Análisis de Datos de Running – Memoria del Proyecto

(Asier Rodríguez – Data Analysis / EDA completo)

---

1. Introducción

En este proyecto he analizado varios datasets de entrenamientos de running procedentes de diferentes fuentes (Garmin, Apple y Kaggle).

El objetivo principal es entender **cómo ciertos factores influyen en la eficiencia al correr**, donde defino eficiencia como:

Para facilitar la comprensión del análisis, se incluyen a continuación algunas definiciones de conceptos habituales en el ámbito del running que son relevantes para este proyecto:

Ritmo (pace):
Tiempo que se tarda en recorrer un kilómetro, expresado en minutos por kilómetro. Un valor menor indica mayor velocidad (por ejemplo, 5:00 min/km es más rápido que 6:00 min/km).

Cadencia (cadence):
Número de pasos por minuto (spm). Es un indicador técnico importante; cadencias más altas suelen estar asociadas a pasos más cortos y una mecánica más eficiente. Algunos dispositivos registran doble cadencia, por lo que se divide entre dos para obtener el valor real.

Longitud de zancada (stride length):
Distancia media recorrida en cada zancada. Suele oscilar entre 0,8 y 1,4 metros en corredores recreativos. Una zancada mayor, manteniendo buena técnica, tiende a correlacionarse con mejor eficiencia.

Volumen semanal:
Cantidad total de kilómetros recorridos en una semana. Se utiliza para estudiar la acumulación de carga y su impacto en la eficiencia futura.

Eficiencia (definida para este proyecto):
En este análisis se define como:
`efficiency = pace_s_km × avg_hr`
Donde `pace_s_km` es el ritmo expresado en segundos por kilómetro y `avg_hr` es la frecuencia cardíaca media. Un valor menor implica mayor eficiencia, ya que refleja la capacidad de mantener buen ritmo con menor esfuerzo cardíaco.

Además, quería comprobar varias hipótesis personales basadas en mi experiencia como corredor:

1. **Existe un rango óptimo de cadencia que maximiza la eficiencia.**
2. **El rendimiento es mejor por la mañana que por la tarde o la noche.**
3. **Las semanas con más kilómetros correlacionan con mejor eficiencia la semana siguiente.**
4. **Una zancada más larga suele asociarse con una eficiencia mayor.**

Para validarlas, he limpiado, transformado y fusionado múltiples datasets, generando después un EDA lo bastante sólido como para responder a cada hipótesis con datos reales.

---

2. Fuentes de datos utilizadas

Trabajé con cinco datasets diferentes:

Asier

Mis propios entrenamientos exportados desde Garmin.

Ari

Entrenamientos de otra corredora exportados también de Garmin. Es el que más me ha costado limpiar, venía en formato JSON y con un formato numérico diferente.

Garmin (Kaggle)

Dataset más grande, con cientos de entrenamientos similares al estilo de Garmin, cogidos de Kaggle.

Log (Kaggle)

Dataset con muchísimos registros; aporta variedad y volumen, cogidos de Kaggle.

Ganix (descartado)

Aunque intenté incluirlo, su dataset **no tenía variables esenciales** para este análisis:

- No tenía **elevation gain** → imposible separar trail de ruta.
- No tenía **stride length** → afecta directamente a la hipótesis 4.
- No tenía **hora del día** → afecta a la hipótesis 2.

Por tanto, aunque limpié parte del dataset, finalmente decidí **excluirlo del merge final** por insuficiencia de datos relevantes.

---

3. Proceso de Limpieza de Datos (Data Cleaning)

La parte más larga del proyecto fue la limpieza. Cada dataset venía con su propio formato, columnas distintas, unidades diferentes e incluso variables mal medidas.

3.1 Normalización de columnas

- Pasé todos los nombres a **snake_case**.
- Unifiqué nombres entre datasets para poder mergearlos sin problemas.

3.2 Conversión de formatos

Incluye:

- Tiempos como `"hh:mm:ss"` o `"mm:ss"` → **segundos** con función propia.

- Distancia en metros → **km**.
- Elevación como `"1,054"` → conversión usando `str.replace` + cast a float.
- Fechas → `datetime`.

3.3 Limpieza de valores incorrectos

- `"--"` → `NaN`.
- Eliminación de registros sin HR o sin cadencia.
- Eliminación de entrenamientos muy cortos (<2 km).
- Filtrado de ritmos imposibles (<2 min/km o >8.5 min/km).

3.4 Creación de nuevas variables

- `efficiency`
- `pace_min_km`
- `time_of_day` → función propia (mañana/tarde/noche).
- `week_year` para análisis semanal
- Bins de cadencia y zancada para heatmaps
- `date_num` → ordinal para LOWESS

3.5 Merge final

Fusioné **Asier**, **Ari**, **Garmin** y **Log**. El dataset final quedó homogéneo y útil para el análisis.
Problema encontrado:

- Aparición de columnas `runner_x` y `runner_y`
  → solucionado renombrando antes de hacer merge.

---

4. Exploratory Data Analysis (EDA)

Una vez tuve el dataset limpio, empecé con el análisis exploratorio.

4.1 Análisis univariante

- Distribución de ritmos
- Distribución de HR
- Distribución de cadencia
- Detección de outliers en zancada
- Variación de eficiencia por corredor

4.2 Análisis bivariante

- Correlaciones entre cadencia, HR, distancia, zancada y eficiencia.

- Regresiones LOWESS para ver tendencias no lineales.

4.3 Análisis multivariante

- Heatmap de **cadencia** × **zancada** para ver zonas de máxima eficiencia.

---

5. Análisis por hipótesis

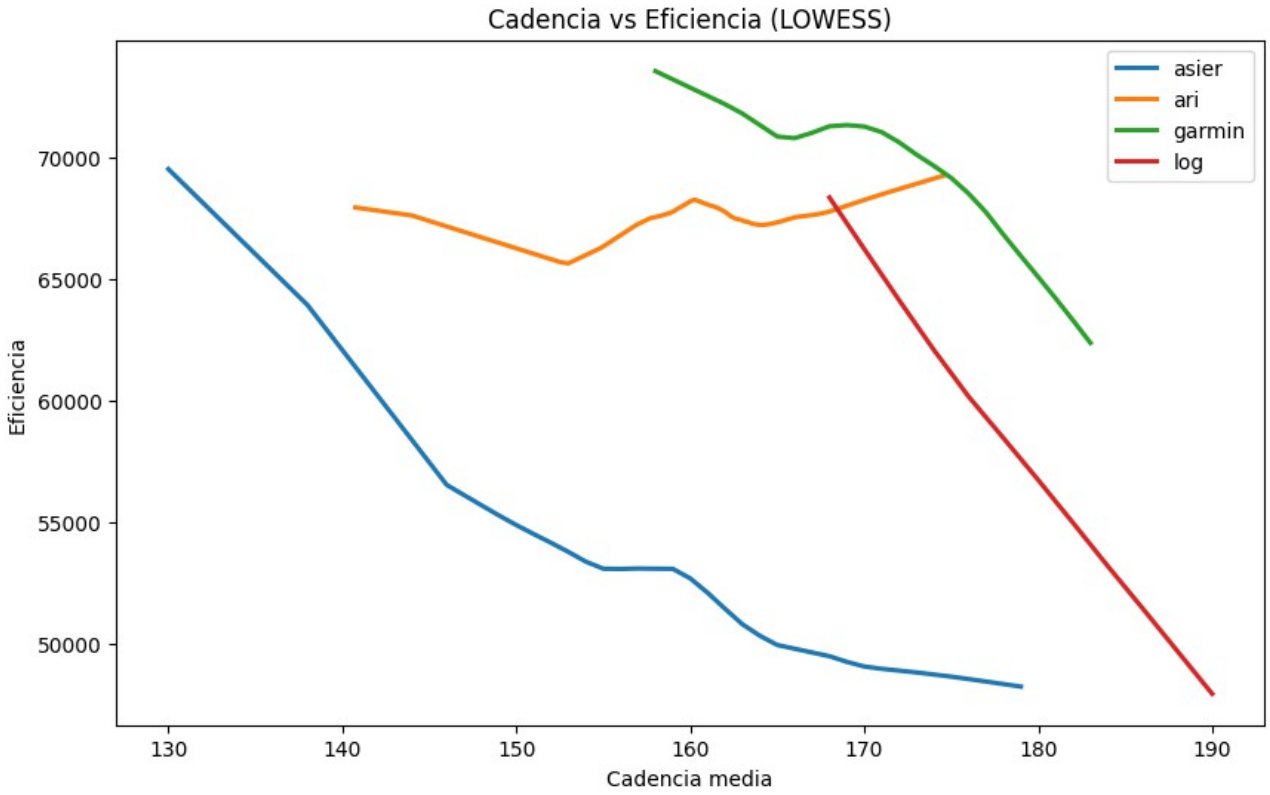
Aquí está lo más interesante del proyecto: comprobar si lo que intuitivamente creemos… **es cierto**.

Hipótesis 1: Existe un rango óptimo de cadencia

Gráfica usada: Cadencia vs Eficiencia (LOWESS) por corredor.

Conclusiones:

En todos los corredores, la eficiencia mejora de forma casi lineal conforme aumenta la cadencia. La hipótesis original planteaba un "rango óptimo", pero los datos indican que dentro del rango observado **más cadencia implica más eficiencia**, sin un límite evidente.



Hipótesis 2: Rendir mejor por la mañana

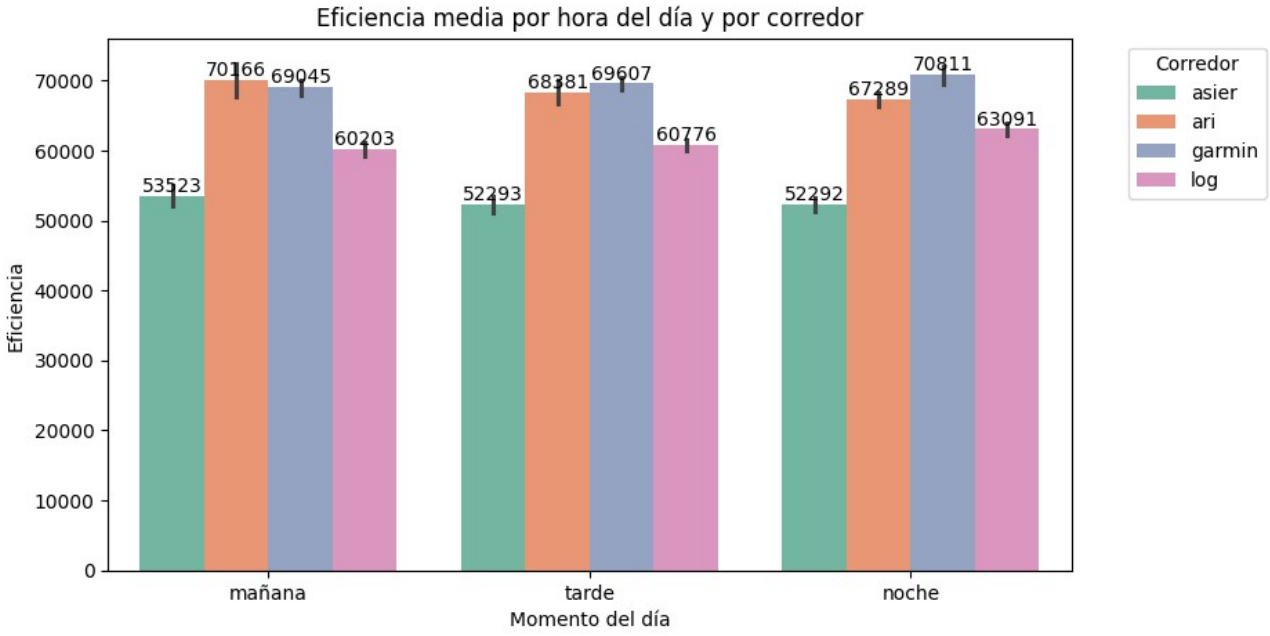
Gráficas utilizadas:

- Eficiencia media por momento del día.
- Eficiencia por momento del día *por corredor.*

Resultados:

- **Asier** rinde claramente mejor **por la mañana**.
- Ari rinde mejor **por la tarde**.
- El dataset Garmin rinde mejor **por la noche**.

Hipótesis parcialmente confirmada. El momento óptimo es **personal y no universal**.



Hipótesis 3: Relación entre km semanales y eficiencia

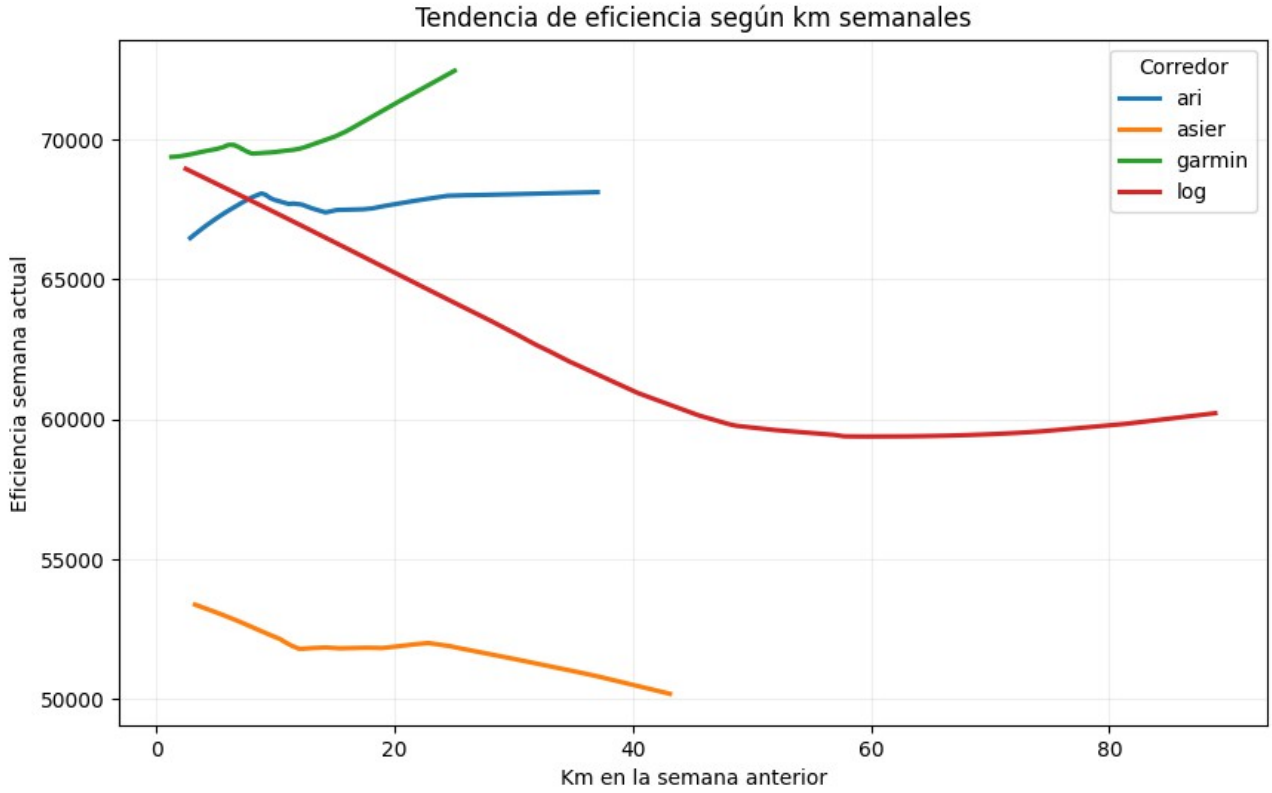
Gráfica:

- Km de la semana anterior vs eficiencia actual.

Resultados:

- En mi caso (Asier), la tendencia es muy clara:
  más km → mejor eficiencia.
- En Ari y Garmin se observa lo contrario, seguramente por fatiga.
- El dataset Log tiene mucha variabilidad, demasiado ruido y dispersión.

La hipótesis **se cumple en mi caso (Asier) y en log hasta ciertos kilometros**.
En otros corredores ocurre lo contrario, lo cual demuestra que el impacto del volumen semanal es **altamente individual** y está condicionado por la tolerancia al entrenamiento y la gestión de la fatiga.



Hipótesis 4: Mayor zancada = mejor eficiencia

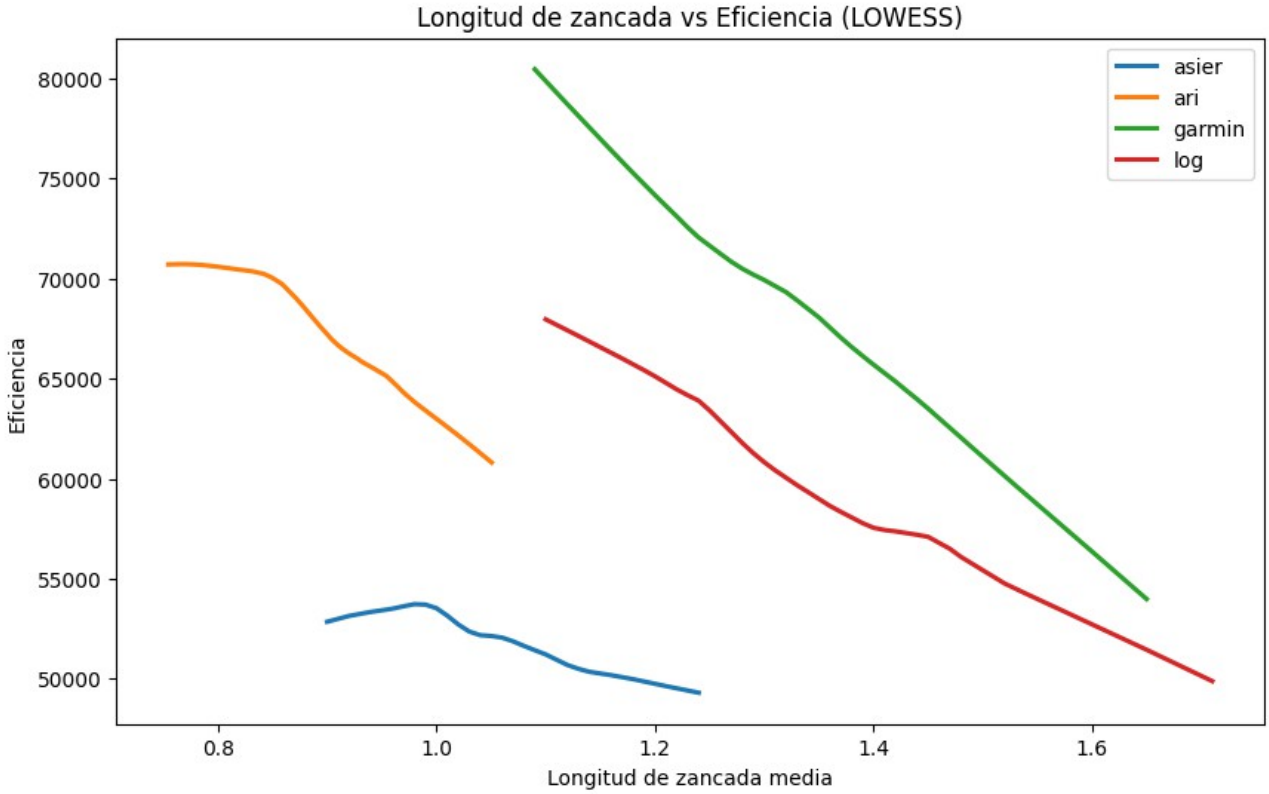
Gráficas:

- Heatmap cadencia × zancada × eficiencia
- LOWESS longitud zancada vs eficiencia

Resultados:

- A mayor zancada, mejor eficiencia.
- La tendencia es **robusta** y se mantiene en todos los datasets.
- El heatmap muestra claramente “zonas óptimas”.

Hipótesis confirmada con fuerza.



---

6. Dificultades encontradas

Durante el proyecto me encontré con varios retos interesantes:

Formatos completamente diferentes entre datasets

Cada corredor exportaba datos de forma distinta.

Tipo numérico desigual entre datasets

Diferencias en los tipos de datos entre datasets (strings, enteros, floats) que obligaron a castear las columnas antes de fusionar.

Tiempos mal formateados (“--”, “3:5”, “1,054”, etc.)

Requirió funciones personalizadas.

Datos incompletos en algunos datasets

El caso más significativo fue **Ganix**, que tuve que descartar.

Clasificación de horarios

Tuve que construir mi propia función mañana/tarde/noche.

Unión de datasets grandes

Aparecían columnas duplicadas o de distinto tipo.

---

7. Conclusiones

Después del análisis, puedo decir que:

- **Las hipótesis se cumplen parcialmente**, dependiendo del corredor y de la variable analizada.
- **La cadencia no tiene un punto óptimo**, sino que en todos los casos **más cadencia implica mejor eficiencia** dentro del rango analizado.
- **La zancada es el predictor más fuerte de eficiencia**: cuanto mayor es, mejor rendimiento.
- **El momento del día influye**, pero de manera completamente **individual** en cada corredor.
- **Los km semanales no son un buen indicador universal**:
    - En mi caso (Asier) sí mejoran la eficiencia,
    - pero en Ari y Garmin ocurre lo contrario, probablemente por **fatiga**.

Además:
Este proyecto demuestra que un EDA real exige **limpieza de datos exhaustiva**, decisiones justificadas y una integración cuidadosa entre datasets muy distintos.
La calidad del dato es clave para poder sacar conclusiones fiables.

---

8. Trabajo futuro

Si continúo el proyecto, estas serían las líneas clave para ampliarlo:

- Integrar **datos meteorológicos** (temperatura, viento, humedad).
- Normalizar la eficiencia según **tipo de sesión** (easy, series, long run, trail…).
- Explorar **modelos predictivos** para estimar eficiencia o ritmo futuro.
- Comparar muchos más corredores.

---

9. Referencias

- Garmin Connect
- Apple
- Kaggle – Dataset *log y Garmin*
- Librerías: pandas, numpy, seaborn, matplotlib
- Statsmodels (LOWESS)