Análisis de los resultados del profiling de los dataframes construidos

Introducción

Este pequeño documento contiene un análisis de la descriptiva realizada con pandas profiling sobre los dataframes construidos. Los dataframes, se construyen con base en los posibles identificadores únicos que se encontraron con el fin de ver si había tablas descriptivas que se deban convertir en dimensiones e identificar las tablas de hechos, con el fin de en pasos anteriores elaborar un modelo de datos, que le permita a los analistas de datos, realizar reportes con indicadores que contesten a las preguntas de negocio.

1.Análisis del dataframe de "embedded_show"

Overview

Este dataframe se construyó tomando los campos que sirvieron para armar la dimensión de show. Pero el profiling se hizo sobre el dataframe en crudo, en el que solo se hizo la selección de las columnas de la tabla original. Con el perfilamiento se detecto que el dataframe quedo con 17 variables, de estas cinco son de tipo numérico, nueve de tipo categórico, dos tipos fecha y a una variable no le detecto tipo de dato, que fue la de "_embedded.show.network", ya que, esta venía completamente vacía.

El reporte informo 11283 celdas vacías, es decir, el 21,5% del dataframe no tiene datos. Además, se encontró que el 16.6% del dataframe son de registros duplicados, es decir, 511 registros, el número de duplicados fue un indicativo de que esta tabla podría ser considerada una dimensión, entonces, se hizo una llave por todas las columnas y se evidencio que la llave mencionada, tenía la misma cardinalidad que el "_embedded.show.id". Lo que sirvió para tomar la decisión de normalizar esta tabla, de la sabana inicial.

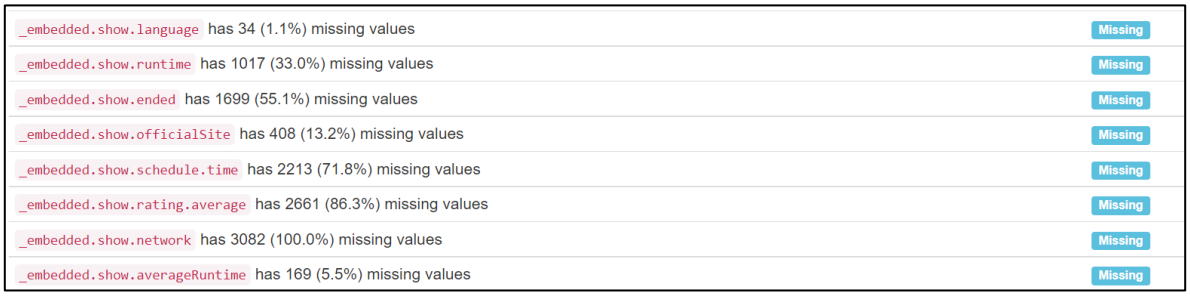
Descriptiva Variables

Con relación a la parte de variables se encontró, que las variables de la **Figura 1**, tienen una alta cardinalidad, es decir, casi todos sus registros son diferentes. (Para un informe real, habría hecho la tabla para que fuese más legible, pero por un tema de tiempo colocaré las imágenes).

Figura 1. Variables con alta cardinalidad del dataframe de embedded_show

En la **Figura 2**, coloco las variables con valores vacíos y su respectivo porcentaje, donde se evidencia que la columna "embedded.show.network", no tiene datos, entonces, habría que validar con el negocio si esto siempre es así y sería una candidata a no incluir en el modelo de datos.

Figura 2. Variables con valores vacíos del dataframe de embedded_show



Correlations

Las **Figuras** de la **3** a la **5**, muestran los coeficientes de correlación para variables numéricas, acá es bueno mencionar que entre más cercano a uno sea el valor indica una correlación positiva entre las variables, es decir, a medida que una crece la otra también lo hace. Y cuando este coeficiente es cercano a menos uno indica una correlación negativa, esto significa que a medida que una crece la otra disminuye. Los coeficientes de las **Figuras** tanto **3** como **5**, se utilizan para identificar correlación no lineal entre las variables. Y el coeficiente de la **Figura 4** (Pearson), se utiliza para determinar correlación lineal. Por último, es importante aclarar, que el hecho de que exista una correlación entre variables no significa que haya una relación de causalidad entre estas.

Las Figuras **6** y **7**, muestran respectivamente los coeficientes de Cramér y Phik, donde un valor de cero indica independencia entre las variables y un valor de uno una asociación o correlación perfecta, pero este tipo de coeficientes se utilizan es para variables categóricas y nominal.

El análisis de correlación es un paso muy importante, cuando se hacen proyectos de machine learning o inteligencia artificial, para comprender los datos desde una perspectiva de como interactúan entre ellos.

Figura 3. *Correlación de Spearman*

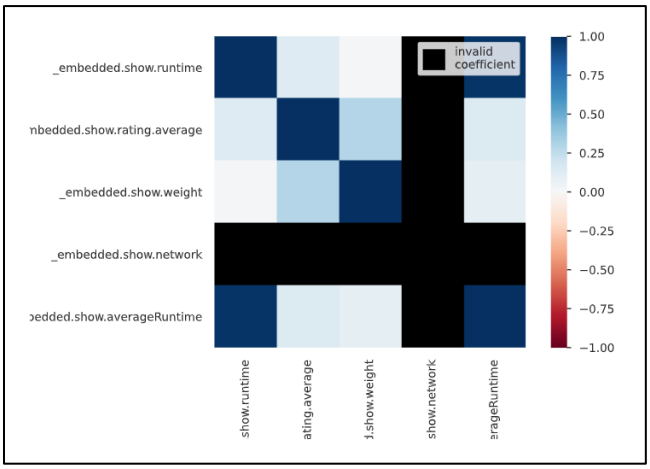


Figura 4. Correlación de Pearson

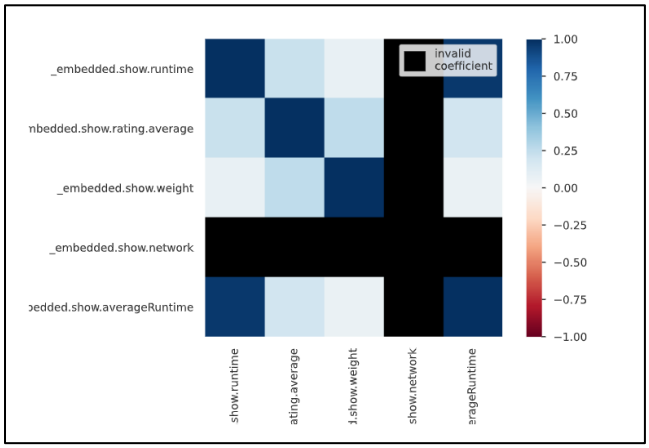


Figura 5. Correlación de Kendall’s

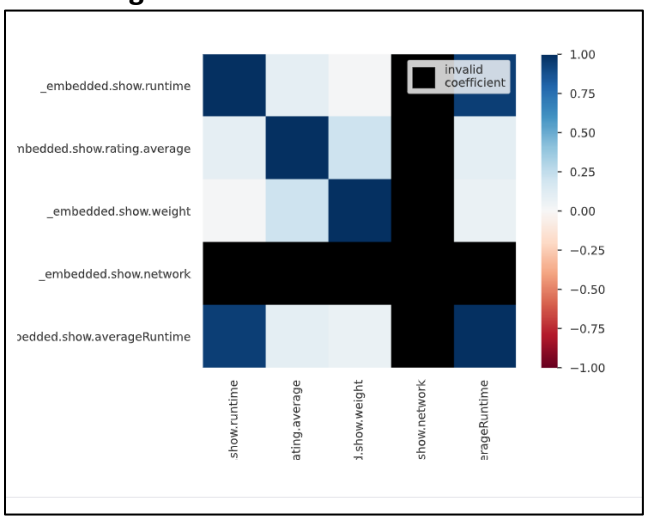


Figura 6. Correlación de Cramer

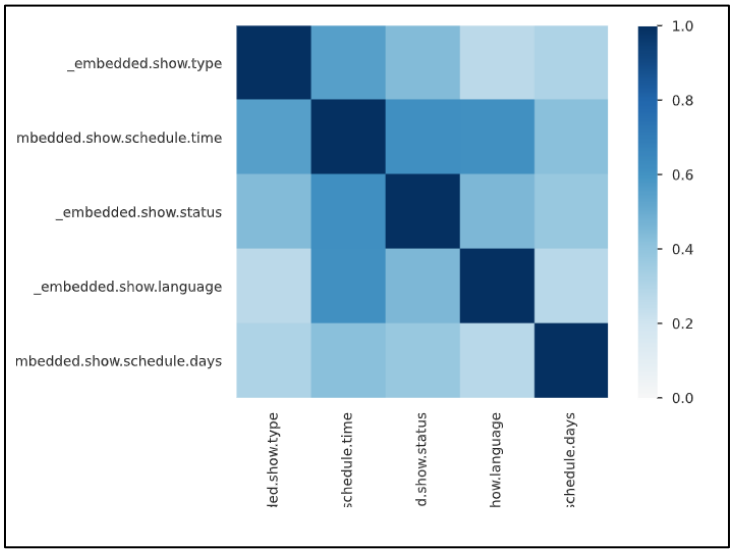
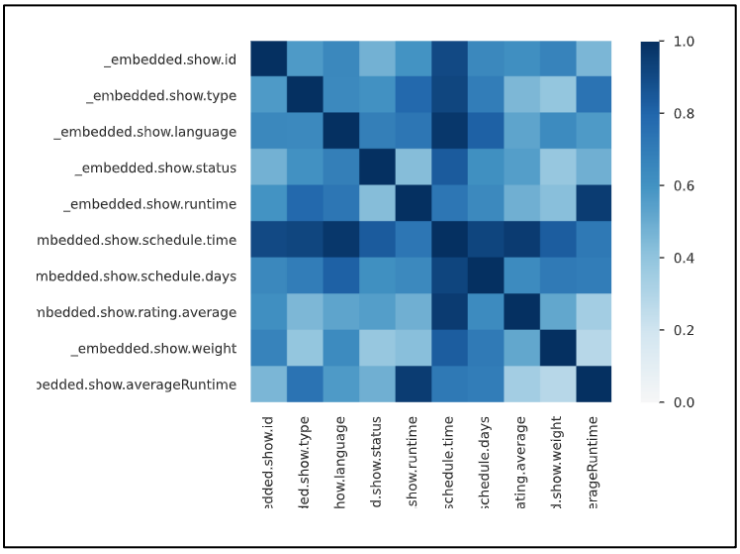


Figura 7. Correlación de Phik



Valores faltantes

Por último, en la Figura 8 y 9, se muestran los valores faltantes del dataframe, donde se aprecian rápidamente que la columna "embedded.show.network", esta completamente vacía.

Figura 8. Valores faltantes vista diagrama de barras

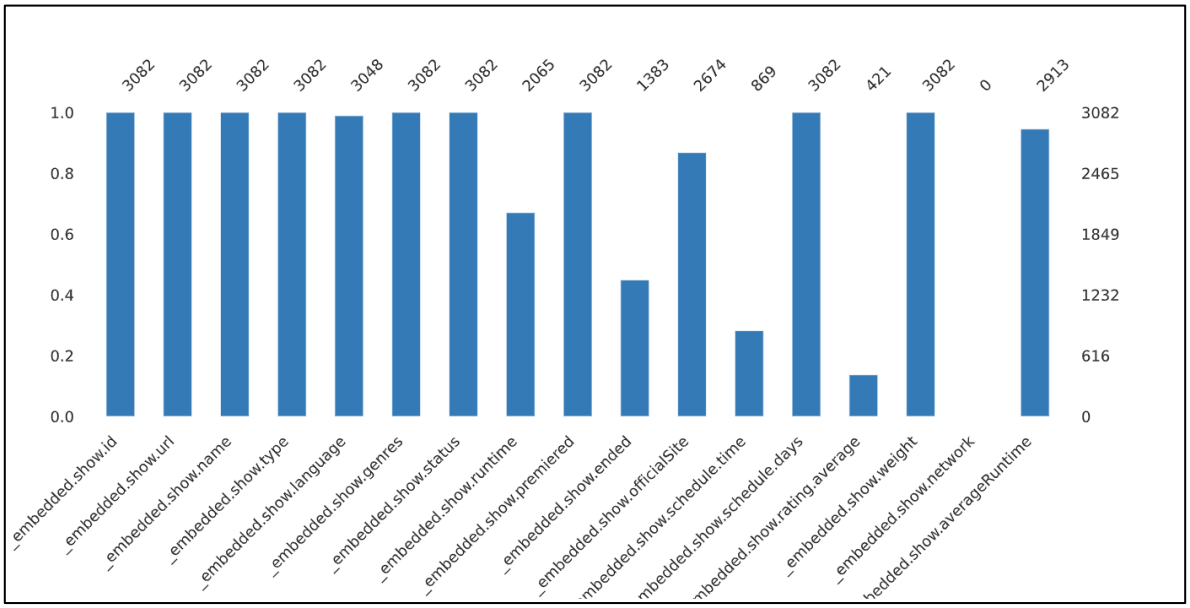
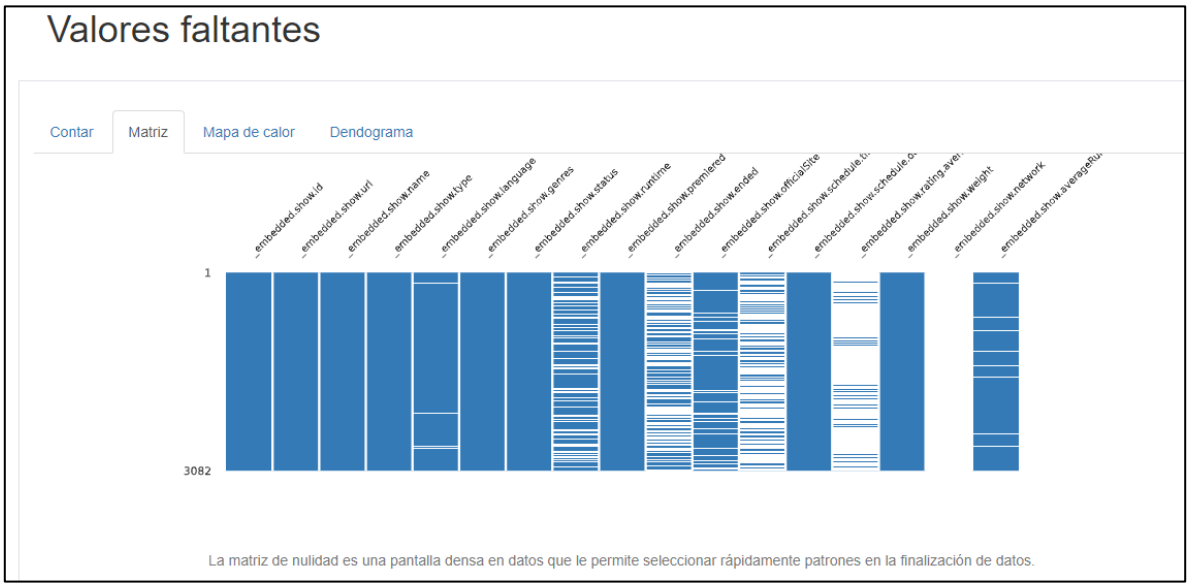


Figura 9. Valores faltantes vista matriz



Para los siguientes dataframes, haré una descripción más corta con el fin de no volver el documento repetitivo, ya que, las ideas de la correlación y los gráficos descritos anteriormente son las mismas solo que cambian los valores obtenidos, y estos es posible consultarlos en los reportes .html generados.

2.Análisis del dataframe de "embedded_show_net_work"

Este dataframe se construyó tomando los campos que sirvieron para armar la dimensión de net_work. Pero el profiling se hizo sobre el dataframe en crudo, en el que solo se hizo la selección de las columnas de la tabla original. Con el perfilamiento se detectó que el dataframe quedo con 6 variables, de estas una es de tipo numérico y cinco de tipo categórico.

El reporte informo 17400 celdas vacías, es decir, el 94.1 % del dataframe no tiene datos. Además, se encontró que el 1,26% del dataframe son de registros duplicados, es decir, 36 registros, el número de duplicados fue un indicativo de que esta tabla podría ser considerada una dimensión, entonces, se hizo una llave por todas las columnas y se evidencio que la llave mencionada, tenía la misma cardinalidad que el "_embedded.show.network.id". Lo que sirvió para tomar la decisión de normalizar esta tabla, de la sabana inicial.

3.Análisis del dataframe de "embedded_show_web_Channel"

Este dataframe se construyó tomando los campos que sirvieron para armar la dimensión de web_Channel. Pero el profiling se hizo sobre el dataframe en crudo, en el que solo se hizo la selección de las columnas de la tabla original. Con el perfilamiento se detectó que el dataframe quedo con 6 variables, de estas una es de tipo numérico y cinco de tipo categórico.

El reporte informo 6161 celdas vacías, es decir, el 33.3 % del dataframe no tiene datos. Además, se encontró que el 4,2 % del dataframe son de registros duplicados, es decir, 130 registros, el número de duplicados fue un indicativo de que esta tabla podría ser considerada una dimensión, entonces, se hizo una llave por todas las columnas y se evidencio que la llave mencionada, tenía la misma cardinalidad que el "_embedded.show.network.id". Lo que sirvió para tomar la decisión de normalizar esta tabla, de la sabana inicial.

4.Dataframe Series

Este dataframe, se creo por el campo de "id", que viene siendo el identificador de la serie y, por ende, este marco de datos fue el insumo que posteriormente se tomo para construir la tabla transaccional, este se construyó tomando la sabana original y quitándole las columnas de los dataframes, que se describieron anteriormente.

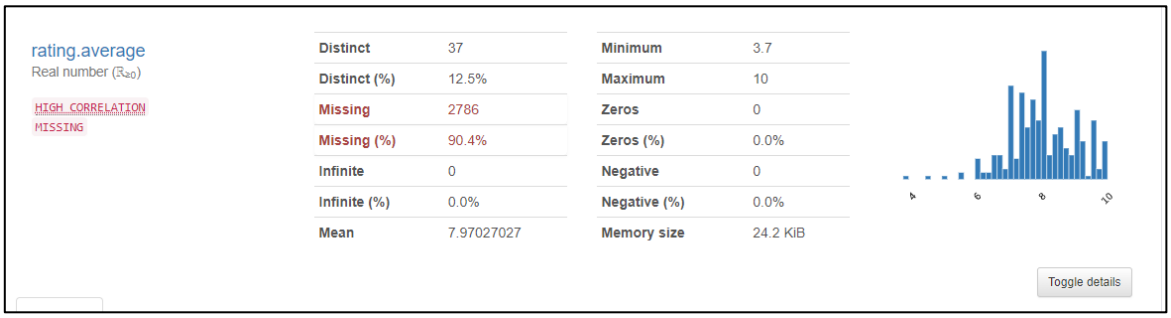
Con el perfilamiento se detectó que el dataframe quedo con 36 variables, de estas 11 de tipo numérico, 17 categóricas, tres tipo fecha y cinco que no se les detecto el tipo de dato, que fueron Image, “embedded.show.dvdCountry”, “_embedded.show.webChannel”, “_embedded.show.Image” y “_embedded_show.webChannel.country”, estas variables, vienen vacías desde la fuente de datos, así que habría que revisar si es por el API, que cuando se abren quedan vacías, o por que solo se extrajeron datos para Diciembre, y con base en esa validación tomar la decisión de eliminarlas, en un ejercicio real, para el ejercicio de prueba técnica que presento, estas fueron eliminadas.

El reporte informo 48008 celdas vacías, es decir, el 43.3 % del dataframe no tiene datos. Además, se encontró que no hay registros duplicados, debido al id de la serie. Entonces, más el entendimiento de los datos oriento la decisión de utilizar este dataframe como insumo para la construcción de la tabla transaccional.

El profiling, también ofrece la posibilidad de analizar estadísticos descriptivos por cada variable, como ejemplo colocaré la variable de rating.average.

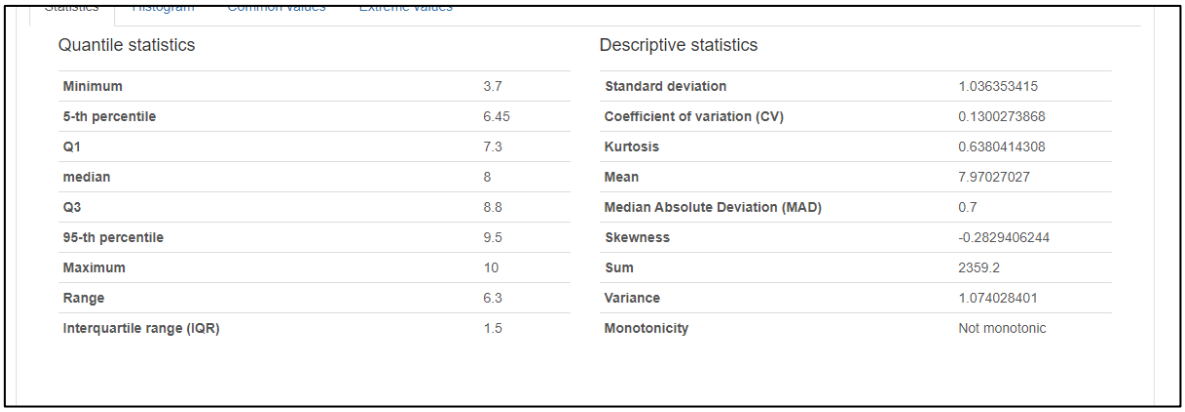
En la **Figura 10**, se aprecia que la variable “rating average”, tiene 37 valores distintos, 2786 valores perdidos y una media o valor promedio de 7.97.

Figura 10. Información de rating average



Y si deseamos profundizar un poco más en las estadísticas descriptivas de las variables, se aprecian los datos de la **Figura 11**.

Figura 11. Estadísticas descriptivas de la variable “rating average”



Quantile statistics

Minimum	3.7
5-th percentile	6.45
Q1	7.3
median	8
Q3	8.8
95-th percentile	9.5
Maximum	10
Range	6.3
Interquartile range (IQR)	1.5

Descriptive statistics

Standard deviation	1.036353415
Coefficient of variation (CV)	0.1300273868
Kurtosis	0.6380414308
Mean	7.97027027
Median Absolute Deviation (MAD)	0.7
Skewness	-0.2829406244
Sum	2359.2
Variance	1.074028401
Monotonicity	Not monotonic

En la **Figura 11**, se ven estadísticos descriptivos, como la media, la desviación estándar, el rango y entre otros.