DOCUMENTACIÓN EDA HARRY POTTER

Contexto

El presente informe documenta el proceso y los hallazgos obtenidos en el Análisis Exploratorio de Datos (EDA) sobre un dataset relacionado con los personajes del universo de Harry Potter. Este dataset incluye información sobre género, casas, especies, estatus sanguíneo, colores de pelo y ojos, y lealtades, entre otros atributos. El objetivo principal del análisis es identificar patrones, relaciones y posibles inconsistencias en los datos, proporcionando una base para futuros estudios o aplicaciones.

Estructura inicial del dataset

El dataset original constaba de:

- **140 filas**
- **15 columnas**, incluyendo:
    - **Variables categóricas**: Gender, House, Species, Blood status, etc.
    - **Variables numéricas**: Id (identificador).
    - **Columnas con valores faltantes**: Job, House, Loyalty, entre otras.

Tabla descriptiva original

Nombre	Tipo de dato	Descripción	Relevancia en el análisis
Id	int64	Identificador del personaje	baja
Name	object	Nombre	Baja
Gender	object	Hombre o Mujer	Alta
Job	object	Trabajo del personaje	Baja
House	object	Casa a la que pertenece	Alta
Wand	object	Varita que usa	Baja
Patronus	object	Patronus de la persona	Baja
Species	object	Especie del personaje	Alta
Blood status	object	Origen sanguíneo	Alta
Hair colour	object	Color de pelo	Media
Eye colour	object	Color de ojos	Media
Loyalty	object	Lealtad a alguien o algo	Alta
Skills	object	Habilidades que posee	Baja
Birth	object	Fecha de nacimiento	Baja
Death	object	Fecha de defunción	Baja

Preguntas claves del análisis

Sobre características generales

1. ¿Cuál es la distribución de personajes por género?
2. ¿Cómo se distribuyen los tipos de sangre entre los personajes?

Sobre atributos específicos

1. ¿Cuáles son los colores de pelo y ojos más comunes entre los personajes? ¿Se relacionan con alguna casa o tipo de sangre?
2. ¿Qué especies predominan entre los personajes y cómo influyen en sus lealtades?

Preguntas avanzadas (bivariantes y multivariantes)

1. ¿Cómo se relacionan las lealtades con la especie o el tipo de sangre de los personajes?
2. ¿Qué combinaciones de atributos parecen ser únicas o recurrentes?
3. ¿Qué casas presentan mayor diversidad en **Species** y **Loyalty**?
4. ¿Cómo interactúan **Hair colour**, **Eye colour**, y **House** entre sí?

Preparación de los datos (ETL)

Análisis de valores únicos por columna

1. Id **y** Name**:**
    - Son identificadores únicos (140 valores en 140 filas). Probablemente no necesiten transformaciones.
2. Gender**:**
    - Solo tiene 2 categorías. Es una columna bien estructurada y lista para el análisis.
3. House**:**
    - Seis categorías únicas: probablemente las cuatro casas de Hogwarts más otras dos (Sin casa o Desconocido). Deberíamos verificar la consistencia.
4. Species**:**
    - Diez especies únicas parecen razonables. Deberíamos revisar si todas están correctamente representadas.

5. Blood status**:**
    - Quince categorías son más de las esperadas. Es posible que haya inconsistencias (por ejemplo, diferencias entre "Puro" y "Sangre pura"). Requiere revisión.

6. Hair colour**,** Eye colour**:**
    - Con 36 y 25 categorías únicas, es probable que necesiten limpieza (e.g., "negro" y "oscuro" podrían referirse a lo mismo).

7. Loyalty**:**
    - 19 categorías: verificar si hay nombres duplicados con variaciones menores.

Acciones realizadas

1. **Columnas eliminadas**: Birth, Death, Skills, Job, Wand y Patronus por no ser relevantes para el análisis.
2. **Detección y eliminación de duplicados**: Se eliminaron filas duplicadas.
3. **Tratamiento de valores inconsistentes**:
    - **House**: Se corrigieron valores para limitar las casas a Gryffindor, Slytherin, Hufflepuff y Ravenclaw. Los valores faltantes se reemplazaron por "Unknown".
    - **Hair colour y Eye colour**: Se unificaron colores similares y se rellenaron valores faltantes con "Unknown".
    - **Species, Loyalty y Blood status**: Se agruparon valores en listas cerradas de categorías relevantes y se rellenaron los valores faltantes con “Unknow”.
    - **Gender**: Se rellenó el único valor faltante con la moda.

Estructura final del dataset

- **159 filas**
- **9 columnas**

Tabla descriptiva final

Nombre	Tipo de dato	Descripción	Relevancia en el análisis

Id	int64	Identificador del personaje	baja
Name	object	Nombre	baja
Gender	object	Hombre o Mujer	Alta
House	object	Casa a la que pertene	Alta
Species	object	Especie del personaje	Alta
Blood status	object	Origen sanguíneo	Alta
Hair colour	object	Color de pelo	Media
Eye colour	object	Color de ojos	Media
Loyalty	object	Lealtad a alguien o algo	Alta

Análisis

1. Análisis univariante

Se realizaron gráficos de barras para explorar la distribución de frecuencias de cada columna categórica. Los principales hallazgos incluyen:

- **Gender**: La mayoría de los personajes son hombres (aproximadamente un 60%).
- **House**: Gryffindor tiene la mayor cantidad de personajes (35%). Slytherin ocupa el segundo lugar con un 18%.
- **Species**: Los humanos son la especie predominante (79%), seguidos de los fantasmas.
- **Blood status**: "Sangre pura" es el tipo de sangre más frecuente (29%), seguido de "mestizos" (19%).
- **Hair colour**: Negro, marrón y pelirrojo son los colores más comunes, respectivamente.
- **Eye colour**: Marrón es el color más común conocido.
- **Loyalty**: La lealtad a Howgarts y a la Orden del Fenix son las categorías predominantes.

2. Análisis bivariante

Se analizaron combinaciones entre columnas categóricas para identificar patrones y relaciones. Ejemplos destacados:

- **House vs. Gender**: Se observó una distribución relativamente balanceada en la representación de géneros dentro de las casas de Ravenclaw y Hufflepuff. Se observó una diferencia de hombres más elevada, de entre un 40 – 58%, en comparación con la cantidad de mujeres que hay dentro de las casas de Gryffindor y Slytherin.
- **House vs. Blood status**: Los "sangre pura" tienen una mayor representación en Slytherin y Gryffindor. Los "mestizos" tienen una mayor representación en Hufflepuff y en Ravenclaw.
- **House vs. Loyalty**: Se observó que en Slytherin predomina la lealtad hacia los Mortífagos y en el resto de las casas predomina Howgarts y la Orden del Fénix.

3. Análisis multivariante

Los análisis muestran que no hay ningún tipo de correlación fuerte entre todas las combinaciones de variables.

Conclusiones

En general hay que destacar la gran falta de datos en varias columnas, por lo que la precisión de este análisis podría verse mermada. Yo he decidido juntar todo lo que no conozco en una categoría, y trabajar con lo que sí conozco.

Distribución de Género

- La distribución de género está equilibrada, con una ligera mayoría masculina. Esto sugiere que la narrativa de *Harry Potter* da un rol destacado tanto a personajes masculinos como femeninos.

Distribución de Casas

- Las casas más representadas son **Gryffindor** y **Slytherin**, lo cual es consistente con la importancia narrativa de estas casas en la saga.

Distribución de especies

- La gran mayoría de personajes son humanos.

Distribución de estatus sanguíneo

- La mayor parte de los personajes son “sangre pura”, seguidos por los “mestizos”.

Distribución de colores de pelo

- El color de pelo más común es el **negro**, seguido del rojo y el marrón.

Distribución de colores de ojos

- El color de ojos más común es el **marrón**, seguido del azul.

Relación entre el género y las casas

- No queda claro que haya alguna relación entre la casa y el género, pero **Gryffindor y Slytherin** presentan una mayor proporción de personajes masculinos, mientras que **Ravenclaw y Hufflepuff** tiene una distribución más equitativa entre géneros.

Relación entre el estatus sanguíneos y las casas

- En **Gryffindor y Slytherin** hay una mayor cantidad de “sangre pura”, mientras que en **Ravenclaw y Hufflepuff** hay una mayor concretación de “mestizos”.

Relación entre las lealtades y las casas

- En **Slytherin** predomina la lealtad a los **Mortífagos,** y en el **resto de casas** la lealtad que más predomina es a **Howgarts y a la Orden del Fénix**.

En cuanto al **análisis multivariante**, habría que destacar la **no correlación** entre varias variables a la vez.