Taller 2: *Visualización, Ciencia de Datos y Big Data*

Tabla de contenido

Big Data, Ciencia de datos y Visualización:

Punto 1

Presente una investigación pequeña donde responda la relación e importancia que tiene la visualización de datos con: Big data, Análisis de datos, Inteligencia de Datos y Ciencia de datos.

La visualización de datos es el "puente" entre todo lo que los datos representan y el entendimiento humano. La visualización es el resultado de un procesamiento que se aplica sobre los datos con tal de ser claros y representativos.

Relación con Big data

Como sabemos, la Big data se identifica por las 5Vs: (Volumen, Valor, Veracidad, Velocidad, Variedad). Aquí uno de los mayores problemas sería el volumen, cosa que la visualización puede solucionar, ya que gracias a esta podemos simplificar grandes cantidades de datos en patrones visuales más claros para cualquier persona que interactúe con la visualización.

Relación con análisis de datos

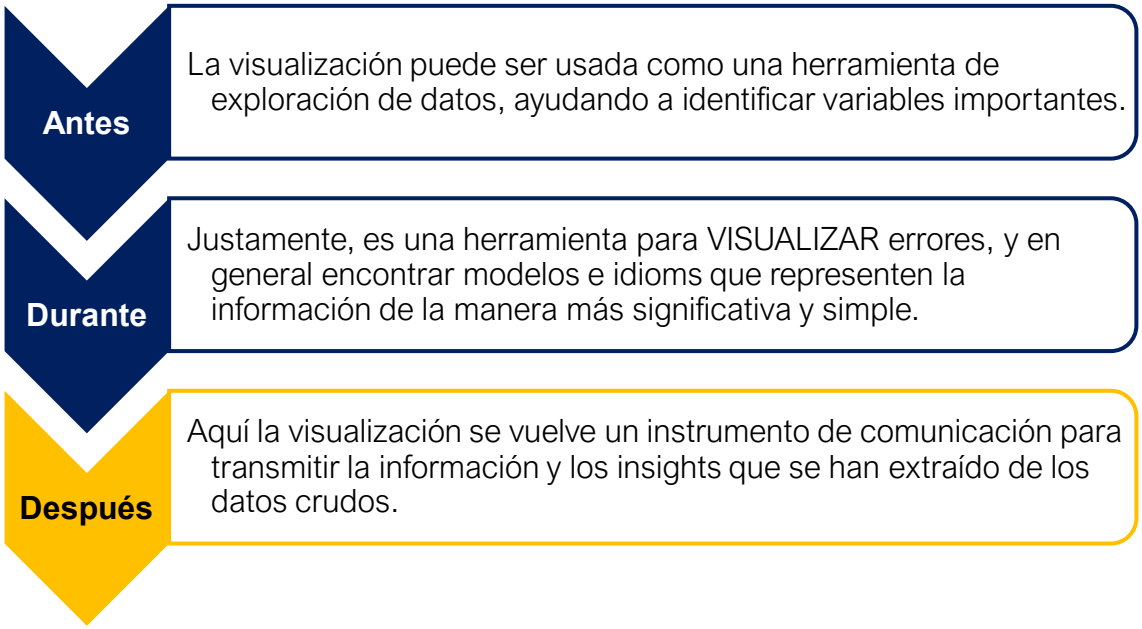
Más allá de las estadísticas y los "números", la visualización permite identificar anomalías con mayor facilidad, la fluctuación de los datos, lo que significan y que tan estrecha es la relación entre las variables. En pocas palabras, la visualización puede "hacer parte" del análisis en sí.

Relación con Inteligencia de datos

La inteligencia de datos (*Data Intelligence*) es el área donde los datos ya comienzan a generar conocimiento y el conocimiento es uno de los pilares principales para sustentar las decisiones, es la visualización de datos un medio que brinda apoyo gráfico para tomar decisiones a nivel empresarial (*Business Intelligence*).

Relación con ciencia de datos

La ciencia de datos es una disciplina que combina matemáticas, estadísticas, programación, es un proceso, y la visualización está presente siempre:



Punto 2

Incluya los niveles de abstracción de la información (del dato a la sabiduría) y la clasificación de los datos: cómo se encuentran en el mundo real y los tipos de datos existentes.

En el contexto de *Big Data* y Ciencia de Datos se reconoce una escala que va desde el dato hasta la sabiduría. El dato corresponde a mediciones aisladas y sin contexto; la información surge cuando dichos datos se organizan y se enmarcan en un contexto específico, permitiendo responder preguntas como qué, cuándo o dónde. El conocimiento se forma al identificar patrones y relaciones entre conjuntos de información, mientras que la sabiduría consiste en utilizar ese conocimiento para tomar decisiones acertadas y diseñar estrategias o políticas adecuadas.

En el mundo real, los datos se generan a partir de diversas fuentes, como sensores e infraestructura *IoT* (por ejemplo, dispositivos de localización y medidores inteligentes), sistemas transaccionales (registros de ventas, operaciones financieras, historias clínicas), interacciones digitales (registros de navegación, clics, uso de aplicaciones, redes sociales) y contenidos multimedia (imágenes, audio y video). Un caso ilustrativo es el de un sistema de transporte público, en el cual se combinan validaciones de tarjetas, coordenadas GPS de los vehículos, grabaciones de cámaras y comentarios de usuarios en redes sociales para analizar la congestión y la calidad del servicio.

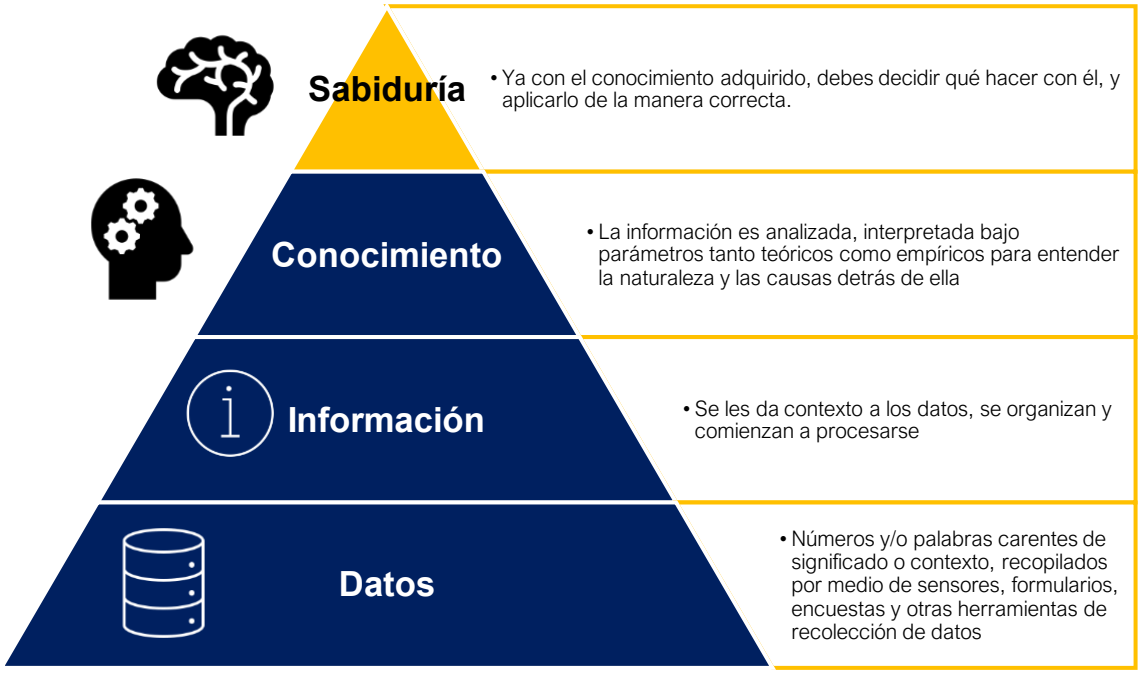
Desde el punto de vista de su estructura, los datos se clasifican en estructurados, semiestructurados y no estructurados.

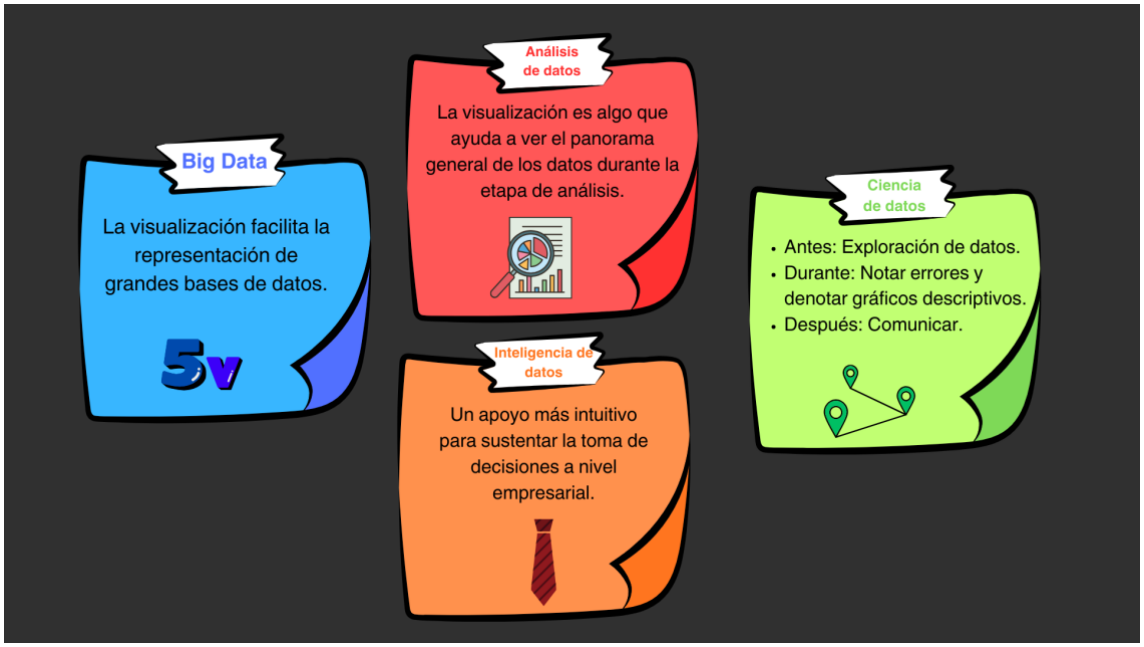
- Los datos **estructurados** se almacenan en tablas con un esquema fijo y bien definido, lo que facilita su gestión y consulta mediante lenguajes como SQL.
- Los datos **semiestructurados**, como documentos JSON o ciertos registros de log, presentan etiquetas y jerarquías pero un esquema más flexible.
- Los datos **no estructurados**, entre los que se incluyen textos libres, imágenes, audios y videos, carecen de un formato predefinido y requieren técnicas especializadas de análisis, como procesamiento de lenguaje natural y métodos de visión por computador, para extraer información relevante.

Punto 3

Genere una imagen o grafico que represente la información que presentó anteriormente.

Una ilustración muy común para representar el nivel de abstracción de la información es la Pirámide *DIKW* (*Data*, *Information*, *Knowledge*, *Wisdom*)




Big Data
La visualización facilita la representación de grandes bases de datos.
5v
Análisis de datos
La visualización es algo que ayuda a ver el panorama general de los datos durante la etapa de análisis.
Inteligencia de datos
Un apoyo más intuitivo para sustentar la toma de decisiones a nivel empresarial.
Ciencia de datos
Antes: Exploración de datos.
Durante: Notar errores y denotar gráficos descriptivos.
Después: Comunicar.

Análisis de visualizaciones:

Busque una visualización que tenga como mínimo dos modismos, debe ser interactiva.

De acuerdo con el *framework* visto, indique el *WHAT* para cada atributo (que es cada atributo). De acuerdo con el *framework* visto, indique el *WHY* para cada *idiom*, es decir, indique cual es el objetivo de cada uno de los *idioms*.

Genere conclusiones sobre la visualización y con argumentos indique si es una buena o una mala visualización.

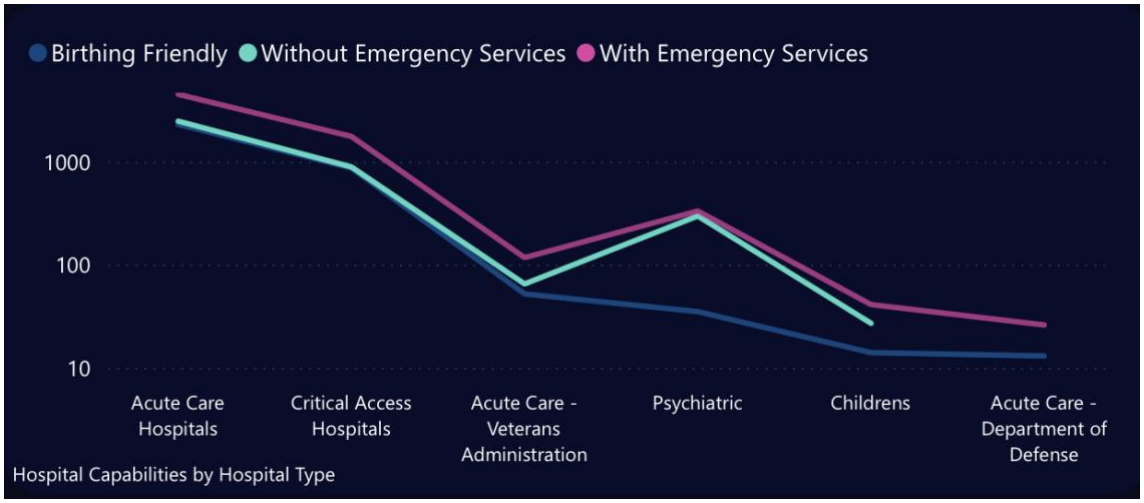
Trabajamos con este *dashboard* de PowerBI que se encuentra disponible en la *Fabric Comunity* de Microsoft.

El dashboard Hospital Quality & Ownership Analysis U.S. ofrece una panorámica estratégica del sistema hospitalario estadounidense, estructurando la información en tres niveles visuales claros: en la parte superior utiliza gráficos de anillo (Donut Charts) para resumir la calificación promedio de los pacientes según el tipo de propiedad (Gubernamental, Con y Sin Ánimo de Lucro); en la sección central detalla la infraestructura mediante comparativas de capacidades y propiedad por tipo de hospital (agudos, críticos, psiquiátricos, etc.); y en la parte inferior desglosa geográficamente la distribución de estas entidades mediante gráficos de líneas por estado. En conjunto, la herramienta permite identificar rápidamente que el sector "Nonprofit" domina en volumen y mantiene calificaciones competitivas, mientras revela brechas de servicios críticos en hospitales especializados, todo ello

presentado con una estética oscura de alto contraste que prioriza la comparación de tendencias y distribuciones sobre la precisión del dato individual.

Gráfico 1

Analicemos a profundidad el siguiente *idiom* del dashboard:



Este gráfico intenta mostrar la distribución de servicios críticos (emergencias y maternidad) a través de diferentes tipologías hospitalarias. Su meta es revelar si ciertos tipos de hospitales (como los de veteranos o niños) carecen sistemáticamente de servicios de emergencia en comparación con los hospitales generales.

Usando la metodología *What-Why-How* vista en clase podemos decir lo siguiente:

What (Datos):

Tipo de *Dataset*: Tabla multidimensional con atributos cuantitativos y categóricos.

Atributos:

- **Categórico (Eje X):** Tipo de hospital (*Acute Care*, *Critical Access*, *Psychiatric*, *Childrens*, etc.).
- **Cuantitativo (Eje Y):** Conteo de hospitales (escala logarítmica).
- **Categórico (Series/Color):** Capacidad del servicio (*Birthing Friendly*, *Without Emergency Services*, *With Emergency Services*).

Dirección de orden:

- Secuencial para el conteo de hospitales (eje y), mientras que el tipo de hospital (eje x) si bien está organizado de una manera secuencial, no tiene un orden en particular.

Why (Tarea):

Acción: *Analyze -> Consume -> Discover* (descubrir patrones generales) y *Search -> Explore* (explorar diferencias entre tipos).

Objetivo (Target): *Compare trends* (comparar la tendencia de conteo entre diferentes capacidades dentro de cada tipo de hospital) e *Identify features* (identificar qué tipo de hospital tiene más o menos servicios de emergencia).

How (Codificación Visual):

Idiom: Gráfico de líneas multi-serie (Line Chart).

Marcas: Líneas que conectan puntos.

Canales:

- **Posición Vertical (Y):** Magnitud (cantidad de hospitales), usando escala logarítmica.
- **Posición Horizontal (X):** Categoría nominal (tipos de hospital).
- **Color (*Hue*):** Categoría de servicio (Azul oscuro, Cian, Magenta).

Crítica de Diseño y Mejoras

Problema con el Eje X: Se usa un gráfico de líneas para datos categóricos nominales en el eje X. Las líneas implican continuidad entre "Acute Care" y "Critical Access", lo cual es semánticamente incorrecto porque no hay una transición gradual entre ser un hospital psiquiátrico y uno de niños.

- **Mejora:** Usar un Gráfico de Barras Agrupadas (*Grouped Bar Chart*). Las barras separan claramente las categorías discretas y permiten comparar las magnitudes lado a lado sin sugerir una secuencia falsa.

Escala Logarítmica: El eje Y usa una escala logarítmica (10, 100, 1000) sin *gridlines* claros o etiquetas intermedias suficientes. Esto dificulta la comparación precisa de alturas para una audiencia no técnica.

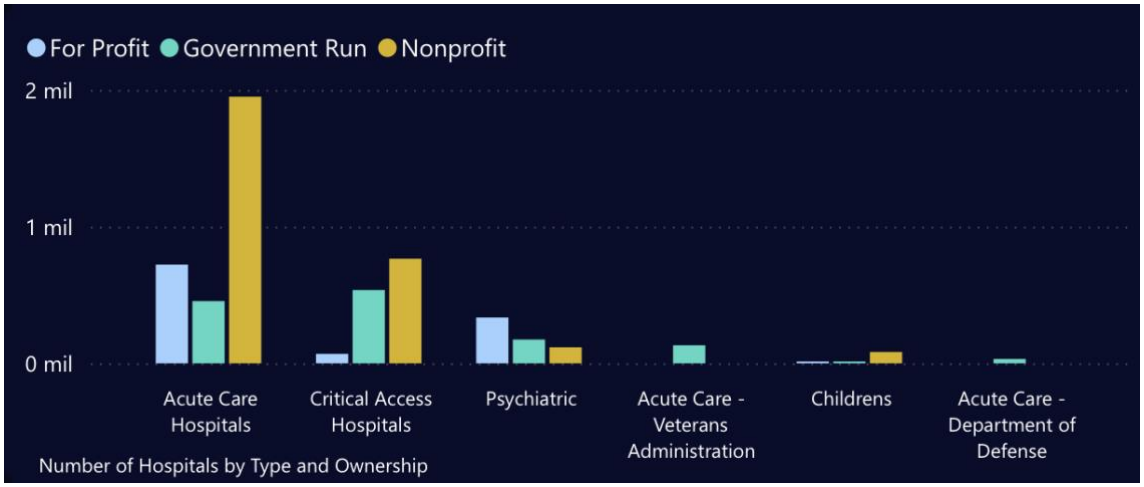
Colores: El contraste entre el azul oscuro y el fondo negro es bajo, dificultando la lectura.

Conclusiones de los Datos

- La gran mayoría de los hospitales de "*Acute Care*" y "*Critical Access*" cuentan con servicios de emergencia (línea magenta alta).
- Existe una caída drástica de servicios de emergencia en hospitales psiquiátricos y de veteranos.
- Los hospitales "*Birthing Friendly*" (amigables para partos) son significativamente menos comunes que los hospitales con emergencias generales en casi todas las categorías.

Gráfico 2

Veamos este otro *idiom* del mismo dashboard:



El objetivo es visualizar la estructura de propiedad del sistema de salud. Permite entender rápidamente quiénes son los dueños mayoritarios (privados, gobierno o sin ánimo de lucro) de la infraestructura crítica.

What (Datos):

Tipo de *Dataset*: Tabla de agregación.

Atributos:

- Categórico (Eje X): Tipo de hospital.
- Cuantitativo (Eje Y): Número de hospitales (escala lineal, miles: "2 mil").
- Categórico (Color): Propiedad (*For Profit*, *Government Run*, *Nonprofit*).

Dirección de orden:

- El número de hospitales (eje y) es secuencial (de 0 mil a 2 mil), el tipo de hospital (eje x) podría ser considerado también como secuencial solo que no tiene un orden, lo cual en si no es crítico (en este caso) para hacer la visualización más clara.

Why (Tarea):

Acción: *Analyze -> Consume -> Present* (presentar la distribución de propiedad).

Objetivo (*Target*): *Lookup* (buscar valores específicos) y *Compare* (comparar la dominancia de tipos de propiedad en cada categoría hospitalaria).

How (Codificación Visual):

Idiom: Gráfico de Barras Agrupadas (*Grouped Bar Chart*).

Marcas: Líneas (barras rectangulares).

Canales:

- **Longitud (Y):** Cantidad de hospitales.
- **Posición (X):** Categorías separadas espacialmente.
- **Color (*Hue*):** Tipo de propiedad (Azul, Cian, Amarillo).

Crítica de Diseño y Mejoras

Etiquetado del Eje Y: La etiqueta "2 mil" y "1 mil" es poco estándar y precisa ("2k" o "2,000" sería mejor). Además, muchas barras son tan pequeñas (cercanas a 0) que se vuelven invisibles en esta escala lineal dominada por la barra amarilla más alta.

- **Mejora:** Un gráfico de barras apiladas al 100% podría ser útil si el interés es comparar la proporción de propiedad en lugar de los totales absolutos, o dividir el gráfico en paneles (*Small Multiples*) para las categorías con menos hospitales (*Childrens*, *Veterans*) que quedan visualmente aplastadas.

Uso del Espacio: Las categorías menores (*Psychiatric*, *Veterans*, *Childrens*) tienen barras casi indistinguibles. Esto hace que el gráfico sea inútil para analizar esos sectores específicos.

Conclusiones de los Datos

- El sector "*Nonprofit*" (sin ánimo de lucro, barras amarillas) domina masivamente el mercado de hospitales de cuidados agudos (*Acute Care*), superando por mucho a los gubernamentales y con fines de lucro.
- Los hospitales de "*Critical Access*" también son mayoritariamente "*Nonprofit*" o gubernamentales, con muy poca participación privada con fines de lucro.
- La propiedad gubernamental (*Government Run*) tiene una presencia notable en hospitales psiquiátricos en comparación con otras categorías.