analyse_visualization

Buisness Task and relevent questions:

When starting the analysis, it is only natural to define what we are looking for, depending on the business task set by our stakeholders. In this case, I have been asked to analyse smart device usage data to find out how customers use competitors' devices. Then, based on the key findings I present, I will select the products to be developed.

Let's break down the main question:

1. What are some trends in smart device usage?

- How are users presenting themselves in terms of physical activities?
- What anomalies are there in the data?
- What kind of struggles or unhealthy behavior do users exhibit?
- What part of their active life needs to be corrected or changed?

2. How could these trends apply to Bellabeat customers?

- What features can we add to existing products to help our users?
- How can we change users' view of and perception of daily activities?
- How can we incorporate the device into people's lives?

3. How could these trends influence the Bellabeat marketing strategy?

- Presenting solutions that help users become the best versions of themselves.
- Promoting certain behaviors in a light and approachable way.
- Focus on aspects of assisting users.

Packages used in analysis.

```
library(janitor)
library(skimr)
library(tidyverse)
library(lubridate)
library(caTools)
```

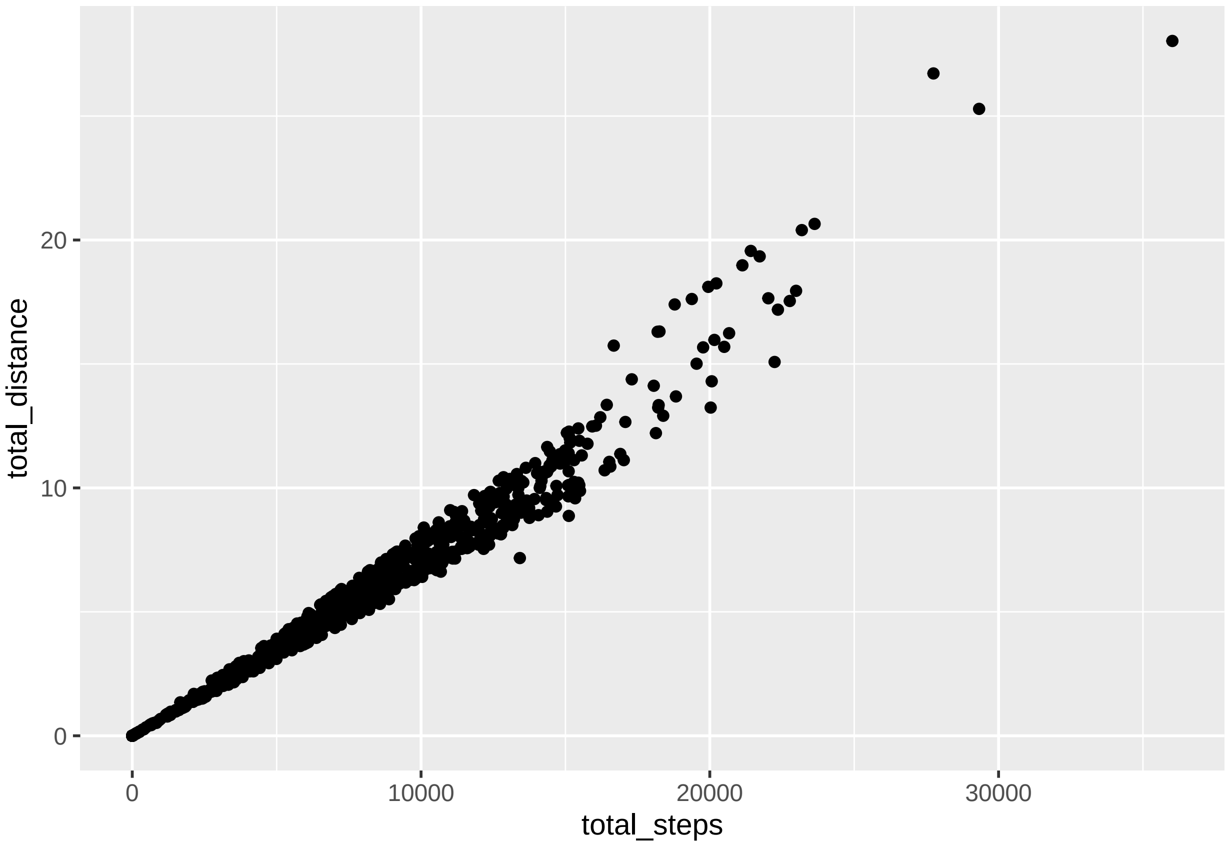
```
load(".RData")
```

Data sense-check:

Before I start analysing the data in more depth, I would like to perform a visual inspection to check that my reasoning is correct.

First I will check Distance and number of steps correlation

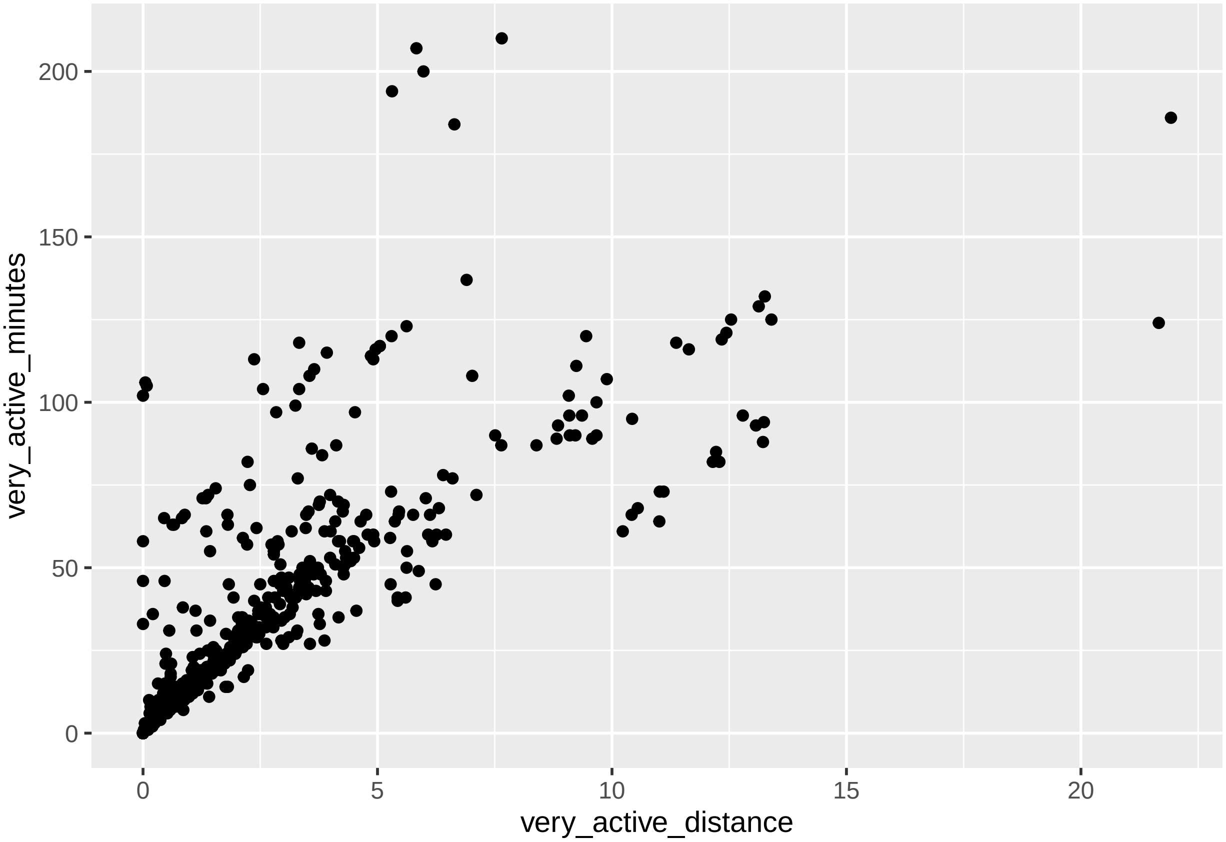
```
ggplot2::ggplot(data = activity) +
  geom_point(mapping = aes(x = total_steps, y = total_distance))
```



Everything seems correct the correlation is apparent

I wanted to check if the types of activity in terms of distance and minutes are correlated. The types of each activity are slightly different in terms of the distance and time spent on the activity. That's why I had to check if these values are correlated with each other.

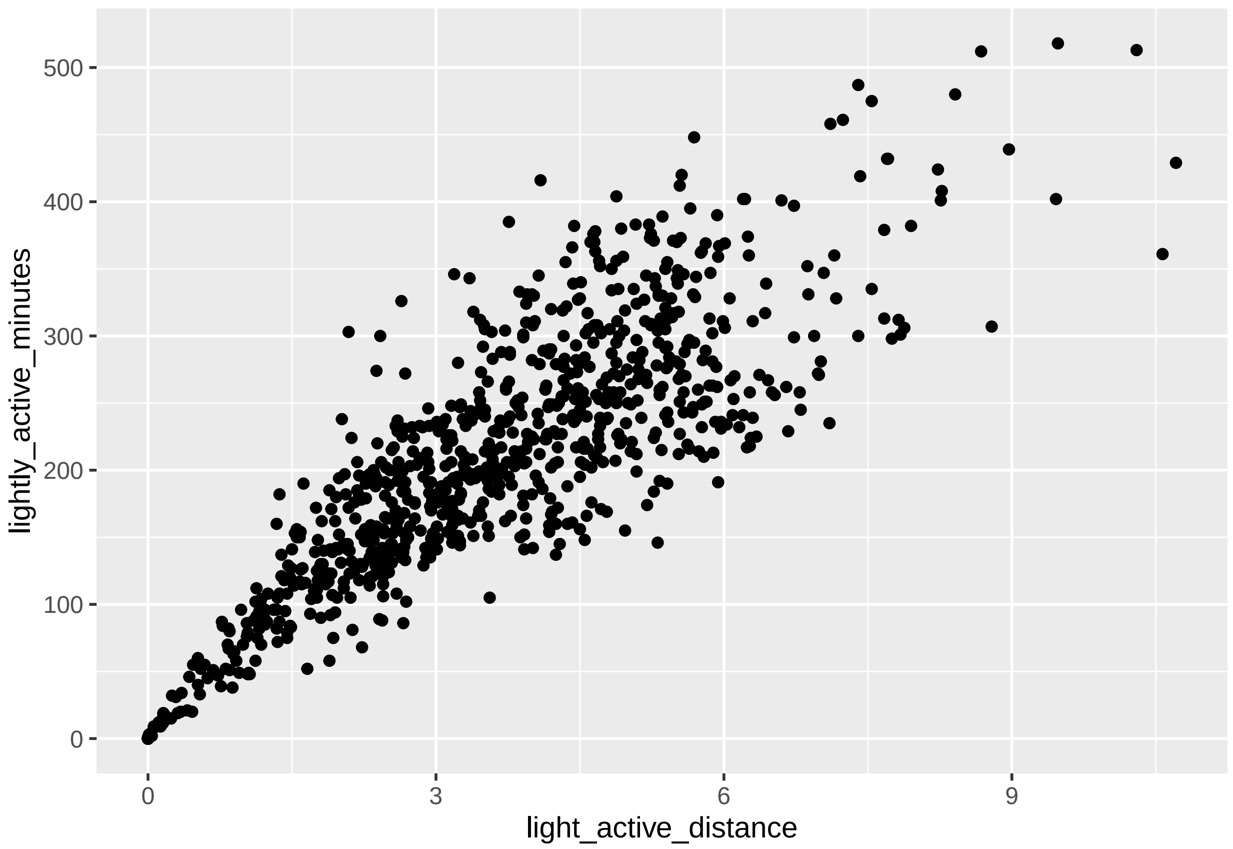
```
ggplot(data = activity) +
  geom_point(mapping = aes(x = very_active_distance, y = very_active_minutes))
```



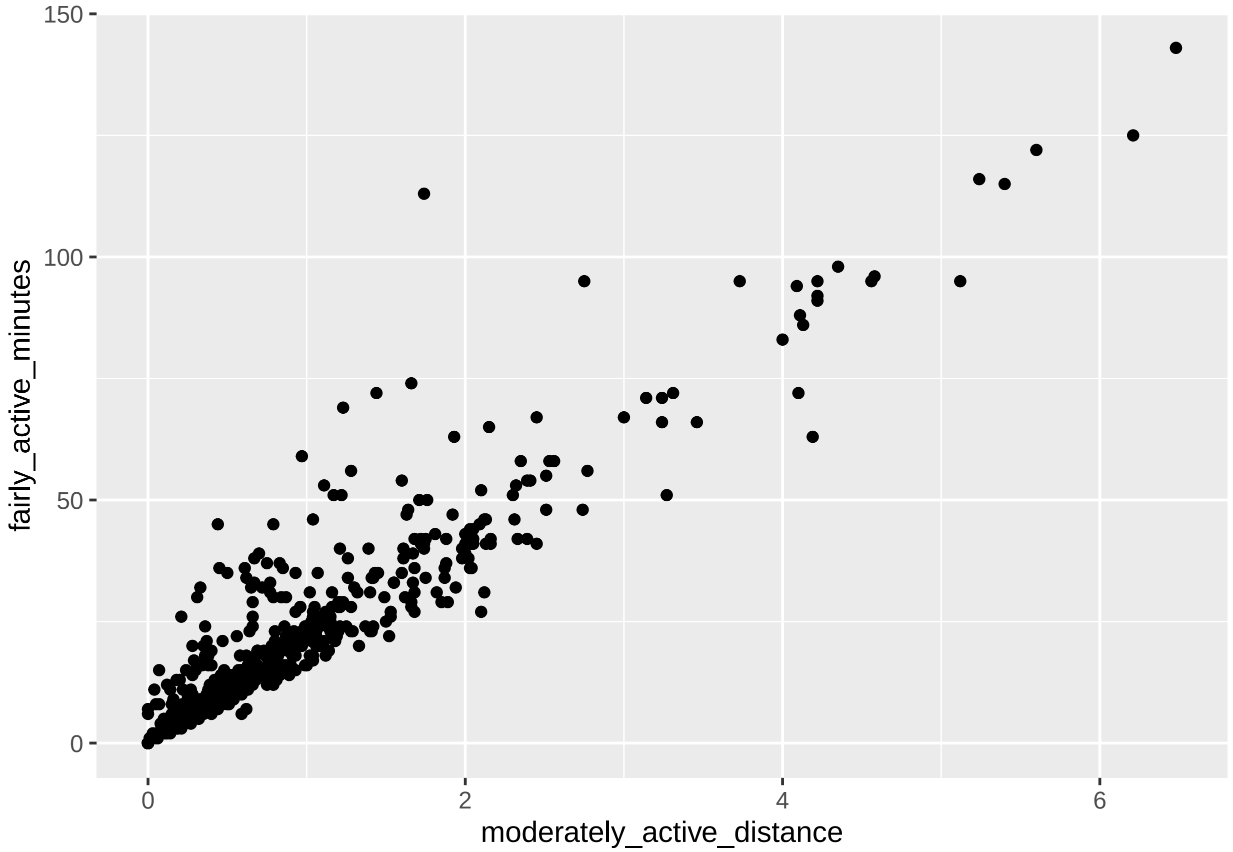
Initially, distance and active minutes are correlated with how normal people perform. Later, however, we can see that the data becomes more dispersed. This is very simple to explain. For example, a very active person can cover a longer distance in less time. Conversely, some people are not physically capable of doing this.

I have done the same thing with other activity categories and their respective distance counterparts.

```r
ggplot(data = activity) +
  geom_point(mapping = aes(x = light_active_distance, y = lightly_active_minutes))
```



```r
ggplot(data = activity) +
  geom_point(mapping = aes(x = moderately_active_distance, y = fairly_active_minutes))
```



Here, the correlation is visible a little better because we are at a lower level of physical exertion, so the performance will be similar for each user.

Daily data analysis:

The data grouped by day is what we're going to take a look at now.

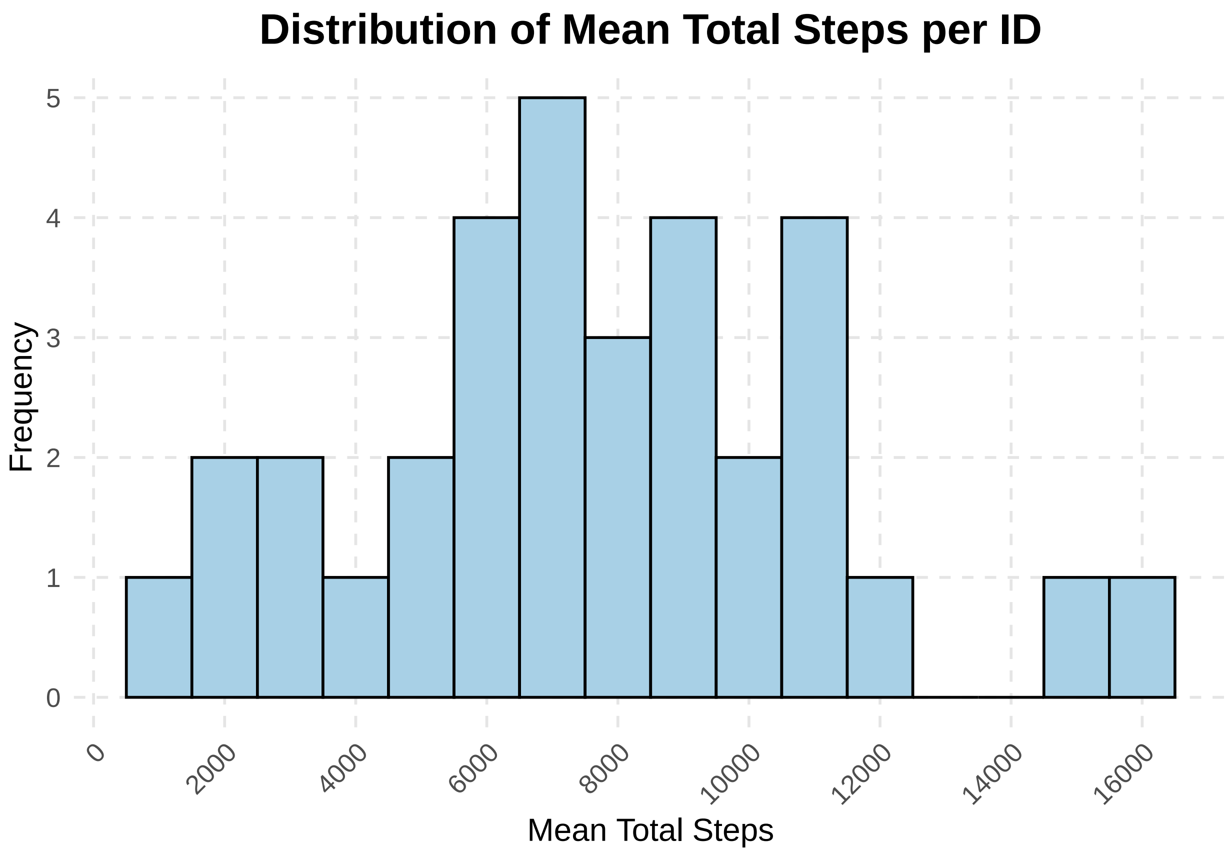
Distribution of mean total steps among users.

Firstly, I would like to find out what kind of users I am dealing with. Is this a sports group, or just regular people who want to exercise more using products that will help them?

```
activity %>%
  select(total_steps, sedentary_minutes) %>%
  summary()
```

```
##   total_steps    sedentary_minutes
##  Min.   :    0   Min.   :   0.0  
##  1st Qu.: 3790   1st Qu.: 729.8  
##  Median : 7406   Median :1057.5  
##  Mean   : 7638   Mean   : 991.2  
##  3rd Qu.:10727   3rd Qu.:1229.5  
##  Max.   :36019   Max.   :1440.0
```

```
# Calculate the mean total number of steps that will be used in the histogram.

activity %>%
  group_by(id) %>%
  summarise(mean_total_steps = mean(total_steps, na.rm = TRUE)) %>%
  arrange(mean_total_steps) %>%

# Plotting the histogram is the next step.

  ggplot(aes(x = mean_total_steps)) +
  geom_histogram(binwidth = 1000, fill = "#A8D0E6", color = "black") +
  labs(title = "Distribution of Mean Total Steps per ID",
       x = "Mean Total Steps",
       y = "Frequency") +
  scale_x_continuous(breaks = seq(0, max(activity$total_steps, na.rm = TRUE) + 2000, by = 2000)) + # se

# Making the histogram more visually appealing.

  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    panel.grid.major = element_line(color = "grey90", linetype = "dashed"),
    panel.grid.minor = element_blank())
```



On average, the observed group takes 7,638 steps per day. This is too low to provide any health benefits. According to this article "How many steps/day are enough? for adults", 10,000 steps/day is a reasonable target for healthy adults. Most users make an average of between 6,000 and 11,000 steps per day.

We can use this information to inform users of their average daily movement and motivate them to move a little more throughout the day.

Notifications about potential diseases that result from not exercising enough would be helpful, but I think it would be a bit of a stretch and could potentially damage the desire for change in habits.

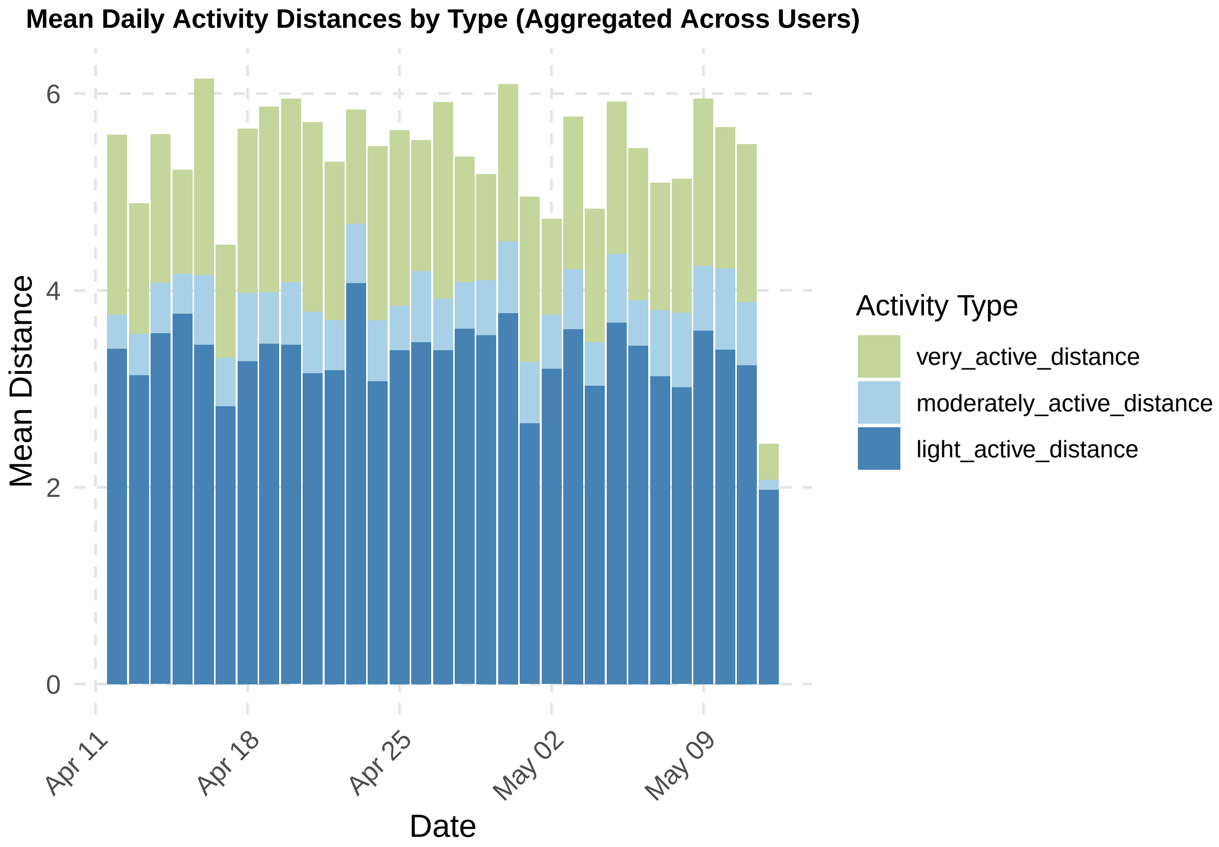
Mean Daily Activity Distances by Type:

Since we have a general idea of how many steps users are taking in general. Let's check what kind of activities they are keen on. This will also define the group. The most popular type of activity will determine what features and product developments we can propose to users.

```
longer_distance_type <- activity %>%
  pivot_longer(
    cols = c("very_active_distance", "moderately_active_distance", "light_active_distance"),
    names_to = "distance_type",
    values_to = "distance"
  ) %>%
  select(id, activity_date, distance_type, distance)
```

```
# Calculate the mean distances and store them in a new data frame
```

```r
mean_distances_per_id_type <- longer_distance_type %>%
  group_by(activity_date, distance_type) %>%
  summarise(mean_total_distance = mean(distance, na.rm = TRUE), .groups = 'drop') %>% # setting .groups
  arrange(activity_date, distance_type)

# A new colour palette is being introduced

my_image_colors <- c(
  "very_active_distance" = "#C4D69B",
  "moderately_active_distance" = "#A8D0E6",
  "light_active_distance" = "#4682B4"
)

# Ensure that the activity type is displayed properly

mean_distances_per_id_type$distance_type <- factor(
  mean_distances_per_id_type$distance_type,
  levels = c("very_active_distance", "moderately_active_distance", "light_active_distance")
)

# Plot the summarized data
ggplot(data = mean_distances_per_id_type, aes(x = activity_date, y = mean_total_distance, fill = distanc
  geom_bar(stat = "identity", position = "stack") + # set up 'position = "stack"' for a stacked bar cha
  labs(
    title = "Mean Daily Activity Distances by Type (Aggregated Across Users)",
    x = "Date",
    y = "Mean Distance",
    fill = "Activity Type"
  ) +

  # Making the histogram more visually appealing

  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 10),
    axis.title = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    panel.grid.major = element_line(color = "grey90", linetype = "dashed"),
    panel.grid.minor = element_blank()) +
  scale_fill_manual(values = my_image_colors)
```



As we can see, the average daily activity distance among app users remains at a low level of effort. Users engage mostly in light activity throughout the day. This suggests that, when developmentally improving products, we should focus on features that enhance 'light active' activities, or in other words, promote pushing yourself to a higher level of physical exertion.

In terms of the time users spend on each activity, we can encourage them to be more active at a higher level of exertion by sending messages informing them that their tempo is too slow and advising them to take longer steps and move faster to maintain a moderate or very active distance score.

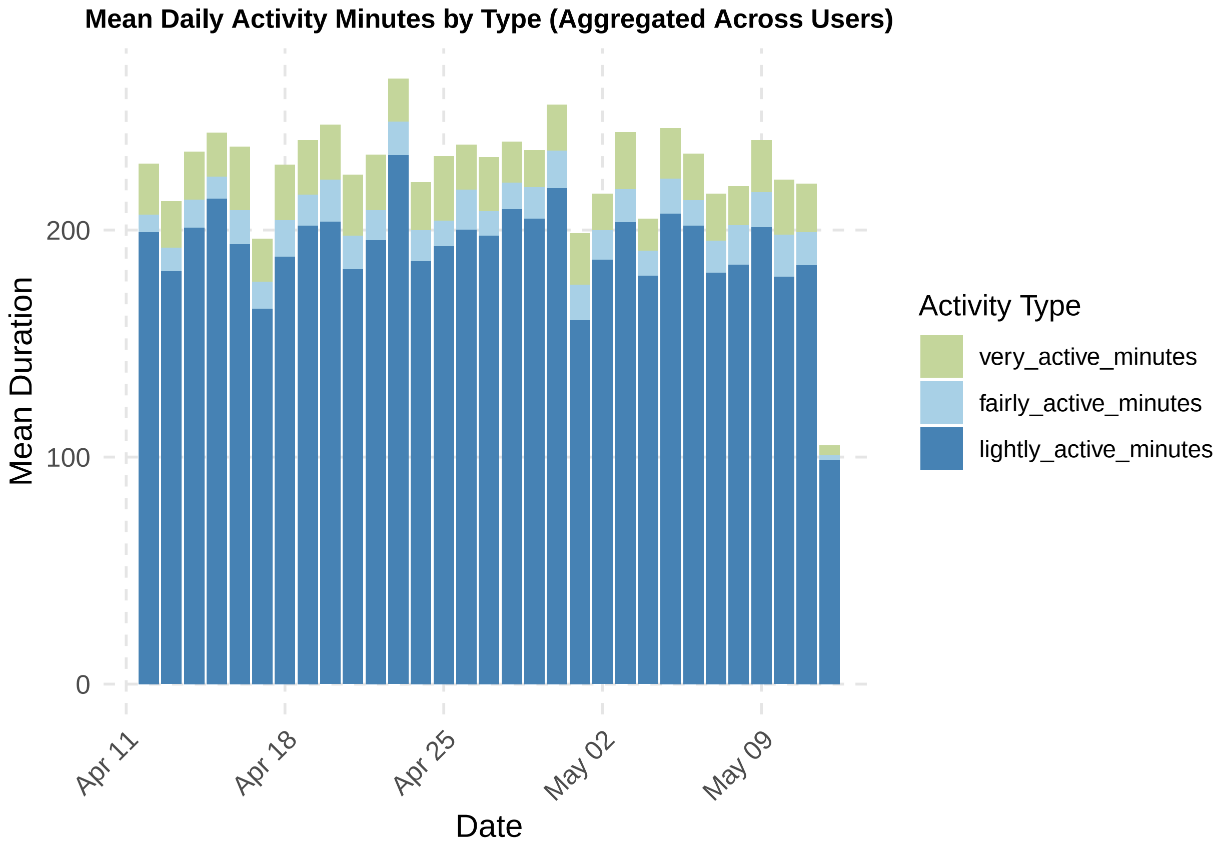
Mean daily activity minutes by type:

We have checked how active users are over a given distance. However, when performing certain exercises, you do not move much, so the distance covered is unlikely to be significant. That's why it is necessary to check whether the time spent on activity resembles any changes compared to the previous visualization.

```
longer_active_minutes_type <- activity %>%
  pivot_longer(
    cols = c("very_active_minutes", "fairly_active_minutes", "lightly_active_minutes"),
    names_to = "active_minute_type",
    values_to = "duration"
  ) %>%
  select(id, activity_date, active_minute_type, duration)
```

```
# Calculate the mean distances and store them in a new data frame

mean_minutes_per_id_type <- longer_active_minutes_type %>%
```

```r
  group_by(activity_date, active_minute_type) %>%
  summarise(mean_total_minutes = mean(duration, na.rm = TRUE), .groups = 'drop') %>%
  arrange(activity_date, active_minute_type)

# The same colour palette is used.

my_image_colors <- c(
  "very_active_minutes" = "#C4D69B",
  "lightly_active_minutes" = "#4682B4",
  "fairly_active_minutes" = "#A8D0E6"
)

# Ensure that the activity type is displayed properly.

mean_minutes_per_id_type$active_minute_type <- factor(
  mean_minutes_per_id_type$active_minute_type,
  levels = c("very_active_minutes", "fairly_active_minutes", "lightly_active_minutes")
)

# Plot the summarized data

ggplot(data = mean_minutes_per_id_type, aes(x = activity_date, y = mean_total_minutes, fill = active_mi
  geom_bar(stat = "identity", position = "stack") +
  labs(
    title = "Mean Daily Activity Minutes by Type (Aggregated Across Users)",
    x = "Date", # Corrected x-axis label to reflect activity_date
    y = "Mean Duration",
    fill = "Activity Type"
  ) +

  # Making the histogram more visually appealing.

  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 10),
    axis.title = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    panel.grid.major = element_line(color = "grey90", linetype = "dashed"),
    panel.grid.minor = element_blank()) +
  scale_fill_manual(values = my_image_colors)
```



The problem is that users need to reduce the time they spend on daily activities in order to achieve a high daily activity score. We need to implement tips on how to do this in the app. We should inform users what kind of exercises can easily boost the score and how long they would need to do them for to change their activity pattern during the day.

People enjoy games and achievements. It would be worth considering precalculated goals in the app to motivate users. This would provide a clear indication of how much work is needed to achieve specific results.

Sleep data analysis:

Just as we have to put effort into regulating our activity and exercise habits, we also have to be mindful of our sleep habits. Let's take a closer look at users' sleep patterns.

Brief look at sleep data.

First, we can summarise the data using the 'summary()' function.

```
sleep %>%
  select(total_sleep_records, total_minutes_asleep, total_time_in_bed) %>%
  summary()
```

```
##  total_sleep_records total_minutes_asleep total_time_in_bed
##  Min.   :1.000       Min.   : 58.0        Min.   : 61.0
##  1st Qu.:1.000       1st Qu.:361.0        1st Qu.:403.0
##  Median :1.000       Median :433.0        Median :463.0
##  Mean   :1.119       Mean   :419.5        Mean   :458.6
##  3rd Qu.:1.000       3rd Qu.:490.0        3rd Qu.:526.0
##  Max.   :3.000       Max.   :796.0        Max.   :961.0
```

```
sleep %>%
  group_by(total_sleep_records) %>%
  summarise(count_of_ids = n()) %>%
  arrange(total_sleep_records)
```

```
## # A tibble: 3 x 2
##   total_sleep_records count_of_ids
##                 <dbl>        <int>
## 1                   1          367
## 2                   2           43
## 3                   3            3
```

From this simple summary, we can see that sleep deprivation is not a major issue for this group. People tend to have one sleep session per day. However, the difference between the mean number of minutes slept and time spent in bed is concerning.I would like to see which days have the greatest difference in total time in bed and total sleep minutes.

Average daily sleep metrics by day of the week:

```
# Fix the date format

sleep$sleep_day=as.POSIXct(sleep$sleep_day, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
sleep$time <- format(sleep$sleep_day, format = "%H:%M:%S")
sleep$date <- format(sleep$sleep_day, format = "%m/%d/%y")
```

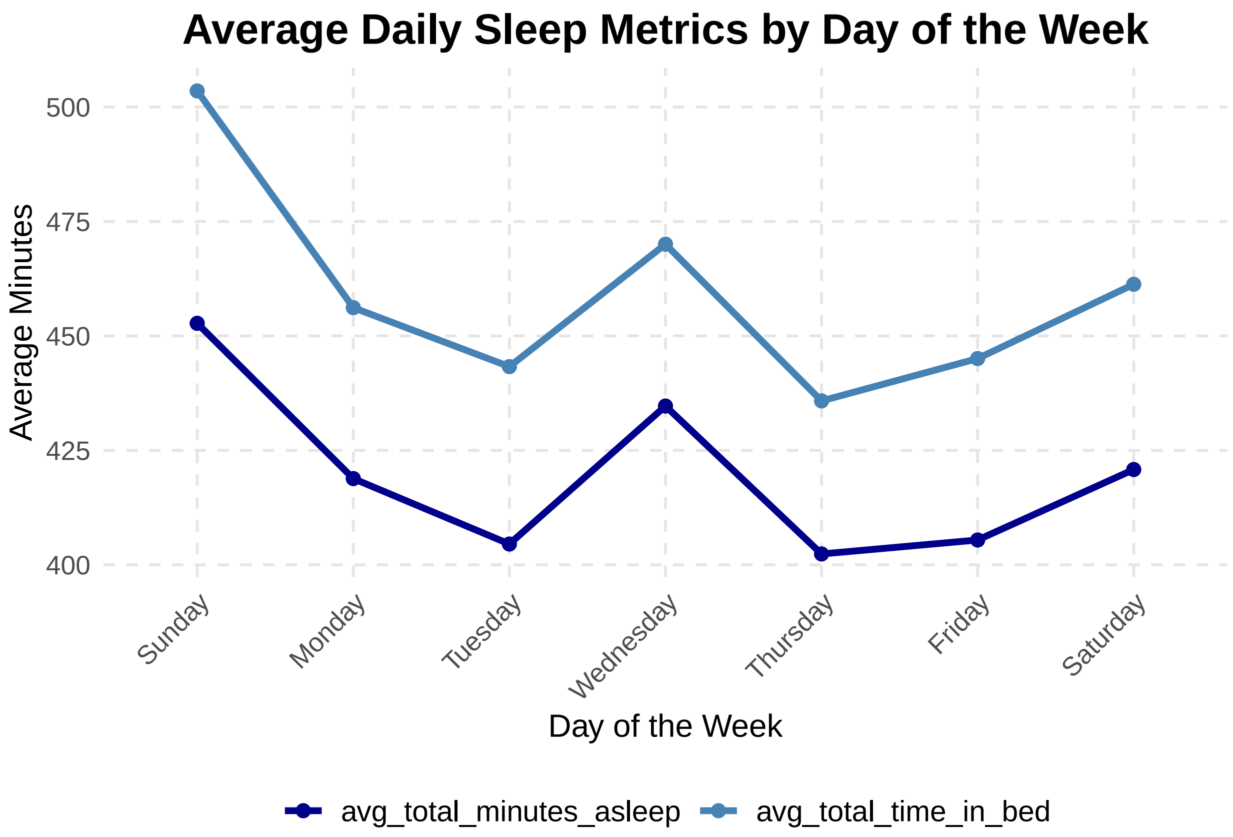
```
#  Use the wday() function to name the day

sleep<- sleep %>%
  mutate(
    day_of_week = wday(sleep_day, label = TRUE, abbr = FALSE), # Full day name
    week_number = isoweek(sleep_day) # ISO week number
  )
```

```
# Preparing data for the plot.

daily_sleep_averages <- sleep %>%
  group_by(day_of_week) %>%
  summarise(
    avg_total_minutes_asleep = mean(total_minutes_asleep, na.rm = TRUE), # After the cleaning phase, I
    avg_total_time_in_bed = mean(total_time_in_bed, na.rm = TRUE)
  ) %>%

  # Ensure the days are ordered correctly for the plot

  mutate(day_of_week = factor(day_of_week, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thur
  arrange(day_of_week)
```

```r
# Change the data format to ensure correct and easy plotting.

plot_data_sleep_longer <- daily_sleep_averages %>%
  pivot_longer(
    cols = c(avg_total_minutes_asleep, avg_total_time_in_bed),
    names_to = "metric",
    values_to = "minutes"
  )
```

```r
#  Plotting the prepared data

ggplot(plot_data_sleep_longer, aes(x = day_of_week, y = minutes, color = metric, group = metric)) +
  geom_line(size = 1.2) + # add lines
  geom_point(size = 2) + # add points for each day
  labs(
    title = "Average Daily Sleep Metrics by Day of the Week",
    x = "Day of the Week",
    y = "Average Minutes",
    color = "Metric"
  ) +
  scale_color_manual(values = c("avg_total_minutes_asleep" = "darkblue", "avg_total_time_in_bed" = "#468

  # Let's keep it consistent. ... Making the histogram more visually appealing.

  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    legend.position = "bottom",
    legend.title = element_blank(),
    legend.text = element_text(size = 11),
    panel.grid.major = element_line(color = "grey90", linetype = "dashed"),
    panel.grid.minor = element_blank()
  )
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



With the plot at our disposal, the only thing missing is the exact numbers to view.

```
weekly_difference <- sleep %>%
  mutate(
    difference = total_time_in_bed - total_minutes_asleep
  ) %>%
  group_by(day_of_week) %>%
  summarise(
    avg_weekly_difference = mean(difference, na.rm = TRUE)
  )

print(weekly_difference)
```

```
## # A tibble: 7 x 2
##   day_of_week avg_weekly_difference
##   <ord>                       <dbl>
## 1 Sunday                       50.8
## 2 Monday                       37.3
## 3 Tuesday                      38.8
## 4 Wednesday                    35.3
## 5 Thursday                     33.4
## 6 Friday                       39.6
## 7 Saturday                     40.5
```

In terms of sleep patterns, I can pinpoint a couple of things:

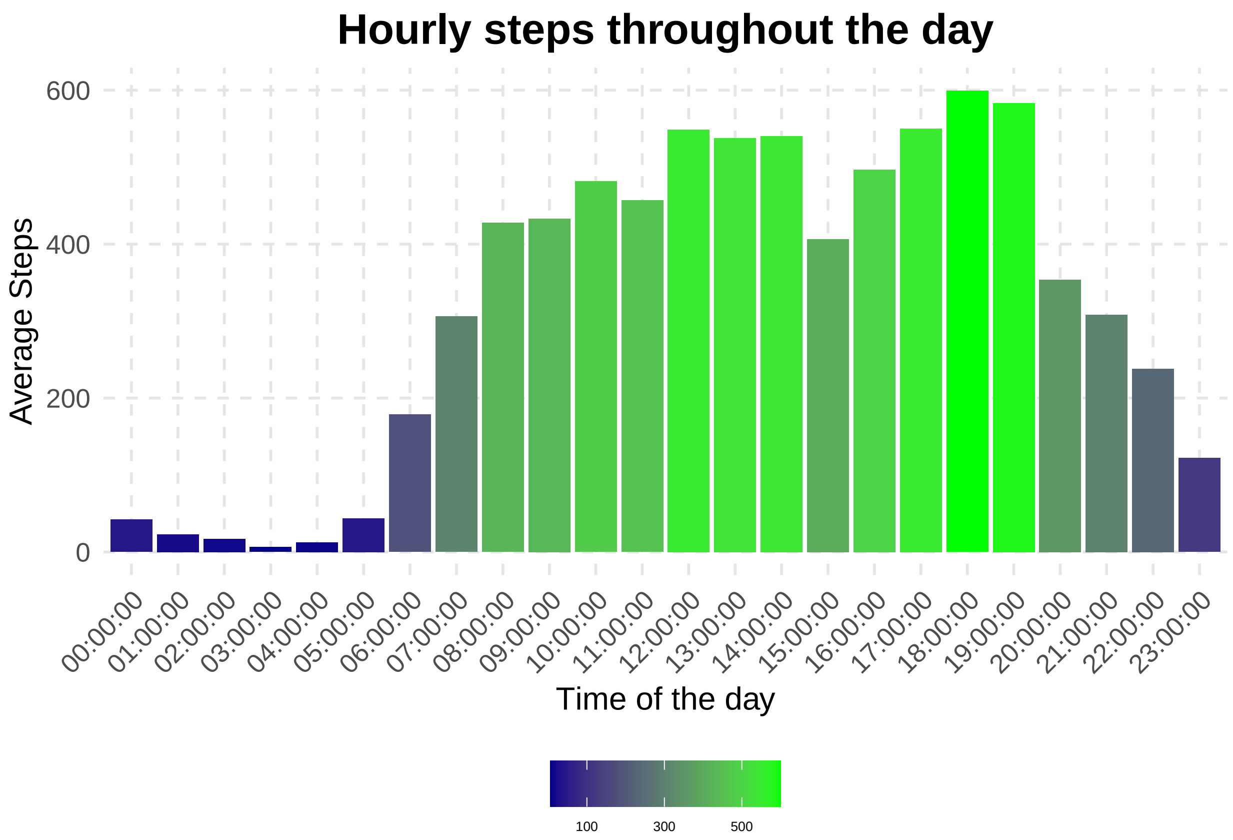
1. Users take over 30 minutes to fall asleep, which is an issue. This could be a sign of insomnia.
2. The sleep pattern is irregular – users do not spend the same amount of time in bed every day, which could lead to sleep deprivation. It is better to go to bed and wake up at the same time every day.
3. Users generally spend less time sleeping during a working day. An interesting pattern emerges. On Monday and Tuesday, users' time in bed decreases. Due to this, they need to spend more time in bed on Wednesday to make up for the previous weekday. This pattern reappears until the weekend.
4. Actual sleep time does not exceed 7.5 hours.

We can help users maintain healthy sleep habits. Additional features include setting preferred sleep hours, receiving a notification an hour before sleep time to remind them not to use their phone, and tips on how to calm down after a stressful day.

Hourly data analysis:

Using hourly data, I want to find out which part of the day users are most active.

```
# Preparing data for the plot

hourly_cis %>%
  group_by(time) %>%
  summarize(average_steps = mean(step_total)) %>%

# Plotting the prepared data

  ggplot() +
  geom_col(mapping = aes(x=time, y = average_steps, fill = average_steps)) +
  labs(
    title = "Hourly steps throughout the day",
    x="Time of the day", y="Average Steps",
    color = "Metric") +
  scale_fill_gradient(low = "darkblue", high = "green", breaks = c(100, 300, 500), )+
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    legend.position = "bottom",
    legend.title = element_blank(),
    legend.text = element_text(size = 5),
    panel.grid.major = element_line(color = "grey90", linetype = "dashed"),
    panel.grid.minor = element_blank())
```

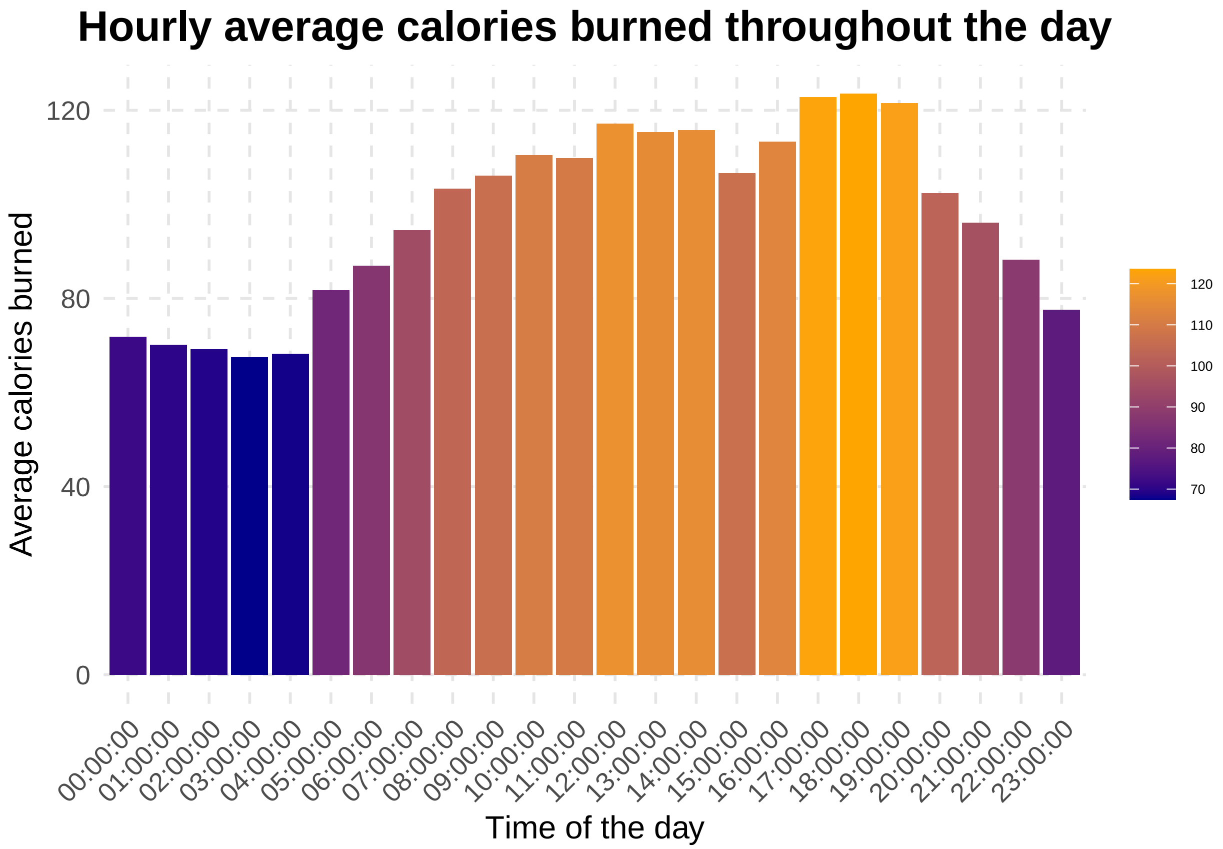


Users are most active, judging by their steps, between 8 AM and 7 PM. This makes sense, as it covers typical work hours and commutes.

We see noticeable step boosts during lunchtime (12 PM - 2 PM), likely from people walking to grab food or take a quick break. There's another big jump in the evenings (5 PM - 7 PM), probably due to the commute home and then daily tasks like grocery shopping or picking up kids.

However, it fails to address when users prefer to exercise. All exercise burns calories, even if it doesn't involve intense movement.

```
# Preparing data for the plot

hourly_cis %>%
  group_by(time) %>%
  summarize(average_calories = mean(calories)) %>%

# Plotting the prepared data

  ggplot() +
  geom_col(mapping = aes(x=time, y = average_calories, fill = average_calories)) +
  labs(
    title = "Hourly average calories burned throughout the day",
    x="Time of the day", y="Average calories burned",
    color = "Metric") +
  scale_fill_gradient(low = "darkblue", high = "orange", )+
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    legend.title = element_blank(),
    legend.text = element_text(size = 5),
    panel.grid.major = element_line(color = "grey90", linetype = "dashed"),
    panel.grid.minor = element_blank())
```



We obtained similar results: it seems that the time of day when people exercise and are most active is between 5 pm and 7 pm.

Therefore, it makes sense to target the 5–7 pm window for promoting activity goals, as this aligns with peak exercise times.

Minutes data analysis:

In order to finalise our findings, let's take a closer look at the minutes data.

```
minute_ciMs %>%
  select(calories, me_ts, steps) %>%
  summary()
```

```
##     calories           me_ts              steps
##  Min.   : 0.0000   Min.   :  0.00   Min.   :  0.000
##  1st Qu.: 0.9357   1st Qu.: 10.00   1st Qu.:  0.000
##  Median : 1.2176   Median : 10.00   Median :  0.000
##  Mean   : 1.6231   Mean   : 14.69   Mean   :  5.336
##  3rd Qu.: 1.4327   3rd Qu.: 11.00   3rd Qu.:  0.000
##  Max.   :19.7499   Max.   :157.00   Max.   :220.000
```

I’ve analyzed the data regarding calorie expenditure and activity levels. The average calorie burn among users is 1.62 calories per minute, with an average of 5 steps taken per minute.

A notable concern arises from the average Metabolic Equivalent of Task (MET) value, which is stated as average 15 METs per minute. This figure is unusually high, as even very vigorous activities typically do not reach this intensity. This discrepancy raises questions about the accuracy of the app’s measurement methodology.

For products with measurement instruments, it’s crucial to ensure accuracy and user understanding.

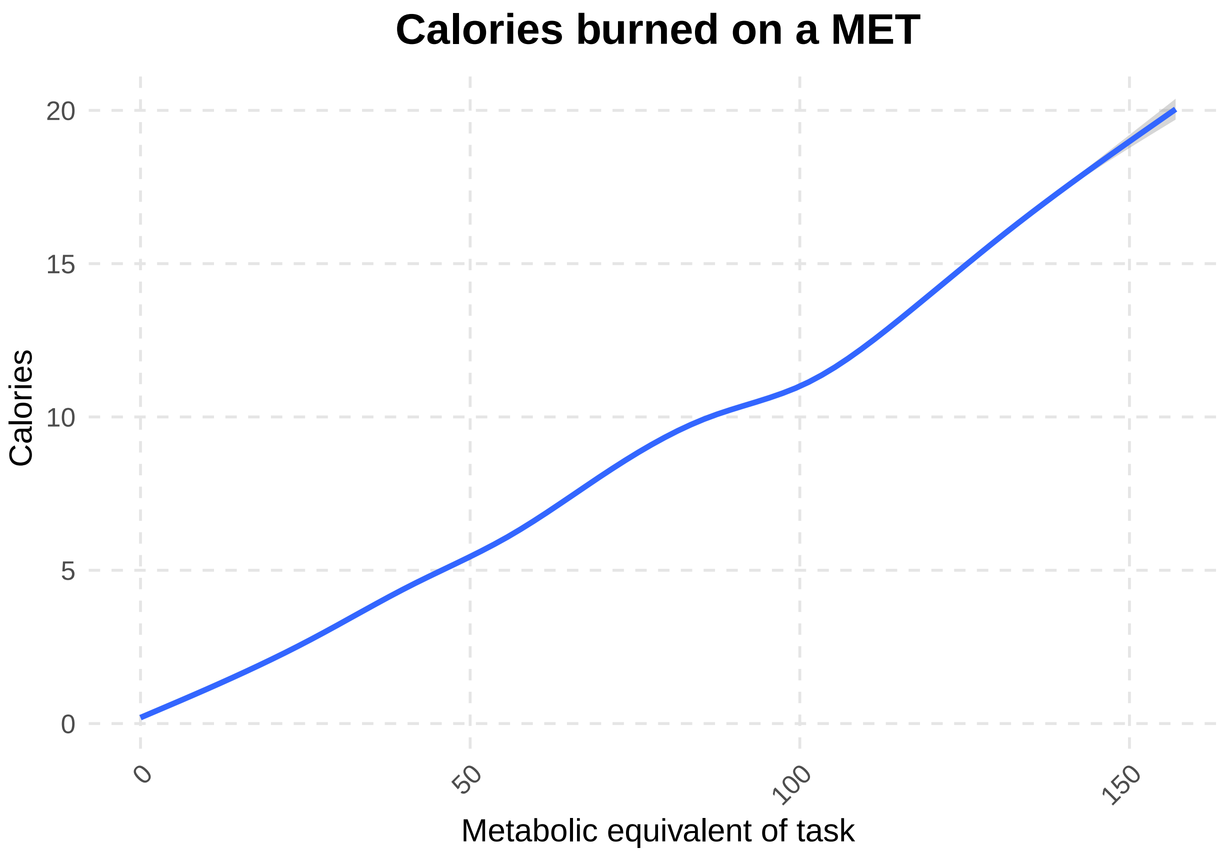
I recommend consistently checking saved measurement results. If these results appear abnormal, users should be prompted to visit a service point for device inspection.

Furthermore, if the measurements are calculated in an unconventional way, this must be explicitly stated in the device instructions. This ensures transparency and helps users interpret the data correctly.

However, it is a logical assumption that a higher MET value should correlate with a greater number of calories burned. Let’s investigate this relationship further.

```
# Plotting
ggplot(data = minute_ciMs) +
  geom_smooth(mapping = aes(x = me_ts, y = calories)) +
  labs(
    title = "Calories burned on a MET",
    x = "Metabolic equivalent of task",
    y = "Calories",
    color = "Metric"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    legend.position = "bottom",
    legend.title = element_blank(),
    legend.text = element_text(size = 11),
    panel.grid.major = element_line(color = "grey90", linetype = "dashed"),
    panel.grid.minor = element_blank()
  )
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula = ’y ~ s(x, bs = "cs")’
```



As expected calories are positively correlated with MET.

Summary:

Analysis of smart device data reveals several significant trends in user behavior related to physical activity and sleep. On average, users exhibit a low daily step count, averaging 7,638 steps, which falls short of the recommended 10,000 steps for optimal health benefits. The majority of their physical activity is categorized as light exertion, suggesting a lack of sustained, moderate-to-vigorous exercise. Peak activity periods are observed between 8 AM and 7 PM, with noticeable increases during lunchtime (12 PM - 2 PM) and in the evenings (5 PM - 7 PM), aligning with commutes and daily errands. The 5-7 PM window also appears to be a preferred time for more focused exercise.

A concerning anomaly in the data is an unusually high average Metabolic Equivalent of Task (MET) value of 15 METs per minute, which raises questions about the accuracy of the device's measurement methodology. This discrepancy needs to be addressed to ensure data integrity and user trust.

Regarding sleep patterns, users frequently experience sleep onset insomnia, taking around 30 minutes to fall asleep. Their sleep schedules are irregular, and there's a noticeable weekday sleep deprivation pattern, where users sleep less on Mondays and Tuesdays, compensating on Wednesdays, leading to overall inconsistent rest. Actual sleep time rarely exceeds 7.5 hours, and a significant difference exists between time spent in bed and actual sleep minutes.

These observations highlight user struggles with insufficient exercise, inefficient activity patterns, and poor sleep hygiene, all of which need correction to foster healthier lifestyles.

These trends offer valuable insights for enhancing Bellabeat's product offerings. To address the low activity levels, Bellabeat can introduce personalized step goals and exertion prompts to encourage users to increase their activity intensity. Real-time feedback advising users on tempo and stride, coupled with efficient exercise tips for boosting activity scores, would be highly beneficial. Implementing gamification features like pre-calculated goals and achievement badges could significantly motivate users.

For sleep improvement, features such as customizable sleep schedules, pre-sleep reminders (e.g., an hour before bedtime to avoid phone use), and relaxation techniques within the app would help users establish healthier sleep habits.

Crucially, Bellabeat must prioritize data accuracy and transparency. This includes consistent measurement validation, prompting users for device inspection if abnormal results persist, and clearly stating measurement methodologies in device instructions, especially concerning the high MET values.

Bellabeat can influence users' perception of daily activities by emphasising incremental progress and making higher exertion levels more accessible. Promoting holistic well-being by connecting activity and sleep to overall health will reinforce the value proposition. Integrating the device into daily life can be achieved through contextual reminders during peak activity times and transforming raw data into actionable, personalized recommendations.

These insights can refine Bellabeat's marketing approach. The strategy should focus on empowerment, positioning Bellabeat as a tool that helps users take control of their health through actionable insights. Marketing should highlight "Achievable Health," showcasing how Bellabeat helps users integrate healthy habits seamlessly into their lives, making health goals less daunting.

Promoting healthy behaviors should be done in a light and approachable way. Campaigns can leverage gamified challenges to highlight the fun and rewarding aspects of health goals.

Finally, the marketing strategy should focus on assisting users, positioning Bellabeat as a "Personal Health Coach" that provides guidance, reminders, and motivation. Emphasizing proactive features that go "Beyond Tracking" will differentiate Bellabeat in the market, and transparent communication regarding data accuracy will build trust.