# Bellabeat Case Study

Caleb

2022-09-19

## Bellabeat Case Study

Ask Prepare Process Analyze Share Act

### Process/Clean Data in Excel and in R

Started off cleaning data in excel. 1. reformating data 2. spliting time and dates into multiple columns 3. reviewing data for spaces and null entries 4. rename columns

### Install packages

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("tibble")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("stringr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("scales")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("ggrepel")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(janitor)
```

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

```
library(dplyr)
library(readr)
library(tibble)
library(stringr)
library(ggrepel)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

### uploaded clean spreadsheets to R studios

```
daily_activity_clean <- read.csv("Fitabase Data 4.12.16-5.12.16/daily_activity_clean.csv")
sleep_daily_clean <- read.csv("Fitabase Data 4.12.16-5.12.16/sleep_day_clean.csv")
hourly_int_clean <- read.csv("Fitabase Data 4.12.16-5.12.16/hourly_int_clean.csv")
sleep_day_clean <- read_csv("Fitabase Data 4.12.16-5.12.16/sleep_day_clean.csv")
```

```
## Rows: 413 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

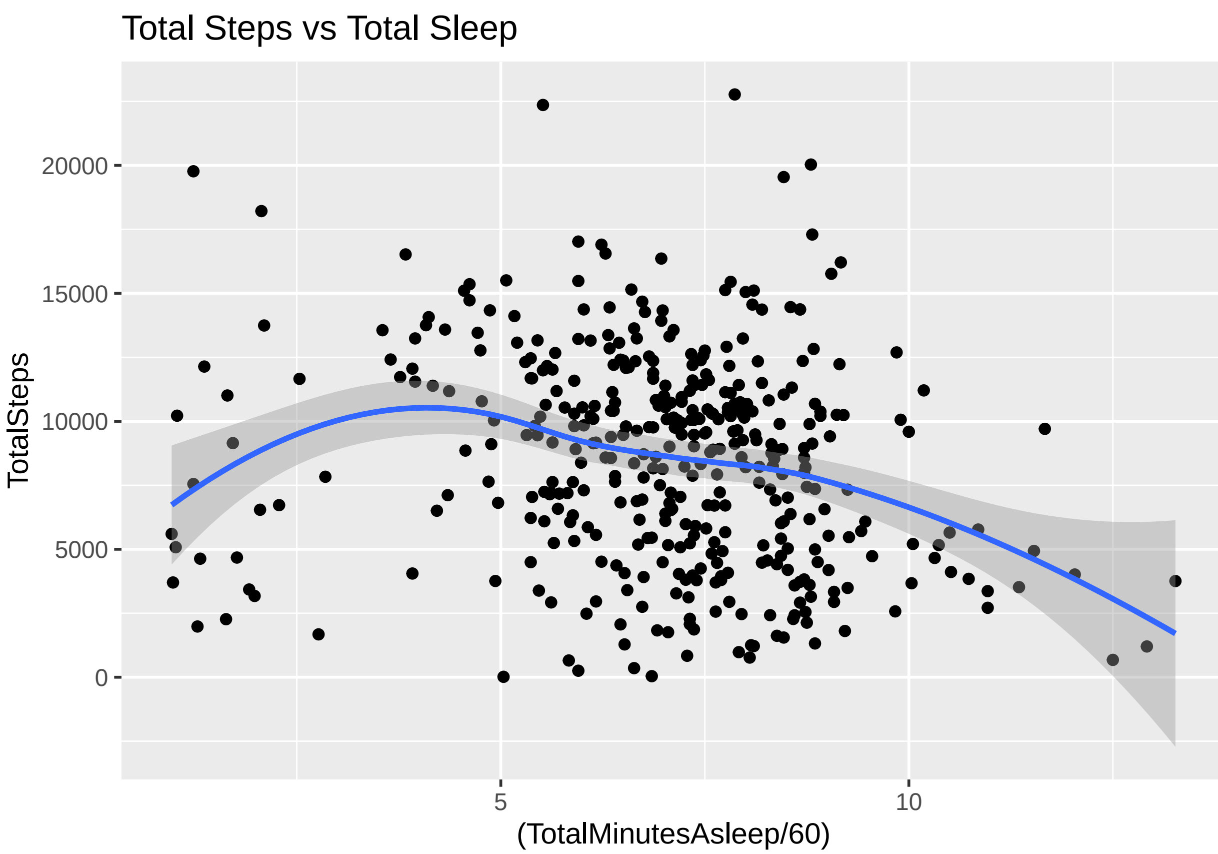
**Lets look at sleep compaired to total steps taken**

```
merged_data <- merge(daily_activity_clean, sleep_day_clean, by = c('Id','Date'))
```

## Analyze Data

```
ggplot(data = merged_data, aes(x=(TotalMinutesAsleep/60), y=TotalSteps))+
  geom_point()+
  geom_smooth()+
  labs(title = "Total Steps vs Total Sleep")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



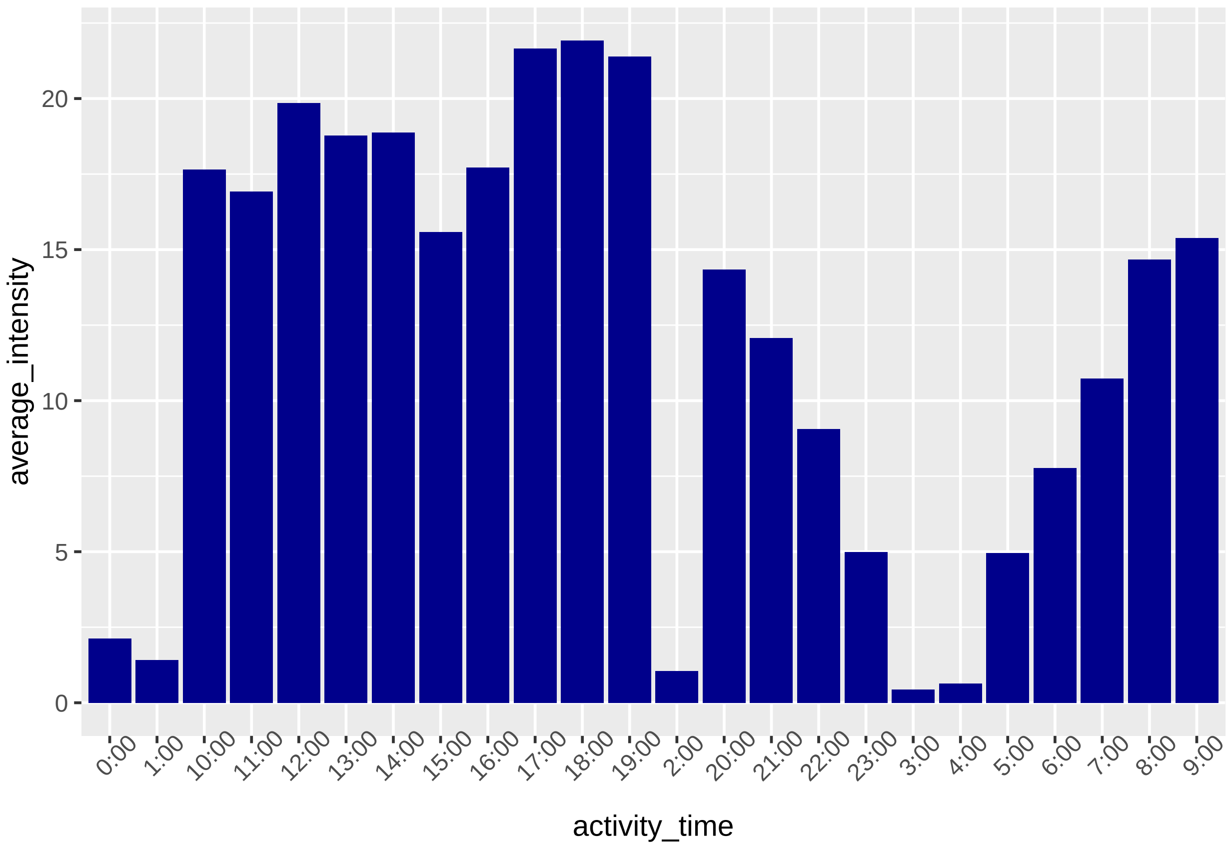
**This shows that participants that slept between 5-8 hours were more likly to achieve 10,000 steps.**

**worthy note: if you slept more then 10 hours you were less active than any other group**

```
int_new <- hourly_int_clean %>%
  group_by(activity_time) %>%
  drop_na() %>%
  summarise(average_intensity = mean(total_intensity))

view(int_new)
```

```
ggplot(data=int_new, aes(x=activity_time, y=average_intensity))+
  geom_histogram(stat="identity", fill= 'darkblue')+
  theme(axis.text.x = element_text(angle = 45))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



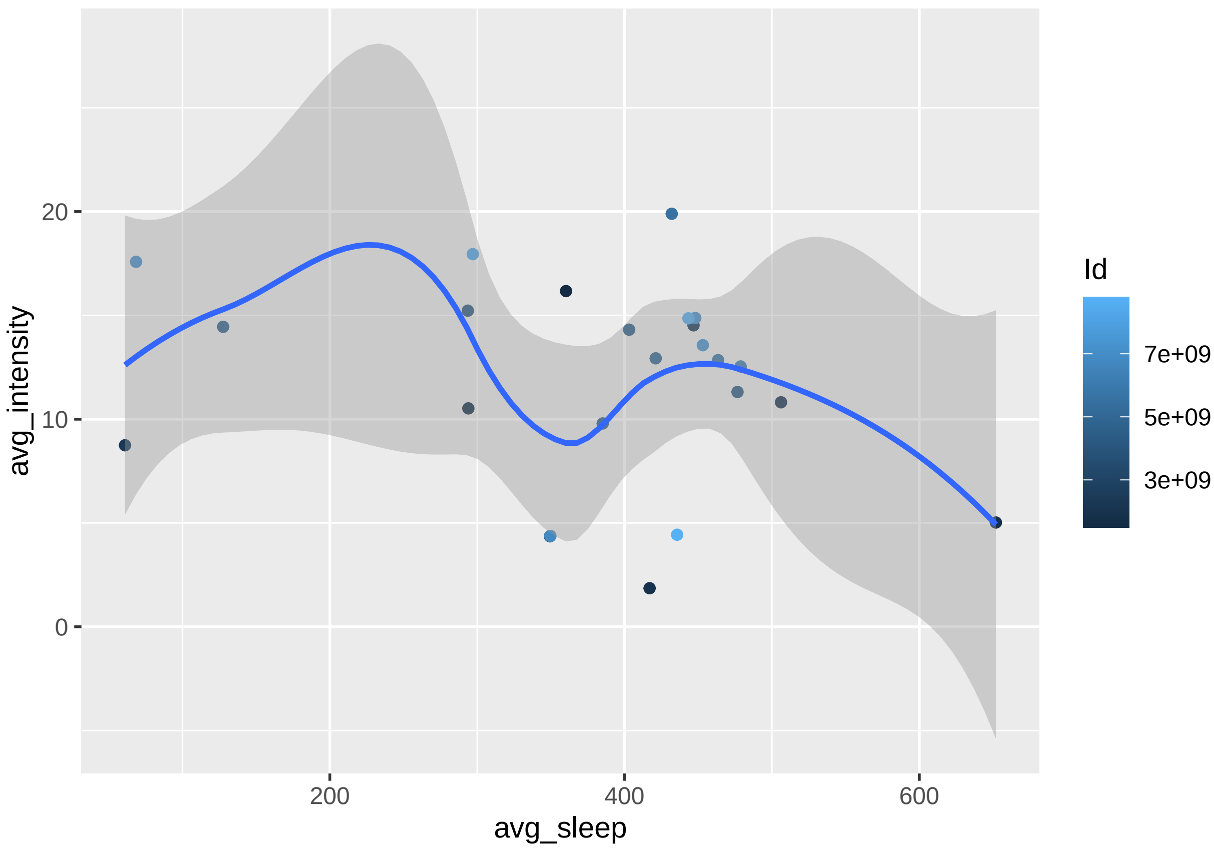

**for this we are looking at the average intensity and when participants are likely to conduct their workouts/activity during a given 24 hours.**

**Two major time frames to focus on lunch time and after work 5-8pm.**

```
int_id <- hourly_int_clean %>%
  group_by(Id) %>%
  drop_na() %>%
  summarise(avg_intensity = mean(total_intensity))
```

```
view(int_id)

sleep_id <- merged_data %>%
  group_by(Id) %>%
  drop_na() %>%
  summarise(avg_sleep = mean(TotalMinutesAsleep))

view(sleep_id)

sleep_int_merge <- merge(sleep_id, int_id, by=('Id'))

view(sleep_int_merge)

ggplot(data = sleep_int_merge, aes(x=avg_sleep,y=avg_intensity,color=Id))+
  geom_point()+
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

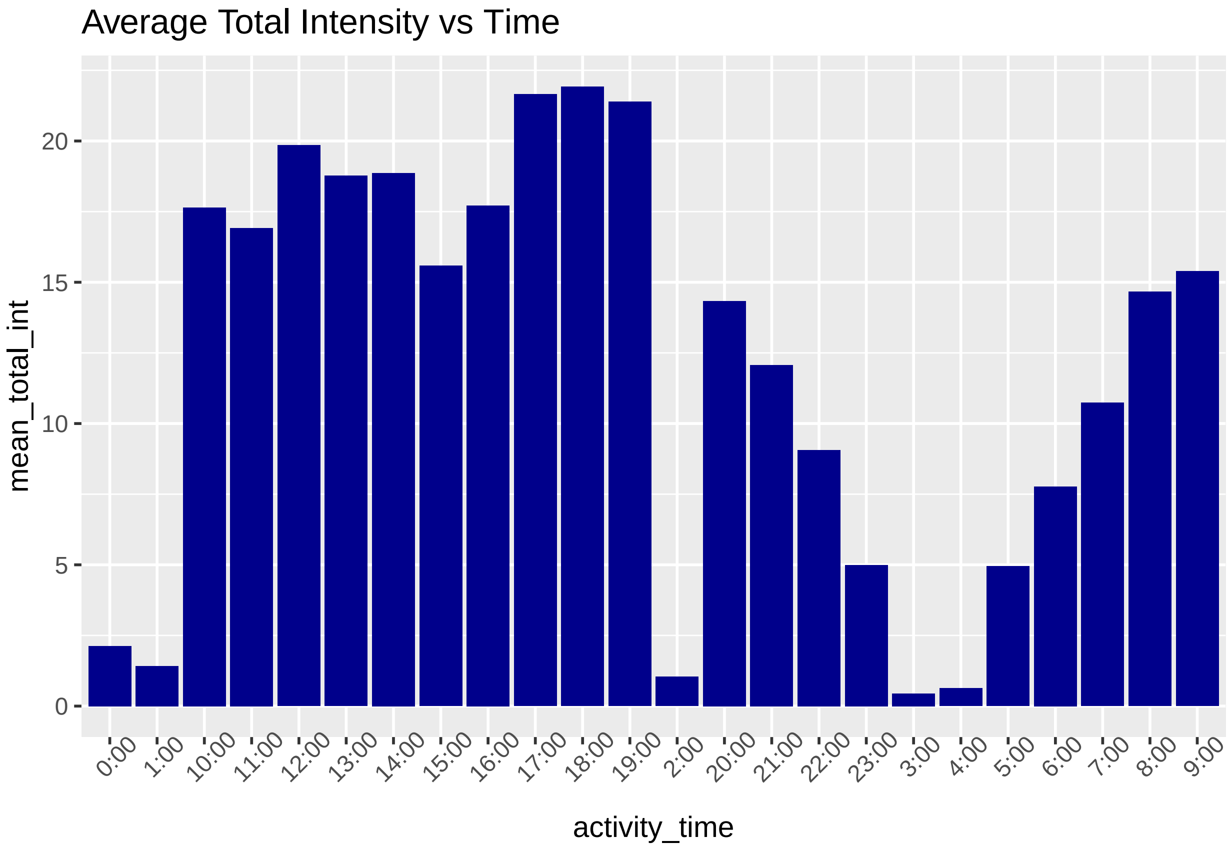



**even tho members that obtain 5-8 hours of sleep are more active member that only get 5-6 have more intense activity**

```
  hist_int_new <- hourly_int_clean %>%
  group_by(activity_time) %>%
  drop_na() %>%
  summarise(mean_total_int = mean(total_intensity))
```

```r
ggplot(data = hist_int_new, aes(x=activity_time, y=mean_total_int))+
  geom_histogram(stat = "identity",fill='darkblue')+
  theme(axis.text.x = element_text(angle = 45))+
  labs(title = "Average Total Intensity vs Time")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

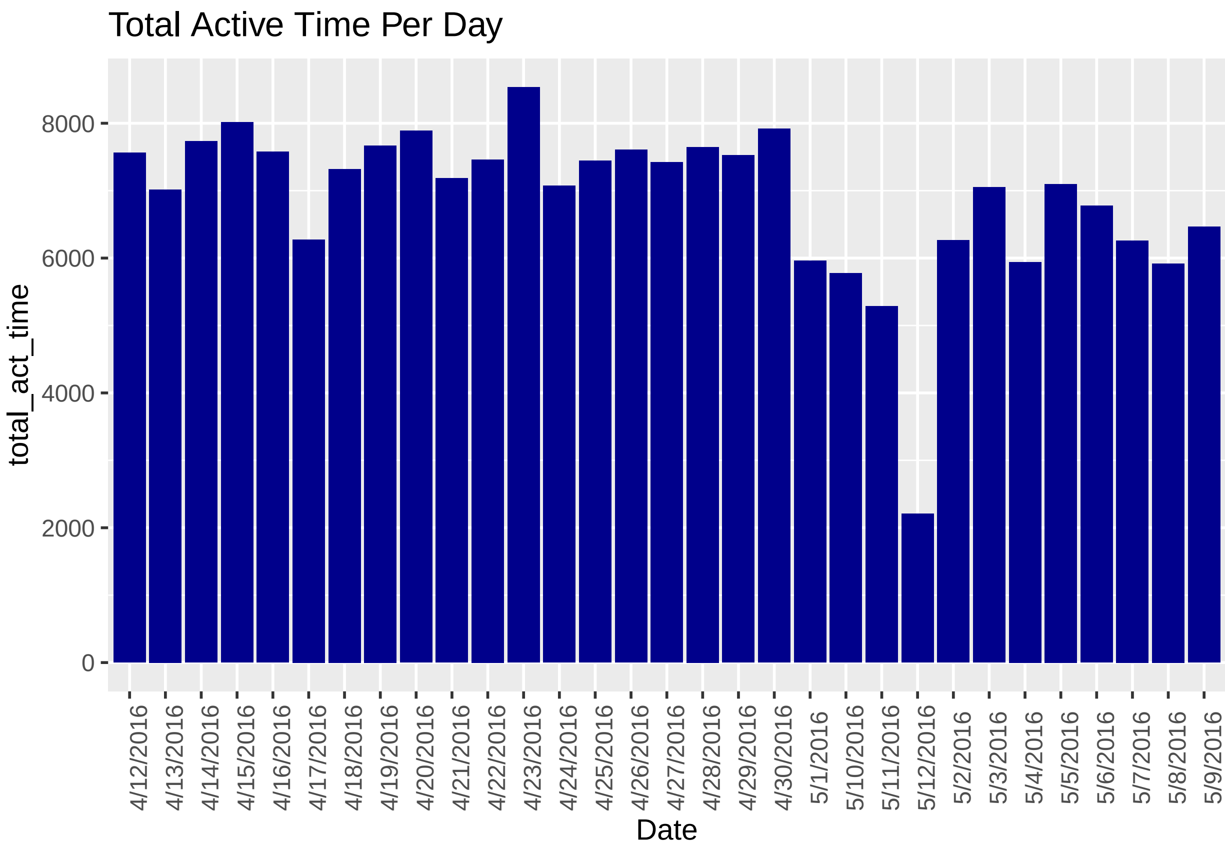



**Taking a look at the average total intensity over time of all participants to see if there is a specific time of day that participants are more likely to workout at.**

```r
  new_daily_activity <- daily_activity_clean %>%
  group_by(Date) %>%
  drop_na() %>%
  summarise(total_act_time = sum(VeryActiveMinutes,FairlyActiveMinutes,LightlyActiveMinutes),total_vam =
```

```r
ggplot(data = new_daily_activity, aes(x=Date, y=total_act_time))+
  geom_histogram(stat = "identity",fill='darkblue')+
  theme(axis.text.x = element_text(angle = 90))+
  labs(title = "Total Active Time Per Day")
```

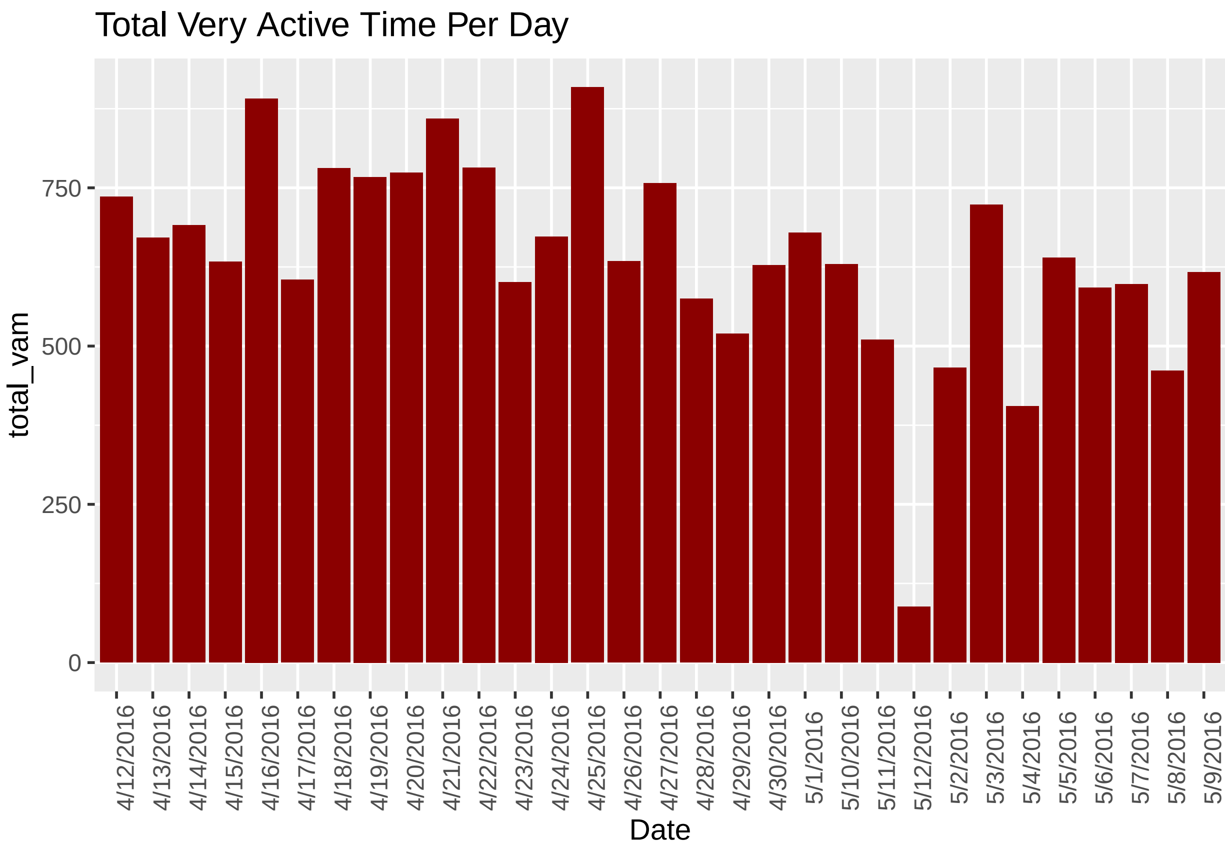
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




**Taking a look at the average total intensity per day of all participants to see if there is a specific day in the week that participants are more likely to workout at.**

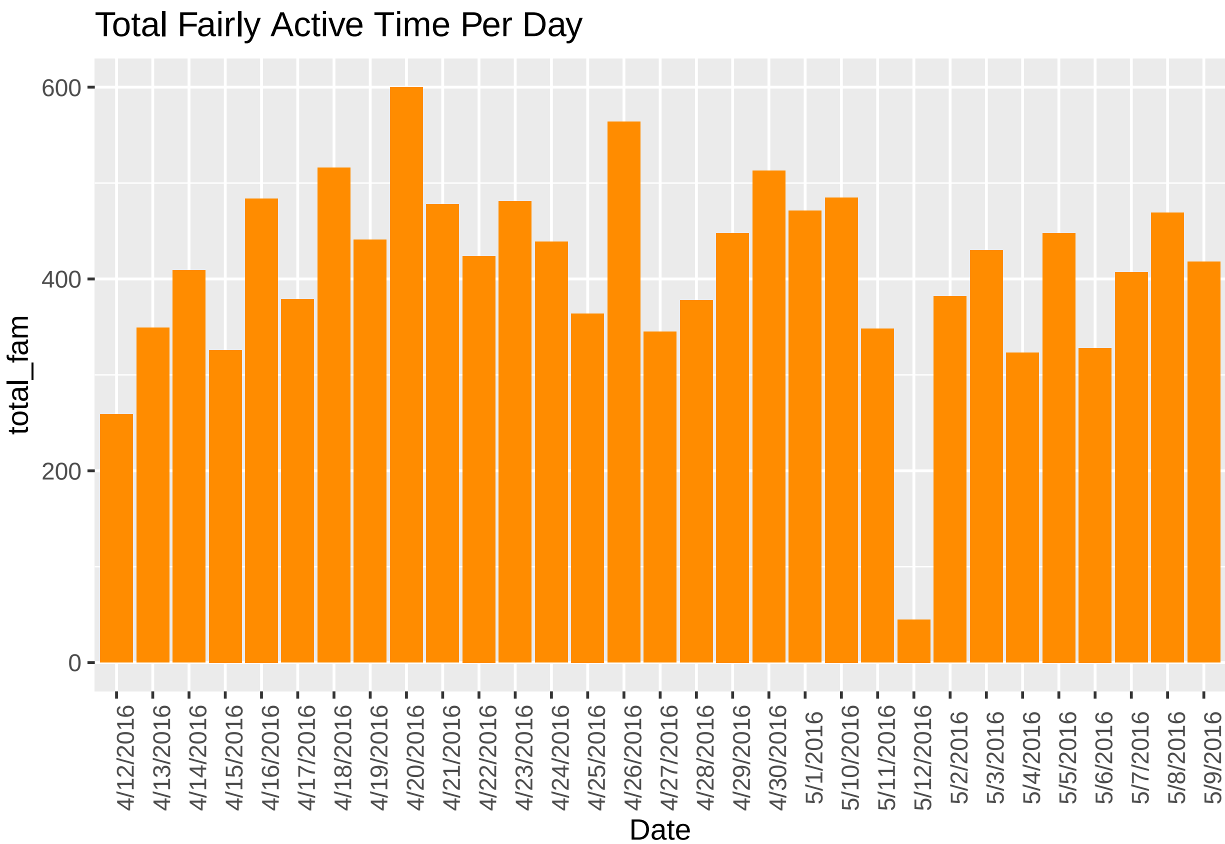
```
ggplot(data = new_daily_activity, aes(x=Date, y=total_vam))+
  geom_histogram(stat = "identity",fill='darkred')+
  theme(axis.text.x = element_text(angle = 90))+
  labs(title = "Total Very Active Time Per Day")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



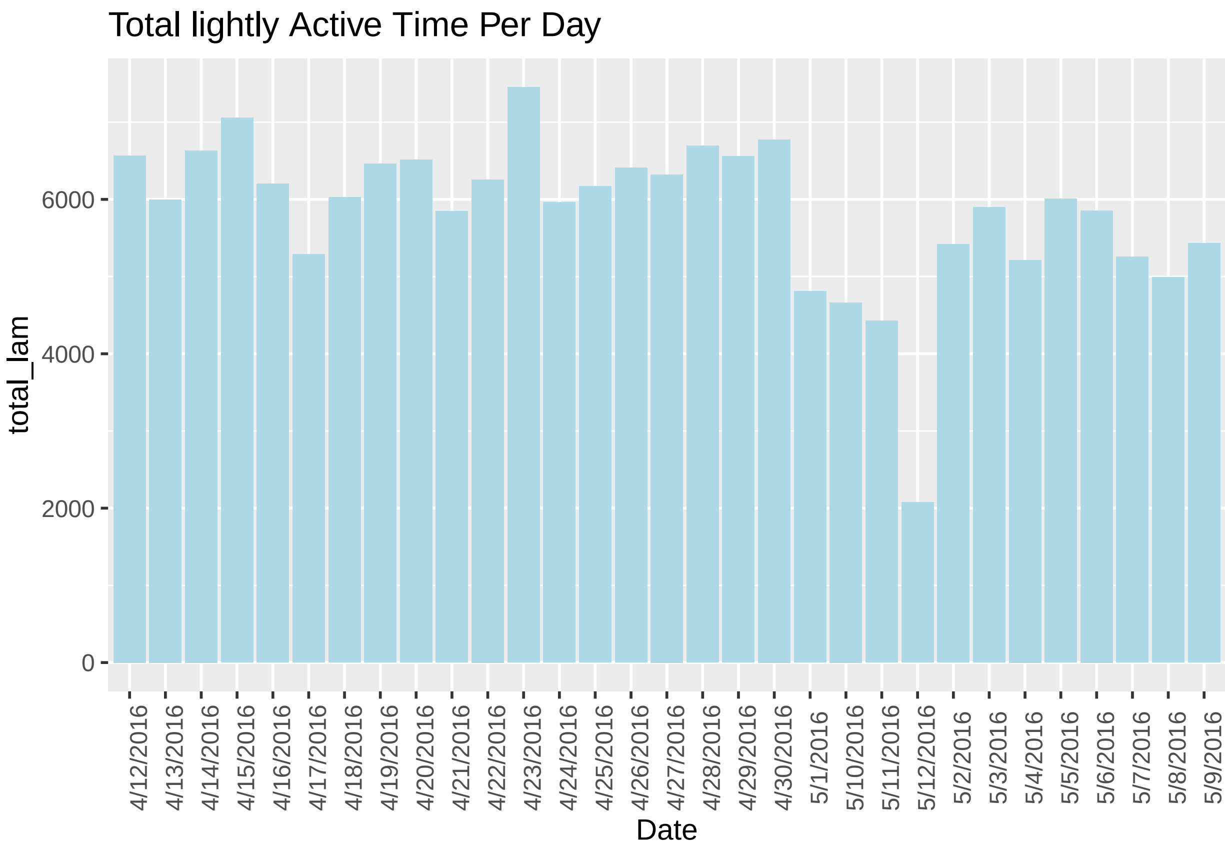

```
ggplot(data = new_daily_activity, aes(x=Date, y=total_fam))+
  geom_histogram(stat = "identity",fill='darkorange')+
  theme(axis.text.x = element_text(angle = 90))+
  labs(title = "Total Fairly Active Time Per Day")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
ggplot(data = new_daily_activity, aes(x=Date, y=total_lam))+
  geom_histogram(stat = "identity",fill='lightblue')+
  theme(axis.text.x = element_text(angle = 90))+
  labs(title = "Total lightly Active Time Per Day")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




**Based on the above data the majority of participants partake in lightly active activity. These are defined by regular daily activity and walks. followed by very active activity and then fairly active activity.**

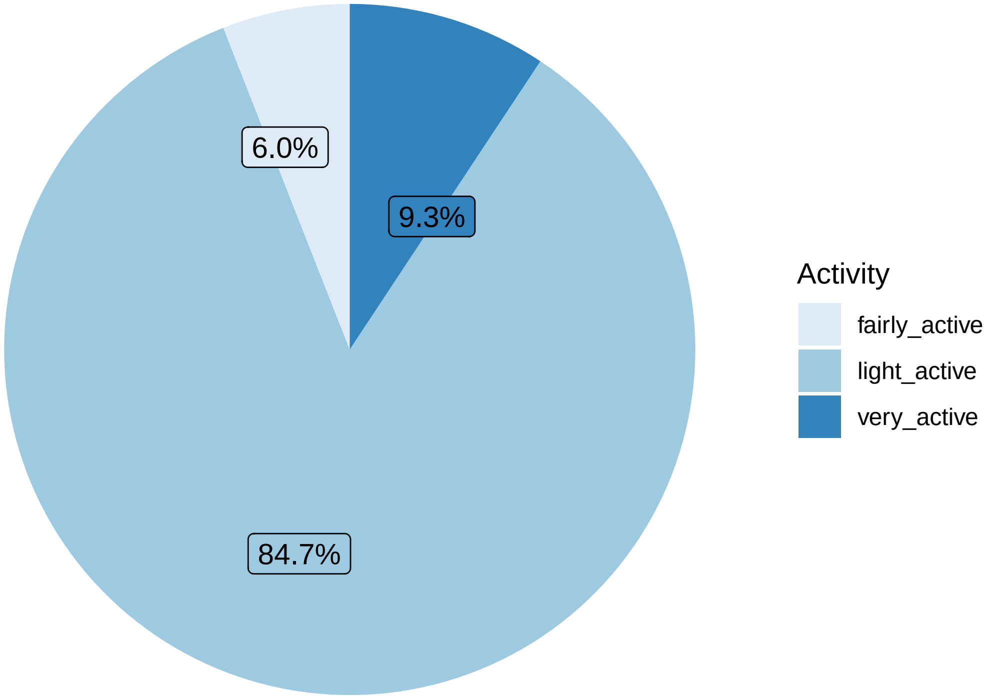
```
sum_activity <- daily_activity_clean%>%
  summarise(light_active = sum(LightlyActiveMinutes), fairly_active = sum(FairlyActiveMinutes),very_acti

view(sum_activity)
```

```
df <- data.frame(
     activity = c("light_active","fairly_active", "very_active"),
     value = c(181244, 12751, 19895))

view(df)
```

```
blank_theme <- theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    plot.title=element_text(size= 15, face="bold"))
```

```
ggplot(df, aes(x=1, y=value, fill=activity)) +
  geom_col() +
  scale_fill_brewer("Activity")+
```

```
  geom_label_repel(aes(label =percent(value/sum(value), size=5)), position = position_stack(vjust = 0.5
  coord_polar(theta = "y") +
  theme_void()
```




**above chart shows that the majority (excluding sedentary activity) is light activity followed by very active activity and close behind that is fairly active activity. It is a good thing to note that sedentary activity is the largest type of activity recorded and was excluded to show a better comparison for the other three types of activity.**

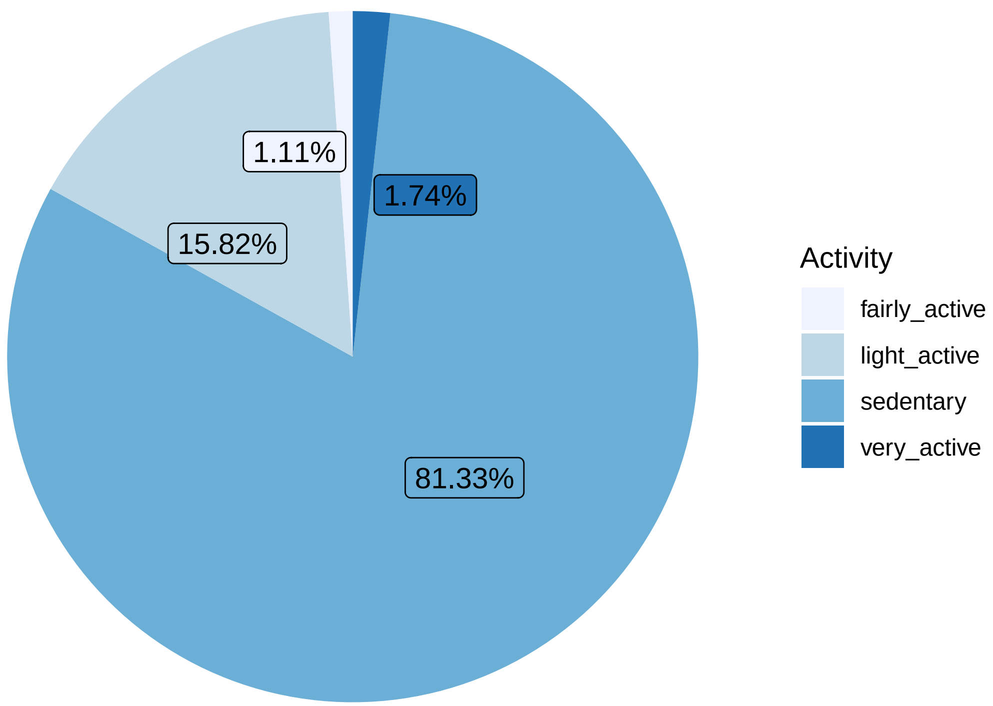
```
sum_activity2 <- daily_activity_clean%>%
  summarise(sedentary = sum(SedentaryMinutes), light_active = sum(LightlyActiveMinutes), fairly_active =

view(sum_activity2)
```

```
dfs <- data.frame(
  activity = c("sedentary", "light_active", "fairly_active", "very_active"),
  value = c(931738, 181244, 12751, 19895))
```

```
head(dfs)
```

```
##        activity  value
## 1     sedentary 931738
## 2  light_active 181244
## 3 fairly_active  12751
## 4   very_active  19895
```

```
ggplot(dfs, aes(x=1, y=value, fill=activity)) +
  geom_col() +
  scale_fill_brewer("Activity")+
  geom_label_repel(aes(y=value, label =percent(value/sum(value), size=5)), position = position_stack(vju
  coord_polar(theta = "y") +
  theme_void()
```


1.11%
1.74%
15.82%
81.33%
Activity
fairly_active
light_active
sedentary
very_active