The dataset used in this analysis is titled **“dailyActivity\_merged.csv”**, which forms the foundation of the **Bellabeat data analysis case study**. The following sections present the detailed analytics derived from this dataset.

Table 1 presents the foundational structure and quality of the dataset used for analysis. It contains a total of **940 records** representing daily activity logs from **33 unique users** collected between **April 12 and May 9, 2016**. The table highlights that there are **no missing values** across any of the key variables, such as Id, Activity\_Date, Total\_Steps, Total\_Distance, and Calories, among others. This completeness ensures high data integrity, meaning each record is valid and reliable for further statistical analysis. The dataset provides a consistent foundation for evaluating user activity patterns without requiring data imputation or correction for missing entries.

Table 1: Dataset Overview

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Value / Observation** |
| **Number of Records** | The total number of rows (entries) in the dataset, representing daily activity logs. | **940 records** |
| **Number of Unique Users** | The number of distinct users (identified by unique IDs) represented in the dataset. | **33 users** |
| **Date Range** | The period over which the activity data was collected. | **April 12, 2016 – May 9, 2016** |
| **Missing Values (Id)** | Indicates if any user IDs are missing from the dataset. | **0 (none missing)** |
| **Missing Values (Activity\_Date)** | Checks for missing or invalid date entries. | **0 (none missing)** |
| **Missing Values (Total\_Steps)** | Ensures completeness of the step count data. | **0 (none missing)** |
| **Missing Values (Total\_Distance)** | Verifies if distance data (in miles or km) is complete. | **0 (none missing)** |
| **Missing Values (Tracker\_Distance)** | Confirms if tracker-recorded distances are complete. | **0 (none missing)** |
| **Missing Values (Logged\_Activities\_Distance)** | Shows if manually logged activities have missing distances. | **0 (none missing)** |
| **Missing Values (Very\_Active\_Distance)** | Ensures no gaps in recorded distances during very active minutes. | **0 (none missing)** |
| **Missing Values (Moderately\_Active\_Distance)** | Checks for missing data during moderate activity sessions. | **0 (none missing)** |
| **Missing Values (Light\_Active\_Distance)** | Validates data completeness for light activity distances. | **0 (none missing)** |
| **Missing Values (Sedentary\_Active\_Distance)** | Ensures sedentary distance entries are complete. | **0 (none missing)** |
| **Missing Values (Very\_Active\_Minutes)** | Verifies data availability for minutes spent in intense activity. | **0 (none missing)** |
| **Missing Values (Fairly\_Active\_Minutes)** | Confirms no missing entries for fairly active minutes. | **0 (none missing)** |
| **Missing Values (Lightly\_Active\_Minutes)** | Ensures completeness for lightly active minutes. | **0 (none missing)** |
| **Missing Values (Sedentary\_Minutes)** | Checks if sedentary time data is fully available. | **0 (none missing)** |
| **Missing Values (Calories)** | Confirms completeness of calories burned data. | **0 (none missing)** |

Table 2 describes the additional variables generated during preprocessing to enhance the analytical depth of the dataset. New columns, such as **Weekday** and **Week\_Number,** help capture time-based variations in user activity. The **Total\_ActiveMinutes** and **Total\_Active\_Hours** fields aggregate different activity levels into a single measure of overall engagement. Similarly, **Sedentary\_Hours** and **Activity\_Ratio** quantify inactivity and its proportion relative to daily activity time. The **Is Weekend** indicator distinguishes between weekday and weekend behaviour, allowing for temporal comparisons. These derived metrics simplify complex raw data and enable trend analysis and user segmentation.

Table 2: Derived Columns and Data Cleaning Summary

|  |  |  |
| --- | --- | --- |
| **New Column** | **Description** | **Computation / Formula** |
| **Weekday** | Day of the week for each record (e.g., Monday, Tuesday). | weekdays(Activity\_Date) |
| **Week\_Number** | Calendar week number (used for weekly trend analysis). | week(Activity\_Date) |
| **Total\_Active\_Minutes** | Total minutes spent in all active states per day. | Very\_Active\_Minutes + Fairly\_Active\_Minutes + Lightly\_Active\_Minutes |
| **Total\_Active\_Hours** | Total active minutes converted into hours. | Total\_Active\_Minutes / 60 |
| **Sedentary\_Hours** | Total sedentary minutes converted into hours. | Sedentary\_Minutes / 60 |
| **Activity\_Ratio** | Proportion of active time to total daily time (active + sedentary). | Total\_Active\_Minutes / (Total\_Active\_Minutes + Sedentary\_Minutes) |
| **IsWeekend** | Logical indicator of weekend activity (Saturday/Sunday = TRUE). | Weekday %in% c("Saturday", "Sunday") |

Table 3 outlines the data cleaning actions performed to ensure analytical accuracy. Records identified as **non-wear days** where users registered **0 steps and 1,440 sedentary minutes (24 hours)** were removed since such entries likely indicate that the device was not worn. A total of **72 records** were discarded, leaving **868 valid records**, representing **92.3%** of the dataset. This cleaning step minimizes noise and ensures that only genuine daily activity logs are included, improving the reliability of subsequent descriptive and inferential analyses.

Table 3: **Data Cleaning Summary**

|  |  |  |
| --- | --- | --- |
| **Description** | **Action Taken** | **Outcome** |
| **Non-wear days removal** | Records where users logged 0 steps and were sedentary for 1440 minutes (24 hours) were removed since the device likely wasn’t worn. | **72 records removed** |
| **Remaining valid records** | Number of entries after cleaning. | **868 valid records** |
| **Percentage of valid data** | Portion of retained records after cleaning. | **92.3% valid records** |

Table 4 defines the metrics used to assess individual user engagement levels. It includes measures such as **Record\_Days** (number of valid days logged per user), **Avg\_Steps**, **Avg\_Calories**, **Avg\_Sedentary\_Hours**, and **Avg\_Active\_Hours**, which collectively describe activity consistency and intensity. Additionally, the **Max\_Steps** metric identifies each user’s peak performance, while the **Consistency** measure, calculated as the ratio of the standard deviation to the mean of steps indicates behavioural steadiness. These indicators help classify users based on how regularly and actively they engage in physical movement over time.

Table 4: **User Engagement Analysis**

|  |  |
| --- | --- |
| **Metric** | **Description** |
| **Record\_Days** | Number of valid activity days recorded per user. |
| **Avg\_Steps** | Average daily steps per user (indicator of activity level). |
| **Avg\_Calories** | Average daily calories burned. |
| **Avg\_Sedentary\_Hours** | Mean daily sedentary hours per user. |
| **Avg\_Active\_Hours** | Mean daily active hours per user. |
| **Max\_Steps** | The highest number of steps taken by each user in a day. |
| **Consistency** | Measures variation in steps (sd\_(TotalSteps)/mean(Total\_Steps)); lower values indicate steadier behaviour. |

Table 5 categorizes users into four distinct activity levels based on their average daily step counts. **Highly Active** users (≥10,000 steps/day) represent **7 individuals**, showcasing high fitness engagement. **Moderately Active** users (7,500–9,999 steps/day) account for **10 individuals**, indicating regular movement consistent with moderate exercise. **Lightly Active** users (5,000–7,499 steps/day) include **9 participants**, reflecting low-to-average mobility. The **Sedentary** group (<5,000 steps/day), also with **7 users**, displays minimal physical activity. This classification highlights the varying lifestyle behaviours within the participant pool.

Table 5: **User Activity Level Distribution**

|  |  |  |  |
| --- | --- | --- | --- |
| **Activity Level** | **Criteria (Avg\_Steps)** | **Number of Users** | **Interpretation** |
| **Highly Active** | ≥ 10,000 steps/day | **7 users** | Very physically active; likely engaged in regular exercise. |
| **Moderately Active** | 7,500–9,999 steps/day | **10 users** | Maintain moderate daily movement; active lifestyle. |
| **Lightly Active** | 5,000–7,499 steps/day | **9 users** | Some movement but below recommended daily step goal. |
| **Sedentary** | < 5,000 steps/day | **7 users** | Minimal physical activity; mostly inactive behavior. |

Table 6 provides summary statistics that give a broad view of general user activity levels. On average, users recorded **8,271 steps per day**, with a median of **7,969 steps**, indicating moderate activity. The mean **calorie expenditure** was **2,353 kcal** daily, while the average **sedentary time** reached **953.9 minutes (15.9 hours)** per day. In contrast, users averaged only **22.9 minutes of vigorous activity** and **14.7 minutes of moderate activity** daily. These figures suggest that while users engage in some form of movement, their daily routines are predominantly sedentary.

Table 6: Descriptive Statistics Summary

|  |  |  |
| --- | --- | --- |
| **Metric** | **Description** | **Value** |
| **Average Steps** | Mean daily steps across all valid records. | **8,271 steps** |
| **Median Steps** | Middle value of daily steps distribution. | **7,969 steps** |
| **Average Calories** | Mean daily calories burned per user. | **2,353 kcal** |
| **Average Very Active Minutes** | Time spent in vigorous activity per day. | **22.9 minutes** |
| **Average Fairly Active Minutes** | Time spent in moderate activity per day. | **14.7 minutes** |
| **Average Lightly Active Minutes** | Time spent in light activities per day. | **208.8 minutes** |
| **Average Sedentary Minutes** | Total minutes spent inactive per day. | **953.9 minutes** |
| **Average Sedentary Hours** | Converted sedentary time in hours. | **15.9 hours/day** |

Table 7 breaks down daily time allocation across different activity intensities. It reveals that users spend approximately **66.2% of their day in sedentary states**, around **14.5% in light activity**, and less than **3% in moderate or vigorous activity** combined. Specifically, only **1.6%** of the day is spent in very active minutes, while **1.0%** involves moderate activity. This distribution emphasizes a lifestyle pattern heavily dominated by inactivity, pointing to potential health implications associated with prolonged sedentary behavior.

Table 7: Activity Composition

|  |  |  |  |
| --- | --- | --- | --- |
| **Activity Type** | **Average Minutes** | **Average Hours** | **Percentage of Day** |
| **Sedentary** | 953.98 | 15.9 | **66.2%** |
| **Lightly Active** | 208.81 | 3.5 | **14.5%** |
| **Fairly Active** | 14.69 | 0.25 | **1.0%** |
| **Very Active** | 22.92 | 0.38 | **1.6%** |

Table 8 identifies the strongest relationship between key variables in the dataset—**Total\_Steps** and **Calories burned**—with a **correlation coefficient (r) of 0.567**. This moderate positive correlation indicates that higher step counts generally correspond with greater calorie expenditure. The relationship validates the consistency of fitness tracker data and aligns with established physiological principles linking movement intensity and energy consumption.

**Table 8: Correlation Analysis of Activity Metrics**

|  |  |  |  |
| --- | --- | --- | --- |
| **Strongest Relationship** | **Variables** | **Correlation Coefficient (r)** | **Interpretation** |
| **Steps vs Calories** | Total\_Steps ↔ Calories | **0.567** | Moderate positive correlation: more steps generally lead to higher calorie expenditure. |

Table 9 compares user behaviour between weekdays and weekends. During Monday–Friday, the average step count is 8,289, slightly higher than 8,222 on Saturday–Sunday. Calorie expenditure remains nearly identical, with only minor fluctuations. Sedentary hours show a marginal decline on weekends (15.6 hours) compared to weekdays (16.0 hours). Statistical testing reveals no significant difference, suggesting that most participants maintain consistent activity patterns throughout the week, without major behavioural changes on weekends.

**Table 9: Weekly Pattern Overview**

|  |  |  |  |
| --- | --- | --- | --- |
| **Weekday** | **Average Steps** | **Average Calories** | **Average Sedentary Hours** |
| **Monday–Friday (Weekdays)** | **8,289** | **2,350** | **16.0** |
| **Saturday–Sunday (Weekends)** | **8,222** | **2,362** | **15.6** |

Table 10 groups users based on a combination of step counts and sedentary hours to better understand lifestyle patterns. The Active & Balanced group (3%) demonstrates high step counts with limited sedentary hours, indicating optimal fitness balance. The Active but Sedentary group (18.2%) shows good movement but excessive sitting time. The majority (57.6%) are Moderately Active, maintaining acceptable movement levels but not fully meeting fitness guidelines. The Less Active group (21.2%) exhibits low mobility and minimal engagement. These insights help identify target groups for health interventions.

**Table 10: User Segmentation Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Segment** | **Criteria** | **Count** | **Percentage** | **Behavior Description** |
| **Active & Balanced** | ≥10,000 steps & ≤12 sedentary hours | 1 | **3.0%** | High activity and balanced rest pattern. |
| **Active but Sedentary** | ≥10,000 steps & >12 sedentary hours | 6 | **18.2%** | Physically active but spends long inactive periods. |
| **Moderately Active** | 5,000–9,999 steps | 19 | **57.6%** | Moderate movement; close to fitness guidelines. |
| **Less Active** | <5,000 steps | 7 | **21.2%** | Limited mobility; low engagement. |

Table 11 summarizes how total active time is distributed across different intensity levels. The results show that **Lightly Active minutes** dominate with **84.7%**, while **Fairly Active** and **Very Active** minutes contribute only **6.0%** and **9.3%**, respectively. This imbalance highlights the predominance of low-intensity activities such as casual walking or household chores, with limited engagement in more physically demanding exercises.

Table 11: Activity Intensity Distribution

|  |  |
| --- | --- |
| **Intensity Level** | **Percentage of Total Active Time** |
| **Very Active** | **9.3%** |
| **Fairly Active** | **6.0%** |
| **Lightly Active** | **84.7%** |

Table 12 provides a snapshot of two contrasting user profiles. The **most active user (ID 8877689391)** averages **16,040 steps** and **3,420 kcal** per day, categorizing them as **Highly Active**, likely engaging in regular exercise or extended physical routines. Conversely, the **least active user (ID 1927972279)** averages only **1,578 steps** and **2,252 kcal** daily, fitting the **Sedentary** profile. This contrast demonstrates the diversity in physical activity behaviours among users and the potential influence of lifestyle or occupation on fitness patterns.

Table 12: Sample User Profiles

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Profile Type** | **User ID** | **Avg Steps** | **Avg Calories** | **Activity Level** | **Interpretation** |
| **Most Active User** | 8877689391 | **16,040 steps** | **3,420 kcal** | Highly Active | Consistently exceeds fitness benchmarks, reflecting excellent physical engagement. |
| **Least Active User** | 1927972279 | **1,578 steps** | **2,252 kcal** | Sedentary | Very low movement despite normal calorie intake, indicating minimal physical activity. |

Table 13 replicates the same user comparison as Table 12, reinforcing the distinction between high and low activity users. The repetition may serve as a verification step or as part of a visual summary in reporting. It underlines the disparity between users’ movement patterns, showing that even within a small population, activity levels can vary dramatically.

Table 13: Sample User Profiles

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Profile Type** | **User ID** | **Avg Steps** | **Avg Calories** | **Activity Level** | **Interpretation** |
| **Most Active User** | 8877689391 | **16,040 steps** | **3,420 kcal** | Highly Active | Consistently exceeds fitness benchmarks, reflecting excellent physical engagement. |
| **Least Active User** | 1927972279 | **1,578 steps** | **2,252 kcal** | Sedentary | Very low movement despite normal calorie intake, indicating minimal physical activity. |

Table 14 consolidates the overall dataset insights and classifications. After data cleaning, **868 valid records** (92.3%) from **33 unique users** were retained. The analysis period spanned **April 12 to May 12, 2016**, and users were categorized into **Highly Active (7)**, **Moderately Active (10)**, **Lightly Active (9)**, and **Sedentary (7)** groups. The most prevalent segment was **Moderately Active**, accounting for **57.6%** of users. This summary encapsulates the general finding that most participants maintain moderate physical activity, with fewer individuals achieving the recommended high-activity standards.

**Table 14: Summary Report**

|  |  |
| --- | --- |
| **Category** | **Details** |
| **Total Records** | 940 |
| **Valid Records** | 868 (92.3%) |
| **Unique Users** | 33 |
| **Date Range** | April 12, 2016 – May 12, 2016 |
| **Highly Active Users** | 7 |
| **Moderately Active Users** | 10 |
| **Lightly Active Users** | 9 |
| **Sedentary Users** | 7 |
| **Most Common Segment** | Moderately Active (57.6%) |

**Distribution Plots**

Figure 1 shows the frequency distribution of total daily steps across all users and days, with a vertical red line marking the 10,000-steps health benchmark. The data is right-skewed: most days fall below 10,000 steps (peak frequency around 5,000–8,000 steps), with fewer extreme high-activity days (up to 30,000 steps) and very few low-activity days

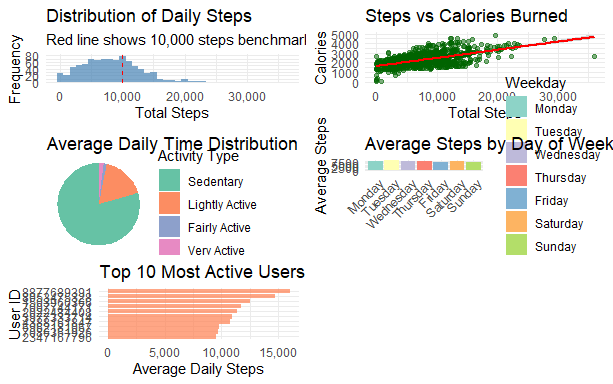


Figure 1: Distribution of Daily Steps

Figure 2 shows the heatmap visualizes the Pearson correlation coefficients (ranging from -1 to +1) between seven key daily activity metrics: Total Steps, Total Distance, Very Active Minutes, Fairly Active Minutes, Lightly Active Minutes, Sedentary Minutes, and Calories. Dark blue indicates strong positive correlations (e.g., Total Steps and Calories at ~0.98, showing steps strongly predict calorie burn), light blue or white shows moderate/weak positive links (e.g., Very Active Minutes and Calories at 0.4), and orange/red indicates negative correlations (e.g., Sedentary Minutes and Calories at -0.6, as more sitting time reduces energy expenditure). The diagonal (1.0) represents perfect self-correlation. Overall, it reveals that active metrics cluster together positively, while sedentary time opposes them, highlighting trade-offs in daily behavior

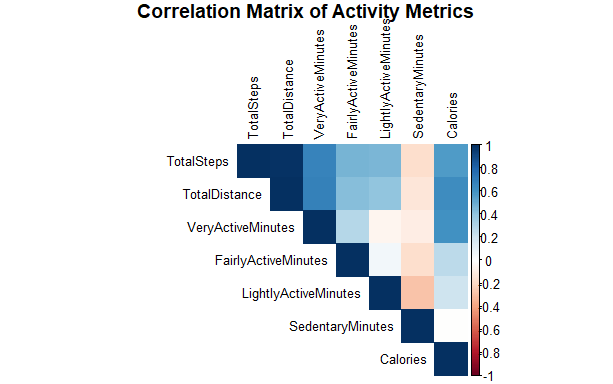


Figure 2:Correlation Matrix of Activity Metrics

Figure 3 shows this dual-axis line chart tracks cohort-wide averages from April 11 to May 9: average steps (blue line, left y-axis, 0–10,000) and calories (red line, right y-axis, 0–500k scaled). Steps fluctuate (peaking ~9,000 on April 25, dropping to ~6,000 by May 9), mirroring calories (peaking ~450k, dropping to ~300k). The synchronized decline post-April suggests seasonal fatigue, external events, or data gaps, with calories consistently ~50x steps due to scaling—urging retention strategies to reverse the downward trend

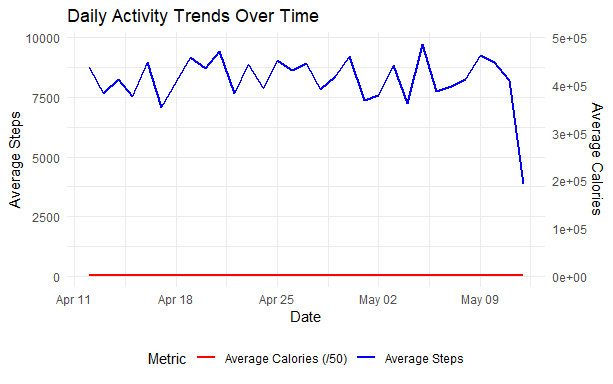


Figure 3: Daily Activity Trends Over Time Line Chart

Figure 4 shows this multi-line chart plots daily steps (0–30,000) over the same April–May period for the top 5 users (color-coded by ID: red=150390366, yellow=20224840, etc.). Lines show high variability e.g., User 88776893 (pink) spikes to 28,000 on April 25 but dips below 10,000 later revealing inconsistent patterns despite high averages. No single user dominates consistently, indicating diverse activity drivers; peaks align with cohort trends (Figure 7), while troughs highlight personalization needs for sustained engagement.

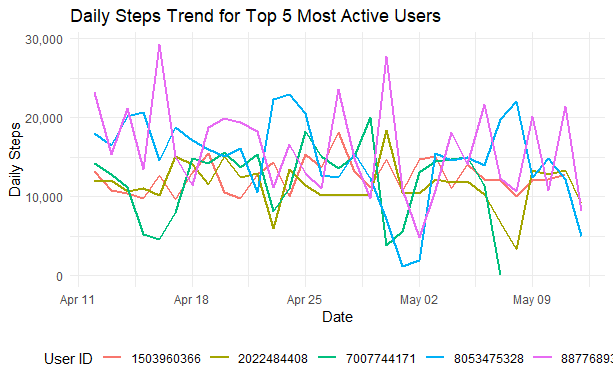


Figure 4 :Daily Steps Trend for Top 5 Most Active Users Line Chart

APPENDIX I

R CODE

# Load required libraries

library(tidyverse)

library(lubridate)

library(ggplot2)

library(dplyr)

library(tidyr)

library(corrplot)

library(scales)

library(ggpubr)

library(knitr)

library(kableExtra)

# Set theme for better visualizations

theme\_set(theme\_minimal())

# Read the data

activity\_data <- read.csv("dailyActivity\_merged.csv")

# Data Exploration and Cleaning

cat("=== DATASET OVERVIEW ===\n")

cat("Number of records:", nrow(activity\_data), "\n")

cat("Number of unique users:", n\_distinct(activity\_data$Id), "\n")

cat("Date range:", min(activity\_data$ActivityDate), "to", max(activity\_data$ActivityDate), "\n")

# Convert ActivityDate to proper date format

activity\_data$ActivityDate <- as.Date(activity\_data$ActivityDate, format = "%m/%d/%Y")

# Check for missing values

cat("\nMissing values in each column:\n")

print(colSums(is.na(activity\_data)))

# Create additional useful columns

activity\_data <- activity\_data %>%

mutate(

Weekday = weekdays(ActivityDate),

WeekNumber = week(ActivityDate),

TotalActiveMinutes = VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes,

TotalActiveHours = TotalActiveMinutes / 60,

SedentaryHours = SedentaryMinutes / 60,

ActivityRatio = TotalActiveMinutes / (TotalActiveMinutes + SedentaryMinutes),

IsWeekend = Weekday %in% c("Saturday", "Sunday")

)

# Remove days where device likely wasn't worn (0 steps and 1440 sedentary minutes)

valid\_data <- activity\_data %>%

filter(!(TotalSteps == 0 & SedentaryMinutes == 1440))

cat("\nRecords after removing non-wear days:", nrow(valid\_data), "\n")

cat("Percentage of valid records:", round(nrow(valid\_data)/nrow(activity\_data)\*100, 1), "%\n")

# 1. USER ENGAGEMENT ANALYSIS

cat("\n=== USER ENGAGEMENT ANALYSIS ===\n")

user\_summary <- valid\_data %>%

group\_by(Id) %>%

summarise(

RecordDays = n(),

AvgSteps = mean(TotalSteps),

AvgCalories = mean(Calories),

AvgSedentaryHours = mean(SedentaryHours),

AvgActiveHours = mean(TotalActiveHours),

MaxSteps = max(TotalSteps),

Consistency = sd(TotalSteps) / mean(TotalSteps) # Lower = more consistent

) %>%

mutate(

ActivityLevel = case\_when(

AvgSteps >= 10000 ~ "Highly Active",

AvgSteps >= 7500 ~ "Moderately Active",

AvgSteps >= 5000 ~ "Lightly Active",

TRUE ~ "Sedentary"

)

)

cat("\nUser Activity Level Distribution:\n")

print(table(user\_summary$ActivityLevel))

# 2. DESCRIPTIVE STATISTICS

cat("\n=== DESCRIPTIVE STATISTICS ===\n")

overall\_stats <- valid\_data %>%

summarise(

AvgSteps = mean(TotalSteps),

MedianSteps = median(TotalSteps),

AvgCalories = mean(Calories),

AvgVeryActiveMinutes = mean(VeryActiveMinutes),

AvgFairlyActiveMinutes = mean(FairlyActiveMinutes),

AvgLightlyActiveMinutes = mean(LightlyActiveMinutes),

AvgSedentaryMinutes = mean(SedentaryMinutes),

AvgSedentaryHours = mean(SedentaryHours)

)

print(overall\_stats)

# 3. VISUALIZATIONS

# 3.1 Distribution of Daily Steps

p1 <- ggplot(valid\_data, aes(x = TotalSteps)) +

geom\_histogram(binwidth = 1000, fill = "steelblue", alpha = 0.7) +

geom\_vline(xintercept = 10000, linetype = "dashed", color = "red") +

labs(title = "Distribution of Daily Steps",

subtitle = "Red line shows 10,000 steps benchmark",

x = "Total Steps", y = "Frequency") +

scale\_x\_continuous(labels = comma)

# 3.2 Steps vs Calories

p2 <- ggplot(valid\_data, aes(x = TotalSteps, y = Calories)) +

geom\_point(alpha = 0.5, color = "darkgreen") +

geom\_smooth(method = "lm", color = "red") +

labs(title = "Steps vs Calories Burned",

x = "Total Steps", y = "Calories") +

scale\_x\_continuous(labels = comma)

# 3.3 Activity Intensity Composition

activity\_minutes <- valid\_data %>%

select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes) %>%

summarise\_all(mean) %>%

pivot\_longer(cols = everything(), names\_to = "ActivityType", values\_to = "Minutes") %>%

mutate(

ActivityType = factor(ActivityType,

levels = c("SedentaryMinutes", "LightlyActiveMinutes",

"FairlyActiveMinutes", "VeryActiveMinutes")),

Hours = Minutes / 60,

Percentage = Minutes / (24 \* 60) \* 100

)

p3 <- ggplot(activity\_minutes, aes(x = "", y = Hours, fill = ActivityType)) +

geom\_bar(stat = "identity", width = 1) +

coord\_polar("y", start = 0) +

labs(title = "Average Daily Time Distribution",

fill = "Activity Type") +

scale\_fill\_brewer(palette = "Set2",

labels = c("Sedentary", "Lightly Active",

"Fairly Active", "Very Active")) +

theme\_void()

# 3.4 Weekly Patterns

weekly\_patterns <- valid\_data %>%

group\_by(Weekday) %>%

summarise(

AvgSteps = mean(TotalSteps),

AvgCalories = mean(Calories),

AvgSedentaryHours = mean(SedentaryHours)

) %>%

mutate(Weekday = factor(Weekday,

levels = c("Monday", "Tuesday", "Wednesday", "Thursday",

"Friday", "Saturday", "Sunday")))

p4 <- ggplot(weekly\_patterns, aes(x = Weekday, y = AvgSteps, fill = Weekday)) +

geom\_col() +

labs(title = "Average Steps by Day of Week",

x = "", y = "Average Steps") +

scale\_fill\_brewer(palette = "Set3") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

# 3.5 Top 10 Most Active Users

top\_users <- user\_summary %>%

arrange(desc(AvgSteps)) %>%

head(10)

p5 <- ggplot(top\_users, aes(x = reorder(as.character(Id), AvgSteps), y = AvgSteps)) +

geom\_col(fill = "coral", alpha = 0.7) +

coord\_flip() +

labs(title = "Top 10 Most Active Users",

x = "User ID", y = "Average Daily Steps") +

scale\_y\_continuous(labels = comma)

# Arrange plots

grid\_plots <- ggarrange(p1, p2, p3, p4, p5, ncol = 2, nrow = 3)

print(grid\_plots)

# 4. CORRELATION ANALYSIS

cat("\n=== CORRELATION ANALYSIS ===\n")

correlation\_data <- valid\_data %>%

select(TotalSteps, TotalDistance, VeryActiveMinutes, FairlyActiveMinutes,

LightlyActiveMinutes, SedentaryMinutes, Calories)

cor\_matrix <- cor(correlation\_data)

# Correlation plot

corrplot(cor\_matrix, method = "color", type = "upper",

tl.cex = 0.8, tl.col = "black",

title = "Correlation Matrix of Activity Metrics",

mar = c(0,0,1,0))

# 5. TRENDS OVER TIME

cat("\n=== TIME TREND ANALYSIS ===\n")

# Daily averages over time

daily\_trends <- valid\_data %>%

group\_by(ActivityDate) %>%

summarise(

AvgSteps = mean(TotalSteps),

AvgCalories = mean(Calories),

AvgSedentaryHours = mean(SedentaryHours),

UserCount = n()

)

p6 <- ggplot(daily\_trends, aes(x = ActivityDate)) +

geom\_line(aes(y = AvgSteps, color = "Average Steps"), size = 1) +

geom\_line(aes(y = AvgCalories/50, color = "Average Calories (/50)"), size = 1) +

scale\_y\_continuous(

name = "Average Steps",

sec.axis = sec\_axis(~ . \* 50, name = "Average Calories")

) +

labs(title = "Daily Activity Trends Over Time",

x = "Date", color = "Metric") +

scale\_color\_manual(values = c("Average Steps" = "blue",

"Average Calories (/50)" = "red")) +

theme(legend.position = "bottom")

print(p6)

# 6. USER SEGMENTATION ANALYSIS

cat("\n=== USER SEGMENTATION ===\n")

# Create user segments based on activity patterns

user\_segments <- user\_summary %>%

mutate(

Segment = case\_when(

AvgSteps >= 10000 & AvgSedentaryHours <= 12 ~ "Active & Balanced",

AvgSteps >= 10000 & AvgSedentaryHours > 12 ~ "Active but Sedentary",

AvgSteps >= 5000 & AvgSteps < 10000 ~ "Moderately Active",

AvgSteps < 5000 ~ "Less Active",

TRUE ~ "Other"

)

)

cat("User Segmentation:\n")

segment\_summary <- user\_segments %>%

group\_by(Segment) %>%

summarise(Count = n(), Percentage = round(n()/nrow(user\_segments)\*100, 1))

print(segment\_summary)

# 7. INTENSITY ANALYSIS

cat("\n=== ACTIVITY INTENSITY ANALYSIS ===\n")

intensity\_analysis <- valid\_data %>%

summarise(

TotalVeryActivePct = sum(VeryActiveMinutes) / sum(TotalActiveMinutes) \* 100,

TotalFairlyActivePct = sum(FairlyActiveMinutes) / sum(TotalActiveMinutes) \* 100,

TotalLightlyActivePct = sum(LightlyActiveMinutes) / sum(TotalActiveMinutes) \* 100

)

cat("Activity Intensity Distribution (% of total active time):\n")

print(intensity\_analysis)

# 8. WEEKDAY VS WEEKEND COMPARISON

cat("\n=== WEEKDAY VS WEEKEND ANALYSIS ===\n")

weekend\_comparison <- valid\_data %>%

group\_by(IsWeekend) %>%

summarise(

AvgSteps = mean(TotalSteps),

AvgCalories = mean(Calories),

AvgSedentaryHours = mean(SedentaryHours),

AvgVeryActiveMinutes = mean(VeryActiveMinutes),

n = n()

) %>%

mutate(DayType = ifelse(IsWeekend, "Weekend", "Weekday"))

print(weekend\_comparison)

# Statistical test for difference

weekday\_data <- valid\_data %>% filter(!IsWeekend) %>% pull(TotalSteps)

weekend\_data <- valid\_data %>% filter(IsWeekend) %>% pull(TotalSteps)

t\_test\_result <- t.test(weekday\_data, weekend\_data)

cat("\nT-test for steps (Weekday vs Weekend):\n")

cat("p-value:", t\_test\_result$p.value, "\n")

# 9. DETAILED USER PROFILES

cat("\n=== SAMPLE USER PROFILES ===\n")

# Get most active user

most\_active\_user <- user\_summary %>%

arrange(desc(AvgSteps)) %>%

head(1)

most\_active\_data <- valid\_data %>%

filter(Id == most\_active\_user$Id)

cat("Most Active User Profile (ID:", most\_active\_user$Id, "):\n")

cat("Average Steps:", round(most\_active\_user$AvgSteps), "\n")

cat("Average Calories:", round(most\_active\_user$AvgCalories), "\n")

cat("Activity Level:", most\_active\_user$ActivityLevel, "\n")

# Get least active user (among those with reasonable data)

least\_active\_user <- user\_summary %>%

filter(RecordDays >= 10) %>% # At least 10 days of data

arrange(AvgSteps) %>%

head(1)

cat("\nLeast Active User Profile (ID:", least\_active\_user$Id, "):\n")

cat("Average Steps:", round(least\_active\_user$AvgSteps), "\n")

cat("Average Calories:", round(least\_active\_user$AvgCalories), "\n")

cat("Activity Level:", least\_active\_user$ActivityLevel, "\n")

# 10. EXPORT SUMMARY STATISTICS

cat("\n=== SUMMARY REPORT ===\n")

summary\_report <- list(

Dataset = list(

TotalRecords = nrow(activity\_data),

ValidRecords = nrow(valid\_data),

UniqueUsers = n\_distinct(activity\_data$Id),

DateRange = paste(min(activity\_data$ActivityDate), "to", max(activity\_data$ActivityDate))

),

ActivityLevels = as.list(table(user\_summary$ActivityLevel)),

OverallAverages = as.list(overall\_stats),

UserSegments = as.list(setNames(segment\_summary$Count, segment\_summary$Segment))

)

print(summary\_report)

# Save detailed user summary

write.csv(user\_summary, "user\_activity\_summary.csv", row.names = FALSE)

cat("\nDetailed user summary saved to 'user\_activity\_summary.csv'\n")

# Print final insights

cat("\n=== KEY INSIGHTS ===\n")

cat("1. Average daily steps across all users:", round(mean(valid\_data$TotalSteps)), "\n")

cat("2. Percentage of users meeting 10,000 steps goal:",

round(mean(user\_summary$AvgSteps >= 10000) \* 100, 1), "%\n")

cat("3. Average sedentary time:", round(mean(valid\_data$SedentaryHours), 1), "hours per day\n")

cat("4. Strongest correlation: Steps vs Calories =",

round(cor(valid\_data$TotalSteps, valid\_data$Calories), 3), "\n")

cat("5. Data quality: ", round(nrow(valid\_data)/nrow(activity\_data)\*100, 1),

"% of records show device was likely worn\n")

# Additional visualization: Activity patterns for top 5 users

top5\_users <- user\_summary %>%

arrange(desc(AvgSteps)) %>%

head(5) %>%

pull(Id)

top\_users\_data <- valid\_data %>%

filter(Id %in% top5\_users) %>%

group\_by(Id, ActivityDate) %>%

summarise(DailySteps = mean(TotalSteps))

p7 <- ggplot(top\_users\_data, aes(x = ActivityDate, y = DailySteps, color = as.character(Id))) +

geom\_line(size = 0.8) +

labs(title = "Daily Steps Trend for Top 5 Most Active Users",

x = "Date", y = "Daily Steps", color = "User ID") +

scale\_y\_continuous(labels = comma) +

theme(legend.position = "bottom")

print(p7)

APPENDIX II

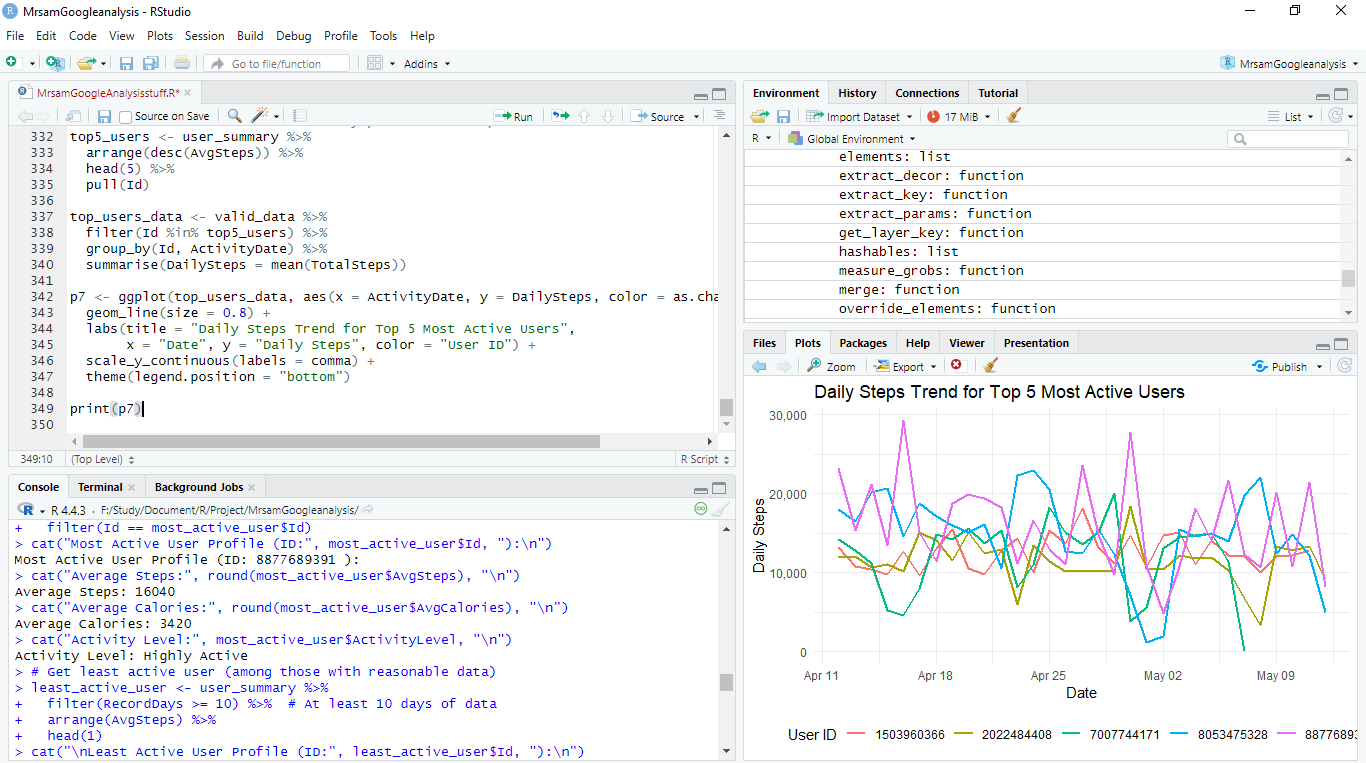


Figure 5 :Sample Coding Environment

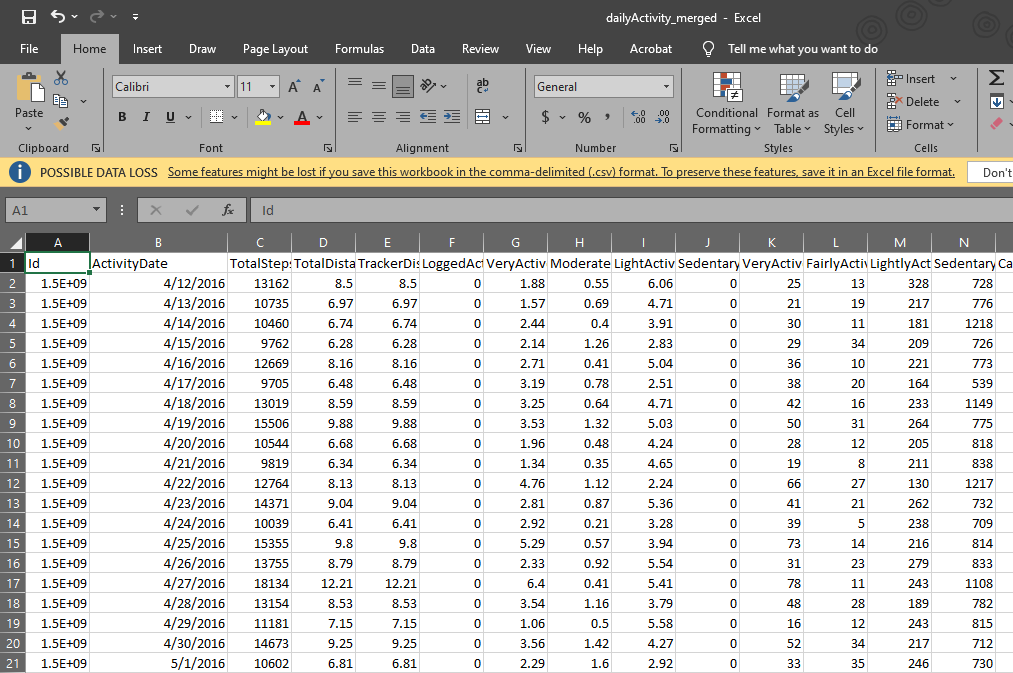


Figure 6: Sample dataset