo da la capacidad de cuantificar y diferenciar entre opciones. Esto es mejor que la escala nominal y la escala ordinal, ya que en éstas no se tienen en cuenta los conocimientos cuantitativos.
  - La escala consiste en variables que existen a lo largo de una escala común a intervalos iguales. Las medidas científicas utilizadas para calcular la distancia entre las variables son altamente confiables.
  - Aunque en esta escala hay ausencia del 0 absoluto, la división de variables no es posible. Entonces, mientras que la escala de intervalo permite analizar una gran cantidad de datos a partir de una pregunta, no tiene la capacidad de calcular proporciones.

Escala de Razón (1)

- Los datos de escala de razón se definen como un tipo de datos cuantitativos que se caracterizan por un punto de cero absoluto, lo que significa que no hay ningún valor numérico negativo. Por ejemplo:
  - Cuatro personas son seleccionadas al azar y se les pregunta ¿cuánto dinero traen? Estos son los resultados: $21, $50, $65 y $300.
    - ¿Existe un orden para estos datos? Sí, $21 < $50 < $65 < $300.
    - ¿Las diferencias entre los valores de datos son significativas? Sí, la persona que tiene $50 tiene $29 más que la persona con $21.
    - ¿Podemos calcular razones en función a estos datos? Sí, porque $0 es la cantidad mínima absoluta de dinero que una persona podría traer con ella.
    - La persona con $300 tienen 6 veces más que la persona con $50.
  - Los datos de escala de razón tienen todas las propiedades de los datos de la escala de intervalo. Por ejemplo, los datos deben tener valores numéricos, la distancia entre los dos puntos es igual, etc. Sin embargo, a diferencia de los datos de intervalo donde el cero es arbitrario, en los datos de una escala de razón el cero es absoluto.

Escala de Razón (2)

- Ejemplos de datos de escala de razón
  - La medición de alturas.
    - La altura puede medirse en centímetros, metros, pulgadas o pies. No es posible tener una altura negativa.
  - Si los comparamos con los datos de una escala de intervalo, por ejemplo, la temperatura puede ser de -10 grados, sin embargo, la altura no puede ser negativa como se mencionó anteriormente.
  - Los datos de escala de razón pueden ser multiplicados y divididos, ésta es una de las principales diferencias entre los datos de escala de razón y los datos de una escala de intervalo, los cuales solo pueden ser sumados y restados.
  - En los datos de escala de razón, la diferencia entre 1 y 2 es la misma que la diferencia entre 3 y 4, pero también aquí 4 es el doble que 2. Esta comparación es imposible en los datos de escala de intervalo.
- Análisis de datos de escala de razón
  - Los datos de escala de razón, junto con los otros 3 niveles de medición, son fundamentalmente un método de captura de datos cuantitativos. Lo que significa que se pueden aplicar todos los tipos de técnicas de análisis estadístico a los datos de razón.

Escala de Razón (3)

- Técnicas de análisis de datos de razón más comunes:
  - **Análisis de tendencias.** Es una técnica utilizada para extraer tendencias e *insights* capturando los datos de encuestas durante un período de tiempo determinado. El análisis de tendencias también juega un papel crítico en el análisis predictivo, en el que se compara y analiza un conjunto de datos con plazos determinados para predecir tendencias futuras.
  - **Análisis FODA o SWOT.** La evaluación FODA y es ampliamente utilizada para evaluar los datos de escala de razón. Las fortalezas y debilidades son aspectos internos de una organización, mientras que las oportunidades y amenazas son externas a una organización.
  - **Análisis Conjoint.** Es una técnica de investigación de mercado de nivel avanzado implementada generalmente para analizar cómo los individuos toman decisiones complicadas en una escala de razón. Qué factores son importantes para los clientes antes de tomar decisiones en las que se tienen múltiples opciones a su disposición. Se pueden probar sitios web, realizar investigaciones de precios o mejorar las características o funciones del producto, utilizando el análisis conjoint o análisis conjunto.
  - **Tablas cruzadas.** Es un método para comprender la relación entre múltiples variables. Esta se utiliza para establecer una correlación entre múltiples variables de datos de escala de razón en un formato tabular. Se pueden tomar decisiones informadas después de analizar los datos de una tabla cruzada. Generalmente se analiza la intención del cliente y el rendimiento del producto utilizando la tabulación cruzada, ya que proporciona una comparación entre dos o más variables.
  - **Análisis TURF.** Significa Análisis de Frecuencia y Alcance Total No Duplicado, es un método que permite analizar el potencial de la investigación de mercado para una combinación de productos y servicios. Evalúa los datos de razón de los clientes abordados mediante una determinada fuente de comunicación y su frecuencia. Esta técnica de análisis es utilizada para comprender si un nuevo producto o servicio será o no bien recibido en el mercado objetivo. Este método de análisis se utilizaba principalmente para diseñar campañas en los medios, pero su uso se ha ampliado para la distribución de productos y el análisis de líneas.

Escala de Razón (4)

- Características de la escala de razón:
  - **Punto de cero absoluto.** Una de las características distintivas de los datos de análisis de razón es el verdadero punto de cero absoluto, el cual hace que los datos sean relevantes y significativos de una manera que es correcto decir "un objeto es dos veces más largo que el otro" o 4 tiene el doble del valor que 2.
  - **Sin valor numérico negativo.** Los datos de escala de razón no tienen ningún valor numérico negativo. Por ejemplo, el peso no puede ser negativo, -20 Kgs no existe.
  - **Cálculo.** Los valores de datos de una escala de razón se pueden sumar, restar, multiplicar y dividir. Se puede realizar un análisis estadístico único para los datos de razón.
- Ejemplos:
  - ¿Cuál es tu peso en kilogramos?
    - Menos de 50 kgs
    - 51-60 kgs
    - 61-70 kgs
    - 71-80 kgs
    - 81-90 kgs
    - Más de 90 kgs

Propiedades fundamentales de los datos (1)

- **Escalas de datos**
  - Las medidas numéricas pueden tener diferentes significados semánticos, incluso si están representadas por los mismos datos numéricos.
    - Dependiendo del significado semántico, diferentes tipos de operaciones matemáticas son apropiadas. Para el significado semántico de la medición numérica, se tienen cuatro escalas.

Escala	Operaciones		Estadística
Razón	.	/	Media generalizada
Intervalo	+	-	Media
Ordinal	>	<	Mediana
Nominal	=	≠	Moda

Propiedades fundamentales de los datos (2)

- **Escalas de datos**
  - Para datos nominales escalados, solo son válidas las pruebas de igualdad o desigualdad.
    - Ejemplos de características nominales son nombres de personas o códigos de objetos. Los datos de una característica nominal se pueden representar por la *moda.*
      - La moda se define como el valor que se presenta con mayor frecuencia.
  - Para datos ordinales escalados son válidas las operaciones "mayor que" y "menor que".
  - Para cada nivel de la escala también son válidas las operaciones y estadística de los niveles inferiores de la escala, por lo que para la escala ordinal tenemos la igualdad, la desigualdad y las combinaciones "mayor o igual" (≥) y "menor o igual" (≤).

Propiedades fundamentales de los datos (3)

- **Escalas de datos**
  - La relación "menor o igual" (≤) define un *orden total*, tal que para cualquier *x*, *y*, *z* se tiene:
    - $(x \leq y) \wedge (y \leq x) \Rightarrow (x = y)$ (antisimetría),
    - $(x \leq y) \wedge (y \leq z) \Rightarrow (x \leq z)$ (transitividad),
    - $(x \leq y) \vee (y \leq x)$ (totalidad).
  - Ejemplos de *características ordinales* son las calificaciones escolares.
  - Los datos de una característica ordinal se pueden representar mediante la *mediana.*
    - Que se define como el valor para el cual existen (casi) tantos valores más pequeños como más grandes.
    - La mediana no es válida para características ordinales, por lo que, por ejemplo, no tiene sentido decir que la calificación escolar promedio es C.
    - Para datos escalados por intervalos, la suma y la resta son válidas. Las entidades escaladas por intervalos tienen puntos cero arbitrarios.
      - Ejemplos: temperaturas en grados Celsius o Fahrenheit, por lo que no tiene sentido decir que 40 °C es el doble que 20 °C.

Propiedades fundamentales de los datos (4)

- **Escalas de datos**
  - Los datos de una característica de intervalo, por ejemplo, dado un conjunto de valores $X = \{x_1, \ldots, x_n\}$, se puede representar mediante la media (aritmética).

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

  - Para datos en escala de proporción, la multiplicación y la división son válidas.
    - Ejemplos de características en escala de proporción son las diferencias de tiempo como edades o temperaturas en la escala Kelvin. Los datos de una característica a escala de intervalo se pueden representar mediante la media generalizada.

$$m_\alpha(X) = \sqrt[\alpha]{\frac{1}{n}\sum_{k=1}^{n} x_k^\alpha}$$

      - Con el parámetro $\alpha \in \mathbb{R} \setminus \{0\}$, que incluye los casos especiales mínimo ($\alpha \to -\infty$), media armónica ($\alpha = -1$), media geométrica ($\alpha \to 0$), media aritmética ($\alpha = 1$), media cuadrática ($\alpha = 2$), y máxima ($\alpha \to \infty$).

Propiedades fundamentales de los datos (5)

- **Mediana**
  - Se define como el número o valor central que está justo en el medio de un conjunto de datos, ordenado de menor a mayor o de mayor a menor.
    - Cuando se tiene un conjunto de datos par, no existe un valor central, entonces es necesario calcular la media aritmética de los valores centrales del conjunto.

$$X = \{1, 2, 3, 4, 5\} \longrightarrow M = 3$$

$$X = \{1, 2, 4, 5\}; \mu_A = \frac{2+4}{2} = 3 \longrightarrow M = 3$$

- **Moda**
  - Se define como el número que está representado más veces dentro de un conjunto de datos; es decir, aquel valor que presenta una mayor frecuencia absoluta dentro de la muestra.
    - La moda se calcula tanto para variables cuantitativas como cualitativas.
    - Moda Unimodal: Cuando el máximo número de repeticiones se da para un solo número.

$$Datos: [3, 5, 5, 6, 8]; M_u = 5; \rightarrow El\ valor\ de\ 5\ se\ repite\ dos\ veces$$

Propiedades fundamentales de los datos (6)

- **Moda**
  - Moda Bimodal: Cuando el máximo número de repeticiones se da para dos números.

$$Datos: [3, 5, 5, 6, 8, 8]; M_B = 5 \ \& \ 8; \rightarrow Ambos\ se\ repiten\ dos\ veces$$

  - Moda Multimodal: Cuando el máximo número de repeticiones se da para tres o más valores.

$$Datos: [3, 3, 5, 5, 6, 8, 8]; M_B = 3, 5 \ \& \ 8; \rightarrow Tres\ valores\ con\ una\ repetición\ de\ dos\ veces$$

- **Varianza**
  - Es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media.
  - La varianza, junto con la desviación estándar, son medidas de dispersión de datos u observaciones.
    - La dispersión de estos datos indica la variedad que estos presentan, es decir, si todos los valores en un conjunto de datos son iguales, entonces no hay dispersión, pero en cambio, sino todos son iguales entonces hay dispersión.

Propiedades fundamentales de los datos (7)

- **Varianza**
  - En resumen, esta dispersión puede ser grande o pequeña, dependiendo de qué tan cercanos sean los valores a la media. Procedimiento de obtención:
    - Calcular la media de conjunto muestra
    - Restar la media a cada número anterior y elevarlo al cuadrado
    - Calcular la media de las diferencias al cuadrado obtenidas en el punto anterior

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

  - Donde:
    - x_i es el número de observaciones de la variable o conjunto, tomando valores de 1 a n.
    - n es el número de observaciones.
    - $\bar{X}$ es la media del conjunto o variable.

Propiedades fundamentales de los datos (8)

- **Desviación estándar**
  - Es una medida que cuantifica la cantidad de dispersión de las observaciones en un conjunto de datos. La baja desviación estándar es un indicador de la cercanía de las puntuaciones a la media aritmética y representa una desviación estándar alta. Las puntuaciones se dispersan en un rango de valores más alto.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

Propiedades fundamentales de los datos (9)

- **Diferencias entre la varianza y la desviación estándar**
  - La varianza es un valor numérico que describe la variabilidad de las observaciones desde su media aritmética. La desviación estándar es una medida de la dispersión de observaciones dentro de un conjunto de datos.
  - La varianza no es más que un promedio de desviaciones al cuadrado. Por otro lado, la desviación estándar es la desviación cuadrática media.
  - La varianza se denota por sigma-cuadrado (σ^2) mientras que la desviación estándar se etiqueta como sigma (σ).
  - La variación se expresa en unidades cuadradas que generalmente son más grandes que los valores en el conjunto de datos dado. A diferencia de la desviación estándar, que se expresa en las mismas unidades que los valores en el conjunto de datos.
  - La varianza mide qué tan lejos están los individuos en un grupo. Por el contrario, la desviación estándar mide la cantidad de observaciones de un conjunto de datos que difiere de su media.

Propiedades fundamentales de los datos (10)

- **Escalas de datos**
  - Las características del conjunto de datos de Iris están en escala de razón. Por tanto, podemos:
    - Estimar aproximadamente el área de la superficie del sépalo multiplicando la longitud del sépalo y el ancho del sépalo.
    - Calcular la moda, la mediana, la media, la varianza y la desviación estándar de cada una de las características del conjunto de datos para las 3 clases de flores.

Proyecto 1. Hacer un programa en R que permita llevar a cabo los cálculos señalados anteriormente sobre el conjunto de datos Iris.