Data Analysis for Bellabeat Project

Data Analyst: Jia Xi Wang

Report to: Urška Sršen , Sando Mur, Bellabeat Marketing Analytics Team

Purpose

The goal of this project is to study the performances of non-Bellabeat smart devices and in this project, FitBit Fitness Tracker Data will be accessed for analysis. After identifying the potential trends from the dataset, the project will study insights from this analysis to further inform Bellabeat marketing strategies. This project will focuse on applying r-programing to assess available dataset according to the five phases below:

- Average hourly steps in a day basis
- Average daily Steps in a weekly basis
- Average steps/day vs. Average Calories consumption/day
- Sleeping quality assessment for subjective group
- Sleeping quality vs. Activity intensity

And finally, the report will be summarized by recommending effective strategies that will do the best for brand growing opportunities.

Scope/Major ProjectActivities:

Activity	Description
Identifying the business task	The purpose of this project is described in the above section: PURPOSE.
Data Preparation/Collection	Dataset made available by Mobius.
Identify trends and relationships within the dataset	R-programming analysis will be involved.
Applying insight into one Bellabeat smart device and create recommendations	Generating Effective recommendations
Deliver final report	Deliver final report and recommendations to Urška Sršen , Sando Mur, Bellabeat Marketing Analytics Team.

Deliverables

Deliverable	description/Details
Project Scope Summary	A clear summary of the business task including project purpose/project scope/project deliverables/Major Milestones.
Final report	Including plots presentation and markdown appendix: A summary of non-Bellabeat device sage trend analysis including supporting visualizations and key findings (trends etc.) and recommendations for applying insights discovered into alleviating one Bellabeat smart device marketing strategies. A description of all data sources used and documentations of any cleaning or manipulation of data

Scheldue Overview/Major Milestone

Milestone	Expected Completion Date	Description
Project Scope Summary	2021-07-15	Review data sources and searching for supplement data source
Data review	2021-07-16	Initial data analysis complete
Data Analysis (Trend/insight discovery)	2021-07-24	Trend for non-Bellabeat smart device usage has been discovered
Visualization Created	2021-07-24	Visualizations created for the purpose of supporting presentation and recommendations
Recommendation’s list	2021-07-25	List of recommendation in improving marketing strategies within company
Final Report	2021-07-26	Final report detailing all work, analysis,

methologies and findings

Analysis

Installing Useful packages

```
library(tidyverse)
library(dplyr)
library(tidyr)
library(ggplot2)
library(lubridate)
library(hms)
library(ggrepel)
library(ggpubr)
```

Loading the dataset needed:

- FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius (https://www.kaggle.com/arashnic)

```
DailyActivity<-read.csv('dailyActivity_merged.csv')
DailySteps<-read.csv('dailySteps_merged.csv')
HourlyCalories<-read.csv('hourlyCalories_merged.csv')
HourlySteps<-read.csv('hourlySteps_merged.csv')
Sleep<- read.csv('sleepDay_merged.csv')
Weight<-read.csv('weightLoginfo_merged.csv')
```

Preview the dataset and corresponding variables

```
library(knitr)
kable(DailyActivity[1:5, ],caption='DailyActivity')
```

DailyActivity

Id	ActivityDate	TotalSteps	TotalDistance	TrackerDistance	LoggedActivitiesDistance	VeryActiveDistance	ModeratelyActiveDistance	LightA
1503960366	4/12/2016	13162	8.50	8.50	0	1.88	0.55	
1503960366	4/13/2016	10735	6.97	6.97	0	1.57	0.69	
1503960366	4/14/2016	10460	6.74	6.74	0	2.44	0.40	
1503960366	4/15/2016	9762	6.28	6.28	0	2.14	1.26	
1503960366	4/16/2016	12669	8.16	8.16	0	2.71	0.41	

```
kable(DailySteps[1:5, ],caption='DailySteps')
```

DailySteps

Id	ActivityDay	StepTotal
1503960366	4/12/2016	13162
1503960366	4/13/2016	10735
1503960366	4/14/2016	10460
1503960366	4/15/2016	9762
1503960366	4/16/2016	12669

```
kable(HourlyCalories[1:5, ],caption='HourlyCalories')
```

HourlyCalories

Id	ActivityHour	Calories
1503960366	4/12/2016 12:00:00 AM	81
1503960366	4/12/2016 1:00:00 AM	61
1503960366	4/12/2016 2:00:00 AM	59
1503960366	4/12/2016 3:00:00 AM	47
1503960366	4/12/2016 4:00:00 AM	48

```
kable(HourlySteps[1:5, ],caption='HourlySteps')
```

HourlySteps

Id	ActivityHour	StepTotal
1503960366	4/12/2016 12:00:00 AM	373
1503960366	4/12/2016 1:00:00 AM	160
1503960366	4/12/2016 2:00:00 AM	151
1503960366	4/12/2016 3:00:00 AM	0
1503960366	4/12/2016 4:00:00 AM	0

```
kable(Sleep[1:5, ],caption='Sleep')
```

Sleep

Id	SleepDay	TotalSleepRecords	TotalMinutesAsleep	TotalTimeInBed
1503960366	4/12/2016 12:00:00 AM	1	327	346
1503960366	4/13/2016 12:00:00 AM	2	384	407
1503960366	4/15/2016 12:00:00 AM	1	412	442
1503960366	4/16/2016 12:00:00 AM	2	340	367
1503960366	4/17/2016 12:00:00 AM	1	700	712

```
kable(Weight[1:5, ],caption = 'Weight')
```

Weight

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

Analyzing average hourly steps in a day basis

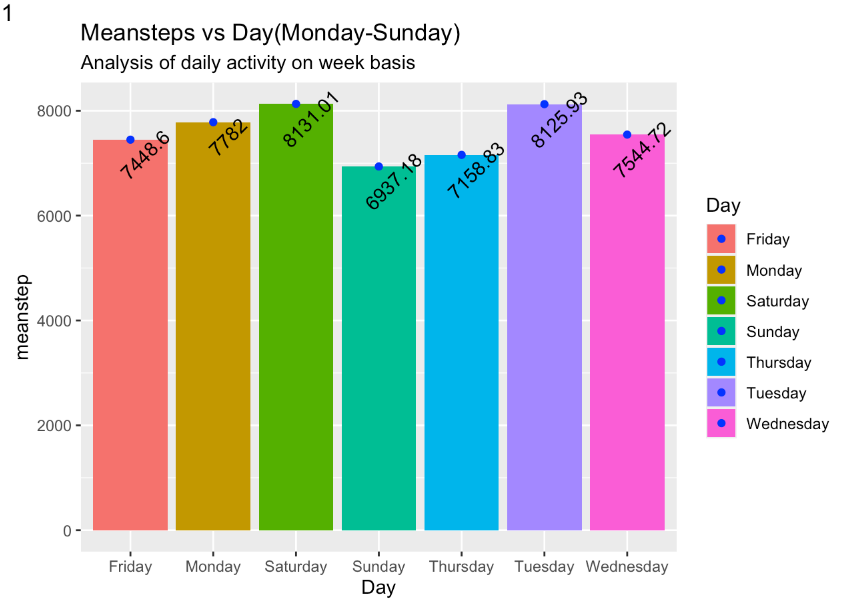
- The average steps generating from the dataset is varied from 6900 to 8200 times amongMonday to Friday
- As we can observe from the figure1 below, the highest step countings occurs on Saturday

```
DailySteps<-DailySteps%>%
            group_by(ActivityDay)%>%
            summarise(meanstep=mean(StepTotal),n=n())%>%
            mutate(Day=weekdays(mdy(ActivityDay)))
##this is the dataset after first groupby

DailySteps<-DailySteps%>%
            group_by(Day)%>%
            summarise(meanstep=mean(meanstep))
DailySteps$meanstep<-round(DailySteps$meanstep,2)
##this is the dataset after second groupby

ggplot(DailySteps,aes(x=Day,y=meanstep,fill=Day))+
  geom_bar(stat="identity")+
  geom_point(stat="identity",color='blue')+
  geom_text(aes(label=meanstep,),vjust=2,angle=45)+
  labs(title='Meansteps vs Day(Monday-Sunday)', subtitle='Analysis of daily activity on week basis',caption = "Da
ta source:CC0:Public Domain, dataset made available through Mpbius",tag='Fig.1')+
  theme(plot.tag.position='topleft',plot.caption.position = 'plot')
```

Fig.1



Meansteps vs Day(Monday-Sunday)

Analysis of daily activity on week basis

Data source:CC0:Public Domain, dataset made available through Mpbius

```
kable(DailySteps[1:7, ],caption = 'Table 1: Daily Steps Summary for a Week Basis')
```

Table 1: Daily Steps Summary for a Week Basis

Day	meanstep
Friday	7448.60
Monday	7782.00
Saturday	8131.01
Sunday	6937.18
Thursday	7158.83
Tuesday	8125.93
Wednesday	7544.72

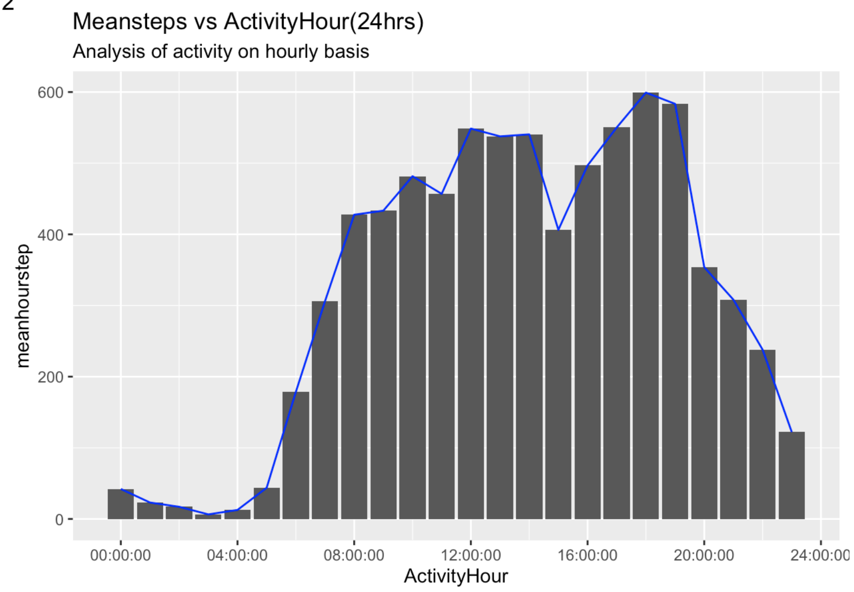
Average daily Steps in a weekly basis

- From the observation in Figure2, the highest step counting occurs around 18:00 which is approximately 600 steps.
- Most of the activity are participated in the afternoon since there is a dramatic contast between the step performance between the morning period and the afternoon period.
- The trend is consistent with human activity.

```
HourlySteps$ActivityHour<-mdy_hms(HourlySteps$ActivityHour)
HourlySteps$ActivityHour<-as_hms(HourlySteps$ActivityHour)
HourlySteps<-HourlySteps%>%
  group_by(ActivityHour)%>%
  summarise(meanhourstep=mean(StepTotal))
HourlySteps$meanhourstep=round(HourlySteps$meanhourstep,2)

ggplot(HourlySteps,aes(x=ActivityHour,y=meanhourstep))+
  geom_bar(stat="identity")+
  geom_line(stat="identity",color='blue')+
  labs(title='Meansteps vs ActivityHour(24hrs)', subtitle='Analysis of activity on hourly basis',caption = "Data
source:CC0:Public Domain, dataset made available through Mpbius", tag='Fig.2')+
  theme(plot.tag.position='topleft',plot.caption.position = 'plot')
```

Fig.2



Data source:CC0:Public Domain, dataset made available through Mpbius

```
kable(HourlySteps[1:7, ],caption = 'Table 2:Average Steps Brief for a Daily
      Basis (Other records can be accessed through dataset HourlySteps')
```

Table 2:Average Steps Brief for a Daily Basis (Other records can be accessed through dataset HourlySteps

ActivityHour	meanhourstep
00:00:00	42.19
01:00:00	23.10
02:00:00	17.11
03:00:00	6.43
04:00:00	12.70
05:00:00	43.87
06:00:00	178.51

Average steps/day vs. Average Calories consumption/day

- In general, the average calories is increasingly assumpted as the number of steps increase. Take 16:00as an example, the steps counting has a sharp rop and the calories value drops as well.
- *NOTE:*The reason that two lines does not show with a similar curvature and the mean hour calories trend has a relatively small slope shown on the graph, it is because of the y-value range on the figure.

```
HourlyCalories$ActivityHour<-mdy_hms(HourlyCalories$ActivityHour)
HourlyCalories$ActivityHour<-as_hms(HourlyCalories$ActivityHour)
HourlyCalories<-HourlyCalories%>%
  group_by(ActivityHour)%>%
  summarise(meanhourcalories=mean(Calories))
HourlyCalories$meanhourcalories<-round(HourlyCalories$meanhourcalories,2)

StepCalories=inner_join(HourlySteps,HourlyCalories,by='ActivityHour')
kable(StepCalories[1:7, ],caption = 'Table 3: A Brief Preview of the Step vs. CaloriesData')
```

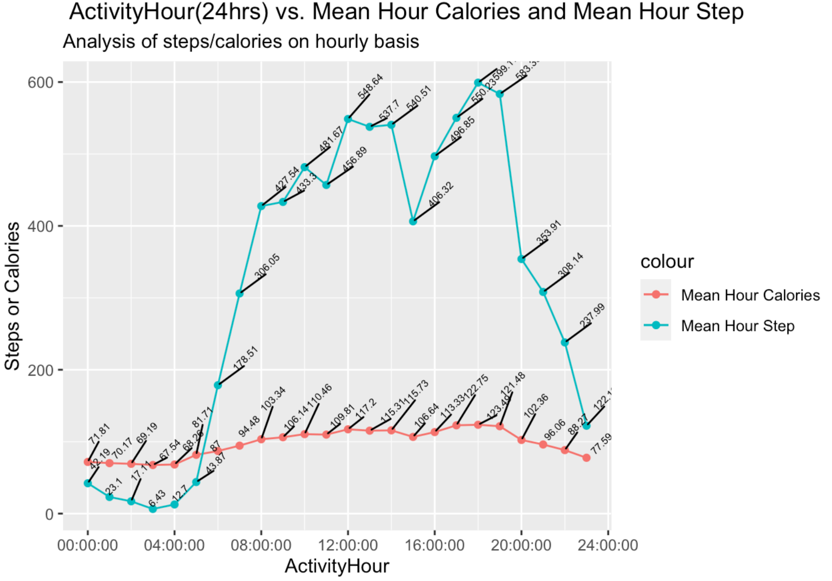
Table 3: A Brief Preview of the Step vs. CaloriesData

ActivityHour	meanhourstep	meanhourcalories
00:00:00	42.19	71.81
01:00:00	23.10	70.17
02:00:00	17.11	69.19
03:00:00	6.43	67.54

04:00:00	12.70	68.26
05:00:00	43.87	81.71
06:00:00	178.51	87.00

```
ggplot(StepCalories,aes(x=ActivityHour))+
  geom_point(mapping=aes(y=meanhourcalories,color='Mean Hour Calories'))+
  geom_line(mapping=aes(y=meanhourcalories,color='Mean Hour Calories'))+
  geom_point(mapping=aes(y=meanhourstep,color='Mean Hour Step'))+
  geom_line(mapping=aes(y=meanhourstep,color='Mean Hour Step'))+
  geom_text_repel(aes(label=meanhourcalories,y=meanhourcalories),hjust=-0.5,
  angle=45,size=2)+
  geom_text_repel(aes(label=meanhourstep,y=meanhourstep),hjust=-0.5,angle=45,
  size=2)+
  labs(x='ActivityHour',y='Steps or Calories',title=' ActivityHour(24hrs) vs. Mean Hour Calories and Mean Hour St
ep', subtitle='Analysis of steps/calories on hourly basis',caption = "Data source:CC0:Public Domain, dataset made
available through Mpbius", tag='Fig.3')+
  theme(plot.tag.position='topleft',plot.caption.position = 'plot')
```

Fig.3



ActivityHour(24hrs) vs. Mean Hour Calories and Mean Hour Step

Analysis of steps/calories on hourly basis

Data source:CC0:Public Domain, dataset made available through Mpbius

Sleeping quality assessment for subjective group

- The sleeping quality is categorized into 3 groups:
  - average sleeping time/ total time in bed >90: Good
  - average sleeping time/ total time in bed >70 and <90: Soso
  - average sleeping time/ total time in bed <70: Bad
- Majority of the subjects (83.3%) has a good sleeping quality. And 8.3% of subjects need to improve their sleeping quality

```
Sleep<-Sleep%>%
      group_by(Id)%>%
      summarise(Asleep=mean(TotalMinutesAsleep),Bed=mean(TotalTimeInBed))
Sleep$Asleep=round(Sleep$Asleep,2)
Sleep$Bed=round(Sleep$Bed,2)
Sleep<-mutate(Sleep,percentage=Asleep/Bed*100)
Sleep<-Sleep%>%mutate(Status=case_when(percentage>90~'good',percentage<70~'bad',
                                       TRUE~'soso'))
Sleep2<-Sleep%>%count(Status)
kable(Sleep2[1:3, ],caption="Table 4:Preview of groups of Sleeping Quality for
                              Following Datafram Construction")
```

Table 4:Preview of groups of Sleeping Quality for Following Datafram Construction

Status	n
bad	2
good	20

soso	2

```
Status<-data.frame(group=c("Good","Soso","Bad"),distribution=c((20/24)*100,
                                                               (2/24)*100,(2/24)
                                                               *100))
kable(Status[1:3, ],caption="Table 5:Preview of groups of Sleeping Quality")
```

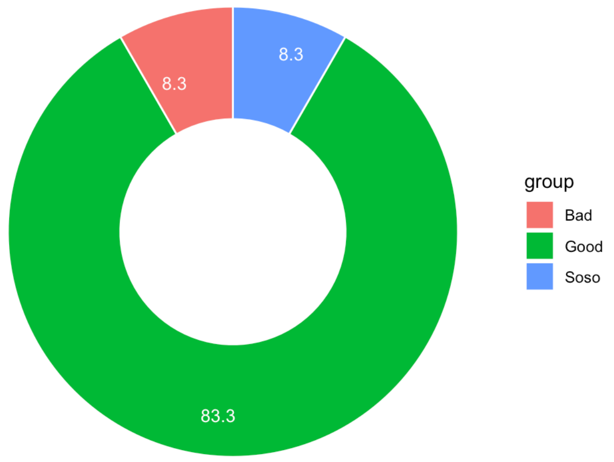
Table 5:Preview of groups of Sleeping Quality

group	distribution
Good	83.333333
Soso	8.333333
Bad	8.333333

```
Status<-Status%>%arrange(desc(group))%>%mutate(yposition=cumsum(distribution)-
                                                 0.5*distribution)
ggplot(Status,aes(x=2,y=distribution,fill=group))+
  geom_bar(stat='identity',width=1,color="white")+
  coord_polar('y',start=0)+
  theme_void()+
  geom_text_repel(aes(label=round(distribution,1),y=yposition),color='white',
                  size=3.55)+xlim(0.5,2.5)+
  labs(title=' Sleep Quality for the Subjective Group',caption = "Datasource:CC0:Public Domain, dataset made avai
lable through Mpbius",tag='Fig.4')+
  theme(plot.tag.position='topleft',plot.caption.position = 'plot')
```

Fig.4

Sleep Quality for the Subjective Group



Datasource:CC0:Public Domain, dataset made available through Mpbius

Seeping Quality vs. Activity intensity

- Figure 5 is the combination of 4 plots between sleeping time and various activity intensities
  - plot A: the effect of taking very intensive activities on sleeping quality
  - plot B: the effect of taking fairly intensive activities on sleeping quality
  - plot C: the effect of taking lightly intensive activities on sleeping quality
  - plot A: the effect of taking sedentary intensive activities on sleeping quality
- As the results generated from figure 5, and having a macro-observation:
  - As long as having very(observed after 60 min)/moderately (observed after 40 min)/lightly activities (observed after 100 min), the sleeping time will decrease in general.
  - For people who have sedentary active intensities, they have relative good sleeping quality and sleeping time drops only after an extremely long sedentary time.

```
DailyActivity<-DailyActivity%>%select(Id,
                                      VeryActiveMinutes,
                                      FairlyActiveMinutes,
                                      LightlyActiveMinutes,
                                      SedentaryMinutes)%>%
               group_by(Id)%>%
               summarise(meanV=mean(VeryActiveMinutes),
                      meanF=mean(FairlyActiveMinutes),
                      meanL=mean(LightlyActiveMinutes),
                      meanS=mean(SedentaryMinutes))

SleepActivity<-inner_join(DailyActivity,Sleep,by='Id')
kable(SleepActivity[1:7, ],caption="Table 6:Preview of Sleeping Quality vs.Activity Intensity")
```

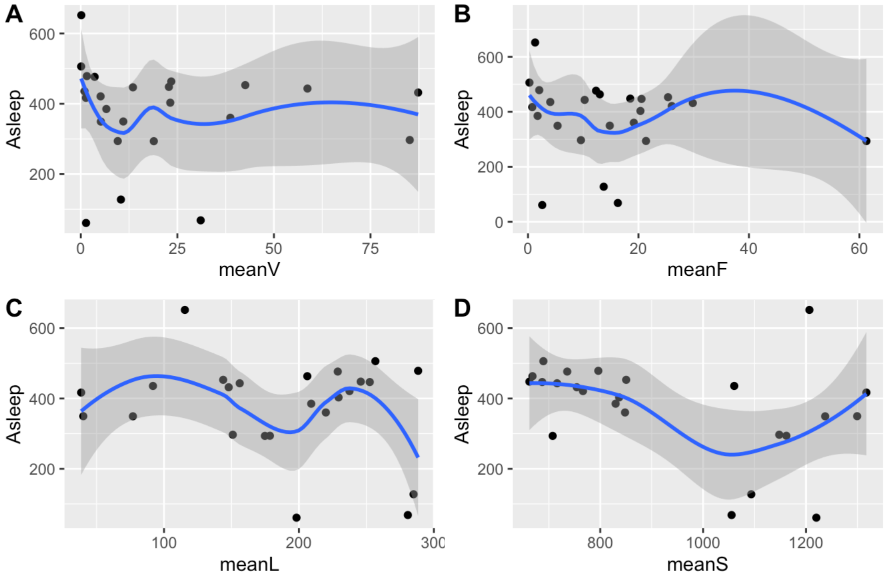
Table 6:Preview of Sleeping Quality vs.Activity Intensity

Id	meanV	meanF	meanL	meanS	Asleep	Bed	percentage	Status
1503960366	38.7096774	19.1612903	219.93548	848.1613	360.28	383.20	94.01879	good
1644430081	9.5666667	21.3666667	178.46667	1161.8667	294.00	346.00	84.97110	soso
1844505072	0.1290323	1.2903226	115.45161	1206.6129	652.00	961.00	67.84599	bad
1927972279	1.3225806	0.7741935	38.58065	1317.4194	417.00	437.80	95.24897	good
2026352035	0.0967742	0.2580645	256.64516	689.4194	506.18	537.64	94.14850	good
2320127002	1.3548387	2.5806452	198.19355	1220.0968	61.00	69.00	88.40580	soso
2347167796	13.5000000	20.5555556	252.50000	687.1667	446.80	491.33	90.93684	good

```
A<-ggplot(SleepActivity,aes(x=meanV,y=Asleep))+
  geom_point()+
  geom_smooth()
B<-ggplot(SleepActivity,aes(x=meanF,y=Asleep))+
  geom_point()+
  geom_smooth()
C<-ggplot(SleepActivity,aes(x=meanL,y=Asleep))+
  geom_point()+
  geom_smooth()
D<-ggplot(SleepActivity,aes(x=meanS,y=Asleep))+
  geom_point()+
  geom_smooth()

figure<-ggarrange(A,B,C,D,labels=c("A","B","C","D"),ncol=2,nrow=2)
annotate_figure(figure,
                top=text_grob("Plots between intensity(minutes) vs.Average Sleep Time",just='center',hjust=0.5,co
lor='black',size=12),
                bottom=text_grob("Data source:CC0:Public Domain, dataset made available through Mpbius",just =NUL
L,hjust=0.1,size=8),
                fig.lab="Fig.5")
```

Fig.5 Plots between intensity(minutes) vs.Average Sleep Time



Data source:CC0:Public Domain, dataset made available through Mpbius

Recommendation List

1. People usully have low level of activity on Sundays (averagely taking 6937 steps), and Bellabeat devices can track customer's activity behavious and send notifications to report customers' performances on a daily/weekly basis. Moreover,encouraging them to participate in sports on those day with lower active intensities.
2. Implementing step vs. calories plot on the tracking device. Providing suggestion when customer have very intensive activities, which will affect their sleeping quality.
3. Tracking the sleeping time over a period of time and generating reports for customers. Encouraging them to make modifications on their fitness plans.