Introducción al Análisis Exploratorio (EDA) - Día 1

Change Status

in-progress · 40 min

Learning Objectives

1. Comprender la filosofía fundamental del análisis exploratorio de datos como aproximación científica
2. Distinguir entre EDA y análisis confirmatorio en la práctica profesional
3. Establecer una metodología estructurada para abordar datasets nuevos de manera sistemática

Actividades y Aprendizajes

Aprende todo sobre funciones y módulos en Python con ejemplos prácticos.

Task 1: La Filosofía del Análisis Exploratorio (15 minutos)

El Análisis Exploratorio de Datos (EDA) representa una **revolución metodológica** en cómo los científicos de datos abordan problemas complejos. Más que una colección de técnicas estadísticas, EDA es una **filosofía de investigación** que prioriza la curiosidad, la flexibilidad y el aprendizaje iterativo sobre las hipótesis rígidas.

De la Hipótesis Deductiva a la Exploración Inductiva

Tradicionalmente, la investigación científica seguía un **modelo hipotético-deductivo**: se formulaba una hipótesis específica, se diseñaba un experimento para probarla, y se aceptaba o rechazaba la hipótesis. Este enfoque, aunque riguroso, tenía limitaciones fundamentales cuando aplicado a datos complejos del mundo real.

EDA introduce un **paradigma complementario**: la **exploración inductiva**. En lugar de empezar con hipótesis rígidas, EDA comienza con **preguntas abiertas**: "¿Qué patrones existen en estos datos?", "¿Qué anomalías puedo identificar?", "¿Qué relaciones inesperadas aparecen?".

Ventajas del enfoque exploratorio:

Descubrimientos inesperados: Patrones que nunca hubiéramos hipotetizado
Preguntas mejores: Los datos sugieren qué preguntas son realmente relevantes
Robustez: Múltiples perspectivas reducen sesgos de confirmación
Adaptabilidad: Se ajusta a la complejidad real de los datos empresariales
EDA vs Análisis Confirmatorio: Dos Filosofías Complementarias

Análisis Confirmatorio:

Propósito: Probar hipótesis específicas con rigor estadístico
Metodología: Experimentos controlados, pruebas de significancia
Herramientas: Pruebas t, ANOVA, regresión, p-values
Contexto: Investigación académica, validación de teorías

Análisis Exploratorio:

Propósito: Generar hipótesis e insights desde los datos
Metodología: Visualización, estadística descriptiva, patrones emergentes
Herramientas: Gráficos, correlaciones, clustering, distribuciones
Contexto: Negocio, productos, optimización operacional

La relación simbiótica: EDA genera hipótesis que el análisis confirmatorio valida. Uno sin el otro es incompleto.

Metodología Estructurada para EDA

John Tukey, padre del EDA moderno, propuso un **marco sistemático** que transforma la exploración caótica en un proceso disciplinados.

Fase 1: Preparación y Contexto

¿Cuál es la fuente de los datos? ¿Son confiables?
¿Qué preguntas de negocio motivan este análisis?
¿Qué variables están disponibles? ¿Qué tipos de datos?
¿Existen restricciones temporales, geográficas o demográficas?

Fase 2: Inspección Inicial

Dimensiones del dataset (filas, columnas)
Tipos de datos y valores faltantes
Estadísticos básicos por variable
Identificación de datos problemáticos

Fase 3: Análisis Univariado

Distribución de cada variable individual
Identificación de outliers y anomalías
Transformaciones necesarias (log, normalización)

Fase 4: Análisis Bivariado/Multivariado

Relaciones entre variables importantes
Patrones de correlación y dependencia
Segmentación natural de los datos

Fase 5: Síntesis e Insights

¿Qué patrones emergen consistentemente?
¿Qué anomalías requieren explicación?
¿Qué hipótesis nuevas se generan?
¿Qué acciones concretas sugieren los datos?

Task 2: Primeros Pasos en EDA con Pandas (10 minutos)

La implementación práctica de EDA requiere **herramientas eficientes** que permitan iteración rápida y visualización inmediata. Pandas, con su integración natural con Jupyter notebooks, se ha convertido en la herramienta estándar para EDA interactivo.

Configuración del Entorno de Análisis

```
# Importaciones estándar para EDA
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Configuración de visualización
plt.style.use('default')  # Estilo de gráficos
pd.set_option('display.max_columns', None)  # Mostrar todas las columnas
pd.set_option('display.width', None)  # Ancho completo
```

Inspección Inicial del Dataset

```
# Carga de datos
df = pd.read_csv('datos_ventas.csv')

# Dimensiones básicas
print(f"Filas: {df.shape[0]}, Columnas: {df.shape[1]}")
```

```
print(f"\nColumnas: {list(df.columns)}")

# Primeras filas para entender estructura
df.head()
```

Análisis de Tipos y Valores Faltantes

```
# Información general
df.info()

# Porcentaje de valores faltantes por columna
missing_percent = (df.isnull().sum() / len(df)) * 100
print("Valores faltantes (%):")
print(missing_percent[missing_percent > 0])
```

Estadísticos Descriptivos Iniciales

```
# Estadísticos básicos para variables numéricas
df.describe()

# Para variables categóricas
df.describe(include=['object'])

# Conteo de valores únicos
df.nunique()
```

Validación de Calidad de Datos

```
# Verificar rangos lógicos
print("Edades fuera de rango:", ((df['edad'] < 0) | (df['edad'] > 120)).sum())
print("Precios negativos:", (df['precio'] < 0).sum())

# Valores duplicados
print("Filas duplicadas:", df.duplicated().sum())
```

Task 3: Preguntas Orientadoras para EDA (5 minutos)

El EDA efectivo requiere **preguntas orientadoras** que guíen la exploración hacia insights relevantes. Estas preguntas evolucionan a medida que se descubre más sobre los datos.

Preguntas Demográficas
¿Cuál es la distribución de edad de nuestros clientes?

¿Cómo se distribuyen geográficamente?
¿Existen segmentos demográficos naturales?
Preguntas de Comportamiento
¿Cuáles son los patrones de compra típicos?
¿Qué productos se compran juntos?
¿Cómo varía el comportamiento por segmentos?
Preguntas de Calidad
¿Qué porcentaje de datos están completos?
¿Existen valores extremos que distorsionen análisis?
¿Los datos son consistentes temporalmente?
Preguntas de Oportunidad
¿Qué patrones no esperados aparecen?
¿Qué segmentos están subexplotados?
¿Qué productos tienen potencial no reconocido?