

## 5. Figuring Stuff Out: Data Analysis

Kelly P. Vincent

Este capítulo de “A Friendly Guide to Data Science” es fundamental porque sienta las bases de lo que significa realmente analizar datos.

### **La Idea Principal: Para ser un Buen Científico de Datos, Primero Debes ser un Buen Analista de Datos**

El capítulo empieza estableciendo una idea crucial: la **Ciencia de Datos** y el **Análisis de Datos** no son lo mismo, pero están íntimamente relacionados. Piénsalo como un continuo:

Análisis de Datos <----- solapamiento -----> Ciencia de Datos

- **Análisis de Datos:** Se enfoca en explorar los datos para entender **lo que pasó** (descriptivo) y **por qué pasó** (diagnóstico). Es la base.
- **Ciencia de Datos:** A menudo va más allá, enfocándose en **lo que pasará** (predictivo) y **qué deberíamos hacer** (prescriptivo), usando técnicas más avanzadas como el *machine learning*.

La tesis del autor es simple: **No puedes hacer buena ciencia de datos si no entiendes los datos, y la única forma de entenderlos es a través de un buen análisis de datos.** Por lo tanto, el análisis de datos es un prerequisito y una parte esencial de cualquier proyecto de ciencia de datos.

---

### **1. La Historia del Análisis de Datos**

El autor nos da un breve recorrido histórico para poner todo en contexto.

- **Desde la Antigüedad:** El análisis de datos ha existido desde que existen los datos. El objetivo siempre ha sido el mismo: entender mejor lo que representan.

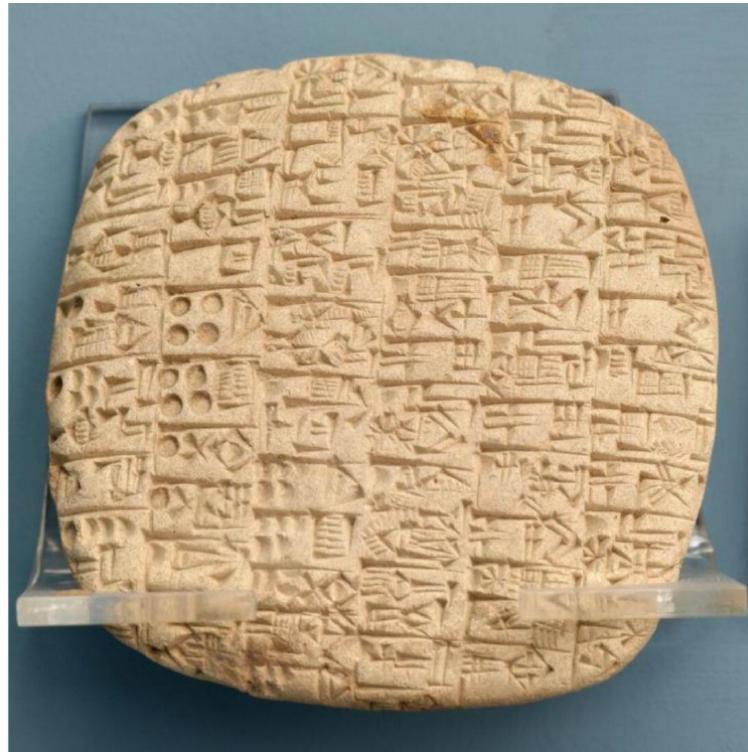


Figura 5-1. Antigua tablilla sumeria de circa 2,500 a.C. que registra cantidades de varias mercancías (cebada, harina, pan y cerveza).

- **La Revolución del Ordenador:** Antes de los ordenadores, el análisis se hacía a mano y con cantidades muy pequeñas de datos. La llegada del ordenador lo cambió todo, permitiendo analizar grandes volúmenes de datos.
  - **Primeras Herramientas:** Se crearon lenguajes y programas específicos para el análisis estadístico, como **SAS** (muy usado en seguros) y **SPSS** (muy usado en ciencias sociales).
  - **La Era del Código Abierto:** Más recientemente, herramientas propietarias y caras como SAS han sido reemplazadas en gran medida por lenguajes de código abierto y gratuitos como **R** (creado por estadísticos) y **Python** (un lenguaje de propósito general con potentes librerías estadísticas), que son los estándares de la industria hoy en día.

## 2. Ejemplos de Análisis de Datos en el Mundo Real

Para ilustrar el poder del análisis, el autor presenta dos ejemplos muy diferentes pero icónicos.

### Ejemplo 1: Moneyball (El Béisbol en la Era de los Datos)

- **El Problema:** En los años 90, el equipo de béisbol Oakland A's perdió a sus jugadores estrella porque no podía competir con los salarios de los equipos ricos.
- **El Método Antiguo:** Los ojeadores se basaban en la intuición y estadísticas tradicionales (como *home runs*) para fichar jugadores.
- **El Nuevo Enfoque (Análisis de Datos):** En lugar de confiar en la intuición, el gerente Billy Beane empezó a analizar estadísticas más oscuras pero más reveladoras (como el porcentaje de veces que un jugador llega a base).
- **El Resultado:** Descubrieron jugadores de alta calidad que estaban infravalorados por el resto de la liga. Pudieron construir un equipo ganador con una fracción del presupuesto de sus rivales. Este enfoque, popularizado por el libro y la película "Moneyball", cambió el béisbol para siempre.

### Ejemplo 2: Deteniendo el Córula en Londres en 1854

- **El Problema:** Una terrible epidemia de cólera asolaba Londres.
- **La Creencia de la Época:** Se pensaba que las enfermedades se propagaban por el "aire malo" (teoría miasmática).
- **El Análisis de Datos de John Snow:** El doctor John Snow tuvo una idea revolucionaria. En lugar de especular, creó un **mapa** de la ciudad y marcó con una barra la ubicación de cada muerte por cólera.
- **El Resultado:** La visualización mostró claramente que las muertes se concentraban masivamente alrededor de una bomba de agua específica en Broad Street. Snow usó esta evidencia para convencer a las autoridades de que clausuraran la bomba. La epidemia se detuvo casi de inmediato. Este es uno de los ejemplos más famosos de la historia de cómo la visualización y el análisis de datos pueden resolver problemas del mundo real y salvar vidas.

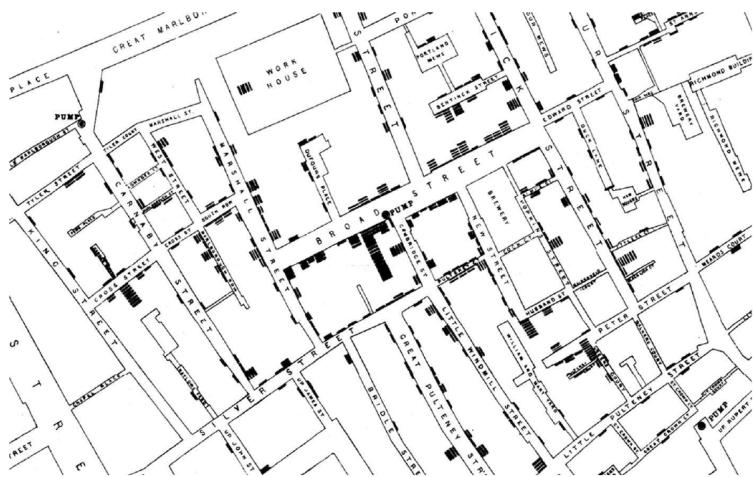


Figura 5-2. El mapa creado por John Snow que muestra las muertes por la epidemia de cólera de 1854.

### 3. Habilidades Fundamentales de un Analista de Datos

Para ser un buen analista (y por extensión, un buen científico de datos), se necesita una combinación de cuatro tipos de habilidades.

#### a) Habilidades Funcionales

Son las habilidades de alto nivel sobre cómo pensar y abordar un problema. Incluyen:

- Pensamiento lógico y sistemático.
- Curiosidad y creatividad para ver los datos desde nuevos ángulos.
- Una base sólida en matemáticas y estadística.
- “Mentalidad de datos” (*data mindset*): Tener una mente abierta, no saltar a conclusiones, y respetar los datos como una representación de la realidad.

#### b) Habilidades Técnicas

Son las herramientas que usas para hacer el trabajo.

- Nivel Básico: Dominio de hojas de cálculo como Microsoft Excel.

- **Nivel Intermedio:** Conocimiento de bases de datos y el lenguaje **SQL** para consultarlas.
- **Nivel Avanzado:** Programación en **Python** o **R**.

c) **Habilidades Blandas (Soft Skills)**

A menudo son las más importantes.

- **Comunicación:** La habilidad de explicar tus hallazgos complejos de una manera simple a personas no técnicas. Si no te entienden, no confiarán en tu trabajo.
- **Gestión del tiempo y priorización:** Saber manejar múltiples proyectos y plazos.
- **Adaptabilidad y mentalidad de crecimiento:** El campo cambia constantemente, así que debes estar siempre aprendiendo.
- **Ética:** Entender el impacto que tu trabajo puede tener.

d) **Conocimiento del Dominio (Domain Knowledge)**

- **¿Qué es?:** Es el **conocimiento experto sobre el área de la que provienen los datos**. Si analizas datos financieros, necesitas saber de finanzas. Si analizas datos de un videojuego, necesitas entender cómo funciona el juego.
  - **¿Por qué es crucial?:** **Sin conocimiento del dominio, es imposible hacer un buen análisis.** No puedes interpretar los datos correctamente ni hacer las preguntas correctas. El ejemplo de John Snow es perfecto: su conocimiento como médico (dominio) le permitió conectar los puntos y darse cuenta de que la fuente de agua era el problema, y no una “nube de aire malo”.
- 

#### 4. CRISP-DM: El Proceso de Análisis de Datos

Aunque cada proyecto es diferente, la mayoría sigue un ciclo de vida general. El más famoso y utilizado en la industria es **CRISP-DM (CRoss-Industry Standard Process for Data Mining)**.

Es un proceso **iterativo**, lo que significa que constantemente vas y vienes entre los pasos. No es una línea recta.

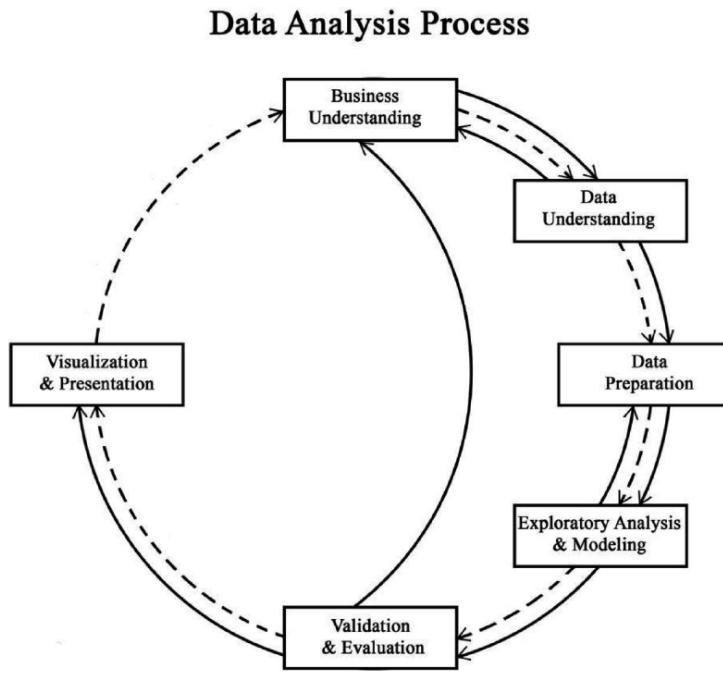


Figura 5-3. El proceso de análisis de datos CRISP-DM.

Los 6 pasos son:

**1. Comprensión del Negocio (Business Understanding):**

- **¿Qué es?**: El punto de partida. Debes entender qué problema de negocio se intenta resolver y definir las preguntas de investigación.
- **Ejemplo**: Un equipo de fútbol quiere saber a qué jugadores defensivos debería traspasar.

**2. Comprensión de los Datos (Data Understanding):**

- **¿Qué es?**: Una vez que sabes lo que buscas, necesitas encontrar y entender los datos disponibles.
- **Ejemplo**: Para el equipo de fútbol, identificamos que necesitamos datos sobre placajes, capturas (*sacks*), intercepciones y balones sueltos (*fumbles*). Luego buscamos en las bases de datos de la organización dónde se encuentra esa información.

### 3. Preparación de los Datos (Data Preparation):

- **¿Qué es?**: Esta suele ser la fase que más tiempo consume. Implica limpiar los datos (ej: corregir errores de texto), manejar valores faltantes, y a veces, **ingeniería de características** (*feature engineering*), que es crear nuevas variables a partir de las existentes para mejorar el análisis.
- **Ejemplo**: Unir tablas para crear una vista única de las estadísticas por jugador, y decidir que los valores nulos se reemplazarán por ceros.

### 4. Análisis Exploratorio y Modelado (Exploratory Analysis and Modeling):

- **¿Qué es?**: ¡La parte divertida! Aquí finalmente “te sumerges en los datos”. Se utiliza la **estadística descriptiva** para generar resúmenes (media, mediana, etc.) y, sobre todo, **visualizaciones** (gráficos de barras, histogramas, etc.) para encontrar patrones, anomalías y *insights*. Esto se conoce como **EDA** (*Exploratory Data Analysis*).
- **Ejemplo**: Crear gráficos para ver la distribución de placajes por jugador, identificar a los jugadores con rendimientos atípicos (*outliers*), y empezar a construir un modelo simple (como una regresión lineal) para dar una “puntuación” a cada jugador.

### 5. Validación y Evaluación (Validation and Evaluation):

- **¿Qué es?**: “Revisar tu trabajo”. Debes asegurarte de que tus cálculos son correctos, que no has malinterpretado los datos y que tus conclusiones son sólidas. También es común que los compañeros revisen tu trabajo (*peer review*).
- **Ejemplo**: Un compañero revisa tu código para asegurarse de que el cálculo de la puntuación de los jugadores es lógicamente correcto y no tiene errores.

### 6. Visualización y Presentación (Visualization and Presentation):

- **¿Qué es?**: El paso final. Compilas todos tus resultados y los presentas a los interesados (*stakeholders*) de una manera clara y convincente.
- **Ejemplo**: Crear una presentación con los gráficos más importantes y una tabla clasificada de los jugadores, explicando tus recomendaciones sobre a quién traspasar. Para un uso continuo, esto podría convertirse en un *dashboard* interactivo.