Análisis Exploratorio de Datos de la Ultra Maratón Transvulcania

Introducción:

El presente informe ejecutivo resume el Análisis Exploratorio de Datos (EDA) realizado sobre un extenso conjunto de datos de carreras de ultramaratón, con un enfoque particular en la emblemática Transvulcania, una prueba de resistencia que se desarrolla en el singular entorno natural de la isla de La Palma, conocida afectuosamente como la "isla bonita" del archipiélago canario .

La Transvulcania no es solo una carrera; es la heredera de antiguos senderos de trashumancia que históricamente articularon la comunicación entre las remotas comunidades de la isla. Lo que antaño fue una red de caminos para el ganado, hoy se ha transformado en un desafiante recorrido de 73 kilómetros que serpentea entre imponentes volcanes, frondosos bosques y las características playas de arena negra de La Palma, otorgándole una personalidad única y marcada en el panorama del trail running.

El objetivo principal de este EDA es comprender en profundidad las dinámicas y características de la carrera Transvulcania a través del análisis de los datos históricos de participación y rendimiento de sus atletas. Buscamos identificar tendencias relevantes, patrones en la participación por género, nacionalidad y club, la evolución del rendimiento a lo largo de los años, y la relación entre variables como la edad y la velocidad. Para llevar a cabo este análisis, se utilizaron diversas librerías de Python especializadas en manipulación y visualización de datos, como Pandas, Seaborn, NumPy, Matplotlib y SciPy.

Findings: Análisis Técnico de los Datos de la Transvulcania

El dataset inicial comprendía un total de 7.461.226 registros de carreras de ultramaratón a nivel mundial, abarcando un amplio periodo desde 1798 hasta 2022, con la participación de 1.641.168 atletas únicos. Tras un filtrado específico, se extrajo la información correspondiente a la Transvulcania, resultando en un conjunto de datos de 10.493 registros.

Limpieza y Transformación de Datos:

Se realizó una exploración inicial para identificar valores nulos y duplicados3 . Posteriormente, se procedió a la renombración de las columnas para mejorar su legibilidad y consistencia4 . Se verificaron y transformaron los tipos de datos de las columnas relevantes. En particular, la columna de fechas del evento se limpió y se convirtió al formato datetime. La distancia de la carrera se extrajo y se convirtió a formato numérico (kilómetros)7 . El rendimiento de los atletas, inicialmente en formato de texto, se transformó a un objeto timedelta para facilitar el análisis temporal. La velocidad media se convirtió a formato numérico.

El tratamiento de los valores nulos se abordó de la siguiente manera:

- Para la columna "club", se identificaron diversos términos que indicaban la ausencia de afiliación (e.g., "unaffiliated", "no club") y se unificaron bajo la categoría "Unaffiliated". Los valores nulos restantes en esta columna también se imputaron con "Unaffiliated".
- Para la columna "birth_year", se analizó la distribución de los datos, observándose una distribución normal con pocos valores atípicos. Los valores nulos se imputaron utilizando una imputación aleatoria dentro del rango intercuartílico (IQR) de los años de nacimiento (1971-1982).
- Se calculó la "age" de los atletas restando el año de nacimiento al año del evento.
- La columna "age_category" presentaba valores nulos y un formato basado en el género ("M" para masculino, "W" para femenino) seguido de un rango de edad (e.g., "M23", "W40", "MU23" para menores de 23). Se creó una función para asignar la categoría de edad basada en la edad calculada y el género, y se aplicó para completar los valores nulos ....Tras la limpieza y transformación, se verificó la ausencia de valores nulos en el DataFrame resultante de la Transvulcania .

Análisis Descriptivo y Visualizaciones:

- **Años Disponibles**: Los datos de la Transvulcania analizados comprenden las ediciones celebradas entre 2009 y 2019.
- **Distribución de Género**: Se observó una predominancia de la participación masculina (9470 registros) frente a la femenina (1023 registros).
- **Número de Finishers**: La visualización del número de finishers a lo largo de los años permite observar la evolución de la participación en la prueba.
- **Performance Media**: El análisis de la performance media (tiempo de finalización) por año y género revela las tendencias en el rendimiento de los atletas a lo largo del tiempo, mostrando una mejora general y diferencias notables entre hombres y mujeres.
- **Participación Femenina**: El análisis de la participación femenina a lo largo de los años muestra la evolución del porcentaje de mujeres que completan la Transvulcania.
- **Países y Clubs con Mayor Representación**: Se identificaron los 10 países y los 10 clubs con mayor número de corredores participantes en la Transvulcania, destacando la importante presencia de corredores nacionales.
- **Relación Edad y Velocidad Media**: El análisis de la relación entre la edad y la velocidad media, diferenciando por género, sugiere posibles patrones en el rendimiento en función de la edad de los atletas.
- **Grupos de Edad con Mejores Resultados**: A través de la comparación de la velocidad promedio por categoría de edad, se puede identificar cuáles son los grupos de edad con un rendimiento medio superior.
- **Diferencia de Velocidad por Género**: La comparación de la distribución de la velocidad promedio entre hombres y mujeres confirma una diferencia significativa en el rendimiento medio entre ambos grupos.

- **Récords de Velocidad**: El seguimiento de los récords de velocidad (mejor tiempo de finalización) por año y género ilustra la mejora del rendimiento élite a lo largo de las ediciones.
- **Participación Múltiple**: Se identificó un número significativo de corredores que han participado en más de una edición de la Transvulcania (1541 atletas). El análisis de la evolución del rendimiento de los 10 corredores con mayor número de participaciones permite observar sus trayectorias individuales en la prueba.

Conclusiones

En resumen, el análisis exploratorio de los datos de la Transvulcania proporciona una visión detallada de la evolución, la demografía de los participantes y las tendencias de rendimiento de esta emblemática carrera de ultramaratón. Los hallazgos obtenidos permiten una mejor comprensión de las características de la prueba y de los atletas que la desafían año tras año.

Se recomienda continuar investigando el impacto de variables externas, como las condiciones climáticas y el estado del terreno, en el desempeño de los atletas. Además, podría ser de interés analizar la preparación y estrategias de los corredores recurrentes para identificar factores de éxito.