Bellabeat Case Study — Prepare Phase

Francesco De Felice

2025-11-15

---

1. Purpose

This phase ensures that all Fitbit datasets are correctly imported, consistent, and ready for processing. The objective is to confirm data completeness, structure, variable types, and potential limitations before cleaning and merging.

---

2. Importing Datasets

All files were stored under: `data/raw`

```
daily_activity <- read_csv("/cloud/project/Bellabeat Case Study/Data/Raw/dailyActivity_merged.csv")
sleep_day      <- read_csv("/cloud/project/Bellabeat Case Study/Data/Raw/sleepDay_merged.csv")
weight_log     <- read_csv("/cloud/project/Bellabeat Case Study/Data/Raw/weightLogInfo_merged.csv")

cat("Datasets imported successfully on", Sys.Date(), "\n")
```

```
## Datasets imported successfully on 20407
```

---

3. Datasets Inspection

```
glimpse(daily_activity)
```

```
## Rows: 940
## Columns: 15
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ ActivityDate             <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~
## $ TotalSteps               <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 13019~
## $ TotalDistance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~
## $ TrackerDistance          <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5~
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3~
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0~
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveMinutes        <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4~
## $ FairlyActiveMinutes      <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21~
## $ LightlyActiveMinutes     <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205, ~
```

```
## $ SedentaryMinutes         <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818~
## $ Calories                 <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203~
```

```
glimpse(sleep_day)
```

```
## Rows: 413
## Columns: 5
## $ Id                 <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150~
## $ SleepDay           <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 AM", "~
## $ TotalSleepRecords  <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TotalMinutesAsleep <dbl> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2~
## $ TotalTimeInBed     <dbl> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3~
```

```
glimpse(weight_log)
```

```
## Rows: 67
## Columns: 8
## $ Id             <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 2873212~
## $ Date           <chr> "5/2/2016 11:59:59 PM", "5/3/2016 11:59:59 PM", "4/13/2~
## $ WeightKg       <dbl> 52.6, 52.6, 133.5, 56.7, 57.3, 72.4, 72.3, 69.7, 70.3, ~
## $ WeightPounds   <dbl> 115.9631, 115.9631, 294.3171, 125.0021, 126.3249, 159.6~
## $ Fat            <dbl> 22, NA, NA, NA, NA, 25, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ BMI            <dbl> 22.65, 22.65, 47.54, 21.45, 21.69, 27.45, 27.38, 27.25,~
## $ IsManualReport <lgl> TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ~
## $ LogId          <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12, 1.461283e+12,~
```

```
summary(daily_activity)
```

```
##        Id            ActivityDate         TotalSteps    TotalDistance
##  Min.   :1.504e+09   Length:940         Min.   :    0   Min.   : 0.000
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 3790   1st Qu.: 2.620
##  Median :4.445e+09   Mode  :character   Median : 7406   Median : 5.245
##  Mean   :4.855e+09                      Mean   : 7638   Mean   : 5.490
##  3rd Qu.:6.962e+09                      3rd Qu.:10727   3rd Qu.: 7.713
##  Max.   :8.878e+09                      Max.   :36019   Max.   :28.030
##  TrackerDistance  LoggedActivitiesDistance VeryActiveDistance
##  Min.   : 0.000   Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 2.620   1st Qu.:0.0000           1st Qu.: 0.000
##  Median : 5.245   Median :0.0000           Median : 0.210
##  Mean   : 5.475   Mean   :0.1082           Mean   : 1.503
##  3rd Qu.: 7.710   3rd Qu.:0.0000           3rd Qu.: 2.052
##  Max.   :28.030   Max.   :4.9421           Max.   :21.920
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.000      Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 1.945      1st Qu.:0.000000
##  Median :0.2400           Median : 3.365      Median :0.000000
##  Mean   :0.5675           Mean   : 3.341      Mean   :0.001606
##  3rd Qu.:0.8000           3rd Qu.: 4.783      3rd Qu.:0.000000
##  Max.   :6.4800           Max.   :10.710      Max.   :0.110000
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0        Min.   :   0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:127.0        1st Qu.: 729.8
##  Median :  4.00    Median :  6.00      Median :199.0        Median :1057.5
##  Mean   : 21.16    Mean   : 13.56      Mean   :192.8        Mean   : 991.2
##  3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.0        3rd Qu.:1229.5
##  Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1440.0
```

```
##     Calories
##  Min.   :   0
##  1st Qu.:1828
##  Median :2134
##  Mean   :2304
##  3rd Qu.:2793
##  Max.   :4900
```

```
summary(sleep_day)
```

```
##        Id              SleepDay         TotalSleepRecords TotalMinutesAsleep
##  Min.   :1.504e+09   Length:413         Min.   :1.000     Min.   : 58.0
##  1st Qu.:3.977e+09   Class :character   1st Qu.:1.000     1st Qu.:361.0
##  Median :4.703e+09   Mode  :character   Median :1.000     Median :433.0
##  Mean   :5.001e+09                      Mean   :1.119     Mean   :419.5
##  3rd Qu.:6.962e+09                      3rd Qu.:1.000     3rd Qu.:490.0
##  Max.   :8.792e+09                      Max.   :3.000     Max.   :796.0
##  TotalTimeInBed
##  Min.   : 61.0
##  1st Qu.:403.0
##  Median :463.0
##  Mean   :458.6
##  3rd Qu.:526.0
##  Max.   :961.0
```

```
summary(weight_log)
```

```
##        Id                Date              WeightKg       WeightPounds
##  Min.   :1.504e+09   Length:67          Min.   : 52.60   Min.   :116.0
##  1st Qu.:6.962e+09   Class :character   1st Qu.: 61.40   1st Qu.:135.4
##  Median :6.962e+09   Mode  :character   Median : 62.50   Median :137.8
##  Mean   :7.009e+09                      Mean   : 72.04   Mean   :158.8
##  3rd Qu.:8.878e+09                      3rd Qu.: 85.05   3rd Qu.:187.5
##  Max.   :8.878e+09                      Max.   :133.50   Max.   :294.3
##
##       Fat              BMI        IsManualReport      LogId
##  Min.   :22.00   Min.   :21.45   Mode :logical   Min.   :1.460e+12
##  1st Qu.:22.75   1st Qu.:23.96   FALSE:26        1st Qu.:1.461e+12
##  Median :23.50   Median :24.39   TRUE :41        Median :1.462e+12
##  Mean   :23.50   Mean   :25.19                   Mean   :1.462e+12
##  3rd Qu.:24.25   3rd Qu.:25.56                   3rd Qu.:1.462e+12
##  Max.   :25.00   Max.   :47.54                   Max.   :1.463e+12
##  NA's   :65
```

Dataset	Rows	Columns	Description	Notes
dailyActivity_merged.csv	940	15	User-level daily summary (steps, distance, calories, activity minutes)	Main dataset
sleepDay_merged.csv	413	5	Sleep duration and time in bed per user/date	Contains duplicates
weightLogInfo_merged.csv	67	8	Weight, BMI, and timestamps	Only a subset of users logged weight

---

4. User Counts & Date Ranges

```
n_users_activity <- n_distinct(daily_activity$Id)
n_users_sleep    <- n_distinct(sleep_day$Id)
n_users_weight   <- n_distinct(weight_log$Id)

cat("Unique users - Activity:", n_users_activity,
    " Sleep:", n_users_sleep,
    " Weight:", n_users_weight, "\n")
```

```
## Unique users - Activity: 33  Sleep: 24  Weight: 8
```

```
# Convert date columns
daily_activity <- daily_activity %>%
  mutate(ActivityDate = mdy(ActivityDate))

sleep_day <- sleep_day %>%
  mutate(SleepDay = mdy_hms(SleepDay))

# Check date ranges
cat("Activity date range:", range(daily_activity$ActivityDate), "\n")
```

```
## Activity date range: 16903 16933
```

```
cat("Sleep date range:", range(sleep_day$SleepDay), "\n")
```

```
## Sleep date range: 1460419200 1463011200
```

---

5. Missing Values & Duplicates

```
# Missing values per column
cat("Missing values in daily_activity:\n")
```

```
## Missing values in daily_activity:
```

```
print(colSums(is.na(daily_activity)))
```

```
##                       Id             ActivityDate                TotalSteps
##                        0                        0                         0
##            TotalDistance          TrackerDistance LoggedActivitiesDistance
##                        0                        0                         0
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                        0                        0                         0
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                        0                        0                         0
##     LightlyActiveMinutes         SedentaryMinutes                 Calories
##                        0                        0                         0
```

```
cat("Missing values in sleep_day:\n")
```

```
## Missing values in sleep_day:
```

```
print(colSums(is.na(sleep_day)))
```

```
##                  Id           SleepDay  TotalSleepRecords TotalMinutesAsleep
##                   0                  0                  0                   0
##     TotalTimeInBed
##                   0
```

```
cat("Missing values in weight_log:\n")
```

```
## Missing values in weight_log:
```

```
print(colSums(is.na(weight_log)))
```

```
##             Id           Date       WeightKg   WeightPounds            Fat
##              0              0              0              0             65
##            BMI IsManualReport          LogId
##              0              0              0
```

```
# Duplicates
dup_sleep <- sum(duplicated(sleep_day))
cat("Number of duplicate rows in sleep_day:", dup_sleep, "\n")
```

```
## Number of duplicate rows in sleep_day: 3
```

Observations:

- No missing values in activity data.
- A few missing entries in weight logs.
- Duplicates found in sleep data → will be handled in Process Phase.

---

6. Data Quality Summary

Check	Result	Action
Unique users	33 (activity), 24 (sleep), 8 (weight)	Documented
Date format	Converted with `mdy()` and `mdy_hms()`	Ok
Date range	March–May 2016	Consistent
Missing values	Minimal (mainly weight)	Note
Duplicates	Found in `sleep_day`	Fix later
User overlap	Not all users appear across datasets	Noted

---

7. Limitations and Notes

- The dataset includes only ~30 participants over ~2 months.
- Limited generalizability to Bellabeat's female user base.
- Missing or incomplete data for certain features (especially weight).
- Despite limitations, data quality is sufficient for trend analysis.

```r
cat("Data successfully validated and ready for cleaning as of", Sys.Date(), "\n")
```

```
## Data successfully validated and ready for cleaning as of 20407
```

Analyst's Reflection

All datasets were imported and validated successfully. Activity and sleep data are robust enough to support meaningful behavioral insights. Weight data will be treated as supplementary due to its limited entries.