SPRINT10-PROYECTO FINAL

ANALISIS DEL IMPACTO DE LA EDAD EN EL RENDIMIENTO DE LA MEDIA MARATON EN BARCELONA

Diliana Bastidas

IT Academic. 2024

Introducción

El Maratón de Barcelona es mucho más que una competición atlética; es una celebración de la pasión, la perseverancia y la conexión entre el deporte y la cultura. Reconocido como uno de los eventos de larga distancia más destacados del mundo, este maratón anual atrae a corredores de todos los niveles, desde atletas élite que buscan batir récords hasta aficionados que persiguen sus propias metas personales. Su recorrido, cuidadosamente diseñado, no solo desafía la resistencia física, sino que también ofrece una experiencia visual inolvidable al pasar por monumentos icónicos como la Sagrada Familia, el Paseo de Gracia y la Plaza de Cataluña, entre otros.

Además de su atractivo deportivo, el Maratón de Barcelona contribuye significativamente al turismo y a la economía local, convirtiéndose en un motor de promoción de la ciudad condal como destino internacional. Los eventos paralelos, como exposiciones deportivas, ferias y actividades familiares, enriquecen la experiencia, haciendo del maratón una auténtica fiesta para residentes y visitantes.

Objetivo

En este análisis, exploraremos en detalle los datos asociados al evento, para un recorrido de 21Km, el objetivo es identificar patrones relevantes, como la distribución de corredores por rango de edad y géneros, los tiempos promedio según categorías (rango de edad).

Hipótesis

- A medida que aumenta la edad en las mujeres corredoras, hay un aumento en los tiempos de carrera.

Metodología

Los datos utilizados en este estudio provienen de la página, www.edreamsmitjabarcelona.com

El data set contiene información sobre la carrera, a continuación se detalló cada variable y su descripción:

1. **Name**: Nombre del corredor.

2. **Bib**: Número de dorsal del corredor (también llamado "Bib number"), que es el número asignado a cada participante en una carrera para identificación.
3. **Category**: Categoría de edad o tipo del corredor, como SENM (sénior masculino) o M35 (masculino de 35 años o más). Esto permite clasificar a los corredores por grupos de edad o competencias específicas.
4. **Poscat**: Posición del corredor dentro de su categoría. Por ejemplo, si un corredor es el mejor en la categoría M35, su Poscat sería 1. Esto permite ver cómo se clasificaron dentro de su grupo de edad o categoría.
5. **Time**: Tiempo total que tardó el corredor en completar la carrera, expresado en formato HH:MM:SS. Este es un valor clave para saber el desempeño de cada participante.
6. **GAP**: Diferencia de tiempo entre el tiempo del corredor y el tiempo del ganador absoluto de la carrera. Un GAP de 00:00:05, por ejemplo, indica que el corredor terminó 5 segundos después del primero.
7. **AVG**: Promedio de ritmo o velocidad del corredor, generalmente expresado en minutos por kilómetro (min/km). Esto indica la rapidez con la que el corredor, y con sus tipos de dato

Campo	Tipo
Pos	float64
Name	object
Bib	float64
Category	object
Poscat	float64
Time	object
GAP	object
AVG	object
Club	object

Se analizaron los diferentes tipos de variables presentes en el conjunto de datos y se decidió convertir algunos valores a tipos numéricos, como **int** y **float**, con el objetivo de prevenir posibles errores de código en el futuro. Asimismo, se eliminaron varias columnas una que contenía únicamente valores **NaN**, y otras que no aportaba información relevante para el análisis.

Se realizó la limpieza de datos respectivas en todas las columnas de mis dos dataframes para los años 2023 y 2024(df, df1). Luego, realicé una concatenación de ambos dataframes, de igual forma, se llevó a cabo una revisión exhaustiva para asegurar que

no existieran valores no permitidos (NaN) o nulos en ninguna de las columnas del dataframe final.

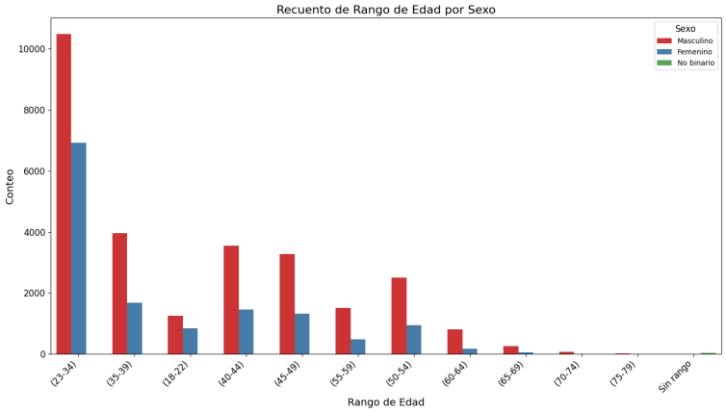
Se crearon nuevas columnas para facilitar el trabajo con la información obtenida, ajustando los valores de Time, GAP y AVG a las unidades correspondientes, de modo que se pudieran trabajar con el mismo tipo de dato.

Además, se creó la columna Sexo para identificar el género de los corredores y una columna adicional para asignar el rango de edad correspondiente a cada corredor según su categoría.

4.- Análisis y visualización

Se llevaron a cabo distintos análisis para comprender la distribución y los patrones de la media maratón.

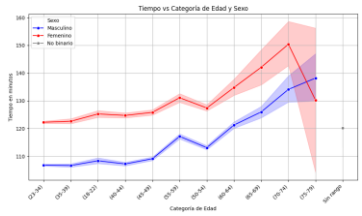
- ✓ Comenzando por el conteo por categorías (rango de edad), donde se visualizó la distribución de los corredores según las diferentes categorías de edad y género. Esto permitió observar la representación numérica de cada grupo en la media maratón, identificando posibles tendencias en la participación.
- ✓ Se contabilizó la cantidad de corredores en cada categoría, proporcionando una visión clara de la distribución y ayudando a entender qué grupos estaban más o menos representados, y se logra observar q la categoría Senior entre (23-34) años de edad es la que tiene mayor participación.

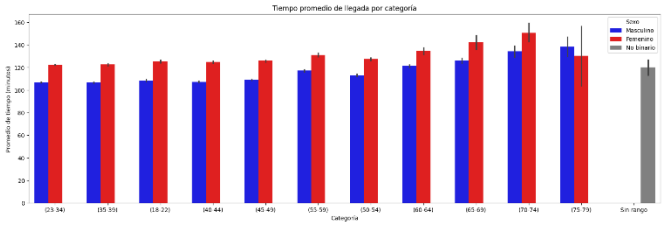


- ✓ Seguidamente, se validó el conteo de mujeres y hombres presentes en la misma, verificando la distribución de los corredores por género, este análisis permitió confirmar la representación de ambos géneros en el evento, observando si existían posibles desequilibrios en la participación. La validación de los datos de género también fue clave para asegurar la correcta asignación de los corredores a sus respectivas categorías de género, garantizando la integridad de los

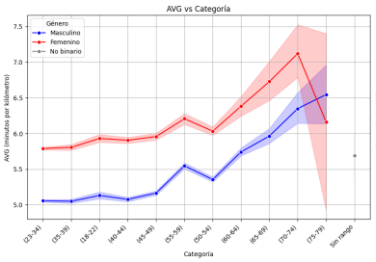
resultados y proporcionando una visión más precisa sobre la diversidad de los participantes en la carrera.

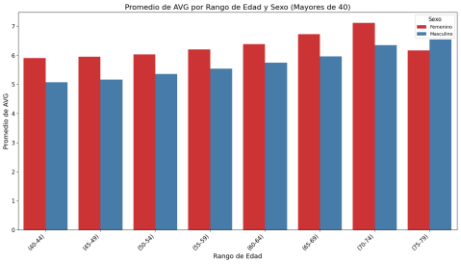
✓ Tiempo de llegada por Rango de edad, el gráfico confirma la tendencia esperada de que los tiempos promedio aumentan a medida que los corredores avanzan en edad. Sin embargo, las diferencias entre los grupos de edad son útiles para comprender mejor las capacidades físicas de los participantes y cómo se distribuyen los rendimientos en la carrera.





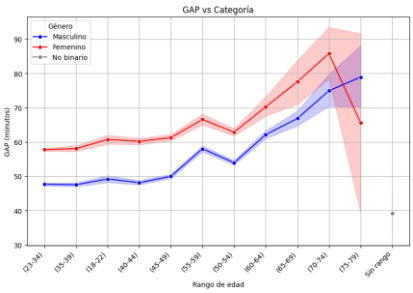
✓ Relación del AVG por categoría y género, se analizó el **tiempo promedio por kilómetro (AVG)** de los corredores, tomando en cuenta tanto su categoría de edad como su género. Este análisis permite identificar patrones en el rendimiento de los corredores en términos de eficiencia, ya que compara el tiempo que cada rango de edad tarda en completar cada kilómetro de la media maratón, en la grafica se tiene el promedio de todas las categorías y las mayores de 40.

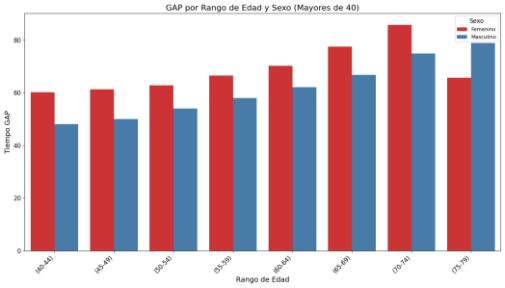




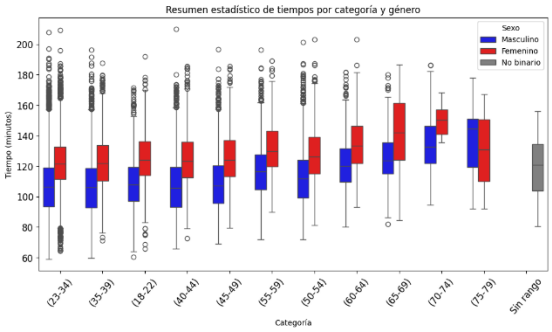
✓ Relación del GAP por categoría y género, se puede ver la variabilidad en los tiempos dentro de cada grupo de edad. En general, las categorías más jóvenes tienen un GAP más pequeño, mientras que las categorías de mayor edad tienen

un GAP más grande debido a la disminución en el rendimiento, en las grafica se tiene el gap de todas las categorías y las mayores de 40



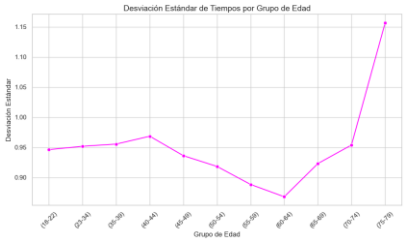


- ✓ Dispersión de los tiempos y detectar valores atípicos entre grupos, la dispersión de los tiempos muestra cómo se distribuyen los tiempos de llegada dentro de cada grupo, permitiendo identificar la variabilidad en el rendimiento de los corredores. Este análisis de dispersión y la identificación de valores atípicos permite comprender mejor los patrones en los tiempos de llegada, la comparación entre los diferentes grupos permite obtener una imagen más clara de cómo varían los rendimientos.



Se realizo la desviación estándar con respecto a los tiempos promedios, La desviación estándar nos muestra que los corredores más jóvenes e intermedios tienen tiempos de carrera más parecidos, lo que sugiere que están más uniformes los tiempo o estado físico similar.

En los corredores mayores, la variabilidad en los tiempos aumenta, probablemente porque la edad influye de forma diferente en cada persona según su condición física o experiencia.



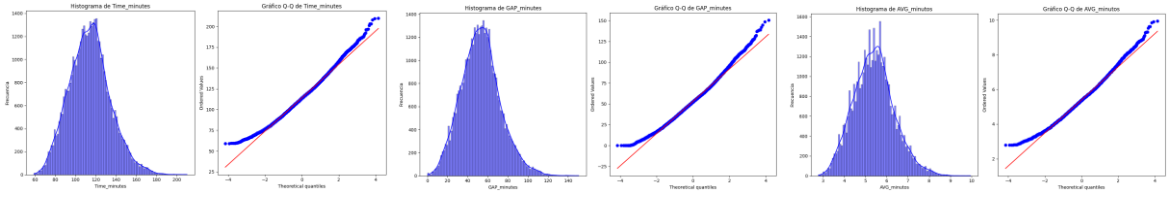
Para probar la hipótesis primero es importante evaluar la normalidad de los datos, ya que esta influye en la elección del método de correlación adecuado

TEST DE NORMALIDAD

Para determinar si mi conjunto de datos sigue una distribución normal se realizó la prueba de distribución de Gauss o campana de Gauss).

Con este Histograma KDE (Kernel Density Estimate), se puede observar cómo están distribuidos los valores de las variables, y visualizar la frecuencia exacta y el KDE para estimar la densidad de probabilidad de los datos, en las gráficas se observó forma de campana para las tres variables, sin embargo, en una de ellas hay ligeras asimetrías o picos, lo que podría indicar que hay más datos en ciertos rangos de tiempo, sin embargo en este caso se puede decir que los tiempos se distribuyen normalmente.,

De igual forma se compara la distribución de un conjunto de datos con una distribución teórica, lo que se confirma por la simetría en el histograma, que los puntos en el grafico Q-Q (Quantile-Quantile) se observan alineados cerca de la línea roja, lo cual se puede decir que los datos siguen una distribución normal.



Debido a que la prueba anterior no es 10% confiable realice **la Prueba de Anderson-Darling**

Prueba de Anderson-Darling para la columna: Time_minutes

Estadístico A-D: 3.662, Nivel de significancia 15.0%: 0.576 (Valor crítico),Nivel de significancia 10.0%: 0.655 (Valor crítico), Nivel de significancia 5.0%: 0.786 (Valor crítico), Nivel de significancia 2.5%: 0.917 (Valor crítico),Nivel de significancia 1.0%: 1.091 (Valor crítico)

Dado que el estadístico A-D es mucho mayor que los valores críticos, podemos concluir que **los datos de Time_minutes, no siguen una distribución normal**.

Prueba de Anderson-Darling para la columna: GAP_minutes

Estadístico A-D: 3.961, Nivel de significancia 15.0%: 0.576 (Valor crítico),Nivel de significancia 10.0%: 0.655 (Valor crítico), Nivel de significancia 5.0%: 0.786 (Valor crítico), Nivel de significancia 2.5%: 0.917 (Valor crítico), Nivel de significancia 1.0%: 1.091 (Valor crítico) La muestra NO sigue una distribución normal.

Prueba de Anderson-Darling para la columna: AVG_minutos

Estadístico A-D: 3.724, Nivel de significancia 15.0%: 0.576 (Valor crítico),Nivel de significancia 10.0%: 0.655 (Valor crítico), Nivel de significancia 5.0%: 0.786 (Valor crítico), Nivel de significancia 2.5%: 0.917 (Valor crítico), Nivel de significancia 1.0%: 1.091 (Valor crítico)

La muestra NO sigue una distribución normal.

2. Test Estadístico

Debido a que mis valores no siguen una distribución normal, realice el TEST de Mann-Whitney, para así determinar mi hipótesis.

La hice para valores estadísticamente iguales

M1(mujeres mayores de 40)= M2(hombres mayores de 40)

Hipótesis

Ho: No hay una diferencia significativa entre hombres y mujeres

H1: Si hay una diferencia significativa entre hombres y mujeres

Me rechazo la hipótesis nula, Estadístico U: 38511499.0, P-valor: 0.0, esto significa que hay una diferencia estadísticamente significativa entre los tiempos de los hombres con respecto a las mujeres

De igual forma para valores estadísticamente mayores, en la media del ritmo de carrera

M1(mujeres mayores de 40) > M2(hombres mayores de 40)

Hipótesis

Ho: No hay una diferencia

H1: Si hay una diferencia significativa entre hombres y mujeres

Me rechazo la hipótesis nula, Estadístico U: 38511499.0, P-valor: 0.0, esto significa que hay una diferencia estadísticamente significativa entre los tiempos de los hombres con respecto a las mujeres.

Conclusiones

El análisis muestra una tendencia conforme aumenta la edad, el tiempo promedio de finalización incrementa, indicando un aumento en los tiempos de carrera

Las categorías de mayores de 40 años tienen un tiempo promedio mayor comparado con los grupos más jóvenes.

La pérdida respecto a los tiempos de carrera con la edad afecta tanto a hombres como a mujeres, aunque las mujeres suelen mostrar una menor variación en tiempos promedio entre categorías en comparación con los hombres.

Bibliografía

- Pagina de media maratón en Barcelona https://edreamsmitjabarcelona.com/resultados/

- Fisiología del corredor 2024 - Sternitz.es https://sternitz.es/running/fisiologia-del-corredor/efectos-de-la-edad-en-la-fisiologia-del-corredor-cambios-y-adaptaciones/

- 17 Statistical hypothesis tests in python (cheat sheet) by Jason Brownlee on (2021 7 November).Machine Learning Mastery. Recuperado 10 de Noviembre 2024 in Statistic. https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

- Estadística para Dummies. Deborah Rumsey 2013