

Data Quality

Haldun Köktaş

2022-11-02

dailyActivity.csv

The dataset contains one “0” column. (LoggedDistance)

There are very low values in SedentaryMinutes column. If they all are 0s, we can conclude that they weren’t measured. But some observations are not all 0 but very low values. (13 minutes of sedentary minutes and 0-9 minutes of others. This means that the person moved only 13 minutes in the whole day.)

There are rows where “Calories” is below 500. (These probably are either wrong measurements or extreme situations.)

There are zero or very low values in TotalSteps columns.

There are very low values regarding sleeping time (below 3 hours). These will regard as outliers.

One person’s bedtime is excessively large (16 hours). This may be due to a disease or an exceptional situation.

Measurement issues disappear when it’s merged with the daily sleep data.

weightLogInfo.csv

There are only eight person’s weight values, some of which have only two records.

One person’s weight is 130 kg. The fact that the record is manually entered, and considering its distance from the average, that might be an outlier.

heartBeat.csv

There are only 14 person’s data.