Unidad: Análisis de Datos

Exploratory Data Analysis (EDA): Enfoque Conceptual

Nicolás Sidicaro
Abril 2025

¿Qué es el EDA?

- **Exploratory Data Analysis (EDA)** es un enfoque para analizar conjuntos de datos con el fin de:
  - Descubrir patrones subyacentes
  - Identificar anomalías
  - Probar hipótesis
  - Verificar supuestos

Importancia del EDA

El EDA es **fundamental** porque:

- Es el **primer paso** crucial antes de cualquier análisis formal
- Permite **entender realmente los datos** antes de aplicar modelos complejos
- Puede **revelar problemas** que afectarían los resultados del análisis
- **Guía la selección** de técnicas apropiadas para el análisis posterior
- Proporciona **contexto valioso** para interpretar resultados

El EDA en el Ciclo de Ciencia de Datos

El EDA es una fase **iterativa** que:

1. Sigue a la **importación y limpieza** inicial de datos
2. **Precede y guía** la modelización formal
3. A menudo lleva a **revisitar** las fases de limpieza de datos
4. **Informa** sobre qué variables considerar en el modelado
5. Ayuda a **interpretar** los resultados de los modelos

Filosofía del EDA

El enfoque "filosófico" del EDA:

- **Escéptico**: No aceptar los datos en su valor nominal
- **Exploratorio**: Buscar sin ideas preconcebidas
- **Iterativo**: Las preguntas llevan a más preguntas
- **Visual**: "Ver para creer" - la visualización es clave
- **Contextual**: Utilizar el conocimiento del dominio

Componentes Principales del EDA

El EDA tiene cuatro componentes principales:

1. **Comprensión del contexto y objetivos**
    - ¿Cuál es el problema de negocio?
    - ¿Qué preguntas estamos tratando de responder?
2. **Análisis univariante**
    - Examinar cada variable individualmente
    - Distribuciones, estadísticas descriptivas, etc.
3. **Análisis bivariante**
    - Relaciones entre pares de variables
    - Correlaciones, tendencias, agrupaciones
4. **Análisis multivariante**
    - Interacciones complejas entre múltiples variables
    - Patrones y estructuras en dimensiones superiores

Etapas del EDA

Un enfoque estructurado para el EDA incluye:

1. **Formular preguntas** sobre los datos
2. **Buscar respuestas** mediante análisis y visualización
3. **Refinar preguntas** basándose en lo descubierto
4. **Generar nuevas preguntas** para profundizar
5. **Comunicar hallazgos** para informar decisiones

1. Comprensión Inicial de los Datos

Antes de cualquier análisis detallado:

- **Explorar la estructura básica**:
  - Número de observaciones y variables
  - Tipos de datos (numérico, categórico, fechas, etc.)
  - Identificadores únicos y claves
- **Examinar las primeras filas**:
  - ¿Los datos tienen sentido a primera vista?
  - ¿Hay problemas evidentes?
- **Comprender el contexto de negocio**:
  - ¿Qué representa cada variable?
  - ¿Cómo se recolectaron los datos?
  - ¿Qué limitaciones podrían tener?

2. Calidad de los Datos

```r
# Código para verificar valores faltantes
colSums(is.na(datos))

# Código para verificar duplicados
sum(duplicated(datos))
```

Aspectos clave a verificar:

- **Valores faltantes**:
  - ¿Cuántos hay y dónde están?
  - ¿Hay patrones en su distribución?
  - ¿Cómo afectan a nuestro análisis?
- **Valores atípicos y extremos**:
  - ¿Hay valores que parecen erróneos?
  - ¿Los outliers son datos reales o errores?

2. Calidad de los Datos

- Inconsistencias:
  - ¿Hay valores contradictorios?
  - ¿Existen registros duplicados?

3. Análisis Univariado: Numéricas

Para cada variable numérica, examinar:

- Estadísticas descriptivas:
  - Centro (media, mediana, moda)
  - Dispersión (rango, varianza, desviación estándar)
  - Forma (asimetría, curtosis)
  - Rango (mínimo, máximo, percentiles)
- Visualizaciones:
  - Histogramas
  - Gráficos de densidad
  - Boxplots
  - QQ plots (para normalidad)

Ejemplo: Análisis Univariado (Num)

```r
# Estadísticas descriptivas
summary(datos$variable_numerica)

# Visualización
par(mfrow=c(2,2))
hist(datos$variable_numerica, main="Histograma")
plot(density(datos$variable_numerica), main="Densidad")
boxplot(datos$variable_numerica, main="Boxplot")
qqnorm(datos$variable_numerica); qqline(datos$variable_numerica)
```

Preguntas clave:

- ¿La distribución es normal, sesgada, multimodal?
- ¿Hay valores atípicos que requieren atención?
- ¿Los valores están dentro de rangos razonables?

4. Análisis Univariado: Categóricas

Para cada variable categórica, examinar:

- Frecuencias y proporciones:
  - ¿Cuántos casos hay en cada categoría?
  - ¿Hay un desbalance significativo?
- Cardinalidad:
  - ¿Cuántas categorías únicas hay?
  - ¿Hay categorías con muy pocos casos?
- Visualizaciones:
  - Gráficos de barras
  - Gráficos circulares (para pocas categorías)
  - Gráficos de Pareto

Ejemplo: Análisis Univariante

```r
# Tabla de frecuencias
table(datos$variable_categorica)
prop.table(table(datos$variable_categorica))

# Visualización
barplot(table(datos$variable_categorica), main="Frecuencias")
pie(table(datos$variable_categorica), main="Proporciones")
```

Preguntas clave:

- ¿Hay categorías dominantes?
- ¿Se necesita agrupar categorías poco frecuentes?
- ¿Las categorías están codificadas de manera consistente?

5. Análisis Bivariado: Relaciones

Examinamos cómo se relacionan pares de variables:

- Numérica vs. Numérica:
  - Correlación (Pearson, Spearman, Kendall)
  - Gráficos de dispersión
  - Heatmaps de correlación
- Categórica vs. Categórica:
  - Tablas de contingencia
  - Pruebas de chi-cuadrado
  - Gráficos de mosaico
- Numérica vs. Categórica:
  - Boxplots agrupados
  - Gráficos de violín
  - ANOVA

Ejemplo: Análisis Bivariado

```
# Correlación numérica
cor(datos$numerica1, datos$numerica2)

# Tabla de contingencia
table(datos$categorica1, datos$categorica2)

# Boxplot agrupado
boxplot(numerica ~ categorica, data = datos)
```

Preguntas clave:

- ¿Qué variables están fuertemente correlacionadas?
- ¿Hay relaciones no lineales que la correlación no captura?
- ¿Las distribuciones varían significativamente entre grupos?

6. Análisis Multivariado: Interacciones

Examinar interacciones entre múltiples variables:

- Técnicas de visualización:
  - Gráficos de pares (pairplots)
  - Gráficos de coordenadas paralelas
  - Heatmaps
  - Gráficos 3D
- Técnicas analíticas:
  - Análisis de componentes principales (PCA)
  - Análisis de clúster
  - Análisis factorial

7. Análisis Temporal

Si los datos tienen componente temporal:

- Tendencias:
  - ¿Hay aumentos o disminuciones sistemáticos a lo largo del tiempo?
- Estacionalidad:
  - ¿Existen patrones que se repiten periódicamente?
- Ciclos:
  - ¿Hay patrones no periódicos más largos?
- Irregularidades:
  - ¿Aparecen eventos inusuales o outliers temporales?
  - ¿Hay cambios estructurales en la serie?

8. Patrones a Buscar

Durante el EDA, debemos estar atentos a:

- **Agrupaciones**: Concentraciones de datos que sugieren segmentos naturales
- **Correlaciones**: Relaciones lineales o no lineales entre variables
- **Tendencias**: Patrones direccionales en los datos
- **Valores atípicos**: Puntos de datos que difieren significativamente del resto
- **Huecos**: Áreas donde faltan datos que podrían ser significativas
- **Distribuciones**: Formas que toman los datos (normal, sesgada, multimodal)

9. Formulación de Hipótesis

El EDA debe generar hipótesis que guíen análisis posteriores:

- Cómo formular hipótesis basadas en datos:
  - Partir de patrones observados
  - Considerar el contexto de negocio
  - Incorporar conocimiento previo del dominio
- Refinamiento iterativo:
  - Comprobar hipótesis iniciales
  - Reformular basándose en la evidencia
  - Generar nuevas hipótesis más específicas
- Documentar hipótesis:
  - Mantener un registro de todas las hipótesis
  - Anotar la evidencia que las apoya o refuta

10. Herramientas para EDA en R

R ofrece numerosas herramientas para realizar EDA:

- **Paquetes básicos**:
  - `base`: Funciones estadísticas fundamentales
  - `stats`: Pruebas estadísticas y modelos
  - `graphics`: Visualizaciones básicas
- **Paquetes del Tidyverse**:
  - `dplyr`: Manipulación de datos
  - `ggplot2`: Visualizaciones avanzadas
  - `tidyr`: Ordenamiento de datos
- **Paquetes especializados**:
  - `DataExplorer`: Automatización de EDA
  - `GGally`: Extensiones de ggplot2 para análisis multivariante
  - `corrplot`: Visualización de matrices de correlación

11. Automatización vs. Exploración

Ventajas y limitaciones de cada enfoque:

- **Automatización**: ✓ Eficiente para conjuntos de datos grandes ✓ Reduce la posibilidad de olvidar verificaciones importantes ✓ Consistente y reproducible ✗ Puede pasar por alto patrones sutiles o particulares ✗ Limitado por lo que está programado para buscar
- **Exploración manual**: ✓ Permite seguir la intuición y el conocimiento del dominio ✓ Facilita descubrir lo inesperado ✓ Adaptable a las peculiaridades de cada conjunto de datos ✗ Más lento y menos sistemático ✗ Susceptible a sesgos del analista

> La combinación de ambos enfoques suele ser lo óptimo

12. Documentación del EDA

La documentación es crucial:

- ¿Por qué documentar?
  - Asegura la reproducibilidad
  - Facilita la comunicación con stakeholders
  - Permite revisar y refinar el análisis
  - Sirve como referencia para futuros proyectos
- ¿Qué documentar?
  - Preguntas iniciales y objetivos
  - Hallazgos principales
  - Decisiones tomadas (y por qué)
  - Visualizaciones clave
  - Hipótesis generadas
  - Limitaciones identificadas

13. Errores Comunes en EDA

Errores a evitar:

- **Saltarse la exploración** e ir directamente al modelado
- **Confiar ciegamente en estadísticas resumidas** sin visualizar
- **No considerar el contexto del dominio**
- **Centrarse solo en tendencias centrales** ignorando la variabilidad
- **No verificar supuestos** de normalidad, independencia, etc.
- **Sobreinterpretar patrones aleatorios**
- **Ignorar valores atípicos** sin investigarlos adecuadamente
- **No documentar el proceso** y los hallazgos

14. Pasos Clave para un EDA Efectivo

1. **Define claramente tus objetivos**
    - ¿Qué preguntas específicas quieres responder?
2. **Comprende la estructura de tus datos**
    - Tipos de variables, dimensiones, granularidad
3. **Evalúa la calidad de los datos**
    - Valores faltantes, outliers, inconsistencias
4. **Explora cada variable individualmente**
    - Distribuciones, estadísticas descriptivas
5. **Analiza relaciones entre variables**
    - Correlaciones, patrones, agrupaciones

14. Pasos Clave para un EDA Efectivo

1. **Genera y refina hipótesis**
    - Documenta tus observaciones y preguntas
2. **Comunica los hallazgos con visualizaciones efectivas**
    - Elije los gráficos apropiados para tus datos
3. **Itera y profundiza** en áreas de interés

15. EDA y Toma de Decisiones

El EDA informa decisiones críticas:

- **Selección de variables** para modelado
  - ¿Qué variables tienen mayor poder predictivo?
  - ¿Cuáles son redundantes?
- **Transformación de datos**
  - ¿Se necesitan normalizar variables?
  - ¿Hay que crear nuevas variables?
- **Gestión de valores atípicos**
  - ¿Deben eliminarse, transformarse o analizarse por separado?
- **Imputación de valores faltantes**
  - ¿Qué método es más apropiado según el patrón observado?
- **Validación cruzada y división de datos**
  - ¿Cómo asegurar que los splits sean representativos?