Análisis *profiling* de *DataFrames* proyecto *Technical Test LB*

Autor: Andrés Mendoza

A continuación, se presenta el análisis realizado sobre el proceso de *profiling* sobre los *DataFrames* creados en el proyecto *Technical Test LB.*

Profiling DataFrame EMISIONES

Como se puede ver en el archivo *.html* del *profiling* de este *DataFrame*, hay una gran cantidad de celdas vacías correspondiente al 12.2% del total de celdas. Al revisar en detalle, se puede encontrar que la gran mayoría de las celdas vacías se encuentran en las columnas *rating.average* y *summary.* Si bien, se puede omitir las celdas vacías de la columna *summary* ya que inicialmente no son relevantes para el modelo de datos que se desea crear; la columna *rating.average* sí representa un problema, ya que es una variable importante en el modelo de datos a generar y sólo se cuenta con el 9.9% de los datos, esto puede significar poca exactitud en el resultado final.

Otro caso similar, pero menos grave, es el de la columna *runtime*, el cual también representa información muy relevante para el modelo y cuenta con un porcentaje de datos faltantes del 7.5%

Con respecto a la integridad de la información, se comprueba que todos los registros son únicos y que no existen correlaciones importantes o lógicas entre las columnas numéricas del *DataFrame*, lo cual reduce el ruido del modelo final.

Profiling DataFrame SERIES

Con respecto al *DataFrame* Series se puede ver que el porcentaje de celdas vacías en menor que el del *DataFrame* Emisiones. En este caso, las dos columnas con más celdas vacías son las columnas *ended* y *runtime*. Al ser este un *DataFrame* que se plantea como un objeto a utilizar para buscar información sobre las emisiones, específicamente para la columna "*País*", la cantidad de celdas vacías las columnas *ended* y *runtime* no representa una pérdida considerable de información.

Se puede ver además una alta correlación entre dos columnas: *runtime* y *averageRuntime*, que es una relación lógica al asumir que *averageRuntime* representa el promedio de la columna *runtime* a través del tiempo.

***Profiling DataFrame* GENEROS**

Como es esperado, para este *DataFrame* construido con la información de la lista de géneros de cada serie, toda la información está disponible, es decir, no hay celdas vacías.

Por otro lado, se observan datos duplicados, los cuales se deberán depurar con un proceso de ETL para obtener información única en el modelo de datos final.

Conclusión con respecto a la limpieza de datos

Si bien existe una gran cantidad de datos vacíos en el *DataFrame* principal y esto supone una inexactitud importante en el modelo de datos final, se decide a no eliminar las filas que tengan algún dato vacío, pues en ellas se puede encontrar información relevante para otro el modelo; además, la existencia de estos datos faltantes puede llevar a considerar otra fuente de datos con información más completa.