# This is a checklist for data cleaning and inspection before data cleaning.

## Check for missing values

|  |  |  |  |
| --- | --- | --- | --- |
| File Name | Done | Method/tool used | Result & remark |
| dailyActivity\_merged.csv | T | Excel filter function | No missing values found. |
| heartrate\_seconds\_merged.csv | T | R isna() function | No missing values found. |
| minuteCaloriesNarrow\_merged.csv | T | No missing values found. |
| minuteCaloriesWide\_merged.csv | T | No missing values found. |
| minuteIntensitiesNarrow\_merged.csv | T | No missing values found. |
| minuteIntensitiesWide\_merged.csv | T | No missing values found. |
| minuteMETsNarrow\_merged.csv | T | No missing values found. |
| minuteSleep\_merged.csv | T | No missing values found. |
| minuteStepsNarrow\_merged.csv | T | No missing values found. |
| minuteStepsWide\_merged.csv | T | No missing values found. |
| sleepDay\_merged.csv | T | No missing values found. |
| weightLogInfo\_merged.csv | T | 65 Missing values found. All of the missing values are in the ‘fat’ column. |

## Unreasonable values (outliers, illegal values)

|  |  |  |  |
| --- | --- | --- | --- |
| File Name | Done | Method/tool used | Remark |
| dailyActivity\_merged.csv | T | Compute quantiles, IQR and upper,lower limits using R. Then any numeric value outside of range (lower limit,upper limit) is identified as potential outliers | Columns with potential outliers and the number of potential outliers :  TotalSteps 32  TotalDistance 35  TrackerDistance 34  LoggedActivitiesDistance 32  VeryActiveDistance 106  ModeratelyActiveDistance 101  LightActiveDistance 21  SedentaryActiveDistance 82  VeryActiveMinutes 98  FairlyActiveMinutes 93  LightlyActiveMinutes 23  SedentaryMinutes 9  Calories 49 |
| heartrate\_seconds\_merged.csv | T | Column Value contains 116658 potential outliers |
| minuteCaloriesNarrow\_merged.csv | T | Column Calories contains 211126 potential outliers |
| minuteCaloriesWide\_merged.csv | T | Shit tons of outliers in every column |
| minuteIntensitiesNarrow\_merged.csv | T | Column Intensity contains 213478 potential outliers |
| minuteIntensitiesWide\_merged.csv | T | Shit tons of outliers in every column |
| minuteMETsNarrow\_merged.csv | T | Column METs contains 250855 potential outliers |
| minuteSleep\_merged.csv | T | Column value contains 16041 potential outliers |
| minuteStepsNarrow\_merged.csv | T | Column Steps contains 199910 potential outliers |
| minuteStepsWide\_merged.csv | T | Shit tons of outliers in every column |
| sleepDay\_merged.csv | T | Columns with potential outliers and the number of potential outliers :  TotalSleepRecords 46  TotalMinutesAsleep 40  TotalMinutesAsleep 40 |
| weightLogInfo\_merged.csv |  | Most values for column “Fat” are missing, outliers depend on how to fill in those values.  A “133.5” in kgs for weight log info, potential outlier or error value. |

## Inconsistent formatting (within & across files)

* **Method used : by inspection**

|  |  |  |
| --- | --- | --- |
| File Name | Done | Remark |
| dailyActivity\_merged.csv | T | Ambiguous date format : Presentation of “ActivityDate” column in Excel varies because some for dates whose day is less than 13, Excel can not determine without ambiguity the format of the date. |
| heartrate\_seconds\_merged.csv | T | Ambiguous date format : “Time” column |
| minuteCaloriesNarrow\_merged.csv | T | Ambiguous date format : “ActivityColumn” column the format of the date. |
| minuteCaloriesWide\_merged.csv | T | Ambiguous date format : ActivityHour” column without ambiguity the format of the date. |
| minuteIntensitiesNarrow\_merged.csv | T | Ambiguous date format : “ActivityMinutes” column |
| minuteIntensitiesWide\_merged.csv | T | Ambiguous date format : “ActivityHour” column |
| minuteMETsNarrow\_merged.csv | T | Ambiguous date format : “ActivityHour” column |
| minuteSleep\_merged.csv | T | Ambiguous date format : “date” column |
| minuteStepsNarrow\_merged.csv | T | Ambiguous date format : “ActivityMinute” column |
| minuteStepsWide\_merged.csv | T | Ambiguous date format : “ActivityHour” column |
| sleepDay\_merged.csv | T | Ambiguous date format : “SleepDay” column |
| weightLogInfo\_merged.csv | T | Ambiguous date format : “Date” column |

## Hard-to-read data fields

|  |  |  |
| --- | --- | --- |
| File Name | Done | Remark |
| dailyActivity\_merged.csv | T | No Hard-to-read data fields found |
| heartrate\_seconds\_merged.csv | T | No Hard-to-read data fields found |
| minuteCaloriesNarrow\_merged.csv | T | No Hard-to-read data fields found |
| minuteCaloriesWide\_merged.csv | T | No Hard-to-read data fields found |
| minuteIntensitiesNarrow\_merged.csv | T | No Hard-to-read data fields found |
| minuteIntensitiesWide\_merged.csv | T | No Hard-to-read data fields found |
| minuteMETsNarrow\_merged.csv | T | No Hard-to-read data fields found |
| minuteSleep\_merged.csv | T | No Hard-to-read data fields found |
| minuteStepsNarrow\_merged.csv | T | No Hard-to-read data fields found |
| minuteStepsWide\_merged.csv | T | No Hard-to-read data fields found |
| sleepDay\_merged.csv | T | No Hard-to-read data fields found |
| weightLogInfo\_merged.csv | T | No Hard-to-read data fields found |

## Potential misspelling & other data entry errors

|  |  |  |
| --- | --- | --- |
| File Name | Done | Remark |
| dailyActivity\_merged.csv | T | No potential data entry errors found |
| heartrate\_seconds\_merged.csv | T | No potential data entry errors found |
| minuteCaloriesNarrow\_merged.csv | T | No potential data entry errors found |
| minuteCaloriesWide\_merged.csv | T | No potential data entry errors found |
| minuteIntensitiesNarrow\_merged.csv | T | No potential data entry errors found |
| minuteIntensitiesWide\_merged.csv | T | No potential data entry errors found |
| minuteMETsNarrow\_merged.csv | T | No potential data entry errors found |
| minuteSleep\_merged.csv | T | No potential data entry errors found |
| minuteStepsNarrow\_merged.csv | T | No potential data entry errors found |
| minuteStepsWide\_merged.csv | T | No potential data entry errors found |
| sleepDay\_merged.csv | T | No potential data entry errors found |
| weightLogInfo\_merged.csv | T | No potential data entry errors found |