**Issue 1 : Missing values**

1. Description : In CSV file weightLogInfo\_merged.csv , 65 Missing values found. All of the missing values are in the ‘fat’ column.
2. Solution : This column states about the users’ body fat percentage, the present value is too less to make any assumption on the missing values, so **this column (fat) should be abandon and not be used in the analysis process.**

**Issue 2 : Potential outliers**

1. Description : Many potential outliers were detected using upper & lower limit from IQR method.
2. Solution :

* For data that contains a small number of potential outliers, human inspection is to be applied to determine the reliability of the value, the outlier values would be replaced by the average value without the outliers.
* For data that contains unreasonable amount of potential outliers, the reason is to be investigated. And judgement must be made on the range of reasonable values, and any outliers outside of this range must be removed or replaced by average value.
* dailyActivity\_merged.csv
* TotalSteps ,TotalDistance ,TrackerDistance ,VeryActiveDistance ,ModeratelyActiveDistance ,LightActiveDistance ,SedentaryActiveDistance ,VeryActiveMinute ,FairlyActiveMinute ,LightlyActiveMinute : All potential outliers seem to be reasonable(possible) outliers and not due to any measurement or data acquisition error.
* SedentaryMinute : Some potential outliers of this column does seem to be unreasonable(such as 0,2 ,13), **needs further investigation**
* Calories : Some potential outliers of this column does seem to be unreasonable(such as 0,52,57), **needs further investigation**
* heartrate\_seconds\_merged.csv : All potential outliers seem to be reasonable(possible) outliers and not due to any measurement or data acquisition error.
* minuteCaloriesNarrow\_merged.csv : All potential outliers seem to be reasonable(possible) outliers and not due to any measurement or data acquisition error.
* minuteCaloriesWide\_merged.csv **\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***
* minuteIntensitiesNarrow\_merged.csv & minuteIntensitiesWide\_merged.csv : values are predetermined qualitative values(0,1,2,3) indicating the intensity, no real outliers present.
* minuteMETsNarrow\_merged.csv : For most people, the reasonable values for METS at maximum would be around 10-11 METs, so there is a huge amount of unreasonable outliers in column METs, indicating potential error in measurement, data acquisition or unit of measuring. **This suggest that METs data is not reasonable & reliable enough to be used in the analysis process.**
* minuteSleep\_merged.csv : values are predetermined qualitative values(0,1,2,3) indicating the intensity, no real outliers present.
* minuteStepsNarrow\_merged.csv & minuteStepsWide\_merged.csv: All potential outliers seem to be reasonable(possible) outliers and not due to any measurement or data acquisition error.
* minuteStepsNarrow\_merged.csv : values of 961 in column “TotalTimeInBed” seem a bit unreasonable as it indicates a 16 hours in bed. But it is not impossible, so they will be considered as normal outliers that are not caused by any measurement or collection error.
* minuteStepsNarrow\_merged.csv : a record of outlier in bodyweight (133.5 kg) present, not enough information to judge as there is only one bodyweight record from the user. Removing it in this case to avoid skewedness in distribution.

**Issue 3 : Inconsistent formatting (Ambiguous date)**

1. Description : For all of the data regarding date, there exist ambiguity of date format.
2. Solution : The ambiguity presents when the day is less than 13, which makes it a potential month value. Judging from the date values where day is larger than 13 and thus make their format clear, we can say that the standard format used for date in this dataset is “mm/dd/yyyy”. To address this issue, we first need to check if it will make any harm to our analysis process, to check if our analysis tool will misidentified the date value. We use R programming language to do the analysis in this case.