Análisis del Perfil de Datos de los programas de TVMaze en Enero de 2024

Introducción

En este informe, se presenta un análisis de perfil de datos detallado sobre la información obtenida de la API de TVMaze, específicamente del endpoint schedule Web/streaming que proporciona datos de episodios emitidos en la web o por streaming para una fecha en específico, en este caso, se analizarán los datos para enero de 2024. El objetivo principal es identificar las características, problemas y tendencias de estas variables, y posteriormente definir una estrategia de limpieza y transformación dentro del proceso ETL, asegurando que los datos finales sean consistentes y de calidad.

El análisis se basó en la generación de un reporte de profiling (similares a los que se generan con la librería ydata-profiling en Python) y otros métodos de inspección de datos, con el fin de resumir el estado actual del dataset y recomendar pasos concretos de remediación.

Análisis general del dataset

1. **Análisis del DataFrame**

   Número total de columnas: 66

   Cantidad de filas: 4991

   Número total de celdas con valores nulos: 129,775

   Porcentaje de celdas faltantes: 39.4%

   Número de filas duplicadas: 0

   Como se observa, contamos con un volumen importante de datos (4991 filas), pero un porcentaje de celdas faltantes cercano al 40%. Esto indica que casi dos quintas partes de nuestro dataset necesitan atención para evitar sesgos o problemas futuros. Además, el hecho de que no existan filas duplicadas (0) es una buena señal: el dataset, al menos, está libre de duplicidad a nivel de filas completas.

2. **Análisis de Tipos de Variables**

   Número de columnas numéricas: 15

   Número de columnas URL: 13

   Número de columnas de texto: 12

   Número de columnas categóricas: 14

   Número de columnas de tipo datetime (fecha y/u hora): 4

   Número de columnas con un tipo no reconocido: 8

   Esta clasificación pone de manifiesto la diversidad de formatos presentes en el dataset. Las columnas no reconocidas (8 en total) pueden deberse a datos anidados, listas, valores mixtos o incluso datos que incluyen HTML. Estas columnas suelen requerir un tratamiento más exhaustivo durante la fase de limpieza y transformación.

3. **Variables con valores únicos**
   - id
   - url
   - _links.self.href

   Estas variables presentan un valor único por registro. No aportan *per se* información analítica (por ejemplo, para modelar ratings o clasificaciones), pero son indispensables para la identificación unívoca de cada episodio y su posterior enlace con información secundaria.

4. **Variables con problemas**

   Luego de un análisis exhaustivo del perfil de datos (1 al 31 de enero de 2024) provisto por la API de TVMaze, se detectaron los siguientes aspectos:

   1. **Altas correlaciones y redundancias**
      Varias columnas muestran información duplicada o redundante. Por

ejemplo:

- _embedded.show.averageRuntime se correlaciona con otros campos de runtime.
- _embedded.show.network.country.code y al menos otras 6 variables tienen una alta correlación.
- _embedded.show.externals.thetvdb y _embedded.show.externals.tvrage (más 9 variables relacionadas) podrían aportar datos muy similares.

Este tipo de duplicidad implica **multicolinealidad** (altamente correlacionadas entre sí) y puede distorsionar análisis estadísticos si no se maneja adecuadamente.

2. **Alto porcentaje de valores perdidos**

- airtime: 51.4% de valores perdidos.
- runtime: 9.5% de valores perdidos.
- summary: 69.2% de valores perdidos. (Debido a que muchos episodios futuros no tienen sinopsis oficial).
- rating.average: 92.9% de valores perdidos (episodios que aún no han sido calificados).

Hay más variables que superan el 30% de datos faltantes, y esto puede **distorsionar** los resultados de cualquier modelado o análisis posterior.

3. **Desequilibrio en variables categóricas**
    - Variables como type o schedule.time se concentran en una o dos categorías dominantes, lo que puede inducir **sesgos** en modelos de clasificación o segmentación.
4. **Valores en cero y outliers**
    - _embedded.show.weight: 2.8% de los valores son cero.
    - Existen variables como season que llegan a valores no esperados (por ejemplo, “2024” en lugar de un simple 1, 2, 3…).

5. **Variables “no soportadas” o complejas**
    - _embedded.show.genres
    - _embedded.show.schedule.days
    - _embedded.show.network

    Estas columnas contienen estructuras de tipo array, listas o anidaciones que requieren transformación a un formato tabular o JSON manejable (por ejemplo, separar géneros en múltiples filas o columnas).

6. **Temporadas atípicas**
    - Se detectan valores fuera de la lógica esperada (por ejemplo, season = 2024). Esto rompe con el patrón de temporadas (1, 2, 3…).

Conclusión

Tras esta inspección, el conjunto de datos para enero de 2024 presenta inconsistencias que deben **corrigirse** antes de integrarlos en una base de datos estructurada o de usarlos para análisis avanzados. A modo de resumen:

- **Redundancia:** Se deben eliminar o combinar columnas altamente correlacionadas o que dupliquen la información.
- **Valores faltantes:** Es esencial implementar una estrategia robusta de imputación o eliminación, dependiendo de la relevancia de cada columna para el objetivo analítico.
- **Desequilibrio categórico:** Requiere técnicas de reagrupamiento (combinar categorías escasas bajo “other”) o muestreo estratificado en caso de modelado.
- **Transformación de datos no estructurados:** Variables que contienen arrays o listas deben aplanarse o separarse en tablas hijas según convenga al modelo de datos.

- **Outliers y valores atípicos:** Detectar y tratar registros como season = 2024 o _embedded.show.weight = 0 cuando no tengan sentido real.

Acciones realizadas

En base a este reporte de perfil de datos, se han tomado las siguientes medidas:

1. **Eliminación de columnas redundantes o con datos no soportados**
    - Se removieron columnas con más del 85% de datos faltantes para mejorar la calidad del dataset y evitar multicolinealidad.
2. **Tratamiento de valores atípicos en la columna "season"**
    - Se eliminaron filas donde la columna season contenía valores incoherentes como 2024, que no se ajustaban a la secuencia esperada de temporadas (1, 2, 3, ...).
3. **Estandarización de fechas**
    - Se convirtió a formato datetime las columnas premiered, ended, airdate y airstamp para asegurar consistencia en el manejo de fechas. Cualquier valor no convertible a fecha fue ignorado sin interrumpir el proceso.
4. **Limpieza de texto en la columna "summary"**
    - Se utilizó BeautifulSoup para eliminar las etiquetas HTML de la columna summary y así facilitar el análisis de texto.
5. **Transformación de estructuras complejas**
    - Columnas como _embedded.show.genres y _embedded.show.schedule.days, que contenían listas, fueron transformadas en texto separado por comas. Esto permite una mayor facilidad de análisis.
6. **Mapeo de días de la semana**
    - Se definió un diccionario days_mapping para transformar los días de la semana en valores numéricos (ej. Monday = 1, Tuesday = 2, ...), lo que facilita su uso en análisis estadísticos.
7. **Imputación de valores faltantes en variables numéricas**
    - Las columnas numéricas como runtime y _embedded.show.averageruntime fueron imputadas con la mediana de sus valores, asegurando que no hubiera sesgos por valores faltantes.
8. **Imputación de valores faltantes en variables categóricas**

- Las columnas categóricas fueron imputadas con la moda (valor más frecuente). En caso de no existir una moda, se utilizó el valor 'unknown'.

9. **Eliminación de filas con menos del 25% de datos completos**
    - Se eliminaron las filas que tenían menos del 25% de datos completos, garantizando que el dataset mantuviera la calidad.
10. **Eliminación de duplicados a nivel de registro**
    - Se confirmaron que no existían filas duplicadas. En caso de encontrar duplicados, fueron eliminados para evitar sesgos en el análisis.
11. **Ajuste de variables categóricas**
    - Las categorías minoritarias en la columna type fueron reagruparlas bajo la etiqueta Other, mejorando la representatividad de las categorías para análisis y modelos.

Cierre

Con estas acciones correctivas, el dataset derivado se alinea mejor con los objetivos del proyecto y se prepara para posteriores usos, ya sea en un sistema de almacenamiento (por ejemplo, Data Warehouse o Data Lake) o en modelos de predicción y análisis de audiencias.

Este proceso permite garantizar consistencia, calidad y confiabilidad en los datos de TVMaze de enero de 2024.