

## Taller 1 Análisis de datos

# Análisis de la dataset de Game of Thrones

**Integrantes:** Matías Ignacio Jara Vargas y Paola Alejandra Rico Merlano

**Curso y profesor:** Análisis de datos para Ingeniería Informática, Prof. Eliecer Peña

**Fecha de entrega:** 13-09-2024

## Contenido

Análisis de la dataset de.....	1
Game of Thrones.....	1
Resumen Ejecutivo .....	1
1. Introducción .....	2
2. Metodología .....	3
3. Análisis de Datos .....	4
4. Discusión .....	17
5. Conclusiones .....	19
6. Referencias.....	20
Bibliografía .....	20

**Link Google colab:** [https://colab.research.google.com/drive/1WBP-xZeLsEa\\_zQWNpC3KOx6iCBskCnHX?usp=sharing](https://colab.research.google.com/drive/1WBP-xZeLsEa_zQWNpC3KOx6iCBskCnHX?usp=sharing)



## Resumen Ejecutivo

Tenemos como objetivo analizar el dataset que nos han proporcionado para aplicar los conceptos vistos en clases, aplicar técnicas de manipulación y usar herramientas que nos permitan realizar el análisis de manera eficiente y correcta, en este caso vamos a utilizar las herramientas power bi y Google collab para analizar profundamente los datasets relacionados con la serie Game of Thrones. Se realizó la fusión de los archivos usando power bi, el tratamiento de valores faltantes, el análisis estadístico descriptivo, la imputación de datos, la detección y tratamiento de outliers, y la visualización de datos. Todo ello con el fin de obtener una mejor comprensión de los datos y presentar los resultados de manera clara y significativa para poder sacar nuestras conclusiones. Usaremos metodologías tanto de investigación como las que mayormente se usan en un proyecto.

Los principales hallazgos que se observan son datos faltantes a la hora de combinar los datasets y los formatos de fechas son diferentes, esto nos puede ocasionar problemas ya que nos podrían dar resultados incoherentes que no aporten valor a nuestros análisis. El análisis permitió identificar tendencias importantes en la calificación de los episodios y la audiencia a lo largo de las temporadas. La imputación de datos y el tratamiento de outliers mejoraron la calidad del dataset, permitiendo realizar un análisis más preciso. Las visualizaciones destacaron las diferencias en la recepción crítica y el seguimiento de la serie a lo largo del tiempo, proporcionando una visión clara de la evolución de Game of Thrones desde su inicio hasta su última temporada.

## 1. Introducción

Se nos presentó un dataset dividido en dos partes, estos datos están relacionados con la serie Game of Thrones (Juego de tronos), es una serie de televisión estadounidense de género dramático y fantástico, en este dataset se muestran datos como la calificación de episodios, los espectadores (por millones), el numero de episodios, las temporadas, entre otras. En el desarrollo de este análisis tiene como objetivo principal el tratamiento de los outliers, los datos faltantes y el manejo de los datos que conforma el dataset, siguiendo el análisis critico y las buenas practicas presentadas en clases.

Los objetivos del estudio son unificar los dos archivos proporcionados utilizando una clave primaria común para obtener un conjunto de datos consolidado y coherente. Se busca calcular medidas descriptivas clave, como la media, mediana, moda, varianza y desviación estándar, para entender la distribución de los datos numéricos. Otro objetivo sería identificar y tratar los valores nulos en las columnas mediante técnicas de imputación adecuadas, como el uso de la media para datos numéricos y la moda para los categóricos. También se pretende detectar valores atípicos, especialmente en la calificación IMDb, y aplicar estrategias de manejo de outliers para mejorar la calidad del análisis. A su vez, se generarán visualizaciones claras y concisas, como gráficos de línea y gráficos de torta, para facilitar la interpretación de las tendencias y patrones en los datos, con especial énfasis en la calificación de los episodios y la audiencia a lo largo de las temporadas. Finalmente, se busca extraer conclusiones relevantes sobre la evolución de la serie Game of Thrones y proporcionar recomendaciones basadas en el análisis realizado.

## 2. Metodología

El dataset utilizado proviene de dos archivos relacionados con la serie Game of Thrones, los cuales contienen información detallada sobre los episodios emitidos a lo largo de sus temporadas. Estos archivos incluyen datos sobre aspectos como el número del episodio, el título, los escritores, directores, fecha de emisión, calificación en IMDb, y la audiencia en millones de espectadores en Estados Unidos. Las columnas principales incluyen tanto variables numéricas como categóricas, permitiendo un análisis integral de la recepción de la serie.

En el estudio, se empleó una variedad de métodos estadísticos para analizar los datos a fondo y obtener información significativa. Se aplicaron estadísticas descriptivas como promedio, valor medio, valor más común, dispersión y desviación típica a los datos numéricos para resumir y comprender su distribución. Las medidas seleccionadas se implementaron para ofrecer un resumen completo de los atributos fundamentales y la distribución de los datos. La puntuación media y mediana muestran el valor habitual, pero el diferencial y cuánto varía con respecto al promedio son claves para determinar qué tan diferente es las clasificaciones de IMDb y cuántas personas las ven. De esta manera, mantenemos el patrón general sin estropear demasiado, se trata de usar gráficos de diagrama de caja para detectar y corregir puntos de datos extraños que podrían arruinar nuestro análisis.

Por último, se usaron técnicas de visualización como gráficos de línea y gráficos de torta para mostrar de manera clara y accesible las tendencias y distribuciones en los datos, lo cual facilita la interpretación de los resultados estadísticos.

### 3. Análisis de Datos

- **Descripción del dataset**

Para iniciar con la importación de los datos importamos los libros de Excel y usamos primero una herramienta llamada power bi, esta nos permite analizar datos de manera profunda, sin la necesidad de utilizar código y también usaremos Google colab para hacer unos puntos específicos del análisis. Una vez que entramos a power bi obtuvimos los datos. Y está estructurado en varias columnas que contienen información detallada sobre los episodios de la serie. Podemos encontrar: `primary_key`, que actúa como identificador único de los episodios; `title`, que indica el título del episodio; `season` y `No. in season`, que detallan la temporada y el número del episodio respectivamente; `Imdb rating`, que contiene la calificación de IMDb en formato numérico; y `U.S. viewers (millions)`, que registra el número de espectadores en EE.UU. en millones. También se incluyen columnas categóricas como `Written by` y `Directed by`, que indican los guionistas y directores de cada episodio. Los tipos de datos en el conjunto incluyen tanto valores numéricos, como la calificación de IMDb y los espectadores, así como valores de texto para nombres, títulos y fechas de emisión. Estos datos permiten un análisis integral de la recepción y el rendimiento de los episodios a lo largo de la serie.

- **Fusión de Archivos**

Pasamos a fusionar los datasets con la opción de transformar datos y los combinamos usando la columna `primary_key` como la clave para la unión. Después de eso exploramos los datos para entender cada una de sus columnas y se usó la unión `inner join` ya que este tipo de unión excluye cualquier fila que tenga datos incompletos en uno de los dos datasets. Esto es esencial en el análisis de datos de series como *Game of Thrones*, donde cada episodio debe estar representado por completo, con la información de ambas fuentes.

## parte\_1

Directed by

Escrito Por

Σ Espectadores(Millones)

📅 Fecha Origen

Σ Imdb rating

Σ Num Temporada

parte\_2.Novel(s) adapted

primary\_key

Σ Temporada

Título

primary_key	Num Temporada	Temporada	Título	Escrito Por	Fecha Origen	parte_2.Novel(s) adapted	Espectadores(Millones)	Imdb rating	Directed by
1	1	1	"Winter Is Coming"	David Benioff & D. B. Weiss	domingo, 17 de abril de 2011	A Game of Thrones	3.22	9.1	Tim Van Patten
2	2	1	"The Kingsroad"	David Benioff & D. B. Weiss	domingo, 24 de abril de 2011	A Game of Thrones	2.2	8.8	Tim Van Patten
3	3	1	"Lord Snow"	David Benioff & D. B. Weiss	domingo, 1 de mayo de 2011	A Game of Thrones	2.44	8.7	Brian Kirk
4	4	1	"Cripples, Bastards, and Broken Things"	Bryan Cogman	domingo, 8 de mayo de 2011	A Game of Thrones	2.43	8.8	Brian Kirk
5	5	1	"The Wolf and the Lamb"	David Benioff & D. B. Weiss	domingo, 15 de mayo de 2011	A Game of Thrones	2.58	9.1	Brian Kirk
6	6	1	"A Golden Crown"	David Benioff & D. B. Weiss	domingo, 22 de mayo de 2011	A Game of Thrones		9.2	Daniel Minahan
7	7	1	"You Win or You Die"	David Benioff & D. B. Weiss		A Game of Thrones	2.4		Daniel Minahan
8	8	1	"The Pointy End"	George R. R. Martin	martes, 5 de junio de 2012	A Game of Thrones	2.72	9	Daniel Minahan
9	9	1	"Baelor"	David Benioff & D. B. Weiss	martes, 12 de junio de 2012	A Game of Thrones	2.66		Alan Taylor
10	10	1	"Fire and Blood"		martes, 19 de junio de 2012	A Game of Thrones	3.04	9.5	Alan Taylor
11	1	2	"The North Remembers"	David Benioff & D. B. Weiss	domingo, 1 de abril de 2012	A Clash of Kings	3.86	8.8	Alan Taylor
12	2	2	"The Night Lands"	David Benioff & D. B. Weiss		A Clash of Kings	3.78	8.3	Alan Taylor
13	3	2	"What Is Dead May Never Die"	Bryan Cogman		A Clash of Kings	3.77	8.8	Alek Safarian
14	4	2	"Garden of Bones"	Veronica Taylor	domingo, 22 de abril de 2012	A Clash of Kings	3.65	8.8	David Petrarca
15	5	2	"The Ghost of Harrenhal"	David Benioff & D. B. Weiss	domingo, 29 de abril de 2012	A Clash of Kings		8.8	David Petrarca
16	6	2	"The Old Gods and the New"	Veronica Taylor	domingo, 6 de mayo de 2012	A Clash of Kings	3.88	9.1	David Nutter
17	7	2	"A Man Without Honor"	David Benioff & D. B. Weiss		A Clash of Kings	3.69	8.8	David Nutter
18	8	2	"The Prince of Winterfell"	David Benioff & D. B. Weiss	domingo, 20 de mayo de 2012	A Clash of Kings		8.8	Alan Taylor
19	9	2	"Blackwater"	George R. R. Martin	domingo, 27 de mayo de 2012	A Clash of Kings	3.38	9.7	Ned Marshall
20	10	2	"Valar Morghulis"	David Benioff & D. B. Weiss	domingo, 3 de junio de 2012	A Clash of Kings	4.2	9.4	Alan Taylor
21	1	3	"Valar Dohaeris"	David Benioff & D. B. Weiss	domingo, 21 de marzo de 2013	A Storm of Swords	4.37	8.8	Daniel Minahan
22	2	3	"Dark Wings, Dark Words"	Veronica Taylor	domingo, 7 de abril de 2013	A Storm of Swords	4.27	8.8	Daniel Minahan
23	3	3	"Walk of Punishment"	David Benioff & D. B. Weiss	domingo, 14 de abril de 2013	A Storm of Swords	4.72	8.9	David Benioff
24	4	3	"And Now His Watch Is Ended"	David Benioff & D. B. Weiss	domingo, 21 de abril de 2013	A Storm of Swords	4.87	9.6	Alex Graves
25	5	3	"Kissed by Fire"		domingo, 28 de abril de 2013	A Storm of Swords	3.33	9	Alex Graves
26	6	3	"The Climb"	David Benioff & D. B. Weiss	domingo, 5 de mayo de 2013	A Storm of Swords	5.5		Alek Safarian
27	7	3	"The Bear and the Maiden Fair"	George R. R. Martin		A Storm of Swords		8.7	Michelle MacLaren
28	8	3	"Second Sons"	David Benioff & D. B. Weiss	domingo, 19 de mayo de 2013	A Storm of Swords	5.13	9	Michelle MacLaren
29	9	3	"The Rains of Castamere"	David Benioff & D. B. Weiss	domingo, 2 de junio de 2013	A Storm of Swords	5.22	9.9	David Nutter
30	10	3	"Mhysa"	David Benioff & D. B. Weiss	domingo, 9 de junio de 2013	A Storm of Swords	5.38	9.2	David Nutter
31	1	4	"Two Swords"	David Benioff & D. B. Weiss	domingo, 6 de abril de 2014	A Storm of Swords	6.64	9.1	D. B. Weiss
32	2	4	"The Lion and the Rose"		domingo, 13 de abril de 2014	A Storm of Swords	6.31	8.7	Alex Graves
33	3	4	"Breaker of Chains"	David Benioff & D. B. Weiss	domingo, 20 de abril de 2014	A Storm of Swords	6.59		Alex Graves
34	4	4	"Lathes and Paper"	Bryan Cogman	domingo, 27 de abril de 2014	A Storm of Swords	6.95	8.8	Michelle MacLaren
35	5	4	"First of His Name"		domingo, 4 de mayo de 2014	A Storm of Swords	7.16	8.8	Michelle MacLaren
36	6	4	"The Laws of Gods and Men"	Bryan Cogman	domingo, 11 de mayo de 2014	A Storm of Swords	6.4	8.7	Alek Safarian
37	7	4	"Mockingbird"	David Benioff & D. B. Weiss		A Storm of Swords	7.2	9.1	Alek Safarian

Imágenes 1 y 2. Fusión de archivos utilizando power bi



ID	No. in season	Temporada	Title
0	1	1	"Winter Is Coming"
1	40	10	"The Children"
2	51	1	"The Red Woman"
3	2	2	"The Kingsroad"
4	56	6	"Blood of My Blood"

	Written by	Original air date
0	David Benioff & D. B. Weiss	17-Apr-11
1	David Benioff & D. B. Weiss	2014-06-15 00:00:00
2	David Benioff & D. B. Weiss	24-Apr-16
3	David Benioff & D. B. Weiss	24-Apr-11
4	Bryan Cogman	2016-05-19 00:00:00

	Novel(s) adapted	espectadores	rating
0	A Game of Thrones	2.22	9.1
1	A Storm of Swords	7.09	9.7
2	Outline from The Winds of Winter and original ...	7.94	8.5
3	A Game of Thrones	2.20	8.8
4	Outline from The Winds of Winter and original ...	NaN	8.4

	Directed by
0	Tim Van Patten
1	Alex Graves
2	Jeremy Podeswa
3	Tim Van Patten
4	Jack Bender

Imágenes 3. Fusión de archivos utilizando Google Colab

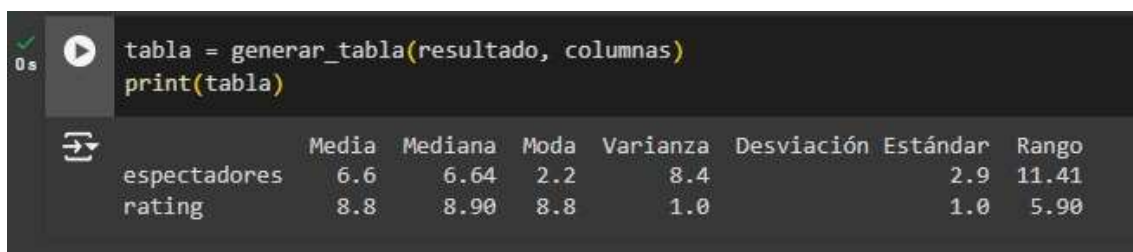
- **Estadísticos Descriptivos:**

En esta fase del análisis, se calculan los estadísticos descriptivos de las columnas numéricas, lo cual incluye la media, mediana, moda, varianza, desviación estándar y el rango. Estos cálculos permiten resumir la información clave de los datos, proporcionando una visión clara de su distribución. La media nos indica el valor promedio de la columna, mientras que la mediana muestra el punto central de los datos, lo que es útil para identificar sesgos si los valores extremos afectan la media. La moda revela el valor más frecuente en la columna, y la varianza junto con la desviación estándar nos ofrecen una medida de la dispersión de los datos, ayudándonos a entender qué tan alejados están los valores del promedio. Por último, el rango permite conocer la diferencia entre el valor más alto y el más bajo, indicando la amplitud de los datos.

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)

columnas = ['espectadores', 'rating']
def calcular_estadisticas(datos):
    me = stats.mode(datos.dropna(), nan_policy='omit')
    moda = me[0] if not me.mode.size == 0 else None
    return {
        'Media': round(datos.mean(), 1),
        'Mediana': datos.median(),
        'Moda': moda,
        'Varianza': round(datos.var(), 1),
        'Desviación Estándar': round(datos.std(), 1),
        'Rango': datos.max() - datos.min()
    }
def generar_tabla(resultado, columnas):
    return pd.DataFrame({col: calcular_estadisticas(resultado[col]) for col in columnas}).T
```

Imágenes 4. Bloque de código que llamamos usando Google Colab para realizar los cálculos



```

0s
[play icon] tabla = generar_tabla(resultado, columnas)
print(tabla)

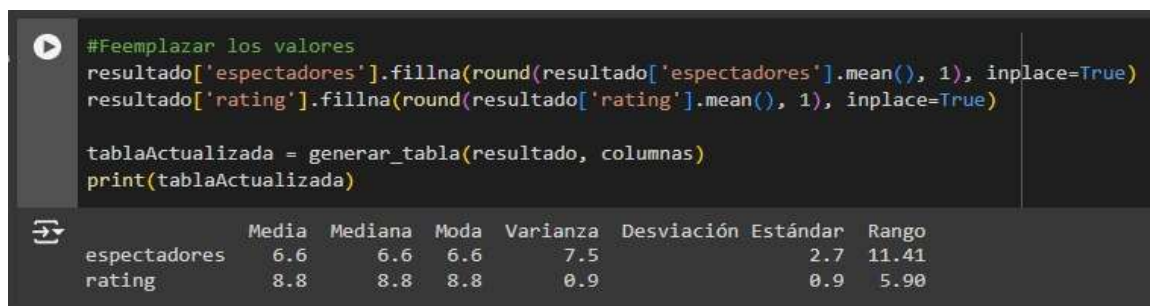
```

	Media	Mediana	Moda	Varianza	Desviación Estándar	Rango
espectadores	6.6	6.64	2.2	8.4	2.9	11.41
rating	8.8	8.90	8.8	1.0	1.0	5.90

Tabla 1. Aquí podemos visualizar los cálculos de la media, mediana, moda, varianza, desviación estándar y rango

- **Imputación de Datos:**

En esta fase nos enfrentamos a un desafío planteado en este dataset, que serian datos faltantes y como trabajar con ellos. Primero identificamos el tipo de valor que tiene la columna, este puede ser numérico (int, float) o categórico (char o string), sabiendo el tipo de datos que tenemos en la columna podemos utilizar diferentes formas de utilizar los datos faltantes. En este caso las columnas de tipo numéricos los nulos o vacíos fueron remplazados por la media aritmética, esto se hace con la finalidad de minimizar el impacto de estos datos faltantes, sin tener la desventaja de perder información (eliminar filas con datos faltantes), por otro lado, para las columnas con datos categóricos los remplazamos por la moda porque un dato categórico no puede tener una media ni una mediana.



```

[play icon] #Reemplazar los valores
resultado['espectadores'].fillna(round(resultado['espectadores'].mean(), 1), inplace=True)
resultado['rating'].fillna(round(resultado['rating'].mean(), 1), inplace=True)

tablaActualizada = generar_tabla(resultado, columnas)
print(tablaActualizada)

```

	Media	Mediana	Moda	Varianza	Desviación Estándar	Rango
espectadores	6.6	6.6	6.6	7.5	2.7	11.41
rating	8.8	8.8	8.8	0.9	0.9	5.90

Tabla 2. Podemos visualizar los cálculos con los valores ya modificados.

```
[7] print(resultado['Written by'])
```

0	David Benioff & D. B. Weiss
1	David Benioff & D. B. Weiss
2	David Benioff & D. B. Weiss
3	David Benioff & D. B. Weiss
4	Bryan Cogman
5	David Benioff & D. B. Weiss
6	NaN
7	NaN
8	Dave Hill
9	NaN
10	David Benioff & D. B. Weiss
11	David Benioff & D. B. Weiss
12	David Benioff & D. B. Weiss
13	NaN
14	David Benioff & D. B. Weiss
15	David Benioff & D. B. Weiss
16	David Benioff & D. B. Weiss
17	David Benioff & D. B. Weiss
18	David Benioff & D. B. Weiss
19	David Benioff & D. B. Weiss
20	Vanessa Taylor
21	Vanessa Taylor

Imagen 3. Datos con valores nulos.

```
moda_written_by = resultado['Written by'].mode()[0] # mode() devuelve una Serie
resultado['Written by'].fillna(moda_written_by, inplace=True)
print(resultado['Written by'])
```

0	David Benioff & D. B. Weiss
1	David Benioff & D. B. Weiss
2	David Benioff & D. B. Weiss
3	David Benioff & D. B. Weiss
4	Bryan Cogman
5	David Benioff & D. B. Weiss
6	David Benioff & D. B. Weiss
7	David Benioff & D. B. Weiss
8	Dave Hill
9	David Benioff & D. B. Weiss
10	David Benioff & D. B. Weiss

Imagen 4. Datos con valores remplazados por la moda.

En esta fase nos podemos preguntar ¿Por qué no sería adecuado reemplazar los valores nulos de la columna "Written by" con la media?, y la respuesta es que la media es adecuada solo para datos numéricos, ya que calcula un promedio que no tiene sentido en datos categóricos como nombres. Los datos categóricos no se pueden promediar, por lo que la moda, que indica el valor más frecuente, es más apropiada para estas columnas. Reemplazar valores nulos con la moda mantiene la coherencia con los datos existentes y evita la pérdida de información significativo.

- **Encoding:**

En este análisis, se aplicó One-Hot Encoding a la columna "Written by", lo que permitió transformar los valores categóricos de los escritores en variables binarias. Esta técnica es útil cuando las categorías no tienen un orden específico, y permite representar cada escritor como una columna independiente, con valores 0 o 1, en nuestro caso es false y true, dependiendo de si el escritor está presente o no en un episodio.

```
# Realizar One-Hot Encoding para la columna "Written by"
resultado_encoded = pd.get_dummies(resultado, columns=['Written by'])
resultado_final = resultado_encoded[['ID']] + [col for col in resultado_encoded.columns if col.startswith('Written by_')]
resultado_final_sorted = resultado_final.sort_values(by='ID')

resultado_final.to_excel('onehotencoding.xlsx', index=False)
```

Imagen 5. Código para realizar One-Hot Encoding.

ID	Written by Bryan Cogman	Written by Dave Hill	Written by David Benioff & D. B. Weiss	Written by George R. R. Martin	Written by Vanessa Taylor
1	FALSO	FALSO	VERDADERO	FALSO	FALSO
2	FALSO	FALSO	VERDADERO	FALSO	FALSO
3	FALSO	FALSO	VERDADERO	FALSO	FALSO
4	VERDADERO	FALSO	FALSO	FALSO	FALSO
5	FALSO	FALSO	VERDADERO	FALSO	FALSO
6	FALSO	FALSO	VERDADERO	FALSO	FALSO
7	FALSO	FALSO	VERDADERO	FALSO	FALSO
8	FALSO	FALSO	FALSO	VERDADERO	FALSO
9	FALSO	FALSO	VERDADERO	FALSO	FALSO
10	FALSO	FALSO	VERDADERO	FALSO	FALSO
11	FALSO	FALSO	VERDADERO	FALSO	FALSO
12	FALSO	FALSO	VERDADERO	FALSO	FALSO
13	VERDADERO	FALSO	FALSO	FALSO	FALSO
14	FALSO	FALSO	FALSO	FALSO	VERDADERO
15	FALSO	FALSO	VERDADERO	FALSO	FALSO
16	FALSO	FALSO	FALSO	FALSO	VERDADERO
17	FALSO	FALSO	VERDADERO	FALSO	FALSO
18	FALSO	FALSO	VERDADERO	FALSO	FALSO
19	FALSO	FALSO	FALSO	VERDADERO	FALSO
20	FALSO	FALSO	VERDADERO	FALSO	FALSO
21	FALSO	FALSO	VERDADERO	FALSO	FALSO
22	FALSO	FALSO	FALSO	FALSO	VERDADERO
23	FALSO	FALSO	VERDADERO	FALSO	FALSO
24	FALSO	FALSO	VERDADERO	FALSO	FALSO
25	FALSO	FALSO	VERDADERO	FALSO	FALSO
26	FALSO	FALSO	VERDADERO	FALSO	FALSO
27	FALSO	FALSO	FALSO	VERDADERO	FALSO
28	FALSO	FALSO	VERDADERO	FALSO	FALSO
29	FALSO	FALSO	VERDADERO	FALSO	FALSO
30	FALSO	FALSO	VERDADERO	FALSO	FALSO
31	FALSO	FALSO	VERDADERO	FALSO	FALSO
32	FALSO	FALSO	VERDADERO	FALSO	FALSO
33	FALSO	FALSO	VERDADERO	FALSO	FALSO
34	VERDADERO	FALSO	FALSO	FALSO	FALSO
35	FALSO	FALSO	VERDADERO	FALSO	FALSO
36	VERDADERO	FALSO	FALSO	FALSO	FALSO
37	FALSO	FALSO	VERDADERO	FALSO	FALSO
38	FALSO	FALSO	VERDADERO	FALSO	FALSO
39	FALSO	FALSO	VERDADERO	FALSO	FALSO
40	FALSO	FALSO	VERDADERO	FALSO	FALSO
41	FALSO	FALSO	VERDADERO	FALSO	FALSO

Tabla 3. Resultado de One-Hot Encoding.

Para la columna "Directed by", se utilizó Label Encoding, que asigna un número entero único a cada director. Esta técnica es más adecuada en este caso, ya que la cantidad de directores es limitada y puede ser manejada de manera más eficiente que con One-Hot Encoding.

```
0 s le = LabelEncoder()
resultado['Directed by Encoded'] = le.fit_transform(resultado['Directed by'])
rf = resultado[['ID', 'Directed by Encoded']]
print(rf)
```

	ID	Directed by Encoded
0	1	19
1	40	1
2	51	12
3	2	19
4	56	11
5	66	0
6	49	9
7	47	17
8	68	9
9	25	1
10	28	16
11	24	1
12	12	0
13	10	0
14	42	15
15	6	5
16	73	8
17	63	13
18	50	9
19	21	5
20	22	5
21	16	9
22	44	13
23	65	14
24	41	15
25	57	13
26	9	0

Imagen 6. Código para realizar Label Encoding y su resultado.

¿Por qué One-Hot Encoding no sería ideal para la columna "Directed by"?: One-Hot Encoding no es una forma optima de procesar los datos porque son demasiadas opciones, en concreto son 20, esto generaría 20 columnas nuevas, aumentando significativamente los datos y de manera innecesaria, en cambio con Label Encoding desacoplamos los directores y les asignamos un numero y nos referimos a este con el numero que se le asigno. Esto disminuye la cantidad de carga que se procesa al cargar un documento y se considera una buena práctica para trabajar columnas con una gran cantidad de opciones.

- **Imputación de Fechas:**

se trató la imputación de valores faltantes en la columna "Original air date", que corresponde a la fecha de emisión original de los episodios. Dado que la coherencia temporal es crucial para mantener el orden cronológico de los

episodios, se optó por rellenar los valores faltantes basándonos en las fechas adyacentes de los episodios previos y posteriores.

```
resultado = resultado.sort_values(by='ID')
fechaformato = '%Y-%m-%d'
resultado['Original air date'] = pd.to_datetime(resultado['Original air date'], format=fechaformato, errors='coerce')

def rellenar_na_con_fecha_mas_cercana(df, columna_fecha):
    df[columna_fecha] = df[columna_fecha].fillna(method='ffill').fillna(method='bfill')
    return df

resultado = rellenar_na_con_fecha_mas_cercana(resultado, 'Original air date')
rx = resultado[['ID', 'Original air date']]

print(rx)
```

Imagen 7. Bloques de código y definiciones correspondientes para trabajar las fechas.

¿Cómo manejaríamos esta situación?: Lo primero que hacemos es ordenar el dataset por nuestra primary key que renombramos como id. Luego, le dimos un formato, después transformamos la columna original air date a tipo de datos fechas con el formato anteriormente designado y los valores faltantes le dimos el valor NaT. El bloque de código rellena los valores faltantes en una columna de fechas usando las fechas más cercanas disponibles. Primero, llena los valores NaT con la fecha más cercana anterior y luego con la más cercana siguiente, asegurando que no queden valores faltantes. Luego se invoca al bloque de código y finalmente para tener una mejor visual creamos un dataframe con la fecha y el id para verificar que siga un orden lógico.

```
print(rx)
```

	ID	Original air date
0	1	2011-05-01
3	2	2011-05-01
67	3	2011-05-01
37	4	2011-05-08
53	5	2011-05-15
15	6	2011-05-22
33	7	2011-05-22
50	8	2012-06-05
26	9	2012-06-12
13	10	2012-06-19
44	11	2012-06-19
12	12	2012-06-19
62	13	2012-06-19
49	14	2012-06-19
30	15	2012-06-19
21	16	2012-05-06
56	17	2012-05-06
32	18	2012-05-28
46	19	2012-05-27
48	20	2012-06-03
19	21	2013-03-31
20	22	2013-03-31
45	23	2013-03-31
11	24	2013-03-31
9	25	2013-03-31
47	26	2013-05-05
42	27	2013-05-05
10	28	2013-05-19
41	29	2013-06-02
69	30	2013-06-09
35	31	2013-06-09
27	32	2013-06-09
64	33	2013-06-09
29	34	2013-06-09
40	35	2014-05-04
71	36	2014-05-11
51	37	2014-05-11
72	38	2014-06-01
54	39	2014-06-08
1	40	2014-06-15
24	41	2014-06-15

Imagen 8. Resultado de las fechas.



- **Tratamiento de Outliers:**

Se utilizó un gráfico de boxplot para visualizar la distribución de las calificaciones de IMDb y detectar outliers en la columna "Imdb rating". Los outliers, o valores atípicos, son aquellos que se encuentran fuera del rango intercuartil definido por el boxplot, lo que puede indicar episodios con calificaciones extremadamente altas o bajas en comparación con la mayoría.

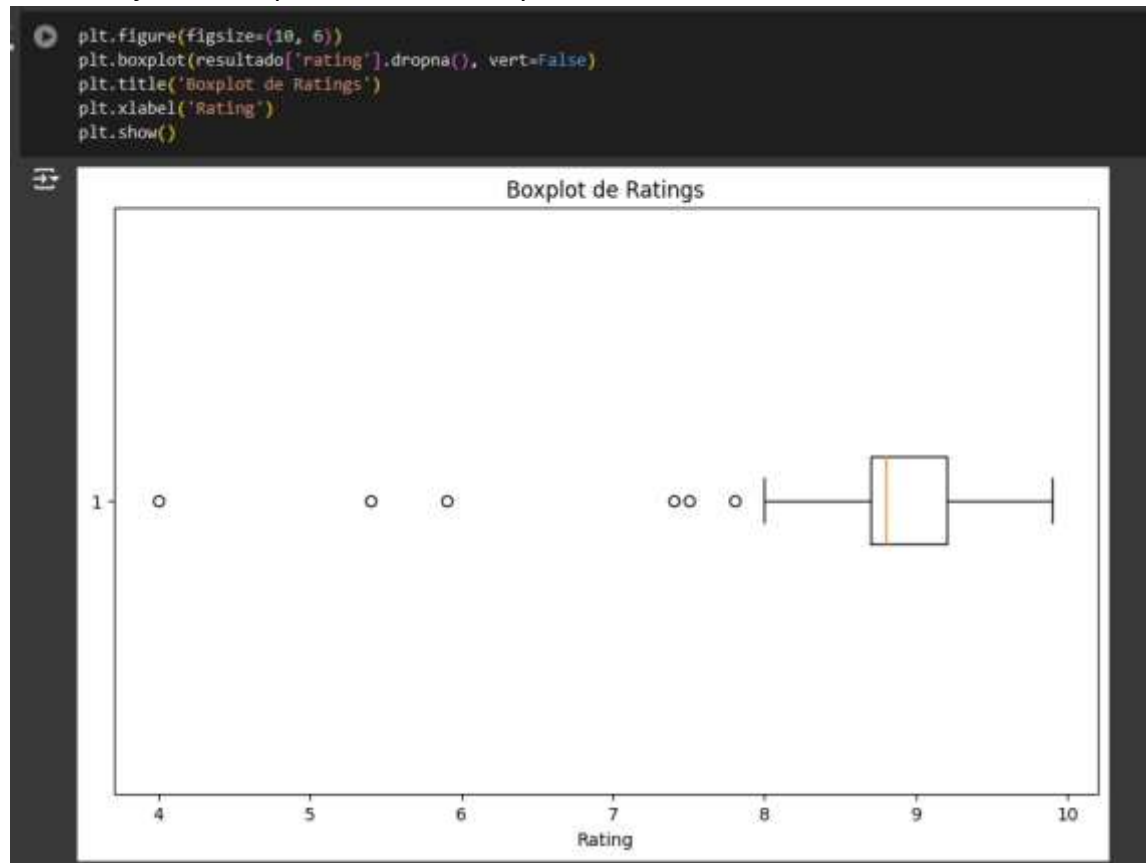


Gráfico Boxplot. Identificando outliers

```

Q1 = resultado['rating'].quantile(0.25)
Q3 = resultado['rating'].quantile(0.75)

IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

outliers = resultado[(resultado['rating'] < lower_bound) | (resultado['rating'] > upper_bound)]
rx = outliers[['ID', 'rating']]
print("Outliers:")
print(rx)

```

Outliers:

ID	rating
8	68
28	69
58	70
70	71
61	72
16	73

Imagen 9. Outliers identificados

Ahora se nos pide elegir una técnica para tratar los outliers, la que elegimos para tratar los outliers fue reemplazar con la mediana. Esta técnica se eligió porque es robusta frente a valores extremos y proporciona una forma simple de manejar los outliers sin asumir una distribución específica de los datos. Reemplazar los outliers con la mediana ayuda a reducir la influencia de los valores extremos en la media y la varianza, haciendo que las estadísticas sean más representativas del conjunto de datos central.

```
media_original = resultado['rating'].mean()
varianza_original = resultado['rating'].var()

Q1 = resultado['rating'].quantile(0.25)
Q3 = resultado['rating'].quantile(0.75)

IQR = Q3 - Q1

lower_bound = Q1 - 2 * IQR
upper_bound = Q3 + 2 * IQR

median_rating = resultado['rating'].median()
resultado['rating'] = resultado['rating'].apply(lambda x: median_rating if x < lower_bound or x > upper_bound else x)

media_ajustada = resultado['rating'].mean()
varianza_ajustada = resultado['rating'].var()

plt.figure(figsize=(10, 6))
sns.boxplot(x=resultado['rating'])
plt.title('Boxplot de Ratings (Con Outliers Reemplazados)')
plt.xlabel('Rating')
plt.show()

print("Media Original:", media_original)
print("Varianza Original:", varianza_original)
print("Media Ajustada:", media_ajustada)
print("Varianza Ajustada:", varianza_ajustada)
```

Imagen 10. Código para el tratamiento de los Outliers

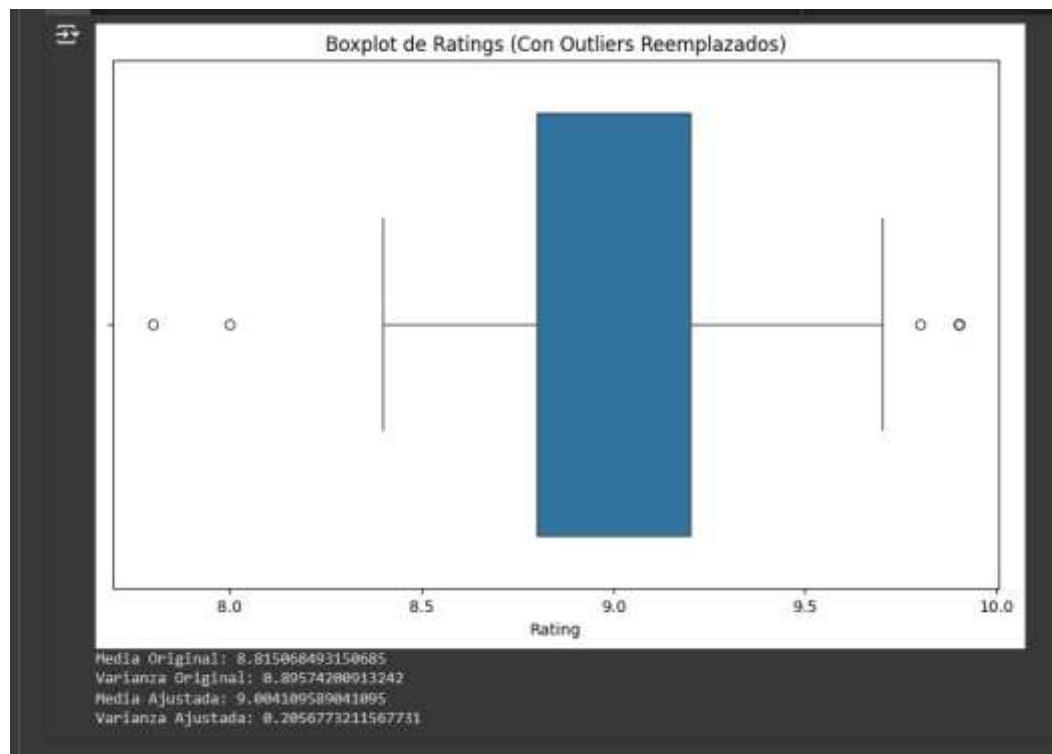


Gráfico Boxplot. Con el tratamiento de los Outliers



- ✓ Media Original: 8.815068493150685
- ✓ Varianza Original: 0.89574200913242
- ✓ Media Ajustada: 9.004109589041095
- ✓ Varianza Ajustada: 0.2056773211567731

Después de tratar los outliers, la media aumentó ligeramente de 8.82 a 9.00, reflejando un valor más representativo del centro de los datos. La varianza disminuyó considerablemente de 0.90 a 0.21, indicando que la dispersión de los datos se redujo al eliminar la influencia de los valores extremos. Esto demuestra que el tratamiento de outliers hace que las estadísticas sean más representativas y menos afectadas por valores extremos.

- **Visualización de Datos:**

Nos solicitan realizar 3 gráficos, a continuación, explicaremos cada uno a detalle y mostraremos sus resultados. Para iniciar mostraremos el código que fue utilizado para cada gráfico.

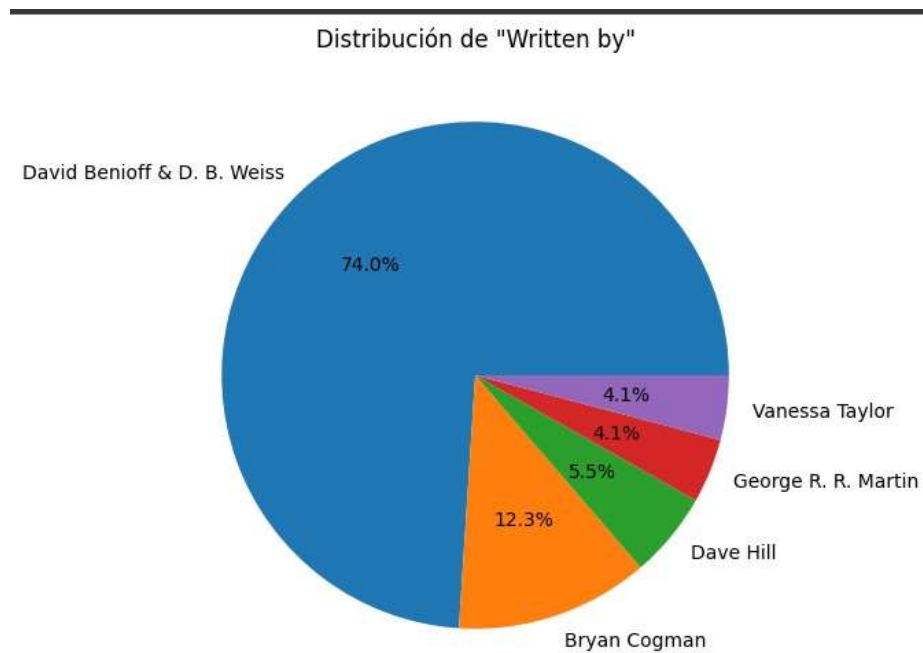
```
# Written by
plt.figure(figsize=(8, 6))
resultado['Written by'].value_counts().plot(kind='pie', autopct='%1.1f%%')
plt.title('Distribución de "Written by"')
plt.ylabel('')
plt.show()

# Rating
plt.figure(figsize=(10, 6))
sns.lineplot(data=resultado.groupby('Temporada')['rating'].mean().reset_index(), x='Temporada', y='rating')
plt.title('Promedio de "Indo rating" por Temporada')
plt.xlabel('Temporada')
plt.ylabel('Promedio de IMDb Rating')
plt.show()

# vistas
plt.figure(figsize=(10, 6))
sns.lineplot(data=resultado.groupby('Temporada')['espectadores'].mean().reset_index(), x='Temporada', y='espectadores')
plt.title('U.S. Viewers (millions) por Temporada')
plt.xlabel('Temporada')
plt.ylabel('U.S. Viewers (millions)')
plt.show()
```

*Imagen 11. Código utilizado para los gráficos.*

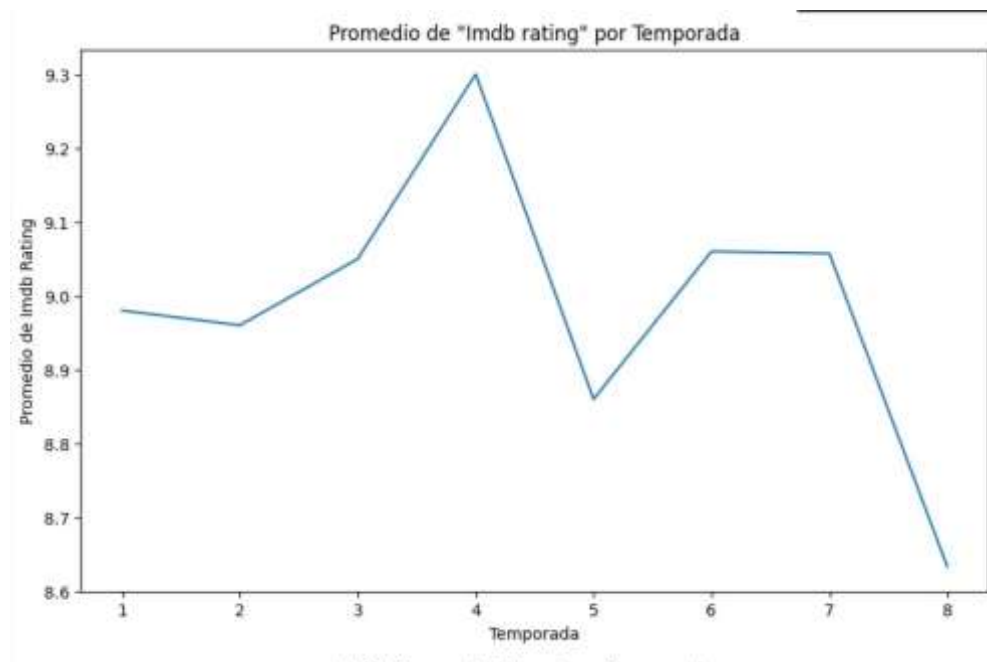
1. Gráfico de torta para "Written by":



*Gráfico torta. "Written by"*

Este gráfico ilustra que el 74% de los episodios fueron escritos por David Benioff y D. B. Weiss, los creadores principales de la serie, destacando su predominante rol en la escritura. Los demás escritores tienen una participación menor.

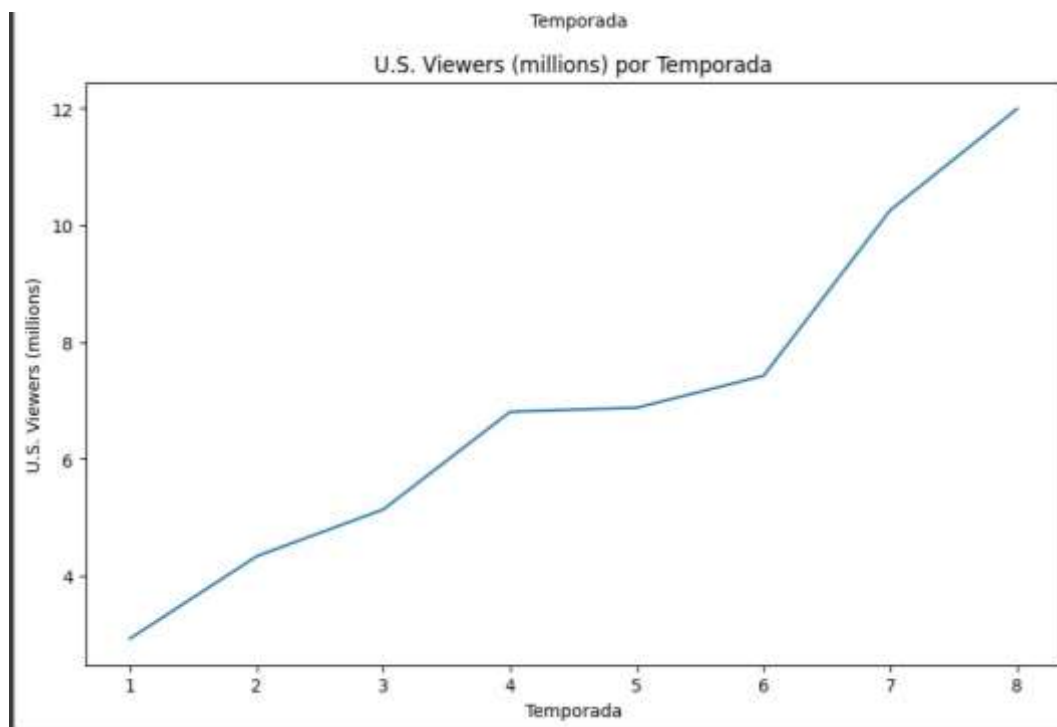
2. Gráfico de línea para el promedio de "Imdb rating" por temporada:



*Gráfico de línea para el promedio de "Imdb rating" por temporada*

Este gráfico muestra la evolución del promedio de calificaciones de IMDb a lo largo de las temporadas de Game of Thrones. Las primeras temporadas tuvieron calificaciones altas, con un notable pico en la cuarta temporada, superando 9.3. A partir de la quinta temporada, las calificaciones comenzaron a descender, con una caída abrupta en la octava y última temporada, reflejando el creciente descontento de los espectadores con el final de la serie.

### 3. Gráfico de línea para "U.S. viewers (millions)" por temporada:

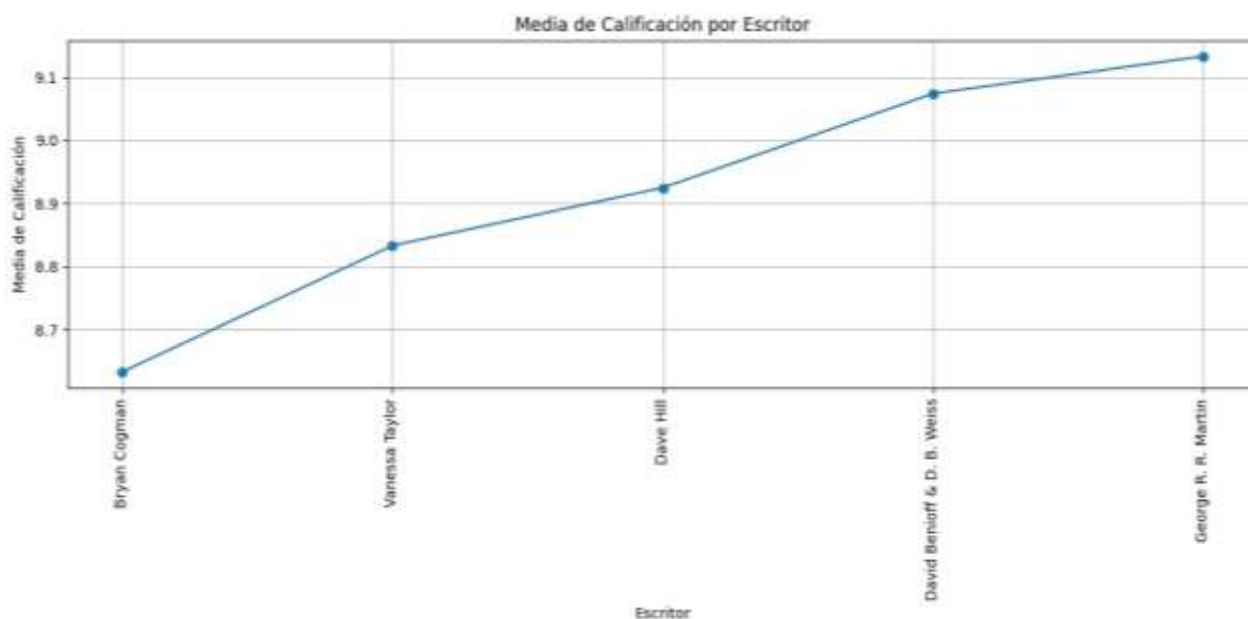


*Gráfico de línea para "U.S. viewers (millions)" por temporada.*

Este gráfico muestra el aumento en las vistas de "Juego de Tronos" a lo largo de sus temporadas, destacando un notable incremento hacia el final de la serie. La creciente anticipación por la última temporada se refleja en el notable aumento de visualizaciones en los episodios finales, evidenciando el entusiasmo y la expectación de los fans. Este fenómeno subraya el impacto duradero y la capacidad de la serie para mantener el interés de su audiencia hasta el último episodio.

- **Visualización de Datos adicionales:**

Elegimos mostrar un gráfico de línea para mostrar la media de calificación por escritor.



*Gráfico de línea para mostrar la media de calificación por escritor.*

El gráfico de líneas es una herramienta ideal para mostrar la media de calificación por escritor, ya que permite visualizar de manera clara las tendencias y comparaciones entre los diferentes escritores. La conexión de los puntos con una línea facilita la identificación de los escritores con las calificaciones más altas y más bajas. En este caso, destaca a George R. R. Martin con la calificación promedio más alta y a Byron C. con la más baja. Este tipo de gráfico es eficiente y directo, proporcionando una representación rápida y efectiva de las diferencias en las calificaciones promedio entre los escritores.

## 4. Discusión

El análisis de los resultados revela distintas tendencias en cómo la participación de los guionistas y la comprensión de la audiencia se han desarrollado a lo largo del tiempo. Weiss escribió la mayor parte del programa, y eso es importante porque lo que se les ocurrió realmente dio forma a toda la historia. Pero luego, decayó un poco, especialmente en la última temporada. La respuesta del público a la historia y su conclusión refleja la desaprobación del público por el final de la serie.

Sin embargo, el análisis tiene algunas limitaciones. La imputación de valores faltantes, particularmente en la columna de fecha y otros datos numéricos, puede no representar con precisión los datos reales porque se basa en métodos estadísticos como el cálculo de la media o la moda. Del mismo modo, encontrar y manejar datos inusuales. Los puntos podrían haber cambiado los resultados, ya que algunos números realmente grandes o pequeños pueden mostrar eventos importantes o únicos. También es clave recordar que cosas

como la cultura o qué más estaba sucediendo cuando se transmitía el programa no se analizaron, y que podría haber cambiado la forma en que la gente veía el espectáculo.

Para hacer mejores conjeturas sobre el futuro, deberíamos utilizar formas más inteligentes de completar la información que falta, como adivinar fechas o quién podría mirar en función de cosas similares. Esto incluye la influencia de las redes sociales y el papel de los profesionales. Echemos un vistazo más de cerca a los escritores. para ver si sus elecciones realmente marcan la diferencia en la popularidad de los episodios, lo que podría decirnos más sobre lo que funciona y lo que no.

## 5. Conclusiones

El proyecto comenzó con la fusión de los datasets utilizando la clave primaria `primary_key`, lo que permitió unificar la información sobre los episodios de la serie Game of Thrones. Esta consolidación fue clave para garantizar que el análisis incluyera datos completos y precisos. A partir de ahí, se realizaron cálculos estadísticos sobre columnas numéricas clave, como el puntaje de IMDb y la audiencia en EE.UU., lo que proporcionó una mejor comprensión de la variabilidad y las tendencias a lo largo de las temporadas.

Posteriormente, se abordó la imputación de valores faltantes, utilizando la media para datos numéricos y la moda para datos categóricos. Este paso fue esencial para completar las columnas con datos ausentes sin afectar significativamente la distribución original. Para la detección de outliers, se utilizó un gráfico de boxplot, lo que permitió identificar valores atípicos, especialmente en la calificación de IMDb, y aplicar técnicas adecuadas para tratarlos. A lo largo del proceso, las visualizaciones de datos jugaron un papel crucial, incluyendo gráficos de línea y torta que ilustraron de manera clara las tendencias de calificación y audiencia.

También se emplearon técnicas avanzadas para analizar datos categóricos, como One-Hot Encoding y Label Encoding. One-Hot Encoding convierte las categorías en columnas binarias, lo que facilita la inclusión de variables categóricas en modelos predictivos sin imponer un orden implícito. Por otro lado, Label Encoding asigna un valor numérico único a cada categoría, lo que es útil para algoritmos que pueden manejar directamente valores numéricos. Estas técnicas se implementaron para asegurar un análisis eficiente y coherente de los datos, permitiendo una mejor integración y comprensión de las variables categóricas en el análisis general.

**Link Google colab:** [https://colab.research.google.com/drive/1WBp-xZeLsEa\\_zQWNpC3KOx6iCBskCnHX?usp=sharing](https://colab.research.google.com/drive/1WBp-xZeLsEa_zQWNpC3KOx6iCBskCnHX?usp=sharing)

## 6. Referencias

### Bibliografía

Ortiz, M. (2023, 14 abril). Game of Thrones: resumen, temporadas, personajes y análisis de la serie. Cultura Genial. [https://www.culturagenial.com/es/game-of-thrones-serie/#:~:text=Game%20of%20Thrones%20\(Juego%20de%20tronos\),%20tambi%C3%A9n%20conocida#:~:text=Game%20of%20Thrones%20\(Juego%20de%20tronos\),%20tambi%C3%A9n%20conocida](https://www.culturagenial.com/es/game-of-thrones-serie/#:~:text=Game%20of%20Thrones%20(Juego%20de%20tronos),%20tambi%C3%A9n%20conocida#:~:text=Game%20of%20Thrones%20(Juego%20de%20tronos),%20tambi%C3%A9n%20conocida)

Estadística, P. Y. (2023, 26 enero). Valores atípicos (outliers). Probabilidad y Estadística. <https://www.probabilidadyestadistica.net/valores-atipicos-outliers/>

Guía completa para el Manejo de Datos Faltantes | Codificando Bits. (s. f.). Codificando Bits. <https://www.codificandobits.com/blog/manejo-datos-faltantes/>

Joshi, S. (2021, 25 febrero). Matplotlib Boxplot Python. Delft Stack. <https://www.delftstack.com/es/howto/matplotlib/matplotlib-boxplot-python/>