

SESIÓN ANÁLISIS EXPLORATORIO DE DATOS

CONTENIDOS:

- ¿Qué es el análisis exploratorio de datos (eda)?
- ¿Qué es el análisis inicial de datos (ida)?
- Contexto en el cual se utiliza el análisis exploratorio de datos.
- Técnicas y herramientas para el análisis exploratorio de datos.
- Análisis univariado y sus objetivos.
- Análisis multivariado y sus objetivos.

¿QUÉ ES EL ANÁLISIS EXPLORATORIO DE DATOS (EDA)?

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés: *Exploratory Data Analysis*) es una etapa fundamental en cualquier proyecto de ciencia de datos. Consiste en la exploración inicial de los datos para comprender su estructura, identificar patrones, detectar anomalías y formular hipótesis. Este proceso no solo ayuda a familiarizarse con los datos, sino que también guía las decisiones sobre qué técnicas de análisis o modelado aplicar.

El Análisis Exploratorio de Datos (EDA) es un enfoque para analizar conjuntos de datos con el fin de resumir sus características principales, a menudo utilizando métodos visuales. Su objetivo es descubrir patrones, relaciones, tendencias y anomalías en los datos antes de aplicar modelos predictivos o análisis más avanzados.

Objetivos del EDA

- **Comprender la estructura y distribución de los datos:** Permite visualizar cómo están organizados los datos y su comportamiento.
- **Identificar valores atípicos (outliers) y datos faltantes:** Detectar valores inusuales o ausencias que puedan afectar el análisis.
- **Detectar relaciones entre variables:** Identificar correlaciones y dependencias entre variables que puedan ser útiles para modelado predictivo.

- **Formular hipótesis para análisis posteriores:** Ayuda a definir preguntas clave para la investigación y pruebas estadísticas.

Diferencia entre EDA y Análisis Confirmatorio

- **EDA:** Enfoque exploratorio sin hipótesis previas, basado en la observación y el descubrimiento de patrones.
- **Análisis Confirmatorio:** Aplicación de métodos estadísticos para probar hipótesis predefinidas con rigor científico.

El EDA es un paso previo al análisis confirmatorio y sirve para generar ideas y detectar problemas antes de aplicar modelos avanzados de análisis de datos.

¿QUÉ ES EL ANÁLISIS INICIAL DE DATOS (IDA)?

El Análisis Inicial de Datos (IDA, *Initial Data Analysis*) es una fase preliminar del EDA que se enfoca en la inspección básica de los datos. Su objetivo es realizar una primera evaluación del conjunto de datos para detectar problemas y asegurar que esté listo para el análisis posterior.

Pasos del IDA

1. **Carga de datos:** Importar el conjunto de datos desde archivos CSV, Excel, SQL, entre otros.
2. **Inspección inicial:** Revisar las primeras filas, tipos de datos y dimensiones del conjunto de datos.
3. **Identificación de problemas:** Detectar valores faltantes, duplicados o inconsistencias en los datos.

Ejemplo de IDA utilizando Pandas:

```
import pandas as pd

# 1 Cargar el conjunto de datos desde un archivo CSV
df = pd.read_csv('datos_ejemplo.csv') # Reemplaza con la ruta de tu archivo

# 2 Inspección inicial: Ver las primeras filas, tipos de datos y dimensiones
print("Primeras filas del dataset:")
print(df.head(), "\n") # Muestra las primeras 5 filas

print("Información del dataset:")
print(df.info(), "\n") # Muestra los tipos de datos y las dimensiones del dataset

# 3 Identificación de problemas:
# a) Valores faltantes
print("Valores faltantes por columna:")
print(df.isnull().sum(), "\n") # Muestra el número de valores faltantes por columna

# b) Duplicados
print("Duplicados en el dataset:")
print(df.duplicated().sum(), "\n") # Muestra la cantidad de filas duplicadas

# c) Estadísticas descriptivas para detectar valores atípicos
print("Estadísticas descriptivas:")
print(df.describe(), "\n") # Muestra estadísticas como media, desviación estándar, etc.

# d) Tipos de datos para revisar posibles conversiones
print("Tipos de datos:")
print(df.dtypes, "\n") # Muestra los tipos de datos de cada columna
```

Ilustración 1 Ejemplo IDA

Explicación del ejemplo:

1. **Carga de datos:** Usamos `pd.read_csv()` para cargar los datos desde un archivo CSV. Asegúrate de que el archivo esté en la misma carpeta que el script o ajusta la ruta del archivo.
2. **Inspección inicial:**
 - `df.head()` nos muestra las primeras 5 filas para entender la estructura del conjunto de datos.

- `df.info()` nos da información sobre el número de filas, columnas y el tipo de datos de cada columna.

3. Identificación de problemas:

- Valores faltantes: `df.isnull().sum()` nos ayuda a ver si alguna columna tiene datos faltantes.
- Duplicados: `df.duplicated().sum()` nos dice si hay filas duplicadas que deben ser eliminadas.
- Estadísticas descriptivas: `df.describe()` nos proporciona una vista general de las estadísticas (promedio, mínimo, máximo, etc.) que nos puede ayudar a detectar valores atípicos.
- Tipos de datos: `df.dtypes` nos muestra los tipos de datos de cada columna, para asegurarnos de que son los adecuados para el análisis posterior.

Datos de archivo `datos_ejemplo.csv`

```
ID,Nombre,Edad,Departamento,Salario
1,Ana,29,Marketing,3000
2,Juan,35,TI,4000
3,María,40,RRHH,3500
4,Pedro,28,Marketing,3200
5,Lucía,25,TI,3800
6,Antonio,38,RRHH,3300
7,Carlos,30,Marketing,3100
8,Sofía,32,TI,3900
9,David,41,RRHH,3400
10,Laura,26,Marketing,3300
```

Técnicas clave dentro del EDA

- **Análisis Univariado:** Se estudia una sola variable a la vez para entender su distribución y comportamiento.
Ejemplo: Histogramas, gráficos de caja (boxplots), medidas estadísticas como media, mediana y moda.
- **Análisis Bivariado:** Se analizan dos variables juntas para detectar relaciones o patrones.
Ejemplo: Diagramas de dispersión (scatter plots), correlación de Pearson, análisis de tablas cruzadas (crosstabs).
- **Análisis Multivariado:** Se estudia la relación entre múltiples variables simultáneamente.
Ejemplo: Matrices de correlación, gráficos de pares (pair plots), regresión múltiple.
- **Visualización de Datos:** Las herramientas de visualización ayudan a interpretar los datos de manera más efectiva.
Ejemplos: Histogramas, diagramas de dispersión, gráficos de barras y mapas de calor (heatmaps).

El Análisis Exploratorio de Datos (EDA) es una fase clave en cualquier proyecto de ciencia de datos. Permite comprender mejor los datos antes de tomar decisiones analíticas y aplicar modelos predictivos.

El Análisis Inicial de Datos (IDA) es el primer paso dentro del EDA y se enfoca en cargar, inspeccionar y detectar problemas en los datos. Con herramientas como Pandas y visualización de datos, podemos hacer un análisis más profundo y garantizar que los datos sean confiables antes de su uso en modelos estadísticos o de aprendizaje automático.



CONTEXTO EN EL CUAL SE UTILIZA EL ANÁLISIS EXPLORATORIO DE DATOS

Aplicaciones del EDA

Ciencia de Datos

- Preparación de datos antes de entrenar modelos de machine learning.
- Detección de valores atípicos (outliers) y datos faltantes.
- Selección de características relevantes para modelado predictivo.

Negocios y Marketing

- Identificación de tendencias en ventas y comportamiento del cliente.
- Segmentación de clientes basada en análisis de patrones.
- Detección de oportunidades de mercado a partir de datos históricos.

Investigación Científica

- Exploración de patrones en datos experimentales.
- Comparación de distribuciones de datos en estudios clínicos o sociales.
- Validación de hipótesis antes de aplicar modelos estadísticos.

Finanzas y Contabilidad

- Detección de fraudes y anomalías en transacciones bancarias.
- Análisis de riesgos financieros y evaluación de portafolios de inversión.
- Seguimiento de tendencias económicas y fluctuaciones de mercado.

Importancia del EDA

- Entender los datos antes de aplicar técnicas avanzadas: Permite descubrir patrones y relaciones clave en los datos.
- Reducir el riesgo de decisiones erróneas: Ayuda a evitar sesgos y suposiciones incorrectas en el análisis.

- Optimizar el tiempo en el modelado predictivo: Una buena exploración reduce la necesidad de ajustes posteriores.
- Facilitar la comunicación de hallazgos: Permite a los analistas explicar los datos de manera clara a tomadores de decisiones y stakeholders.

TÉCNICAS Y HERRAMIENTAS PARA EL ANÁLISIS EXPLORATORIO DE DATOS

Técnicas Comunes

- **Estadísticas Descriptivas**
 - Medidas de tendencia central: Media, mediana, moda.
 - Medidas de dispersión: Rango, varianza, desviación estándar.
 - Resúmenes numéricos para evaluar la distribución de los datos.
- **Visualización de Datos**
 - Histogramas para ver la distribución de los datos.
 - Diagramas de caja (boxplots) para detectar valores atípicos.
 - Diagramas de dispersión para analizar relaciones entre variables.
- **Análisis de Correlación**
 - Medición de la relación entre variables con coeficientes de correlación.
 - Mapas de calor (heatmaps) para visualizar correlaciones en grandes volúmenes de datos.

Herramientas Populares

- **Pandas:** Librería de Python para manipulación de datos tabulares.
- **NumPy:** Librería para cálculos numéricos y manejo eficiente de arreglos.
- **Matplotlib y Seaborn:** Librerías de visualización para gráficos estadísticos.
- **Scipy y Statsmodels:** Herramientas para análisis estadístico y pruebas de hipótesis.

Ejemplo de Uso de Herramientas

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Cargar un dataset de ejemplo
df = pd.read_csv("ventas.csv")

# Estadísticas básicas
print(df.describe())

# Visualización de la distribución de precios
sns.histplot(df["Precio"], bins=10, kde=True)
plt.title("Distribución de Precios")
plt.show()
```

Ilustración 2 Ejemplo de Uso de Herramientas

ANÁLISIS UNIVARIADO Y SUS OBJETIVOS

El Análisis Univariado examina una sola variable a la vez para comprender su comportamiento, distribución y características principales. Este análisis es fundamental en el Análisis Exploratorio de Datos (EDA), ya que nos permite tener una visión clara de los datos antes de proceder a análisis más complejos que involucren múltiples variables.

Objetivos del Análisis Univariado:

1. Comprender la distribución de la variable: Analizar cómo se distribuyen los datos, identificar patrones y verificar si siguen distribuciones conocidas (como la normal).
2. Detectar valores atípicos (outliers): Identificar valores que se desvían significativamente de la tendencia general de los datos y pueden afectar el análisis posterior.

3. Evaluar la centralidad de los datos: Medir la tendencia central de la variable a través de medidas como la media, mediana y moda.
4. Analizar la dispersión de los datos: Estudiar el grado de variabilidad de los datos mediante estadísticas como la desviación estándar, rango o cuartiles.
5. Identificar sesgos en los datos: Verificar si los datos tienen una distribución sesgada, lo que podría influir en el análisis y la interpretación.

Técnicas de Análisis Univariado

- Distribución de Frecuencias
 - Conteo de valores únicos en una variable categórica.
- Medidas de Tendencia Central
 - Media (mean): Promedio de los valores.
 - Mediana (median): Valor central cuando los datos están ordenados.
 - Moda (mode): Valor que más se repite.
- Medidas de Dispersión
 - Rango: Diferencia entre el valor máximo y el mínimo.
 - Varianza y Desviación Estándar: Miden la dispersión de los valores con respecto a la media.

Ejemplo de Análisis Univariado con Pandas

```
# Calcular medidas estadísticas
print("Media de Precios:", df["Precio"].mean())
print("Mediana de Precios:", df["Precio"].median())
print("Desviación Estándar de Precios:", df["Precio"].std())

# Visualización con boxplot
sns.boxplot(x=df["Precio"])
plt.title("Boxplot de Precios")
plt.show()
```

Ilustración 3 Ejemplo de Análisis Univariado con Pandas

ANÁLISIS MULTIVARIADO Y SUS OBJETIVOS

El Análisis Multivariado examina múltiples variables simultáneamente para identificar patrones, correlaciones y relaciones entre ellas. Es útil para evaluar cómo una variable depende de otra y detectar tendencias en conjuntos de datos complejos.

Técnicas de Análisis Multivariado

- Correlación
 - Evalúa la relación entre dos variables numéricas.
 - Se usa el coeficiente de correlación de Pearson.
- Análisis de Componentes Principales (PCA)
 - Técnica de reducción de dimensionalidad para datos con muchas variables.
- Gráficos de Dispersión Matricial (Pair Plots)
 - Visualización de relaciones entre pares de variables en una sola figura.


Ejemplo de Análisis Multivariado

```
# Matriz de correlación
print(df.corr())

# Visualización de un mapa de calor para correlaciones
plt.figure(figsize=(8,6))
sns.heatmap(df.corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Matriz de Correlación")
plt.show()

# Pair plot para analizar relaciones entre variables
sns.pairplot(df, hue="Categoría")
plt.show()
```

Ilustración 4 Ejemplo de Análisis Multivariado



El Análisis Exploratorio de Datos (EDA) es una fase clave en ciencia de datos, permitiendo descubrir patrones ocultos, detectar anomalías y entender la estructura de los datos antes de aplicar modelos más complejos.

El Análisis Univariado se enfoca en estudiar variables de manera individual, mientras que el Análisis Multivariado analiza múltiples variables simultáneamente para entender sus relaciones.

Gracias a herramientas como Pandas, Seaborn y Matplotlib, podemos realizar análisis detallados de datos en Python, optimizando la exploración y mejorando la calidad del análisis.

Dominar el EDA es esencial para cualquier analista de datos o científico de datos, ya que proporciona información clave para la toma de decisiones basada en datos.