Bellabeat-Case-Study2

Malay Raj

2023-04-20

## BELLABEAT CASE STUDY

This analysis is based on the Bellabeat case study “‘Sophisticated, Clear, and Polished’: Bellabeat data and Data Visualization”.

## PREPARE:

## Installing required packages

Installing and loading common packages and libraries.

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.1.8  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(skimr)

## Warning: package 'skimr' was built under R version 4.2.3

## COLLECTING DATA

uploading fitabase datasets(csv files) Here we’ll create a dataframes named ‘daily\_activity’,‘sleep\_day’,‘weight\_log’ and read in the CSV files from the datasets.

daily\_activity <- read.csv("C:/Users/KIIT/OneDrive/Desktop/Bellabeat/Fitabase Data 4.12.16-5.12.16/updated Datasets/dailyActivity\_merged.csv")  
sleep\_day <- read.csv("C:/Users/KIIT/OneDrive/Desktop/Bellabeat/Fitabase Data 4.12.16-5.12.16/updated Datasets/sleepDay - sleepDay\_merged.csv")  
weight\_log <- read.csv("C:/Users/KIIT/OneDrive/Desktop/Bellabeat/Fitabase Data 4.12.16-5.12.16/updated Datasets/weightLogInfo - weightLogInfo\_merged.csv")

## PROCESS:

## EXPLORING A FEW KEY TABLES

Take a look at all the datas.

head(daily\_activity)

## 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

head(sleep\_day)

## Id SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed  
## 1 2026352035 5/12/2016 1 456 485  
## 2 4319703577 5/12/2016 1 302 321  
## 3 4445114986 5/12/2016 1 483 501  
## 4 4702921684 5/12/2016 1 404 442  
## 5 5553957443 5/12/2016 1 438 475  
## 6 6962181067 5/12/2016 1 516 535

head(weight\_log)

## Id Date WeightKg WeightPounds Fat BMI IsManualReport  
## 1 2873212765 5/12/2016 57.3 126.3249 NA 21.69 TRUE  
## 2 6962181067 5/12/2016 61.9 136.4661 NA 24.17 TRUE  
## 3 8877689391 5/12/2016 84.0 185.1883 NA 25.14 FALSE  
## 4 6962181067 5/11/2016 61.9 136.4661 NA 24.17 TRUE  
## 5 8877689391 5/11/2016 85.4 188.2748 NA 25.56 FALSE  
## 6 6962181067 5/10/2016 62.1 136.9071 NA 24.24 TRUE  
## LogId  
## 1 1.463098e+12  
## 2 1.463098e+12  
## 3 1.463035e+12  
## 4 1.463011e+12  
## 5 1.462950e+12  
## 6 1.462925e+12

Identifying all the columns in the datasets.

colnames(daily\_activity)

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

colnames(sleep\_day)

## [1] "Id" "SleepDay" "TotalSleepRecords"   
## [4] "TotalMinutesAsleep" "TotalTimeInBed"

colnames(weight\_log)

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

Note that both datasets have the ‘Id’ field - this can be used to merge the datasets.

## UNDERSTANDING SOME SUMMARY STATISTICS

How many unique participants are there in each dataframe? It looks like there may be more participants in the daily activity dataset than the sleep day dataset and weight log dataset.

n\_distinct(daily\_activity$Id)

## [1] 33

n\_distinct(sleep\_day$Id)

## [1] 24

n\_distinct(weight\_log$Id)

## [1] 8

How many observations are there in each dataframe?

nrow(daily\_activity)

## [1] 940

nrow(sleep\_day)

## [1] 413

nrow(weight\_log)

## [1] 67

What are some quick summary statistics we’d want to know about each data frame? Let’s find out that. For the daily activity dataframe:

daily\_activity %>%   
 select(TotalSteps,  
 TotalDistance,  
 SedentaryMinutes) %>%  
 summary()

## TotalSteps TotalDistance SedentaryMinutes  
## Min. : 0 Min. : 0.000 Min. : 0.0   
## 1st Qu.: 3790 1st Qu.: 2.620 1st Qu.: 729.8   
## Median : 7406 Median : 5.245 Median :1057.5   
## Mean : 7638 Mean : 5.490 Mean : 991.2   
## 3rd Qu.:10727 3rd Qu.: 7.713 3rd Qu.:1229.5   
## Max. :36019 Max. :28.030 Max. :1440.0

For the sleep day dataframe:

sleep\_day %>%   
 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

For the weight log dataframe:

weight\_log %>%   
 select(WeightKg,  
 WeightPounds,  
 BMI) %>%  
 summary()

## WeightKg WeightPounds BMI   
## Min. : 52.60 Min. :116.0 Min. :21.45   
## 1st Qu.: 61.40 1st Qu.:135.4 1st Qu.:23.96   
## Median : 62.50 Median :137.8 Median :24.39   
## Mean : 72.04 Mean :158.8 Mean :25.19   
## 3rd Qu.: 85.05 3rd Qu.:187.5 3rd Qu.:25.56   
## Max. :133.50 Max. :294.3 Max. :47.54

## ANALYZE:

We then run a quick summary on the various data frames by using the skim\_without\_chart() function to provide broader overview of a data frames.

skim\_without\_charts(daily\_activity)

Data summary

|  |  |
| --- | --- |
| Name | daily\_activity |
| Number of rows | 940 |
| Number of columns | 15 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 14 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ActivityDate | 0 | 1 | 8 | 9 | 0 | 31 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1 | 4.855407e+09 | 2.424805e+09 | 1503960366 | 2.320127e+09 | 4.445115e+09 | 6.962181e+09 | 8.877689e+09 |
| TotalSteps | 0 | 1 | 7.637910e+03 | 5.087150e+03 | 0 | 3.789750e+03 | 7.405500e+03 | 1.072700e+04 | 3.601900e+04 |
| TotalDistance | 0 | 1 | 5.490000e+00 | 3.920000e+00 | 0 | 2.620000e+00 | 5.240000e+00 | 7.710000e+00 | 2.803000e+01 |
| TrackerDistance | 0 | 1 | 5.480000e+00 | 3.910000e+00 | 0 | 2.620000e+00 | 5.240000e+00 | 7.710000e+00 | 2.803000e+01 |
| LoggedActivitiesDistance | 0 | 1 | 1.100000e-01 | 6.200000e-01 | 0 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 4.940000e+00 |
| VeryActiveDistance | 0 | 1 | 1.500000e+00 | 2.660000e+00 | 0 | 0.000000e+00 | 2.100000e-01 | 2.050000e+00 | 2.192000e+01 |
| ModeratelyActiveDistance | 0 | 1 | 5.700000e-01 | 8.800000e-01 | 0 | 0.000000e+00 | 2.400000e-01 | 8.000000e-01 | 6.480000e+00 |
| LightActiveDistance | 0 | 1 | 3.340000e+00 | 2.040000e+00 | 0 | 1.950000e+00 | 3.360000e+00 | 4.780000e+00 | 1.071000e+01 |
| SedentaryActiveDistance | 0 | 1 | 0.000000e+00 | 1.000000e-02 | 0 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 1.100000e-01 |
| VeryActiveMinutes | 0 | 1 | 2.116000e+01 | 3.284000e+01 | 0 | 0.000000e+00 | 4.000000e+00 | 3.200000e+01 | 2.100000e+02 |
| FairlyActiveMinutes | 0 | 1 | 1.356000e+01 | 1.999000e+01 | 0 | 0.000000e+00 | 6.000000e+00 | 1.900000e+01 | 1.430000e+02 |
| LightlyActiveMinutes | 0 | 1 | 1.928100e+02 | 1.091700e+02 | 0 | 1.270000e+02 | 1.990000e+02 | 2.640000e+02 | 5.180000e+02 |
| SedentaryMinutes | 0 | 1 | 9.912100e+02 | 3.012700e+02 | 0 | 7.297500e+02 | 1.057500e+03 | 1.229500e+03 | 1.440000e+03 |
| Calories | 0 | 1 | 2.303610e+03 | 7.181700e+02 | 0 | 1.828500e+03 | 2.134000e+03 | 2.793250e+03 | 4.900000e+03 |

skim\_without\_charts(sleep\_day)

Data summary

|  |  |
| --- | --- |
| Name | sleep\_day |
| Number of rows | 413 |
| Number of columns | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SleepDay | 0 | 1 | 8 | 9 | 0 | 31 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1 | 5.000979e+09 | 2.06036e+09 | 1503960366 | 3977333714 | 4702921684 | 6962181067 | 8792009665 |
| TotalSleepRecords | 0 | 1 | 1.120000e+00 | 3.50000e-01 | 1 | 1 | 1 | 1 | 3 |
| TotalMinutesAsleep | 0 | 1 | 4.194700e+02 | 1.18340e+02 | 58 | 361 | 433 | 490 | 796 |
| TotalTimeInBed | 0 | 1 | 4.586400e+02 | 1.27100e+02 | 61 | 403 | 463 | 526 | 961 |

skim\_without\_charts(weight\_log)

Data summary

|  |  |
| --- | --- |
| Name | weight\_log |
| Number of rows | 67 |
| Number of columns | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| logical | 1 |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 0 | 1 | 8 | 9 | 0 | 31 | 0 |

**Variable type: logical**

| skim\_variable | n\_missing | complete\_rate | mean | count |
| --- | --- | --- | --- | --- |
| IsManualReport | 0 | 1 | 0.61 | TRU: 41, FAL: 26 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1.00 | 7.009282e+09 | 1.950322e+09 | 1.503960e+09 | 6.962181e+09 | 6.962181e+09 | 8.877689e+09 | 8.877689e+09 |
| WeightKg | 0 | 1.00 | 7.204000e+01 | 1.392000e+01 | 5.260000e+01 | 6.140000e+01 | 6.250000e+01 | 8.505000e+01 | 1.335000e+02 |
| WeightPounds | 0 | 1.00 | 1.588100e+02 | 3.070000e+01 | 1.159600e+02 | 1.353600e+02 | 1.377900e+02 | 1.875000e+02 | 2.943200e+02 |
| Fat | 65 | 0.03 | 2.350000e+01 | 2.120000e+00 | 2.200000e+01 | 2.275000e+01 | 2.350000e+01 | 2.425000e+01 | 2.500000e+01 |
| BMI | 0 | 1.00 | 2.519000e+01 | 3.070000e+00 | 2.145000e+01 | 2.396000e+01 | 2.439000e+01 | 2.556000e+01 | 4.754000e+01 |
| LogId | 0 | 1.00 | 1.461772e+12 | 7.829948e+08 | 1.460444e+12 | 1.461079e+12 | 1.461802e+12 | 1.462375e+12 | 1.463098e+12 |

## Exploring Avg sleep time $ Avg time in bed

Avg\_minutes\_asleep <- sleep\_day %>% summarize(avg\_sleeptime = mean(TotalMinutesAsleep))

Avg\_minutes\_asleep

## avg\_sleeptime  
## 1 419.4673

Avg\_TimeInBed <- sleep\_day %>%  
summarize(avg\_TimeInBed = mean(TotalTimeInBed))

Avg\_TimeInBed

## avg\_TimeInBed  
## 1 458.6392

Above exploration showes the average participants stayed up 40 minutes in bed before they fell asleep.

I converted the ActivityDate column to days of the week (Monday-Friday), from the daily\_activity data set

daily\_activity <- daily\_activity %>%   
 mutate(weekday1 = weekdays(as.Date(ActivityDate, "%m/%d/%Y")))  
  
glimpse(daily\_activity)

## Rows: 940  
## Columns: 16  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036…  
## $ ActivityDate <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/…  
## $ TotalSteps <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019…  
## $ TotalDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…  
## $ TrackerDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…  
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ VeryActiveDistance <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…  
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…  
## $ LightActiveDistance <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…  
## $ SedentaryActiveDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ VeryActiveMinutes <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…  
## $ FairlyActiveMinutes <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…  
## $ LightlyActiveMinutes <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, …  
## $ SedentaryMinutes <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…  
## $ Calories <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203…  
## $ weekday1 <chr> "Tuesday", "Wednesday", "Thursday", "Friday",…

daily\_activity$weekday1 <- ordered(daily\_activity$weekday1, levels=c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"))

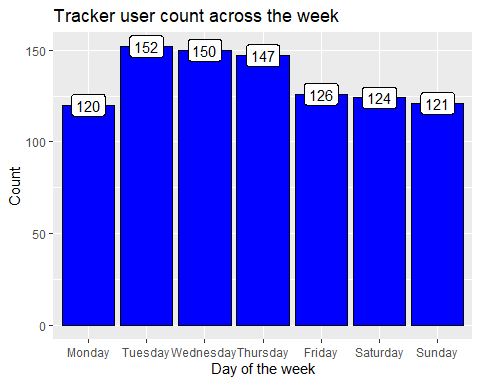
activity\_data <- daily\_activity %>%   
 group\_by(weekday1) %>%   
 summarize(count\_of = n())  
  
glimpse(activity\_data)

## Rows: 7  
## Columns: 2  
## $ weekday1 <ord> Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday  
## $ count\_of <int> 120, 152, 150, 147, 126, 124, 121

## VISUALIZATION

Let’s see the number of times the users track their activities in a week.

ggplot(activity\_data, aes(x=weekday1, y=count\_of)) +  
 geom\_bar(stat="identity",color="black",fill="#0000FF") +  
 labs(title="Tracker user count across the week", x="Day of the week", y="Count") +  
 geom\_label(aes(label=count\_of),color="black")



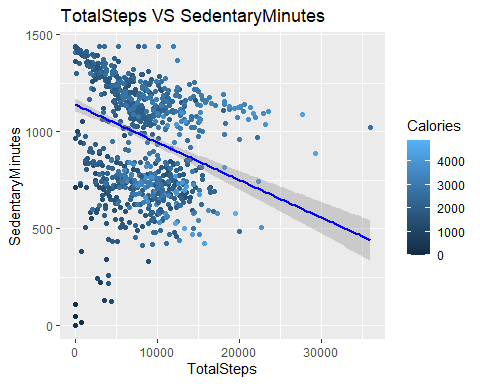
The results shows that most of the activities recorded were on Tuesday, Wednesday and Thursday.

## Plotting a few other explorations.

Let’s find out what’s the relationship between steps taken in a day and sedentary minutes.

ggplot(data=daily\_activity, aes(x=TotalSteps, y=SedentaryMinutes, color=Calories)) + geom\_point() + ggtitle("TotalSteps VS SedentaryMinutes") + geom\_smooth(method="lm",color="blue")

## `geom\_smooth()` using formula = 'y ~ x'



Observing relationship between steps taken and calories burned.

mean\_steps <- mean(daily\_activity$TotalSteps)  
mean\_steps

## [1] 7637.911

mean\_calories <- mean(daily\_activity$Calories)  
mean\_calories

## [1] 2303.61

ggplot(data=daily\_activity, aes(x=TotalSteps, y=Calories, color=Calories)) + geom\_point() + ggtitle("Calories burned for every step taken") + geom\_smooth(method="lm", color="black")

## `geom\_smooth()` using formula = 'y ~ x'

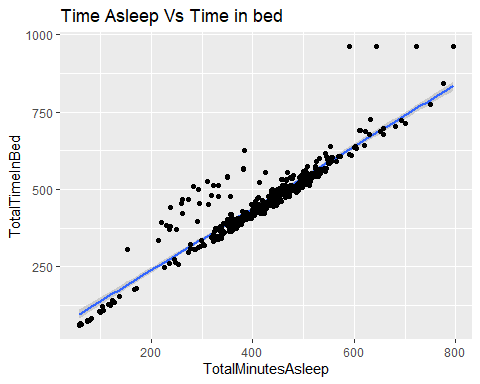


The visualization above shows a positive correlation between the steps taken and the calories burnt.

Let’s find out what’s the relationship between minutes asleep and time in bed.

ggplot(data=sleep\_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom\_point() + ggtitle("Time Asleep Vs Time in bed") + geom\_smooth(method="lm") + geom\_jitter()

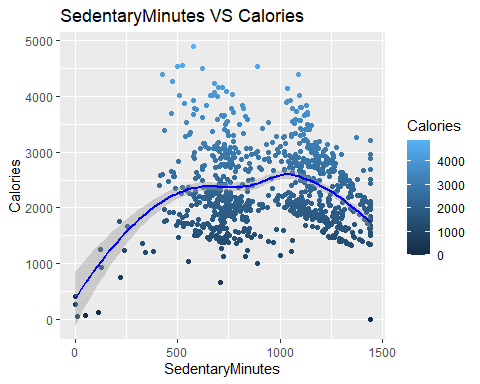
## `geom\_smooth()` using formula = 'y ~ x'



Relationship between being sedentary, light, fairly light, very active minutes and calories burned.

ggplot(data=daily\_activity, aes(x=SedentaryMinutes, y=Calories, color=Calories)) + geom\_point() + ggtitle("SedentaryMinutes VS Calories") + geom\_smooth(method="loess",color="blue")

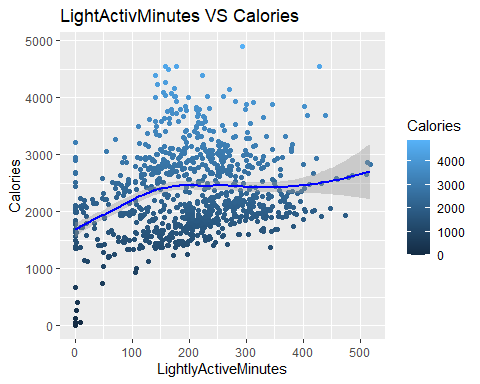
## `geom\_smooth()` using formula = 'y ~ x'



As we can see that there are lesser burned calories as sedentary minutes increased.

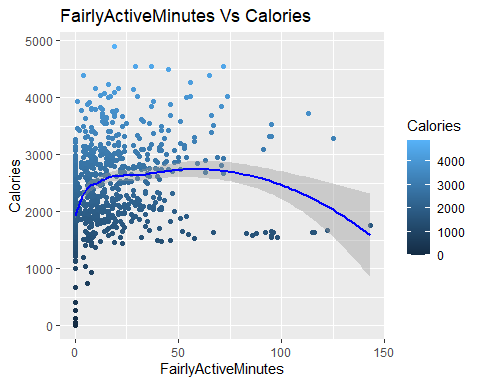
ggplot(data=daily\_activity, aes(x=LightlyActiveMinutes, y=Calories, color=Calories)) + geom\_point() + ggtitle("LightActivMinutes VS Calories") + geom\_smooth(method="loess",color="blue")

## `geom\_smooth()` using formula = 'y ~ x'



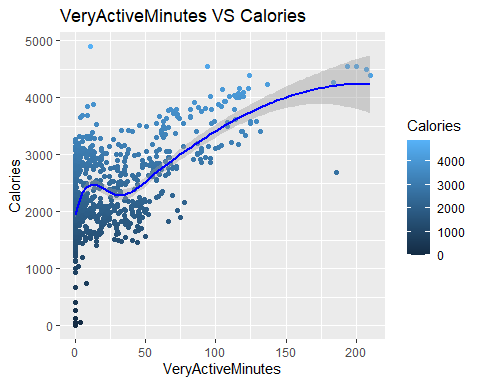
ggplot(data=daily\_activity, aes(x=FairlyActiveMinutes, y=Calories, color=Calories)) + geom\_point() + ggtitle("FairlyActiveMinutes Vs Calories") + geom\_smooth(method="loess",color="blue")

## `geom\_smooth()` using formula = 'y ~ x'



ggplot(data=daily\_activity, aes(x=VeryActiveMinutes, y=Calories, color=Calories)) + geom\_point() + ggtitle("VeryActiveMinutes VS Calories") + geom\_smooth(method="loess",color="blue")

## `geom\_smooth()` using formula = 'y ~ x'



Above vizualization is showing positive correlation between very active minutes and calories burned.