Profiling Análisis

En este archivo se expone el análisis de los *Profiling Reports* realizados a los dataframes (DFs) construidos con base en la revisión de los datos extraídos de la API y los datos de interés dados por el cliente.

(1) ***df_air_basic_db_VF***:

En este dataframe están incorporados 8 variables entre las cuales hay 1 variable con formato Datetime, 2 variables numéricas y 5 categóricas. Hay 3082 registros.

La variable que entrega información relevante es la de Fecha_alaire que permite ver el comportamiento del conjunto de las series de tv durante las semanas del mes de Diciembre del año 2020. Este comportamiento sugiere que el día en el que mas series de tv están al aire son los días Viernes y Jueves. Entre Sábado y Domingo empieza con un numero moderado de series que van aumentando paulatinamente inmediatamente avanza la semana (Fig.1).

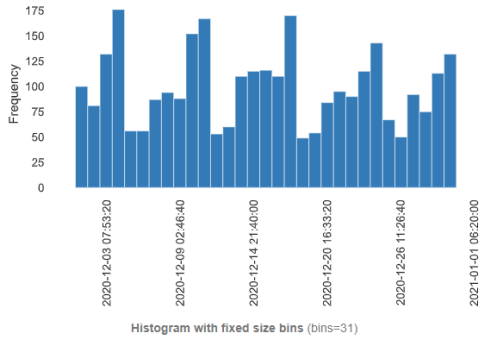

Fig. 1. Frecuencia de las series de televisión al aire durante Diciembre 2020.

En este DF no hay datos nulos y no hay duplicados. Esto se debe a que las variables en este DF hacen parte de las variables principales que entrega el dataset estudiado. Sin embargo, si hay valores que se repiten en variables como name_complete que tienden a sugerir la posible existencia de duplicados (Fig. 2).

Fig. 2. name_complete posibles duplicados.

Al revisar con más detenimiento se encuentra que, a pesar de la cercanía entre valores, variables como url hacen que este DF no tenga duplicados porque su porcentaje de distinción entre datos es 100% (por ello esta variable, url, tiene alta cardinalidad, pues tiene muchos valores únicos; Fig. 3).

Fig. 3. url características.

Revisando la correlación entre variables de este dataframe es relativamente baja, teniendo el valor más alto el correspondiente a id_serie y Tipo (< 0.4; Fig. 4).

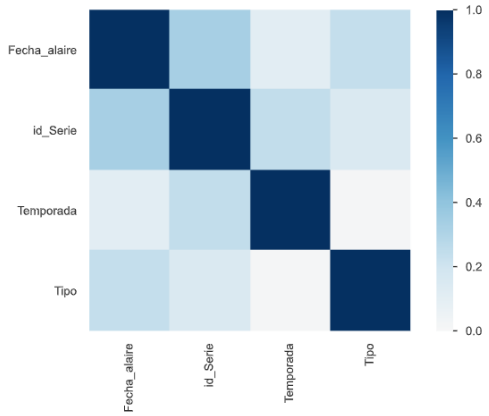

Fig. 4. Phi_k de las variables del *df_air_basic_db_VF.*

Adicionalmente, la variable Serie muestra que 'the-yo' es la que más se repite. Tiene 38 episodios que fueron televisados en Diciembre del 2020 (Fig. 5).

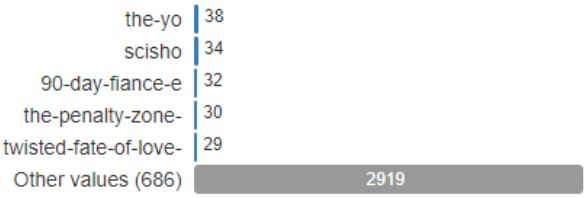

Fig. 5. Series con mas episodios televisados en el dataset estudiado.

(2) ***df_genre_status_db_VF***:

En este dataframe hay 8 variables: 3 numéricas, 4 categóricas y 1 con formato DateTime. Hay 3082 registros.

Aquí las variable más importantes son Genero y Estado que tienen ambas correlación entre sí y con otras variables como id_Serie. Por ejemplo, el coeficiente Phi_K muestra que las variables Genero Estado tiene una correlación cercana a 0.7, sugiriendo que hay una relación matemática positiva entre las dos variables (Fig. 6).

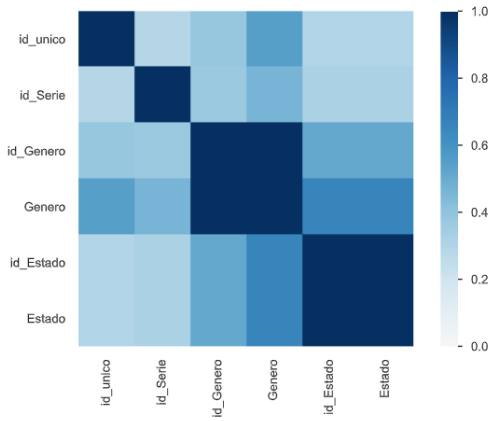

Fig. 6. Phi_k de las variables del *df_genre_status_db_VF*.

Revisando con más detalle estas dos variables, el Género que más se registró (Fig. 7) en el mes de Diciembre del 2020 fue el Genero Drama (35.7%), seguido por sin Genero definido (25.3%), Comedy (17.5%), Action (3.9%) y Crime (3.7%).

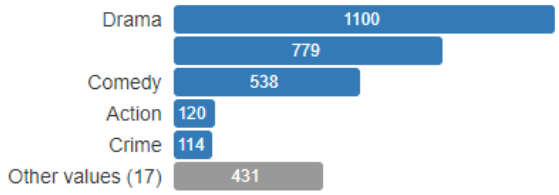

Fig. 7. Géneros registrados de las series televisadas en Diciembre 2020.

En el caso de Estado (Fig. 8), hay un balance entre Ended (que ya finalizó; 44.8%) y Running (que todavía lo están presentado; 42.8%). To Be Determined (Está por definirse) equivale al 22% aproximadamente de los registros restantes.

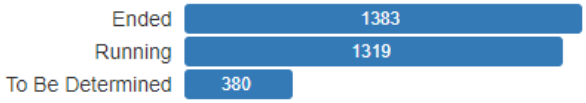

Fig. 8. Estado de las series televisadas en Diciembre 2020.

La relación positiva entre Genero y Estado sugiere que hay una alta posibilidad que las películas en Drama, sin Genero y Comedy estén distribuidas en los dos Estados con más fuerza Ended y Running.

En este dataframe no hay duplicados. Sin embargo, hay datos sin categoría definida para el género equivalente al 25%. Esto se debe revisar y corregir porque son datos nulos.

(3) ***df_country_db_VF:***

En este dataframe hay 5 variables: 2 numéricas y 3 categóricas. Hay 3082 registros.

Este dataframe no hay duplicados. Sin embargo hay 1573 registros que no tienen un País asociado (51.0%), situación que se debe corregir porque estos son datos nulos. Esto puede estar sucediendo porque esta identificando a la categoría vacía como una categoría valida (Fig.).

Las variable que más aporta información es la variable País (Fig. 9). Esta indica que los países con mayor registro en el dataset estudiado de Diciembre del 2020, son China (15.7%), USA (6.3%), Norway (5.7%) y la Republica de Korea (4.6%).

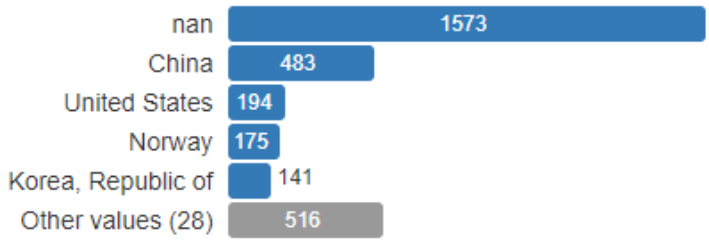

Fig. 9. Países donde se televiso las series de tv registradas en Diciembre 2020.

En términos de correlaciones, el coeficiente Phi_K muestra (Fig. 10) que hay una correlación positiva entre los id_serie y el País. Esto sugiere que hay series que están ligadas a los países en los que mas se ven o transmiten.

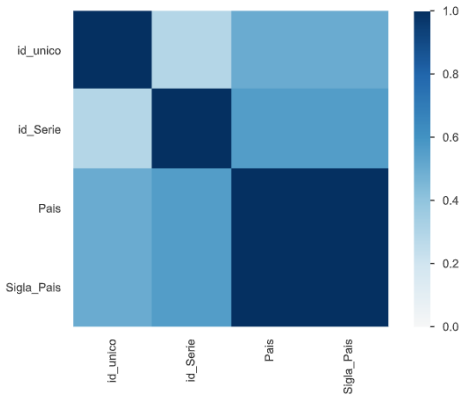

Fig. 10. Phi_k de las variables del *df_country_db_VF*.

(4) ***df_runtime_db_VF:***

En este dataframe hay 6 variables: 5 numéricas y 1 categórica. Hay 2817 registros, por lo que se retiraron 265 registros vacíos.

La variable que más aporta información de este dataframe es la variable runtime. El registro que mas se repite es el equivalente a 45 (17.0% de los datos; Fig. 11).

Value	Count	Frequency (%)
45	478	17.0%
30	183	6.5%
60	139	4.9%
20	136	4.8%
12	97	3.4%
15	82	2.9%
25	82	2.9%
120	81	2.9%
23	72	2.6%
5	62	2.2%
Other values (102)	1405	49.9%

Fig. 11. Valores que mas se repiten y su porcentaje para la variable runtime.

El dataset no indica si este valor son minutos horas o días. Por lo general el runtime hace referencia a la duración del episodio en minutos. Esto quiere decir que 17% de los episodios registrados duran aproximadamente 45 min, seguidos por 30 min (6.5%).

Se puede decir por las interacciones observadas runtime y Avg_runtime (Fig. 12) son variables con aproximadamente los mismos valores. Por lo que se debe escoger una a la hora de analizar los datos.

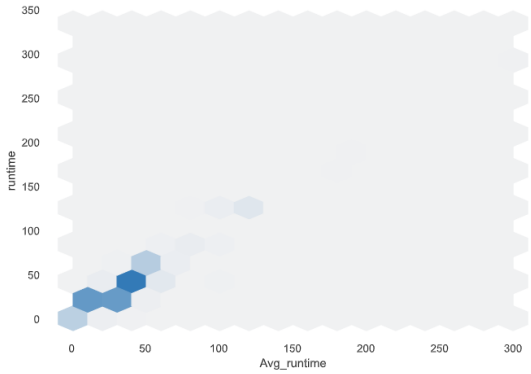

Fig. 12. Runtime vs. Avg_runtime

Por último el coeficiente Phi_K sugiere una relación positiva media entre el id_serie y el runtime (0.4 aproximadamente; Fig.).

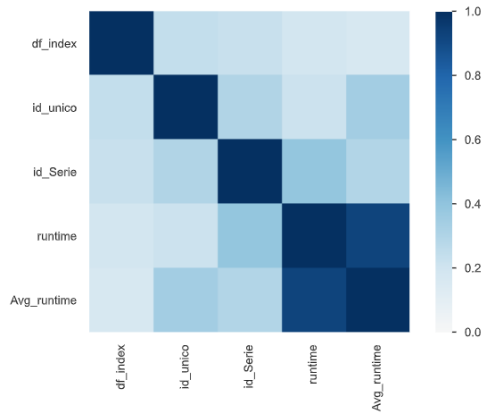

Fig. 13. Phi_k de las variables del *df_runtime_db_VF*.

(5) ***df_ratings_db_VF***

En este dataframe hay 6 variables 5 numéricas y 1 categórica. Este es el DF con el menor numero de registros 250. Esto porque existían muchas filas con valores nulos (2832).

La variable que más aporta información es Avg_rating con valores: mínimo de 3.7 y máximo de 10.

Embedded average rating (E_Avg_rating) es muy diferente a Avg_rating (Fig. 14). Tienen sus puntos relacionados totalmente distribuidos por el espacio, sin formar una distribución o tendencia clara. Importante revisar cuál de los dos tiene mas peso y porque escoger uno o el otro.

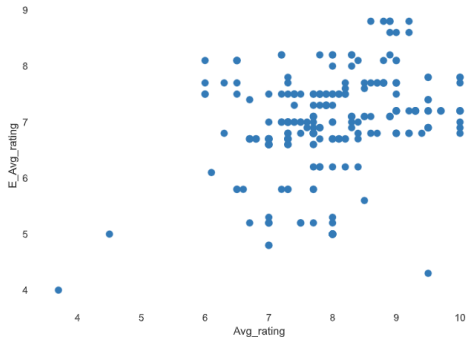

Fig. 14. E_Avg_rating vs. Avg_rating

El coeficiente Phi_k sugiere que hay una relación positiva relevante entre la serie y el rating obtenido (cercana a 1). Esto se presta para hacer una revisión mas de cerca de las series que quedaron en este dataframe.

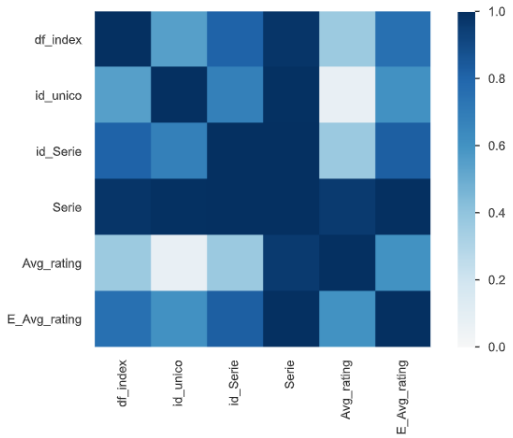

Fig. 15. Phi_k grafico de las variables del *df_ratings_db_VF*.