Capstone Project for the Google Data Analytics Professional Certificate

Bellabeat is a high-tech manufacturer of health-focused products for women. As a junior data analyst working with marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. I have performed analysis on data to give recommendations.

1. Ask

The business task is to analyze smart device usage data to gain insights, identify trends and understand how users use these products. The insights will be used to make data-driven decisions by applying them to one of Bellabeat's products and help guide marketing strategy.

The key stakeholders are as follows:

- Urška Sršen: Bellabeat's co-founder and Chief Creative Officer
- Sando Mur: Mathematician and Bellabeat's cofounder; a key member of Bellabeat's executive team
- Bellabeat's marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

Business objectives are as follows:

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat's marketing strategy?

2. Prepare

The data set is Fitbit Fitness Tracker Data taken from Kaggle which contains personal fitness trackers from thirty Fitbit users. It contains 18 CSV files. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits. Generated by respondents from a survey via Amazon Mechanical Turk between 12 March 2016 to 12 May 2016.

The dataset is organized in long and wide formats. The dataset is bad quality as we can check for the Dataset to be ROCCC as follows.

- Reliable: Low, as there are only 30 respondents
- Original: Low as the third-party provider (Amazon Mechanical Turk)
- Comprehensive: Medium, as matches Bellabeat product data
- Current: Low, data is 5 years old
- Cited: Low, Data collected from third party

3. Process, Analyze, Share, Act

Excel will be used for the data cleaning process and for removing errors from it. R programming language will be used for analysis and visualization.

Following cleaning and manipulation of data are handled. Null data, misspelled words, mistyped numbers, extra spaces and characters, duplication, mismatched data types, inconsistent strings, and date formats, misleading column names, business logic, and truncated/missing data.

The data has been uploaded to R and analyzed.

import library and datasets.

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

```
## Installing package into 'C:/Users/tashf/AppData/Local/R/win-library/4.2
'
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\tashf\AppData\Local\Temp\RtmpiEpR9a\downloaded_packages
```

```
library(tidyverse)
```

```
## ── Attaching packages
## ─────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✓ ggplot2 3.4.0      ✓ purrr   1.0.1
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────── tidyverse_confl
icts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

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

```
## Warning: package 'dplyr' is in use and will not be installed
```

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

```
## Installing package into 'C:/Users/tashf/AppData/Local/R/win-library/4.2
'
## (as 'lib' is unspecified)
```

```
## package 'janitor' successfully unpacked and MD5 sums checked
##
```

```
## The downloaded binary packages are in
##  C:\Users\tashf\AppData\Local\Temp\RtmpiEpR9a\downloaded_packages

library(janitor)

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

library(ggplot2)
library(lubridate)

## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Importing data

```
daily_activity <- read.csv("C:\\Users\\tashf\\OneDrive\\Desktop\\Case Stud
y\\Google_Project\\Fitabase_Data_4.12.16-5.12.16\\dailyActivity_merged.csv
")
daily_sleep <- read.csv("C:\\Users\\tashf\\OneDrive\\Desktop\\Case Study\\
Google_Project\\Fitabase_Data_4.12.16-5.12.16\\sleepDay_merged.csv")
weight_log <- read.csv("C:\\Users\\tashf\\OneDrive\\Desktop\\Case Study\\G
oogle_Project\\Fitabase_Data_4.12.16-5.12.16\\weightLogInfo_merged.csv")
daily_calories <- read.csv("C:\\Users\\tashf\\OneDrive\\Desktop\\Case Stud
y\\Google_Project\\Fitabase_Data_4.12.16-5.12.16\\dailyCalories_merged.csv
")
daily_intensities <- read.csv("C:\\Users\\tashf\\OneDrive\\Desktop\\Case S
tudy\\Google_Project\\Fitabase_Data_4.12.16-5.12.16\\dailyIntensities_merg
ed.csv")
daily_steps <- read.csv("C:\\Users\\tashf\\OneDrive\\Desktop\\Case Study\\
Google_Project\\Fitabase_Data_4.12.16-5.12.16\\dailySteps_merged.csv")
```

Data is joined using id and date columns

```
merge_1 <- merge(daily_activity, daily_sleep, by = c("Id" ))
merge_2 <- merge(daily_steps, daily_calories, by = c("Id" ))
merge_3 <- merge(daily_intensities, weight_log, by = c("Id" ))
merge_4 <- merge(merge_1, merge_2, by = c("Id" ))
```

Convert Date from charater format to Date format

```
merge_4$Date <- mdy(merge_4$Date)
```

summarize the data.

```
summary(merge_4)
```

```
##        Id                 Date                 Time              WeightK
g
##  Min.   :1.504e+09   Min.   :2016-04-12   Length:1043982     Min.   : 5
2.60
##  1st Qu.:6.962e+09   1st Qu.:2016-04-18   Class :character   1st Qu.: 6
1.20
##  Median :6.962e+09   Median :2016-04-28   Mode  :character   Median : 6
1.50
##  Mean   :6.475e+09   Mean   :2016-04-26                      Mean   : 6
2.83
##  3rd Qu.:6.962e+09   3rd Qu.:2016-05-04                      3rd Qu.: 6
1.90
##  Max.   :6.962e+09   Max.   :2016-05-12                      Max.   :13
3.50
##
##   WeightPounds        Fat              BMI        IsManualReport
##  Min.   :116.0   Min.   :22.0     Min.   :22.65   Mode :logical
##  1st Qu.:134.9   1st Qu.:22.0     1st Qu.:23.89   FALSE:28205
##  Median :135.6   Median :25.0     Median :24.00   TRUE :1015777
##  Mean   :138.5   Mean   :23.5     Mean   :24.40
##  3rd Qu.:136.5   3rd Qu.:25.0     3rd Qu.:24.17
##  Max.   :294.3   Max.   :25.0     Max.   :47.54
##                  NA's   :994971
##      LogId            SleepDay         TotalSleepRecords TotalMinutesAs
leep
##  Min.   :1.46e+12   Length:1043982     Min.   :1.000     Min.   : 59.0
##  1st Qu.:1.46e+12   Class :character   1st Qu.:1.000     1st Qu.:411.0
##  Median :1.46e+12   Mode  :character   Median :1.000     Median :442.0
##  Mean   :1.46e+12                      Mean   :1.092     Mean   :437.5
##  3rd Qu.:1.46e+12                      3rd Qu.:1.000     3rd Qu.:476.0
##  Max.   :1.46e+12                      Max.   :3.000     Max.   :750.0
##
##  TotalTimeInBed  ActivityDay.x        StepTotal     ActivityDay.y
##  Min.   : 65.0   Length:1043982     Min.   :    0   Length:1043982
##  1st Qu.:424.0   Class :character   1st Qu.: 5908   Class :character
##  Median :457.0   Mode  :character   Median :10320   Mode  :character
##  Mean   :456.3                      Mean   : 9658
##  3rd Qu.:497.0                      3rd Qu.:12207
##  Max.   :775.0                      Max.   :20031
##
##     Calories
##  Min.   :   0
##  1st Qu.:1850
##  Median :2039
##  Mean   :2010
##  3rd Qu.:2173
##  Max.   :4552
##
```

check data for unique id

```
 n_distinct(daily_activity$Id)
```

```
## [1] 8
```

```
nrow(daily_activity)
```

```
## [1] 67
```

should be 30 as survey was of 30 people

```
summary(merge_3)
```

```
##        Id            ActivityDay        SedentaryMinutes LightlyActiveM
inutes
##  Min.   :1.504e+09   Length:2076        Min.   :   0.0   Min.   :  0.0
##  1st Qu.:6.962e+09   Class :character   1st Qu.: 680.0   1st Qu.:210.8
##  Median :6.962e+09   Mode  :character   Median : 837.0   Median :235.5
##  Mean   :7.010e+09                      Mean   : 887.8   Mean   :240.8
##  3rd Qu.:8.878e+09                      3rd Qu.:1122.2   3rd Qu.:288.0
##  Max.   :8.878e+09                      Max.   :1440.0   Max.   :448.0
##
##  FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
##  Min.   : 0.00       Min.   :  0.00    Min.   :0.000000
##  1st Qu.: 4.00       1st Qu.:  7.00    1st Qu.:0.000000
##  Median :12.00       Median : 29.00    Median :0.000000
##  Mean   :14.45       Mean   : 37.63    Mean   :0.005039
##  3rd Qu.:22.00       3rd Qu.: 61.00    3rd Qu.:0.000000
##  Max.   :74.00       Max.   :210.00    Max.   :0.110000
##
##  LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
##  Min.   : 0.000      Min.   :0.000            Min.   : 0.0000
##  1st Qu.: 3.610      1st Qu.:0.080            1st Qu.: 0.1375
##  Median : 4.660      Median :0.410            Median : 1.7400
##  Mean   : 4.701      Mean   :0.659            Mean   : 3.3042
##  3rd Qu.: 5.890      3rd Qu.:1.070            3rd Qu.: 3.9000
##  Max.   :10.710      Max.   :2.390            Max.   :21.6600
##
##      Date              WeightKg       WeightPounds        Fat
##  Length:2076        Min.   : 52.60   Min.   :116.0   Min.   :22.0
##  Class :character   1st Qu.: 61.40   1st Qu.:135.4   1st Qu.:22.0
##  Mode  :character   Median : 62.50   Median :137.8   Median :23.5
##                     Mean   : 72.03   Mean   :158.8   Mean   :23.5
##                     3rd Qu.: 85.10   3rd Qu.:187.6   3rd Qu.:25.0
##                     Max.   :133.50   Max.   :294.3   Max.   :25.0
##                                                      NA's   :2014
##       BMI        IsManualReport      LogId              X
##  Min.   :21.45   Mode :logical   Min.   :1.460e+12   Mode:logical
##  1st Qu.:23.96   FALSE:805       1st Qu.:1.461e+12   NA's:2076
##  Median :24.39   TRUE :1271      Median :1.462e+12
##  Mean   :25.18                   Mean   :1.462e+12
##  3rd Qu.:25.56                   3rd Qu.:1.462e+12
##  Max.   :47.54                   Max.   :1.463e+12
##
```

quick summary statistics

```
colnames(daily_activity)
```

```
## [1] "Id"             "Date"           "Time"           "WeightKg"
## [5] "WeightPounds"   "Fat"            "BMI"            "IsManualReport"
## [9] "LogId"
```

```
daily_sleep %>%
  select(TotalSleepRecords,
  TotalMinutesAsleep,
  TotalTimeInBed) %>%
  summary()
```

```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  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
```

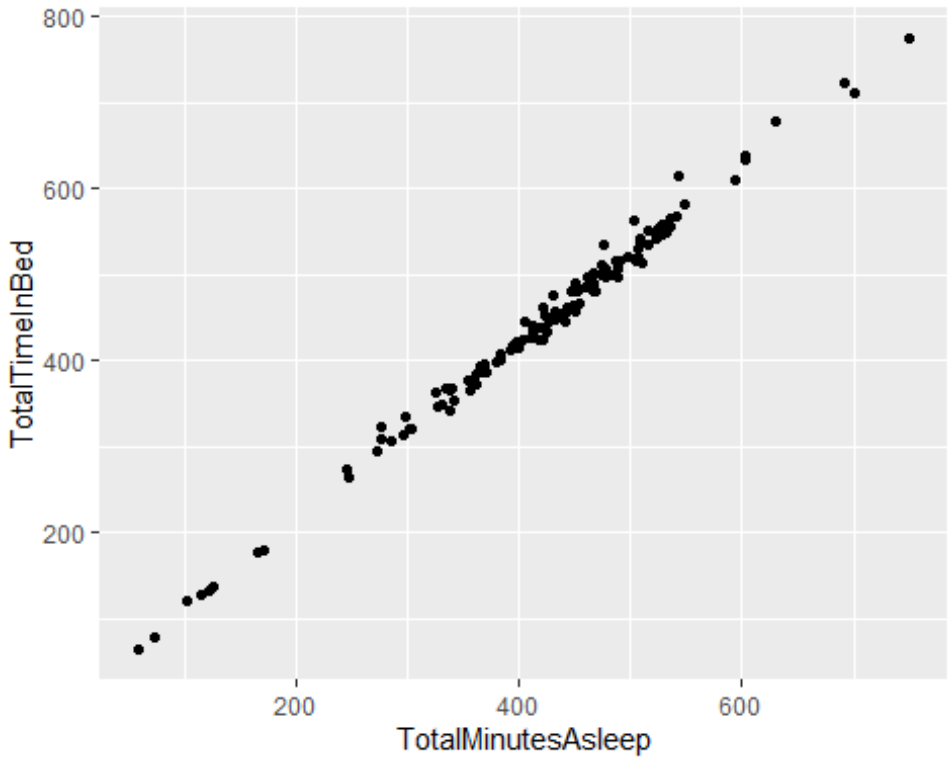
```
colnames(merge_3)
```

```
##  [1] "Id"                       "ActivityDay"
##  [3] "SedentaryMinutes"         "LightlyActiveMinutes"
##  [5] "FairlyActiveMinutes"      "VeryActiveMinutes"
##  [7] "SedentaryActiveDistance"  "LightActiveDistance"
##  [9] "ModeratelyActiveDistance" "VeryActiveDistance"
## [11] "Date"                     "WeightKg"
## [13] "WeightPounds"             "Fat"
## [15] "BMI"                      "IsManualReport"
## [17] "LogId"                    "X"
```

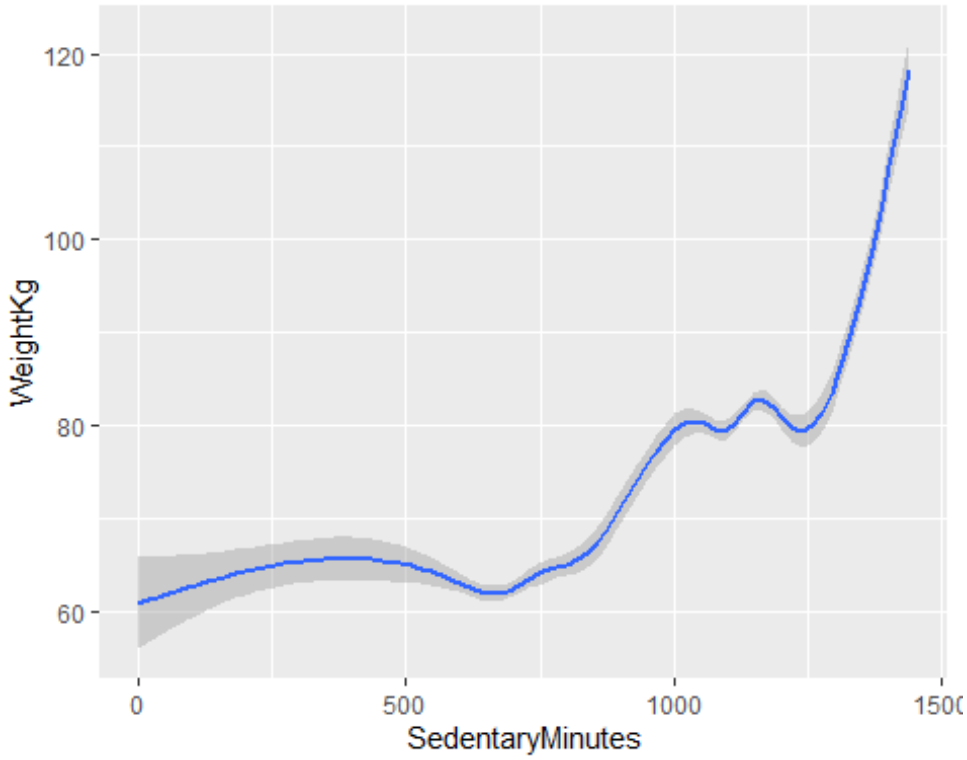
```
colnames(merge_4)
```

```
##  [1] "Id"                 "Date"               "Time"
##  [4] "WeightKg"           "WeightPounds"       "Fat"
##  [7] "BMI"                "IsManualReport"     "LogId"
## [10] "SleepDay"           "TotalSleepRecords"  "TotalMinutesAsleep"
## [13] "TotalTimeInBed"     "ActivityDay.x"      "StepTotal"
## [16] "ActivityDay.y"      "Calories"
```

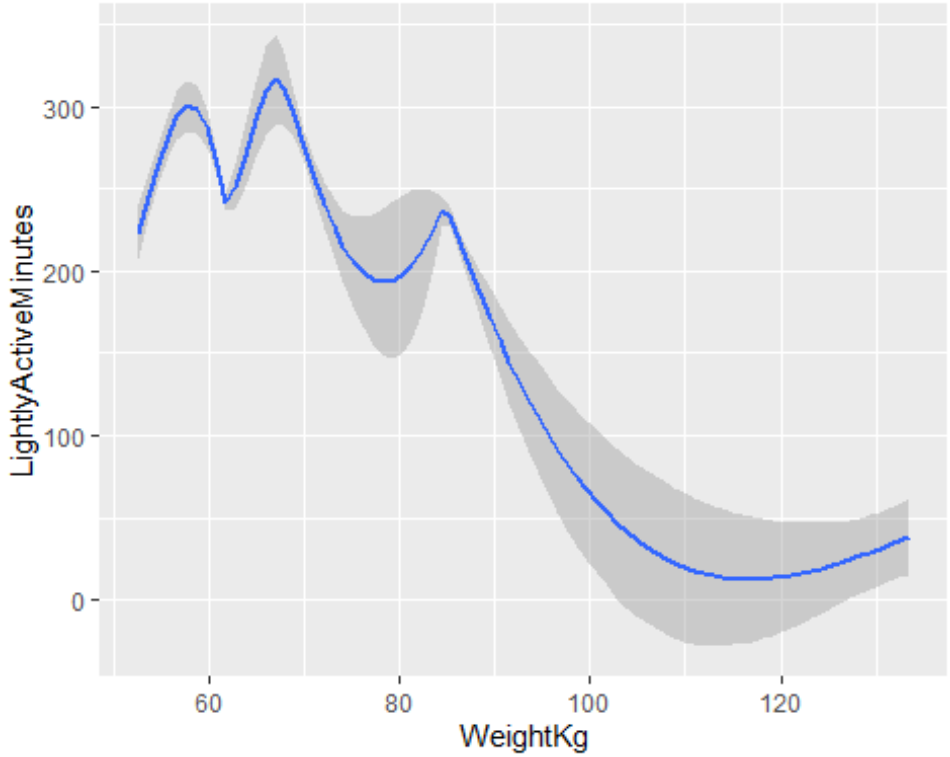
```
ggplot(data=merge_4) +geom_point(mapping=aes(x=TotalMinutesAsleep,y=TotalT
imeInBed))
```



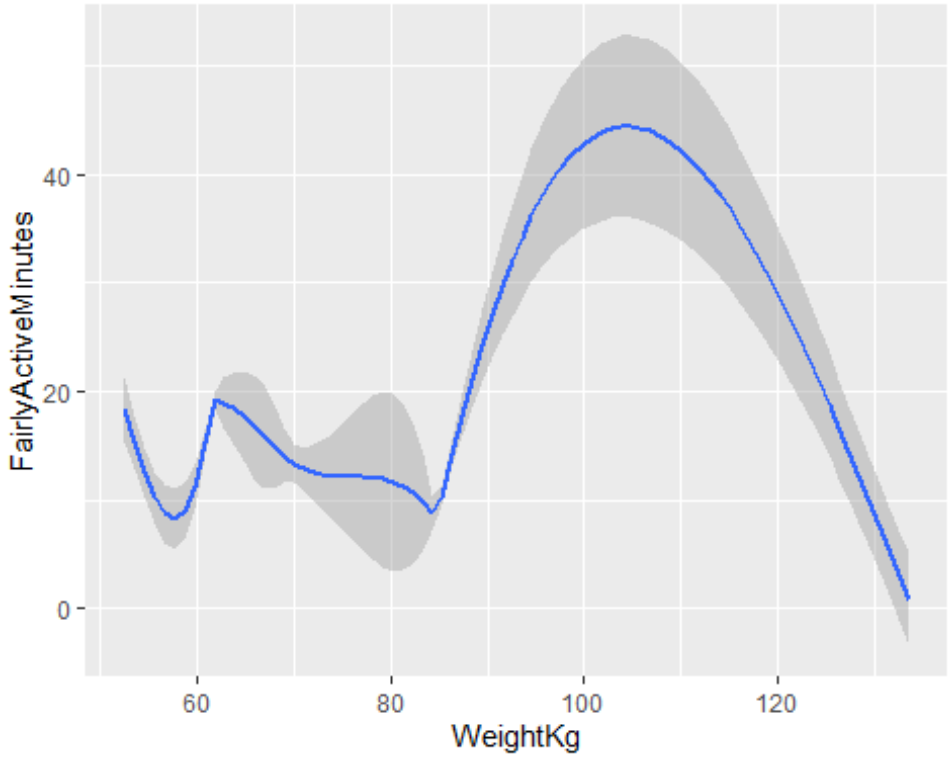

```
ggplot(data=merge_3) +geom_smooth(mapping=aes(y=WeightKg, x = SedentaryMin
utes))

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



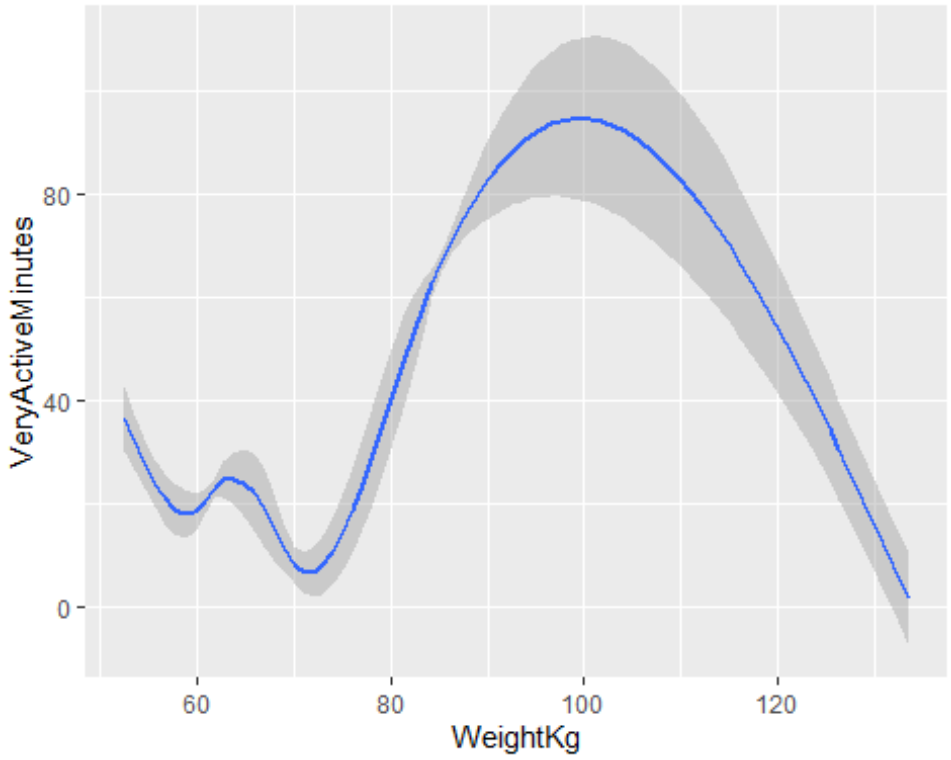
```
ggplot(data=merge_3) +geom_smooth(mapping=aes(x=WeightKg, y = LightlyActiv
eMinutes))

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




```
ggplot(data=merge_3) +geom_smooth(mapping=aes(x=WeightKg, y = FairlyActive
Minutes))

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

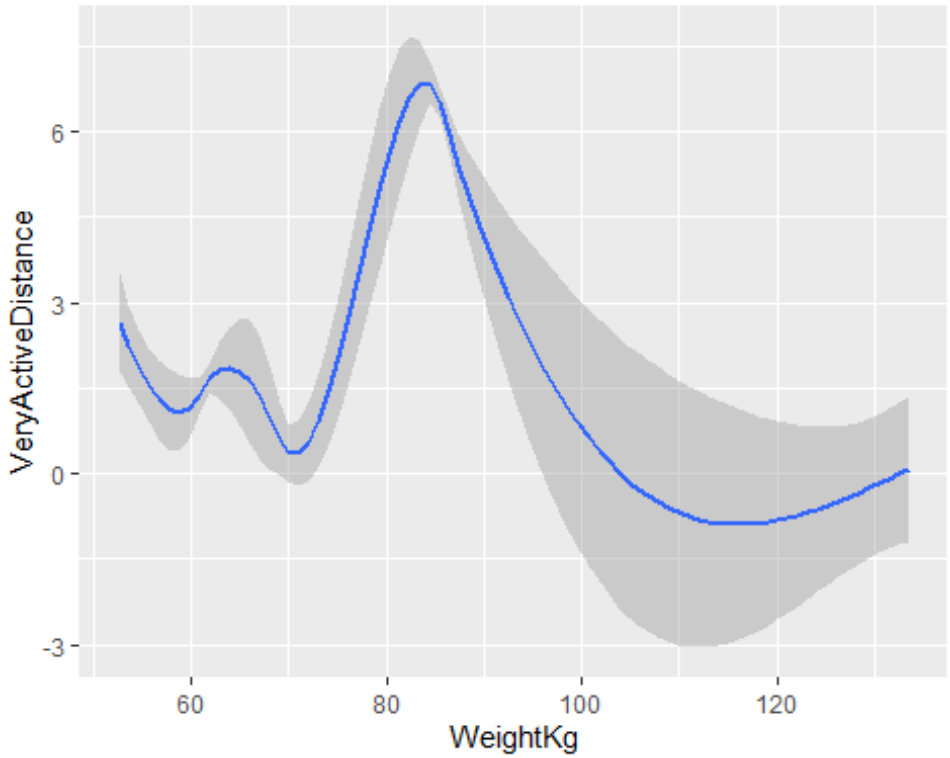



```
ggplot(data=merge_3) +geom_smooth(mapping=aes(x=WeightKg, y = VeryActiveMi
nutes))

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




```
ggplot(data=merge_3) +geom_smooth(mapping=aes(x=WeightKg, y = VeryActiveDi
stance))
```

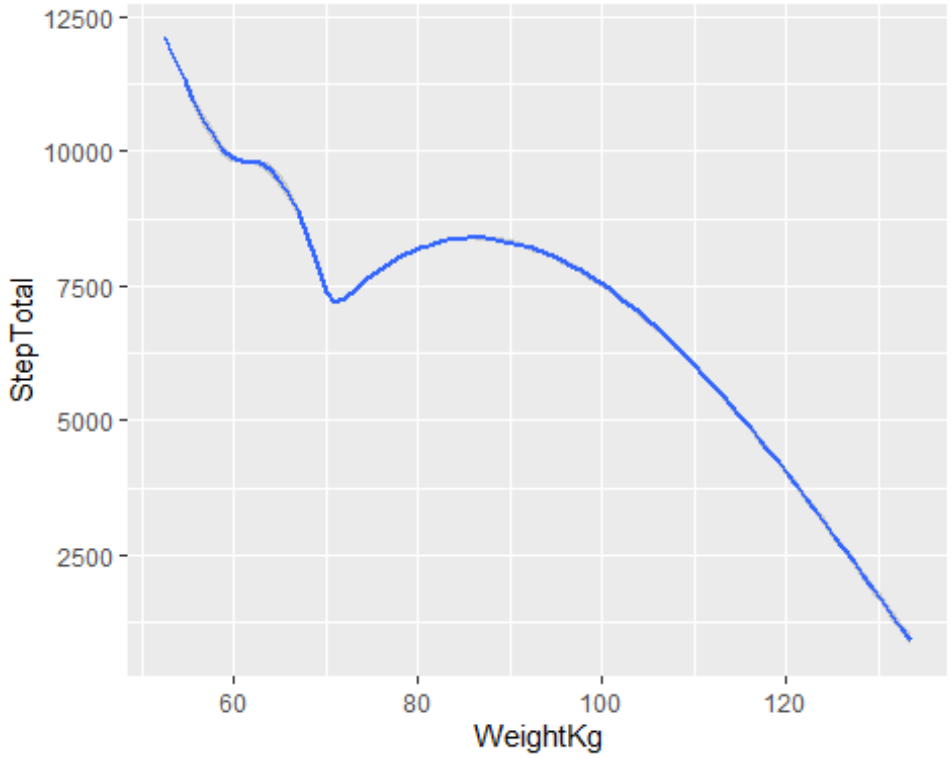
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")
'
```



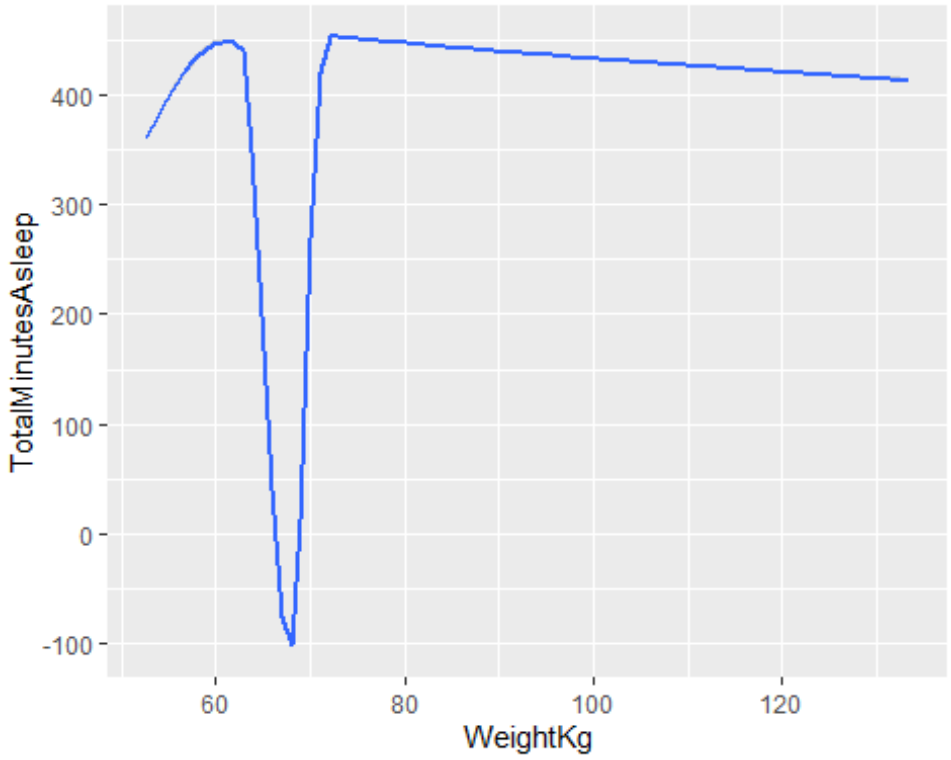
From these plot we can see higher weight people are more sedentary. so we should target specifically above 70kg between 90 and 120kg people are very or fairly active- seems like high weight people are trying to lose weight and exercise more than normal people but they have less very active distance which means they run/jog less and are using indoor activities to stay active such as gym

```
ggplot(data=merge_4) +geom_smooth(mapping=aes(x=WeightKg, y = StepTotal))
```

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

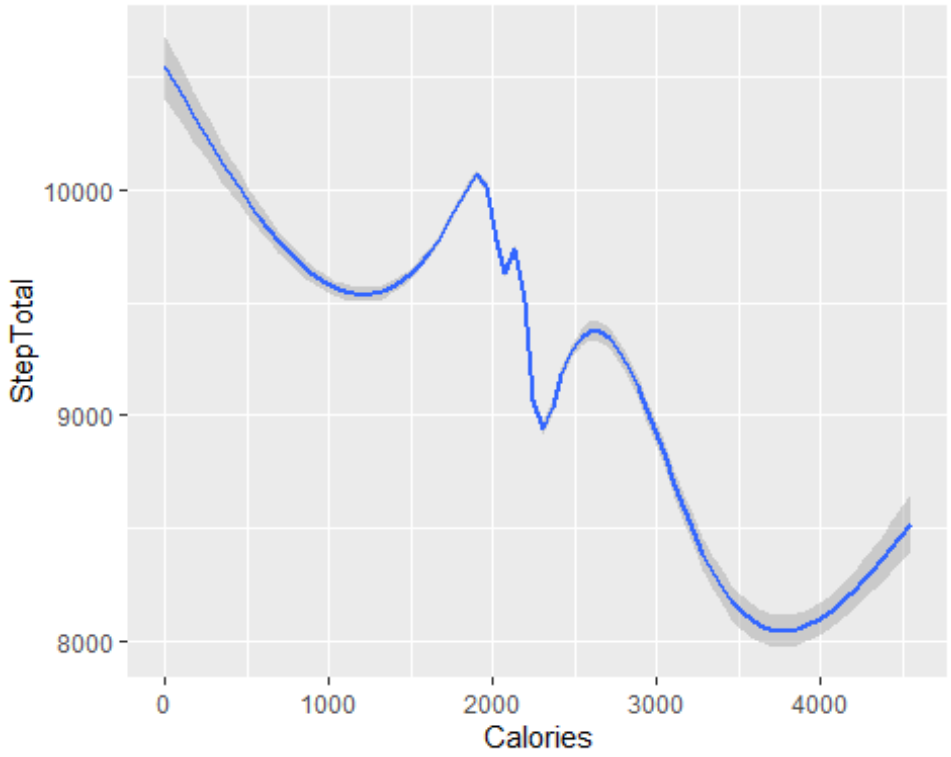



```
ggplot(data=merge_4) +geom_smooth(mapping=aes(x=WeightKg, y = TotalMinutes
Asleep))

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




```
ggplot(data=merge_4) +geom_smooth(mapping=aes(x=Calories, y = StepTotal))
```

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




```
heart_rate <- read.csv("heartrate_seconds_merged.csv")
hourly_calories <- read.csv("hourlyCalories_merged.csv")
hourly_steps <- read.csv("hourlySteps_merged.csv")
```

column names

```
colnames(heart_rate)
```

```
## [1] "Id"    "Time"  "Value"
```

```
colnames(hourly_calories)
```

```
## [1] "Id"           "ActivityHour" "Calories"
```

```
colnames(hourly_steps)
```

```
## [1] "Id"           "ActivityHour" "StepTotal"
```

unique ids

```
n_distinct(heart_rate$Id)
```

```
## [1] 14
```

```
n_distinct(hourly_calories$Id)
```

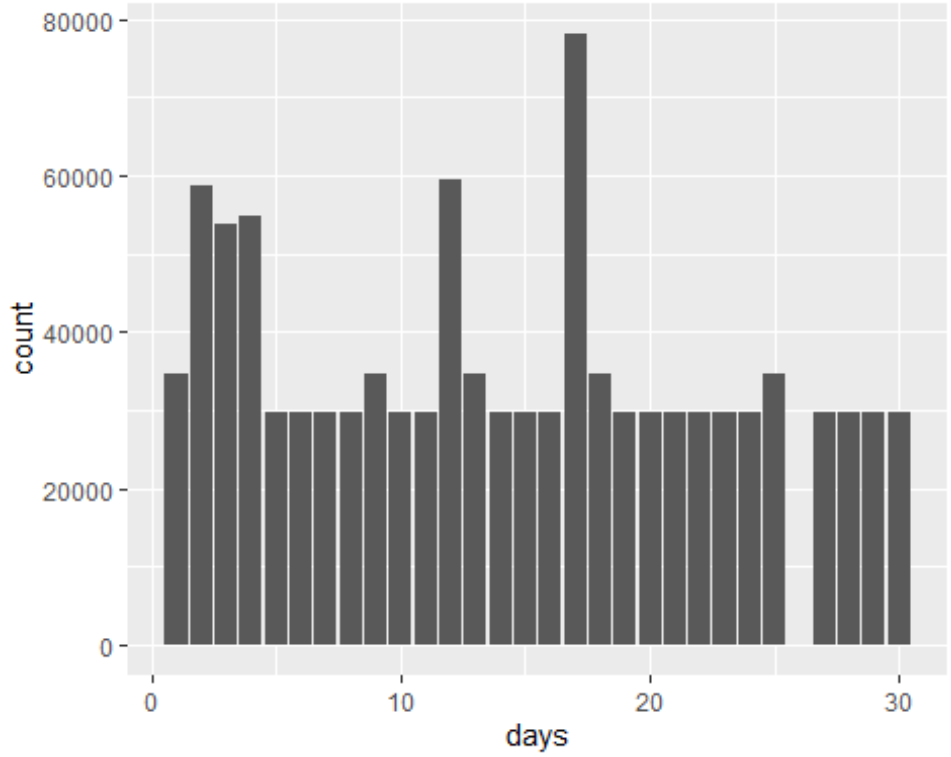
```
## [1] 33
```

```
n_distinct(hourly_steps$Id)
```

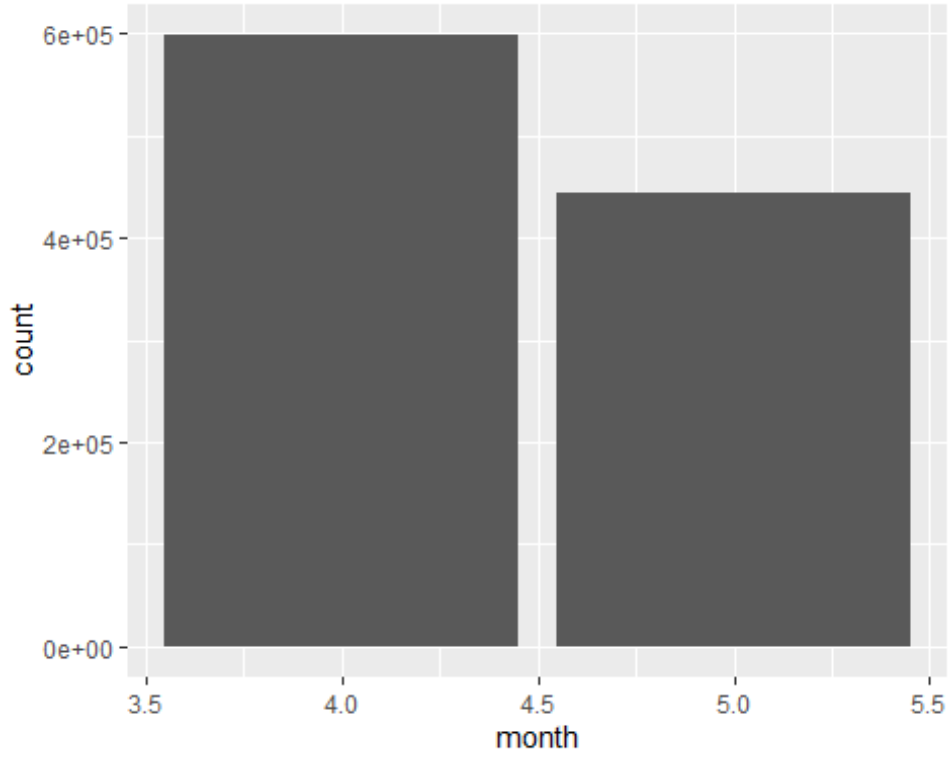
```
## [1] 33
```

Unique id should be 33

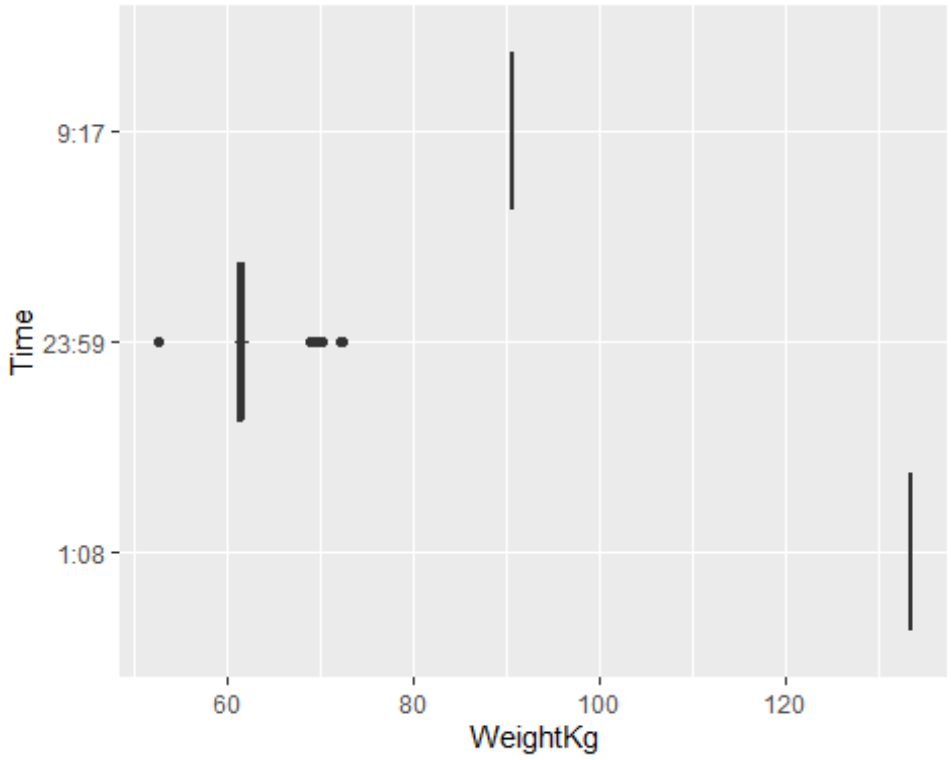
```
merge_4$month <- month(merge_4$Date)
merge_4$days <- day(merge_4$Date)

ggplot(data=merge_4)+ geom_bar(mapping=aes(x=days))
```




```
ggplot(data=merge_4)+ geom_bar(mapping=aes(x=month))
```



```
ggplot(data=merge_4)+ geom_boxplot(mapping=aes(x=WeightKg, y= Time))
```



4. Conclusion

- We can see higher weight people are more sedentary. so we should target specifically below 70kg but between 90 and 120kg people are very or fairly active - seems like high weight people are trying to lose weight and exercise more than normal people but they have less (very active distance) which means they run/jog less and are using indoor activities to stay active such as gym
- less than 70kg as they are active but won't be willing to pay a lot because they are not passionate, they have more active distance though meaning they run/walk more. however, between 90 and 120kg people are passionate and would be willing to spend more money
- Generally, people's data was collected at 12 mid night, 1pm noon and 9pm night.
- People are most active in the start of month and middle of month while most active months are April and May.
- Steps decreases with increase in weight, number of calories increase with decrease in step.
- Bellabeat's marketing team can encourage users by educating and equipping them with knowledge about fitness benefits, suggest different types of exercises, calories intake and burn rate information on Bellabeat's application.
- Most people use fitbit to track steps and calories burned, people don't use to track sleep much. I will suggest focusing on steps, calories and sleep tracking more in application.
- The relation between steps taken vs calories burned and very active minutes vs calories burned shows positive correlation. So, this can be a good marketing strategy.

- If users want to lose weight, it’s probably a good idea to control daily calorie consumption. Bellabeat’s can suggest some ideas for low-calorie lunch and dinner.
- The Bellabeat app can recommend reducing sedentary time.