FitBit Fitness Tracker Dataset: Initial Data Exploration

As there are two time periods included in this dataset, having an inequal number of data files, we must reconcile these disparities and ensure that we have usable, corresponding data from each time period. In the case that we don't, we must determine whether to use data that is only included in one of the two time periods. If so, we must also determine how to effectively use this data. In order to complete these tasks, we must first complete a detailed exploration of the information included in each time period.

Contents of Data Files:

File Name	Records Time Interval	Columns	Match Between Time Periods
dailyActivity_merged	days	id, activity date, total steps, total distance, tracker distance, logged activity, distance per activity level (very active, moderately active, light active, sedentary active), minutes per activity level (very active, moderately active, light active, sedentary active), calories	match
heartrate_seconds_merged	seconds	id, date and time, value (heartrate)	match
hourlyCalories_merged	hours	id, date and time, calories	match
hourlyIntensities_merged	hours	id, date and time, total intensity, average intensity	match
hourlySteps_merged	hours	id, date and time, step total	match
minuteCaloriesNarrow_merged	minutes	id, date and time, calories	match
minuteIntensitiesNarrow_merged	minutes	id, date and time, intensity	match
minuteMETsNarrow_merged	minutes	id, date and time, METs (energy used)	match
minuteSleep_merged	minutes	id, date and time, value, logID	match

minuteStepsNarrow_merged	minutes	id, date and time, steps	match
weightLogInfo_merged	days	id, date, weight kg, weight lb, fat, BMI, IsManualReport, LogID	match
dailyCalories_merged	days	id, date, calories	only 2nd period
dailyIntensities_merged	days	id, date, minutes per activity level (very active, moderately active, light active, sedentary active), distance per activity level (very active, moderately active, light active, sedentary active)	only 2nd period
dailySteps_merged	days	id, date, step total	only 2nd period
minuteCaloriesWide_merged	minutes	id, date and time, calories00 - calories59	only 2nd period
minuteIntensitiesWide_merged	minutes	id, date and time, intensity00 – intensity59	only 2nd period
minuteStepsWide_merged	minutes	id, date and time, steps00 – steps 59	only 2nd period
sleepDay_merged	days	id, date, total sleep records, total minutes asleep, total time in bed	only 2nd period

By exploring the contents of each data file in each of the two time periods, we were able to identify which files included corresponding data between the two time periods. In doing so, we discovered that the dataset for the second time period included seven extra data files that did not have a corresponding data file in the first time period. However, upon closer inspection, it seems that these seven extra files simply store repeated data, though in different time intervals. For example, while both time periods include a hourlyCalories file, the second time period also includes a minuteCalories file, which breaks down the calories burnt by minute rather than hour.

These findings can be confirmed during data cleaning – sanity check! For example, in the minuteCalories file, we can find the sum of calories burnt by minute for a specific user in a specific hour and compare this to the data for the same user in the same hour in the hourlyCalories file. While this may seem to be an unnecessary use of time, it is important to confirm that we have a proper understanding of the contents and organization of our dataset before we proceed to analysis.

Continuing the exploration of this dataset, we will count the number of unique users and the number of days included in each time period, keeping a lookout for any overlaps or missing values. We will also note the datatypes of each of the columns in our dataset, which will determine whether we need to convert any datatypes during the data cleaning process.