Análisis – Profiling

episodes_df:

1. La columna type está desbalanceada, lo que significa que la mayoría de sus valores corresponden a la misma categoría, por lo tanto, se puede considerar una constante lo que afectaría el modelo que se vaya a construir, por ende, se eliminará la columna para futuros análisis.
2. Más del 50% de los valores son nulos en las columnas airtime, rating_average,image,summary, por lo tanto, se eliminarán debido a que el porcentaje de valores faltantes es alto y no es recomendable remplazar esos valores utilizando métodos estadísticos.
3. La columna runtime tiene el 9.4% de valores faltantes, esta columna es cuantitativa y describe el tiempo de ejecución de los programas, por ende, se completarán esos valores faltantes con la mediana.
4. La columna id que es extraída directamente desde la fuente de los datos es un valor único, por lo tanto, será la llave principal para relacionarla con las demás tablas.

links_episodes_df:

1. Se tienen 692 show diferentes
2. No existen columnas con valores faltantes, por lo tanto, no se realizará ninguna limpieza.

embedded_df:

1. La columna schedule_time no muestra que tiene valores faltantes, porque en los datos originales aparecen como si tuviera un objeto, por lo tanto, estos valores se convertirán a None, para que el programa los tome como datos faltantes. A pesar de que la columna tiene el 57.8% de valores que se consideran faltantes, por el momento no se eliminará ya que se considera importante y dependiendo de la aplicación se tomará la decisión si se mantiene o si se elimina.
2. El 7.2% de los valores de language son faltantes, por lo tanto, se remplazará estos valores por la moda, en este caso será por el valor English, adicional se transformará los valores a minúscula.
3. La columna runtime se eliminará debido a que tiene el 74.8% de los datos como valores faltantes, adicional, en este mismo dataframe tiene la columna de averageRuntime que muestra el promedio de este valor.
4. Los valores faltantes de averageRuntime se remplazarán por el valor de la columna runtime del dataframe de episodes, teniendo en cuenta el id del episodio que será la llave.
5. La columna ended tiene el 64.1% de datos faltantes, por lo tanto, no sería bueno mantenerla ya que puede afectar el rendimiento de los modelos que se vayan a aplicar, por ende, se eliminará.

6. La columna officialSite, tiene el 10.1% de los datos, al ser una columna de texto son pocas las herramientas de ciencia de datos que permitan remplazar estos valores faltantes, por lo tanto, se mantendrá tal y cómo está ya que muestra información relevante.
7. Las columnas rating_average, network, dvdCountry, image, webChannel, tienen más del 80% de valores faltantes, por lo tanto, se eliminarán dichas columnas.

genres_df:

1. Este dataset tiene 6682 observaciones y no cuenta con valores faltantes. No se realizará ninguna modificación.

days_df:

1. En este dataframe se observan 3460 distintas observaciones de la columna id_episodes, debido a que no guarda la información de los episodios que tienen el atributo days en blanco.

webChannel_df:

1. Se eliminará la columna country
2. Cuenta con 4691 observaciones y 9 columnas
3. Este dataframe relaciona 32 diferentes países, pero el 31.8% de los datos de country_name son nulos.

image_df:

1. No se realizará ninguna limpieza de los datos.
2. Cuenta con 4554 observaciones, 2 columnas numéricas y 2 de texto.
3. La columna médium cuenta con 662 observaciones únicas.

externals_df:

1. Se eliminarán las observaciones que tienen valores faltantes todas de las siguientes columnas: tvrage, thetvdb, imdb.
2. Cuenta con 4807 observaciones, 4 columnas con datos numéricos y 1 columnas de texto.

links_show_df:

1. No se realizará ninguna limpieza de los datos.
2. Se cuenta con 4807 observaciones y 5 columnas, de las cuales 2 son numéricas y 3 de tipo texto

Nota: Los nombres de los dataset que están en rojo, significa que hay que realizarle una limpieza de datos.