¿Qué es un EDA?

Introducción

El **Análisis Exploratorio de Datos** (EDA, por sus siglas en inglés: **Exploratory Data Analysis**) es un enfoque para analizar conjuntos de datos con el fin de resumir sus características principales, a menudo con la ayuda de representaciones visuales. Es una etapa fundamental en el proceso de análisis de datos, ya que permite conocer la estructura, las tendencias y los patrones en los datos antes de realizar análisis más profundos o construir modelos predictivos.

El objetivo del EDA no es solo encontrar patrones, sino también identificar cualquier error, inconsistencia o anomalía en los datos. Esta etapa permite a los científicos de datos, analistas y otras personas involucradas en el proceso de análisis comprender mejor los datos, lo que les ayuda a tomar decisiones informadas sobre qué técnicas y métodos se deben utilizar para su análisis posterior.

1. Objetivos del EDA

El Análisis Exploratorio de Datos tiene varios objetivos clave:

- **Observar la distribución de los datos**: Analizar cómo se distribuyen las variables dentro del dataset.
- **Detectar patrones y relaciones**: Buscar correlaciones, tendencias y agrupamientos entre las variables.
- **Identificar valores atípicos (outliers)**: Encontrar datos anómalos que puedan distorsionar los resultados del análisis o indicar problemas en la recolección de datos.
- **Evaluar la calidad de los datos**: Detectar valores faltantes, errores tipográficos y otros problemas que pueden necesitar corrección antes de realizar un análisis más profundo.
- **Formular hipótesis**: A través de los resultados del EDA, se pueden generar preguntas adicionales o hipótesis que se pueden probar con técnicas más avanzadas.

2. Etapas del EDA

El proceso de EDA se puede dividir en varias etapas, cada una enfocada en obtener una comprensión más profunda de los datos:

a) Revisión de la estructura del dataset

- Examinar el conjunto de datos para comprender qué variables contiene, cuántas filas tiene y qué tipo de datos (números, texto, fechas, etc.) presenta cada columna.

- Determinar si las variables están completas o si tienen valores faltantes.
- Analizar si las variables están en el formato adecuado y si son consistentes.

b) Análisis univariado

- **Distribución de variables numéricas**: Se analizan de forma individual, utilizando estadísticas descriptivas como la media, la mediana, la desviación estándar, los cuartiles, y los histogramas. Los histogramas permiten visualizar la forma de la distribución y detectar posibles sesgos o distribuciones anómalas.
- **Análisis de variables categóricas**: Se utilizan tablas de frecuencia o gráficos de barras para mostrar cómo se distribuyen los valores de las variables categóricas y qué clases son más comunes.

c) Análisis bivariado

- Se exploran las relaciones entre dos variables. Por ejemplo, un gráfico de dispersión (scatter plot) puede ser útil para observar la relación entre dos variables numéricas, mientras que una tabla de contingencia o un gráfico de barras agrupadas puede mostrar la relación entre una variable categórica y una numérica.
- **Correlación**: Para las variables numéricas, se puede calcular el coeficiente de correlación de Pearson para observar la relación lineal entre ellas.

d) Detección de valores atípicos (outliers)

- Los valores atípicos son datos que se desvían significativamente del resto del conjunto de datos. El EDA puede identificar estos puntos usando gráficos como diagramas de caja (box plots) o mediante el análisis de valores estadísticos como el rango intercuartil (IQR).
- Es importante identificar estos valores, ya que pueden afectar la precisión de los análisis o representar errores en la recolección de datos.

e) Tratamiento de datos faltantes

- Durante el EDA, se analizan los valores faltantes en el dataset. Se pueden usar diversos enfoques para manejar los datos faltantes, como eliminar las filas o columnas con datos faltantes, o imputar valores utilizando la media, la mediana o algoritmos más avanzados.

3. Herramientas para realizar un EDA

En la actualidad, existen muchas herramientas y bibliotecas que facilitan el proceso de EDA, especialmente en lenguajes como Python y R. Algunas de las herramientas más populares incluyen:

- **Python**:
  - **Pandas**: Para manipulación de datos y análisis estadísticos.
  - **Matplotlib y Seaborn**: Para la creación de gráficos y visualización de datos.
  - **Plotly**: Para crear visualizaciones interactivas.
  - **NumPy**: Para operaciones matemáticas y estadísticas.
- **R**:
  - **ggplot2**: Para la creación de gráficos avanzados.
  - **dplyr**: Para manipulación de datos.
  - **summarytools**: Para obtener estadísticas descriptivas rápidas.
- **Herramientas visuales**:
  - **Tableau**: Para crear visualizaciones interactivas y de fácil comprensión.
  - **Power BI**: Herramienta de análisis empresarial con fuertes capacidades de visualización.

4. Técnicas comunes usadas en un EDA

Algunas de las técnicas más comunes que se utilizan en el Análisis Exploratorio de Datos incluyen:

- **Histogramas**: Para ver la distribución de una variable numérica.
- **Diagramas de caja (Box plots)**: Para visualizar la dispersión y detectar valores atípicos.
- **Gráficos de dispersión (Scatter plots)**: Para estudiar las relaciones entre dos variables numéricas.
- **Matriz de correlación**: Para observar las relaciones entre varias variables numéricas y encontrar correlaciones significativas.
- **Mapas de calor (Heatmaps)**: Para representar visualmente matrices de correlación o cualquier otra matriz de datos.
- **Diagramas de barras**: Para analizar la frecuencia de variables categóricas.

5. Importancia del EDA en los negocios

El EDA no solo es crucial en el ámbito académico o de la investigación, sino que también tiene un impacto significativo en el mundo de los negocios. A través del EDA, las empresas pueden:

- **Identificar tendencias**: El análisis exploratorio puede revelar patrones en los datos que pueden ser útiles para predecir comportamientos futuros, como la demanda de productos o los patrones de compra.

- **Tomar decisiones informadas**: Los resultados del EDA permiten que los analistas y directivos tomen decisiones basadas en datos, minimizando los riesgos y mejorando la efectividad de las estrategias empresariales.
- **Optimizar recursos**: Detectar eficiencias en los procesos y mejorar áreas con menor rendimiento es más fácil gracias al EDA.
- **Mejorar la experiencia del cliente**: Al comprender las preferencias y el comportamiento de los clientes a través de los datos, las empresas pueden mejorar la personalización de sus productos y servicios.

Conclusión

El Análisis Exploratorio de Datos (EDA) es un paso fundamental en cualquier proyecto de análisis de datos. Nos ayuda a entender los datos, identificar patrones, detectar errores, y a preparar el camino para el análisis más profundo o la construcción de modelos predictivos. Es una técnica clave para científicos de datos y analistas, ya que proporciona una base sólida sobre la cual construir conclusiones y tomar decisiones estratégicas.