ANALISIS EXPLORATORIO DE DATOS

CONTEXTO

Swetro es una aplicación en la que los usuarios se registran para competir en retos deportivos de running, ciclismo y caminata, ya sea para ganar premios o simplemente por el espíritu de competencia. Para participar, los usuarios registran sus actividades mediante relojes inteligentes de marcas como Garmin, Suunto, Wahoo, IgpSport, Polar y Apple Watch, o a través de las apps móviles oficiales de estas marcas desde un dispositivo móvil.

Uno de los desafíos recurrentes que enfrenta Swetro es la detección de actividades sospechosas o incoherentes, las cuales pueden surgir por diversas razones, tales como:

- Errores en los sensores de los dispositivos, lo que genera registros incorrectos en métricas cómo la velocidad, distancia o ritmo cardíaco.
- Registros humanamente imposibles, como velocidades extremas o distancias cubiertas en tiempos irrealistas.
- Uso inadecuado del dispositivo, por ejemplo, dejarlo encendido todo el día sin realizar actividad física real.
- Intento de fraude, donde un usuario registra una actividad, pero en realidad realizó otra (ejemplo: marcar una caminata como un trote).
- Aprovechamiento de medios externos, como subirse a un vehículo o utilizar una bicicleta eléctrica para obtener mejores tiempos y distancias.

Por lo tanto, para evitar que estas irregularidades afecten la competencia justa y la validez de los premios, es necesario desarrollar un sistema de detección de actividades atípicas o sospechosas.

Para este estudio, se utilizará un dataset de los últimos 5 meses (octubre 2024 - febrero 2025) con registros de actividades deportivas:

Cada fila representa una actividad registrada por un usuario.

Cada columna corresponde a una variable o característica de la actividad, como distancia, tiempo, velocidad promedio, ritmo cardíaco promedio, elevación ganada, etc.

Dado que las irregularidades pueden variar según la disciplina deportiva, este análisis se centrará exclusivamente en los registros de running, buscando patrones que indiquen posibles fraudes o datos erróneos.

Descripción de los campos

Las variables que describen cada actividad son los siguientes:

- **UserId:** Identificador único del usuario en la base de datos de la empresa.
- **Type:** Tipo de actividad registrada (*Running, Cycling, Walking, Other*).
- **Name:** Nombre asignado a la actividad, ya sea por el usuario o automáticamente por la aplicación.
- **StartTimeUtc:** Fecha y hora de inicio de la actividad en UTC (*Tiempo Universal Coordinado*).
- **DurationInSeconds:** Duración total de la actividad medida en segundos.
- **DistanceInMeters:** Distancia recorrida durante la actividad, expresada en metros.
- **Steps:** Número total de pasos registrados durante la actividad.
- **AverageSpeedInMetersPerSecond:** Velocidad promedio alcanzada durante la actividad, expresada en metros por segundo.
- **AveragePaceInMinutesPerKilometer:** Ritmo promedio de la actividad, expresado en minutos por kilómetro.
- **TotalElevationGainInMeters:** Suma total de la elevación ganada en la actividad, expresada en metros.
- **TotalElevationLossInMeters:** Suma total de la elevación perdida en la actividad, expresada en metros.
- **AverageHeartRateInBeatsPerMinute:** Frecuencia cardíaca promedio durante la actividad, medida en latidos por minuto (*BPM*).
- **SourceType:** Marca del dispositivo con el cual se registró la actividad (*Garmin, Suunto, Wahoo, IgpSports, Polar, Apple Watch*).
- **SourceName:** Modelo del dispositivo o aplicación que registró la actividad (ejemplo: *Garmin Forerunner 945, fēnix 3 HR, etc*).
- **Warnings:** Indicadores de posibles anomalías detectadas en la actividad. Se generan alertas en los siguientes casos:
    1. Si la duración de la actividad es menor a 5 minutos.
    2. Si es una actividad de ciclismo y la elevación ganada supera los 16 metros por minuto (1,000 metros por hora).
    3. Si es una actividad de running y la elevación ganada supera los 8 metros por minuto (500 metros por hora).

4. Si es una actividad de running con un ritmo promedio menor a 3.5 minutos por kilómetro (3:30 min/km).
5. Si la actividad no registra distancia recorrida.

- **CreationTime:** Fecha y hora en la que el registro de la actividad fue creado en la base de datos.

Pregunta SMART

¿Es posible identificar actividades de running atípicas o sospechosas con datos de los últimos 5 meses (10/2024 - 02/2025) utilizando métricas cómo velocidad, distancia, elevación, tiempo y frecuencia cardiaca?

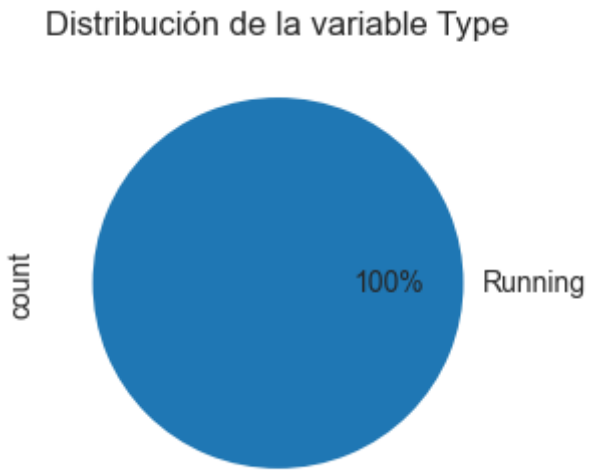
ANALISIS EXPLORATORIO

Se puede ver que hay datos vacíos, para las variables DistanceInMeters, Steps, AverageSpeedInMetersPerSeconds, AveragePaceInMinutesPerKilometer, TotalElevationGainInMeters,TotalElevationLossInMeters, AverageHeartRateInBeatsPerMinute, SourceName, ActiveKilocalories y Warnings.

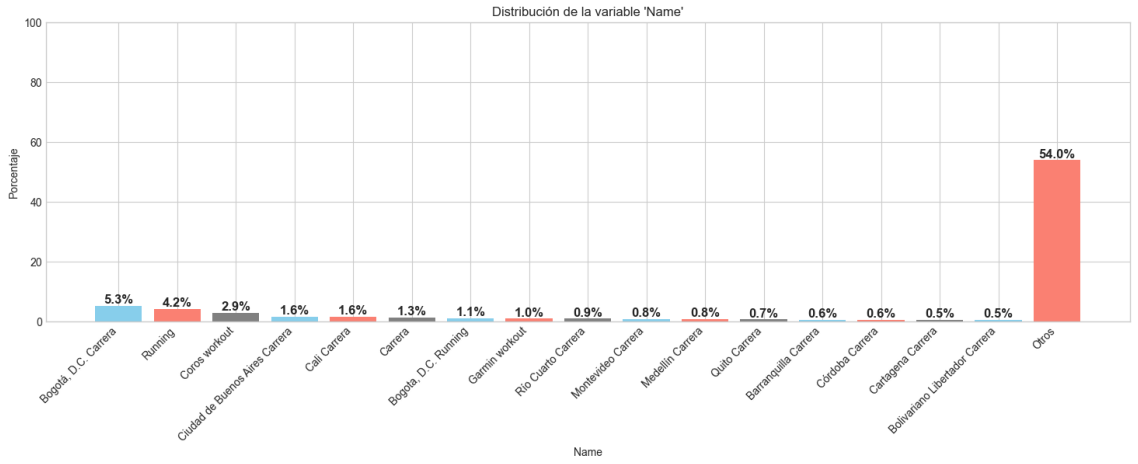
En el proceso de limpieza se decidirá qué hacer con estas observaciones.

Podemos observar que el dataset cuenta con las siguientes variables categóricas: Type, Name, StartTimeUtc, SourceType, SourceName, Warnings y CreationTime.

ANÁLISIS DE VARIABLES CATEGÓRICAS



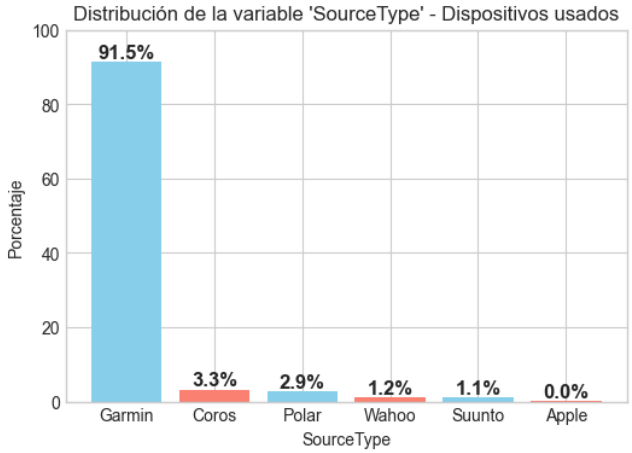
Debido a que Type representa el tipo de actividad deportiva y en el presente análisis sólo se tienen registros de Running, en el proceso de limpieza se eliminará está columna ya que no aporta información valiosa para el modelo.



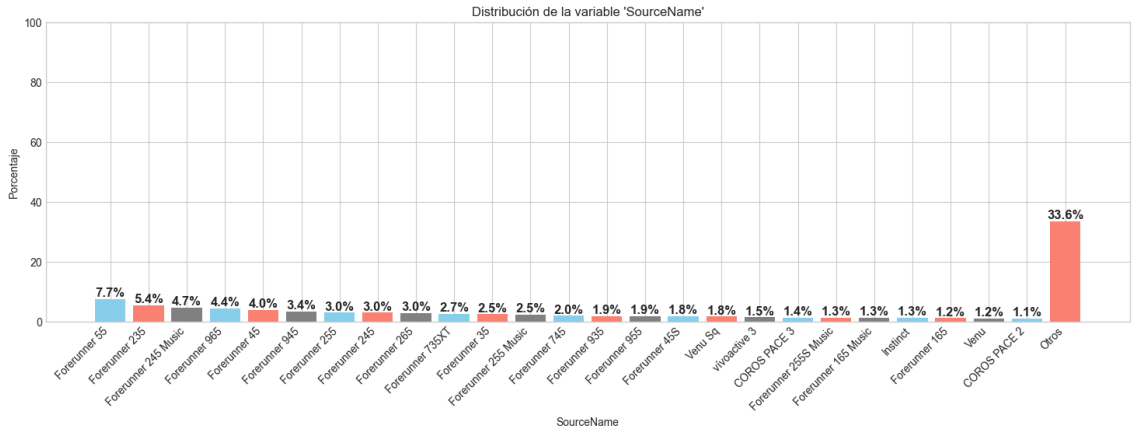
En la anterior gráfica se enseñan algunos de los nombres que representan al menos el 0.5% de los nombres encontrados, en cualquier otro caso se agregarían a la categoría de otros para poder mejorar la visualización de los datos.

Esta variable representa el nombre de la actividad que le coloca el reloj o el usuario manualmente, esta variable no es relevante para el análisis por lo que en el proceso de limpieza se eliminará.

La variable StartTimeUtc representa la fecha y hora en la cual se inició la actividad, en los siguientes pasos se podría transformar en nuevas columnas que representen cada valor por separado para poder realizar un mejor análisis de su valor.

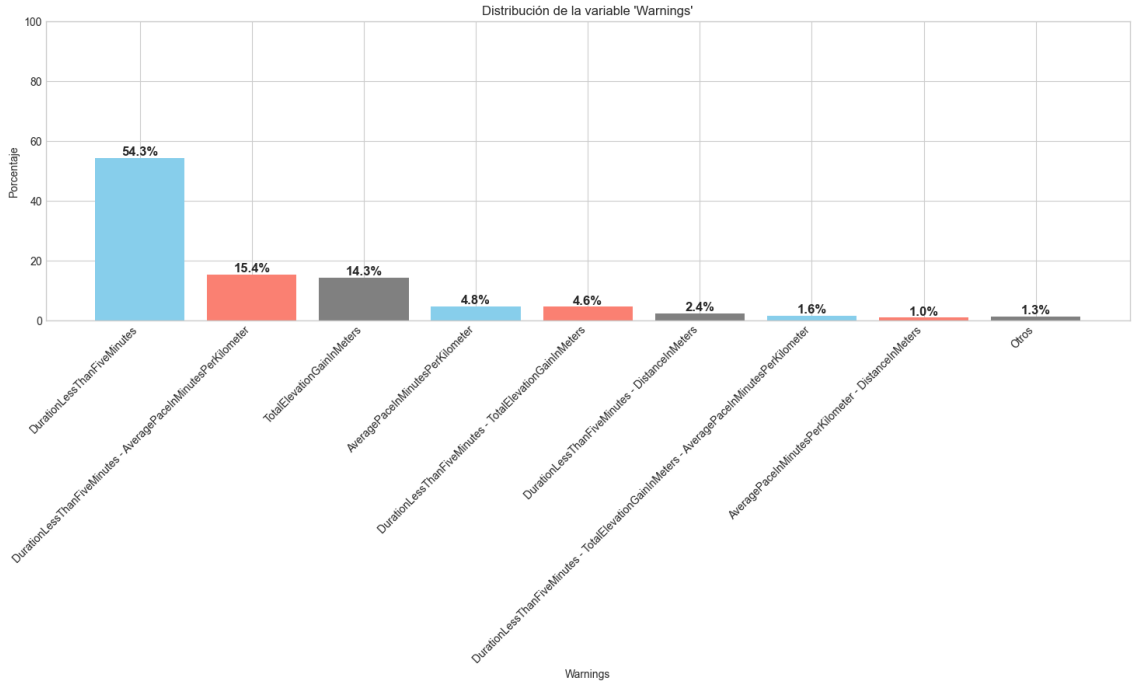


En la gráfica anterior, se observa que Garmin domina ampliamente como la marca más utilizada. Esto podría introducir sesgos en el modelo, ya que la distribución está fuertemente inclinada hacia esta marca. Por ello, durante el proceso de limpieza, evaluaremos qué acciones tomar con esta variable, dado que el dataset muestra un claro desbalance.



En la anterior gráfica se enseñan algunos de los modelos que representan al menos el 1% de los modelos encontrados, en cualquier otro caso se agregarían a la categoría de otros para poder mejorar la visualización de los datos.

En el proceso de limpieza se podría evaluar si se pudiera eliminar la variable debido a que ya se encuentra representada en la variable sourceType, o se podría realizar algún tratamiento para representar de otra forma.

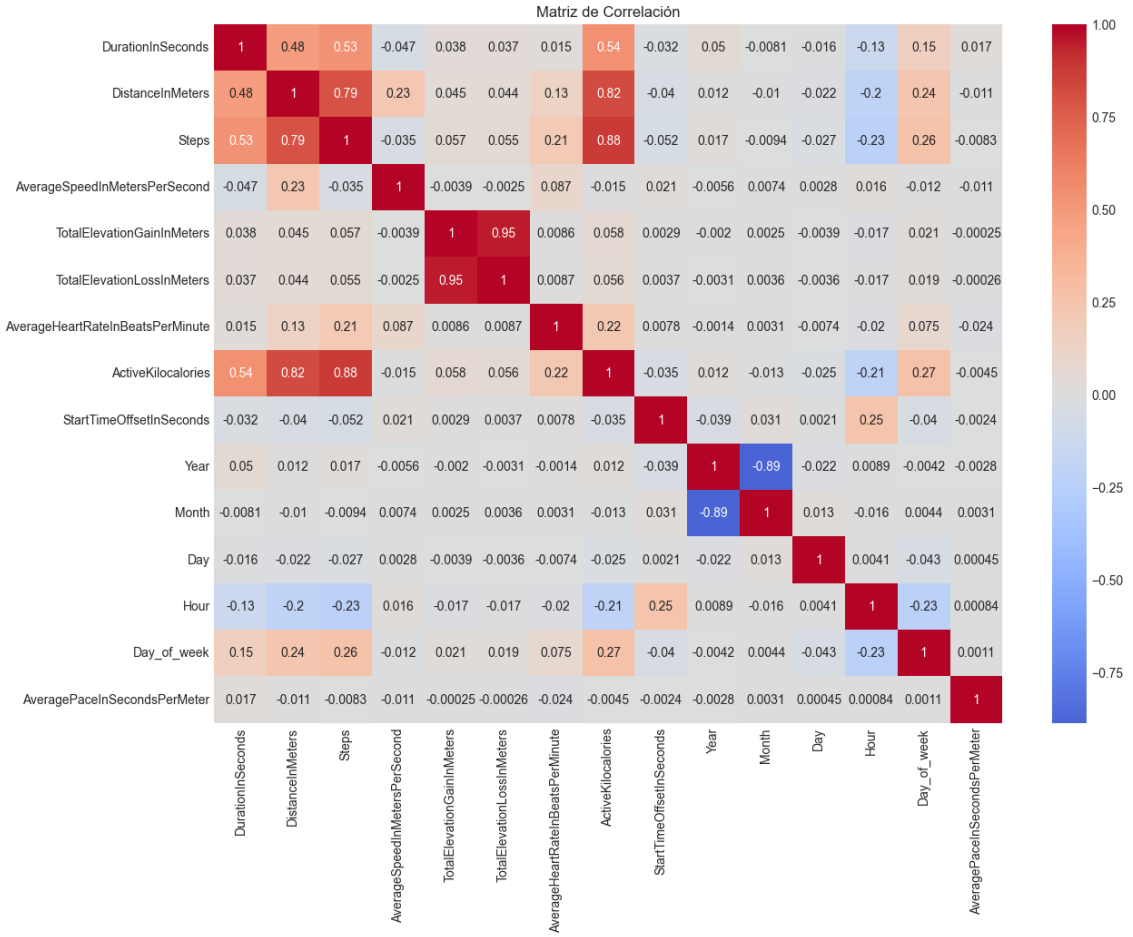


La variable warnings representa la clasificación manual que realiza la empresa para identificar datos atípicos basados en reglas del negocio, por esta razón en el proceso de limpieza se procederá a eliminar esta columna debido a que puede representar un data leakage y por lo tanto sesgaría el modelo.

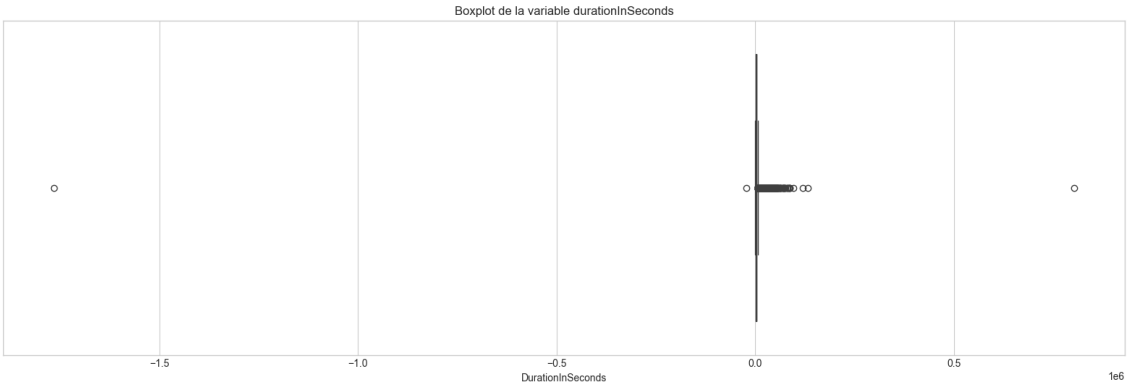
Al igual que la variable StartTimeUtc, la variable CreationTime representa una fecha; sin embargo, en este caso, indica la fecha y hora en que la actividad fue registrada

en la base de datos de la empresa. Dado que esta información no aporta valor al análisis, se eliminará durante el proceso de limpieza.

ANÁLISIS DE VARIABLES CONTINUAS

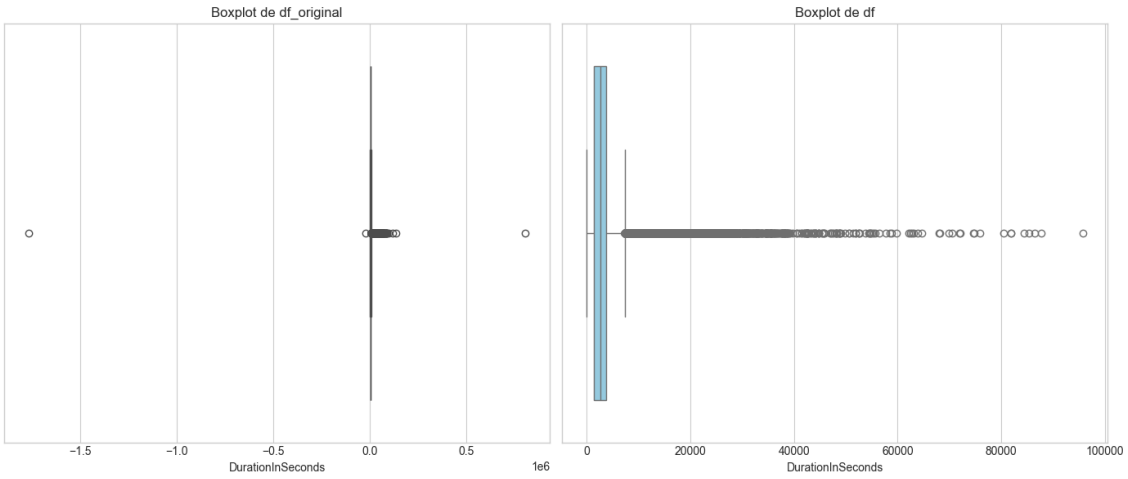


En la matriz de confusión se puede observar una alta correlación positiva entre las variables Steps y DistancesInMeters por lo que se puede apreciar una relación lineal. Además, se ve una correlación muy alta entre las variables TotalElevationGainInMeters y TotalElevationLossInMeters, está correlación alta puede mostrar una posible colinealidad debido a que ambas representan la misma variable, pero medida en perdida o ganancia por lo que se deberá realizar algún tratamiento para evitar la colinealidad. Además, también se puede observar correlación positiva, pero en menor medida entre Steps y DurationInSeconds, DistanceInMeters y DurationInSeconds. Las anteriores correlaciones positivas indican que si una variable aumenta la otra también.

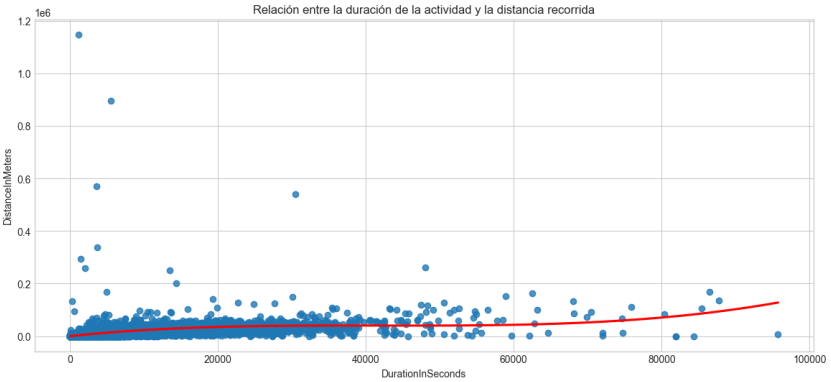


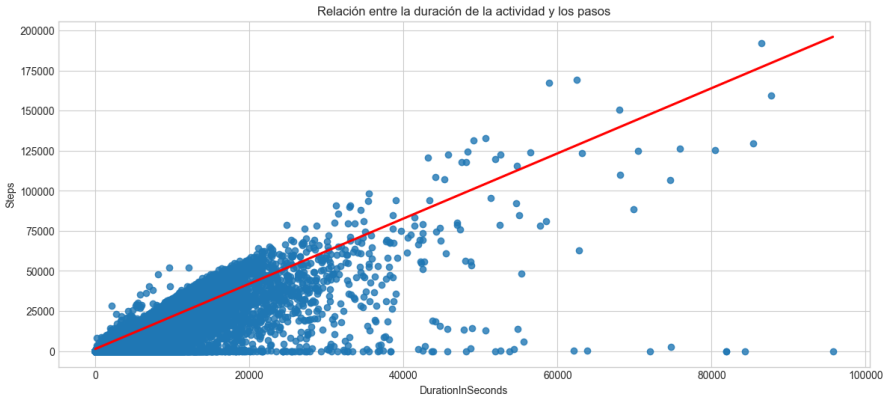
Podemos observar que la variable de duración contiene valores negativos y valores iguales a cero. Además, se puede observar la presencia de datos atípicos que se salen de la media.

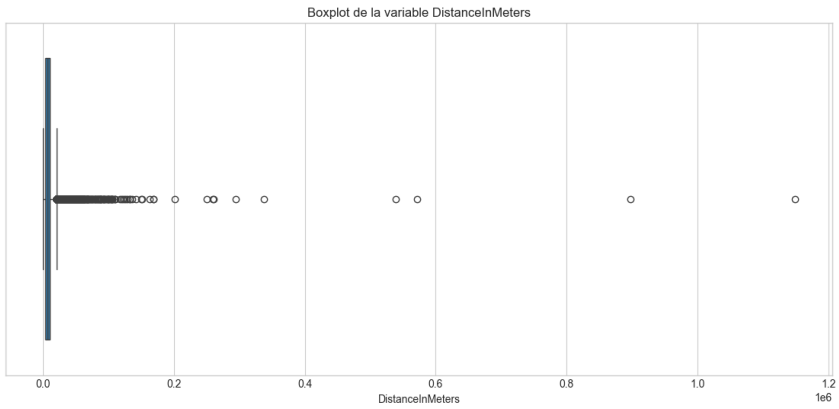
Tiempos nulos o negativos no tienen sentido práctico en nuestro análisis, así que se eliminarán del dataset



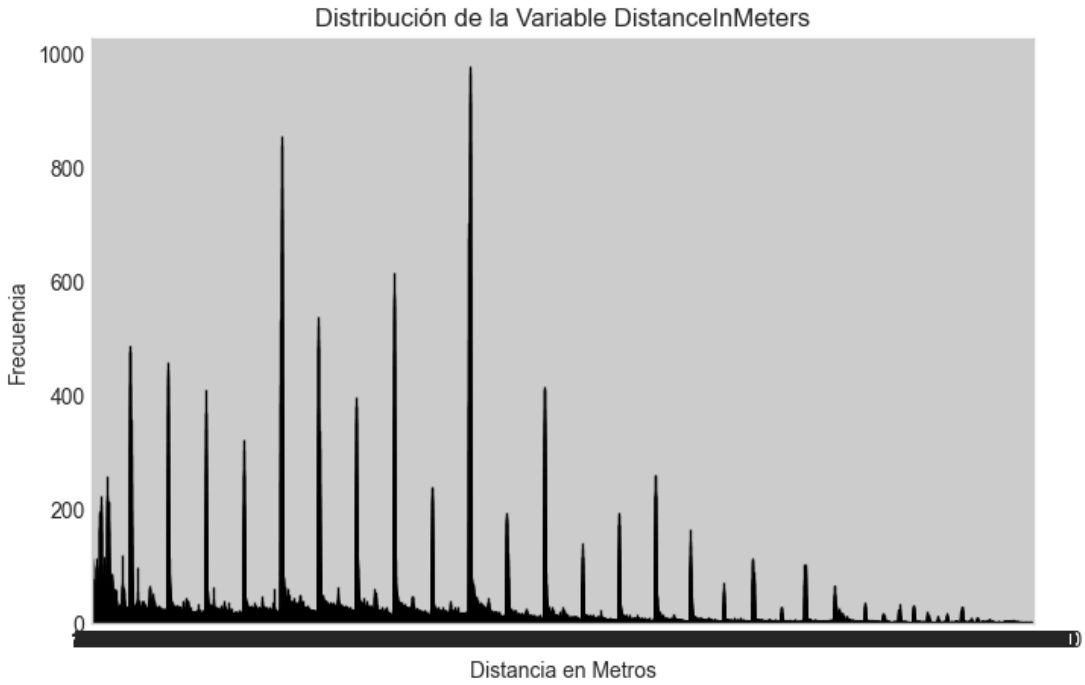
En el gráfico podemos observar que parece que hay muchos valores atípicos, pero esto se debe a que una gran cantidad de registros tiene una duración muy corta.

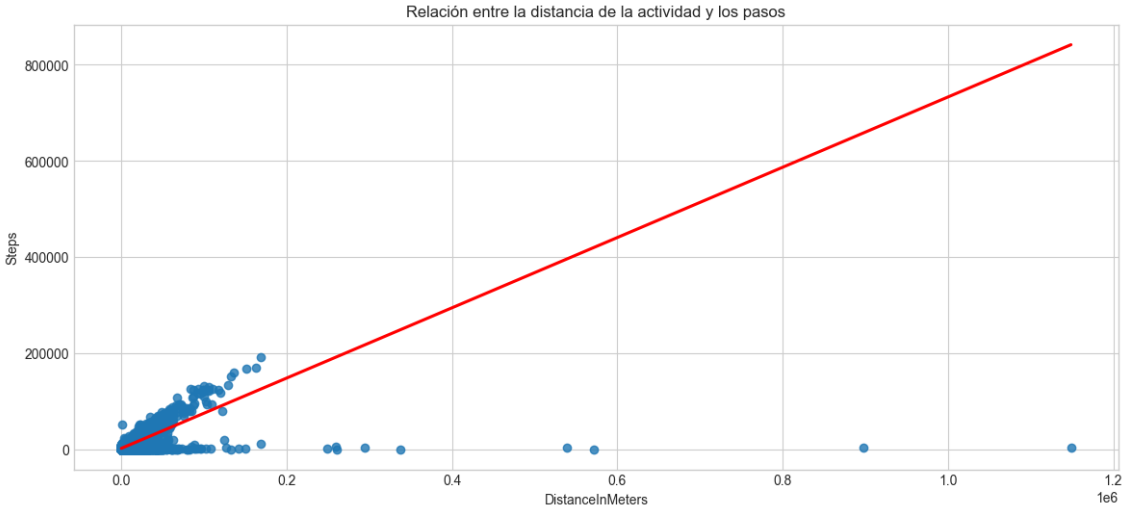


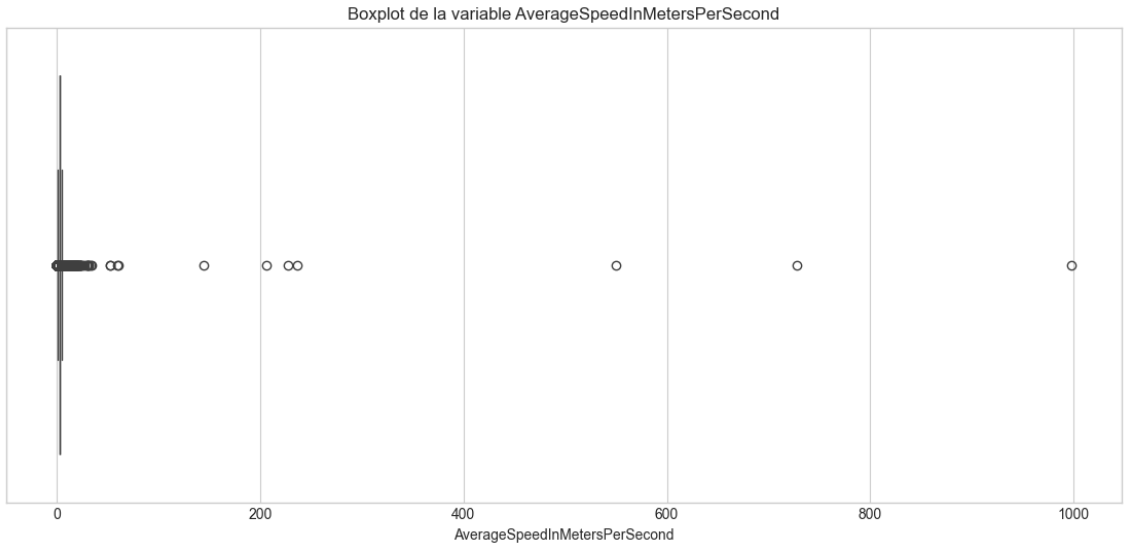




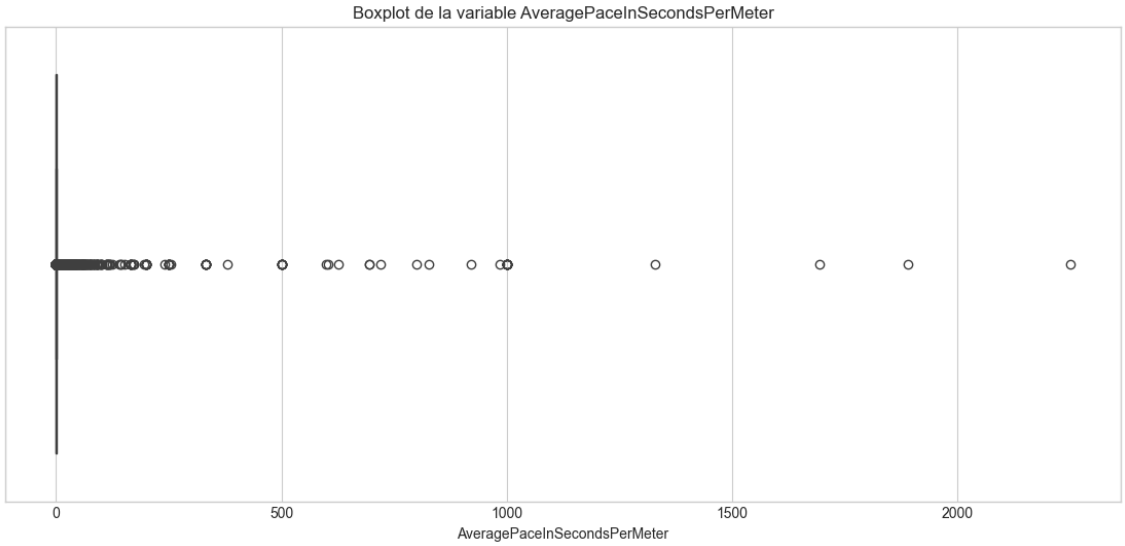
Esta variable tiene 508 registros menores o iguales a 0. Considerando que estamos analizando una actividad deportiva de tipo Running, registros sin distancia válida no serán útiles para el análisis.



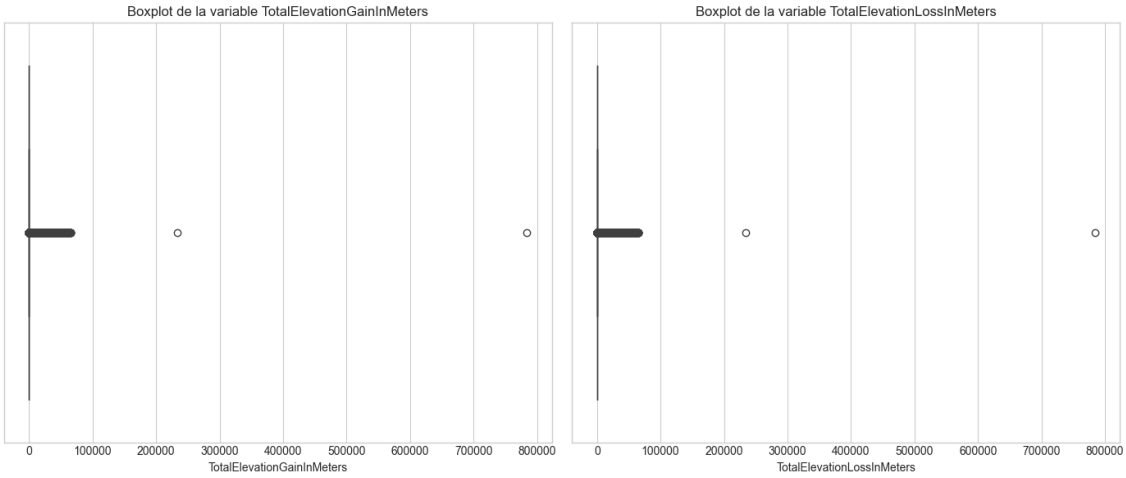




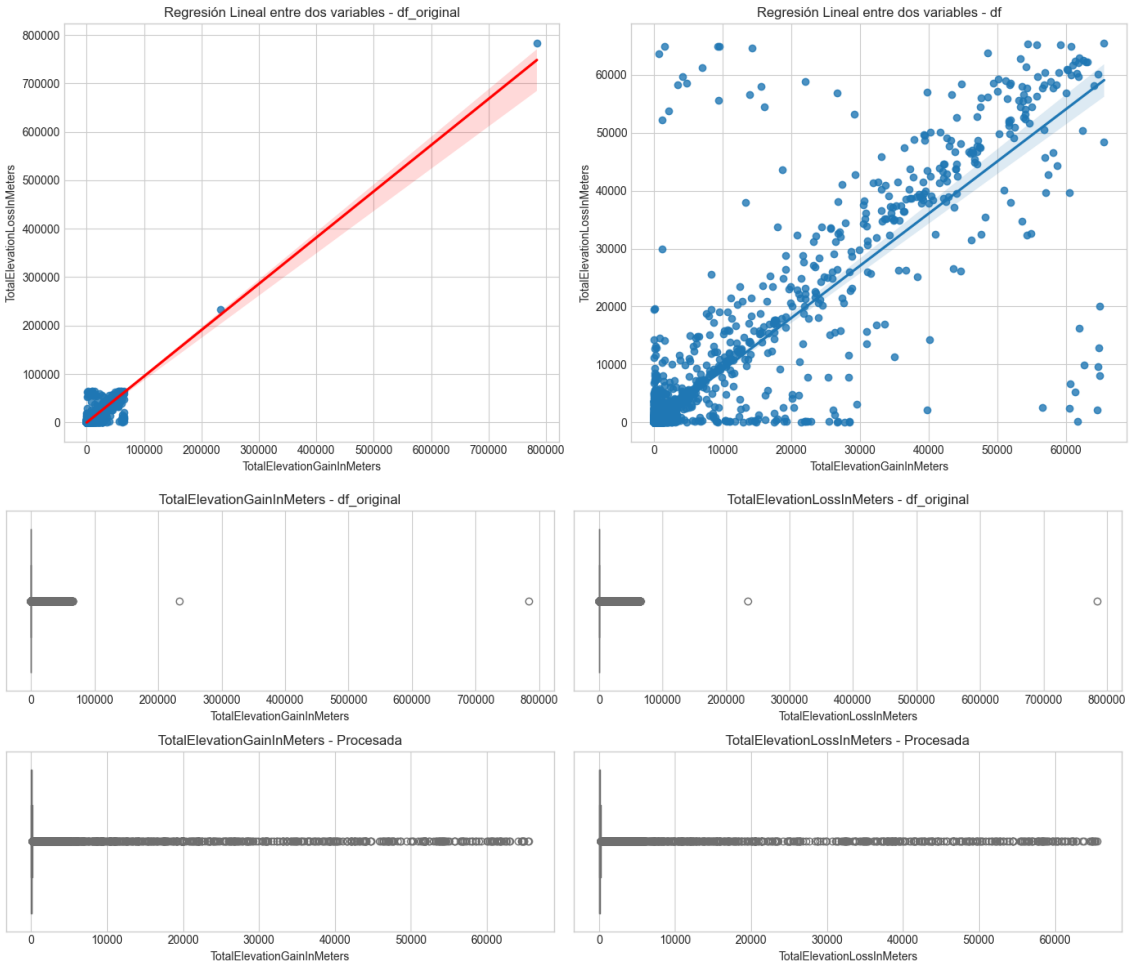
Se pueden observar 16 registros de velocidad menores o iguales a cero por lo que se eliminarán debido a que no tendrían validez para el análisis realizado.



Para el caso del ritmo los valores atípicos se observan de dos formas, la primera cuando el ritmo es muy bajo debido a que entre menor sea el ritmo significa que la persona corre más rápido; por otro lado, si el ritmo es muy alto significa que la persona está caminando o está en reposo.



Eliminar los 2 outliers que entorpecen el entendimiento de las variables TotalElevationGainInMeters y TotalElevationLossInMeters



Debido a que las variables manejadas individualmente tienen un alto porcentaje de nulos:

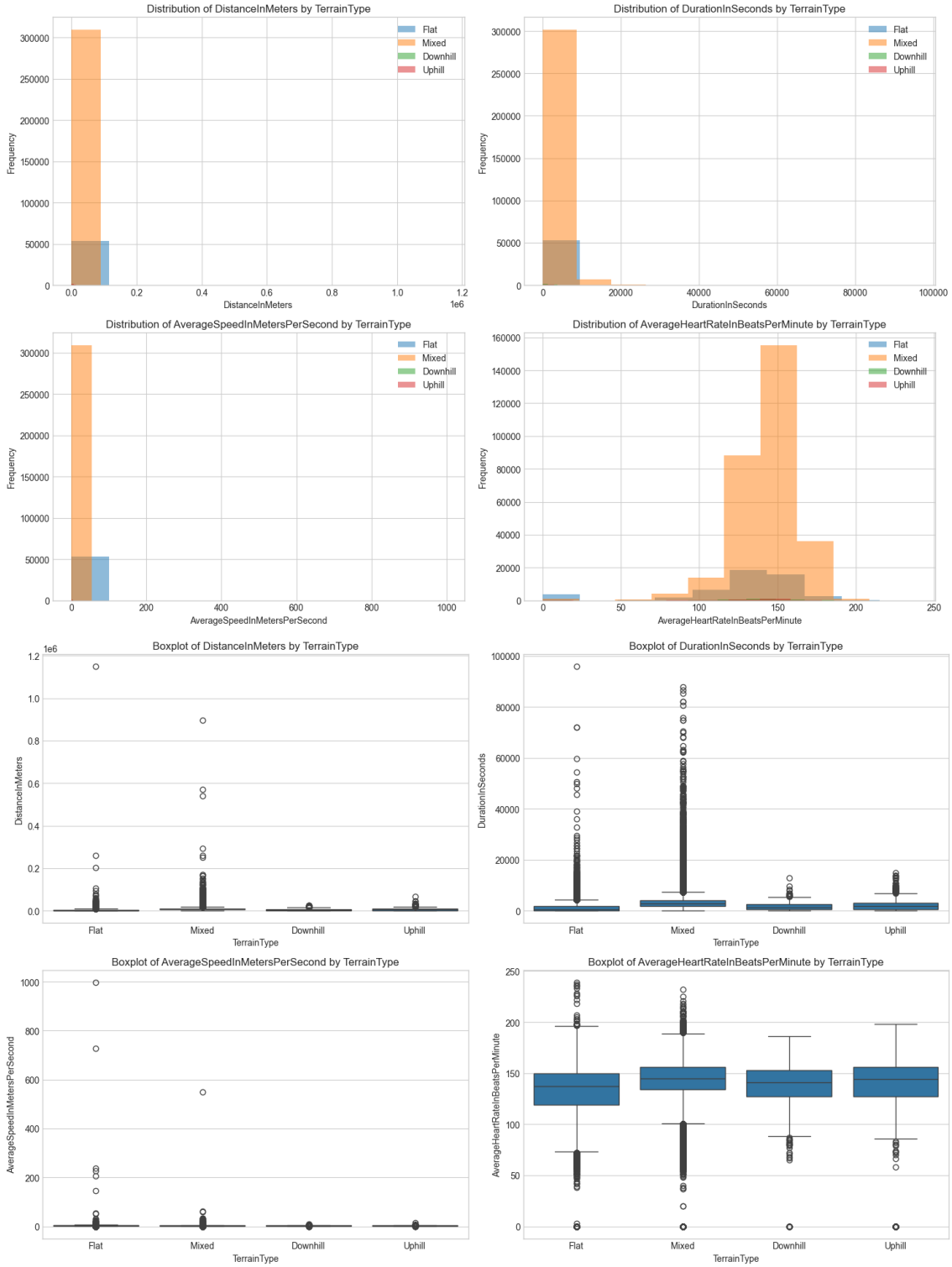
TotalElevationGainInMeters: 34886 registros nulos

TotalElevationLossInMeters: 34401 registros nulos

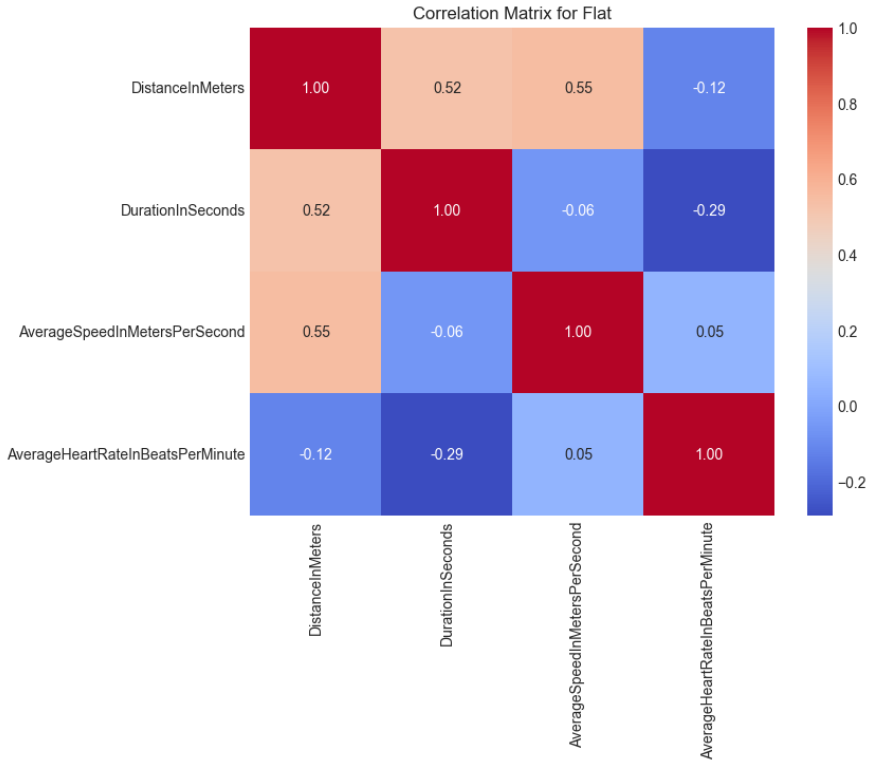
Y se consideran importantes para el sistema de detección, se hace una transformación clasificando el terreno recorrido por el usuario como:

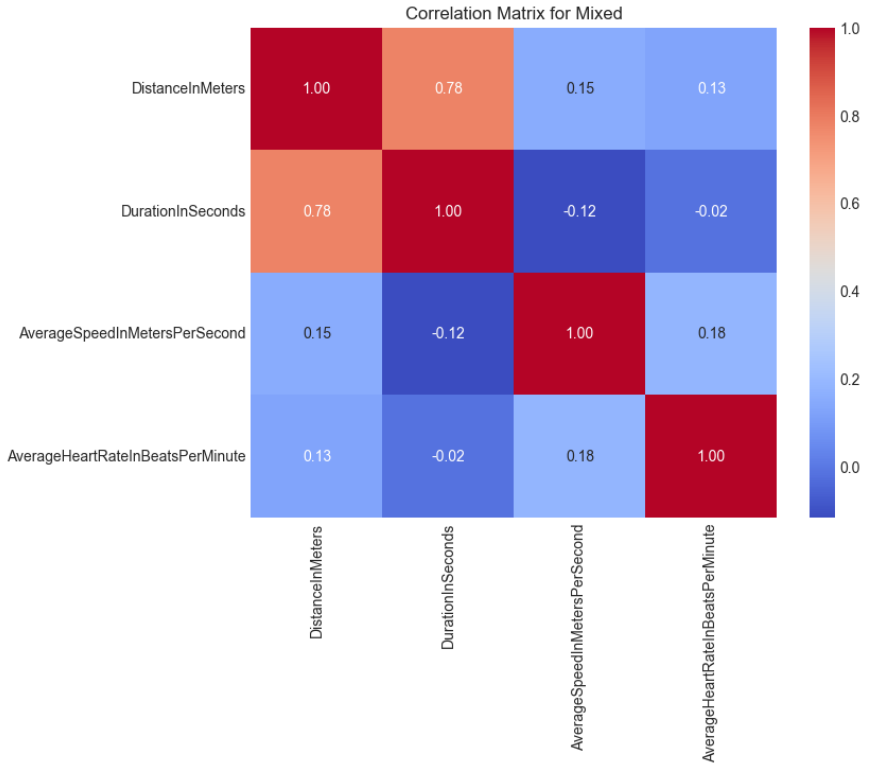
- **Flat**: Si TotalElevationGainInMeters y TotalElevationLossInMeters son 0 o nulos
- **Uphill**: Si TotalElevationGainInMeters es mayor a 0 y TotalElevationLossInMeters es 0
- **Downhill**: Si TotalElevationGainInMeters es 0 y TotalElevationLossInMeters es mayor a 0
- **Mixed**: Si TotalElevationGainInMeters y TotalElevationLossInMeters son mayores a 0

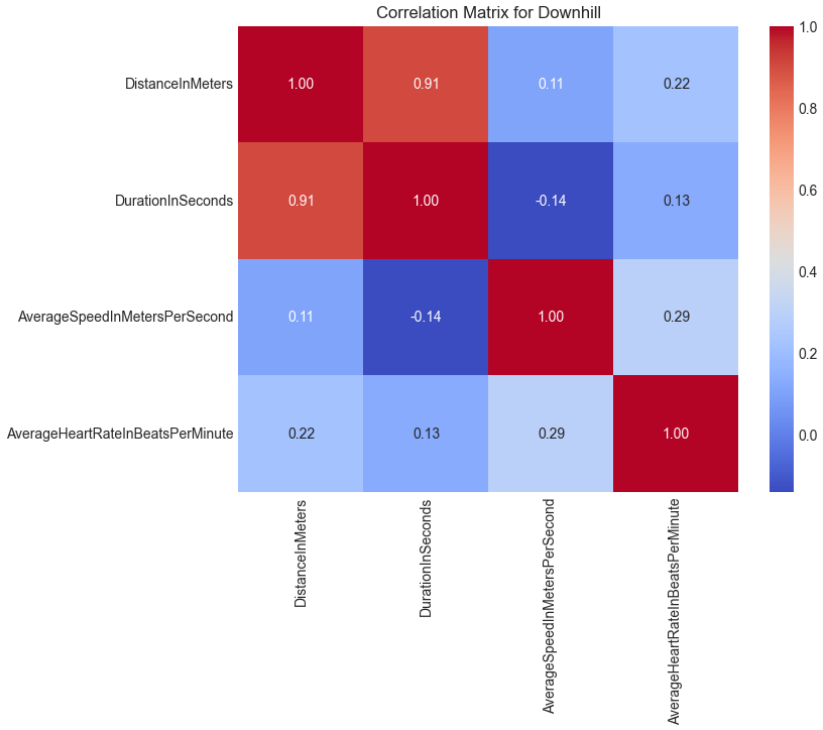
Debido a que se observa una alta correlación entre ambas variables lo que nos indica que si se usan como predictoras para el modelo podemos tener problemas de colinealidad.

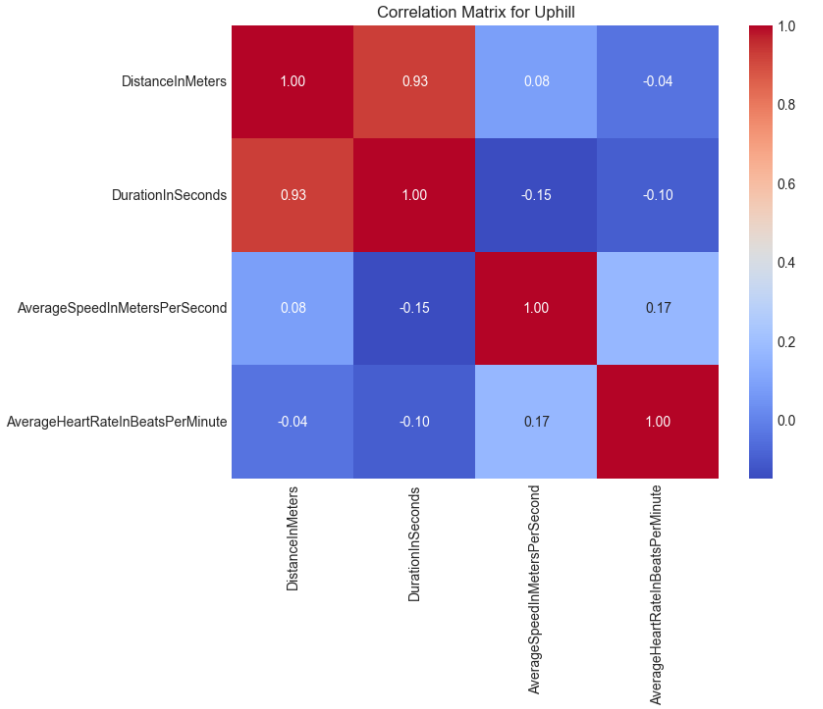


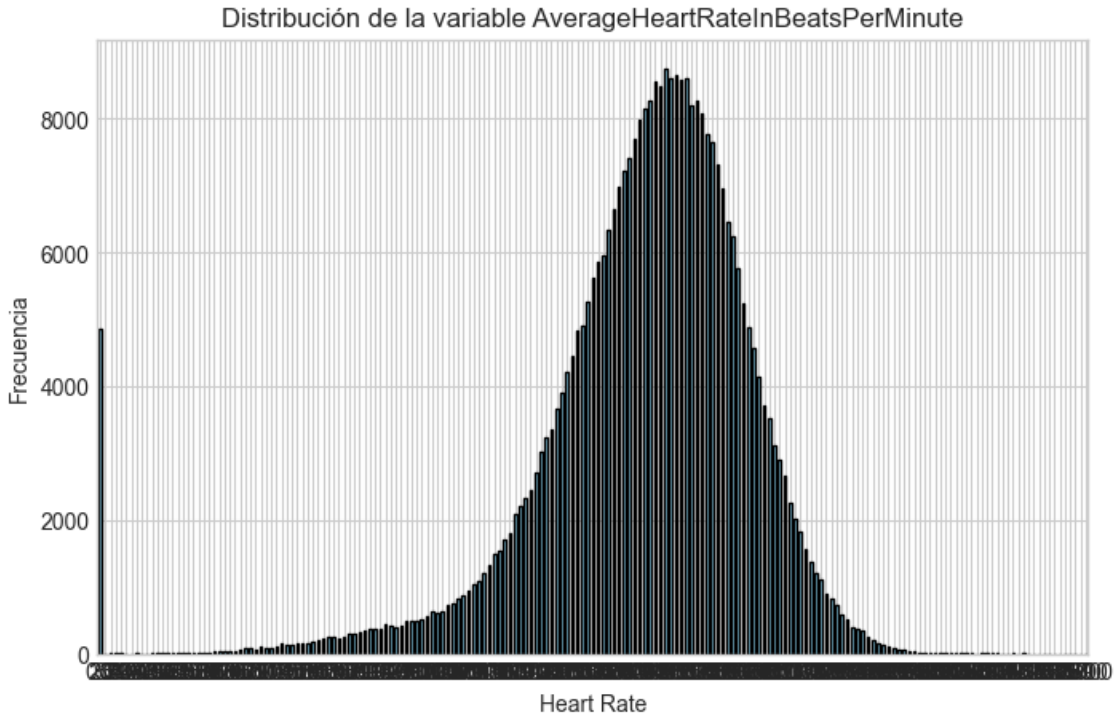
En los Boxplot se evidencia que con la clasificación de TerrainType en 4 categorías: Flat, Uphill, Downhill y Mixed, se observa que los terrenos Mixed y Flat presentan una mayor probabilidad de contener datos atípicos en cada una de las variables.

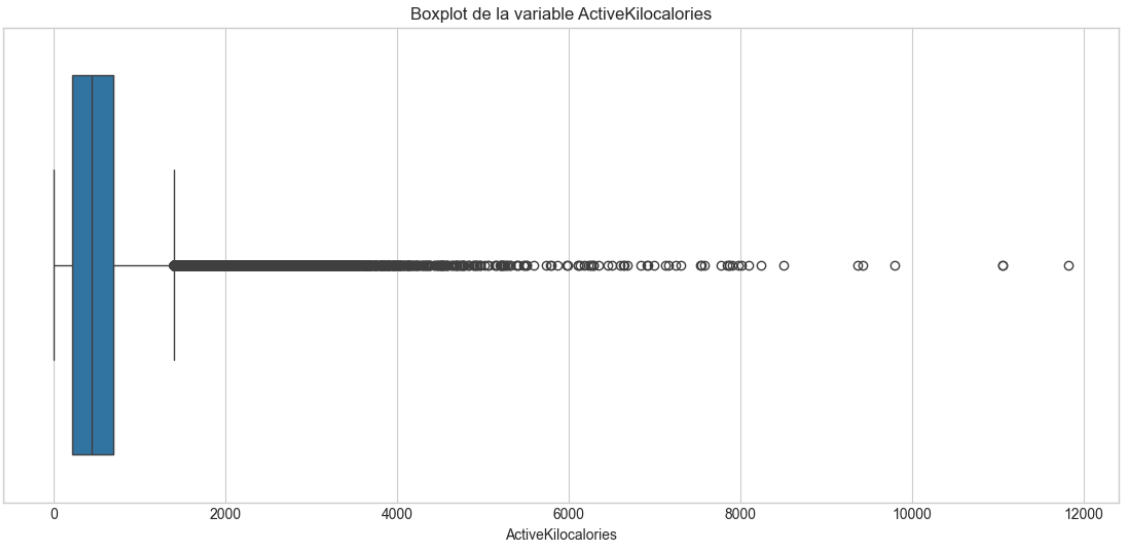

Correlation Matrix for Flat
DistanceInMeters
DurationInSeconds
AverageSpeedInMetersPerSecond
AverageHeartRateInBeatsPerMinute
1.00
0.52
0.55
-0.12
0.52
1.00
-0.06
-0.29
0.55
-0.06
1.00
0.05
-0.12
-0.29
0.05
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
DistanceInMeters
DurationInSeconds
AverageSpeedInMetersPerSecond
AverageHeartRateInBeatsPerMinute

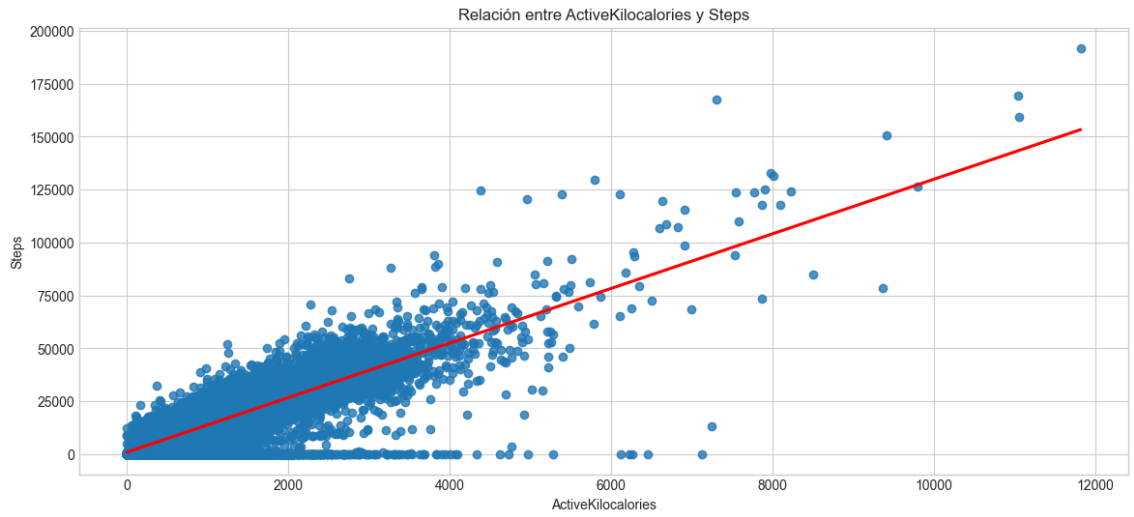

Correlation Matrix for Mixed
DistanceInMeters
DurationInSeconds
AverageSpeedInMetersPerSecond
AverageHeartRateInBeatsPerMinute
1.00
0.78
0.15
0.13
0.78
1.00
-0.12
-0.02
0.15
-0.12
1.00
0.18
0.13
-0.02
0.18
1.00
1.0
0.8
0.6
0.4
0.2
0.0
DistanceInMeters
DurationInSeconds
AverageSpeedInMetersPerSecond
AverageHeartRateInBeatsPerMinute

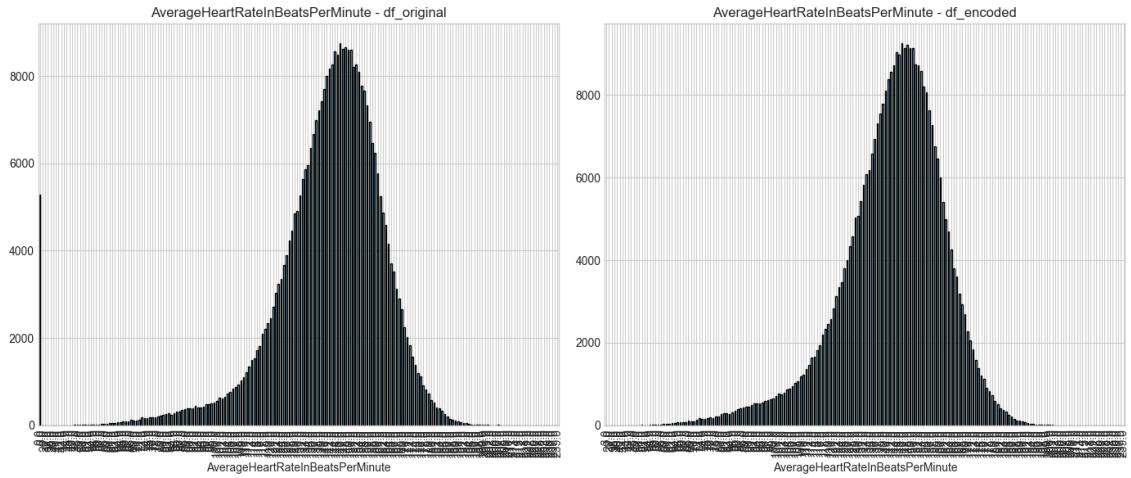

Correlation Matrix for Downhill
DistanceInMeters
DurationInSeconds
AverageSpeedInMetersPerSecond
AverageHeartRateInBeatsPerMinute
1.00
0.91
0.11
0.22
0.91
1.00
-0.14
0.13
0.11
-0.14
1.00
0.29
0.22
0.13
0.29
1.00
1.0
0.8
0.6
0.4
0.2
0.0


Correlation Matrix for Uphill
DistanceInMeters
DurationInSeconds
AverageSpeedInMetersPerSecond
AverageHeartRateInBeatsPerMinute
1.00
0.93
0.08
-0.04
0.93
1.00
-0.15
-0.10
0.08
-0.15
1.00
0.17
-0.04
-0.10
0.17
1.00
1.0
0.8
0.6
0.4
0.2
0.0


Distribución de la variable AverageHeartRateInBeatsPerMinute
Frecuencia
8000
6000
4000
2000
0
Heart Rate

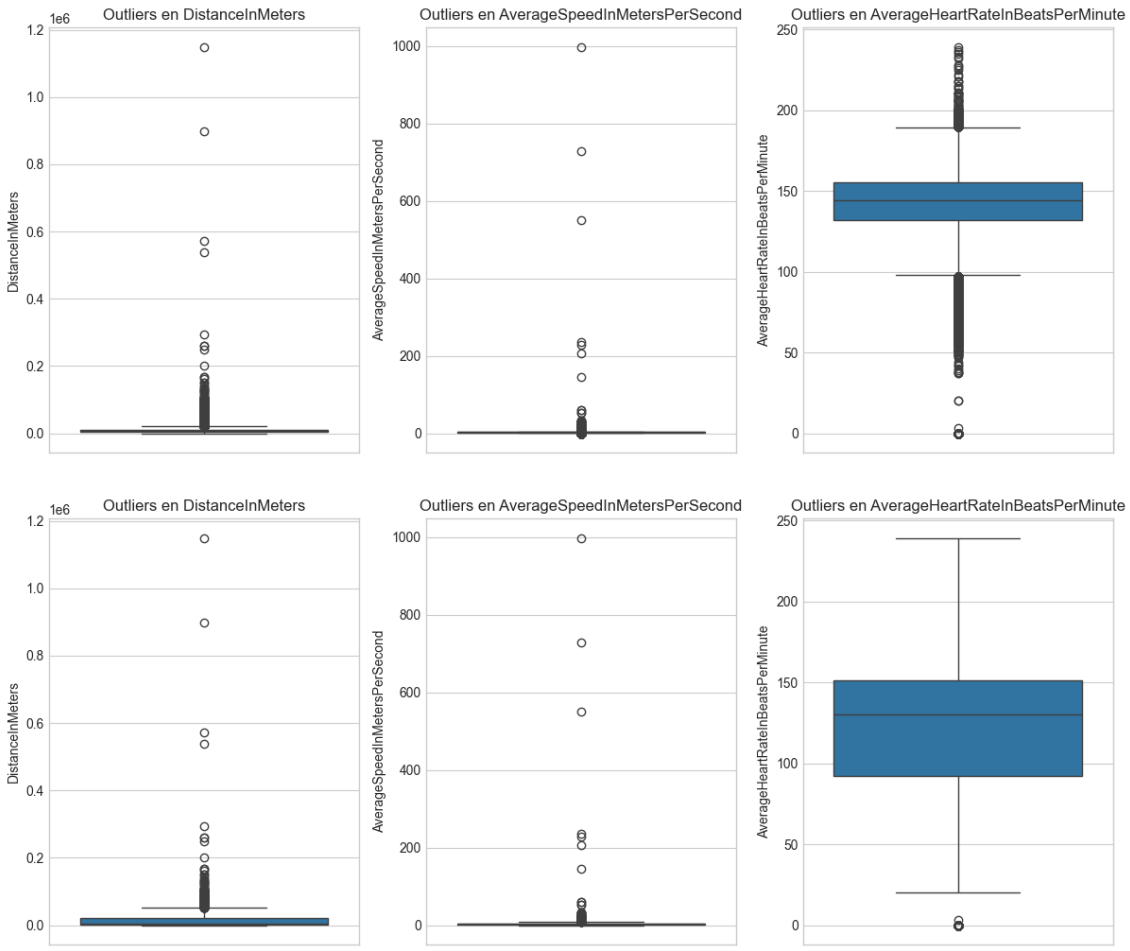

Boxplot de la variable ActiveKilocalories
0
2000
4000
6000
8000
10000
12000
ActiveKilocalories





ANÁLISIS DE OUTLIERS

- DistanceInMeters: Distancia recorrida durante la actividad, expresada en metros.
- AverageHeartRateInBeatsPerMinute: Frecuencia cardíaca promedio durante la actividad, medida en latidos por minuto (BPM).
- AveragePaceInMinutesPerKilometer: Ritmo promedio de la actividad, expresado en minutos por kilómetro.



CONCLUSIONES DE ANÁLISIS EXPLORATORIO

1. Reveló la existencia de datos inconsistentes como valores nulos, negativos o ceros en variables como DurationInSeconds, DistanceInMeters, AverageSpeedInMetersPerSecond y AverageHeartRateInBeatsPerMinute. Además, se identificaron valores atípicos (outliers) en variables como DistanceInMeters, AverageSpeedInMetersPerSecond y AverageHeartRateInBeatsPerMinute, lo que indica posibles errores de medición, actividades sospechosas

2. Se observó un desbalance significativo en la variable SourceType, donde la marca Garmin domina ampliamente como la fuente de datos. Esto podría introducir sesgos en el modelo, ya que la distribución está fuertemente inclinada hacia esta marca.

3. Se identificaron correlaciones entre diversas variables, como Steps y DistanceInMeters, TotalElevationGainInMeters y TotalElevationLossInMeters, Steps y DurationInSeconds, y DistanceInMeters y DurationInSeconds. Estas correlaciones son esperables en el contexto de actividades de running, pero es importante considerarlas

durante la selección de características para el modelo, ya que la colinealidad entre variables puede afectar la precisión del modelo.

CONCLUSIONES DE MODELAMIENTO

- Resultados generales de cada modelo:
    - KMeans clustering detected anomalies: 18381
    - DBSCAN detected anomalies: 364881
    - One-Class SVM detected anomalies: 7347
    - Isolation Forest detected anomalies: 73521
- A partir del modelo de K-Means, se identificaron 18381 registros atípicos. Estos fueron detectados al analizar los datos que se alejan significativamente de su centroide. Para determinar las anomalías, se consideraron como outliers aquellos puntos cuya distancia al centroide superaba el percentil 95, marcando así el 5% de los puntos más alejados de su cluster. Además, se observó que los registros atípicos se presentaban principalmente en terrenos flat y mixed, con mayor porcentaje en mixed. En cuanto a los dispositivos, Garmin fue el que registró la mayor cantidad de datos atípicos. La hora con mayor presencia de anomalías fue las entre las 6:00 y 7:00 AM. Al analizar la combinación de terreno y dispositivo, se identificó que la mayor cantidad de registros atípicos correspondía a Garmin en terreno Mixed.
- Al analizar los clusters generados por K-Means, se observa que los clusters 0 y 8 presentan altos valores de distancia, destacando especialmente el cluster 0, donde se identifican posibles valores atípicos. Además, tanto en el cluster 0 como en el 12 se evidencian valores inusuales en la métrica de duración.
- DBSCAN detectó la mayor cantidad de atípicos, acercándose al total de los datos, lo cual sugiere que la configuración actual del modelo no es adecuada y requiere ajustes para obtener resultados más precisos.
- Isolation Forest depende del porcentaje de anomalías esperado dentro del conjunto de datos, lo que significa que al aumentar este valor, también se incrementa la cantidad de datos detectados como atípicos. Por lo tanto, al establecer el parámetro de contamination en 0.2, estamos indicando que se espera que el 20% de los datos sean anomalías.