



## RETO 2. Julián Alberto Uribe Gómez. ITM Instituto Tecnológico Metropolitano. Medellín. Equipo ACE

¿Qué perspectivas e inferencias se pueden hacer sobre los contenidos de TV pública, de forma que ayuden a tomar mejores decisiones en sus contenidos y franjas, de acuerdo con sus variables de interés?

### Tabla de contenido

<b>Conjunto de datos</b> .....	<b>1</b>
<b>Herramienta utilizada</b> .....	<b>1</b>
<b>Procedimiento</b> .....	<b>2</b>
<b>Análisis y propuesta de solución del reto</b> .....	<b>3</b>
Duración Total .....	3
Categorización.....	4
Horarios .....	5
Tendencias .....	5
Franjas .....	7
Población sorda .....	7
<b>Conclusión</b> .....	<b>7</b>

### Conjunto de datos

Para la propuesta, se utilizó el conjunto de datos: Parrilla de programación de televisión abierta.

#### Parrillas de programación televisión abierta

<https://www.postdata.gov.co/dataset/parrillas-de-programación-televisión-abierta>  
<https://www.postdata.gov.co/dataset/parrillas-de-programación-televisión-abierta/resource/1895fbee-42f5-41b5-bba7-a2d03cb0f723#{}>

### Herramienta utilizada

Se utilizó para el análisis de los datos *Python*, *Jupyter Notebook* y *Excel*  
Las librerías de *Python* utilizadas fueron:

- `import pandas as pd`
- `import numpy as np`
- `import matplotlib.pyplot as plt`



- `import seaborn as sb`
- `import plotly.express as px`
- `from datetime import datetime`
- `plt.rcParams['figure.figsize'] = (16, 9)`
- `plt.style.use('ggplot')`

## Procedimiento

1. Descargar localmente el archivo de datos: parrilla de programación televisión abierta.
2. Lectura jupyter archivo parrilla. El archivo utilizado es:
  - `Parrilla_programacion.csv`
3. Separar del conjunto de datos anterior las variables [GENERO y TIPO].
4. Limpiar el nuevo conjunto de datos localmente y generar 4 variables nuevas de agrupación [GENERO1, TIPO1, AGRUPACION1, AGRUPACION2]. El archivo utilizado es:
  - `Dfgenerotipo.csv`
5. Concatenar el conjunto de datos inicial con nuevo conjunto de datos creado. Los archivos concatenados son:
  - `Parrilla_programacion.csv`
  - `Dfgenerotipo.csv`
6. Atomizar variable FECHA y obtener nuevas variables como [mes, día, semana, dia\_semana, nombre\_dia, nombre\_mes].
7. Atomizar variable HORA\_INICIO.
8. Atomizar variable DURACION y transformar en int64 numérica continua.

$$\begin{aligned} & Duracion\ total\ en\ minutos \\ &= (hora\ duracion) * 60 + (minuto\ duracion) \\ &+ \frac{segundo\ duracion}{60} \end{aligned}$$

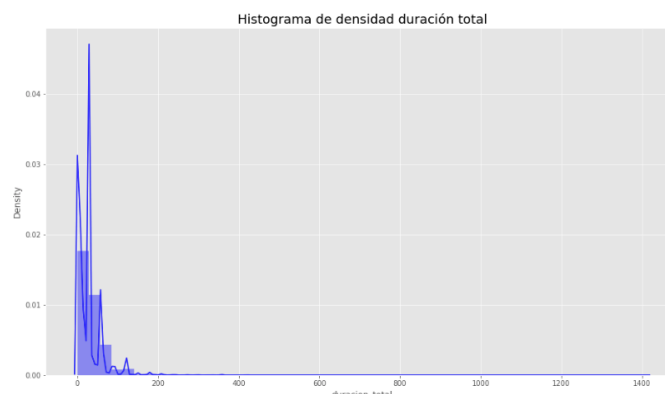
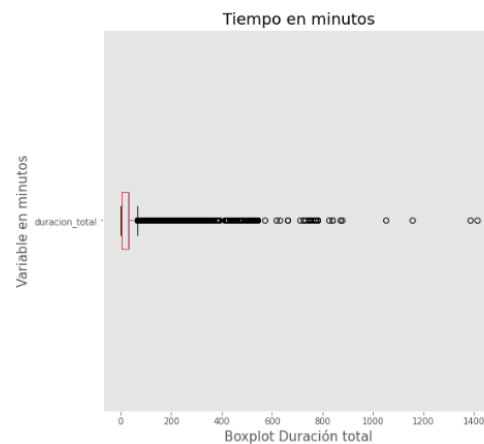
9. Realizar análisis sobre la duracion\_total y días y meses.
10. Realizar análisis de categorización.
11. Realizar análisis de horarios y franjas.
12. Realizar análisis de tendencias.
13. Realizar análisis programas subtítulos.

## Análisis y propuesta de solución del reto

### Duración Total

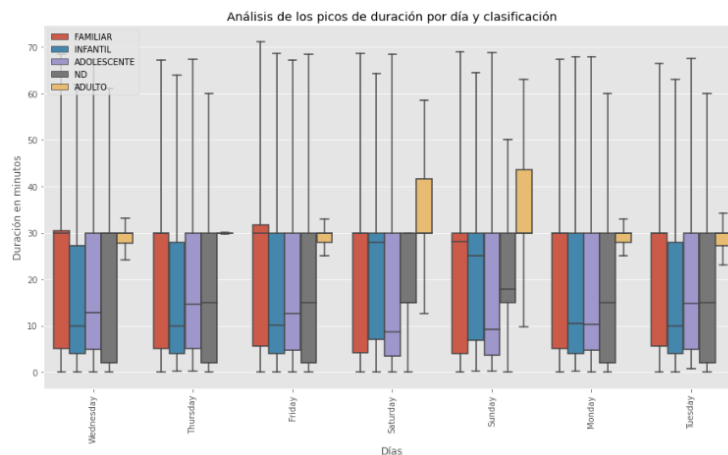
cuenta	687803.0
promedio	29.084137
desviación estándar	34.628287
mínimo	0
25%	5
50%	28.466667
75%	30
máximo	1412.083333

En promedio los programas duran 29 minutos, sin embargo se tiene una desviación superior a la media, esto indica que los datos de duración se encuentran fuera de control estadístico al ser el coeficiente de variación mayor al 20%, adicional se encuentra un programa con un máximo de 1412 minutos, y en general el 75% de los programas tienen 30 minutos de duración.

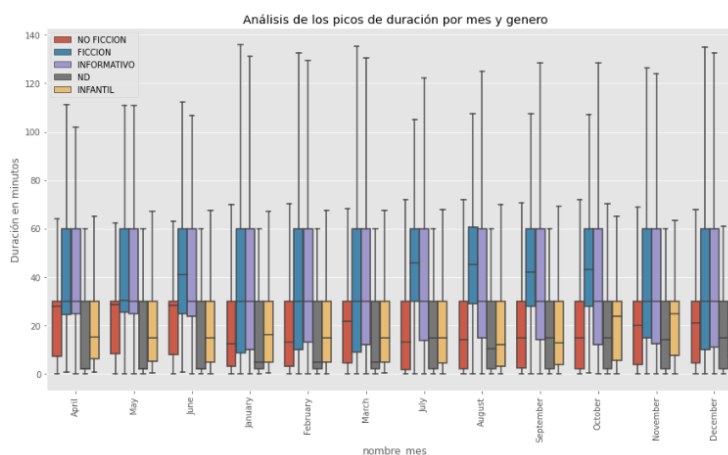


Al analizar el boxplot y el diagrama de distribución sobre la variable duración, estos indican que la distribución está sesgada positivamente hacia la derecha, lo

cual corresponde a ambas gráficas vistas anteriormente. Por otro lado se encuentran una gran cantidad de datos de duración fuera de la caja, algunos de ellos considerados máximos o valores atípicos de la distribución.



Un análisis estadístico descriptivo sobre la duración de los programas por clasificación y día muestra que, en general los fines de semana los programas clasificados como "adultos" tienen una duración promedio mayor que en semana. Se encuentra una clasificación "ND" con una fuerte representatividad durante toda la semana. En general, para todas las otras categorías durante la semana el 75% de los programas duran en 30 minutos.



Al analizar la duración de los programas por GENERO1 (variable limpia obtenida del conjunto de datos separado del original) y mes muestra que, en general todos los géneros han tenido comportamientos similares en todos los meses. Se encuentra que los programas de ficción e informativos tienen duraciones en un cuartil 75 de 60 minutos, donde en todos los casos tienen valores máximos que llegan hasta más de 130 minutos.

## Categorización

De acuerdo a la Clasificación por operador, se generan mayor cantidad de contenidos familiares en TV pública regional abierta y TV sin ánimo de lucro local abierta. Por otro lado, de acuerdo al genero por operador, se halla mayor cantidad de contenidos de no ficción en TV pública regional abierta y mayor contenido informativo en TV sin ánimo de lucro local abierta.

## Horarios

De acuerdo con todos los horarios empleados desde las 0 horas a 23 horas, la TV pública regional abierta y TV sin ánimo de lucro local abierta tienen mayor participación en toda la franja horaria, donde prevalece un mayor numero de programas Familiares, de No ficción e informativos.

Se aplico un código condicional sobre la hora\_inicio donde se clasifican los horarios en las siguientes franjas:

- 0 horas a 6 am (madrugada)
- 6 horas a 12 m (mañana)
- 12 m a 18 horas (tarde)
- 18 horas a 0 horas (noche)

De acuerdo con las franjas horarias generadas mediante la condición, la TV pública regional abierta y TV sin ánimo de lucro local abierta tienen mayor participación en toda la franja horaria, donde prevalece un mayor numero de programas Familiares, de No ficción en franjas de la madrugada.

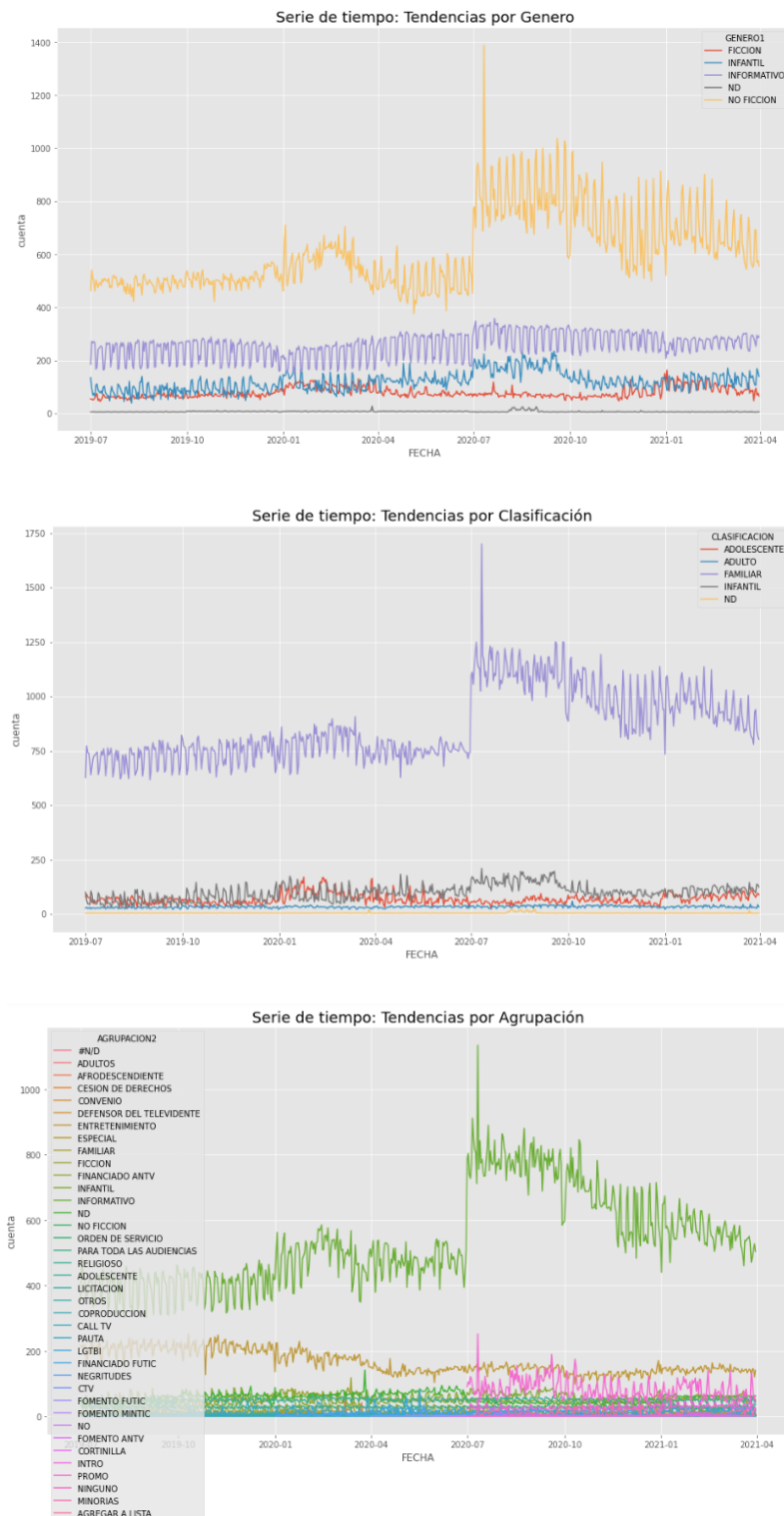
## Tendencias

En este punto se utilizo el conjunto de datos df4 generado del df3, que contiene las franjas horarias en la variable Hora\_inicio.

Adicional, en este punto se utiliza la variable única generada como respuesta a las variables [GENERO, TIPO] de este reto, la cual es AGRUPACION2, se explica que la variable GENERO esta incluida en la variable TIPO, por otro lado la variable TIPO tiene celdas vacías que son reemplazadas por los datos de GENERO. Un análisis sobre estas variables reflejan que se pueden generar 3 categorías de agrupación:

- informativo: [entrevistas, reportajes, documentales, opinión, tertulia, actualidad, culturales, políticos y deportivos]
- ficción: [series, miniseries, telenovelas, movies, cines]
- entretenimiento: [magazin, galas, concursos, reality]

No obstante se incluyen otras categorías que por su cantidad e importancia se dejaron y no se agruparon en estas categorías mencionadas.



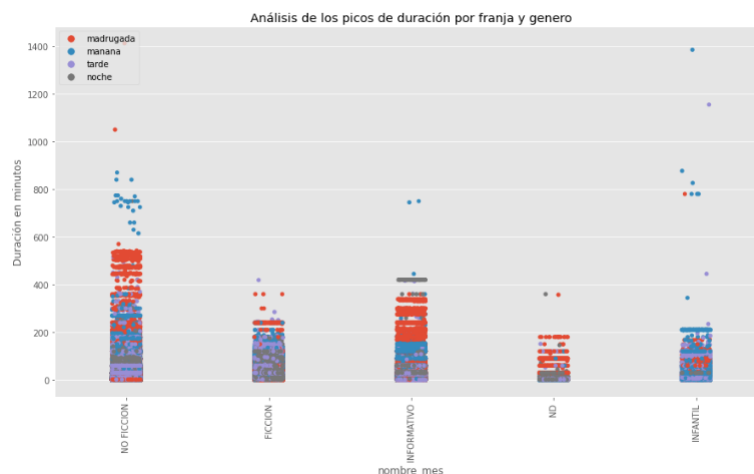
Existe un comportamiento mayor y mas amplio de programas de No ficción durante todo el tiempo de estudio. Los programas de ficción son menos representativos y sin incrementos significativos durante toda la serie de tiempo. El comportamiento temporal de la clasificación familiar es mucho más alto que otras clasificaciones. El contenido adulto permanece constante y sin muchos programas, mientras que los contenidos infantiles tienen alzas diferenciadas de los contenidos para adolescentes, sin embargo siguen siendo bajos.

De acuerdo a la variable [AGRUPACION2] generada como variable respuesta de la unión de GÉNERO y TIPO, al igual que las anteriores series, los datos más representativos son No Ficción y Familiar.

## Franjas

En este punto se utilizó el conjunto de datos df4 generado del df3, que contiene las franjas horarias en la variable Hora\_inicio.

En general, el análisis muestra que durante todos los días se presentan similitudes en la cantidad de programas por franja horaria, excepto leves cambios, no muy significativos. Las franjas por mes, detectaron menos cantidad de programas en las franjas de abril, mayo y junio, esto se debe a la base de datos, ya que estos meses solo cuentan con 1 registro para el año 2020 y ninguno para el año 2019. Al analizar nuevamente los picos de duración por género y franja se encuentra el dato atípico de 1400 minutos en el género infantil en la franja de la mañana, este género tiene diversos datos atípicos en su análisis.



## Población sorda

En este punto se utilizó el conjunto de datos df4 generado del df3, que contiene las franjas horarias en la variable Hora\_inicio.

Para las variables del conjunto de datos analizado: Closed\_Caption, Lengua\_senas y Subtitulado, se observa que existe la transcripción manual y automática en clasificaciones familiares, y en géneros de no ficción. En lengua de señas hay una cantidad muy pequeña de programas que cuentan con ella, en general solo familiares en proporción pequeña, al igual que por géneros donde no hay mayor representatividad. Adicional hay una proporción muy pequeña de programas subtitulados en géneros informativos familiares.

## Conclusión

La televisión pública es una importante manera de acercar a las personas a una comunicación sana y plural, sin matices y con total libertad e independencia, es por esta razón que conocer sus actores, contenidos, franjas, duración y demás resulta ser un ejercicio necesario para planear esfuerzos y mejorar el sistema y contenidos de la televisión pública. En general la mayoría de los programas son familiares, informativos y de no ficción, por otro lado la TV pública regional abierta y TV sin ánimo de lucro local abierta son los grandes actores del proceso de entrega de programas de calidad, con gran participación y expansión en toda la franja horaria.

Como punto negativo, se encontró una gran cantidad de datos atípicos en la duración de los programas, dando a entender que aunque existe un estándar en el tiempo de programación, muchos programas tienen tiempos de duración por fuera de la norma.

Se encontró muy poca participación de lenguaje de señas y programas subtítulos en la mayoría de programas, géneros, tipos y clasificaciones, lo que indica que se debe fortalecer este aspecto para la inclusión.

Se adicionan archivos de datos que se revisaron de forma local para la consolidación de una variable que agrupe GENERO y TIPO y que se utilizó en los análisis presentados.