**Análisis de datos sobre entrenamiento en la población 🏋️**

🪧**Descripción del proyecto:**

Este proyecto realiza un análisis exploratorio acerca de un set de datos sobre el entrenamiento en la población general.

Este estudio, tiene como objetivo el análisis de las diferencias en los hábitos de entrenamiento en la población general así como según el género de la persona que lo realiza. Además, también se estudiará qué tipos de entrenamientos tienen un mayor impacto en la salud de las personas midiendo factores como el pulso en reposo, las calorías quemadas o las emociones después del entrenamiento.

Para llevar a cabo el estudio se utilizarán técnicas básicas para el análisis de datos y se creará un Dashboard para obtener una representación visual de los cálculos.

**🏛️Estructura del proyecto:**

Para llevar a cabo el análisis de los datos se ha utilizado la siguiente estructura de carpetas.

l—Data #Datos crudos

l—Análisis #Documento con el EDA realizado y los diferentes gráficos obtenidos así como el Dashboard

l—Documentación #Readme con la descripción del proyecto y análisis paso a paso.

**🧰Instalaciones y requisitos:**

Para este proyecto se ha utilizado como herramienta tanto para el análisis como para la creación del Dashboard, Microsoft Excel. Por lo que será la única herramienta necesaria para poder tener acceso al estudio.

🧪**Análisis y creación del Dashboard:**

***Paso 1:*** Se hace una carga en una nueva hoja excel del set de datos workout\_fitness\_tracker\_data. Para hacer el import de dicho set de datos se utiliza Data → Get Data → From file → From Csv

**1. Transformación y limpieza de los datos**

***Paso 2:*** Se convierten los datos a formato tabla para facilitar su análisis.

***Paso 3:*** Se realiza un análisis de las categorías que representa cada una de las variables que contienen información para saber de qué hablan y que estén correctamente categorizadas en excel.

* Se convierte a números sin decimales todas las columnas que contienen números.
* Se convierte la variable **Age (Columna B)** en un rango en la columna C llamada Rango de edad.
* Se comprueba que no haya registros en blanco en ninguna de las columnas con categorías de datos y se extraen las primeras conclusiones a simple vista:
* No hay registros en blanco en ninguna de las columnas de datos de los registros.
* La columna **V02 Max (Columna Q)** tiene el mismo dato para todos los registros y como es poco probable que esto sea así en la realidad no se tendrá en cuenta para el análisis.
* La columna **Body fat (%) (Columna R)** tiene el mismo dato para todos los registros y como es poco probable que esto sea así en la realidad no se tendrá en cuenta para el análisis.
* La columna **Water intake (Columna N)** tiene el mismo dato para todos los registros y como es poco probable que esto sea así en la realidad no se tendrá en cuenta para el análisis.
* Una vez comprendido a nivel general todos los datos y hemos descartado los datos que van a ser inservibles para nosotros, vamos a analizar en mayor profundidad las columnas que contienen datos que impactan en este estudio:

**User ID (Columna A):**

Representa el ID unívoco de cada uno de los registros de las personas que respondieron a la encuesta.

* Tipo de variable → Variable cuantitativa discreta.
* Rango de datos: 0-10000
* Tipo de dato: Numerico

**Age (Columna B):**

Esta variable representa la edad de la persona que responde a la encuesta.

* Tipo de variable → Variable cuantitativa discreta
* Rango de datos: 18 - 59 años
* Tipo de dato: Numerico

**Gender (Columna C):**

La columna género representa el género de la persona que responde a la encuesta.o.

* Tipo de variable → Variable cualitativa nominal
* Rango de datos: Hombre, Mujer, Otro.
* Tipo de dato: Categórico

**Height (cm) (Columna D)**

Esta variable representa la altura de la persona que responde a la encuesta.

* Tipo de variable → Variable cuantitativa continua
* Rango de datos → 150-199 cm
* Tipo de dato: Numerico

**Weight (kg) (Columna E)**

Esta variable representa el peso de la persona que responde a la encuesta.

* Tipo de variable → Variable cuantitativa continua
* Rango de datos → 50-119 kg
* Tipo de dato: Numerico

**Workout type (Columna F)**

Esta variable representa el tipo de entrenamiento que realiza la persona que responde a la encuesta.

* Tipo de variable → Variable cualitativa nominal
* Rango de datos → Cycling, Cardio, Strength, Yoga, Running, HIIT.
* Tipo de dato: Categórico

**Workout duration (Columna G):**

Esta variable representa la cantidad de tiempo en minutos que invierte en su entrenamiento la persona que responde a la encuesta:

* Tipo de variable → Variable cuantitativa continua
* Rango de datos: 10-119 minutos
* Tipo de dato: Numérico

**Calories burned (Columna H)**

Esta variable representa la cantidad de calorías que ha quemad la persona que responde en su sesión de entrenamiento:

* Tipo de variable → Variable cuantitativa continua
* Rango de datos: 100 - 999 calorias quemadas
* Tipo de dato: Numérico

**Heart Rate (bpm) (Columna I)**

Esta variable representa el ritmo cardiaco medio de la persona que responde a la encuesta en su entrenamiento:

* Tipo de variable → Variable cuantitativa continua
* Rango de datos: 80 -179 bpm
* Tipo de dato: Numérico

**Steps taken (Columna J)**

Esta variable representa la cantidad de pasos que ha dado la persona que responde a la encuesta en su entrenamiento:

* Tipo de variable → Variable cuantitativa continua
* Rango de datos: 1000 - 19998 pasos
* Tipo de dato: Numérico

Esta columna es probable que incluya outliers y como la variable no afecta a nuestro estudio se excluirá del mismo.

**Distance (Km) (Columna K)**

Esta variable representa la cantidad de kilometros que ha dado la persona que responde a la encuesta en su entrenamiento:

* Tipo de variable → Variable cuantitativa continua
* Rango de datos: 5 - 1499 kilometros
* Tipo de dato: Numérico

Al analizar esta variable se considera imposible que una persona recorra más de 500 kilómetros en casos extremos por lo que esta variable no será considerada tampoco en este estudio. Además es probable que en esta columna haya outliers.

**Workout intensity (Columna L)**

Esta variable representa la intensidad del entrenamiento de la persona que responde a la encuesta en su entrenamiento

* Tipo de variable → Variable cualitativa nominal
* Rango de datos: High, Medium, Low
* Tipo de dato: Categórico

**Sleep hours (Columna M)**

Esta variable representa la cantidad de horas que duerme de media la persona que responde a la encuesta en su entrenamiento.

* Tipo de variable → Variable cuantitativa continua
* Rango de datos: 40 - 100
* Tipo de dato: Numérico

Esta variable está mal descrita ya que no nos especifica la frecuencia en la que se produce ese registro es decir, semanal, mensual… Al observar esto, decidimos excluir también esta variable de nuestro estudio.

**Daily calories intake (Columna O)**

Esta variable representa la cantidad de caloría que consume de media al dia la persona que responde a la encuesta en su entrenamiento.

* Tipo de variable → Variable cuantitativa continua
* Rango de datos: 1500 - 3999 calorias
* Tipo de dato: Numérico

En esta variable es posible que haya Outliers es por eso por lo que tendremos que tenerlo en cuenta a la hora de realizar nuestros cálculos.

**Resting Heart Rate (bpm) (Columna P)**

Esta variable representa el ritmo cardiaco medio en reposo de la persona que responde a la encuesta en su entrenamiento.

* Tipo de variable → Variable cuantitativa continua
* Rango de datos: 50 - 89 bpm
* Tipo de dato: Numérico

**Mood before workout (Columna S)**

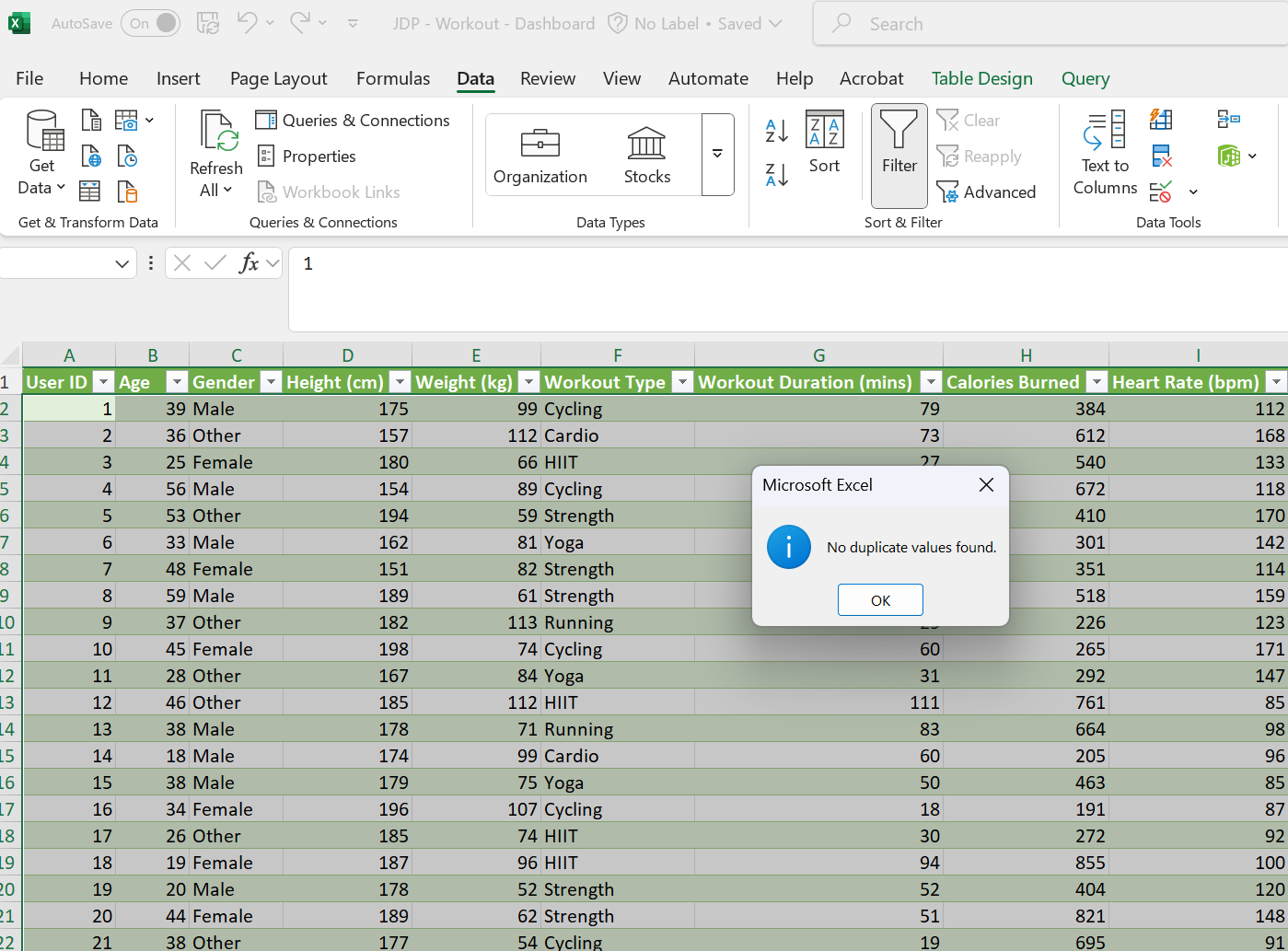
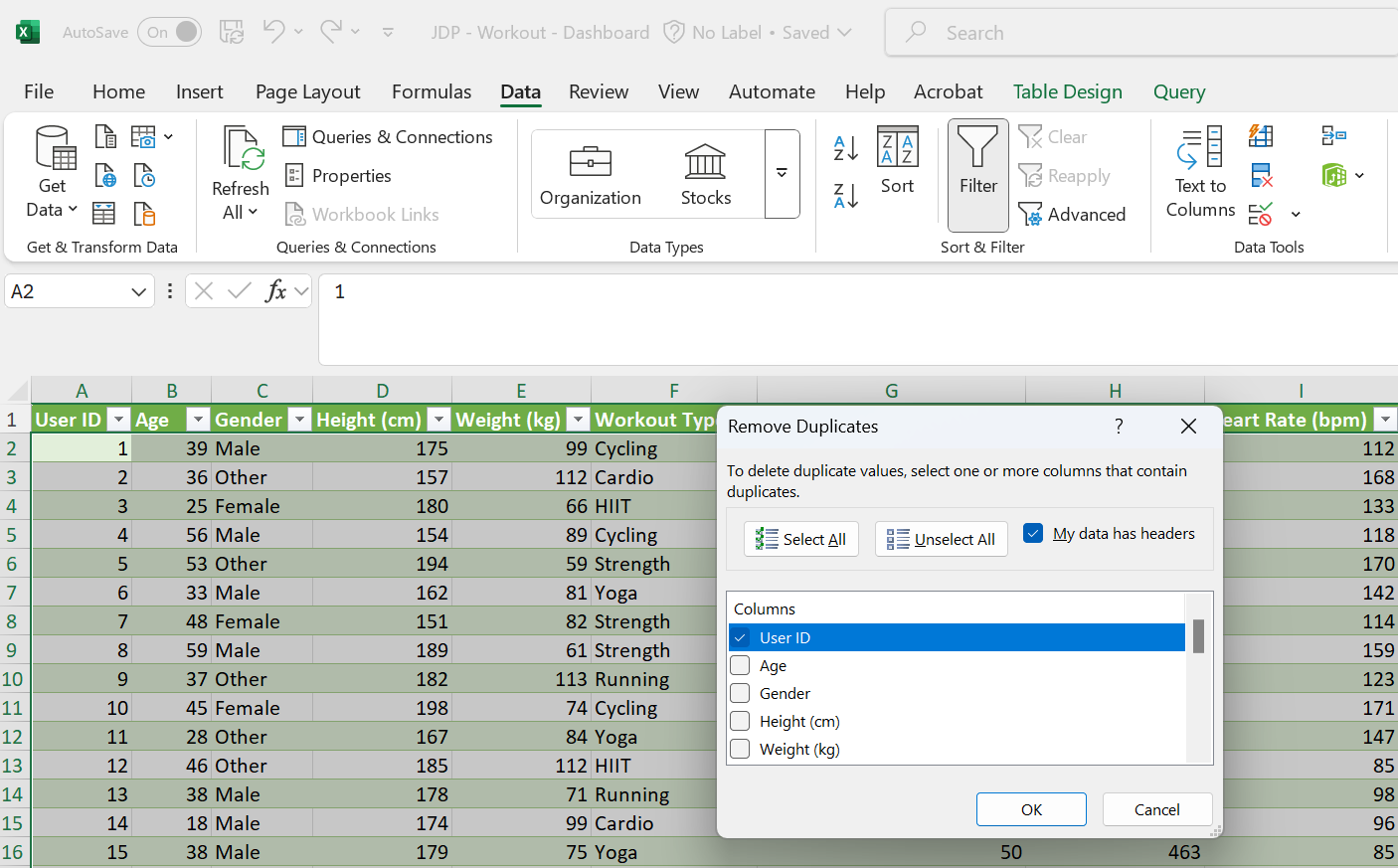
Esta variable representa el estado de ánimo antes del entrenamiento de la persona que responde a la encuesta en su entrenamiento.

* Tipo de variable → Variable cualitativa nominal
* Rango de datos: Happy, Neutral, Stressed, Tired
* Tipo de dato: Categórico

**Mood after workout (Columna T)**

Esta variable representa el estado de ánimo después del entrenamiento de la persona que responde a la encuesta en su entrenamiento.

* Tipo de variable → Variable cualitativa nominal
* Rango de datos: Happy, Neutral, Stressed, Tired
* Tipo de dato: Categórico

***Paso 4:*** Se analiza la columna “User ID” utilizando la función eliminar duplicados para comprobar que no haya ningún registro duplicado.

Tras comprobar que no hay valores duplicados ni outliers en nuestros registros, podemos proseguir con el análisis.

Ahora, seleccionaremos las variables que serán de utilidad en nuestro estudio que son las siguientes:

1. UserID
2. Gender
3. Workout type
4. Calories Burned
5. Daily Calories Intake
6. Resting Heart Rate (bpm)
7. Heart rate (bpm)
8. Workout duration

***Paso 5:*** Se confecciona la plantilla que se usará posteriormente para la creación del dashboard. (Ver como resultado final del dashboard)

***Paso 6:*** Se generan las tablas dinámicas y gráficas necesarias para poder llevar a cabo el análisis (Ver en la pestaña Análisis en el fichero: JDP - Workout - Dashboard

***Paso 7:*** Se añaden las gráficas obtenidas y los KPIs a nuestro Dashboard.

***Paso 8:*** Una vez obtenidos los gráficos, creamos los segmentadores necesarios y realizamos la conexión con el resto de tablas dinámicas de manera que cuando nosotros elijamos un filtro u otro este afecte a todas las gráficas.

**🔍Resultados y Conclusiones**

* Encontramos que el deporte preferido entre los hombres de este estudio aunque por un pequeño margen es el yoga, el de las mujeres el cycling y el de la gente que no especificó su género el entrenamiento de fuerza.
* En cuánto a la duración de los entrenamientos, observamos que los entrenamientos de cardio aunque de nuevo por un pequeño margen son los más duraderos.
* Respecto del estudio calórico realizado, observamos que en los deportes en los que hay una mayor demanda calórica los usuarios tienden a ingerir de media una mayor cantidad de calorías. La excepción a la regla es el running deporte que tiene de media una exigencia calórica menor al resto pero los deportistas que practican este deporte son los que más calorías consumen.
* Respecto del estudio del ritmo cardíaco de los deportistas es prácticamente imposible sacar ninguna conclusión ya que en cada deporte el corazón tiende a presentar un comportamiento muy similar presentado de base unas pulsaciones en reposo menores a las de la población entre la gente que practica estos deportes.
* Por último, la conclusión que obtenemos en el estudio emocional de los deportistas es que antes del entrenamiento un 24,25% de los participantes decía estar estresado y un 25,53% decía estar cansado. Al salir del entrenamiento los sujetos afirman estar energizados, fatigados o neutros por lo que podemos afirmar que el entrenamiento mejora el estado de ánimo general así como los niveles de energía.

🐾 **Proximos pasos**

Los datos utilizados en este estudio parecen presentar sesgos bastante grandes al mostrar todos una distribución bastante equitativa entre todas las posibles respuestas de todas las variables por lo que no recomiendo continuar con el estudio.

De continuarlo, mi recomendación sería empezar por realizar un estudio profundo de las variables para entender estos sesgos y las posibles correlaciones entre variables para obtener conclusiones más claras. También intentaría encontrar el sesgo en todos los datos y realizaría un estudio de las distribuciones probabilísticas de cada una de las variables.

🤝 **Contribuciones**

Las contribuciones al estudio son bienvenidas, para contribuir al estudio por favor observar las recomendaciones que se ofrecen en el apartado anterior.

✍️**Autores**

Jorge Donderis Pla ([jorgedonderis@gmail.com](mailto:jorgedonderis@gmail.com)) (<https://github.com/jdonderis>)