Análisis del Perfil de Datos

Tras analizar el perfil de datos generado para los datos del 1 al 31 de enero de la API de TVMAZE, se identificaron varios problemas en el conjunto de datos, tales como variables con altas correlaciones, valores perdidos, datos no soportados y desequilibrios en variables categóricas. A continuación, se detallan las conclusiones.

1. Se han detectado múltiples variables con altas correlaciones entre sí, columnas como:

   - _embedded.show.averageRuntime con _embedded.show.network.country.code y otras 6 variables.
   - _embedded.show.externals.thetvdb con _embedded.show.externals.tvrage y otras 9 variables.

Estas columnas pueden duplicar la información y ser redundantes.

2. Varias variables presentan un alto porcentaje de valores perdidos:

- airtime: 51.4% valores perdidos.
- runtime: 9.5% valores perdidos.
- summary: 69.2% valores perdidos.
- rating.average: 92.9% valores perdidos.

Hay más variables con más del 30% de valores perdidos, lo que puede distorsionar los resultados y las conclusiones del análisis.

airtime has 2460 (51.4%) missing values	Missing
runtime has 453 (9.5%) missing values	Missing
summary has 3312 (69.2%) missing values	Missing
rating.average has 4448 (92.9%) missing values	Missing
image.medium has 3561 (74.4%) missing values	Missing
image.original has 3561 (74.4%) missing values	Missing
_embedded.show.language has 330 (6.9%) missing values	Missing
_embedded.show.runtime has 3582 (74.8%) missing values	Missing
_embedded.show.averageRuntime has 309 (6.5%) missing values	Missing
_embedded.show.ended has 3066 (64.0%) missing values	Missing
_embedded.show.officialSite has 473 (9.9%) missing values	Missing
_embedded.show.rating.average has 4048 (84.6%) missing values	Missing
_embedded.show.network has 4787 (100.0%) missing values	Missing
_embedded.show.webChannel.id has 112 (2.3%) missing values	Missing
_embedded.show.webChannel.name has 112 (2.3%) missing values	Missing
_embedded.show.webChannel.country.name has 1602 (33.5%) missing values	Missing

3. Desequilibrio en variables categóricas:

La variable type y schedule.time se encuentra desequilibradas con una categoría predominando sobre las demás. Lo que puede provocar que los modelos se inclinen a la categoría mayoritaria.

type is highly imbalanced (96.3%)	Imbalance
_embedded.show.schedule.time is highly imbalanced (50.6%)	Imbalance

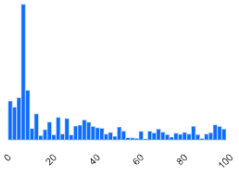
4. Variables con valores en cero como: _embedded.show.weight, donde el 2.8% de los valores son cero

_embedded.show.weight

Real number (ℝ)

High correlation Zeros

Distinct	101	**Minimum**	0
Distinct (%)	2.1%	**Maximum**	100
Missing	0	**Zeros**	134
Missing (%)	0.0%	**Zeros (%)**	2.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	30.856695	**Memory size**	37.5 KiB



More details

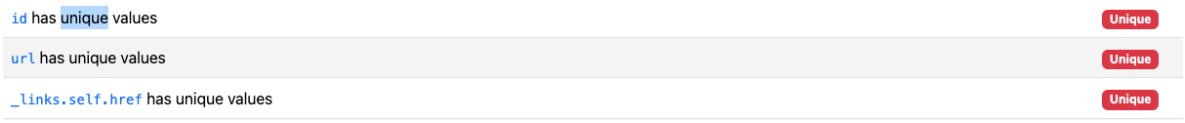
5. Varias variables están marcadas como no soportadas o requieren limpieza. Entre ellas se encuentran las siguientes, las cuales son fundamentales para la información de los shows y episodios:

- _embedded.show.genres
- _embedded.show.schedule.days
- _embedded.show.network

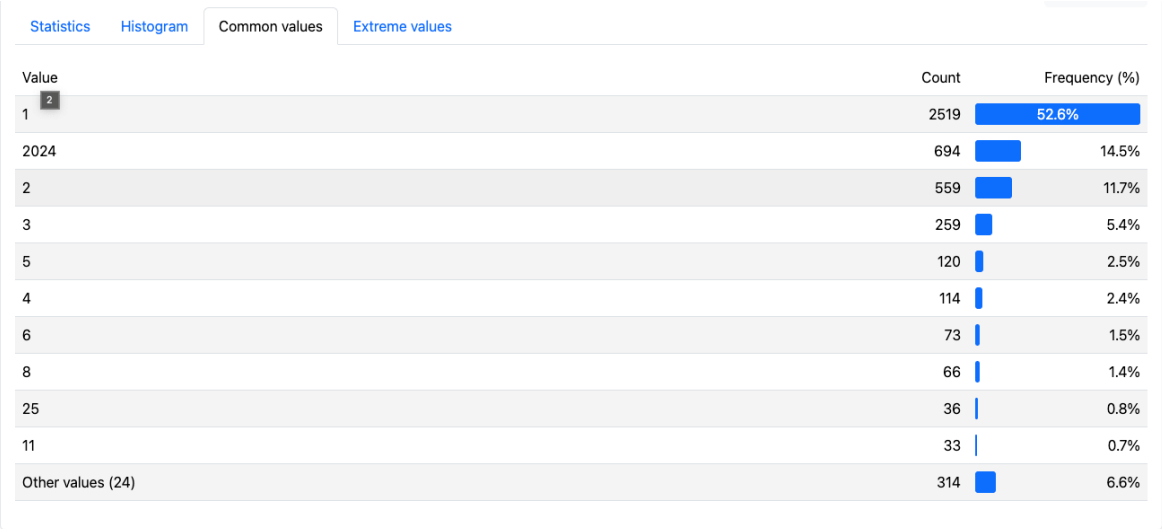
_embedded.show.genres	_embedded.show.schedule.days	_embedded.show.network
[Comedy, Crime]	[Monday, Tuesday, Wednesday, Thursday]	NaN
[Thriller, Mystery]	[Thursday]	NaN
[]	[Thursday]	NaN
[]	[Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday]	NaN
[Drama]	[]	NaN
[Drama]	[]	NaN
[Drama]	[]	NaN
[Drama]	[]	NaN
[Drama]	[]	NaN
[Drama]	[]	NaN

- Como se puede observar las variables: embedded.show.genres y _embedded.show.schedule.days almacenan datos en forma de listas o vectores, por lo que deben ser tranformadas.
- El campo NaN puede significar valores faltantes.

6. id, url, _links.self.href: Tienen valores únicos para cada registro. Siendo útiles para rastrear los episodios y todos los datos. No aportan información analítica, pero son esenciales para realizar las relaciones de los elementos.



7. Columnas como season presentan valores atípicos, se puede observar que los números van del 1 al 25 y una distorsión en 2024.



Conclusión

El conjunto de datos presenta inconsistencias que deben solucionarse antes de incorporarlos en un modelo estructurado de base de datos o utilizarlos en análisis posteriores.

- Redundancia de Datos: Se debe manejar la redundancia de datos para evitar problemas de multicolinealidad. Es recomendable eliminar o combinar variables altamente correlacionadas.

- Valores Faltantes: Es crucial manejar adecuadamente los valores faltantes para mantener la integridad del análisis.

- Desequilibrio en Variables Categóricas: Las variables categóricas desequilibradas requieren ajustes para garantizar su confiabilidad.

- Preprocesamiento de Variables No Estructuradas: Es necesario preprocesar y transformar variables no estructuradas y no soportadas para su uso efectivo. En particular, las variables que contienen arrays, como géneros y días de emisión, deben ser transformadas en un formato adecuado para la carga y el análisis.

Acciones realizadas

1. faltantes o contienen datos no soportados e irrelevantes para el análisis. Se eliminaron estas columnas identificadas para reducir la redundancia y mejorar la calidad del conjunto de datos.

2. Se eliminaron las filas donde el valor de la columna season es 2024, considerado un valor atípico no esperado, ya que las temporadas suelen numerarse de manera secuencial, adicional se puede evidenciar que van del 1 al 25, ese valor es incorrecto.

3. Se seleccionaron las columnas _embedded.show.genres, _embedded.show.schedule.days y _embedded.show.network que contienen arrays o estructuras complejas. Estas columnas se convirtieron en cadenas de texto separadas por comas para facilitar su manipulación y análisis. Aunque los datos de la columna _embedded.show.network son escasos, se decidió incluirlos en el modelo para no perder información valiosa.

4. Se definió un diccionario days_mapping para convertir los nombres de los días de la semana a números, con el fin de estandarizar los datos y facilitar los análisis numéricos y el modelado.

5. Se rellenaron los valores faltantes de columnas numéricas como runtime y _embedded.show.averageRuntime con la mediana de cada columna, manteniendo así la consistencia en la distribución de los datos.

6. Se eliminaron las filas duplicadas en el DataFrame para asegurar que cada registro sea único y evitar sesgos en el análisis.

7. Se ajustaron las variables categóricas, reemplazando las categorías con pocas apariciones por 'Other', con el objetivo de equilibrar la distribución de categorías y mejorar la confiabilidad de los modelos.

8. Se convierten variables categóricas en variables dummy para análisis y modelado.

9. Se eliminaron las filas del conjunto de datos que tienen más del 20% de valores faltantes (NaN), mejorando así la calidad y confiabilidad de los datos restantes para el análisis.