Tipología y Ciclo de Vida de los Datos

Práctica 2: Limpieza y análisis de datos

Javier Gómez de Diego

In [1]:
```
import numpy as np
import pandas as pd
import sklearn
import matplotlib.pyplot as plt
import re
```

In [2]:
```
a = pd.read_csv('activities.csv', sep=',')

a.columns
```

Out[2]:
```
Index(['ID de actividad', 'Fecha de la actividad', 'Nombre de la actividad',
       'Tipo de actividad', 'Descripción de la actividad',
       'Tiempo transcurrido', 'Distancia', 'Ritmo cardiaco máximo',
       'Esfuerzo Relativo', 'Desplazamiento', 'Equipamiento de la actividad',
       'Nombre del archivo', 'Peso del deportista', 'Peso de la bicicleta',
       'Tiempo transcurrido.1', 'Tiempo en movimiento', 'Distancia.1',
       'Velocidad máxima', 'Velocidad promedio', 'Desnivel positivo',
       'Pérdida de desnivel', 'Desnivel bajo', 'Desnivel alto',
       'Pendiente máxima', 'Pendiente promedio', 'Pendiente positiva promedio',
       'Pendiente negativa promedio', 'Cadencia máx.', 'Cadencia promedio',
       'Ritmo cardiaco máximo.1', 'Ritmo cardiaco promedio', 'Vatios máx.',
       'Vatios promedio', 'Calorías', 'Temperatura máx.',
       'Temperatura promedio', 'Esfuerzo Relativo.1', 'Trabajo total',
       'Número de carreras', 'Tiempo en ascenso', 'Tiempo en descenso',
       'Otro tiempo', 'Esfuerzo Percibido',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.type">Type</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.start_time">Start Time</span>',
       'Potencia promedio ponderada', 'Recuento de potencia',
       'Usar Esfuerzo Percibido', 'Esfuerzo Relativo percibido',
       'Desplazamiento.1', 'Peso total levantado', 'Desde la carga',
       'Distancia ajustada en pendientes', 'Hora de observación meteorológica',
       'Condición meteorológica', 'Temperatura meteorológica',
       'Temperatura aparente', 'Punto de rocío', 'Humedad',
       'Presión meteorológica', 'Velocidad del viento', 'Ráfaga de viento',
       'Dirección del viento', 'Intensidad de precipitación',
       'Hora del amanecer', 'Hora del atardecer', 'Fase lunar', 'Bicicleta',
       'Equipamiento', 'Probabilidad de precipitación',
       'Tipo de precipitación', 'Nubosidad', 'Visibilidad meteorológica',
       'Índice UV', 'Ozono meteorológico',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.jump_count">Jump Count</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.total_grit">Total Grit</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.avg_flow">Avg Flow</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.flagged">Flagged</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.avg_elapsed_speed">Avg Elapsed Speed</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.dirt_distance">Dirt Distance</span>'],
      dtype='object')
```

In [3]:
```
a.columns[a.columns.str.contains('translation_missing')]
```

Out[3]:
```
Index(['<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.type">Type</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.start_time">Start Time</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.jump_count">Jump Count</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.total_grit">Total Grit</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.avg_flow">Avg Flow</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.flagged">Flagged</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.avg_elapsed_speed">Avg Elapsed Speed</span>',
       '<span class="translation_missing" title="translation missing: es-ES.lib.export.portability_exporter.activities.horton_values.dirt_distance">Dirt Distance</span>'],
      dtype='object')
```

In [4]:
```
structure = '<span (?:.*?)>(.*?)</span>'
newc = {}
for i in a.columns[a.columns.str.contains('translation_missing')]:
    newc[i] = re.findall(structure, i)[0]
a.rename(columns = newc, inplace = True)
a.columns
```

Out[4]:
```
Index(['ID de actividad', 'Fecha de la actividad', 'Nombre de la actividad',
       'Tipo de actividad', 'Descripción de la actividad',
       'Tiempo transcurrido', 'Distancia', 'Ritmo cardiaco máximo',
       'Esfuerzo Relativo', 'Desplazamiento', 'Equipamiento de la actividad',
       'Nombre del archivo', 'Peso del deportista', 'Peso de la bicicleta',
       'Tiempo transcurrido.1', 'Tiempo en movimiento', 'Distancia.1',
       'Velocidad máxima', 'Velocidad promedio', 'Desnivel positivo',
       'Pérdida de desnivel', 'Desnivel bajo', 'Desnivel alto',
       'Pendiente máxima', 'Pendiente promedio', 'Pendiente positiva promedio',
       'Pendiente negativa promedio', 'Cadencia máx.', 'Cadencia promedio',
       'Ritmo cardiaco máximo.1', 'Ritmo cardiaco promedio', 'Vatios máx.',
       'Vatios promedio', 'Calorías', 'Temperatura máx.',
       'Temperatura promedio', 'Esfuerzo Relativo.1', 'Trabajo total',
       'Número de carreras', 'Tiempo en ascenso', 'Tiempo en descenso',
       'Otro tiempo', 'Esfuerzo Percibido', 'Type', 'Start Time',
       'Potencia promedio ponderada', 'Recuento de potencia',
       'Usar Esfuerzo Percibido', 'Esfuerzo Relativo percibido',
       'Desplazamiento.1', 'Peso total levantado', 'Desde la carga',
       'Distancia ajustada en pendientes', 'Hora de observación meteorológica',
       'Condición meteorológica', 'Temperatura meteorológica',
       'Temperatura aparente', 'Punto de rocío', 'Humedad',
       'Presión meteorológica', 'Velocidad del viento', 'Ráfaga de viento',
       'Dirección del viento', 'Intensidad de precipitación',
       'Hora del amanecer', 'Hora del atardecer', 'Fase lunar', 'Bicicleta',
       'Equipamiento', 'Probabilidad de precipitación',
       'Tipo de precipitación', 'Nubosidad', 'Visibilidad meteorológica',
       'Índice UV', 'Ozono meteorológico', 'Jump Count', 'Total Grit',
       'Avg Flow', 'Flagged', 'Avg Elapsed Speed', 'Dirt Distance'],
      dtype='object')
```

In [5]:
```
a.to_csv('activities2.csv', index=False, sep=',')  # Save as CSV
```