Bellabeat Case Study : Data Transformation Report

M.S.KADRI

2022-07-03

Report summary:

1. Bellabeat Case Study
2. Data Processing Tools
3. Data Examination & Diagnosis
4. Data Transformation

1. Bellabeat Case Study

The Bellabeat case study is a data analysis project for a a high-tech manufacturer of health-focused products for women. The main objective of this data analysis project is to conduct a full and thorough analysis of smart device data to gain insight into how consumers are using their smart devices. The conducted analysis aims for extracting insights from the provided data to guide future marketing strategy for the Bellabeat marketing analytics team.

3. Data Processing Tools

For the data processing and analysis phases of this project a selection of data manipulation tools where used to examine the data, verify its integrity, and conduct data transformation operations to prepare it for the analysis phase.

The tools are: * **RStudio**: an IDE for R, a programming language for statistical computing and graphics. * **BigQuery**: a fully-managed, serverless data warehouse that enables scalable analysis over petabytes of data. * **Google Sheets**: a spreadsheet program included as part of the free, web-based Google Docs Editors suite offered by Google.

4. Data Examination & Diagnosis:

All of the 18 CSV files where examined and processed using RStudio, BigQuery, and Google Sheets. A summary of the diagnosis process is detailed in the table below.

Table 1: Table 1: Diagnosis Summary

Dataset file	missing values	duplicate values	inconsistent data type
dailyActivity_merged.csv	0	0	940
dailyCalories_merged.csv	0	0	940
dailyIntensities_merged.csv	0	0	940
dailySteps_merged.csv	0	0	940
heartrate_seconds_merged.csv	0	0	2483658
hourlyCalories_merged.csv	0	0	22099
hourlyIntensities_merged.csv	0	0	22099
hourlySteps_merged.csv	0	0	22099

Dataset file	missing values	duplicate values	inconsistent data type
minuteCaloriesNarrow_merged.csv	0	0	1325580
minuteCaloriesWide_merged.csv	0	0	21645
minuteIntensitiesNarrow_merged.csv	0	0	1325580
minuteIntensitiesWide_merged.csv	0	0	21645
minuteMETsNarrow_merged.csv	0	0	1325580
minuteSleep_merged.csv	0	0	188521
minuteStepsNarrow_merged.csv	0	0	1325580
minuteStepsWide_merged.csv	0	0	21645
sleepDay_merged.csv	0	0	413
weightLogInfo_merged.csv	65	0	67

Dataset file: dailyActivity_merged.csv
format: CSV file
Number of observations: 940
Number of variables: 15
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityDate" column observations are "char" type instead of "date" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: dailyCalories_merged.csv
format: CSV file
Number of observations: 940
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityDay" column observations are "char" type instead of "date" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: dailyIntensities_merged.csv
format: CSV file
Number of observations: 940
Number of variables: 10
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityDay" column observations are "char" type instead of "date" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: dailySteps_merged.csv
format: CSV file
Number of observations: 940
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityDay" column observations are "char" type instead of "date" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: heartrate_seconds_merged.csv
format: CSV file
Number of observations: 2483658
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "Time" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: hourlyCalories_merged.csv
format: CSV file
Number of observations: 22099
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityHour" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: hourlyIntensities_merged.csv
format: CSV file
Number of observations: 22099
Number of variables: 4
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityHour" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: hourlySteps_merged.csv
format: CSV file
Number of observations: 22099
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityHour" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: minuteCaloriesNarrow_merged.csv
format: CSV file
Number of observations: 1325580
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityMinute" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: minuteCaloriesWide_merged.csv
format: CSV file
Number of observations: 21645
Number of variables: 62
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityHour" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: minuteIntensitiesNarrow_merged.csv
format: CSV file
Number of observations: 1325580
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: “ActivityMinute” column observations are “char” type instead of “datetime” type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: minuteIntensitiesWide_merged.csv
format: CSV file
Number of observations: 21645
Number of variables: 62
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: “ActivityMinute” column observations are “char” type instead of “datetime” type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: minuteMETsNarrow_merged.csv
format: CSV file
Number of observations: 1325580
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: “ActivityMinute” column observations are “char” type instead of “datetime” type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: minuteSleep_merged.csv
format: CSV file
Number of observations: 188521
Number of variables: 4
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "date" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: minuteStepsNarrow_merged.csv
format: CSV file
Number of observations: 1325580
Number of variables: 3
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityMinute" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: minuteStepsWide_merged.csv
format: CSV file
Number of observations: 21645
Number of variables: 62
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityHour" column observations are "char" type instead of "datetime" type.

Data Range Validation
No issues.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: sleepDay_merged.csv
format: CSV file
Number of observations: 413
Number of variables: 5
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "ActivityHour" column observations ar "char" type instead of "datetime" type.

Data Range Validation
1 issue: - All Id observations has incomplete data range for the sleepday variable column. - The Id number 6962181067 has only one observation missing for the sleepday variable column.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

Dataset file: weightLogInfo_merged.csv
format: CSV file
Number of observations: 67
Number of variables: 8
Metadata availability: No

Data Structure Validation
No issues.

Data Type Validation
1 issue: "Date" column observations ar "char" type instead of "date" type.

Data Range Validation
1 issue: - All Id observations has incomplete data range for the sleepday variable column.

Data constraints Validation
No issues.

Data Consistency Validation
No issues.

---

4. Data Transformation

To fix the redundant issue of the invalid data type for the date/time columns in all the datasets, the dataset files were loaded to RStudio to change the datatype manually. Another major reason for using RStudio is the fact that BigQuery won't accept schemas with wrong data types format. And since BigQuery will be used for further data aggregation and extraction, the process is fixing the data type issue with R was the most practical approach.

Additionally, RStudio was used to verify the integrity of data and to check it for duplicates and missing values. The code snippet below was used to detect the presence of duplicates and missing values in the DailyActivity dataset and to get a quick examination of the data and its structure and a summary statistic.

```
DailyActivity_merged <- read.csv("data/dailyActivity_merged.csv")

DailyActivity_fixed <- DailyActivity_merged
DailyActivity_fixed$ActivityDate <- as.Date(DailyActivity_fixed$ActivityDate, format='%m/%d/%Y')

str(DailyActivity_fixed)
```

```
## 'data.frame':    940 obs. of  15 variables:
##  $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : Date, format: "2016-04-12" "2016-04-13" ...
##  $ TotalSteps               : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...
##  $ TotalDistance            : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance          : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance      : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes      : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes     : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes         : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ Calories                 : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

```
summary(DailyActivity_fixed)
```

```
##        Id              ActivityDate          TotalSteps    TotalDistance
##  Min.   :1.504e+09   Min.   :2016-04-12   Min.   :    0   Min.   : 0.000
##  1st Qu.:2.320e+09   1st Qu.:2016-04-19   1st Qu.: 3790   1st Qu.: 2.620
##  Median :4.445e+09   Median :2016-04-26   Median : 7406   Median : 5.245
##  Mean   :4.855e+09   Mean   :2016-04-26   Mean   : 7638   Mean   : 5.490
##  3rd Qu.:6.962e+09   3rd Qu.:2016-05-04   3rd Qu.:10727   3rd Qu.: 7.713
##  Max.   :8.878e+09   Max.   :2016-05-12   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.: 2.053
##  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.: 4.782      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
```

```
cat("Number of NA values: ", sum(is.na(DailyActivity_fixed)), "\n")
```

```
## Number of NA values:  0
```

```
cat("Number of distinct IDs: ", n_distinct(DailyActivity_fixed$Id), "\n")
```

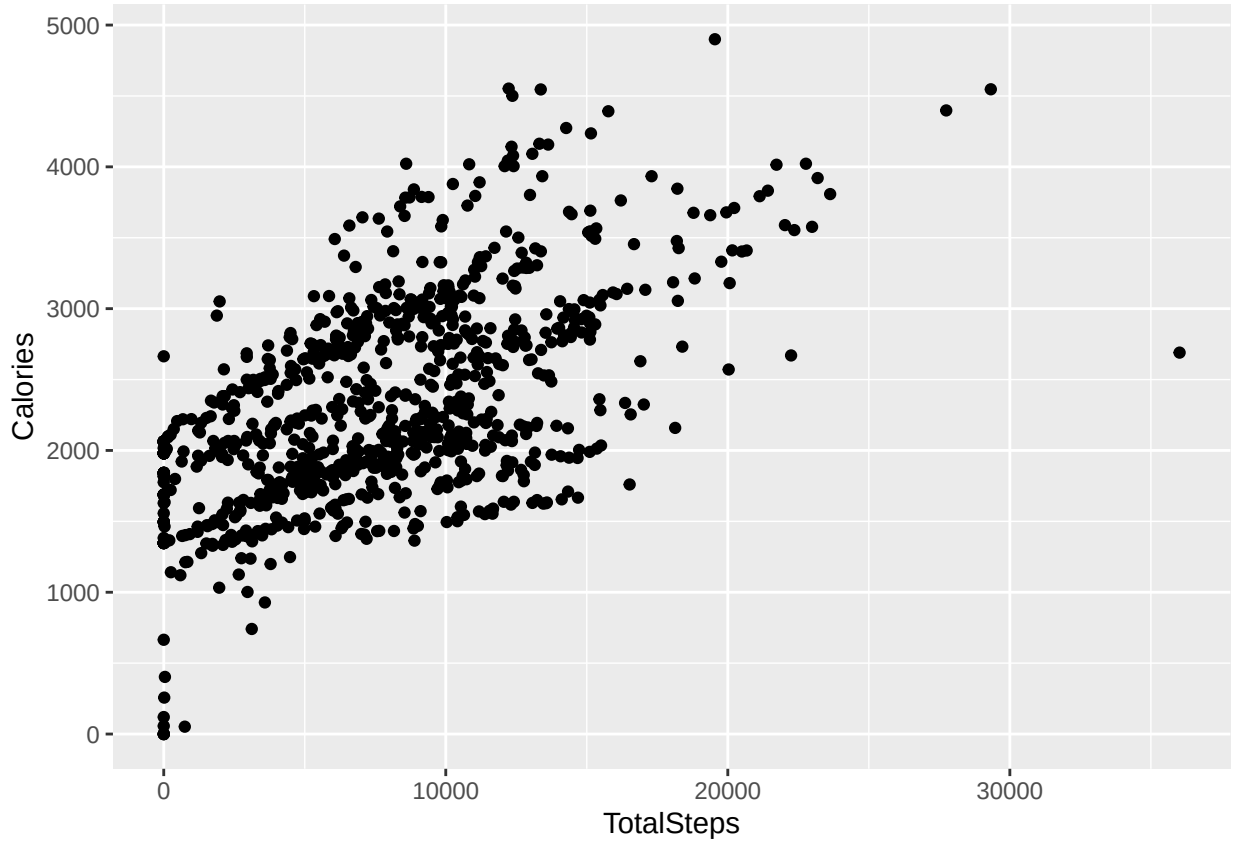
```
## Number of distinct IDs:  33
```

```
v1 <- duplicated(DailyActivity_fixed)
cat("Number of duplicate: ", length(v1[v1==TRUE]), "\n")
```

```
## Number of duplicate:  0
```

```
plot1 <- ggplot(data=DailyActivity_fixed, aes(x=TotalSteps, y=SedentaryActiveDistance)) + geom_point()
plot2 <- ggplot(data=DailyActivity_fixed, aes(x=TotalSteps, y=VeryActiveDistance)) + geom_point()
plot3 <- ggplot(data=DailyActivity_fixed, aes(x=TotalSteps, y=Calories)) + geom_point()

plot3
```



After fixing the dataset files and saving the fixed versions in a different folder, the fixed datasets were loaded to bigQuery and Google Spreadsheets for further analysis.

Consequently, and after the data extraction process, Google Spreadsheet was used to generate pivot tables and charts for data visualization.

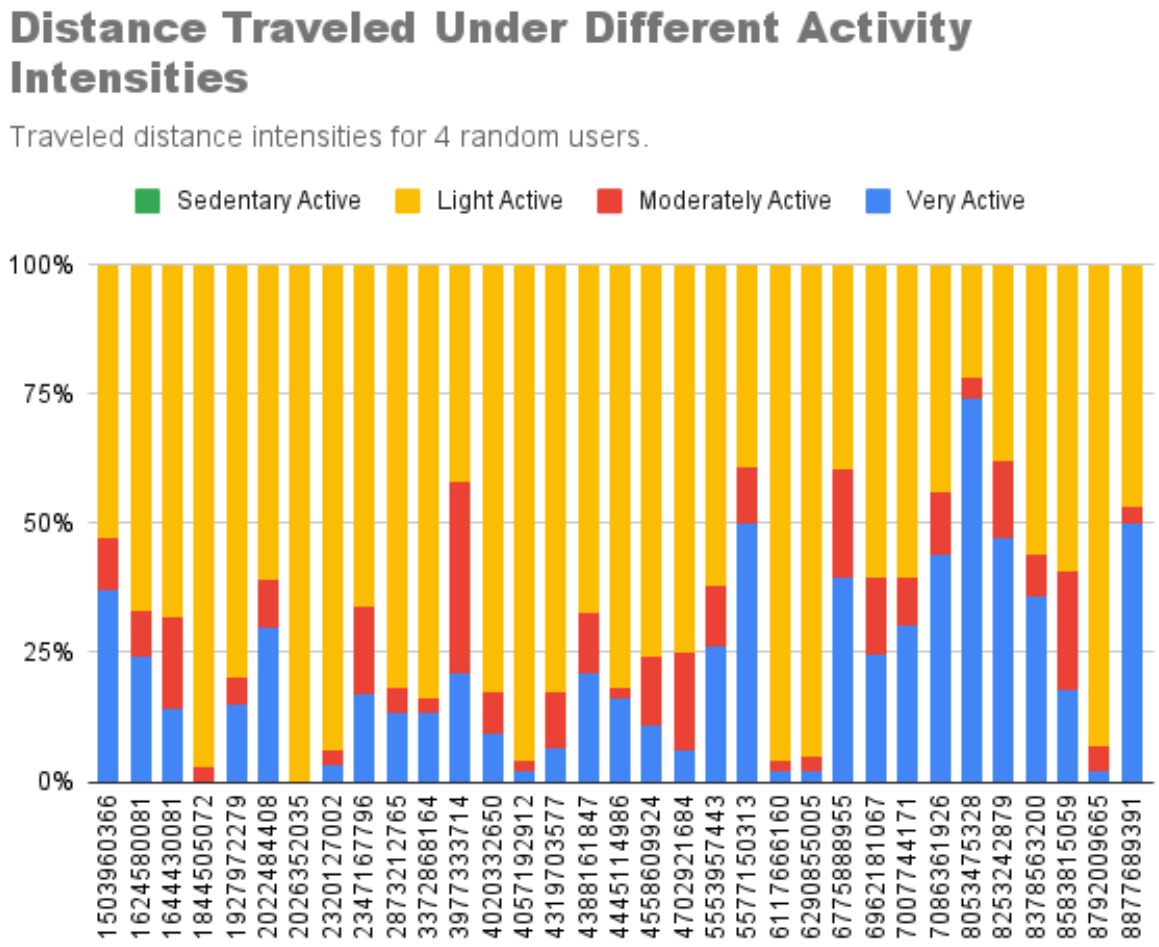


Figure 1: Chart generated using Google Spreadsheet