Chapter 1: Exploratory Data Analysis Fundamentals

Suresh Kumar Mukhiya, Usman Ahmed

Este capítulo introductorio de "Hands-On Exploratory Data Analysis with Python" es clave porque define qué es el Análisis Exploratorio de Datos (EDA) y por qué es tan importante.

La Idea Principal: Conoce Tus Datos Antes de Construir Nada

Imagina que eres un chef y te dan una caja de ingredientes misteriosos. No te pondrías a cocinar un plato complejo de inmediato, ¿verdad? Primero, abrirías la caja, olerías los ingredientes, quizás probarías un poco, verías qué texturas tienen. En resumen, **explorarías** lo que tienes.

El **Análisis Exploratorio de Datos (EDA)** es exactamente eso, pero con datos. Es el proceso de **"familiarizarse" con un conjunto de datos** para descubrir sus características principales. En lugar de saltar directamente a construir un modelo de *machine learning* sofisticado, primero usas estadísticas y visualizaciones simples para:

- Descubrir patrones.
- Detectar anomalías y errores (datos sucios).
- Probar hipótesis iniciales.
- Entender la estructura de los datos.

La filosofía, popularizada por el estadístico John Tukey, es simple: deja que los datos te hablen y te cuenten su historia **antes** de que intentes forzarlos a encajar en un modelo.

---

1. Entendiendo la Ciencia de Datos (y el Papel del EDA)

El autor comienza situando el EDA dentro del campo más amplio de la Ciencia de Datos. El proceso general de la ciencia de datos (similar al framework CRISP-DM) incluye varias etapas, y el EDA es una de las más importantes:

1. **Requisitos de los Datos**: Definir qué datos se necesitan para resolver un problema.

2. **Recolección de Datos**: Juntar los datos de diversas fuentes.

3. **Procesamiento de Datos**: Estructurar y organizar los datos.

4. **Limpieza de Datos**: ¡Una de las más importantes! Corregir errores, eliminar duplicados, manejar valores faltantes. No puedes hacer un buen análisis con datos "sucios".

5. **EDA (Análisis Exploratorio de Datos)**: Aquí es donde **finalmente empiezas a entender los datos**. Usas estadísticas descriptivas y visualizaciones para ver qué tienes entre manos. **Este libro se centra en esta etapa**.

6. **Modelado y Algoritmos**: Ahora sí, construyes modelos (ej: de regresión, clasificación) para hacer predicciones.

7. **Producto de Datos**: Creas una aplicación que usa el modelo (ej: un sistema de recomendación).

8. **Comunicación**: Presentas tus hallazgos a los interesados, a menudo a través de visualizaciones.

---

2. Pasos Fundamentales del Proceso de EDA

El autor desglosa el proceso de EDA en cuatro grandes pasos:

1. **Definición del Problema**: Antes de tocar cualquier dato, debes entender claramente el objetivo del negocio. ¿Qué pregunta estamos tratando de responder? ¿Cuál es el objetivo final?

2. **Preparación de los Datos**: Esto implica encontrar las fuentes de datos correctas, limpiarlas y transformarlas en un formato útil para el análisis.

3. **Análisis de Datos**: Este es el corazón del EDA. Se usan diversas técnicas para resumir los datos, encontrar correlaciones y relaciones ocultas, y desarrollar modelos predictivos iniciales.

4. **Desarrollo y Representación de Resultados**: Presentar los hallazgos de una

manera que sea fácil de interpretar para los interesados, usando gráficos, tablas y mapas.

---

3. Dando Sentido a los Datos (Tipos de Datos)

Para analizar los datos correctamente, primero debes saber qué tipo de datos tienes. El autor los divide en dos grandes grupos:

a) Datos Numéricos (Cuantitativos)

Representan una medida o una cantidad. Se pueden subdividir en:

- **Datos Discretos**: Son contables y tienen un número finito de valores. Piensa en ellos como "números enteros".
  - **Ejemplo**: El número de caras al lanzar una moneda 200 veces (puede ser 0, 1, 2, ..., 200, pero no 1.5). El número de estudiantes en una clase.
- **Datos Continuos**: Pueden tomar cualquier valor dentro de un rango. Piensa en ellos como "números con decimales".
  - **Ejemplo**: La temperatura de tu ciudad (puede ser 25.1°C, 25.11°C, etc.). El peso de una persona.

b) Datos Categóricos (Cualitativos)

Representan una característica o una etiqueta. No se pueden medir numéricamente.

- **Ejemplos**: Género (Hombre, Mujer, Otro), estado civil, género de una película (Acción, Comedia, Drama).
- **Subtipos**:
  - **Binarios (Dicotómicos)**: Solo dos valores posibles (ej: sí/no, verdadero/falso, éxito/fracaso).
  - **Politómicos**: Más de dos valores posibles (ej: estado civil).

---

4. Escalas de Medición

Este es un concepto estadístico clave que te dice qué tipo de operaciones matemáticas puedes hacer con tus datos. Entender esto es crucial para elegir el análisis y la visualización correctos.

1. **Escala Nominal**:

   - **¿Qué es?**: Son etiquetas puras. Los valores no tienen un orden ni una jerarquía.
   - **Ejemplos**: Género, color de pelo, país de origen.
   - **¿Qué puedes hacer?**: Solo contar frecuencias (ej: "hay un 60% de mujeres y un 40% de hombres"). No puedes calcular un promedio.
   - **Visualización**: Gráficos de barras, gráficos de pastel.

2. **Escala Ordinal**:

   - **¿Qué es?**: Las etiquetas tienen un **orden** o una jerarquía, pero la distancia entre ellas no es uniforme ni medible.
   - **Ejemplos**: La **escala de Likert** ("Totalmente en desacuerdo", "En desacuerdo", "Neutral", "De acuerdo", "Totalmente de acuerdo"). Niveles de satisfacción ("Muy insatisfecho", "Insatisfecho", "Satisfecho"). Tallas de ropa (S, M, L).
   - **¿Qué puedes hacer?**: Contar frecuencias y calcular la **mediana** o la moda. **No puedes calcular el promedio** (no tiene sentido promediar "Satisfecho" y "Neutral").

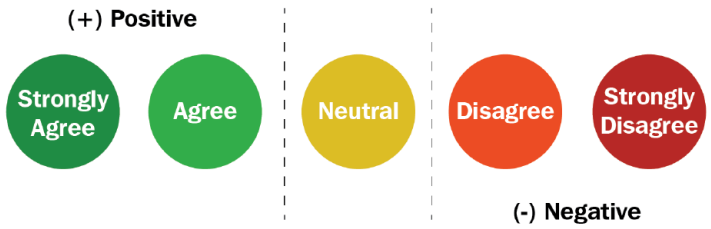

Figura: Ejemplos de la escala de Likert utilizada para medir sentimientos o satisfacción en una escala ordinal.

How do you feel today?	How satisfied are you with our service?
● 1 - Very Unhappy	● 1 - Very Unsatisfied
○ 2 - Unhappy	○ 2 - Somewhat Unsatisfied
○ 3 - OK	○ 3 - Neutral
○ 4 - Happy	○ 4 - Somewhat Satisfied
○ 5 - Very Happy	○ 5 - Very Satisfied

Figura: Otro ejemplo de la escala de Likert.

3. **Escala de Intervalo**:

   - **¿Qué es?**: Los valores tienen un orden y la **distancia entre ellos es uniforme y medible**, pero **no hay un cero absoluto**.

   - **Ejemplos**: La temperatura en grados Celsius o Fahrenheit. La diferencia entre 10°C y 20°C es la misma que entre 20°C y 30°C. Pero 0°C no significa "ausencia de temperatura". Por eso, no puedes decir que 20°C es "el doble de caliente" que 10°C.

   - **¿Qué puedes hacer?**: Sumar, restar, y calcular el promedio, la mediana y la moda.

4. **Escala de Ratio**:

   - **¿Qué es?**: La escala más completa. Tiene orden, distancia uniforme y un **cero absoluto y significativo**.

   - **Ejemplos**: Altura, peso, edad, ingresos. Un peso de 0 kg significa "ausencia

de peso". Aquí sí puedes decir que alguien que gana $100,000 gana "el doble" que alguien que gana $50,000.

- **¿Qué puedes hacer?**: Todas las operaciones matemáticas.

Provides:	Nominal	Ordinal	Interval	Ratio
The "order"of values is known		✓	✓	✓
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓

Tabla: Resumen de los tipos de datos y las operaciones permitidas para cada escala de medición (Nominal, Ordinal, Intervalo, Ratio).

---

5. Comparando el EDA con Otros Enfoques de Análisis

El autor compara el EDA con los enfoques clásico y bayesiano para resaltar su característica única.

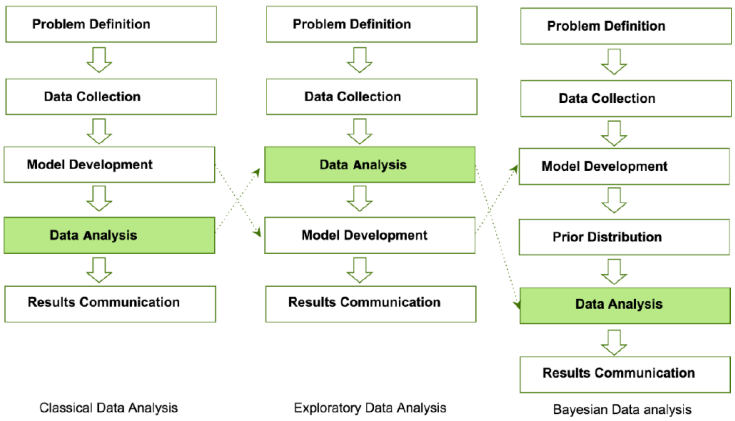

Figura: Diagramas de flujo que comparan los pasos de los enfoques de análisis de datos Clásico, Exploratorio y Bayesiano.

- **Análisis Clásico**: `Definición del Problema -> Recolección de Datos -> Desarrollo del Modelo -> Análisis de Datos`. Aquí, el modelo se impone *antes* de analizar los datos en profundidad.

- **Análisis Exploratorio (EDA)**: `Definición del Problema -> Recolección de Datos -> Análisis de Datos -> Desarrollo del Modelo`. **¡El cambio clave!** El análisis de los datos (la exploración) viene *antes* de decidir qué modelo usar. El foco está en los datos mismos, no en un modelo predeterminado.

- **Análisis Bayesiano**: Es similar al clásico, pero incorpora un paso de "Distribución a Priori", que es el conocimiento o creencia previa que se tiene sobre el problema.

---

6. Herramientas de Software para EDA

Finalmente, el capítulo menciona algunas herramientas de código abierto populares para realizar EDA, destacando las que se usarán en el libro:

- **Python**: El lenguaje de programación más popular para ciencia de datos, con librerías como Pandas, NumPy, Matplotlib y Seaborn.

- **R**: Un lenguaje diseñado por y para estadísticos, muy potente para el análisis y la visualización.

- **Weka y KNIME**: Herramientas con interfaces gráficas que permiten hacer análisis de datos sin necesidad de escribir código.