# **Profiling de los DataFrames**

Por Cristian Julián Mancera Bolivar

*Prueba de Ingreso Lulo Bank*

Tabla de contenido

[Profiling de los DataFrames 1](#_Toc110210821)

[Dataframe Chapters 1](#_Toc110210822)

[Variables 2](#_Toc110210823)

[Airdate: 2](#_Toc110210824)

[Airtime 3](#_Toc110210825)

[Number 4](#_Toc110210826)

[Runtime 5](#_Toc110210827)

[Rating\_average 5](#_Toc110210828)

[Valores nulos 6](#_Toc110210829)

[Dataframe Shows 7](#_Toc110210830)

[Variables 7](#_Toc110210831)

[Average\_runtime 8](#_Toc110210832)

[Language 9](#_Toc110210833)

[Schedule\_days 9](#_Toc110210834)

[Type 10](#_Toc110210835)

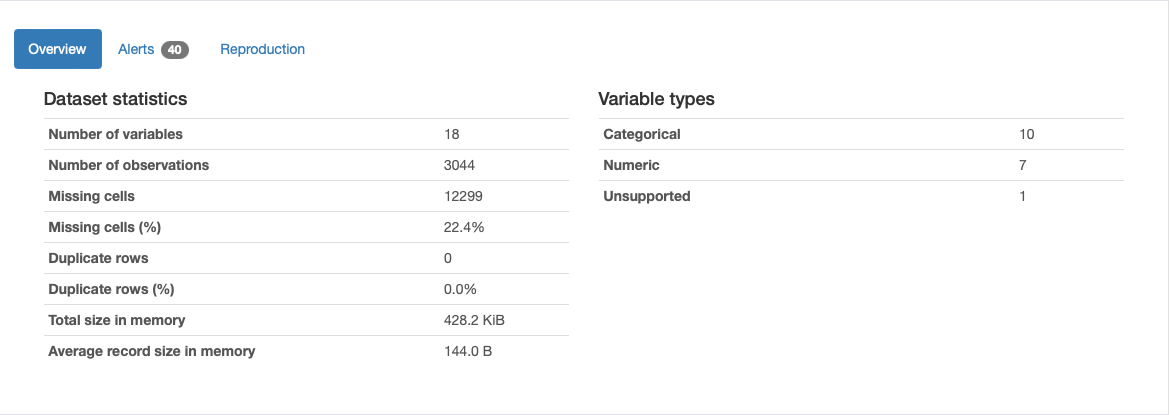
[Valores nulos 10](#_Toc110210836)

[Otros Dataframes 10](#_Toc110210837)

## Dataframe Chapters

Este dataframe describe los capítulos de los shows presentados en diciembre del 2020 por lo cual cada registro presenta un capítulo de su respectivo show

Para este Dataframe Podemos encontrar a continuación las estadísticas generales donde en rasgos generales podemos evidenciar que tenemos 18 variables de las cuales 10 son categóricas 7 numéricas y 1 no soportada (esta es la variable que almacena las url de las imágenes)



### 

### Variables

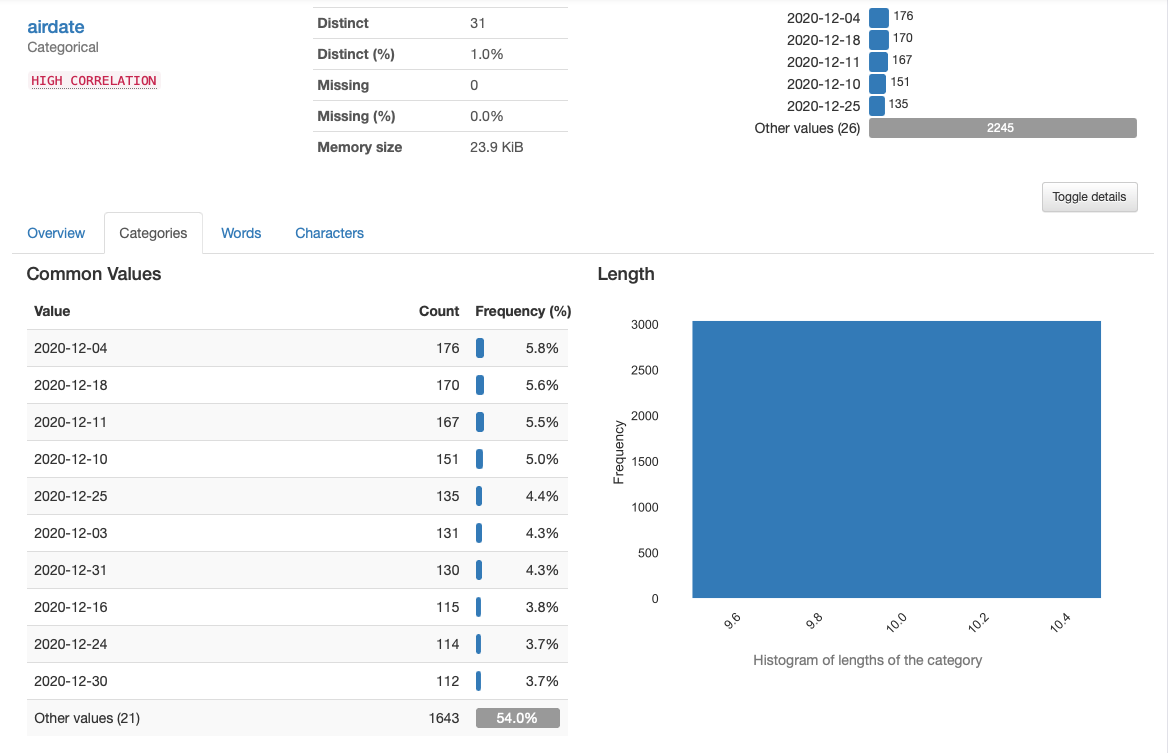
Las 18 variables de este dataframe son las siguientes:

* Id: código único del episodio.
* url: Contiene la url de TV Maze con la información del episodio al que se hace referencia en el registro.
* name: Nombre del episodio.
* season: Temporada de la serie o show a la que pertenece el episodio
* number: Número del episodio.
* type: es una variable categórica con 3 opciones las cuales son: regular, significant special y insignificant special.
* airdate: Fecha del estreno.
* Airtime: hora del estreno.
* Airstamp: No identificado
* Runtime: tiempo del episodio en minutos
* Image: url de la imagen del episodio
* Summary: resumen del episodio.
* rating\_average: rating promedio del episodio
* \_links\_self\_href: Link a la información del episodio dada por el API.
* image\_medium: imagen del episodio en tamaño mediano.
* image\_original: imagen del episodio en tamaño original.
* show\_id: id único del show o serie al que pertenece el episodio.

Como podemos evidenciar hay muchas variables con información de texto o categóricas que a pesar de ser numéricas representan una categoría y no una correlación con otras variables, por tal motivo seleccioné para analizar las variables que nos pueden dar una información comparativa más significativa de los episodios

#### Airdate:

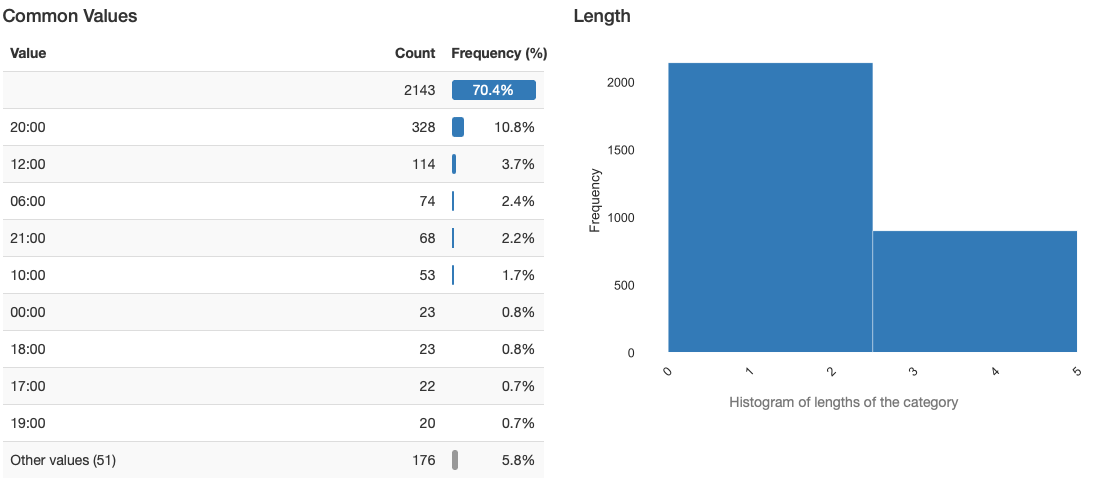
Esta variable nos puede dar información valiosa de que día del mes se estrenan más episodios



Como podemos evidenciar por una diferencia muy corta el día en el que más se lanzaron episodios es el 4 de diciembre del 2020, siguiéndole el 18 y el 11, los 3 fueron días viernes lo cual puede hacernos ver que este día es uno de los favoritos para el estreno de episodios.

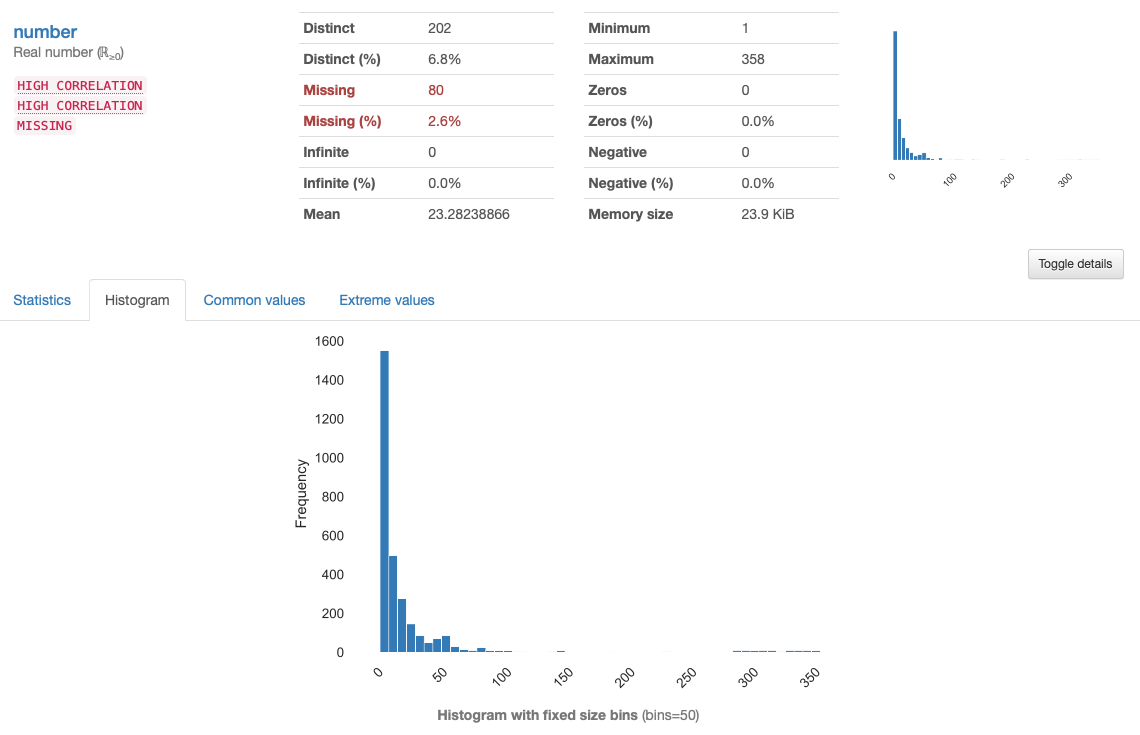
#### Airtime

A continuación evidenciamos que no hay información del 70.4% de los episodios por lo cual el resultado del análisis podría verse sesgado, de igual forma dentro de los datos que tenemos se evidencia una tendencia a que los capítulos inicien a las 8 pm , lo cual podría indicar que es un tiempo ideal para captar la atención del televidente, es el tiempo que se denomina prime time.



#### Number

Esta variable nos habla del número del episodio que se está presentando

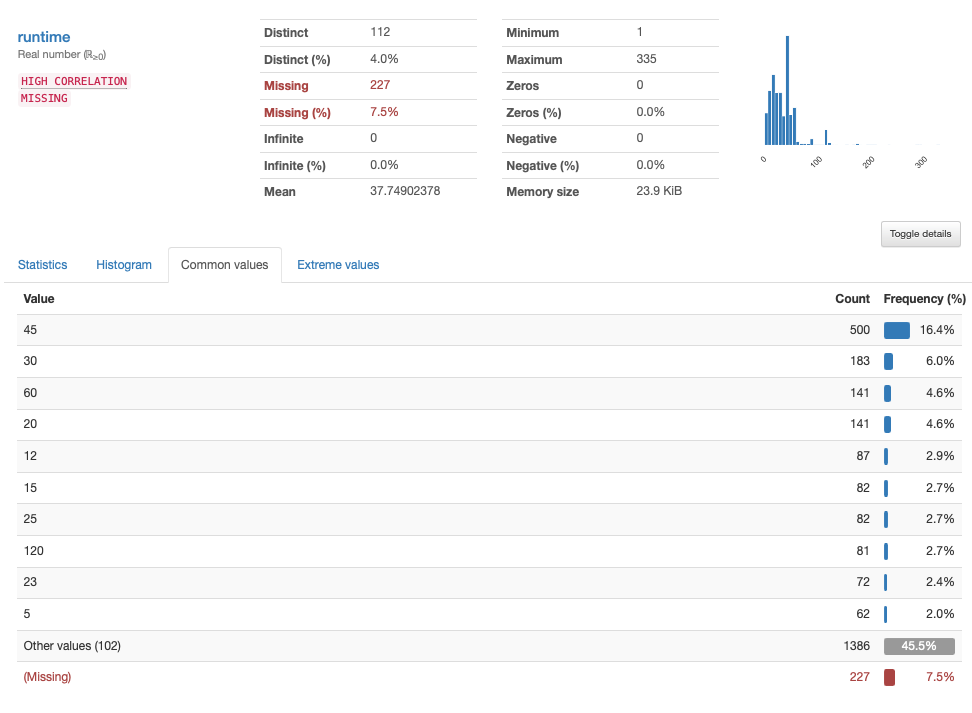


En el histograma es interesante ver para este mes se estrenan en su mayoría los primeros episodios de muchas series.

Dentro de los primeros 10 valores se encuentra la cantidad de episodios que se respalda en el histograma viendo que en diciembre muchos programas sacaron sus primeros episodios, tal vez se deba a la temporada de navidad en las que las familias se unen o a las vacaciones de colegios e instituciones educativas, lo que hace que las personas tengan más tiempo disponible para ver series, para validar si esta hipótesis es cierta se tendría que contrastar la data con la de otros meses.

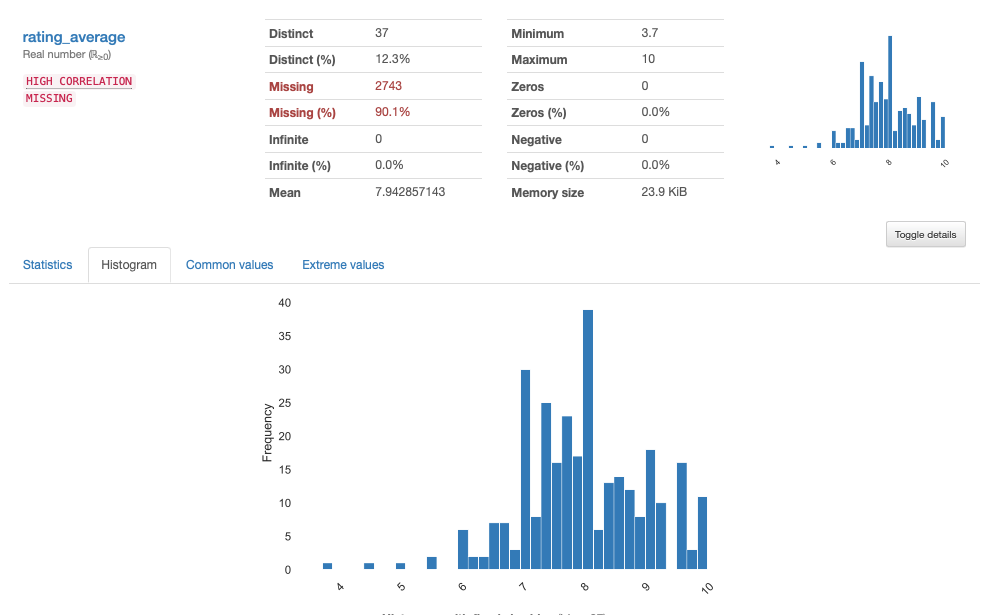


#### Runtime

En esta variable podemos ver que la frecuencia más alta de duración de una serie es de 45 min y el promedio es de 37.74 min, es un tiempo coherente ya que series que no es muy común que los episodios de las series duren más de 1 hora.

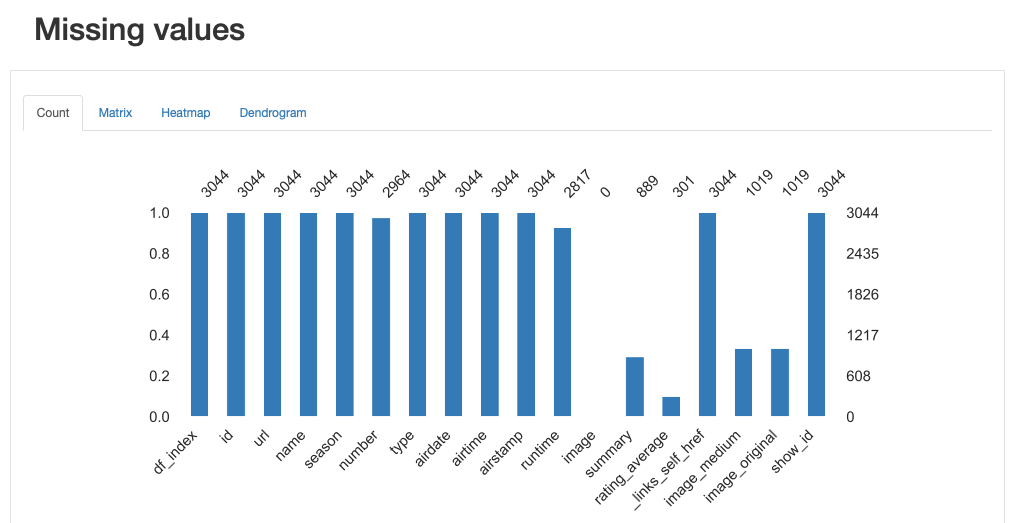
#### Rating\_average

Esta variable nos dice que la media es de 7.94 y que los valores más altos están entre 7 y 8, los resultados que nos arroja no son muy confiables ya que el 90.1% de los campos están vacíos, esto se puede deber a inconvenientes en el flujo de los datos en la plataforma.



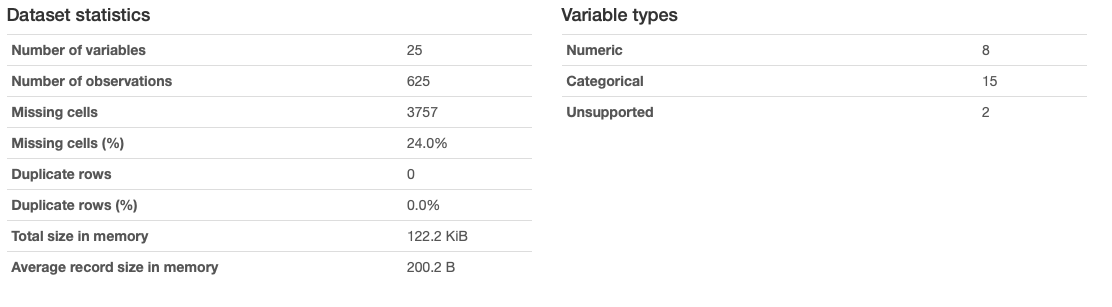
### Valores nulos

En este apartado podemos encontrar que las variables summary que nos da el resumen del episodio y rating average que nos da el rating promedio no son muy confiables por su falta de datos por lo cual los análisis realizados con estas variables podrían verse altamente sesgados



## Dataframe Shows

En este dataframe podremos encontrar la información general de los shows o series transmitidos en el mes de diciembre del 2020, podemos evidenciar que se transmitieron en total 625 series y de cada una de estas se capturaron 25 variables, al totalizar la información evidenciamos que el 24% de la misma, está vacía, más adelante veremos cuales han sido las variables más afectadas.



### Variables

A continuación una descripción de cada una de las variables del dataframe:

* show\_id: Id Unico del show.
* url: Link a la información del show.
* name: Nombre del show.
* Type: Tipo de serie.
* Language: Lenguaje original de la serie.
* Genres: Géneros de la serie.
* Status: Estado de la serie en el periodo de la proyección del episodio.
* Runtime: Tiempo de duración de cada episodio del show.
* averageRuntime: Tiempo de duración promedio del show.
* Premiered: Fecha de estreno del Show.
* Ended: Fecha del final del Show.
* officialSite: url del Show.
* schedule\_time: Hora a la que se proyecta el Show.
* schedule\_days: Días en los que se proyecta el Show.
* rating\_average: Rating promedio del Show.
* Weight: Peso del Show.
* Network: Campo vacio.
* Summary: Resumen del Show.
* Updated: Última actualización del show.
* Image: Imagen del Show.
* links\_self\_href: Link a la información del episodio.
* links\_previousepisode\_href: Link a API para información de un episodio.
* links\_nextepisode\_href: Link a API para información de un episodio.
* webChannel\_id: Id del web channel donde se transmite.
* network\_id: id de la cadena donde se transmite.

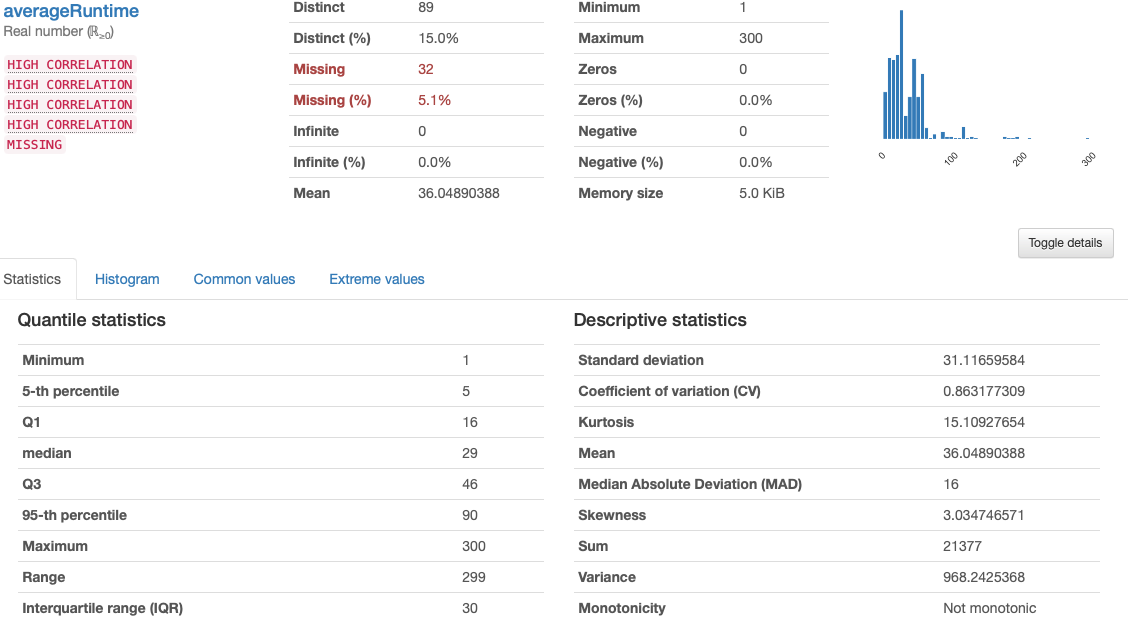
Para el caso de las variables **links\_previousepisode\_href** y **links\_nextepisode\_href** los links no nos llevan al episodio anterior ni al siguiente episodio a pesar de que así lo indica el nombre de la misma, al validar lleva a información de otros episodios random.

Igual que en el dataframe Chapters hay muchas variables informativas que no aportan insights significativos, por tal motivo he decidido analizar las que muestran hallazgos que se podrían considerar importantes

#### Average\_runtime

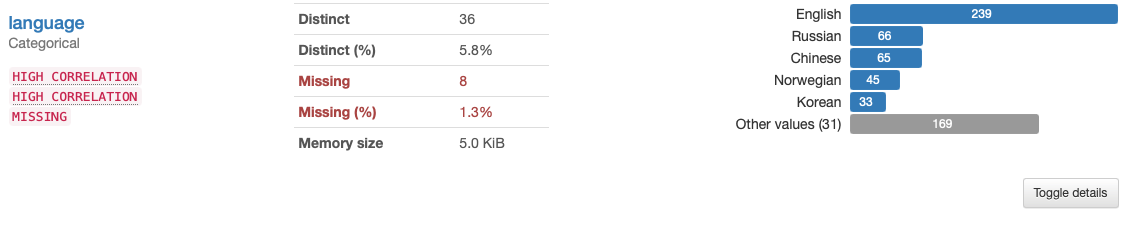
Tenemos una megia general de 36.04 en este caso tenemos que la mayoria de shows tiene una duración por episodio de 30 a 40 min. Varia un poco con la información por capitulo ya que algunos shows pueden tener más o menos episodios que otros.





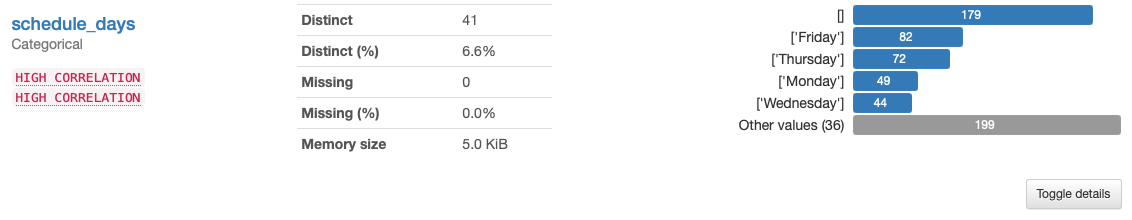
#### Language

Podemos evidenciar que la mayoría de los shows son en lenguaje inglés, lo cual no es sorpresivo ya que es la primera economía del mundo, lo interesante es ver la participación de series rusas y chinas y cómo estás tienen valores significativos. Pareciera que fueran unos mercados emergentes en este sector económico.



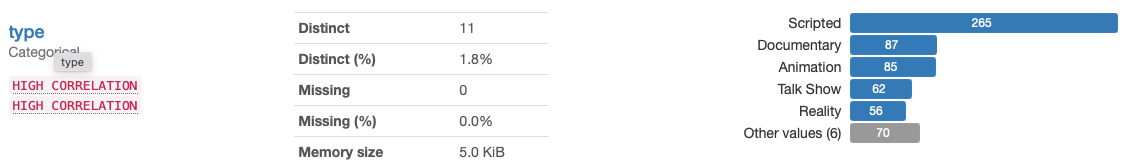
#### Schedule\_days

Para esta variable hay un valor significativo de espacios en blanco, aún así se puede evidenciar una tendencia a lanzar los Show los días Viernes y Jueves, quizá esto se pueda asociar a la temporada de festividades pero para validar esto se tendría que comparar la información con otros meses



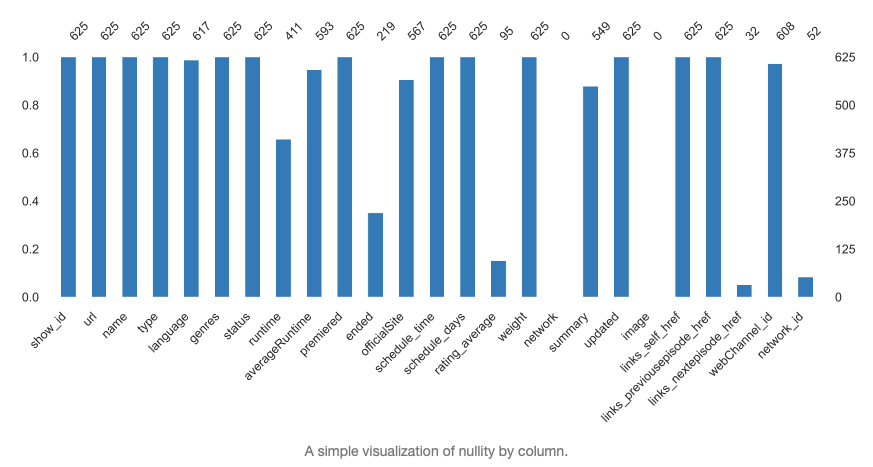
#### Type

En esta variable se puede ver que la mayoría de las series proyectadas son guionizadas o actuadas, en segundo puesto están los documentales y seguido las animaciones. Seria interesante saber si esto tiene que ver con el mes o si este es el comportamiento general.



### Valores nulos

En este caso podemos evidenciar que las variables network y image son descartadas ya que no incluyen datos, por otro lado Rating\_average es una de las más incompletas así que los análisis realizados a esta variable son poco confiables



## Otros Dataframes

El resto de Dataframes son tablas que se crearon para las relaciones del modelo de datos con variables informativas, sus respectivos archivos profiling se encuentran adjuntos en la carpeta, a continuación se describe cada uno de estos

* Countries: Dataframe de países en donde se exponen las series.
* DVD: información de los dvd’s de algunas series
* External: Información categórica de otras fuentes (Ej:imdb)
* Image: links con imágenes de los Shows
* Network: Información de las cadenas donde se lanzan los shows.
* Web\_Channel: Nombre de las plataformas web donde se transmiten los Shows.