Análisis de Datasets: Episodes y Shows

- Episodes: Contiene información relacionada con episodios de series de televisión obtenidos de la API de TV Maze.
- Shows: Contiene información relacionada con shows o series completas obtenidas de la misma API.

El análisis se lleva a cabo mediante un perfilado de datos utilizando herramientas como pandas-profiling o ydata-profiling, y se enfoca en identificar problemas de calidad de datos que deben ser tratados antes de realizar un análisis profundo o aplicar modelos predictivos.

Dataset: Episodes

Estadísticas Generales

- Número de variables: 15
- Número de observaciones: 4,996
- Celdas faltantes: 7,875 (10.5%)
- Filas duplicadas: Ninguna (0%)
- Tamaño en memoria: 3.5 MiB

Alertas detectadas

- **Alta Correlación:**
  - episode_id ↔ show_id
  - episode_number ↔ episode_type
  - episode_type ↔ episode_number
  - show_id ↔ episode_id

Imbalance:

- episode_type está altamente desbalanceada (95.5%).

Valores faltantes:

- episode_airtime: 53.0% (2,646 registros).

- episode_runtime: 11.2% (559 registros).
- episode_rating_average: 92.7% (4,632 registros).

Recomendaciones

- Eliminar variables con más del 90% de valores faltantes (episode_rating_average).
- Revisar variables con correlación alta y decidir si mantener o eliminar según su relevancia.
- Considerar eliminar variables que no aportan valor analítico (URLs e IDs únicos).

Dataset: Shows

Estadísticas Generales

- Número de variables: 52
- Número de observaciones: 729
- Celdas faltantes: 18,335 (48.4%)
- Filas duplicadas: Ninguna (0%)
- Tamaño en memoria: 1.9 MiB

Alertas detectadas

Alta correlación:

- Variables con correlación perfecta o alta, especialmente entre identificadores y métricas relacionadas (show_runtime, show_average_runtime).

Valores faltantes:

- Varias columnas tienen más del 90% de celdas faltantes (show_network_name, show_network_country_name, etc.).
- Columnas con 100% de valores faltantes (show_network, show_image, _embedded.show.dvdCountry).

Variables únicas:

- show_id, show_url, show_updated, show_self_href.

Variables no soportadas:

- Al menos 8 columnas con formato no compatible (image, show_genres, etc.).

Recomendaciones:

- Eliminar columnas con más del 90% de celdas faltantes o aquellas que son irrelevantes para el análisis.
- Imputar valores en variables numéricas donde sea posible.
- Normalizar variables numéricas y transformar variables categóricas relevantes.
- Revisar correlaciones y eliminar variables redundantes.

Conclusiones

Ambos datasets presentan problemas significativos de calidad de datos. El dataset de shows tiene un alto porcentaje de celdas faltantes (48.4%) y múltiples variables con más del 90% de nulos. Por otra parte, el dataset de episodios presenta alta correlación en algunas variables y columnas con más del 90% de datos faltantes.