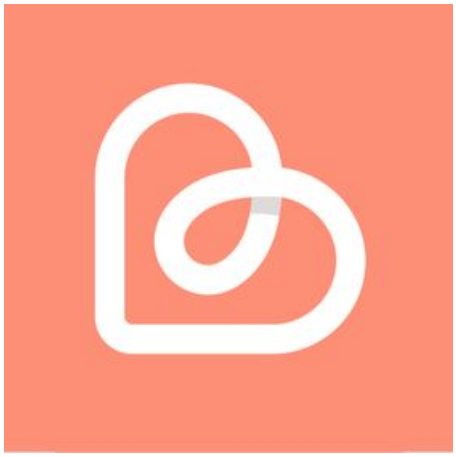
Bellabeat Case Study

Love Kumar

09/05/2021



Introduction

In this case study, I will perform real world analyses for Bellabeat, a high-tech manufacturer of health-focused smart devices for women. In order to answer the key business questions of the company I would follow the steps of data analysis process i.e Ask, Prepare, Process, Analyze, Share and Act.

Key StakeHolders:

- Urška Sršen: Bellabeat's cofounder and Chief Creative Officer
- Sando Mur: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team

Business task

- To Analyze non Bellabeat (FitBit) smart devices data to gain insights and trends about user's usage of the smart devices
- To give high level recommendations for marketing strategy based on the analyses done.

Data Preparation

I was provided with a dataset of Thirty FitBit users which is publically available at https://www.kaggle.com/arashnic/fitbit

Dataset consists of 18 csv files about the activities of users like Distance covered , Total Steps , Sleep , Active Minutes . This data was collected in the period of 03/12/2016 - 05/12/2016 .

I started with " dailyActivity_merged.csv " file .

Lets have a look at the dataset and different attributes in this file using R.

```
fit_data<-read.csv("dailyActivity_merged.csv")
colnames(fit_data)
```

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

```
str(fit_data)
```

```
## 'data.frame':    940 obs. of  15 variables:
##  $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ TotalSteps               : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...
##  $ TotalDistance            : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance          : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance      : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes      : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes     : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes         : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ Calories                 : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

```
head(fit_data)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              1.88                     0.55
## 2                        0              1.57                     0.69
## 3                        0              2.44                     0.40
## 4                        0              2.14                     1.26
## 5                        0              2.71                     0.41
## 6                        0              3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

Data Cleaning and Manipulation

For data cleaning and manipulation I use spreadsheet software (Google sheets)

data cleaning procedure

- Changed the format of specific columns to have only 2 decimal points in observations
- Filtered the data to find rows with all zeros observations.
- Deleted the rows with all zeros obsevations .
- Deleted the blanks rows .

data manipulation

There are atrributes named TotalDistance,LightActiveDistance,ModeratelyActiveDistance and VeryActiveDistance

Id

D	E	F	G	H
TotalDistance	LoggedActivitie	VeryActiveDistance	ModeratelyActiveDistance	LightActiveDistance
8.50	0	1.88	0.55	6.06
6.97	0	1.57	0.69	4.71
6.74	0	2.44	0.40	3.91
6.28	0	2.14	1.26	2.83
8.16	0	2.71	0.41	5.04
6.48	0	3.19	0.78	2.51
8.59	0	3.25	0.64	4.71
9.88	0	3.53	1.32	5.03
6.6				4.24
6.3				4.65
8.1				2.24
9.0				5.36
6.4				3.28
9.80	0	5.29	0.57	3.94
8.79	0	2.33	0.92	5.54

TotalDistance=(VeryActiveDistance) + (ModeratelyActiveDistance)+ (LightActiveDistance)

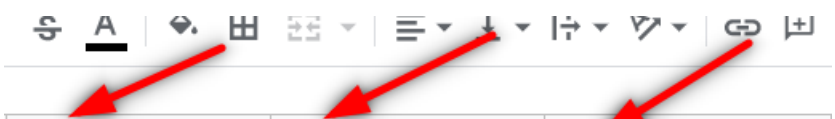
I Created Calculated Fields that contains percentage values of LightActiveDistance , ModeratelyActiveDistance And VeryActiveDistance over the TotalDistance covered

N	O	P
PercentLightDistance	PercentModerateDistance	PercentVeryActiveDistance
58.01	5.93	36.20
45.06	20.06	34.08
61.76	5.02	33.21
38.73	12.04	49.23
54.83	7.45	37.83
50.91	13.36	35.73
63.47	7.19	29.34
		21.14
		58.55
		31.08
		45.55
		53.98
63.03	10.47	26.51
44.31	3.36	52.42
44.43	13.60	41.50
78.04	6.99	14.83
46.16	15.35	38.49

PercentLightDistance = (LightlyActiveDistance/ TotalDistance) * 100

Calculated percentages

I calculated Total Active Minutes by adding the columns named VeryActiveMinutes ,fairlyActiveMinutes and LightlyActiveMinutes



I	J	K
VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes
25	13	328
21	19	217
30	11	181
29	34	209
36	10	221
38	20	164
42	16	233
50	31	264
28	12	205
19	8	211
66	27	130
41	21	262

Then I created a column containing Percentage values of LightlyActiveMinutes and VeryActiveMinutes as follows

Q	R	S
TotalActiveMins	PercentLightlyActiveMins	PercentVeryActiveMins
222	81.53	13.51
272	76.84	10.66
267	82.77	13.48
222	73.87	17.12
291	80.07	14.43
345	76.52	14.49
245	83.67	11.43
238	88.66	7.98
223	58.30	29.60
324	80.86	12.65
282	84.40	13.83
303	71.29	24.09

Now in the “sleepDay_merged.csv” file I converted the Total Minutes of Sleep to the Total Hours of sleep

D	E	F
TotalMinutesAsleep	TotalTimeInBed	TotalHoursAsleep
327	346	8.175
384	407	9.6
412	442	10.3
340	367	8.5
700	712	17.5
304	320	7.6
360	377	9
325	364	8.125
361	384	9.025
430	449	10.75
277	323	6.925
245	274	6.125
366	393	9.15
341	354	8.525
404	425	10.1

Analyses and Visualizations

Installing Tidyverse package and loading ggplot

```
install.packages("tidyverse",repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/shiva/OneDrive/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## Warning: unable to access index for repository http://cran.us.r-project.org/src/contrib:
##   cannot open URL 'http://cran.us.r-project.org/src/contrib/PACKAGES'
```

```
## Warning: package 'tidyverse' is not available (for R version 4.0.2)
```

```
## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/4.0:
##   cannot open URL 'http://cran.us.r-project.org/bin/windows/contrib/4.0/PACKAGES'
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

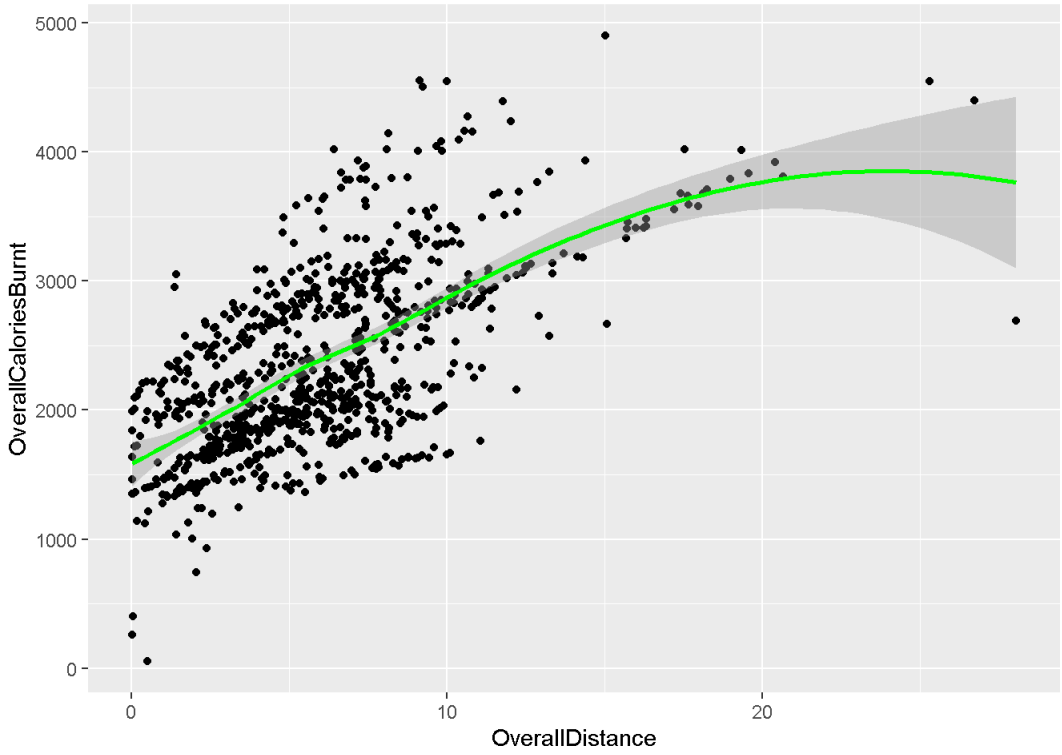
Analysis and visualizations based on Distance covered :

- Using Cleaned version of “dailyActivity_merged.csv”

Total Distance Vs Overall Calories Burnt

```
fit<-read.csv("dailyActivity_merged_cleaned.csv")
ggplot(data=fit)+geom_point(mapping = aes(x=TotalDistance,y=Calories))+
  geom_smooth(mapping = aes(x=TotalDistance,y=Calories),color="green")+
  xlab("OverallDistance")+ylab("OverallCaloriesBurnt")
```

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

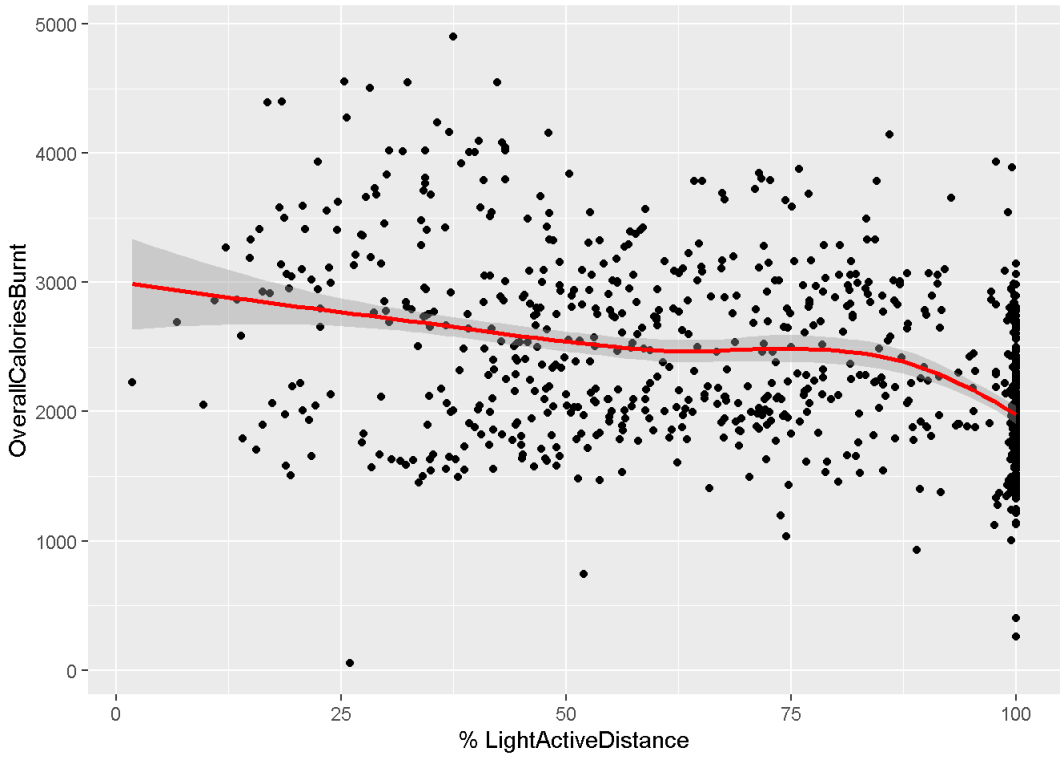


It clearly shows a positive correlation between Total Distance and Calories Burnt . Hence there is a upward Trendline . But! Lets dive deeper , and have a look at

Percentage of Lightly Active distance Vs Total Calories Burnt

```
ggplot(data=fit)+geom_point(mapping = aes(x=PercentLightDistance,y=Calories))+
  geom_smooth(mapping = aes(x=PercentLightDistance,y=Calories),color="red")+
  xlab("% LightActiveDistance")+ylab("OverallCaloriesBurnt")
```

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



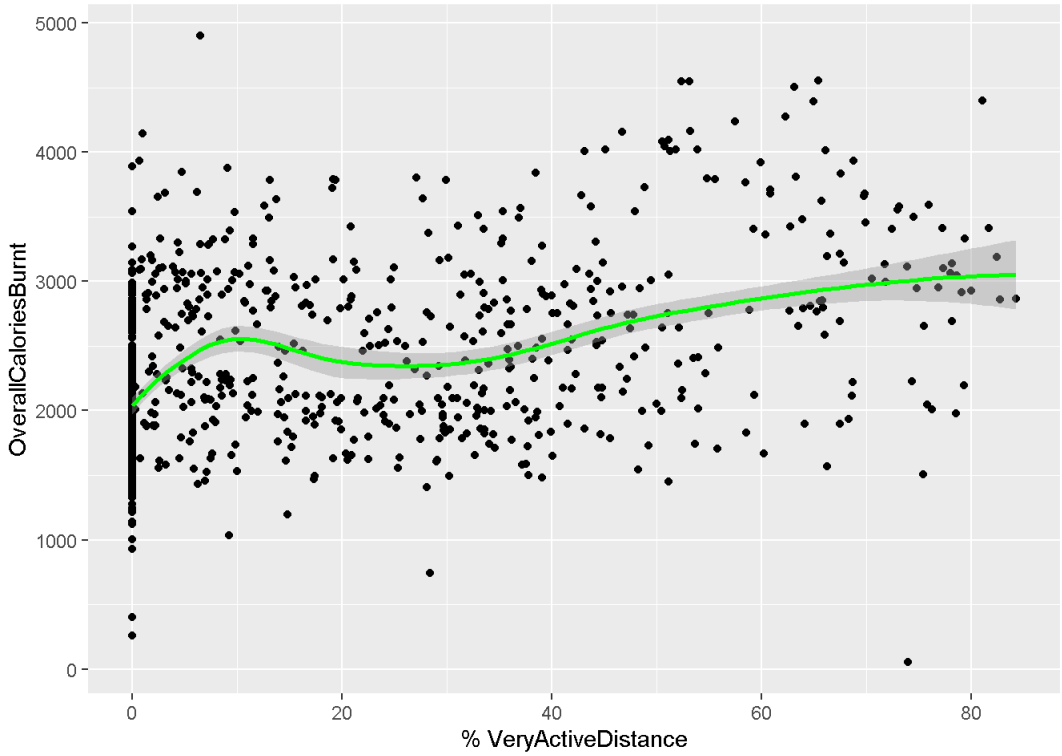
This shows that more the percentage of lightly active covered distance in total distance covered lesser the overall calories burnt. There is a declining trendline.

##On the Other Hand

Percentageof very active dictance Vs Total Calories Burnt

```
ggplot(data=fit)+geom_point(mapping = aes(x=PercentVeryActiveDistance,y=Calories))+
  geom_smooth(mapping = aes(x=PercentVeryActiveDistance,y=Calories),color="green")+
  xlab("% VeryActiveDistance")+ylab("OverallCaloriesBurnt")
```

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



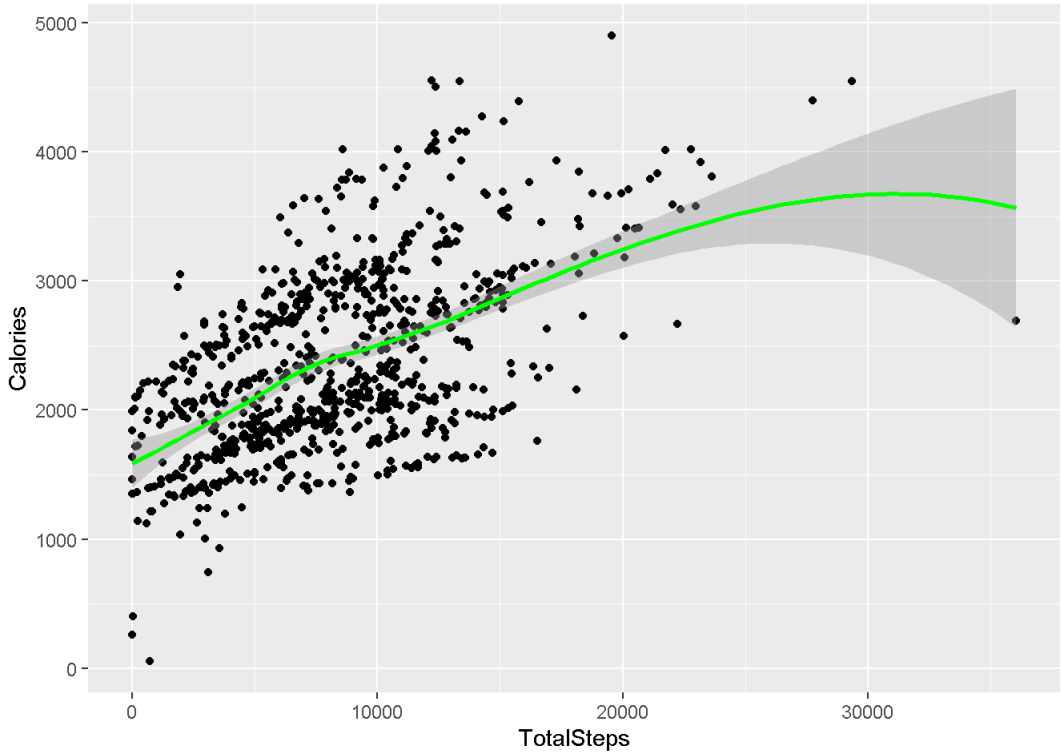
This Clearly Shows that more the percentage of very actively covered distance in total distance covered more the overall calories burnt

Analysis and visualizations based on Total Steps Taken :

Total Steps Vs Overall Calories Burnt

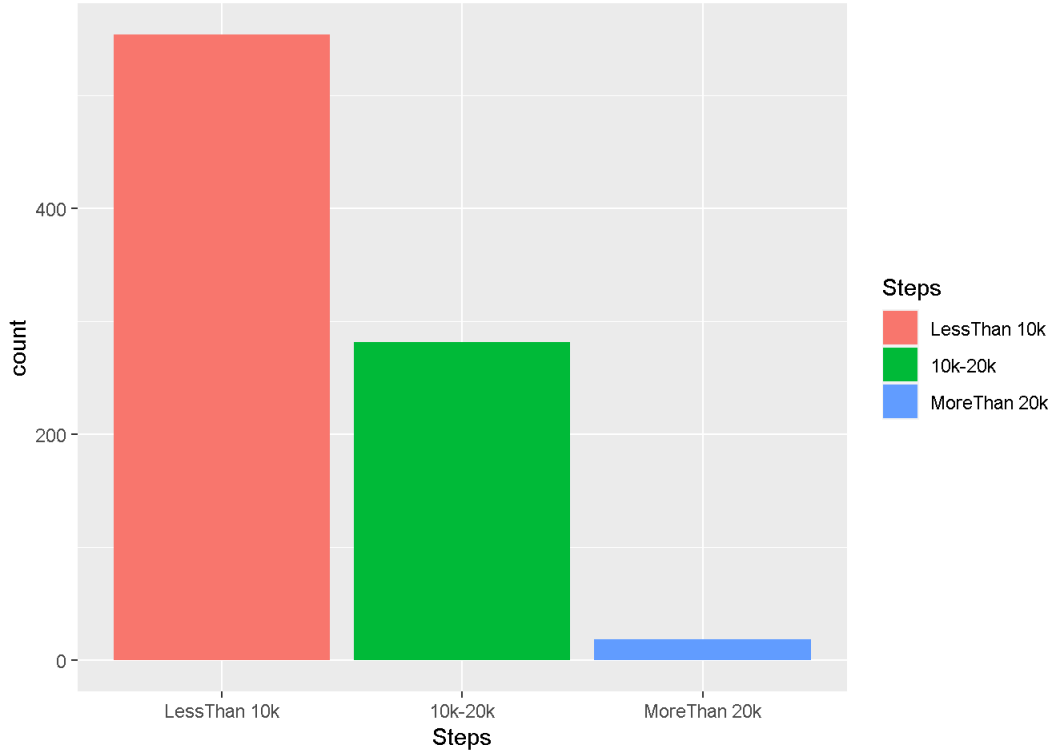
```
library(ggplot2)
ggplot(data=fit)+
  geom_point(mapping = aes(x=TotalSteps,Calories))+
  geom_smooth(mapping = aes(x=TotalSteps,Calories),color="green")
```

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



It is clear that more steps burns more calories . But According to the data , 64.8 % of the observations are less than 10k steps .

```
fit$Steps<-fit$TotalSteps
fit$Steps<-cut(fit$Steps,breaks = c(0,9999,19999,37000),labels = c("LessThan 10k","10k-20k","MoreThan 20k"))
ggplot(data=fit)+geom_bar(mapping = aes(x=Steps,fill=Steps))
```

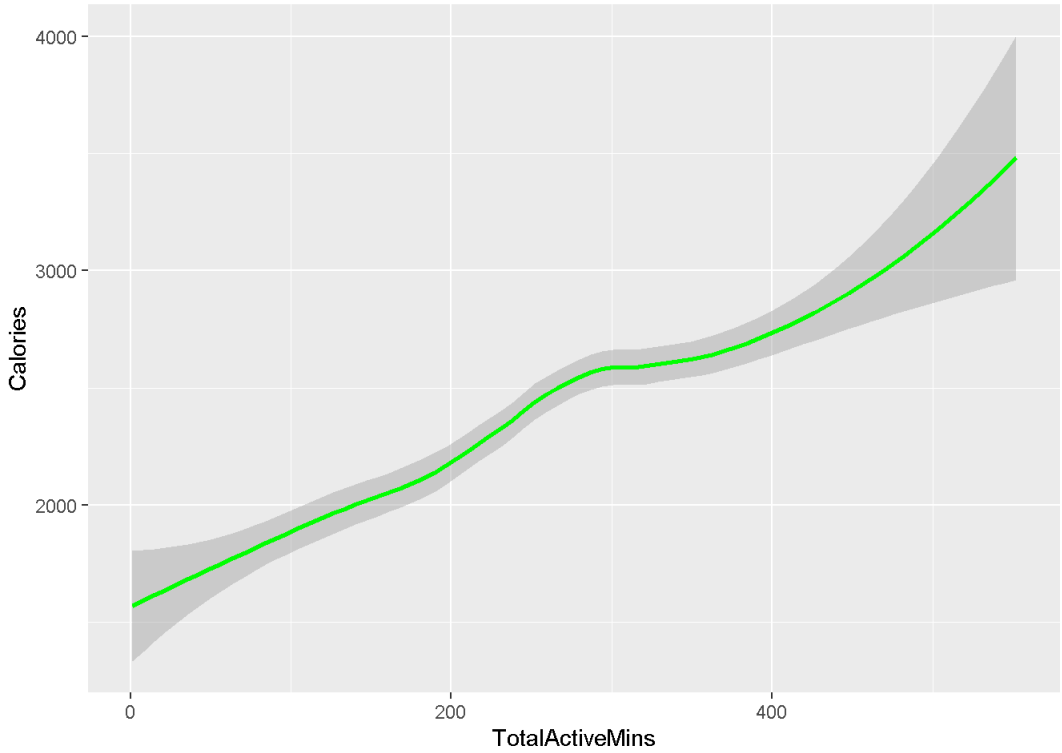


Analysis based on Active Minutes

It is obvious that more the total active minutes in a day more the overall calories burnt

```
ggplot(data=fit)+geom_smooth(mapping = aes(x=TotalActiveMins,y=Calories),color="green")
```

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

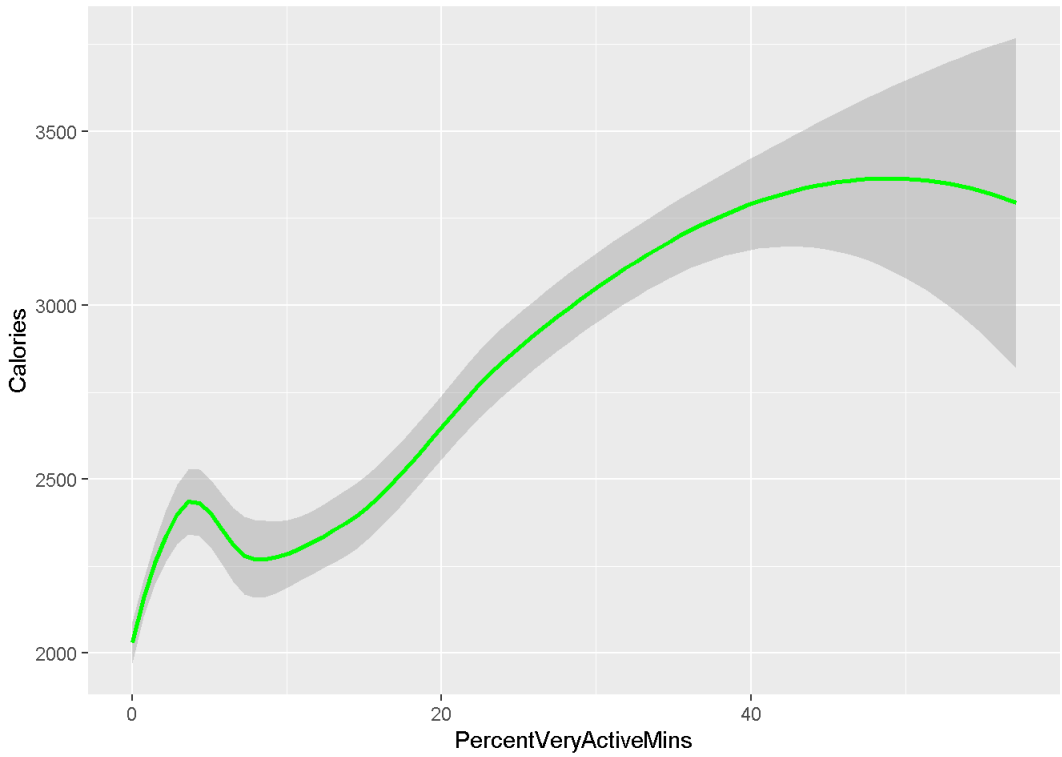


But these active minutes can be Lightly Active Minutes or Very Active Minutes. Lets see , how the percentage values Very Active Minutes is Related to the overall Calories burnt

Percentage of Very Active Minutes Vs Calories Burnt

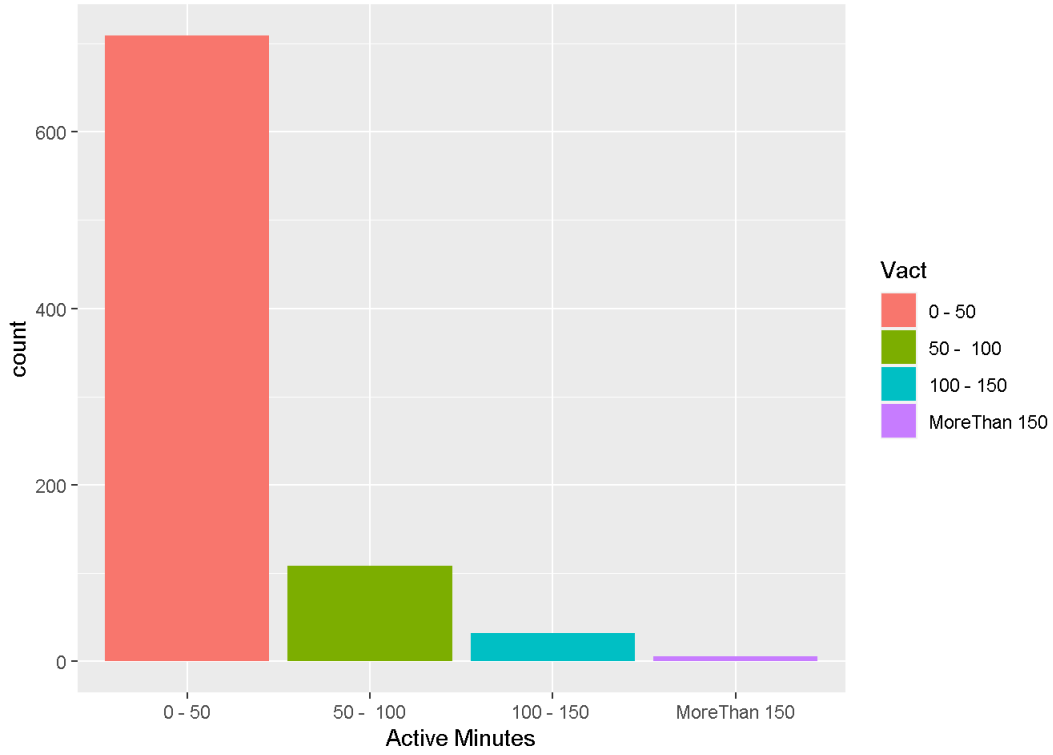
```
ggplot(data=fit)+
  geom_smooth(mapping = aes(x=PercentVeryActiveMins,y=Calories),color="green")
```

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



More the percentage of Very Active Minutes in total active minutes , more the overall calories burnt . But In the data Most of the Oservations have very less “Very Active Minutes”

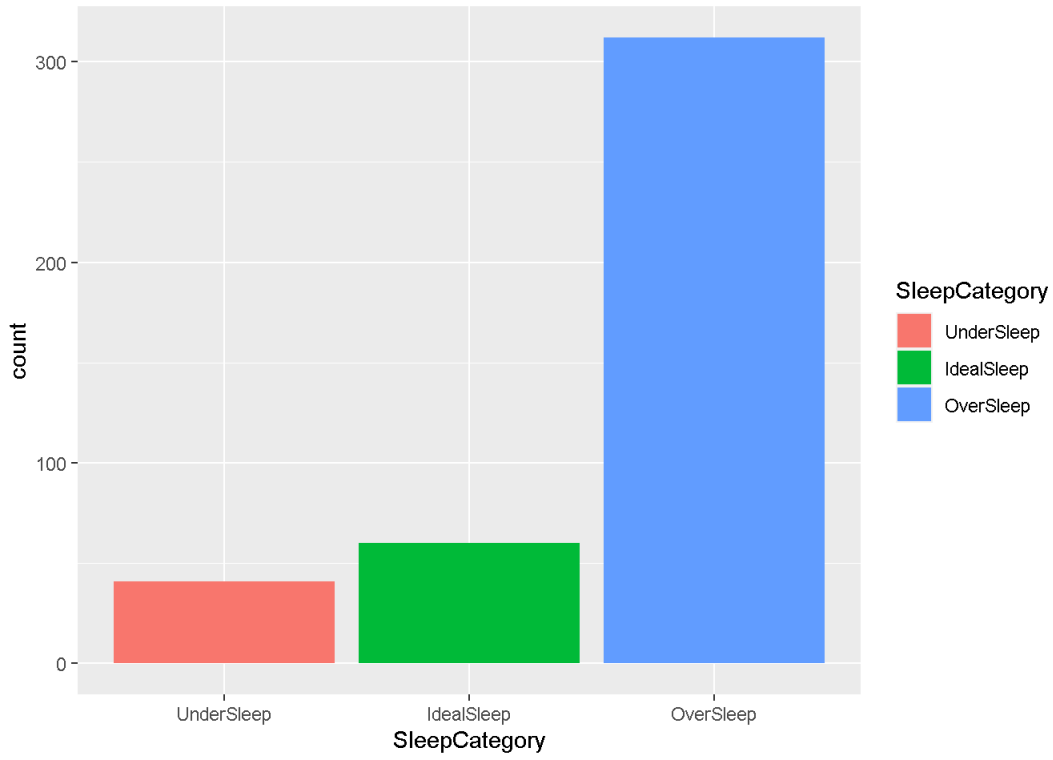
```
fit$Vact<-fit$VeryActiveMinutes
fit$Vact<-cut(fit$Vact,breaks = c(-1,50,100,150,250),labels = c("0 - 50","50 -  100","100 - 150 ","MoreThan 150"))
ggplot(data = fit)+geom_bar(mapping = aes(x=Vact,fill=Vact))+xlab("Active Minutes")
```



Analysis on Sleep

On the basis of the data and my analysis, 75.5% of the observations falls under the category of “OverSleep”

```
sleep<-read.csv("sleepDay_merged.csv")
sleep$SleepCategory<-sleep$TotalHoursAsleep
sleep$SleepCategory<-cut(sleep$SleepCategory,breaks = c(1,7,9,20),labels = c("UnderSleep","IdealSleep","Over
Sleep"))
ggplot(data=sleep)+geom_bar(mapping = aes(x=SleepCategory,fill=SleepCategory))
```



Some recommendations based on above analyses :

- Add and advertise a feature that give daily targets of steps to take to the users . (increase the target weekly by some some steps)
- Add and advertise a feature that challenge the users to walk faster while the users are walking .
- Add and advertise a feature that categorize the user based on their sleep hours and motivates them to fall under “Ideal Sleeper”

category.