Bellabeat Case Study — Process (Clean) Phase

Francesco De Felice

2025-11-15

Contents

1. Purpose

This phase focuses on cleaning and preparing the datasets for analysis. Key actions include: - Removing duplicates - Handling missing values - Merging core datasets (daily_activity + sleep_day) - Creating clean, ready-to-analyze variables

Preload Datasets & Convert them for Safety

```
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")

# Convert date formats

daily_activity <- daily_activity %>%
mutate(ActivityDate = mdy(ActivityDate))
sleep_day <- sleep_day %>%
mutate(SleepDay = mdy_hms(SleepDay),
SleepDay = as_date(SleepDay))
```

2. Remove Duplicates

Check duplicates in sleep data

```
sum(duplicated(sleep_day))
```

```
## [1] 3
```

Remove duplicate rows

```
sleep_day <- sleep_day %>% distinct()
```

Confirm duplicates removed

```
sum(duplicated(sleep_day))
```

```
## [1] 0
```

All duplicates removed from the sleep dataset.

3. Handle Missing Values

Count missing values

```
# Check for missing values

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
```

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

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

If any NA found, drop or impute (none expected here)

```
daily_activity <- daily_activity %>% drop_na()
sleep_day      <- sleep_day %>% drop_na()
```

No significant missing data found — all clean.

4. Merge Key Datasets

We'll merge by matching each user's ID and corresponding date.

```
merged_data <- daily_activity %>%
rename(Date = ActivityDate) %>%
inner_join(
sleep_day %>% rename(Date = SleepDay),
by = c("Id", "Date")
)

glimpse(merged_data)
```

```
## Rows: 410
## Columns: 18
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ Date                     <date> 2016-04-12, 2016-04-13, 2016-04-15, 2016-04-~
## $ TotalSteps               <dbl> 13162, 10735, 9762, 12669, 9705, 15506, 10544~
## $ TotalDistance            <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68, 6.3~
## $ TrackerDistance          <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68, 6.3~
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.14, 2.71, 3.19, 3.53, 1.96, 1.3~
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 1.26, 0.41, 0.78, 1.32, 0.48, 0.3~
## $ LightActiveDistance      <dbl> 6.06, 4.71, 2.83, 5.04, 2.51, 5.03, 4.24, 4.6~
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveMinutes        <dbl> 25, 21, 29, 36, 38, 50, 28, 19, 41, 39, 73, 3~
## $ FairlyActiveMinutes      <dbl> 13, 19, 34, 10, 20, 31, 12, 8, 21, 5, 14, 23,~
## $ LightlyActiveMinutes     <dbl> 328, 217, 209, 221, 164, 264, 205, 211, 262, ~
## $ SedentaryMinutes         <dbl> 728, 776, 726, 773, 539, 775, 818, 838, 732, ~
## $ Calories                 <dbl> 1985, 1797, 1745, 1863, 1728, 2035, 1786, 177~
## $ TotalSleepRecords        <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TotalMinutesAsleep       <dbl> 327, 384, 412, 340, 700, 304, 360, 325, 361, ~
## $ TotalTimeInBed           <dbl> 346, 407, 442, 367, 712, 320, 377, 364, 384, ~
```

Merged dataset now contains both activity and sleep information.

5. Create Derived Variables

To enhance the analysis, create some useful features.

```
clean_data <- merged_data %>%
mutate(
TotalActiveMinutes = VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes,
SleepEfficiency = TotalMinutesAsleep / TotalTimeInBed,
Weekday = weekdays(Date)
) %>%
drop_na(TotalSteps, TotalMinutesAsleep, SleepEfficiency)

summary(select(clean_data, TotalActiveMinutes, SleepEfficiency))
```

```
##  TotalActiveMinutes SleepEfficiency
##  Min.   :  2.0      Min.   :0.4984
##  1st Qu.:206.5      1st Qu.:0.9118
##  Median :263.5      Median :0.9426
##  Mean   :259.5      Mean   :0.9165
##  3rd Qu.:315.5      3rd Qu.:0.9606
##  Max.   :540.0      Max.   :1.0000
```

New columns created: TotalActiveMinutes, SleepEfficiency, and Weekday.

6. Save Cleaned Dataset

```
write_csv(clean_data, "/cloud/project/Bellabeat Case Study/Data/Raw/clean_data.csv")
cat("  Cleaned dataset with weight info saved successfully on", Sys.Date(), "\n")
```

```
##   Cleaned dataset with weight info saved successfully on 20407
```

7. Quick Sanity Check

```
n_users <- n_distinct(clean_data$Id)
date_range <- range(clean_data$Date)
cat("Users:", n_users, "\n")
```

```
## Users: 24
```

```
cat("Date range:", date_range, "\n")
```

```
## Date range: 16903 16933
```

Clean dataset ready for analysis — consistent dates and users confirmed.

Analyst's Reflection The data is now clean, consistent, and merged. Derived variables add behavioral insight potential for the upcoming Analyze Phase. The dataset is ready for descriptive statistics, visualizations, and correlation checks.