Universidad Tecnológica de Chihuahua

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

V.1. Reporte de investigación de técnicas de visualización

Enrique Mascote
RICARDO ALONSO RIOS MONRREAL

2. Introducción

En el contexto actual de la ingeniería de datos y la inteligencia de negocios, la capacidad de recolectar y procesar grandes volúmenes de información (Big Data) ha dejado de ser el único diferenciador competitivo. El verdadero valor reside en la capacidad de interpretar estos datos y comunicarlos efectivamente para la toma de decisiones estratégicas. Aquí es donde la visualización de datos actúa como el puente crítico entre el análisis técnico y la cognición humana.

La visualización de datos no es simplemente una cuestión estética; es una disciplina que combina principios matemáticos, estadísticos y de psicología cognitiva para representar información abstracta de manera gráfica. Un conjunto de datos crudo, por más valioso que sea, es ininteligible para el cerebro humano sin un procesamiento previo. Las representaciones visuales aprovechan la capacidad pre-atenta del cerebro para detectar patrones, tendencias y anomalías en milisegundos.

El objetivo del presente reporte es investigar y documentar exhaustivamente las metodologías actuales para la representación de información. El alcance abarca desde los fundamentos teóricos y principios de diseño, pasando por el proceso de *Storytelling* con datos, hasta el análisis técnico de las herramientas de Business Intelligence (BI) y bibliotecas de programación más utilizadas en la industria tecnológica moderna.

3. Fundamentos de Visualización de Datos

3.1 Conceptos básicos

La creación de una visualización efectiva se basa en reglas fundamentales que aseguran la integridad de los datos y la claridad del mensaje.

- **Sistemas de coordenadas:**
    - *Cartesiano:* El estándar más común (ejes X e Y perpendiculares). Ideal para comparaciones precisas de longitud y posición.
    - *Polar:* Utiliza un ángulo y un radio desde el centro. Útil para datos cíclicos, aunque más difícil de interpretar con precisión para el ojo humano.
    - *Geográfico:* Proyecciones (como Mercator) utilizadas para mapear datos sobre la superficie terrestre.
- **Tipos de ejes:**
    - *Cuantitativos (Continuos):* Representan valores numéricos donde la distancia entre puntos es matemáticamente consistente (ej. temperatura, ingresos).
    - *Categóricos (Discretos):* Representan etiquetas o grupos sin un orden numérico inherente (ej. nombres de países, departamentos).

    - *Temporales:* Un caso especial de eje cuantitativo dedicado al flujo del tiempo.

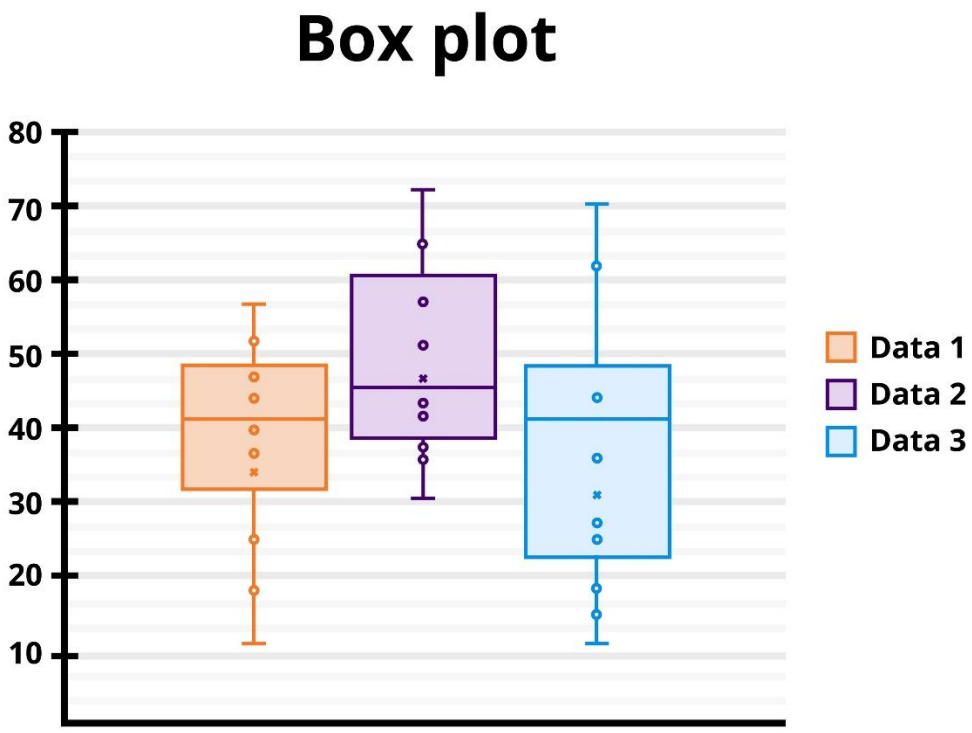
- **Esquemas de colores y psicología:** El color es una herramienta funcional, no decorativa. Se clasifica en tres esquemas:
    1. *Secuencial:* Gradientes de un solo tono (de claro a oscuro) para representar magnitud en datos ordenados (ej. densidad de población).
    2. *Divergente:* Dos colores contrastantes que convergen en un punto medio neutro (generalmente cero o la media). Vital para visualizar desviaciones positivas y negativas.
    3. *Cualitativo (Categórico):* Colores distintivos para separar grupos que no tienen orden intrínseco (ej. rojo para A, azul para B). *Psicología:* Se debe considerar el daltonismo (evitar combinaciones rojo-verde) y las asociaciones culturales (rojo suele implicar "peligro" o "pérdida" en contextos financieros occidentales).
- **Principios de diseño visual:** Se destaca el concepto de *Data-Ink Ratio* de Edward Tufte, que postula que la "tinta" utilizada en un gráfico debe dedicarse tanto como sea posible a los datos y no a elementos decorativos (bordes, fondos oscuros, rejillas innecesarias). Menos ruido visual equivale a mayor comprensión.

3.2 Tipos de representación gráfica

A. Visualización de cantidad Comparación de magnitudes entre categorías.

- **Gráficos de barras/columnas:** Barras rectangulares con longitud proporcional al valor.
    - *Uso:* Comparar valores discretos. Es la visualización más precisa para el ojo humano.
- **Pictogramas:** Uso de iconos para representar cantidades.
    - *Uso:* Infografías o audiencias no técnicas. Menos preciso pero alto impacto visual.

B. Visualización de distribución



Cómo se dispersan los datos alrededor de una media o mediana.

- **Histogramas:** Agrupan datos continuos en "bins" o intervalos.
    - *Uso:* Ver la forma de la distribución (normal, sesgada, bimodal).
- **Box Plots (Diagrama de Caja y Bigotes):** Muestran la mediana, cuartiles (Q1, Q3) y valores atípicos (outliers).
    - *Uso:* Comparar distribuciones estadísticas entre varios grupos sin saturar la vista.
- **Gráficos de violín:** Combinan un box plot con una estimación de densidad de kernel (KDE).
    - *Uso:* Cuando se necesita ver la densidad de probabilidad de los datos, no solo los cuartiles.

C. Visualización de proporción Relación de las partes con el todo.

- **Gráficos de pastel/donas:** Dividen un círculo en sectores.

    - *Uso:* Mostrar participaciones de mercado simples (máximo 3-4 categorías). *Advertencia:* El cerebro juzga mal las áreas y ángulos; usar con precaución.
- **Treemaps (Mapas de árbol):** Rectángulos anidados.
    - *Uso:* Visualizar datos jerárquicos y proporciones cuando hay muchas categorías.

D. Visualización de relación XY Correlaciones entre dos o tres variables.

- **Diagrama de dispersión (Scatter Plot):** Puntos en coordenadas cartesianas.
    - *Uso:* Detectar correlación, clusters o outliers entre dos variables numéricas.
- **Gráfico de burbujas:** Scatter plot donde el tamaño del punto representa una tercera variable.

E. Visualización de datos geoespaciales

- **Mapas de calor (Heatmaps):** Intensidad de color según la concentración de valores.
- **Coropletas:** Regiones geográficas (países, estados) coloreadas según una variable estadística.

F. Visualización de incertidumbre

- **Gráficos de error / Intervalos de confianza:** Barras sobrepuestas en puntos o columnas que indican el margen de error estándar o desviación típica. Fundamental en ingeniería y ciencia para mostrar la fiabilidad del dato.

4. Proceso de Storytelling con Datos

4.1 Definición y componentes

El *Storytelling con datos* es la habilidad de construir una narrativa convincente basada en análisis complejos, guiando a la audiencia hacia una conclusión o acción específica. No se trata de "inventar" historias, sino de estructurar la evidencia.

- **Componentes:**
    1. *Datos:* La evidencia veraz y analizada.
    2. *Visualización:* La representación que hace visible el patrón.
    3. *Narrativa:* El hilo conductor que explica el "por qué" y el "qué sigue".

4.2 Etapas del proceso

1. **Comprensión de la audiencia:** ¿A quién me dirijo? (Ej. Un CFO busca rentabilidad; un ingeniero busca eficiencia técnica). Conocer su nivel de alfabetización de datos.
2. **Definición del mensaje clave:** Si tuvieras solo 3 minutos, ¿qué deberían saber? Este es el "Big Idea".
3. **Selección de datos relevantes:** Filtrar el ruido. No mostrar todos los datos exploratorios, solo los explicativos.
4. **Diseño de la narrativa visual:** Elegir el gráfico correcto y resaltar (con color o texto) solo lo importante.
5. **Presentación efectiva:** El orden de las diapositivas debe seguir una lógica: Problema -> Contexto -> Análisis -> Solución.

4.3 Mejores prácticas

- **Estructura:** Inicio (Contexto), Medio (El conflicto/análisis), Fin (La resolución/recomendación).
- **Errores comunes:** Saturar con efectos 3D (distorsionan la realidad), falta de etiquetas en ejes, usar demasiados colores sin significado, y no concluir nada.
- **Caso exitoso:** El periodismo de datos del *New York Times* o *Financial Times*, donde gráficos interactivos guían al lector paso a paso por la noticia, en lugar de soltar una tabla estática.

5. Herramientas de Visualización (Business Intelligence)

A. Tableau

- **Descripción:** Plataforma líder en el mercado de BI, conocida por su motor gráfico "VizQL" que traduce acciones de arrastrar y soltar en consultas de base de datos.
- **Características:** Interfaz intuitiva *drag-and-drop*, capacidad de manejar millones de filas, dashboards interactivos robustos.
- **Ventajas:** Potencia visual superior, gran comunidad, conexión a casi cualquier fuente de datos.
- **Desventajas:** Curva de aprendizaje moderada-alta para funciones avanzadas, costo de licencia elevado.
- **Gráficas:** Soporta todas las mencionadas en la sección 3 y permite mapas personalizados complejos.

B. Power BI (Microsoft)

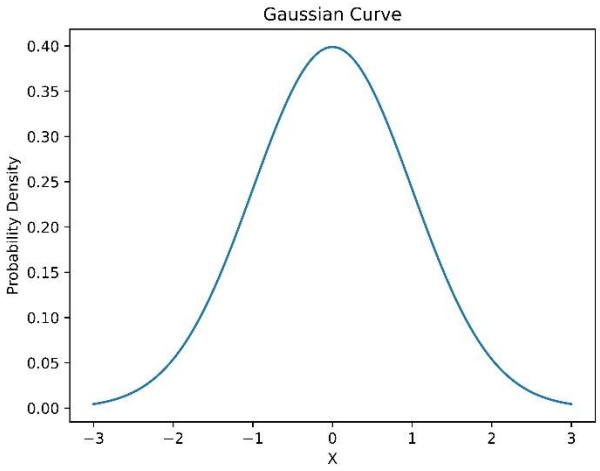
- **Descripción:** Servicio de análisis de negocios de Microsoft que proporciona visualizaciones interactivas con capacidades de autoservicio.
- **Características:** Integración nativa con Excel y Azure, lenguaje DAX (Data Analysis Expressions) para cálculos complejos.
- **Ventajas:** Costo accesible (versión desktop gratuita), familiaridad para usuarios de Excel, actualizaciones mensuales.
- **Desventajas:** La versión gratuita tiene límites de datos en la nube; DAX puede ser complejo de dominar.
- **Gráficas:** Amplia galería nativa y un "Marketplace" de visualizaciones creadas por la comunidad.

C. Google Data Studio (Looker Studio)

- **Descripción:** Herramienta gratuita basada en la nube para convertir datos en informes informativos y personalizables.
- **Características:** Colaboración en tiempo real (similar a Google Docs), conexión nativa a Google Analytics, Sheets y Ads.
- **Ventajas:** 100% web, gratuito, curva de aprendizaje muy baja.
- **Desventajas:** Capacidades de modelado de datos limitadas en comparación con Tableau/PowerBI, lento con grandes volúmenes de datos.
- **Gráficas:** Gráficos estándar, series temporales, tablas dinámicas y mapas de Google.

6. Bibliotecas de Visualización (Programación)

A. Matplotlib (Python)



- **Filosofía:** Es la biblioteca "abuelo" de la visualización en Python. Busca replicar la capacidad de graficado de MATLAB.
- **Características:** Control total sobre cada píxel del gráfico.
- **Nivel de complejidad:** Medio-Alto (muy verboso).
- **Ejemplo de código:**

Python

```
import matplotlib.pyplot as plt

x = [1, 2, 3, 4]

y = [10, 20, 25, 30]

plt.plot(x, y)

plt.title("Gráfico Lineal Simple")

plt.show()
```

B. Seaborn (Python)

- **Filosofía:** Construida sobre Matplotlib, orientada a la visualización de datos estadísticos atractivos por defecto.
- **Características:** Funciona excelentemente con DataFrames de Pandas. Integra cálculos estadísticos (regresiones) automáticamente.
- **Nivel de complejidad:** Bajo-Medio.
- **Ejemplo de código:**

Python

```
import seaborn as sns

# Crea un gráfico de dispersión con línea de regresión

sns.lmplot(x="total_bill", y="tip", data=tips_dataset)
```

C. ggplot2 (R)

- **Filosofía:** Basada en *The Grammar of Graphics* (Leland Wilkinson). Construye gráficos por capas (datos + geometría + estética).
- **Características:** Sintaxis declarativa muy potente. Es el estándar de oro en la estadística académica.
- **Nivel de complejidad:** Medio (requiere entender la gramática de capas).

- **Ejemplo de código:**

R

```
library(ggplot2)
ggplot(data, aes(x=mpg, y=hp)) +
  geom_point() +
  theme_minimal()
```

D. D3.js (JavaScript)

- **Filosofía:** *Data-Driven Documents*. Manipula el DOM (Document Object Model) web basándose en datos.
- **Características:** Permite crear cualquier visualización imaginable, interactiva y animada para la web. No es una biblioteca de gráficos *per se*, sino de manipulación de elementos SVG.
- **Nivel de complejidad:** Muy Alto.
- **Ejemplo de código:** (Fragmento conceptual)

JavaScript

```
d3.select("body").selectAll("p")
  .data([4, 8, 15, 16, 23, 42])
  .enter().append("p")
  .text(function(d) { return "Valor: " + d; });
```

7. Conclusiones

La realización de esta investigación permite concluir que la visualización de datos es una competencia transversal crítica en la ingeniería. No basta con generar modelos predictivos precisos si sus resultados no pueden ser comunicados a los *stakeholders*.

Al comparar herramientas (BI) versus bibliotecas (Código), se identifica un compromiso claro: las herramientas como Tableau o Power BI ofrecen velocidad y facilidad de uso, ideales para reportes corporativos ágiles y usuarios de negocio. Por otro lado, bibliotecas como D3.js o Matplotlib ofrecen flexibilidad infinita y reproducibilidad científica, siendo preferibles para el desarrollo de software a medida o investigación académica rigurosa.

La elección de la visualización adecuada (barras vs. pastel, lineal vs. dispersión) puede alterar completamente la percepción de la realidad. Como ingenieros, tenemos la

responsabilidad ética de elegir representaciones que no solo sean estéticas, sino veraces y exentas de manipulación visual.

Referencias

Getting Started with Tableau. (n.d.). Tableau. https://www.tableau.com/learn/get-started

JulCsc. (n.d.). *Documentación de Power BI - Power BI*. Microsoft Learn. https://learn.microsoft.com/es-es/power-bi/

Python for data science, AI & development. (2020, December 1). Coursera. https://www.coursera.org/learn/python-for-applied-data-science-ai/paidmedia?utm_medium=sem&utm_source=gg&utm_campaign=b2c_latam_ibm-data-analyst_ibm-skills-network_ftcof_professional-certificates_px_dr_bau_gg_sem_pr-bd_s1-v2_en_m_hyb_24-06_x&campaignid=21345975099&adgroupid=162526353105&device=c&keyword=&matchtype=&network=g&devicemodel=&creativeid=701297776269&assetgroupid=&targetid=dsa-2382539636588&extensionid=&placement=&gad_source=1&gad_campaignid=21345975099&gbraid=0AAAADdKX6YpgStoUsx5vTOUoxC548nA5&gclid=CjwKCAiA3L_JBhAlEiwAlcWO531awx0ejWycHdqlqWn8JZPc5pO4fl4XD2vaZXoc7TqWzqffFbWPlhoCxhsQAvD_BwE

Storytelling with Data. (n.d.). Google Books. https://books.google.com.mx/books?id=retRCgAAQBAJ&printsec=copyright#v=onepage&q&f=false