# Análisis del *profiling* realizado a los dos *dataframe* generados

Teniendo en cuenta que se crearon 6 dataframes diferentes, se ha utilizado la librería *pandas_profiling* para generar un perfilamiento de cada uno de dichos dataframes de forma más ágil. La herramienta permite generar un archivo *html* por cada uno de los perfilamientos generando un detalle general en donde se encuentran las estadísticas del dataset como numero de variables, observaciones, datos faltantes, duplicados y los tipos de variables, entre otros.

Hay, dentro de la vista general, un resumen de alertas encontrado por la herramienta, en donde se pueden encontrar datos relevantes, como por ejemplo que variables tienen mayor número de datos faltantes, cuales son valores únicos, cuales tienen altas correlaciones.

Así mismo detalla todas las variables y genera reportes dependiendo del tipo de datos que contengan. Seguido de esto muestra las interacciones que puede haber entre las variables numéricas, así como un mapa de calor de correlaciones entre las mismas. Finalmente se contabilizan en un gráfico los datos faltantes por variables y se imprime una muestra de los datos.

## Dataframe *episodes_df*

A nivel general se encuentra que hay en total 12 variables, las cuales se habían establecido con anterioridad. Hay un total de 3991 observaciones y hay a su vez un total de 3070 datos faltantes en todo el dataframe. No se encuentran valores duplicados.

4 variables son de tipo numérico (`id, season, number y runtime`), 4 variables son de tipo texto (`url, name, _links_self_href y _links_show_href`), 1 variable es categórica (`type`) y 2 son tipo fecha (`airdate, airtime y airstamp`). A pesar de esta lectura que hace la librería, al realizar una visualización manual de las variables del dataframe en el notebook, encontramos que las variables tienen otros Dtype, por lo que se requerirá realizar algunos cambios.

En los tipos de alertas se encuentra información como la alta correlación entre las variables *`type`* y *`number`*, sin embargo, esto puede no informar mucho, pues se encuentra que la variable *`type`* está desbalalceada (tal como se podrá ver más adelante). Se resumen las variables com mayor número de datos faltantes y aquellos con datos únicos.


Se procederá a realizar un análisis de algunas variables individuales.

*Variable `id`*

En primera instancia, al ver que efectivamente el número de valores únicos es igual al número de observaciones, significa que hay un único id para cada episodio, por lo que, para facilitar el trabajo se hace necesario realizar cambio de tipo de variable para que sea tipo índice o *index*. No se encuentran valores negativos, valores nulos, ni tampoco ceros por lo que se reafirma el cambio.

*Variable `url`*

Esta variable presenta el URL de cada episodio emitido, hay tantos episodios como links diferentes, por lo que no hay duplicados ni valores nulos. No se hace necesario cambios.

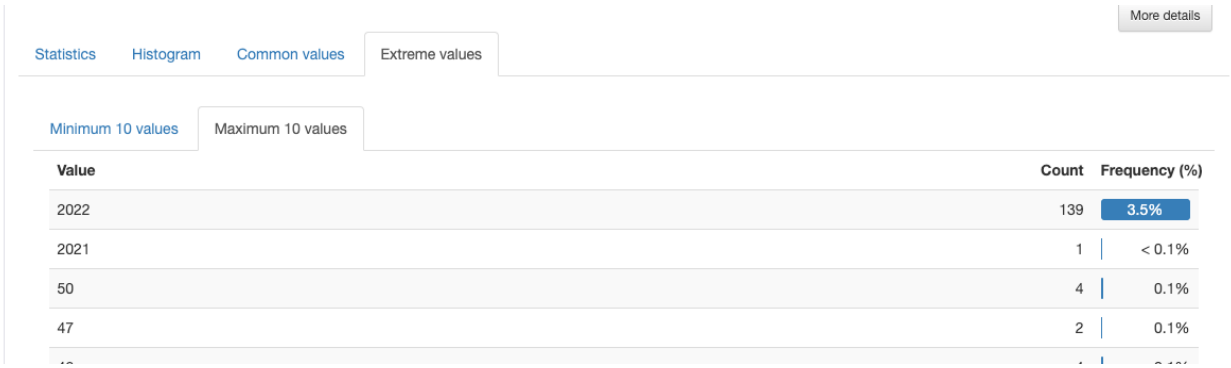
*Variable `name`*

Corresponde al nombre de los capítulos. La variable cuenta con datos duplicados, al revisar en el notebook de Python, utilizando la función de *value_counts* de pandas, encontramos que los valores duplicados se dan porque el título comienza con la parabra “Episodio #”, esto en diferentes idiomas. Adicionalmente no hay datos nulos, por lo que no se realizan modificaciones a la variable.

*Variable `season`*

Es visible que, al ser una variable numérica que refleja la temporada en la que se encuentra el capítulo, hay muchos valores iguales (solo 31 distintos). El mínimo es 1 y el máximo 2022. En este sentido podemos ver que hay probablemente valores atípicos, podemos validar en el *html* correspondiente que la temporada 2022 se repite 139 veces y la temporada 2021 se repite 1 vez, correspondiendo a los valores más diferentes de los demás. Por esto se

procederá a realizar la eliminación de dichos datos atípicos (pues no se tiene certeza de la temporada a la que pertenecen).




*Variable `number`*

Corresponde al número del episodio emitido. Se puede ver que los datos tienen una distribución no uniforme, pues la mayoría (el 52,9%) se encuentran entre los números del 1 al 8. El resto de los datos se distribuyen entre el numero 9 y el número 328. Se encuentra también que hay 84 valores nulos que pueden llegar a intervenir a la hora de hacer posteriores análisis. Debido a esto, y teniendo en cuenta el bajo porcentaje (2,1% del total) se procede a eliminarlos.

*Variable `type`*

Esta variable muestra el tipo de episodio. Hay 3 tipos, *regular, insignificant_special y significant_special*, sin embargo aproximadamente el 98% de los datos tienen a "regular" por tipo. El notebook de Python reconoce esta variable como `object`, por lo que se procederá a realizar el cambio a `category`.

*Variable `airdate`*

Es la fecha en la que salió al aire el episodio. No se presentan datos nulos, sin embargo, se procederá a eliminarse, puesto que la variable *airstamp* contiene la misma información más la hora.

*Variable `airtime`*

Es la hora en la que se emite el episodio. Al momento de revisar el profiling, se encuentra que hay muchos datos nulos, pues se ubican en el 68.4% de los datos totales. Si bien es cierto que es una variable que puede ser interesante de analizar, se procederá a eliminarla por completo.

*Variable `airstamp`*

Esta variable proporciona información de la fecha y hora en la que se emite el episodio, por lo que no solo es redundante con las dos anteriores, sino que se encuentra mas completa al no contener datos nulos. Por esto se procede a cambiar el formato a datetime para poder realizar trabajos con pandas de manera más sencilla en un futuro (utilizando por separado los días del mes o la hora o minutos únicamente).

*Variable `runtime`*

Hace referencia a la duración del episodio en minutos. Se encuentran datos nulos, por lo que será necesario eliminarlos, debido a que son solo 256, del total de datos. Hay ciertos valores atípicos con duraciones incluso mayores a los 180 minutos (3 horas) o valores poco realistas, con episodios de menos de 10 minutos. Debido a la cantidad en cada uno, se procederá a eliminar los episodios con 1 minuto (solo 5) y aquellos con más de 180 minutos (6 episodios).

*Variable `_links_self_href`*

Esta columna proporciona el enlace único de cada episodio. No hay valores repetidos. No se realizará modificaciones.

*Variable `_links_show_href`*

Corresponde al link del programa general que contiene los episodios de que trata el dataframe. No se realizará modificaciones.

## Dataframe *episode_ratings_df*

El dataframe cuenta con 2 variables, `id` y `rating_average`, ambas variables numéricas. A pesar de la baja cantidad de variables, hay un total de 3492 datos faltantes que corresponden en su totalidad a la variable de ratings. De esta manera se considera no eliminar ni tratar estos datos faltantes, toda vez que la eliminación podría hacer que estadísticas tan básicas como la media, reflejen datos alejados de la realidad.

**Dataset statistics**

| | |
|---|---|
| Number of variables | 2 |
| Number of observations | 3991 |
| Missing cells | 3492 |
| Missing cells (%) | 43.7% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 62.5 KiB |
| Average record size in memory | 16.0 B |

**Variable types**

| | |
|---|---|
| Numeric | 2 |

## Dataframe *shows_df*

El dataframe hace referencia a los programas o shows y cuenta con 15 variables diferentes. Hay un total de 2835 datos nulos, sin embargo este corresponde solo al 4.7% del total de datos que contiene el dataframe. Hay 640 filas duplicadas, por lo que la primera modificación que se realizará es la depuración de dichos datos.

La tabla de alertas nos permite encontrar inicialmente algunas variables que pueden necesitar algún tipo de modificación o limpieza. Estas se detallarán más delante de manera individual.

**Alerts**

| | |
|---|---|
| Dataset has 640 (16.0%) duplicate rows | Duplicates |
| `_embedded_show_id` is highly overall correlated with `_embedded_show_weight` | High correlation |
| `_embedded_show_runtime` is highly overall correlated with `_embedded_show_averageRuntime` | High correlation |
| `_embedded_show_averageRuntime` is highly overall correlated with `_embedded_show_runtime` | High correlation |
| `_embedded_show_weight` is highly overall correlated with `_embedded_show_id` | High correlation |
| `_embedded_show_language` has 102 (2.6%) missing values | Missing |
| `_embedded_show_runtime` has 2101 (52.6%) missing values | Missing |
| `_embedded_show_averageRuntime` has 237 (5.9%) missing values | Missing |
| `_embedded_show_officialSite` has 395 (9.9%) missing values | Missing |
| `_embedded_show_genres` is an unsupported type, check if it needs cleaning or further analysis | Unsupported |
| `_embedded_show_weight` has 53 (1.3%) zeros | Zeros |

### *Variable `_embedded_show_id`*

Proporciona el id del programa. Inicialmente será necesario convertir esta variable a índice, esto con el fin de hacer más ágil las operaciones posteriores. Hay 749 datos distintos, esto quiere decir que esos son los shows a los que hace referencia en su totalidad el dataframe. Se convertirá el tipo de dato a índice.

### *Variable `_embedded_show_url`*

Corresponde al enlace del show. Hay, al igual que índices, 749 datos diferentes. No hay datos nulos. No se realizarán modificaciones.

*Variable _embedded_show_name*

Hace referencia al nombre del programa. Hay 749 datos diferentes y no se registran datos faltantes. No requiere limpieza de datos.

*Variable _embedded_show_type*

Esta variable categórica detalla el tipo de show en el que catalogan cada programa. Hay 10 distintas categorías, encontrándose que la mayor es “scripted”, en el que se encuentran 2217 shows en total (el 55.5%). Al verificar las distintas categorías no se encuentra necesidad de realizar cambios a la variable.

*Variable _embedded_show_language*

Detalla el lenguaje en el que se desarrolla el programa. Hay 35 diferentes categorías, sin embargo se encuentra que hay 102 valores faltantes, que solo corresponden al 2.6% del total de datos. Se procederá a eliminar dichos valores nulos, los cuales pueden llegar a ser significativamente menos después de la limpieza de datos que se ha hecho hasta el momento. El más del 61% de los shows se emiten en idioma inglés y chino.

*Variable _embedded_show_genres*

La variable se encuentra no soportada por la herramienta pandas_profiling, esto debido a que los datos que contienen se encuentran almacenados como lista de strings dentro de cada celda (dentro de corchetes cuadrados []) por lo que se procederá a realizar una representación One-Hot, que hará que se creen diferentes columnas por cada uno de las categorías y llenará con 0 o 1 en caso de coincidir o no. De esta manera se puede hacer un manejo más sencillo de todo el dataframe. Posterior a esto se elimina la columna original *_embedded_show_genres*. A pesar del tratamiento, se encuentra la mayoría de datos son nulos, sin embargo al ser una cantidad considerable (982) y ser una variable de interés no se eliminarán. Se muestra un ejemplo de los datos que contiene hasta ahora la variable.

| _embedded_show_genres |
|---|
| [Children] |
| [Children] |
| [Comedy, Crime] |
| [Comedy, Crime] |
| [] |
| [Drama, Thriller, Mystery] |
| [Drama, Thriller, Mystery] |
| [Drama, Thriller, Mystery] |

| _embedded_show_genres | |
|---|---|
| [] | 982 |
| [Drama] | 308 |
| [Comedy] | 205 |
| [Romance] | 167 |
| [Drama, Romance] | 163 |

*Variable _embedded_show_status*

Es una variable categórica que representa el estado del programa, es decir si está aún al aire o si ha finalizado. Adicionalmente tiene una categoría adicional en caso de que aún no se haya determinado si termina o no. Hay por tanto 3 categorías diferentes, no se encuentras datos nulos. Se cambiará a variable categórica.

*Variable _embedded_show_runtime y _embedded_show_averageRuntime*

Se trata de dos variables numéricas que indican la duración de los programas. Ambas tienen una media muy similar (34.2 y 46.4 respectivamente), valores mínimos y máximos parecidos, y están altamente correlacionados (ver heatmap) por lo que podemos inferir que se refieren a la misma información, a saber, el promedio de duración por cada episodio que lo compone. Hay una diferencia fundamental y es que la variable *_embedded_show_runtime* tiene 2101 datos faltantes, mientras que *_embedded_show_averageRuntime* tiene solo 237. Debido a esto último se determina que se eliminará la primera variable y se eliminaran los datos faltantes del segundo.

*Variable _embedded_show_premiered*

Esta variable indica la fecha en la que se estrenó el programa. Hay 465 datos distintos y no se encuentran datos nulos. Al validar el histograma de fechas se encuentran datos atípicos pues hay un dato incluso de 1976. A pesar de esto, se cree conveniente mantener los datos integralmente.

*Variable _embedded_show_officialSite*

Indica el enlace del sitio web oficial en el que se encuentra el programa de televisión. Hay 676 datos diferentes sin embargo hay 395 datos nulos que corresponden al 9.9% del total. Al considerar las limpiezas realizadas hasta el momento, se considera eliminar las líneas de los datos faltantes, ya que la cantidad de estos puede disminuir significativamente.

*Variable _embedded_show_weight*

La variable representa el peso del programa, el cual corresponde a un número entero entre 0 y 100. No se encuentran datos nulos. No se realizará modificaciones.

*Variable _embedded_show_updated*

Esta variable parece representar un valor único (pues hay 749 distintos elementos) por lo que sería casi un índice adicional o un numero único por cada show. Al tener en cuenta que ya tenemos un id para el dataframe, se procederá con su eliminación.

*Variable _embedded_show__links_self_href*

En esta variable se almacenan las url internas del show. Hay 749 distintos elementos, por lo que no hay duplicados ni tampoco datos nulos. No se realizan modificaciones.

*Variable _embedded_show__links_previousepisode_href*

Esta variable, hace referencia al episodio donde se encuentra anidada la información del show. Es decir que no hace referencia al programa en general sino al episodio principal, debido a lo cual se eliminará del dataframe.

## Dataframe *show_schedule_df*

El dataframe hace referencia a los horarios en los cuales se emiten los programas o shows. Tiene 3 variables diferentes, sin embargo, en general tiene 2713 datos faltantes. Una variable es numérica, otra es fecha y hay una final que no es identificada, la cual corresponde a una lista de strings correspondiente a los días en los que se emite.

**Dataset statistics**

| | |
|---|---|
| **Number of variables** | 3 |
| **Number of observations** | 3991 |
| **Missing cells** | 2713 |
| **Missing cells (%)** | 22.7% |
| **Duplicate rows** | 640 |
| **Duplicate rows (%)** | 16.0% |
| **Total size in memory** | 93.7 KiB |
| **Average record size in memory** | 24.0 B |

**Variable types**

| | |
|---|---|
| **Numeric** | 1 |
| **DateTime** | 1 |
| **Unsupported** | 1 |

*Variable _embedded_show_id*

Corresponde al índice del show que se utilizará. Al igual que en el dataframe anterior, hay 749 diferentes datos únicos, por lo que se realizará una limpieza de datos repetidos,

intentando mantener aquellas filas con menos cantidad de datos nulos. Posterior a ello se cambia a tipo índice para mejor manejo del dataframe.

*Variable _embedded_show_schedule_time*

Corresponde a la hora del día en el que se transmite el programa. Hay, sin embargo 2713 datos nulos, los cuales pueden corresponder a los datos duplicados en general. Se considera una variable que puede ser de importancia posterior, por lo que no se realizarán modificaciones.

*Variable _embedded_show_schedule_days*

Esta variable muestra los días de la semana en los que se transmite el programa. Se utiliza la función de value_counts de pandas en Python para poder determinar la estructura de la variable, toda vez que al ser una variable con cadenas de texto almacenadas en listas, el pandas_profiling no puede hacer un análisis. Tiene 1136 datos nulos, sin embargo, no se considera conveniente eliminar los datos faltantes. Se realizará una conversión de las listas a cadenas de texto plano separadas por comas.

| _embedded_show_schedule_days | |
|---|---|
| [] | 1136 |
| [Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday] | 599 |
| [Thursday] | 440 |
| [Friday] | 332 |
| [Wednesday] | 230 |
| [Tuesday] | 151 |
| [Sunday] | 145 |
| [Monday, Tuesday, Wednesday, Thursday, Friday] | 139 |
| [Monday] | 122 |
| [Saturday] | 107 |
| [Monday, Tuesday, Wednesday, Thursday] | 57 |
| [Monday, Tuesday] | 44 |
| [Tuesday, Thursday] | 44 |
| [Monday, Tuesday, Wednesday] | 34 |
| [Friday, Saturday, Sunday] | 32 |
| [Monday, Thursday] | 31 |

## Dataframe *show_ratings_df*

Este dataframe muestra dos variables únicamente, el id de los shows y el promedio de rating que obtienen. Hay en total 3375 datos faltantes, correspondientes al 42.3% del total de datos del dataset, asimismo hay 640 filas duplicadas, por lo que la primera modificación que se realizará es la depuración de dichos datos.

*Variable _embedded_show_id*

En esta variable se alojan los id del dataframe. Hay 749 datos distintos y no se presentan datos nulos. Se procede a cambiar el tipo de variable a index.

*Variable _embedded_show_rating_average*

Corresponde al promedio de rating que tienen los distintos shows. Se encuentra un gran numero de datos faltantes, 3375 para ser exactos, por lo que hay solo 616 lineas, menos que los 749 de índice que se habían considerado antes, es decir que hay 133 filas sin promedio de rating. No se encuentran datos atípicos, teniendo un mínimo de 1 y un máximo de 8.8, por lo que se considera relevante mantener todos los datos sin eliminar datos nulos, esto teniendo en cuenta que puede ser de interés para relacionar próximamente los dataframes.

## Dataframe *webchannel_df*

El dataframe refleja las características que tienen los canales en los que transmiten los programas. Contiene 7 variables diferentes, 2 numericas, 2 en texto y 3 categoricas. Hay 640 datos duplicados, por lo que lo primero que se realizará es esta limpieza, tomando como argumento el id.

**Dataset statistics**

| | |
|---|---|
| **Number of variables** | 7 |
| **Number of observations** | 3991 |
| **Missing cells** | 6649 |
| **Missing cells (%)** | 23.8% |
| **Duplicate rows** | 640 |
| **Duplicate rows (%)** | 16.0% |
| **Total size in memory** | 218.4 KiB |
| **Average record size in memory** | 56.0 B |

**Variable types**

| | |
|---|---|
| **Numeric** | 2 |
| **Text** | 2 |
| **Categorical** | 3 |

*Variable _embedded_show_id*

Corresponden al id de los shows. Hay 749 datos diferentes, por lo que se realizará una limpieza de datos, tal como se mencionó anteriormente, utilizando como argumento este id. De esta manera se mantendrán las filas con menos cantidad de datos faltantes. Después de ello se convierte la variable a tipo índice.

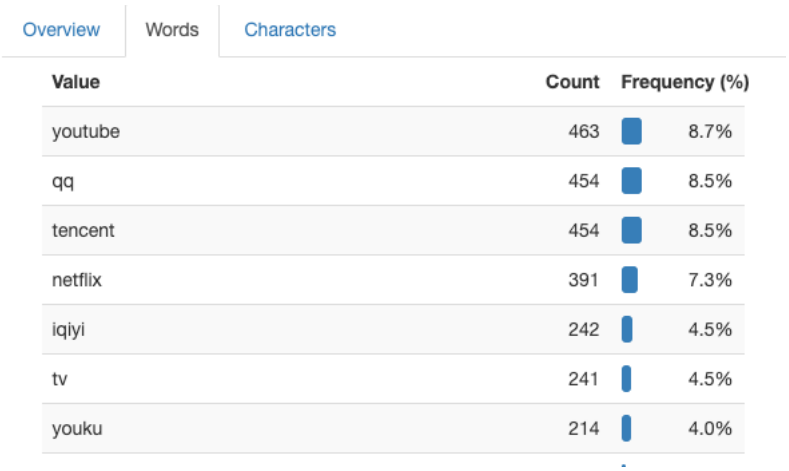
*Variable _embedded_show_webChannel_id*

Se trata del canal web en el que se emiten los programas. Es una variable numérica que va desde el 1 hasta el 573. Tiene 100 datos nulos, por lo que se procederá a eliminarlos, toda vez que corresponden solo al 2.5%.

*Variable _embedded_show_webChannel_name*

En esta variable se almacena información del nombre del canal web donde se emite el programa. Hay 100 datos faltantes, que se eliminarán. Se encuentra que el canal en donde

más se emiten programas es en youtube, qq, Tencent, netfix, entre otros (hay 153 distintos datos).



Overview | Words | Characters

| Value | Count | Frequency (%) |
|---|---|---|
| youtube | 463 | 8.7% |
| qq | 454 | 8.5% |
| tencent | 454 | 8.5% |
| netflix | 391 | 7.3% |
| iqiyi | 242 | 4.5% |
| tv | 241 | 4.5% |
| youku | 214 | 4.0% |

*Variables* *_embedded_show_webChannel_ country_name, _embedded_show_webChannel_country_code y _embedded_show_webChannel_country_timezone*

Se describen las tres variables dado que están todas correlacionadas con índice 1 en el Heatmap, debido a esto se decide solamente trabajar con una de las tres, a saber, country_name, y se eliminan las demás. Country_name tiene 30 distintas categorías, sin embargo, se tienen 1816 datos nulos. Debido a que puede ser una variable de interés futuro y a que la cantidad de datos nulos es muy grande, no se realizaran eliminación por este concepto.

*Variable* *_embedded_show_webChannel_officialSite*

En esta variable indica sitio oficial del canal en el que se transmite el programa o show. Hay 1001 datos faltantes, y considerando además que el link hace referencia a casi la misma información del nombre del sitio (nombre Netflix, sitio oficial www.netflix.com) se procede a eliminar la variable.