Code

Bellabeat Marketing Strategy - Google Data Analytics Capstone

Abdullahi Olapojoye

Bellabeat - A women-centric wellness technology company to beat

Bellabeat is a women-centric wellness technology company that develops wearables and accompanying products that monitor biometric and lifestyle data to help women better understand how their bodies work and make healthier choices.It was founded in 2013 by UrškaSršen and Sando Mur.More information can be found here (https://bellabeat.com/)

Primary Stakeholders:

UrškaSršen: Bellabeat's cofounder and Chief Creative Officer(CCO)

Sando Mur – Mathematician and Bellabeat's cofounder

Secondary Stakeholders:

Bellabeat Marketing Analytics Team: A data analyst team that collects, analyzes, and reports data that helps guide Bellabeat's marketing strategy

Business Task

Analyze data from non-Bellabeat products users to gain valuable insights about users habits that can guide the Bellabeat marketing strategy.

Tools Used:

- **R & SQL** - Data preparation, Cleaning and Analysis
- **R & Google Sheets** - Visualization and Pivot Table

Data Cleaning

The following cleaning activities were carried out:

- Inspect datasets
- Transform data
- Deal with null values
- Deal with inconsistent data formats

Install R packages:

Hide

```
#Packages used
install.packages("ggcorrplot")
library("ggcorrplot")
install.packages("ggpubr")
library("ggpubr")
install.packages("sqldf")
library("sqldf")
install.packages("tidyverse")
library("tidyverse")
```

Load the datasets:

18 csv files from the Fitbit data was loaded

Hide

```
daily_activity_data <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv")
daily_calories_data <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyCalories_merged.csv")
daily_intensities_data <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyIntensities_merged.cs
v")
daily_steps_data <- read_csv("Fitabase Data 4.12.16-5.12.16/dailySteps_merged.csv")
hourly_calories_data <- read_csv("Fitabase Data 4.12.16-5.12.16/hourlyCalories_merged.csv")
hourly_intensities_data <- read_csv("Fitabase Data 4.12.16-5.12.16/hourlyIntensities_merged.c
sv")
hourly_steps_data <- read_csv("Fitabase Data 4.12.16-5.12.16/hourlySteps_merged.csv")
minute_calories_data_narrow <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteCaloriesNarrow_m
erged.csv")
minute_intensities_data_narrow <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteIntensitiesNa
rrow_merged.csv")
minute_steps_data_narrow <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteStepsNarrow_merged.
csv")
minute_calories_data_Wide <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteCaloriesWide_merge
d.csv")
minute_intensities_data_Wide <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteIntensitiesWide
_merged.csv")
minute_steps_data_Wide <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteStepsWide_merged.cs
v")
minute_METS_data_Narrow <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteMETsNarrow_merged.cs
v")
minute_sleep_data <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteSleep_merged.csv")
sleep_day_data <- read_csv("Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv")
heartrate_seconds_data <- read_csv("Fitabase Data 4.12.16-5.12.16/heartrate_seconds_merged.cs
v")
weight_data <- read_csv("Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv")
```

Inspecting the loaded data:

Hide

```
#glimpse of the data
head(daily_activity_data)
head(daily_calories_data)
head(daily_intensities_data)
head(daily_steps_data)
head(sleep_day_data)
head(weight_data)
head(hourly_steps_data)
head(hourly_intensities_data)
head(hourly_calories_data)
head(minute_calories_data_narrow)
head(minute_calories_data_Wide)
head(minute_intensities_data_narrow)
head(minute_intensities_data_Wide)
head(minute_steps_data_narrow)
head(minute_steps_data_Wide)
head(heartrate_seconds_data)
head(minute_METS_data_Narrow)
head(minute_sleep_data)
#check the dimension
dim(daily_activity_data)
dim(daily_calories_data)
dim(daily_intensities_data)
dim(daily_steps_data)
dim(sleep_day_data)
dim(weight_data)
dim(hourly_steps_data)
dim(hourly_intensities_data)
dim(hourly_calories_data)
dim(minute_calories_data_narrow)
dim(minute_calories_data_Wide)
dim(minute_intensities_data_narrow)
dim(minute_intensities_data_Wide)
dim(minute_steps_data_narrow)
dim(minute_steps_data_Wide)
dim(heartrate_seconds_data)
dim(minute_METS_data_Narrow)
dim(minute_sleep_data)
```

Data Transformation:

Hide

```
# Create a new column TotalActiveMinutes which is the sum of columns "VeryActiveMinutes", "Fa
irlyActiveMinutes", "LightlyActiveMinutes":
daily_activity_data <- daily_activity_data %>%
  rowwise() %>%
  mutate(
    TotalActiveMinutes = sum(c(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinute
s)))

# Removing the columns 5 to 13 because they are already figured in the dataset:
daily_activity_final <- daily_activity_data %>%
  select(-c(5:13))

#merging daily_activity_data and sleep_day_data:
daily_data_final <- merge(daily_activity_final, sleep_day_data, by="Id")

#Previewing the new transformed dataframe:
head(daily_data_final)

#merge daily activity with weight:
daily_weight_data <- merge(daily_activity_final, weight_data, by="Id")

#Merge hourly_data:
hourly_data <- merge(hourly_calories_data, hourly_intensities_data)
hourly_data_final <- merge(hourly_data, hourly_steps_data)

#Previewing the new dataframe:
head(hourly_data_final)

#Transforming minute datasets
df <- merge(minute_calories_data_narrow, minute_intensities_data_narrow)
minute_datasets_combined <- merge(df, minute_steps_data_narrow)
minute_data_final <- merge(minute_datasets_combined, minute_METS_data_Narrow)

#Previewing the new dataframe:
head(minute_data_final)
```

```
# Checking each column of the dataframe daily_data_final for null values:

is.null(daily_data_final$Id)
is.null(daily_data_final$ActivityDate)
is.null(daily_data_final$TotalSteps)
is.null(daily_data_final$TotalDistance)
is.null(daily_data_final$SedentaryMinutes)
is.null(daily_data_final$Calories)
is.null(daily_data_final$TotalActiveMinutes)

# Checking each column of the dataframe hourly_data_combined for null values:

is.null(hourly_data_final$Id)
is.null(hourly_data_final$ActivityHour)
is.null(hourly_data_final$Calories)
is.null(hourly_data_final$TotalIntensity)
is.null(hourly_data_final$AverageIntensity)
is.null(hourly_data_final$StepTotal)

# Checking each column of the dataframe minute_data_combined for null values:

is.null(minute_data_final$Id)
is.null(minute_data_final$ActivityMinute)
is.null(minute_data_final$Calories)
is.null(minute_data_final$Intensity)
is.null(minute_data_final$Steps)
is.null(minute_data_final$METs)
```

Reformating the data:

Hide

```
#Reformat the data and extract weekdays from date
daily_data_final$ActDate <- strptime(as.character(daily_data_final$ActivityDate), "%m/%d/%Y")
daily_data_final$NewActivityDate <- format(daily_data_final_v2$ActDate, "%Y-%m-%d")
daily_data_final$Week_day <- wday(daily_data_final$NewActivityDate, label=TRUE, abbr = FALSE)
```

Result of cleaning :

1. No null or missing values was found in the data

4. daily_weight_data

Number of unique Ids:

Hide

```
# Inspecting the disitnct Id's in each of the four dataframes:

n_distinct(daily_data_final$Id)
n_distinct(hourly_data_final$Id)
n_distinct(minute_data_final$Id)
n_distinct(daily_weight_data$Id)
```

Only 24% and 21% of users tracked their weight and heart rate respectively. There were 33 unique users.

Tracker Usage Summary

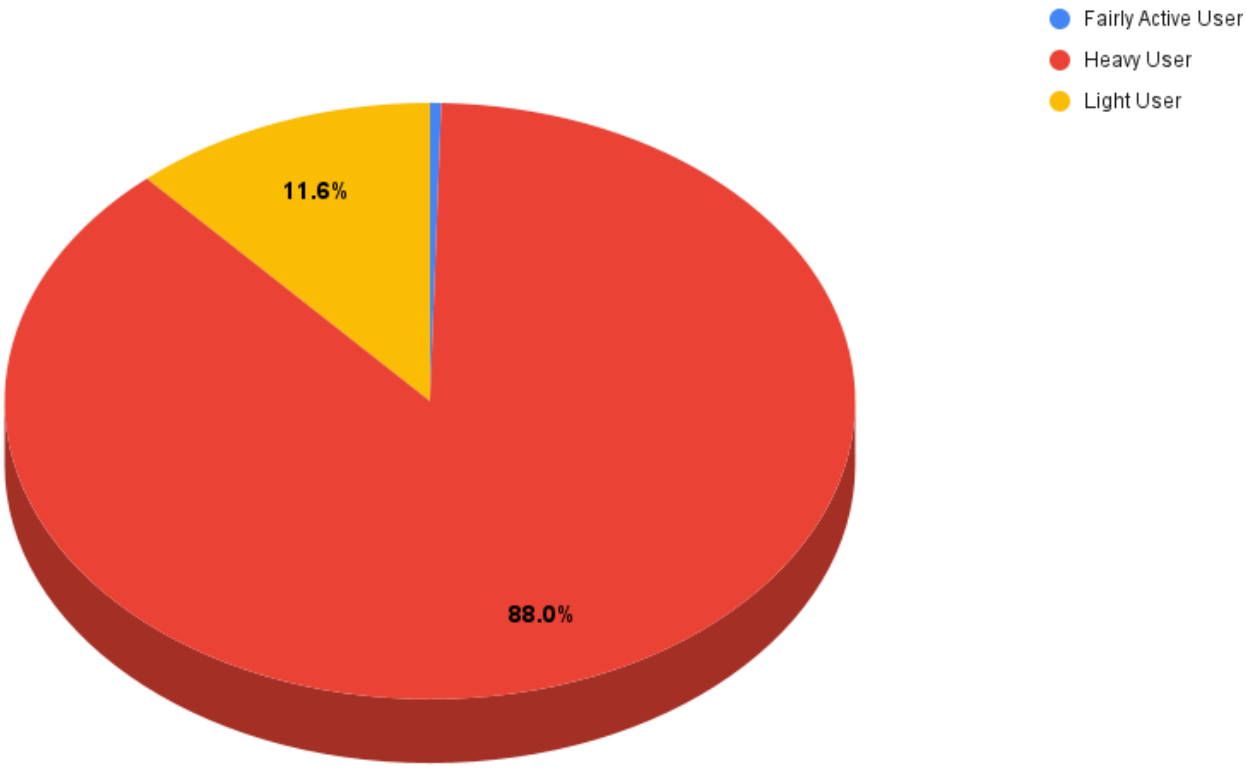
Fitbit users can be classified into four categories

1. Heavy User who used more than 27 days.
2. Light User who used between 11 and 20 days.
3. Fairly Active User who used between 1 and 10 days.
4. Inactive User who used for 0 day.

Hide

```
User_sum <- sqldf("SELECT Id,
          COUNT(Id) AS Total_Logged_Uses,
          CASE
          WHEN COUNT(Id) >27 THEN 'Heavy User'
          WHEN COUNT(Id) BETWEEN 16 and 27 THEN 'Light User'
          WHEN COUNT(Id) BETWEEN 1 and 15 THEN 'Fairly Active User'
          WHEN COUNT(Id) == 0 THEN 'Inactive User'
          END AS User_Summary
          FROM daily_activity_data
          GROUP BY Id")
view(User_sum)
```

Fitness Tracker User Types Chart - Google Sheets



This plot shows that majority of participants use the Fitbit Tracker actively.

Converting the minutes data to hours and creating new columns in daily_data_final dataframe:

Hide

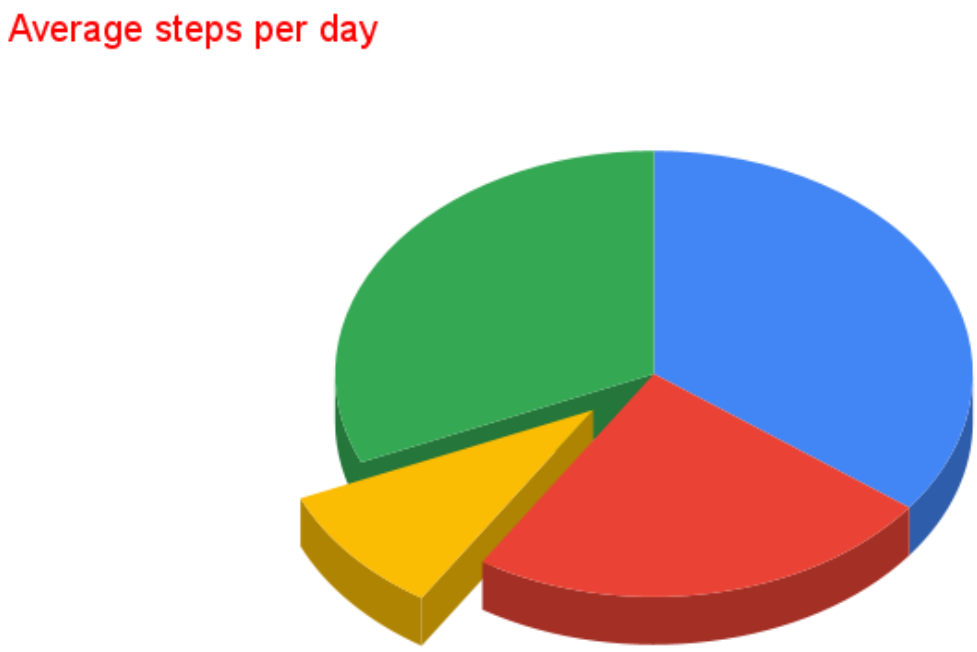
```
# Adding columns "TotalHoursAsleep","TotalActiveHours", "TotalSedentaryHours":

daily_data_final_v2 <- daily_data_final %>%
  mutate(TotalHoursAsleep = round(TotalMinutesAsleep/60, 1),
         SedentaryHours = round(SedentaryMinutes/60, 1),
         TotalActiveHours = round(TotalActiveMinutes/60, 1),
         TotalHoursInBed = round(TotalTimeInBed/60, 1))

# Previewing the transformed dataframe:
head(daily_data_final_v2)
```

```
Avg_Steps_per_day <- sqldf("SELECT Id,
ROUND(AVG(TotalSteps)) AS Avg_Total_Steps,
CASE
WHEN ROUND(AVG(TotalSteps)) < 5000 THEN 'Not recommended'
WHEN ROUND(AVG(TotalSteps)) BETWEEN 5000 AND 7499 THEN 'Fair range'
WHEN ROUND(AVG(TotalSteps)) BETWEEN 7500 AND 9999 THEN 'Recommended range'
WHEN ROUND(AVG(TotalSteps)) > 10000 THEN 'Above recommended range'
END AS User_Type
FROM daily_activity_data
GROUP BY Id")
```

Average Steps per Day Chart - Google sheets (https://docs.google.com/spreadsheets/d/11phEJxBFuiqXRSWfwyQsvuTQ2nc9o8wrl9qWNawsILk/edit?usp=sharing)

Hide

```
knitr::include_graphics("Plots/Average steps per day.png")
```

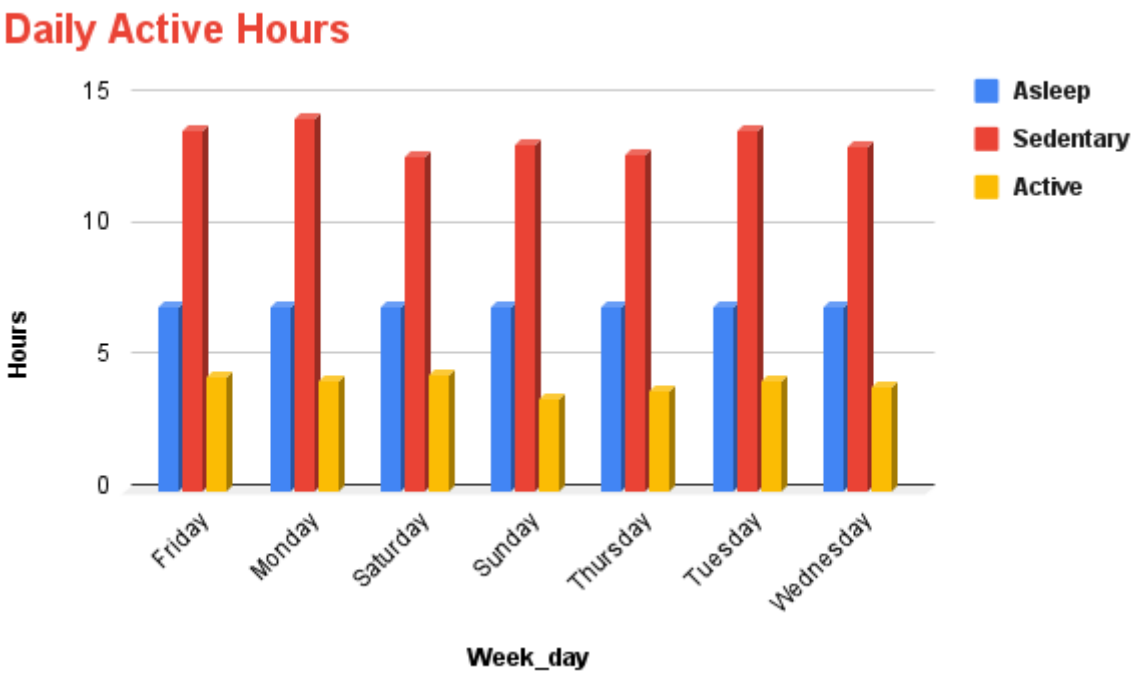


```
daily_data_final_v2$ActDate <- strptime(as.character(daily_data_final_v3$ActivityDate), "%m/%d/%Y")
daily_data_final_v2$NewActivityDate <- format(daily_data_final_v2$ActDate, "%Y-%m-%d")
daily_data_final_v2$Week_day <- wday(daily_data_final_v2$NewActivityDate, label=TRUE, abbr = FALSE)
```

Realationship between Active hours and Weekdays - Google Sheets (https://docs.google.com/spreadsheets/d/1ysW6v63rrpAV1dIHdShGTPqQnBwBZtIldAy649SGPEE/edit?usp=sharing)

Hide

```
knitr::include_graphics("Plots/Daily Active Hours.png")
```



It can be inferred from the plot that users' behavior do not change with the day of the week. ### Average Steps per hour:

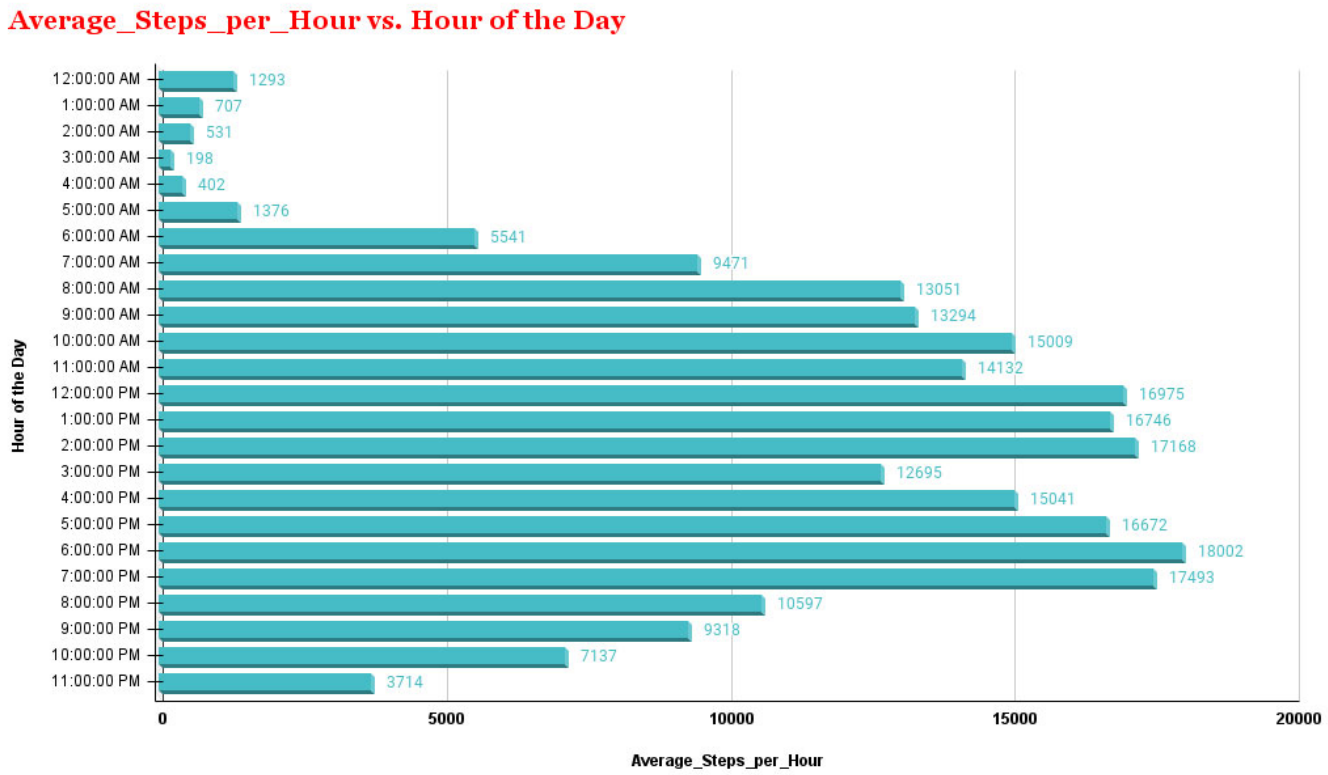
Hide

```
Avg_steps_per_hour <- sqldf("SELECT
```

Average Steps per Hour Chart - Google sheets (https://docs.google.com/spreadsheets/d/1RLp6w67JyP7C0GdCEXbH0lhUcG4kKiTXInmfG4cf8-M/edit?usp=sharing)

Hide

```
knitr::include_graphics("Plots/Average_Steps_per_Hour vs. Hour of the Day.png")
```

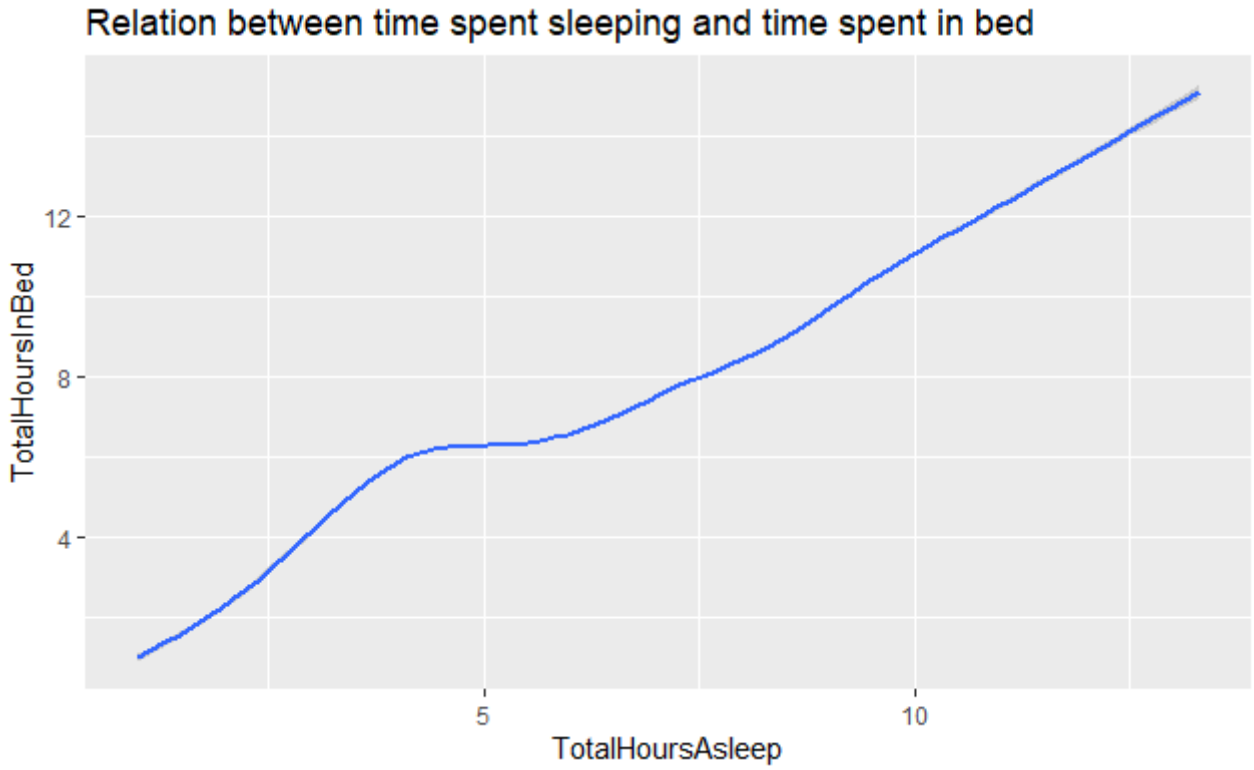


This plot pints to the fact that Fitbit users are more active in the afternoon and evening. Less activity is seen in the night.

Aggregate data on relationships between calories, sleep and steps:

Hide

```
### Case_when() to create new variable to aggregate the data from dataframes: daily_data_comb
ined_v2 and weight_info_v2:

daily_data_final_v3 <- daily_data_final_v2 %>%
  mutate(sleep_summary = case_when(TotalHoursAsleep>=6 & TotalHoursAsleep <=8   ~ "6-8", Tota
lHoursAsleep >8 ~ ">8", TRUE ~ "<6"),
         calories_burned = case_when(Calories>=1800 & Calories <=2900   ~ "Medium(1800-290
0)", Calories > 2900 ~ "High(>2900)", TRUE ~ "Low(0-1800)"),
         total_steps_taken = case_when(TotalSteps>=4600 &  TotalSteps <=8100   ~ "Steps(4600-
8100)", TotalSteps>8100 &  TotalSteps <11300   ~ "Steps(8100-11300)", TotalSteps >= 11300 ~
"Steps(>11300)", TRUE ~ "Steps(<4660)"),
         total_distance_travelled = case_when(TotalDistance>=3 &  TotalDistance <=6   ~ "Dist
ance(3-6)", TotalDistance>6 &  TotalDistance <=8   ~ "Distance(6-8)", TotalDistance > 8 ~ "Di
stance(>8)", TRUE ~ "Distance(<3)"),
         sedentary_hours_spent= case_when(SedentaryHours>=11 & SedentaryHours <=13   ~ "Seden
tary Hours(11-13)",  SedentaryHours > 13 ~ "Sedentary Hours(>13)", TRUE ~ "Sedentary Hours(<1
1)"),
         total_active_hours_spent = case_when(TotalActiveHours>=3.2 & TotalActiveHours <=5.2
~ "Average(3.2-5.2)",  TotalActiveHours > 5.2 ~ "High(>5.2)", TRUE ~ "Low(<3.2)"))

daily_weight_data_v2 <- daily_weight_data %>%
  mutate(weight_category = case_when(WeightKg>62 & WeightKg <85  ~ "62-85", WeightKg >=85 ~
">85", TRUE ~ "<62"))

# Sorting the observations of the above aggregated data frames respectively:
daily_data_final_v3$sleep_summary <- factor(daily_data_final_v3$sleep_summary,levels = c("<
6", "6-8", ">8"))
daily_data_final_v3$total_distance_travelled <- factor(daily_data_final_v3$total_distance_tra
velled,levels = c("Distance(<3)", "Distance(3-6)", "Distance(6-8)", "Distance(>8)"))
daily_data_final_v3$sedentary_hours_spent <- factor(daily_data_final_v3$sedentary_hours_spen
t,levels = c("Sedentary Hours(<11)", "Sedentary Hours(11-13)", "Sedentary Hours(>13)"))
daily_data_final_v3$total_steps_taken <- factor(daily_data_final_v3$total_steps_taken,levels
= c("Steps(<4660)", "Steps(4600-8100)", "Steps(8100-11300)", "Steps(>11300)"))

daily_weight_data_v2$WeightKg <- factor(daily_weight_data_v2$WeightKg, levels = c("<62", "62-
85", ">85"))
```



As expected, there is a positive relationship between the total hours in bed and the total hours asleep as shown in the figure above.

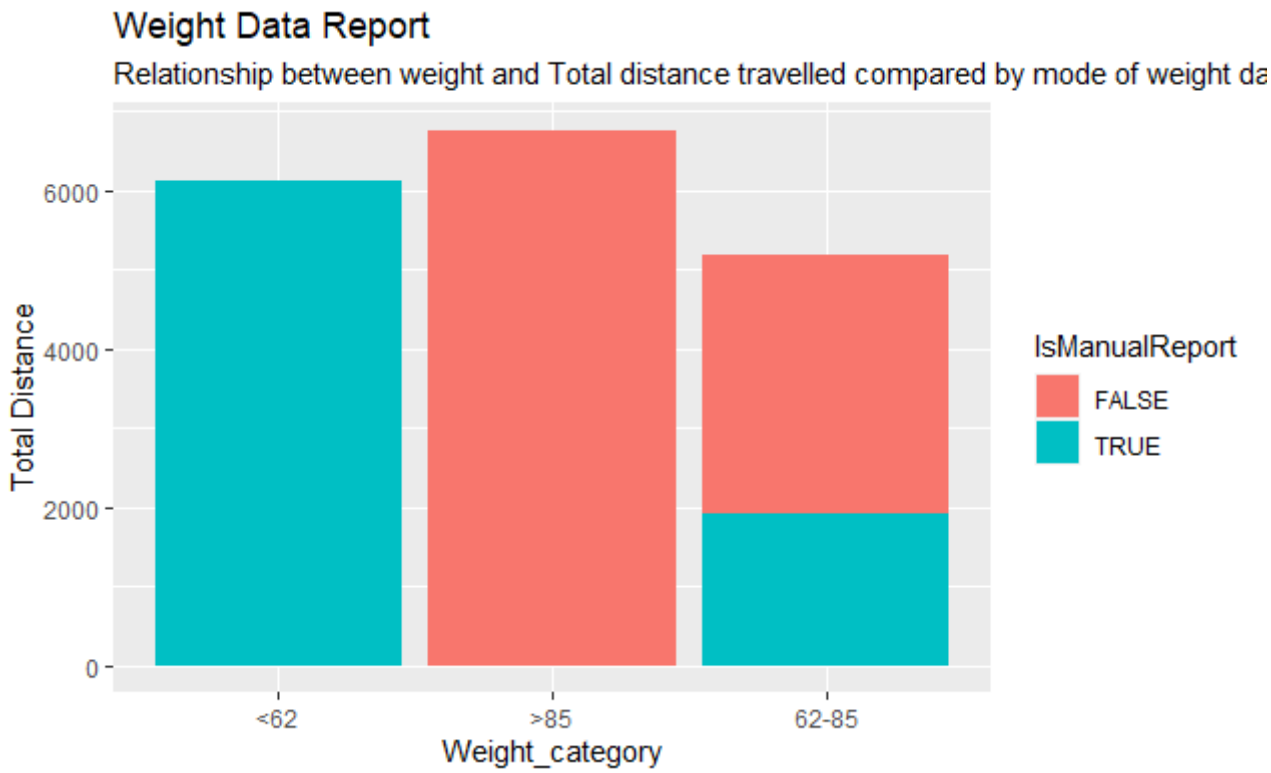
Correlation between steps and weight:

Hide

```
#Relationship between steps and weight
BMI_Steps <- sqldf("SELECT dwd.BMI,ddf.TotalSteps
FROM daily_weight_data AS dwd INNER JOIN daily_activity_data AS ddf ON dwd.Id = ddf.Id")

#view(BMI_Steps)
```

Relationship between weight and total distance compared by mode of report:

Hide

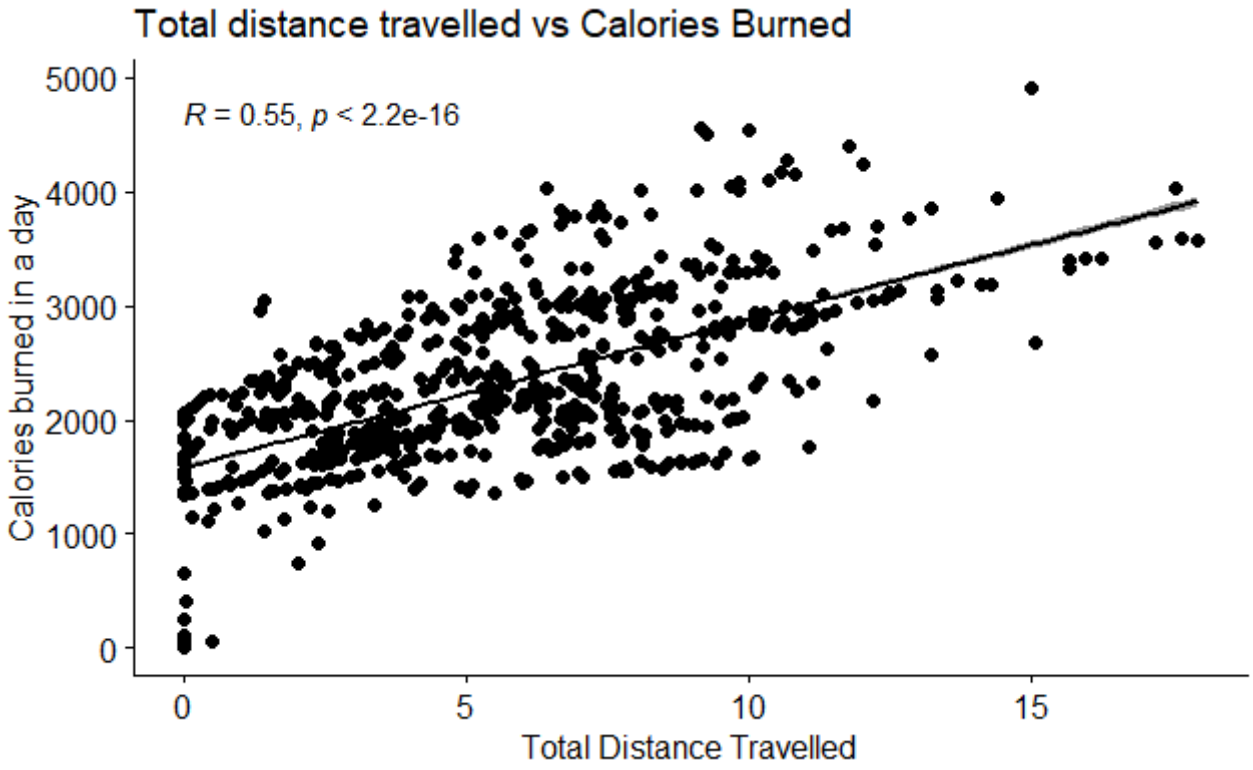


This plot shows that most of the weight data recorded were done manually and most people with weight above 85% do not record their weight data.

Relationship between Distance travelled and Calories burned in a day:

Hide

```
ggscatter(daily_data_final_v3, x = "TotalDistance", y = "Calories",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          title = "Total distance travelled vs Calories Burned",
          xlab = "Total Distance Travelled", ylab = "Calories burned in a day")
```

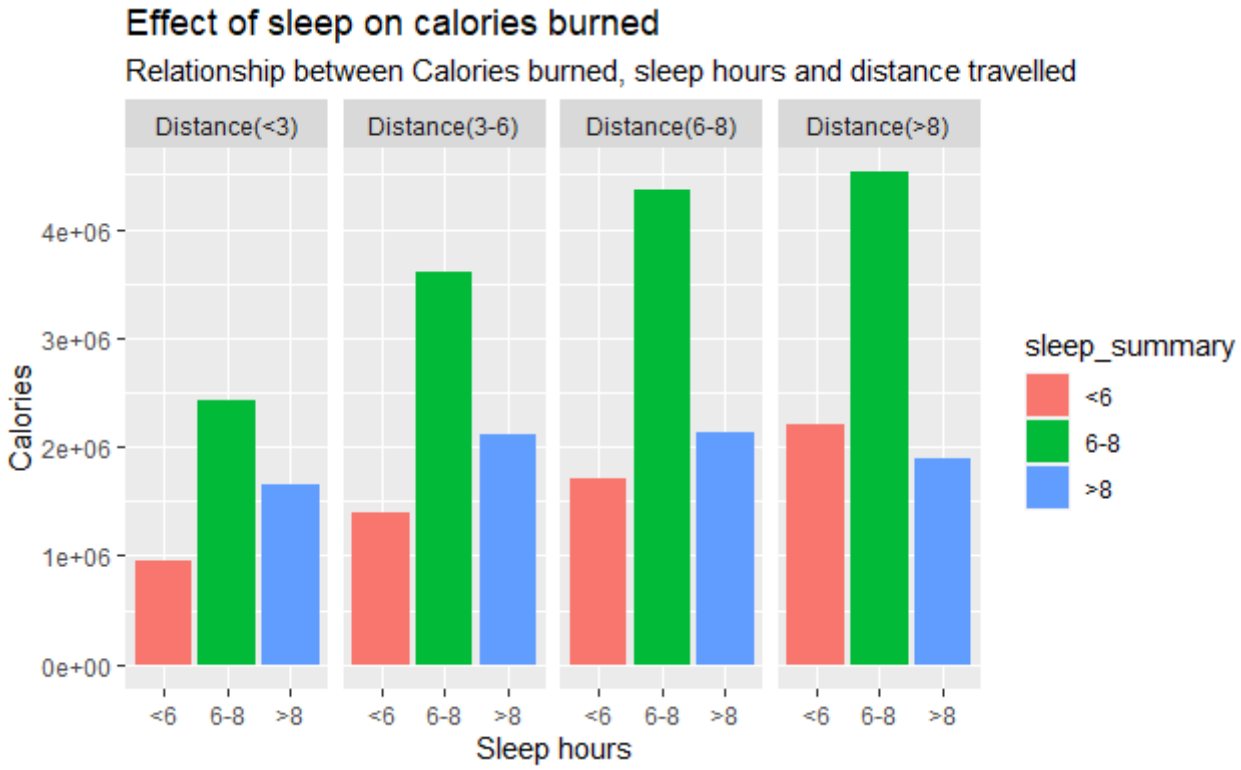


It can be deduced from the plot above that more calories are burned per day astotal distance travelled increases.

Relationship between sleep and calories burned compared by distance travelled:

Hide

```
ggplot(data=daily_data_final_v3)+
  geom_col(mapping=aes(x = sleep_summary, y = Calories, fill=sleep_summary))+
  facet_grid(~total_distance_travelled)+
  labs(title="Effect of sleep on calories burned", subtitle="Relationship between Calories bu
rned, sleep hours and distance travelled",
       x = "Sleep hours", "Calories Burned")
```

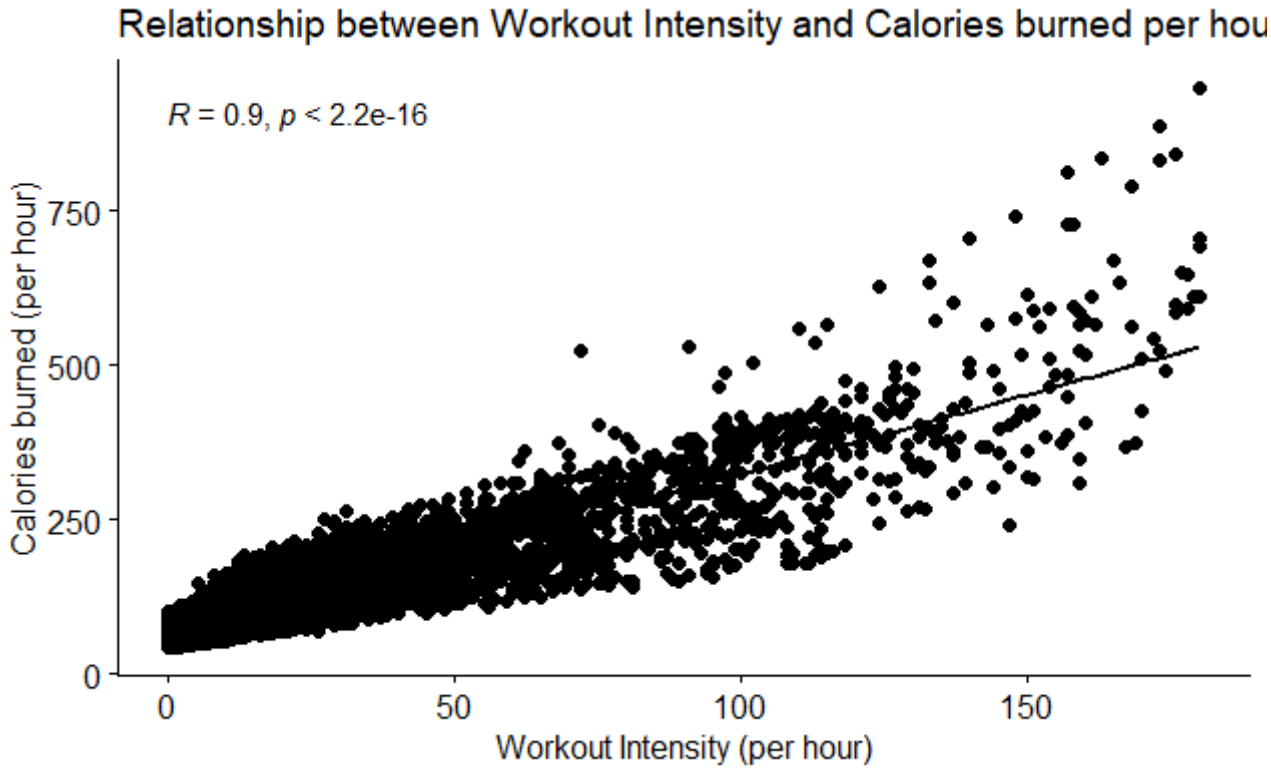


This plot shows that users who takes the recommended amount of sleep (6-8 hours) daily are able to cover more distance and burn more calories.

Relationship between Intensity of workout per hour and calories burned per hour:

Hide

```
ggscatter(hourly_data_final, x = "TotalIntensity", y = "Calories",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          add.params = list(color = "black",
                            fill = "yellow"),
          title = "Relationship between Workout Intensity and Calories burned per hour",
          xlab = "Workout Intensity (per hour)", ylab = "Calories burned (per hour)")
```



The higher the intensity of workout the more the calories burned daily as shown in the plot above.

Insights

- Weight and heartRate are the least tracked by the fitbit users. Only 24% and 21% of users track them.
- Sedentary hours - Most of the participants spent more time inactive.
- The average number of steps during weekdays and weekends are more or less similar signifying that there is no correlation between number of steps in the weekends or weekdays.
- Participants were most active in the evenings, fairly active in the afternoons and less active time at night.
- People who takes the recommended 6-8 hours sleep takes more steps and are able to burn more calories than those who do not.
- Users with healthy Body Mass Index(BMI) recorded more steps than others the users. Participants with high BMI tend to not input their weight data manually. This is important in understanding the users habits.