ANÁLISIS EXPLORATORIO DE DATOS DE ANIME: TENDENCIAS, POPULARIDAD Y CLASIFICACIONES

DESCRIPCIÓN BREVE

Este reporte presenta un análisis detallado de datos relacionados con el anime, explorando tendencias en calificaciones, popularidad, y tipos de producciones a lo largo del tiempo. A través del uso de métricas descriptivas y correlaciones, se identifican patrones clave que explican los factores asociados al éxito y la preferencia de los usuarios.

Denisia Pelu

Data Sience

Contenido

Análisis de Datos de Anime

1. Introducción

Este análisis se centra en el estudio de los registros de animes, evaluando su género y calificación promedio. A través de este análisis, exploraremos qué géneros y tipos de anime destacan más según las valoraciones de los usuarios, así como las fechas de estreno de los animes más populares. Además, identificaremos si existe alguna tendencia en cuanto a las estaciones del año en las que los animes más destacados tienden a estrenarse, como el verano, otoño o invierno.

Para llevar a cabo este análisis, se han utilizado dos conjuntos de datos (DataFrames):

2. DataFrames Utilizados

DataFrame 1 (df_1)

Este conjunto de datos contiene información clave sobre los animes registrados en la plataforma MyAnimeList. Los campos que componen este DataFrame son los siguientes:

- anime_id: Identificador único de MyAnimeList que distingue a cada anime.
- name: Nombre del anime.
- genre: Lista de géneros del anime, separados por comas.
- type: Tipo de anime (puede ser: 'TV', 'Movie', 'Special', 'ONA', 'OVA').
- episodes: Número total de episodios del anime.
- rating: Calificación promedio sobre 10 asignada por los usuarios de MyAnimeList.
- members: Número de miembros que han visto o leído el anime en MyAnimeList.

DataFrame 2 (df_2)

Este conjunto de datos ofrece detalles adicionales sobre los animes, en particular aquellos registrados en otra fuente. Los campos que componen este DataFrame son los siguientes:

- Title: Nombre del anime.
- Rank: Clasificación general del anime en la plataforma.
- Type: Tipo de anime (puede ser: 'TV', 'Movie', 'Special', 'ONA', 'OVA').
- Episodes: Número total de episodios del anime.
- Aired: Mes y año de inicio y fin de la emisión del anime.

- **Members: Número de miembros que han visto o leído el anime en esta plataforma.**
- **page_url: Enlace URL a la página del anime en la plataforma correspondiente.**
- **image_url: Enlace URL a la imagen de portada del anime en la plataforma.**
- **Score: Calificación promedio sobre 10 asignada por los usuarios de la plataforma.**

3. Análisis Exploratorio de Datos (EDA)

3.1 Análisis Exploratorio de Datos (EDA)

El análisis comenzó con la unión de los dos DataFrames utilizando la columna normalizada, lo cual permitió mejorar el procesamiento de los datos, eliminando mayúsculas y caracteres especiales (como "-", "/", ";", ":", "#", "º", "!", "?", "+", ".", "☆", "♡", "()", "{}" y "[]"). Sin embargo, es recomendable revisar el DataFrame en busca de otros caracteres extraños que puedan estar afectando la unión. Esta revisión podría explicar la reducción en el número de filas tras combinar ambos DataFrames.

2.1 Método de Unión

Se optó por utilizar el parámetro how='inner' para realizar la unión, ya que al emplear how='outer' (unión que incluye tanto las filas comunes como las no comunes), se generaban 17,168 filas con un promedio de 5,000 valores NaN por columna. La opción inner garantiza una unión más limpia y coherente, lo cual facilita el análisis de los datos.

3.2 Estructura del DataFrame

El DataFrame resultante tiene las siguientes características:

- **Filas: 8,003**
- **Columnas: 17**
- **Espacio en memoria: Utiliza el tipo de datos np.int64, lo que implica un consumo de 8 bytes por valor.**

3.3 Verificación de Jerarquías o Estructuras Anidadas

Se identificó que las columnas 'genre' y 'Aired' podrían contener estructuras anidadas. Al analizar estas columnas, confirmamos que presentan jerarquías o estructuras que deben ser tratadas con más detalle:

- 'genre': Necesita ser desglosada para analizar cada género de forma independiente.
- 'Aired': Se debe extraer solo el mes y año de inicio, creando una nueva columna 'fecha_ini'.

3.4 Clasificación de Variables y Revisión de Valores Faltantes

Conclusiones:

1. Popularidad según Calificaciones y Miembros:
    - Las calificaciones más altas (ej. 9.37, 9.26) corresponden a animes populares con una gran base de miembros (más de 2 millones). Esto indica una correlación entre popularidad y calificación.
2. Relación con Fechas de Emisión:
    - Las fechas de emisión cubren un rango extenso de décadas, lo que permite explorar cómo el año de estreno puede influir en el éxito del anime.
3. Información Duplicada o Redundante:
    - Columnas como Title/name, episodes/Episodes y Type/type contienen datos redundantes. Esto debe ser revisado y corregido para evitar duplicados.
4. Datos Faltantes:
    - Se observó que algunas entradas contienen valores NaN o Unknown en columnas como rating y episodes, lo cual indica la necesidad de tratar los valores faltantes para evitar sesgos en el análisis.
5. Áreas de Análisis:
    - Identificar tendencias en géneros más calificados o con más miembros.
    - Examinar la distribución de animes destacados por temporada o año de estreno.
    - Evaluar cómo diferentes tipos de animes (TV, Movie, OVA, ONA) se distribuyen en cuanto a calificaciones y popularidad.

3.5 Revisión de Duplicados

Se observó que la columna 'nombre_normalizado' contiene animes repetidos. Es necesario eliminar los duplicados en esta columna.

Se identificaron también inconsistencias en las columnas 'type', 'Type', 'episodes' y 'Episodes':

- 'type' contiene valores NaN y 'Type' contiene valores Unknown.
- 'Episodes' es una columna de tipo string que contiene el valor "?" en algunas filas.
- Se deben limpiar y normalizar estas columnas antes de proceder.

Acciones tomadas:

- Se han combinado las columnas 'type' y 'Type', manteniendo la columna 'Type' para consistencia.
- Se resolvieron los valores faltantes en las columnas 'Type' y 'Episodes'.
- Se eliminaron los duplicados de la columna 'nombre_normalizado' y se mantuvo la fila con el valor más alto de 'Episodes', 'Members' y 'Score'/ 'rating'.

3.6 Resumen Estadístico General del DataFrame

A continuación, se presenta un resumen estadístico general de las principales columnas del DataFrame:

Variable	count	mean	std	min	25%	50%	75%	max
anime_id	7924	12255.25	11063.23	1	2649.75	8286	21409	34525
rating	7792	6.64	0.98	2.00	6.12	6.73	7.32	9.50
members	7924	25000.67	65491.09	22	485.75	2691.50	17121	1013917
Rank	7924	6475.58	3690.56	1	3302.75	6440.50	9628.25	12788
Score	7924	6.44	0.95	1.85	5.83	6.50	7.15	9.10

Otras columnas (como genre, name, Type, Episodes, etc.) también se presentan con estadísticas clave (frecuencias, valores únicos, etc.), lo cual es útil para comprender la distribución y las características de los animes en los datos.

3.7 Tratamiento de la Columna 'Genre'

La columna 'genre' estaba anidada, por lo que fue necesario separarla para analizar cada género de forma individual. Se eliminaron espacios innecesarios y se estandarizó la lista de géneros.

Género con la Mejor Puntuación Promedio:

Genre Score Rating

Josei 7.34 7.49

Género con Más Miembros:

Genre Miembros

Comedy 273,806,862

Conclusiones:

- Comedy es el género con más miembros, pero Josei tiene la mejor puntuación promedio en cuanto a Score y rating.
- Josei tiende a enfocarse en historias de romance y la vida cotidiana, dirigido principalmente a mujeres adultas, lo que sugiere que estos animes son más apreciados por su audiencia, a pesar de ser menos populares que los de Comedy.

4. Limpieza de Datos

Durante el proceso de limpieza de datos, se identificaron varias columnas con valores faltantes (NaN), principalmente en las columnas rating, Episodes, Score, Rank, y members. A continuación, se detallan las estrategias utilizadas para el manejo de estos valores faltantes:

- Relleno de valores faltantes en la columna rating: Se consideraron tres opciones para el relleno de valores nulos en la columna rating:
    1. Rellenar con la calificación de la columna Score.
    2. Rellenar con la media de members agrupada por Type.
    3. Rellenar utilizando los cuartiles de members y agrupando por Type.

La primera opción fue descartada debido a que la calificación está relacionada con el número de visitantes y no con las calificaciones directas. Tras evaluar la segunda opción, se detectó un sesgo, por lo que se optó por la tercera, que consistió en dividir los datos por cuartiles según members y rellenar los valores faltantes de rating en función del cuartil y Type. Sin embargo, algunos valores aún permanecieron nulos y fueron eliminados del conjunto de datos, ya que no fue posible rellenarlos adecuadamente.

- Relleno de valores faltantes en la columna Episodes: Para ciertos valores de Episodes, que contenían NaN, se realizó un relleno manual. Las filas con valores atípicos de Episodes fueron inspeccionadas y corregidas en función de la naturaleza del anime.
- Eliminación de filas con NaN en season: Se eliminaron las filas con valores faltantes en la columna season, ya que eran registros de animes cuya fecha de creación no pudo determinarse.

---

5. Análisis Univariante

En el análisis univariante, se examinaron las distribuciones de las principales variables del conjunto de datos para entender su comportamiento y características. A continuación se presentan las interpretaciones y observaciones clave:

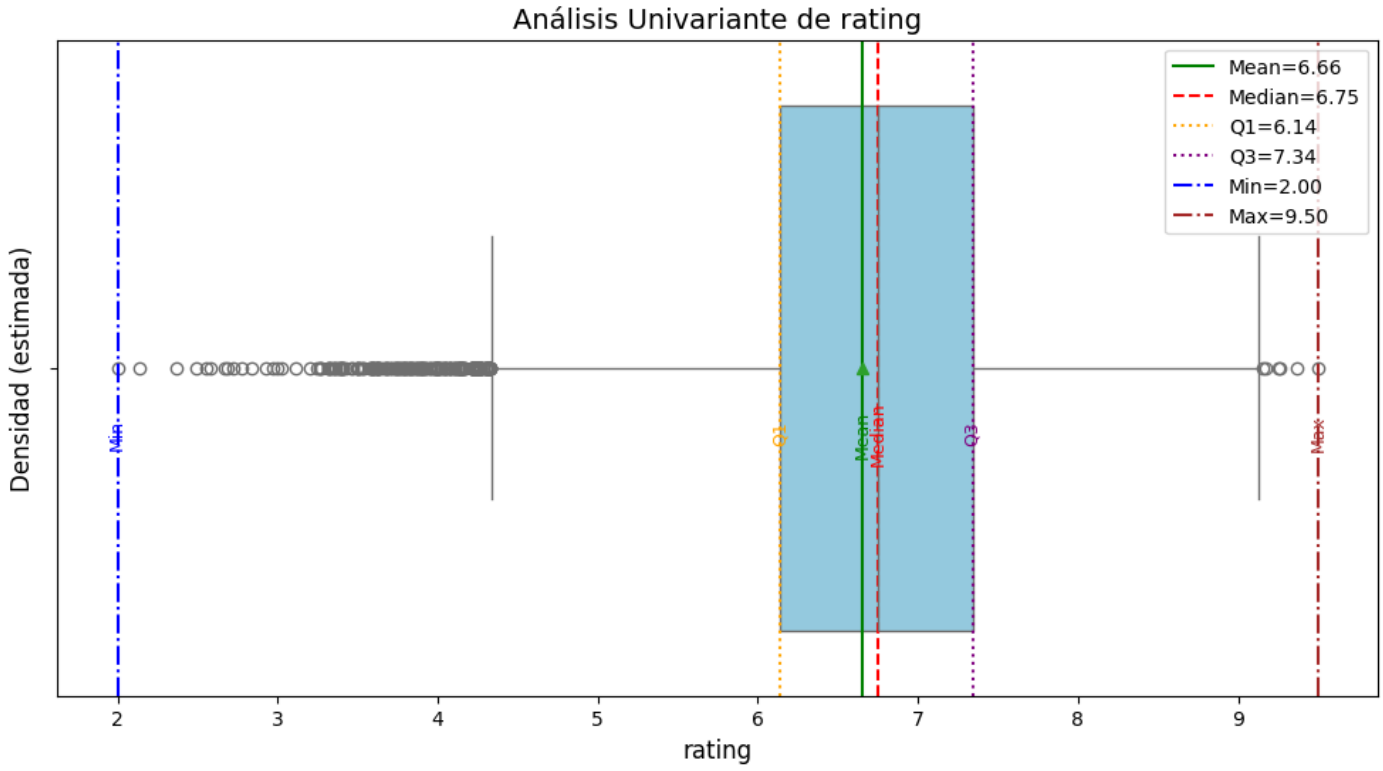
- Distribución de members / Members: La distribución de estas variables muestra una fuerte asimetría positiva, con una cola larga a la derecha. Esto indica que la mayoría de los animes tienen un número relativamente bajo de miembros, pero existen algunos animes con una cantidad excepcionalmente alta de miembros. Esto es característico de distribuciones exponenciales.
- Distribución de rating / Score: Ambas distribuciones son similares, con una tendencia a concentrarse en valores moderados. La distribución de rating se asemeja a una distribución normal, con una ligera cola hacia la derecha, lo que sugiere que la mayoría de los animes tienen calificaciones intermedias, mientras que los valores extremos son menos comunes.
- Distribución de Episodes: La distribución de la variable Episodes presenta un sesgo positivo, lo que significa que la mayoría de los animes tienen un número bajo de episodios, mientras que algunos pocos tienen valores extremos muy altos. Este sesgo puede estar relacionado con animes muy largos que constituyen excepciones.
- Distribución de date_ini: Se observa que la mayoría de los animes comenzaron a emitirse entre 1994 y 2012. Los registros anteriores a 1970 son pocos, y estos valores atípicos podrían necesitar una revisión más detallada.

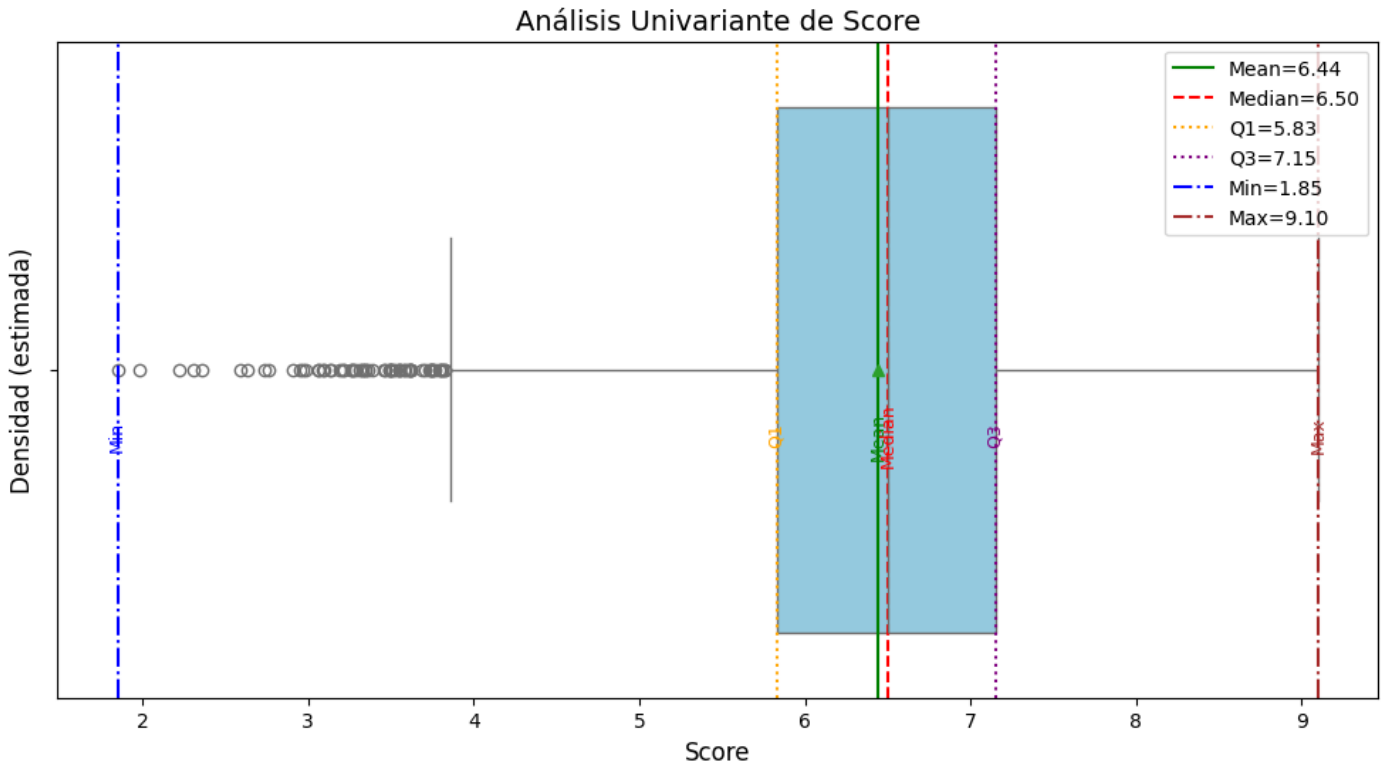
---

6. Medidas de Asimetría y Curtosis

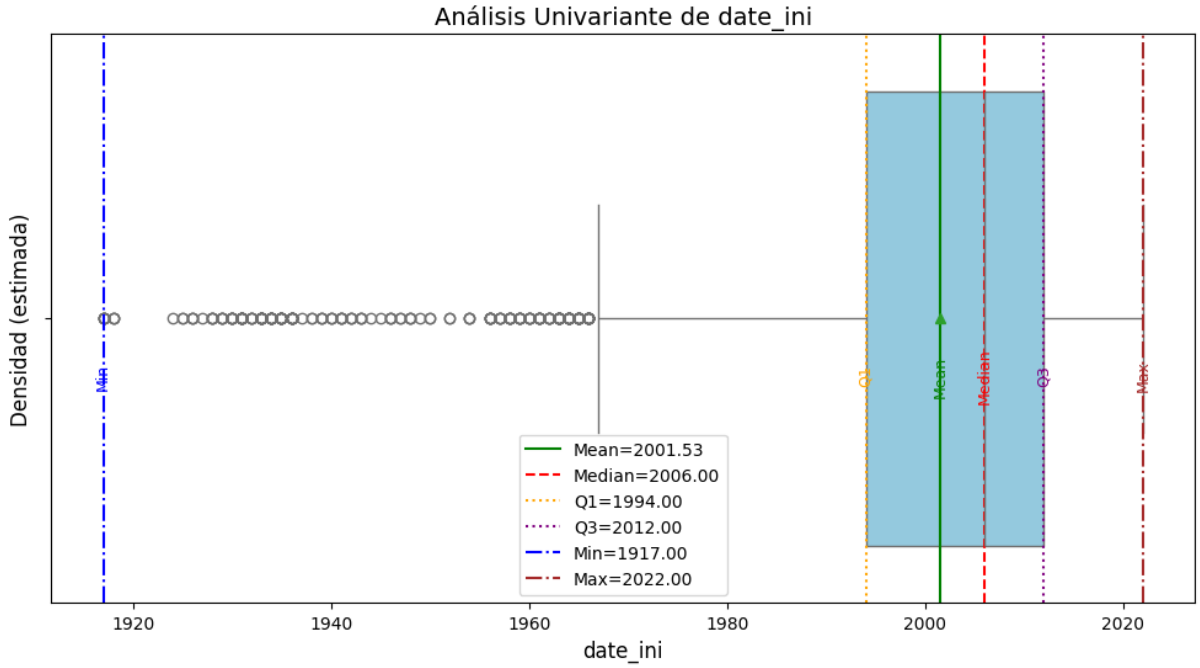
Se calcularon las medidas de asimetría y curtosis para las variables clave con el fin de evaluar el sesgo y la forma de las distribuciones.

- **Episodes:**
    - **Asimetría: 44.46 (Sesgo positivo, distribución sesgada a la derecha).**
    - **Curtosis: 2710.02 (Curtosis mayor que 3, colas más pesadas que una distribución normal).**
- **rating:**
    - **Asimetría: -0.69 (Distribución relativamente simétrica).**
    - **Curtosis: 0.84 (Curtosis menor que 3, colas más ligeras que una distribución normal.**





- **members:**
  - **Asimetría: 5.61 (Sesgo positivo, distribución sesgada a la derecha).**
  - **Curtosis: 43.80 (Curtosis mayor que 3, colas más pesadas que una distribución normal).**
- **date_ini:**
  - **Asimetría: -1.96 (Distribución sesgada a la izquierda).**
  - **Curtosis: 5.54 (Curtosis mayor que 3, colas más pesadas que una distribución normal).**

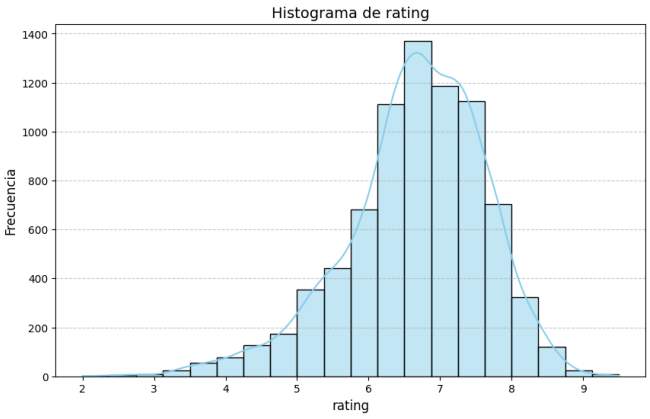


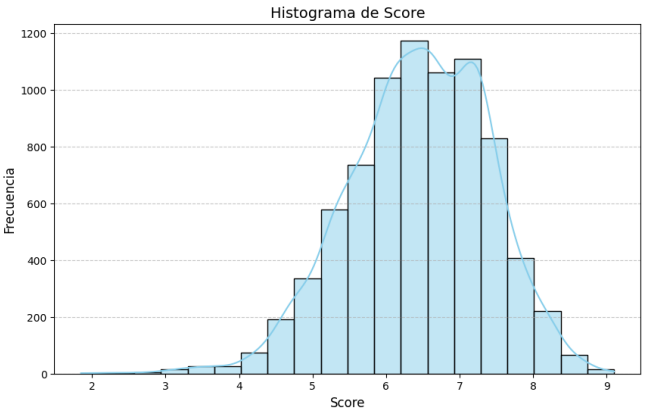
Las medidas de asimetría y curtosis confirman las observaciones previas sobre las distribuciones sesgadas y las colas pesadas en las variables clave como Episodes y members.

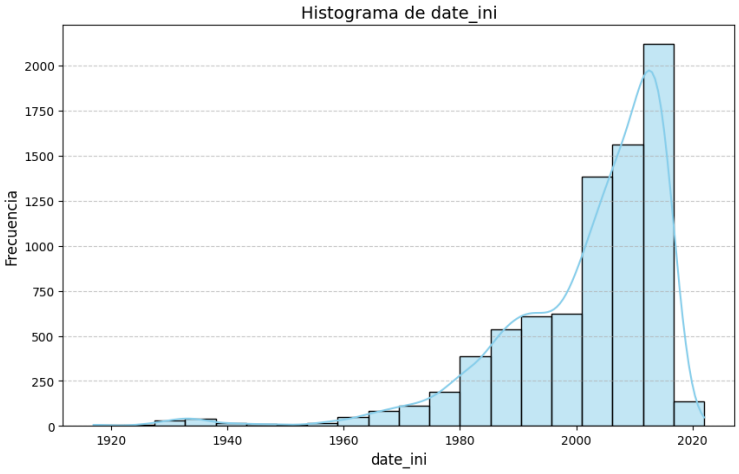
---

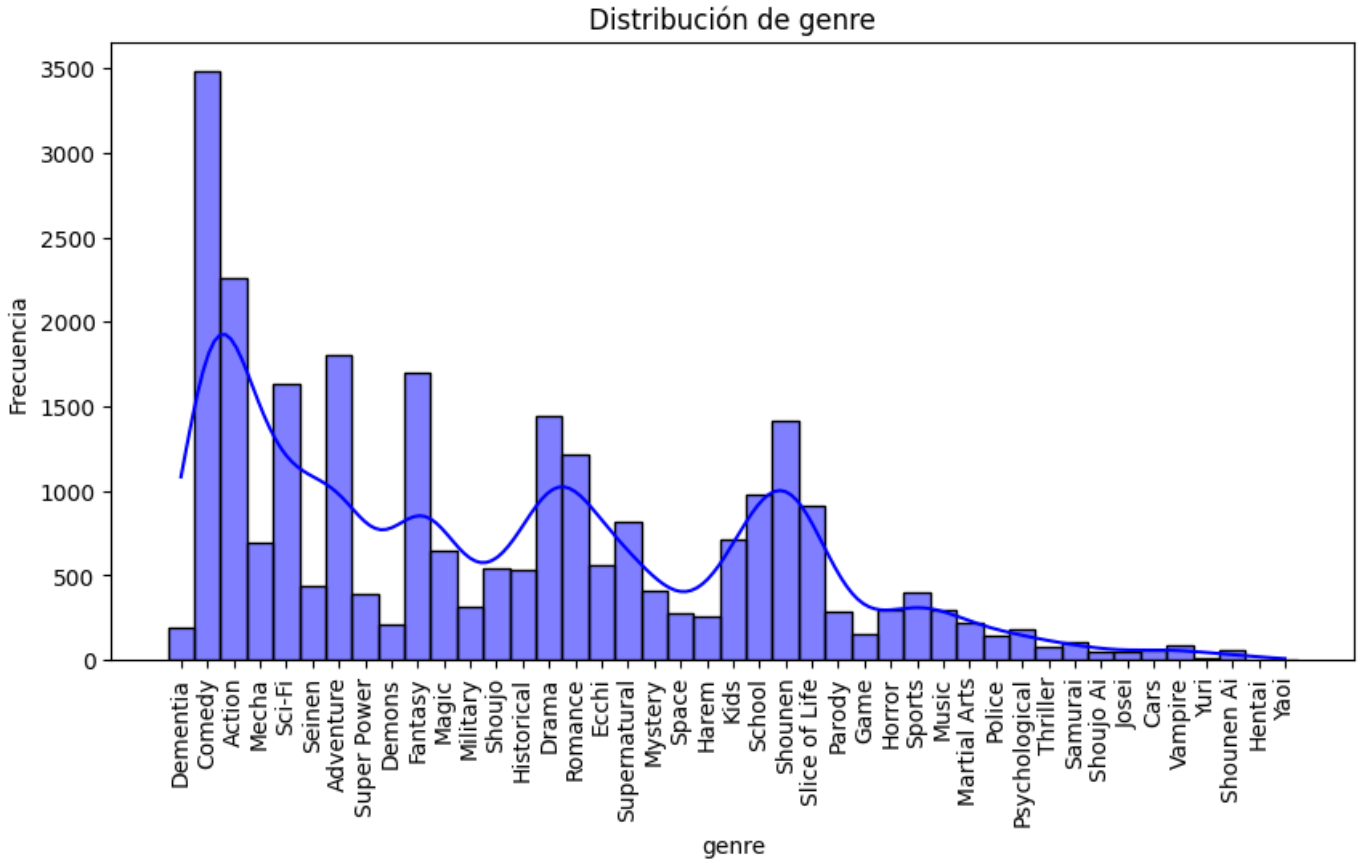
7. Pruebas de Normalidad

Se realizó una prueba de normalidad utilizando el test de Shapiro-Wilk. Los resultados mostraron que la mayoría de las variables no siguen una distribución normal. Este hallazgo es importante, ya que puede afectar la interpretación de la media y la varianza, y sugiere que otras métricas, como la mediana, podrían ser más apropiadas para algunas variables.









8. Conteo y Distribución de Variables Categóricas

Se analizaron las distribuciones de varias variables categóricas importantes:

- **season: La mayoría de los animes fueron lanzados en los meses de abril, octubre y julio, lo que sugiere una concentración de lanzamientos en ciertos períodos del año.**
- **date_ini: Se observó que la mayoría de los animes comenzaron entre los años 2012 y 2014. Este análisis muestra una tendencia clara de crecimiento en la producción de animes en estos años.**
- **Type: Los animes se distribuyen principalmente entre cinco tipos: TV (38.98%), Movie (21.01%), OVA (19.12%), Special (14.51%) y ONA (6.37%).**
- **genre: Los géneros más comunes en los animes son Comedy, Action, Adventure, Fantasy, Drama y Shounen. Se observa que el género Yuri es extremadamente raro en comparación con otros géneros.**

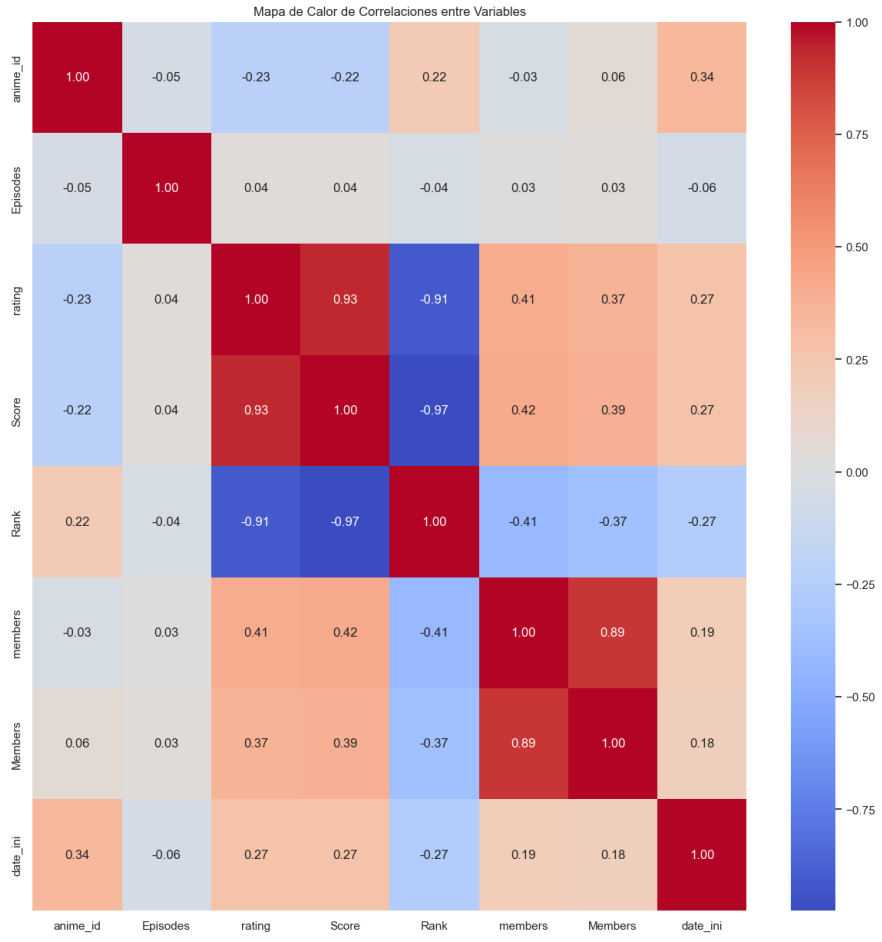
9. Relaciones entre Variables

9.1 Matriz de Correlación

Para explorar las relaciones entre las variables principales del conjunto de datos, se calculó la matriz de correlación, que mide la relación lineal entre pares de variables numéricas. Los valores oscilan entre -1 (relación negativa perfecta) y 1 (relación positiva perfecta), mientras que 0 indica la ausencia de relación lineal.

Resultados Clave:

- **Existe una correlación positiva fuerte (0.93) entre la calificación (rating) y la puntuación (Score), lo que indica que los animes con mayor calificación tienden a tener una puntuación más alta.**
- **La relación entre el Rank y la Score es negativa fuerte (-0.97), lo cual es esperado ya que un rango más bajo (mejor clasificación) está asociado con puntuaciones más altas.**
- **Las variables members y Members presentan una alta redundancia (0.89), lo que sugiere que ambas columnas reflejan información similar.**
- **El tipo de anime (Type) muestra una relación moderada con rating (0.31), lo que indica que el tipo puede influir parcialmente en la calificación.**

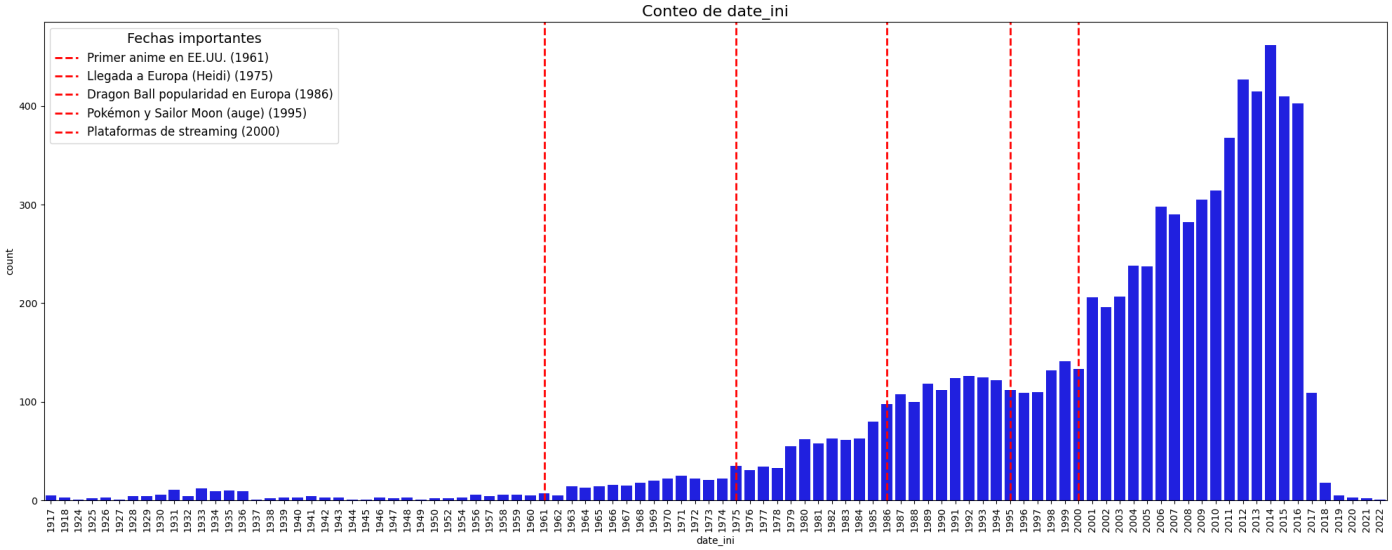


10. Observaciones Adicionales

Además del análisis cuantitativo, se incluyó un contexto histórico sobre la evolución del anime. Este análisis histórico muestra cómo el anime ha crecido y se ha expandido en diferentes mercados a lo largo de las décadas. Los hitos clave incluyen:

- **Década de 1960: Introducción del anime en Estados Unidos.**
- **Década de 1970: Llegada del anime a Europa.**
- **Década de 1980: Aumento de la popularidad del anime en Occidente.**
- **Década de 1990: Auge del anime en Europa.**
- **Década de 2000 en adelante: Expansión del acceso al anime a través de Internet y plataformas de streaming.**

Este contexto histórico es útil para interpretar los picos y valles en la producción de animes a lo largo del tiempo, coincidiendo con la expansión del anime en mercados internacionales.



El análisis exploratorio de datos (EDA) ha proporcionado una visión integral sobre la estructura y características del conjunto de datos de animes. A través del análisis univariante, la exploración de valores atípicos, y la evaluación de distribuciones, se han identificado patrones y tendencias que reflejan el comportamiento de la industria del anime a lo largo del tiempo. Las medidas de asimetría y curtosis confirman las distribuciones sesgadas, mientras que las pruebas de normalidad indican que las distribuciones no son normales. A partir de este análisis, se pueden realizar pasos posteriores para modelar o realizar inferencias más profundas, teniendo en cuenta los sesgos y las distribuciones no normales de las variables clave.

Durante el análisis exploratorio de datos (EDA) se observaron algunos puntos clave en las distribuciones de las variables más importantes. A continuación se detallan algunas conclusiones adicionales que surgieron a partir de los datos:

- Valor máximo de rating: El valor máximo de la columna rating es 9.5, lo que representa un valor excepcionalmente alto dentro del conjunto de datos. Un ejemplo de anime con este rating es *Mogura no Motoro* (ID 23005), que pertenece al género *Slice of Life*, con una calificación de 9.5 y un Score de 5.56. Este anime tiene solo 1 episodio y es de tipo *Movie*, y se emitió entre julio de 1962.
- Cuartil de Score en 9.1: Un anime como *Fullmetal Alchemist: Brotherhood* (ID 5114), clasificado en el cuarto cuartil (Q4) de la columna Score, tiene un rating de 9.26 y un Score de 9.1, con 64 episodios. Este anime es un ejemplo de producción extensa y popular, con un número significativo de miembros (793,665), lo que también resalta la correlación entre la alta calificación y la cantidad de miembros registrados.
- Valores de members y ejemplos destacados: El análisis de la columna members muestra que existen animes con números muy altos de miembros registrados. Un ejemplo es *Death Note* (ID 1535), que tiene un rating de 8.71 y un Score de 8.62, siendo parte del cuartil Q4 de la variable members, con 1,013,917 miembros. De manera similar, *Shingeki no Kyojin* (ID 16498) también pertenece a este cuartil (Q4) con un número de miembros de 3,759,013 y un rating de 8.54. Estos ejemplos reflejan la tendencia de que los animes con más miembros registrados generalmente tienen una alta calificación tanto en rating como en Score.

Estas observaciones refuerzan la idea de que los valores extremos en las variables de rating y members están ligados a ciertos animes que destacan tanto en términos de popularidad como de calificación. Estos animes son representativos de producciones exitosas que logran una gran aceptación entre los espectadores.