

# Case Study Bellabeat\_Report

Aseem Mehrotra

03/09/2021

## Google Data Analytics Capstone Project

### Bellabeat Case Study: How Can a Wellness Technology Company Play it Smart

Bellabeat is a high-tech manufacturer of health-focused products for women. As per the co-founder & CCO of Bellabeat (Urska Srsen), analyzing device fitness data could help unlock new growth opportunities for the company. Contributing to the vision of CCO, this report focuses on the data collected for FitBit Fitness Tracker data to gain insight into how customers are using their smart devices.

In this document following steps of Data Analysis Process are followed:

- Ask
- Prepare
- Process
- Analyze
- Share
- Act

#### Phase 1: Ask

##### Key Objectives:

##### 1. The Business Task:

- a. **Analyse the Fitbit usage data** in order to gain insight into how consumers use non-Bellabeat smart devices
- b. Use these insights to **inform Bellabeat Marketing Strategy** for the product **"Time"** by Bellabeat.

##### 2. Elaborate Business Task:

Given the dataset, the business task is to search for user patterns of usage of their smart devices in order to gain insights that would better orientate marketing decisions. In short, in this report I will try to answer the question **"How do people use the smart devices?"**

##### 3. Key Stake Holders:

The main stakeholders are:

- Co-founder & Chief Creative Officer: Bellabeat

- Co-founder & Mathematician: Bellabeat
- Marketing Analytics Team: Bellabeat

## Phase 2: Prepare

### 1. Data Source:

- The data used in this case study is from the **Kaggle Notebook by the user Mobius**
- According to metadata, the original data comes from **Zenodo**

### 2. Shortcomings of the Given Data

- Data is old: Data was collected 5 years ago (in 2016)
- Data is small: The sample size of data is 33 FitBit users
- Data may be non-inclusive: The identities of the FitBit users are anonymous, so the data may not be inclusive of all ages, genders, economic backgrounds, geography and ethnicities.

### 3. Data Cleaning

- The data cleaned using R through RStudio
- Data cleaning code with outcome is as follows

## Loading Packages

```
library(tidyverse)
```

```
## — Attaching packages — tidyverse 1.3.1 —
—
```

```
## ✓ ggplot2 3.3.5      ✓ purrr 0.3.4
## ✓ tibble 3.1.4       ✓ dplyr 1.0.7
## ✓ tidyr 1.1.3        ✓ stringr 1.4.0
## ✓ readr 2.0.1        ✓ forcats 0.5.1
```

```
## — Conflicts — tidyverse_conflicts() —
—
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(readxl)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':  
##  
## smiths
```

```
library(scales)
```

```
##  
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':  
##  
## discard
```

```
## The following object is masked from 'package:readr':  
##  
## col_factor
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load sh  
ared object '/Library/Frameworks/R.framework/Resources/modules//R_X11.so':  
## dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Librar  
y not loaded: /opt/X11/lib/libSM.6.dylib  
## Referenced from: /Library/Frameworks/R.framework/Versions/4.1/Resources/module  
s/R_X11.so  
## Reason: image not found
```

```
## Could not load tcltk. Will use slower R code instead.
```

```
## Loading required package: RSQLite
```

```
library(janitor)
```

```
##  
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':  
##  
##   chisq.test, fisher.test
```

```
library(lubridate)
```

```
##  
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':  
##  
##   date, intersect, setdiff, union
```

## Reading Unique Data from the Dataset

```
# The code will read unique observations in each dataset. Thus removing the duplicates (if any)  
dailyAct <- unique(read.csv("dailyActivity_merged.csv"))  
dailyCal <- unique(read.csv("dailyCalories_merged.csv"))  
dailyInt <- unique(read.csv("dailyIntensities_merged.csv"))  
dailyStep <- unique(read.csv("dailySteps_merged.csv"))  
heartRate <- unique(read.csv("heartrate_seconds_merged.csv"))  
slpDay <- unique(read.csv("sleepDay_merged.csv"))  
wtInfo <- unique(read.csv("weightLogInfo_merged.csv"))
```

## Identifying NULL Values

```
# Null Values in each Column of Data Frame  
lapply(dailyAct, function(x) { length(which(is.na(x)))})
```

```
## $Id
## [1] 0
##
## $ActivityDate
## [1] 0
##
## $TotalSteps
## [1] 0
##
## $TotalDistance
## [1] 0
##
## $TrackerDistance
## [1] 0
##
## $LoggedActivitiesDistance
## [1] 0
##
## $VeryActiveDistance
## [1] 0
##
## $ModeratelyActiveDistance
## [1] 0
##
## $LightActiveDistance
## [1] 0
##
## $SedentaryActiveDistance
## [1] 0
##
## $VeryActiveMinutes
## [1] 0
##
## $FairlyActiveMinutes
## [1] 0
##
## $LightlyActiveMinutes
## [1] 0
##
## $SedentaryMinutes
## [1] 0
##
## $Calories
## [1] 0
```

```
lapply(dailyCal, function(x) { length(which(is.na(x)))})
```

```
## $Id
## [1] 0
##
## $ActivityDay
## [1] 0
##
## $Calories
## [1] 0
```

```
lapply(dailyInt, function(x) { length(which(is.na(x)))})
```

```
## $Id
## [1] 0
##
## $ActivityDay
## [1] 0
##
## $SedentaryMinutes
## [1] 0
##
## $LightlyActiveMinutes
## [1] 0
##
## $FairlyActiveMinutes
## [1] 0
##
## $VeryActiveMinutes
## [1] 0
##
## $SedentaryActiveDistance
## [1] 0
##
## $LightActiveDistance
## [1] 0
##
## $ModeratelyActiveDistance
## [1] 0
##
## $VeryActiveDistance
## [1] 0
```

```
lapply(dailyStep, function(x) { length(which(is.na(x)))})
```

```
## $Id
## [1] 0
##
## $ActivityDay
## [1] 0
##
## $StepTotal
## [1] 0
```

```
lapply(heartRate, function(x) { length(which(is.na(x)))})
```

```
## $Id
## [1] 0
##
## $Time
## [1] 0
##
## $Value
## [1] 0
```

```
lapply(slpDay, function(x) { length(which(is.na(x)))})
```

```
## $Id
## [1] 0
##
## $SleepDay
## [1] 0
##
## $TotalSleepRecords
## [1] 0
##
## $TotalMinutesAsleep
## [1] 0
##
## $TotalTimeInBed
## [1] 0
```

```
lapply(wtInfo, function(x) { length(which(is.na(x)))})
```

```
## $Id
## [1] 0
##
## $Date
## [1] 0
##
## $WeightKg
## [1] 0
##
## $WeightPounds
## [1] 0
##
## $Fat
## [1] 65
##
## $BMI
## [1] 0
##
## $IsManualReport
## [1] 0
##
## $LogId
## [1] 0
```

*# Weight Log Info database wtInfo has 65 missing or NULL values in "Fat" column.*

**The above code ensures that no duplicates exist in the imported data-sets and null values are also identified in individual data-sets.**

## Summary of Daily Data-sets

```
colnames(dailyAct)
```

```
## [1] "Id" "ActivityDate"
## [3] "TotalSteps" "TotalDistance"
## [5] "TrackerDistance" "LoggedActivitiesDistance"
## [7] "VeryActiveDistance" "ModeratelyActiveDistance"
## [9] "LightActiveDistance" "SedentaryActiveDistance"
## [11] "VeryActiveMinutes" "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes" "SedentaryMinutes"
## [15] "Calories"
```

```
head(dailyAct)
```



```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366 4/12/2016      13162          8.50          8.50
## 2 1503960366 4/13/2016      10735          6.97          6.97
## 3 1503960366 4/14/2016      10460          6.74          6.74
## 4 1503960366 4/15/2016       9762          6.28          6.28
## 5 1503960366 4/16/2016      12669          8.16          8.16
## 6 1503960366 4/17/2016       9705          6.48          6.48
## LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1              0              1.88              0.55
## 2              0              1.57              0.69
## 3              0              2.44              0.40
## 4              0              2.14              1.26
## 5              0              2.71              0.41
## 6              0              3.19              0.78
## LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1              6.06              0              25
## 2              4.71              0              21
## 3              3.91              0              30
## 4              2.83              0              29
## 5              5.04              0              36
## 6              2.51              0              38
## FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1              13              328              728      1985
## 2              19              217              776      1797
## 3              11              181             1218      1776
## 4              34              209              726      1745
## 5              10              221              773      1863
## 6              20              164              539      1728
```

```
colnames(dailyCal)
```

```
## [1] "Id"           "ActivityDay" "Calories"
```

```
head(dailyCal)
```

```
##           Id ActivityDay Calories
## 1 1503960366 4/12/2016      1985
## 2 1503960366 4/13/2016      1797
## 3 1503960366 4/14/2016      1776
## 4 1503960366 4/15/2016      1745
## 5 1503960366 4/16/2016      1863
## 6 1503960366 4/17/2016      1728
```

```
colnames(dailyInt)
```

```
## [1] "Id" "ActivityDay"
## [3] "SedentaryMinutes" "LightlyActiveMinutes"
## [5] "FairlyActiveMinutes" "VeryActiveMinutes"
## [7] "SedentaryActiveDistance" "LightActiveDistance"
## [9] "ModeratelyActiveDistance" "VeryActiveDistance"
```

```
head(dailyInt)
```

```
##           Id ActivityDay SedentaryMinutes LightlyActiveMinutes
## 1 1503960366 4/12/2016           728           328
## 2 1503960366 4/13/2016           776           217
## 3 1503960366 4/14/2016          1218           181
## 4 1503960366 4/15/2016           726           209
## 5 1503960366 4/16/2016           773           221
## 6 1503960366 4/17/2016           539           164
## FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
## 1              13              25              0
## 2              19              21              0
## 3              11              30              0
## 4              34              29              0
## 5              10              36              0
## 6              20              38              0
## LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
## 1              6.06              0.55              1.88
## 2              4.71              0.69              1.57
## 3              3.91              0.40              2.44
## 4              2.83              1.26              2.14
## 5              5.04              0.41              2.71
## 6              2.51              0.78              3.19
```

```
colnames(dailyStep)
```

```
## [1] "Id" "ActivityDay" "StepTotal"
```

```
head(dailyStep)
```

```
##           Id ActivityDay StepTotal
## 1 1503960366 4/12/2016     13162
## 2 1503960366 4/13/2016     10735
## 3 1503960366 4/14/2016     10460
## 4 1503960366 4/15/2016      9762
## 5 1503960366 4/16/2016     12669
## 6 1503960366 4/17/2016      9705
```

**Column names of Daily Calories and Daily Intensities data-sets already exist in Daily Activity Data-**

set

## Summary of other Data-sets (used in this study)

```
colnames(heartRate)
```

```
## [1] "Id"      "Time"    "Value"
```

```
head(heartRate)
```

```
##           Id           Time Value
## 1 2022484408 4/12/2016 7:21:00 AM    97
## 2 2022484408 4/12/2016 7:21:05 AM   102
## 3 2022484408 4/12/2016 7:21:10 AM   105
## 4 2022484408 4/12/2016 7:21:20 AM   103
## 5 2022484408 4/12/2016 7:21:25 AM   101
## 6 2022484408 4/12/2016 7:22:05 AM    95
```

```
colnames(wtInfo)
```

```
## [1] "Id"           "Date"         "WeightKg"      "WeightPounds"
## [5] "Fat"          "BMI"          "IsManualReport" "LogId"
```

```
head(wtInfo)
```

```
##           Id           Date WeightKg WeightPounds Fat   BMI
## 1 1503960366 5/2/2016 11:59:59 PM    52.6    115.9631  22 22.65
## 2 1503960366 5/3/2016 11:59:59 PM    52.6    115.9631  NA 22.65
## 3 1927972279 4/13/2016 1:08:52 AM   133.5    294.3171  NA 47.54
## 4 2873212765 4/21/2016 11:59:59 PM    56.7    125.0021  NA 21.45
## 5 2873212765 5/12/2016 11:59:59 PM    57.3    126.3249  NA 21.69
## 6 4319703577 4/17/2016 11:59:59 PM    72.4    159.6147  25 27.45
##  IsManualReport      LogId
## 1             True 1.462234e+12
## 2             True 1.462320e+12
## 3            False 1.460510e+12
## 4             True 1.461283e+12
## 5             True 1.463098e+12
## 6             True 1.460938e+12
```

## Phase 3: Process

### 1. Compare different data-sets

- Different data-sets are compared using `sqldf()` & `nrow()` functions to identify the number of rows in different data sets

```
# Compare Daily Activity and Daily Calories data-sets for number of rows
dailyActCal <- dailyAct %>%
  select(Id, ActivityDate, Calories)
check1a <- nrow(sqldf("SELECT * FROM dailyActCal INTERSECT SELECT * FROM dailyCal")
)
check1b <- nrow(dailyAct)
check1c <- nrow(dailyCal)
check1a
```

```
## [1] 940
```

```
check1b
```

```
## [1] 940
```

```
check1c
```

```
## [1] 940
```

**Daily Activity, Daily Calorie & new data-set of INTERSECTION of these 2 contain the similar number of rows: 940.**

```
dailyActInt <- dailyAct %>%
  select(Id, ActivityDate, SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, VeryActiveMinutes, SedentaryActiveDistance, LightActiveDistance, ModeratelyActiveDistance, VeryActiveDistance)
check2a <- nrow(sqldf("SELECT * FROM dailyActInt INTERSECT SELECT * FROM dailyInt")
)
check2b <- nrow(dailyAct)
check2c <- nrow(dailyInt)
check2a
```

```
## [1] 940
```

```
check2b
```

```
## [1] 940
```

```
check2c
```

```
## [1] 940
```

**Daily Activity, Daily Intensities & new data-set of INTERSECTION of these 2 contain the similar number of rows: 940.**

**Daily Activity data-set can be used out of these 3 data-sets for analysis as Daily Activity data-set already has the observations which are present in Daily Calories & Daily Intensities Data-sets**

## 2. Join Data-Sets

### Left Join Daily Activity Data-Set with Heart Rate, Sleep & weight Data-Sets

```
dailyHR <- sqldf("SELECT Id, Value AS HeartRate, TRIM(SUBSTR(Time,1,9)) AS Date FROM heartRate")
dailyActHR <- sqldf("SELECT A.*, B.HeartRate FROM dailyAct AS A LEFT JOIN dailyHR AS B ON A.Id = B.Id AND A.ActivityDate = B.Date")

dailySlp <- sqldf("SELECT Id, TRIM(SUBSTR(SleepDay,1,9)) AS Date, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed FROM slpDay")
dailyActHRSlp <- sqldf("SELECT A.*, B.TotalSleepRecords, B.TotalMinutesAsleep, B.TotalTimeInBed FROM dailyActHR AS A LEFT JOIN dailySlp AS B ON A.Id = B.Id AND A.ActivityDate = B.Date")

wtLog <- sqldf("SELECT Id, TRIM(SUBSTR(Date,1,9)) AS Date, WeightKg, BMI FROM wtInfo")
dailyActHRSlpWt <- sqldf("SELECT A.*, B.WeightKg, B.BMI FROM dailyActHRSlp AS A LEFT JOIN wtLog AS B ON A.Id = B.Id AND A.ActivityDate = B.date")
```

**The data-set named “dailyActHRSlpWt” will be used for further analysis in this report.**

## 3. Clean Data Names in the Data-set

```
dailyActHRSlpWt <- dailyActHRSlpWt %>%
  clean_names()
```

## 4. Data Types Change

```
# Change Activity Date Column to Date Data Type
dailyActHRSlpWt <- dailyActHRSlpWt %>%
  mutate(activity_date = mdy(activity_date))

# Change ID column from Numeric to Character for ease in the analysis
dailyActHRSlpWt <- dailyActHRSlpWt %>%
  mutate(id = as.character(id))
```

## 5. Summary of the Data

```
names(dailyActHRSlpWt)
```

```
## [1] "id" "activity_date"
## [3] "total_steps" "total_distance"
## [5] "tracker_distance" "logged_activities_distance"
## [7] "very_active_distance" "moderately_active_distance"
## [9] "light_active_distance" "sedentary_active_distance"
## [11] "very_active_minutes" "fairly_active_minutes"
## [13] "lightly_active_minutes" "sedentary_minutes"
## [15] "calories" "heart_rate"
## [17] "total_sleep_records" "total_minutes_asleep"
## [19] "total_time_in_bed" "weight_kg"
## [21] "bmi"
```

```
head(dailyActHRSslpWt)
```

```
##          id activity_date total_steps total_distance tracker_distance
## 1 1503960366 2016-04-12      13162          8.50          8.50
## 2 1503960366 2016-04-13      10735          6.97          6.97
## 3 1503960366 2016-04-14      10460          6.74          6.74
## 4 1503960366 2016-04-15       9762          6.28          6.28
## 5 1503960366 2016-04-16      12669          8.16          8.16
## 6 1503960366 2016-04-17       9705          6.48          6.48
## logged_activities_distance very_active_distance moderately_active_distance
## 1              0              1.88              0.55
## 2              0              1.57              0.69
## 3              0              2.44              0.40
## 4              0              2.14              1.26
## 5              0              2.71              0.41
## 6              0              3.19              0.78
## light_active_distance sedentary_active_distance very_active_minutes
## 1              6.06              0              25
## 2              4.71              0              21
## 3              3.91              0              30
## 4              2.83              0              29
## 5              5.04              0              36
## 6              2.51              0              38
## fairly_active_minutes lightly_active_minutes sedentary_minutes calories
## 1              13              328              728      1985
## 2              19              217              776      1797
## 3              11              181             1218      1776
## 4              34              209              726      1745
## 5              10              221              773      1863
## 6              20              164              539      1728
## heart_rate total_sleep_records total_minutes_asleep total_time_in_bed
## 1          NA              1              327              346
## 2          NA              2              384              407
## 3          NA             NA              NA              NA
## 4          NA              1              412              442
## 5          NA              2              340              367
## 6          NA              1              700              712
## weight_kg bmi
## 1          NA  NA
## 2          NA  NA
## 3          NA  NA
## 4          NA  NA
## 5          NA  NA
## 6          NA  NA
```

```
summarise(dailyActHRSlpWt)
```

```
## data frame with 0 columns and 1 row
```

```
glimpse(dailyActHRSlpWt)
```

```
## Rows: 2,484,264
## Columns: 21
## $ id          <chr> "1503960366", "1503960366", "1503960366", "...
## $ activity_date <date> 2016-04-12, 2016-04-13, 2016-04-14, 2016-0...
## $ total_steps   <int> 13162, 10735, 10460, 9762, 12669, 9705, 130...
## $ total_distance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9...
## $ tracker_distance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9...
## $ logged_activities_distance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ very_active_distance <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3...
## $ moderately_active_distance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1...
## $ light_active_distance <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5...
## $ sedentary_active_distance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ very_active_minutes <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66,...
## $ fairly_active_minutes <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, ...
## $ lightly_active_minutes <int> 328, 217, 181, 209, 221, 164, 233, 264, 205...
## $ sedentary_minutes <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 8...
## $ calories <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2...
## $ heart_rate <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ total_sleep_records <int> 1, 2, NA, 1, 2, 1, NA, 1, 1, 1, NA, 1, 1, 1...
## $ total_minutes_asleep <int> 327, 384, NA, 412, 340, 700, NA, 304, 360, ...
## $ total_time_in_bed <int> 346, 407, NA, 442, 367, 712, NA, 320, 377, ...
## $ weight_kg <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ bmi <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

```
summary(dailyActHRSlpWt)
```



```

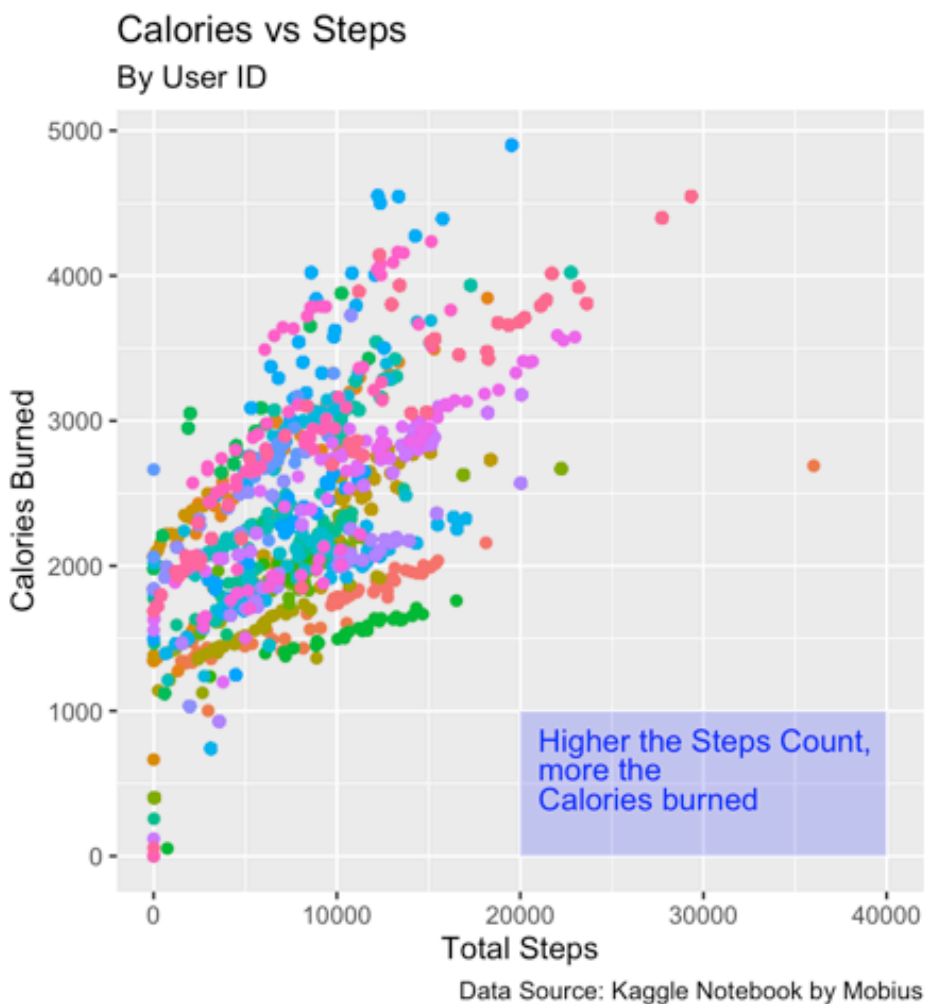
##      id      activity_date      total_steps      total_distance
## Length:2484264      Min.   :2016-04-12      Min.   :    0      Min.   : 0.000
## Class :character      1st Qu.:2016-04-19      1st Qu.: 6083      1st Qu.: 4.200
## Mode  :character      Median :2016-04-26      Median :10140      Median : 7.100
##                                     Mean   :2016-04-26      Mean   : 9891      Mean   : 7.177
##                                     3rd Qu.:2016-05-04      3rd Qu.:12346      3rd Qu.: 8.740
##                                     Max.   :2016-05-12      Max.   :36019      Max.   :28.030
##
## tracker_distance logged_activities_distance very_active_distance
## Min.   : 0.000      Min.   :0.0000      Min.   : 0.00
## 1st Qu.: 4.200      1st Qu.:0.0000      1st Qu.: 0.00
## Median : 7.100      Median :0.0000      Median : 0.98
## Mean   : 7.142      Mean   :0.1674      Mean   : 2.05
## 3rd Qu.: 8.680      3rd Qu.:0.0000      3rd Qu.: 3.00
## Max.   :28.030      Max.   :4.9421      Max.   :21.92
##
## moderately_active_distance light_active_distance sedentary_active_distance
## Min.   :0.0000      Min.   : 0.000      Min.   :0.000000
## 1st Qu.:0.0000      1st Qu.: 3.150      1st Qu.:0.000000
## Median :0.4400      Median : 4.330      Median :0.000000
## Mean   :0.6759      Mean   : 4.395      Mean   :0.003235
## 3rd Qu.:1.0400      3rd Qu.: 5.580      3rd Qu.:0.000000
## Max.   :6.4800      Max.   :10.710      Max.   :0.110000
##
## very_active_minutes fairly_active_minutes lightly_active_minutes
## Min.   : 0.00      Min.   : 0.00      Min.   : 0.0
## 1st Qu.: 0.00      1st Qu.: 0.00      1st Qu.:195.0
## Median : 17.00      Median : 14.00      Median :238.0
## Mean   : 32.79      Mean   : 17.61      Mean   :243.6
## 3rd Qu.: 50.00      3rd Qu.: 29.00      3rd Qu.:294.0
## Max.   :210.00      Max.   :143.00      Max.   :518.0
##
## sedentary_minutes      calories      heart_rate      total_sleep_records
## Min.   : 0.0      Min.   : 0      Min.   : 36.00      Min.   :1.0
## 1st Qu.: 662.0      1st Qu.:2086      1st Qu.: 63.00      1st Qu.:1.0
## Median : 775.0      Median :2629      Median : 73.00      Median :1.0
## Mean   : 841.1      Mean   :2686      Mean   : 77.33      Mean   :1.1
## 3rd Qu.:1065.0      3rd Qu.:3089      3rd Qu.: 88.00      3rd Qu.:1.0
## Max.   :1440.0      Max.   :4900      Max.   :203.00      Max.   :3.0
##                                     NA's      :606      NA's      :854695
## total_minutes_asleep total_time_in_bed      weight_kg      bmi
## Min.   : 58.0      Min.   : 61.0      Min.   : 52.6      Min.   :21.5
## 1st Qu.:385.0      1st Qu.:408.0      1st Qu.: 61.4      1st Qu.:24.0
## Median :436.0      Median :461.0      Median : 62.1      Median :24.2
## Mean   :425.9      Mean   :455.3      Mean   : 71.5      Mean   :24.9
## 3rd Qu.:478.0      3rd Qu.:526.0      3rd Qu.: 85.1      3rd Qu.:25.6
## Max.   :796.0      Max.   :961.0      Max.   :133.5      Max.   :47.5
## NA's      :854695      NA's      :854695      NA's      :2007285      NA's      :2007285

```

# Phase 4: Analysis

## 1. Calories Burned vs Total Steps Grouped by User Id

```
# Calories Burned vs Total Steps by ID: Scatter Plot
dailyActHRS1pWt %>%
  ggplot(mapping = aes(x=total_steps, y=calories))+
  geom_point(mapping = aes(color = id))+
  labs(x="Total Steps", y= "Calories Burned", title = "Calories vs Steps", subtitle =
= "By User ID", caption = "Data Source: Kaggle Notebook by Mobius", color = "ID")+
  annotate("rect", xmin = 20000, xmax = 40000, ymin = 0, ymax = 1000, alpha = 0.2,
fill = "blue")+
  annotate("text", x= 21000, y= 800, label = "Higher the Steps Count, ", color = "
blue", hjust = "left")+
  annotate("text", x= 21000, y= 600, label = "more the ", color = "blue", hjust = "
left")+
  annotate("text", x= 21000, y= 400, label = "Calories burned", color = "blue", hju
st = "left")+
  theme(legend.box.background = element_rect(color = "Blue", size = 2), legend.text
= element_text(face = "bold"))
```



ID	
1503960366	4558609924
1624580081	4702921684
1644430081	5553957443
1844505072	5577150313
1927972279	6117666160
2022484408	6290855005
2026352035	6775888955
2320127002	6962181067
2347167796	7007744171
2873212765	7086361926
3372868164	8053475328
3977333714	8253242879
4020332650	8378563200
4057192912	8583815059
4319703577	8792009665
4388161847	8877689391
4445114986	

```
# Calories vs Steps by User ID: Box Plot
```

```
dailyActHRS1pWt %>%
```

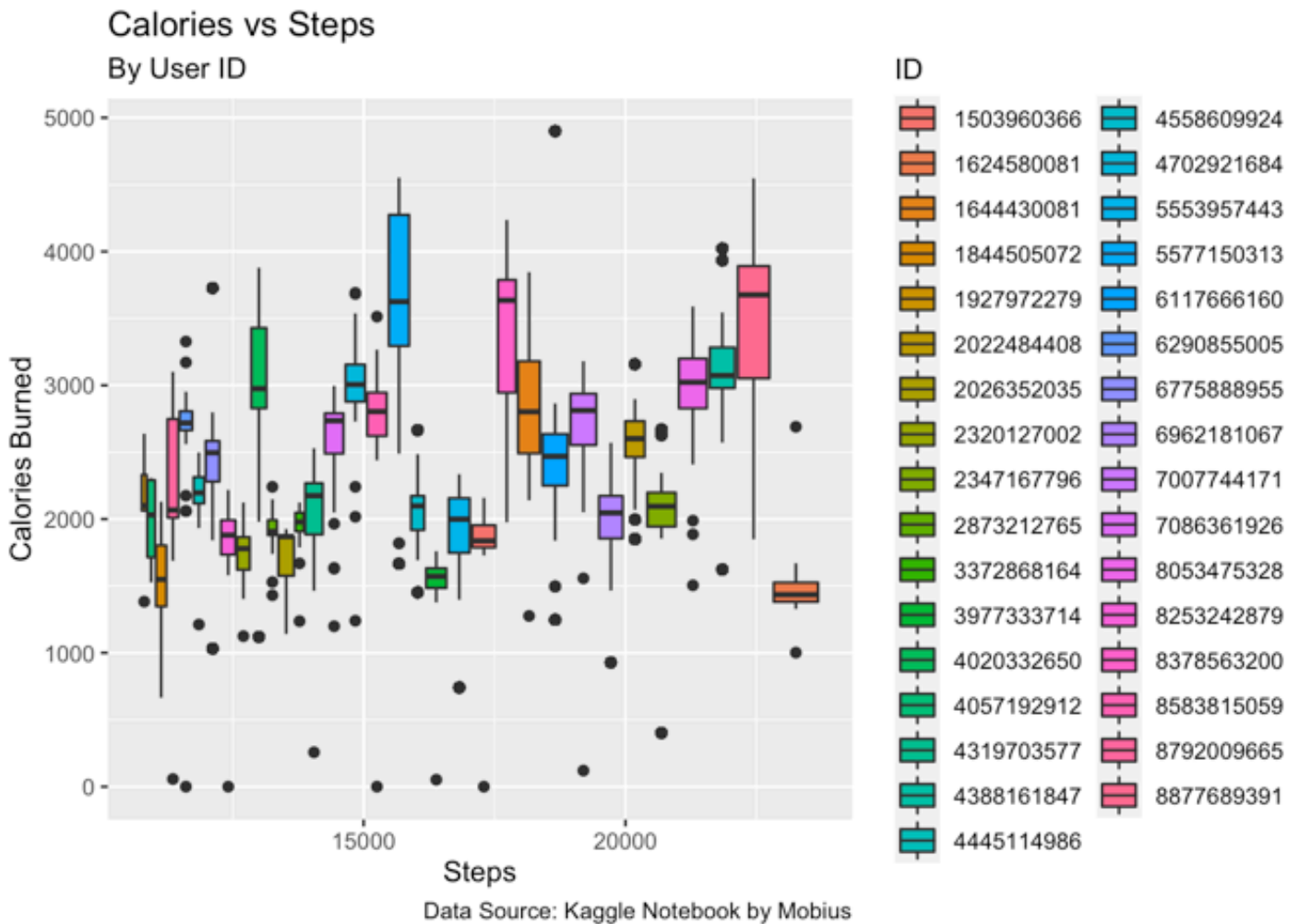
```
  select(id, total_steps, calories) %>%
```

```
  group_by(id) %>%
```

```
  ggplot(mapping = aes(x = total_steps, y= calories))+
```

```
  geom_boxplot(aes(fill=id))+
```

```
  labs(x= "Steps", y= "Calories Burned",title = "Calories vs Steps", subtitle = "By  
User ID", caption = "Data Source: Kaggle Notebook by Mobius", fill = "ID")
```



Analyzing the above plot it is evident that **"Higher the Steps count, more the Calories Burned"**

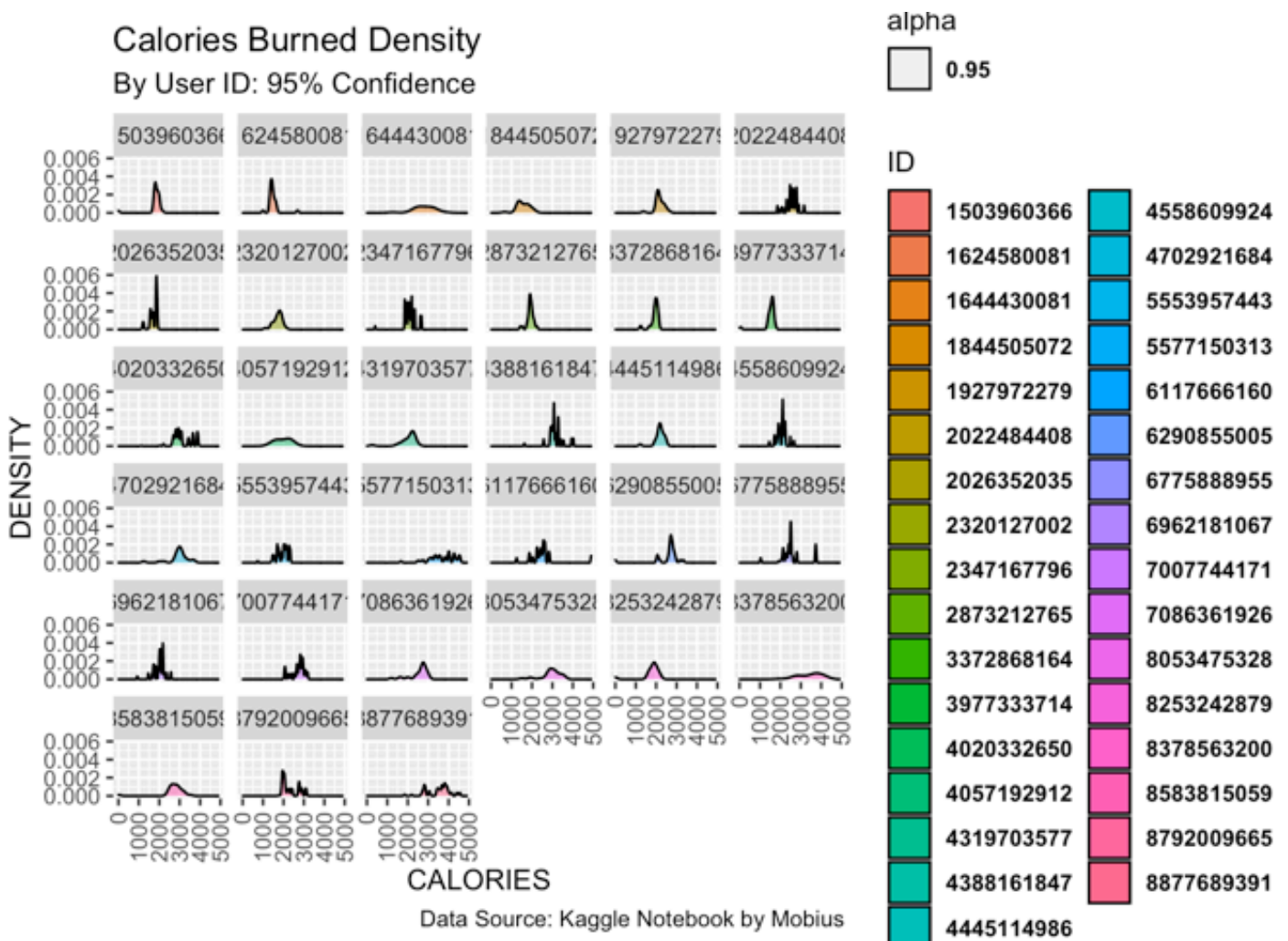
## 2. Calories Burned by Users

## # Calories Burned by User ID: Density Plot

```

dailyActHRSlpWt %>%
  select(id, total_steps, calories) %>%
  group_by(id) %>%
  ggplot(aes(x= calories))+
  geom_density(aes(fill=factor(id), alpha = 0.95))+
  labs(x="CALORIES", y= "DENSITY", title = "Calories Burned Density", subtitle = "By
User ID: 95% Confidence", caption = "Data Source: Kaggle Notebook by Mobius", sci
entific = FALSE, fill = "ID")+
  theme(legend.box.background = element_rect(color = "Blue", size = 0), legend.text
= element_text(face = "bold"), axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust =1))+
  facet_wrap(~id)

```

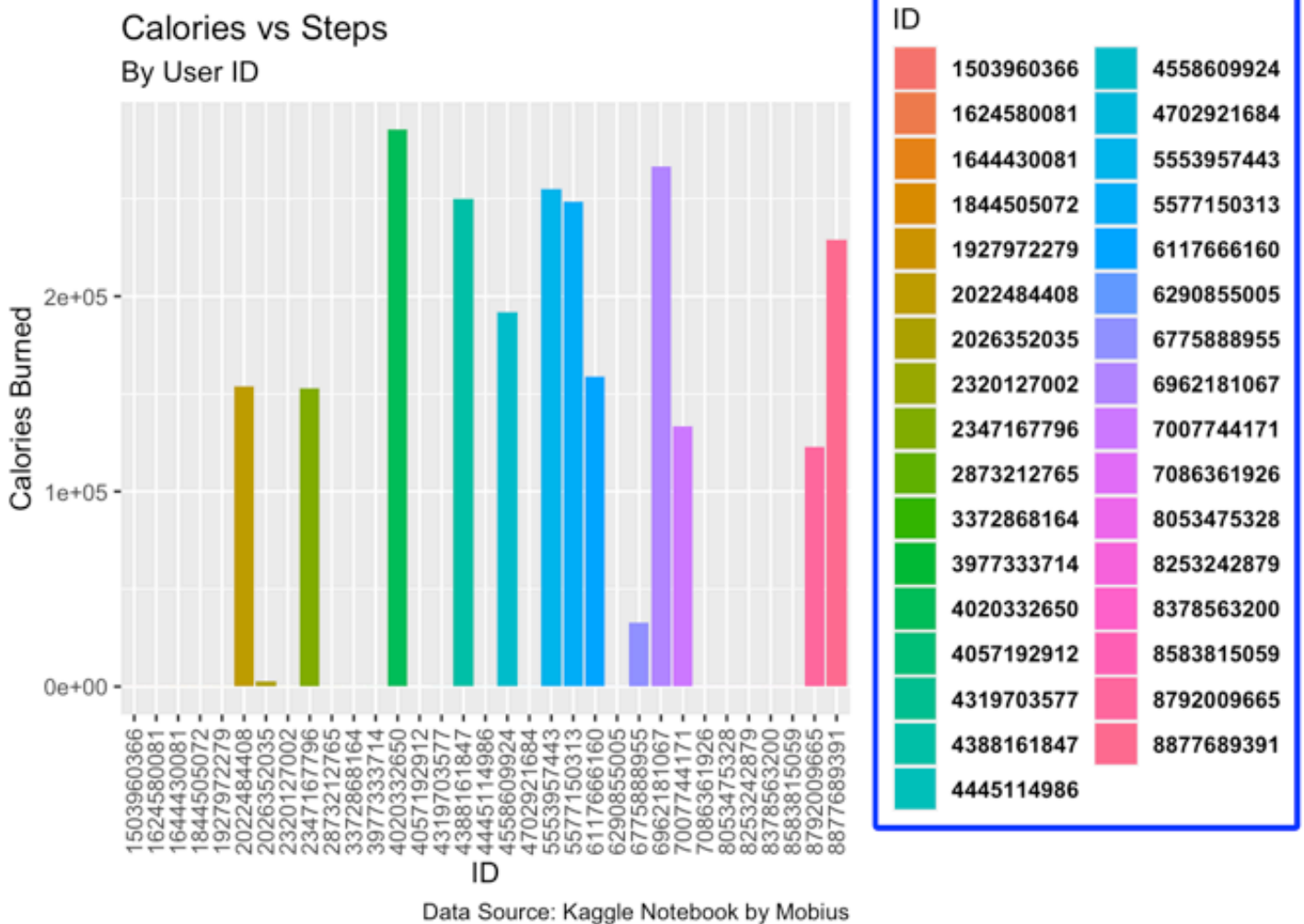


**# Calories Burned by User ID: Bar Graph**

```

dailyActHRS1pWt %>%
  group_by(id) %>%
  select(id, calories) %>%
  ggplot(mapping = aes(id))+
  geom_bar(mapping = aes(fill = id))+
  labs(x="ID", y= "Calories Burned", title = "Calories vs Steps", subtitle = "By User ID", caption = "Data Source: Kaggle Notebook by Mobius", scientific = FALSE, fill = "ID")+
  theme(legend.box.background = element_rect(color = "Blue", size = 2), legend.text = element_text(face = "bold"), axis.text.x = element_text(angle = 90, vjust = 0.5, hjust =1))

```



**Almost all users use the FitBit device for Calories Burned tracking**

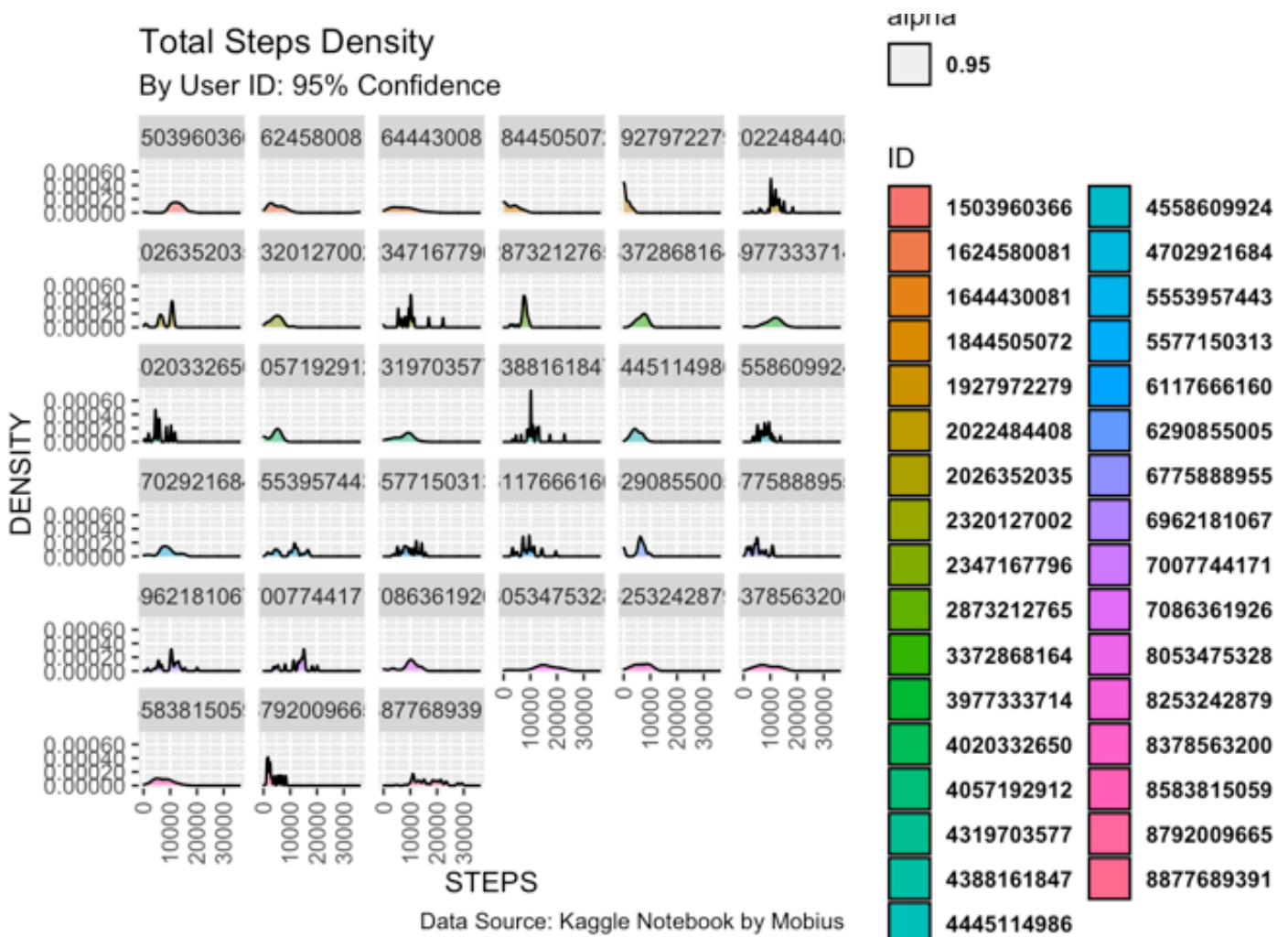
### 3. Steps Taken by Users

## # Total Steps by User ID: Density Plot

```

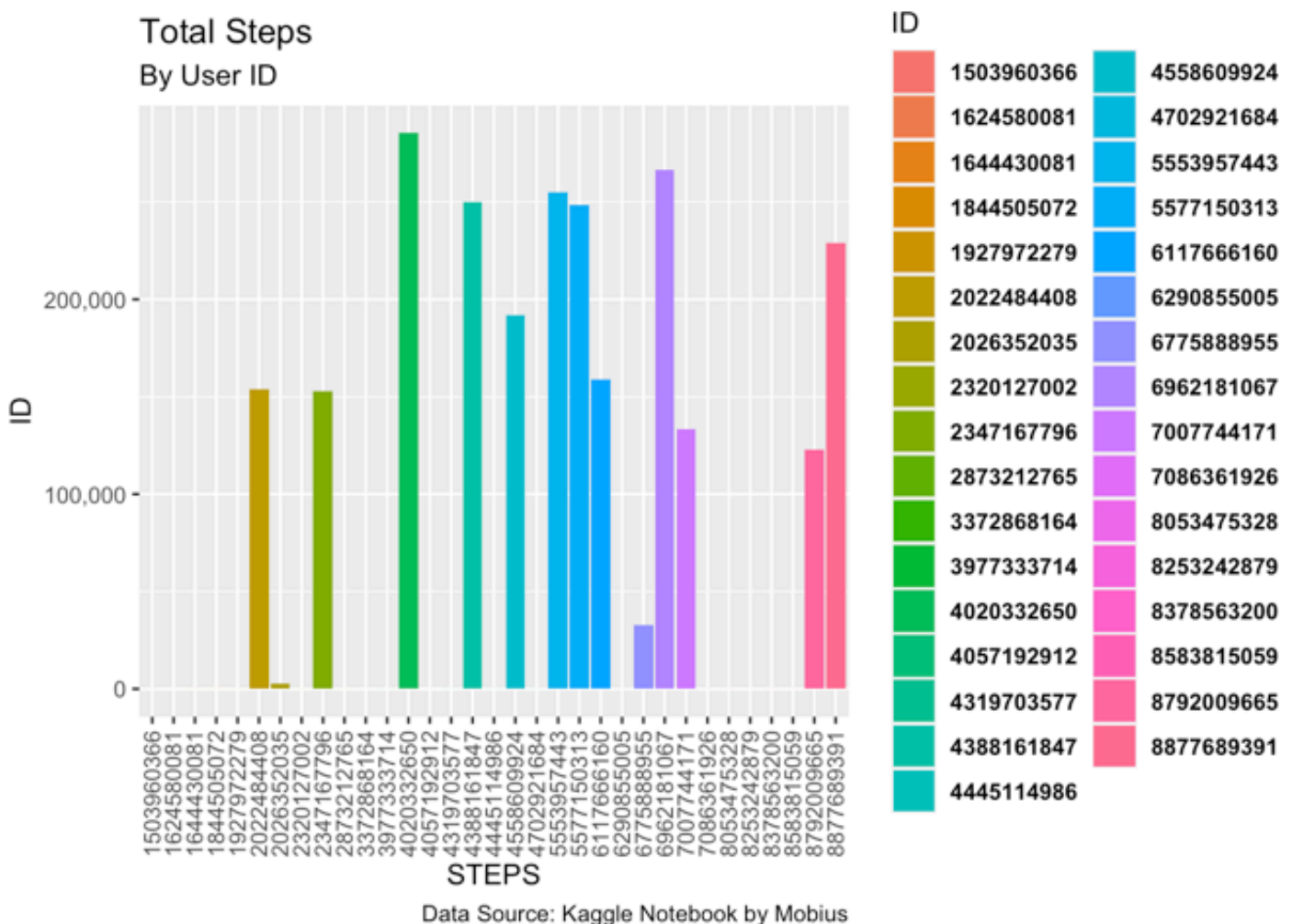
dailyActHRSlpWt %>%
  select(id, total_steps, calories) %>%
  ggplot(aes(x= total_steps))+
  scale_y_continuous(labels= scales::comma)+
  geom_density(aes(fill=factor(id), alpha = 0.95))+
  labs(x="STEPS", y= "DENSITY", title = "Total Steps Density", subtitle = "By User ID: 95% Confidence", caption = "Data Source: Kaggle Notebook by Mobius", scientific = FALSE, fill = "ID")+
  theme(legend.box.background = element_rect(color = "Blue", size = 0), legend.text = element_text(face = "bold"), axis.text.x = element_text(angle = 90, vjust = 0.5, hjust =1))+
  facet_wrap(~id)

```





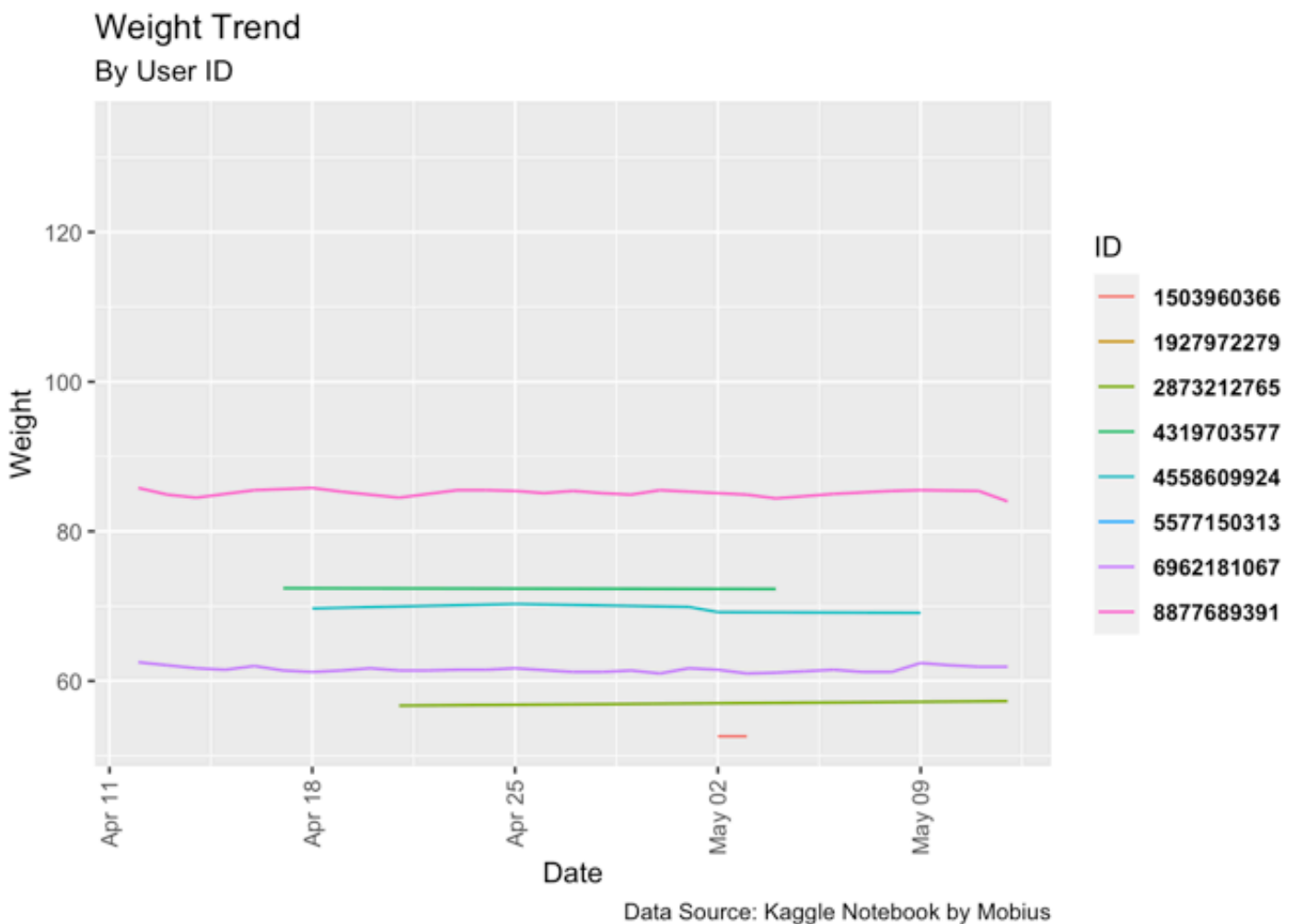
```
# Total Steps by User ID: Bar Graph
# Total Steps by User ID: Bar Graph
dailyActHRSlpWt %>%
  select(id, total_steps) %>%
  ggplot(aes(x= id))+
  scale_y_continuous(labels= scales::comma)+
  geom_bar(aes(fill=factor(id)))+
  labs(x="STEPS", y= "ID", title = "Total Steps", subtitle = "By User ID", caption =
= "Data Source: Kaggle Notebook by Mobius", scientific = FALSE, fill = "ID")+
  theme(legend.box.background = element_rect(color = "Blue", size = 0), legend.text
= element_text(face = "bold"), axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust =1))
```



Steps tracking is utilized by all the users, although, some users are very active & some are sedentary.

#### 4. Weight Logs

```
# Weight Trend By User ID
dailyActHRSlpWt %>%
  select(id, activity_date, weight_kg) %>%
  drop_na() %>%
  ggplot(mapping = aes(activity_date, weight_kg))+
  geom_line(aes(color = id))+
  labs(x="Date", y= "Weight", title = "Weight Trend", subtitle = "By User ID", caption = "Data Source: Kaggle Notebook by Mobius", color = "ID")+
  theme(legend.box.background = element_rect(color = "Blue", size = 0), legend.text = element_text(face = "bold"), axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



**Weight Log** is not famous among the users and only 8 out of 33 users logged the weight info in the samrt device.

## 5. Sleep Trends



```

dailyActHRSlpWt %>%
  select(id, activity_date, total_minutes_asleep) %>%
  drop_na() %>%
  ggplot(mapping = aes(activity_date, total_minutes_asleep)) +
  geom_point(aes(color = id)) +
  labs(x = "Date", y = "Sleep", title = "Sleep Trend", subtitle = "By User ID", caption = "Data Source: Kaggle Notebook by Mobius", color = "ID") +
  theme(legend.box.background = element_rect(color = "Blue", size = 0), legend.text = element_text(face = "bold"), axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  facet_wrap(~id)

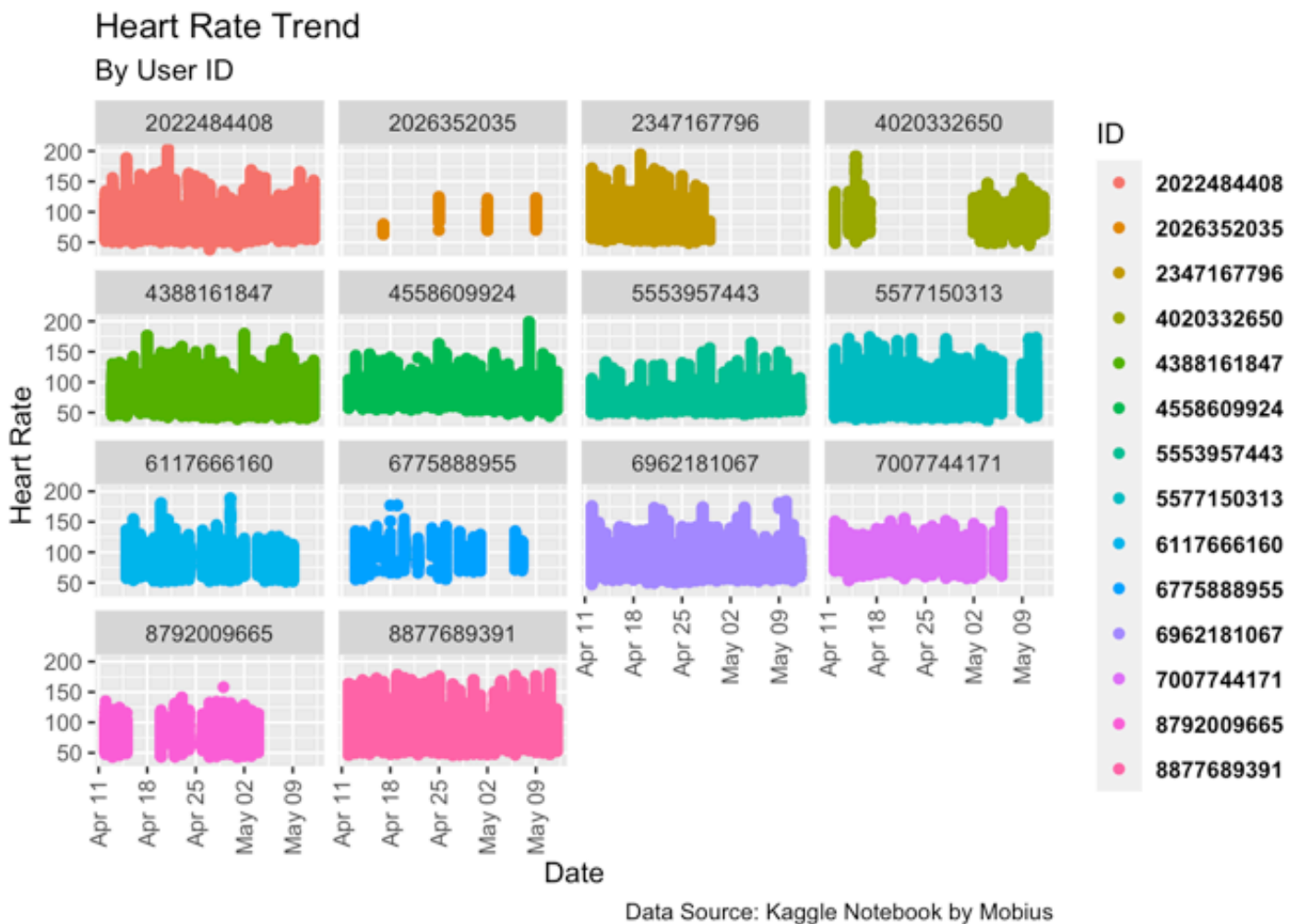
```



Sleep Tracking is famous feature of smart device among the users. 24 users out of 33 used Fitbit smart device for Sleep tracking.

## 6. Heart Rate Trends

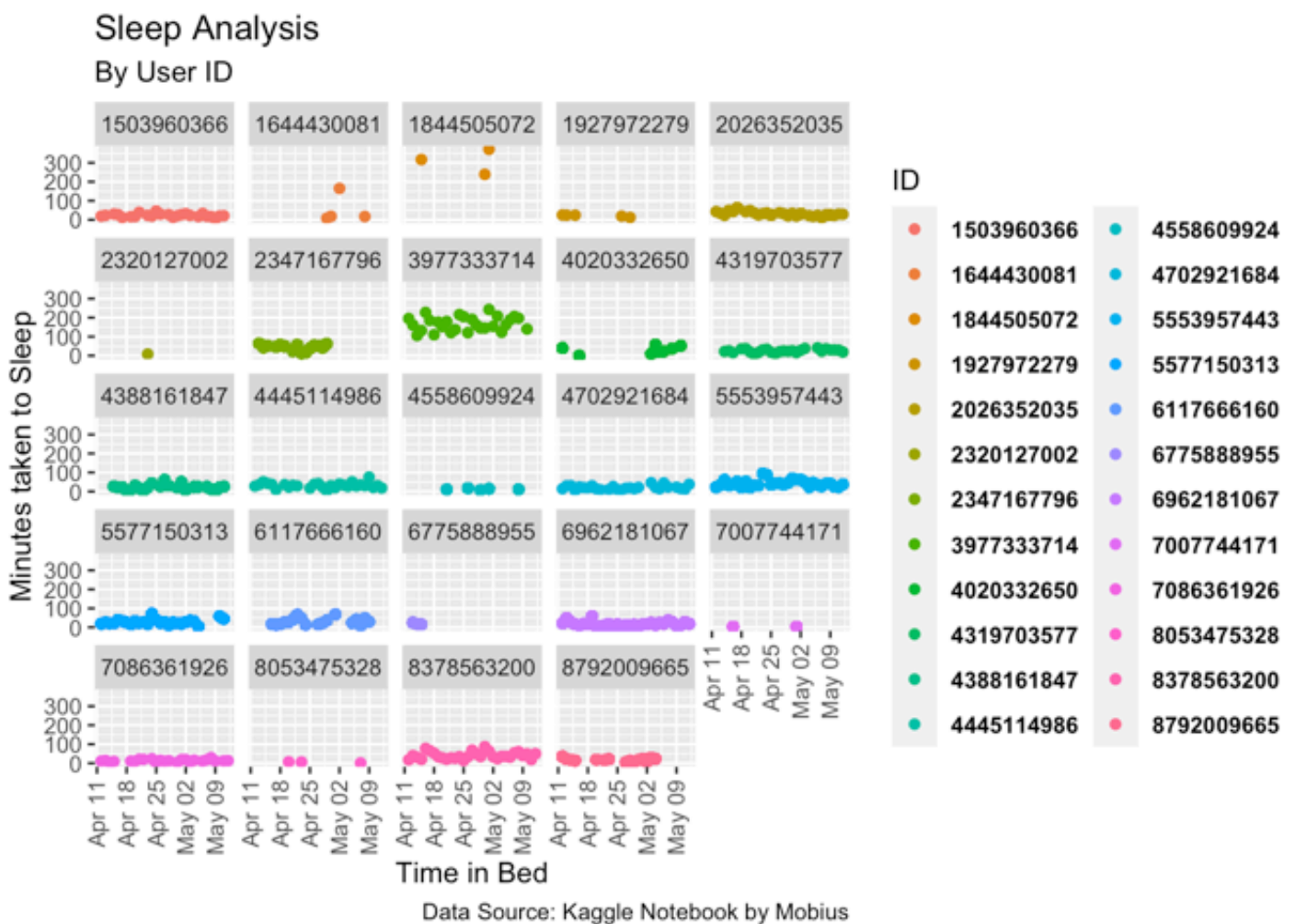
```
# Heart Rate Trend By User ID
dailyActHRSlpWt %>%
  select(id, activity_date, heart_rate) %>%
  drop_na() %>%
  ggplot(mapping = aes(activity_date, heart_rate))+
  geom_point(aes(color = id))+
  labs(x="Date", y= "Heart Rate", title = "Heart Rate Trend", subtitle = "By User I
D", caption = "Data Source: Kaggle Notebook by Mobius", color = "ID")+
  theme(legend.box.background = element_rect(color = "Blue", size = 0), legend.text
= element_text(face = "bold"), axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust =1))+
  facet_wrap(~id)
```



Almost 42% FitBit users are utilizing the Heart Rate track feature as per the data available.

## 7. Time to Fall Asleep

```
# Time to get sleep
dailyActHRSslpWt %>%
  select(id, activity_date, total_minutes_asleep, total_time_in_bed) %>%
  drop_na() %>%
  ggplot(mapping = aes(activity_date, total_time_in_bed-total_minutes_asleep)) +
  geom_point(aes(color = id)) +
  labs(x="Time in Bed", y= "Minutes taken to Sleep", title = "Sleep Analysis", subtitle = "By User ID", caption = "Data Source: Kaggle Notebook by Mobius", color = "ID") +
  theme(legend.box.background = element_rect(color = "Blue", size = 0), legend.text = element_text(face = "bold"), axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  facet_wrap(~id)
```



Time taken to get asleep is higher (nearly 45 minutes). This might be due to stress.

**Use of Fitbit for Calories Burned, Steps Taken, Sleep & Heart Rate tracking is Fairly high as compared to Weight Tracking**

**Phase 5: Share**

This report is available in HTML format on GitHub.

## Phase 6: Act

# Recommendations based on analysis:

1. Marketing can use the connection between **Steps Taken & Calorie Burned** to advertise **Steps Taken** as key for being fit by burning **More Calories**.
2. **Mindfulness Tracking** can be advertised to reduce the stress and can be linked to better sleep.
3. Features like “**Menstrual Tracking**” will enable women to have value addition by using the product from **Bellabeat**