Bellabeat Case Study

Jayoda

2024-12-14

Data Cleaning and Transformation

```
# Ensure libraries are loaded
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all confl
icts to become errors
```

```
# Load CSV files
daily_activity <- read.csv("C:/Users/Jayoda Kulatunga/Desktop/Data Analytics certif
icates/Case Study Files/CaseStudy02/mturkfitbit_export_3.12.16-4.11.16/Fitabase Dat
a 3.12.16-4.11.16/dailyActivity_merged.csv")
sleep_day <- read.csv("C:/Users/Jayoda Kulatunga/Desktop/Data Analytics certificate
s/Case Study Files/CaseStudy02/mturkfitbit_export_4.12.16-5.12.16/Fitabase Data 4.1
2.16-5.12.16/sleepDay_merged.csv")
```

Explore a few key tables

Take a look at the daily_activity data.

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    3/25/2016      11004          7.11            7.11
## 2 1503960366    3/26/2016      17609         11.55           11.55
## 3 1503960366    3/27/2016      12736          8.53            8.53
## 4 1503960366    3/28/2016      13231          8.93            8.93
## 5 1503960366    3/29/2016      12041          7.85            7.85
## 6 1503960366    3/30/2016      10970          7.16            7.16
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               2.57                     0.46
## 2                        0               6.92                     0.73
## 3                        0               4.66                     0.16
## 4                        0               3.19                     0.79
## 5                        0               2.16                     1.09
## 6                        0               2.36                     0.51
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                4.07                       0                33
## 2                3.91                       0                89
## 3                3.71                       0                56
## 4                4.95                       0                39
## 5                4.61                       0                28
## 6                4.29                       0                30
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  12                  205              804     1819
## 2                  17                  274              588     2154
## 3                   5                  268              605     1944
## 4                  20                  224             1080     1932
## 5                  28                  243              763     1886
## 6                  13                  223             1174     1820
```

Identify all the columns in the daily_activity data.

```
colnames(daily_activity)
```

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

Take a look at the sleep_day data.

```
head(sleep_day)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

Identify all the columns in the daily_activity data.

```
colnames(sleep_day)
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

Data Cleaning

```
# Inspect data for missing or duplicate values
sum(is.na(daily_activity))  # Check for missing values in daily_activity
```

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

```
sum(is.na(sleep_day))       # Check for missing values in sleep_day
```

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

```
# Remove duplicates
daily_activity <- daily_activity %>% distinct()
sleep_day <- sleep_day %>% distinct()
```

```
# Check for duplicate IDs
daily_activity %>% group_by(Id) %>% summarise(count = n()) %>% filter(count > 1)
```

```
## # A tibble: 35 × 2
##            Id count
##         <dbl> <int>
##  1 1503960366    19
##  2 1624580081    19
##  3 1644430081    10
##  4 1844505072    12
##  5 1927972279    12
##  6 2022484408    12
##  7 2026352035    12
##  8 2320127002    12
##  9 2347167796    15
## 10 2873212765    12
## # ℹ 25 more rows
```

```
sleep_day %>% group_by(Id) %>% summarise(count = n()) %>% filter(count > 1)
```

```
## # A tibble: 23 × 2
##            Id count
##         <dbl> <int>
##  1 1503960366    25
##  2 1644430081     4
##  3 1844505072     3
##  4 1927972279     5
##  5 2026352035    28
##  6 2347167796    15
##  7 3977333714    28
##  8 4020332650     8
##  9 4319703577    26
## 10 4388161847    23
## # ℹ 13 more rows
```

```
# Ensure consistent date formatting
sleep_day$ActivityDate <- as.Date(sleep_day$SleepDay, format = "%m/%d/%Y")
daily_activity$ActivityDate <- as.Date(daily_activity$ActivityDate, format = "%m/%
d/%Y")
```

```
# Check and align column data types
str(daily_activity)
```

```
## 'data.frame':    457 obs. of  15 variables:
##  $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : Date, format: "2016-03-25" "2016-03-26" ...
##  $ TotalSteps               : int  11004 17609 12736 13231 12041 10970 12256 1226
2 11248 10016 ...
##  $ TotalDistance            : num  7.11 11.55 8.53 8.93 7.85 ...
##  $ TrackerDistance          : num  7.11 11.55 8.53 8.93 7.85 ...
##  $ LoggedActivitiesDistance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num  2.57 6.92 4.66 3.19 2.16 ...
##  $ ModeratelyActiveDistance: num  0.46 0.73 0.16 0.79 1.09 ...
##  $ LightActiveDistance      : num  4.07 3.91 3.71 4.95 4.61 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : int  33 89 56 39 28 30 33 47 40 15 ...
##  $ FairlyActiveMinutes      : int  12 17 5 20 28 13 12 21 11 30 ...
##  $ LightlyActiveMinutes     : int  205 274 268 224 243 223 239 200 244 314 ...
##  $ SedentaryMinutes         : int  804 588 605 1080 763 1174 820 866 636 655 ...
##  $ Calories                 : int  1819 2154 1944 1932 1886 1820 1889 1868 1843 1
850 ...
```

```
str(sleep_day)
```

```
## 'data.frame':    410 obs. of  6 variables:
##  $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr  "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/1
5/2016 12:00:00 AM" "4/16/2016 12:00:00 AM" ...
##  $ TotalSleepRecords : int  1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: int  327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : int  346 407 442 367 712 320 377 364 384 449 ...
##  $ ActivityDate      : Date, format: "2016-04-12" "2016-04-13" ...
```

Understanding some summary statistics

How many unique participants are there in each dataframe?

It looks like there may be more participants in the daily activity

dataset than the sleep dataset.

```
n_distinct(daily_activity$Id)
```

```
## [1] 35
```

```
n_distinct(sleep_day$Id)
```

```
## [1] 24
```

How many observations are there in each dataframe?

```
nrow(daily_activity)
```

```
## [1] 457
```

```
nrow(sleep_day)
```

```
## [1] 410
```

What are some quick summary statistics we'd want to know about each data frame?

For the daily activity dataframe:

```
daily_activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :  32.0
##  1st Qu.: 1988   1st Qu.: 1.410   1st Qu.: 728.0
##  Median : 5986   Median : 4.090   Median :1057.0
##  Mean   : 6547   Mean   : 4.664   Mean   : 995.3
##  3rd Qu.:10198   3rd Qu.: 7.160   3rd Qu.:1285.0
##  Max.   :28497   Max.   :27.530   Max.   :1440.0
```

For the sleep dataframe:

```
sleep_day %>%
  select(TotalSleepRecords,
         TotalMinutesAsleep,
         TotalTimeInBed) %>%
  summary()
```

```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.00      Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.00      1st Qu.:361.0      1st Qu.:403.8
##  Median :1.00      Median :432.5      Median :463.0
##  Mean   :1.12      Mean   :419.2      Mean   :458.5
##  3rd Qu.:1.00      3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.00      Max.   :796.0      Max.   :961.0
```

What does this tell us about how this sample of people's activities?

Plotting a few explorations

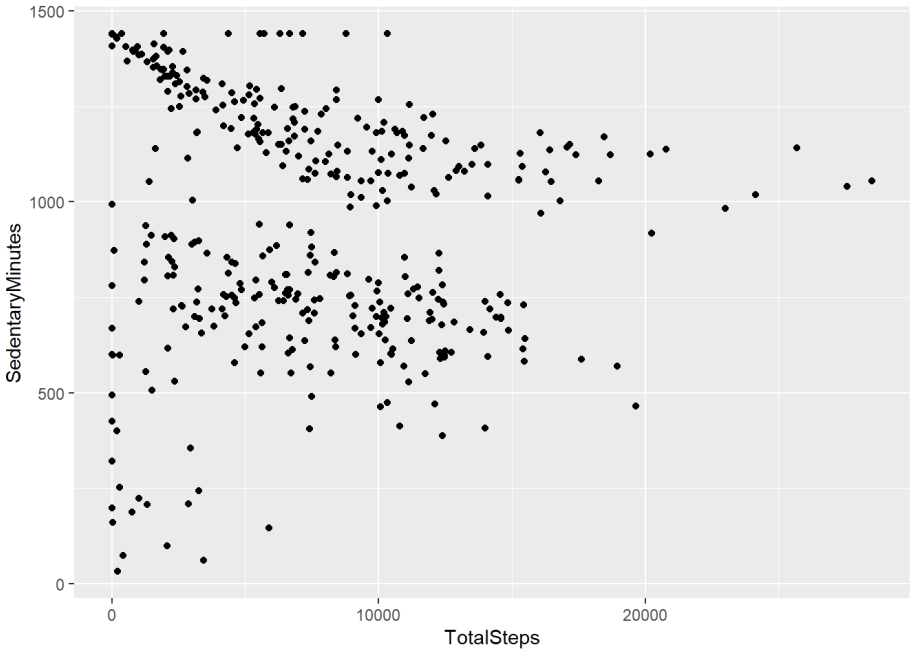
What's the relationship between steps taken in a day and sedentary minutes?

How could this help inform the customer segments that we can market to?

E.g. position this more as a way to get started in walking more?

Or to measure steps that you're already taking?

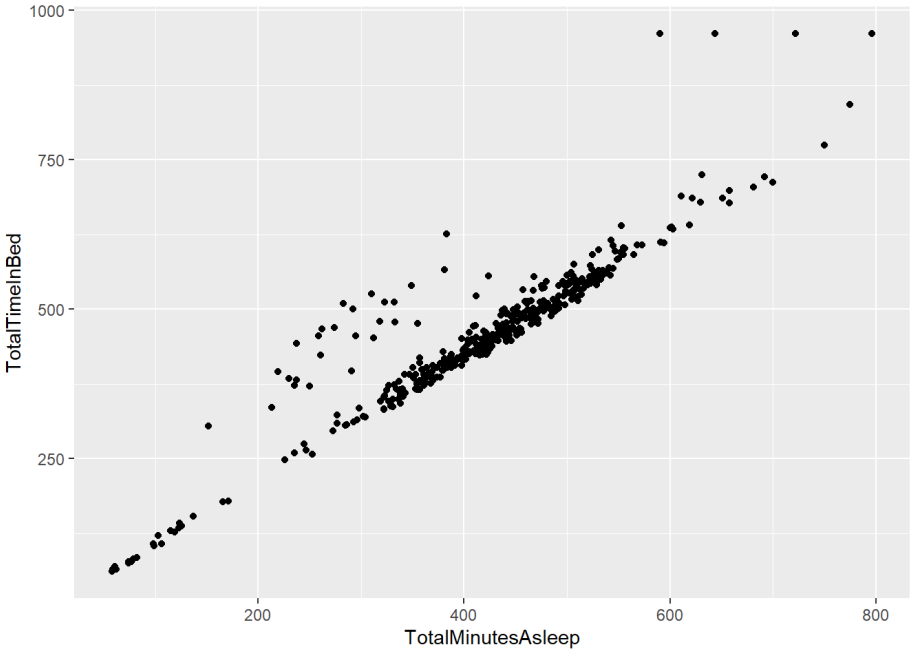
```
ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom_point()
```



What's the relationship between minutes asleep and time in bed?

You might expect it to be almost completely linear - are there any unexpected trends?

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom_point()
```



What could these trends tell you about how to help market this product? Or areas where you might want to explore further?

Merging these two datasets together

```
combined_data <- merge(sleep_day, daily_activity, by="Id")
```

```
head(combined_data)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/12/2016 12:00:00 AM                 1                327
## 3 1503960366 4/12/2016 12:00:00 AM                 1                327
## 4 1503960366 4/12/2016 12:00:00 AM                 1                327
## 5 1503960366 4/12/2016 12:00:00 AM                 1                327
## 6 1503960366 4/12/2016 12:00:00 AM                 1                327
##   TotalTimeInBed ActivityDate.x ActivityDate.y TotalSteps TotalDistance
## 1            346     2016-04-12     2016-04-09      12432          8.10
## 2            346     2016-04-12     2016-04-12        224          0.14
## 3            346     2016-04-12     2016-04-10      10057          6.98
## 4            346     2016-04-12     2016-03-26      17609         11.55
## 5            346     2016-04-12     2016-04-08      12521          7.94
## 6            346     2016-04-12     2016-03-27      12736          8.53
##   TrackerDistance LoggedActivitiesDistance VeryActiveDistance
## 1            8.10                       0               2.59
## 2            0.14                       0               0.00
## 3            6.98                       0               4.00
## 4           11.55                       0               6.92
## 5            7.94                       0               3.31
## 6            8.53                       0               4.66
##   ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
## 1                     0.59                4.92                       0
## 2                     0.00                0.13                       0
## 3                     0.49                2.48                       0
## 4                     0.73                3.91                       0
## 5                     0.90                3.74                       0
## 6                     0.16                3.71                       0
##   VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
## 1                32                  15                  248              738
## 2                 0                   0                    9               32
## 3                44                  13                  168              737
## 4                89                  17                  274              588
## 5                46                  22                  212             1160
## 6                56                   5                  268              605
##   Calories
## 1     1883
## 2       50
## 3     1755
## 4     2154
## 5     1895
## 6     1944
```

Take a look at how many participants are in this data set.

```
n_distinct(combined_data$Id)
```

```
## [1] 24
```

Merging Datasets

```
# Merge datasets on 'Id' and 'ActivityDate'
combined_data <- merge(
  x = sleep_day,
  y = daily_activity,
  by = c("Id"),
  all = TRUE  # Keeps all rows for analysis; use 'all=FALSE' for inner join
)
```

```
# Remove redundant or unnecessary columns
combined_data <- combined_data %>% select(-c(SleepDay))  # Removing duplicate time
field
```

Data Transformation

```
# Create derived metrics
combined_data <- combined_data %>%
  mutate(
    SleepEfficiency = TotalMinutesAsleep / TotalTimeInBed, # Calculate sleep effici
ency
    ActivityToSedentaryRatio = TotalSteps / SedentaryMinutes # Activity-to-sedentar
y ratio
  )
```

```
# Remove rows with invalid or extreme outlier values
combined_data <- combined_data %>%
  filter(TotalSteps >= 0 & TotalMinutesAsleep >= 0 & SedentaryMinutes >= 0)
```

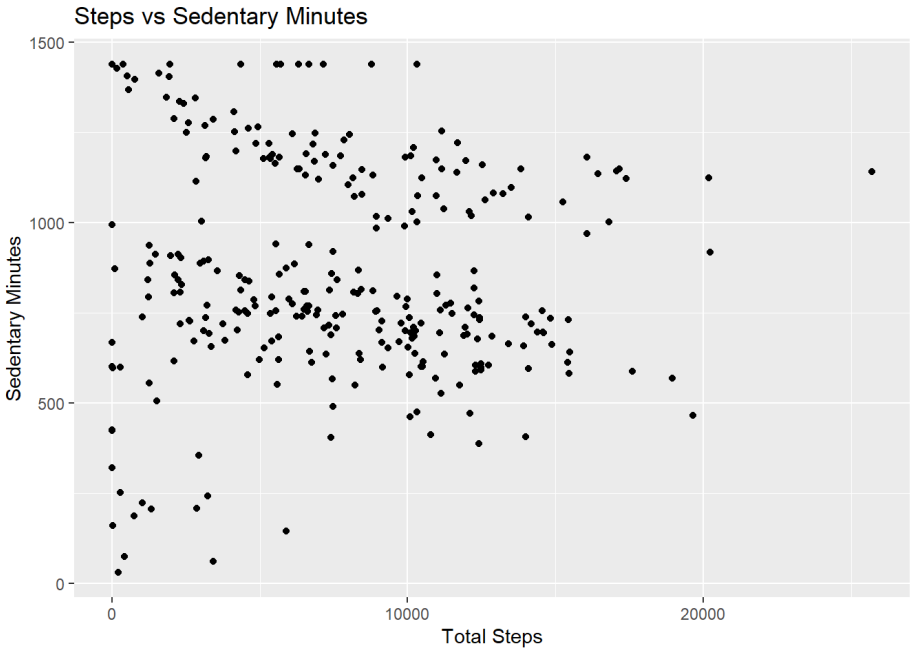
```
# Aggregate data to summarize by user (e.g., average values per user)
user_summary <- combined_data %>%
  group_by(Id) %>%
  summarise(
    AvgSteps = mean(TotalSteps, na.rm = TRUE),
    AvgSleepMinutes = mean(TotalMinutesAsleep, na.rm = TRUE),
    AvgSedentaryMinutes = mean(SedentaryMinutes, na.rm = TRUE),
    AvgSleepEfficiency = mean(SleepEfficiency, na.rm = TRUE)
  )
```

Verification

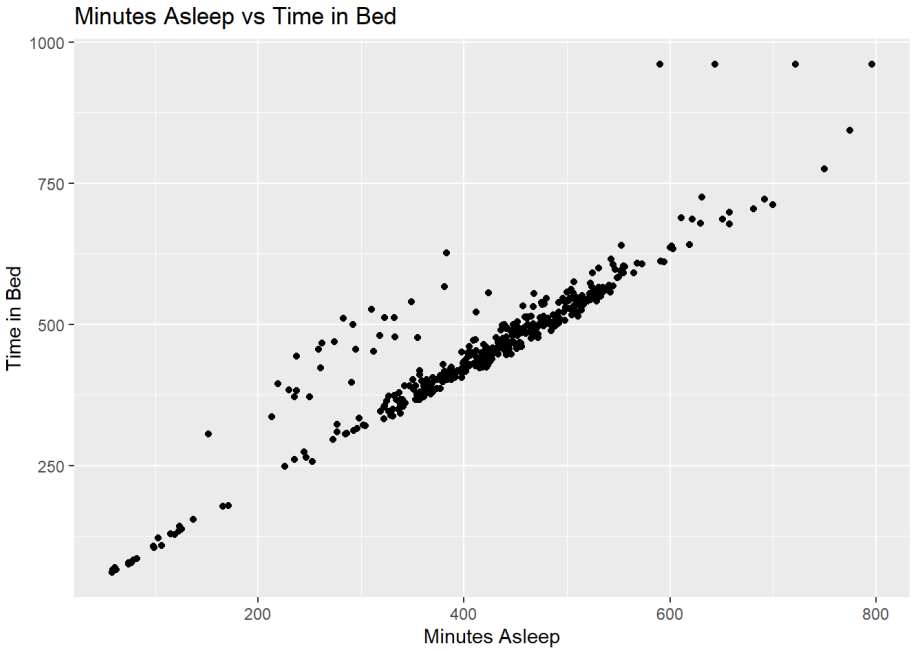
```
# Summary statistics
summary(combined_data)
```

```
##        Id            TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.504e+09   Min.   :1.000     Min.   : 58.0      Min.   : 61.0
##  1st Qu.:3.977e+09   1st Qu.:1.000     1st Qu.:357.0      1st Qu.:398.0
##  Median :4.445e+09   Median :1.000     Median :427.0      Median :459.0
##  Mean   :4.840e+09   Mean   :1.114     Mean   :415.4      Mean   :453.7
##  3rd Qu.:6.776e+09   3rd Qu.:1.000     3rd Qu.:485.0      3rd Qu.:522.0
##  Max.   :8.792e+09   Max.   :3.000     Max.   :796.0      Max.   :961.0
##  ActivityDate.x       ActivityDate.y         TotalSteps    TotalDistance
##  Min.   :2016-04-12   Min.   :2016-03-12   Min.   :    0   Min.   : 0.000
##  1st Qu.:2016-04-19   1st Qu.:2016-04-02   1st Qu.: 3436   1st Qu.: 2.390
##  Median :2016-04-27   Median :2016-04-05   Median : 7583   Median : 5.550
##  Mean   :2016-04-26   Mean   :2016-04-04   Mean   : 7577   Mean   : 5.297
##  3rd Qu.:2016-05-04   3rd Qu.:2016-04-09   3rd Qu.:11107   3rd Qu.: 7.710
##  Max.   :2016-05-12   Max.   :2016-04-12   Max.   :25701   Max.   :20.140
##  TrackerDistance  LoggedActivitiesDistance VeryActiveDistance
##  Min.   : 0.000   Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 2.390   1st Qu.:0.0000           1st Qu.: 0.000
##  Median : 5.550   Median :0.0000           Median : 0.290
##  Mean   : 5.273   Mean   :0.1818           Mean   : 1.209
##  3rd Qu.: 7.710   3rd Qu.:0.0000           3rd Qu.: 2.140
##  Max.   :20.140   Max.   :5.4569           Max.   :16.820
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.000      Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 1.960      1st Qu.:0.000000
##  Median :0.3600           Median : 3.500      Median :0.000000
##  Mean   :0.6596           Mean   : 3.351      Mean   :0.001225
##  3rd Qu.:0.9200           3rd Qu.: 4.690      3rd Qu.:0.000000
##  Max.   :6.4000           Max.   :12.510      Max.   :0.100000
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0        Min.   :  32.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:139.0        1st Qu.: 654.0
##  Median :  5.00    Median : 10.00      Median :205.0        Median : 738.0
##  Mean   : 20.83    Mean   : 14.93      Mean   :193.4        Mean   : 792.2
##  3rd Qu.: 33.00    3rd Qu.: 24.00      3rd Qu.:257.0        3rd Qu.: 868.0
##  Max.   :202.00    Max.   :141.00      Max.   :586.0        Max.   :1440.0
##     Calories    SleepEfficiency  ActivityToSedentaryRatio
##  Min.   :   0   Min.   :0.4984   Min.   : 0.000
##  1st Qu.:1819   1st Qu.:0.9113   1st Qu.: 5.111
##  Median :2133   Median :0.9420   Median :10.134
##  Mean   :2208   Mean   :0.9179   Mean   :11.285
##  3rd Qu.:2624   3rd Qu.:0.9607   3rd Qu.:15.781
##  Max.   :4430   Max.   :1.0000   Max.   :56.328
```

```
# Check relationships with scatterplots
ggplot(data = combined_data, aes(x = TotalSteps, y = SedentaryMinutes)) +
  geom_point() +
  labs(title = "Steps vs Sedentary Minutes", x = "Total Steps", y = "Sedentary Minu
tes")
```




```
ggplot(data = combined_data, aes(x = TotalMinutesAsleep, y = TotalTimeInBed)) +
  geom_point() +
  labs(title = "Minutes Asleep vs Time in Bed", x = "Minutes Asleep", y = "Time in
Bed")
```



Save Cleaned Data

```
# Save cleaned and transformed data for further analysis
write.csv(combined_data, "cleaned_combined_data.csv", row.names = FALSE)
write.csv(user_summary, "user_summary.csv", row.names = FALSE)
```