Bellabeat Case Study

Caleb

2023-01-13

#Loading Packages

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

## package 'tidyverse' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\caleb\AppData\Local\Temp\Rtmp6zMBMO\downloaded\_packages

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

## package 'rmarkdown' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\caleb\AppData\Local\Temp\Rtmp6zMBMO\downloaded\_packages

install.packages("ggplot2", repos = "http://cran.us.r-project.org")

## package 'ggplot2' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\caleb\AppData\Local\Temp\Rtmp6zMBMO\downloaded\_packages

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

## package 'dplyr' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'dplyr'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying  
## D:\Program Files\R-4.2.2\library\00LOCK\dplyr\libs\x64\dplyr.dll to D:\Program  
## Files\R-4.2.2\library\dplyr\libs\x64\dplyr.dll: Permission denied

## Warning: restored 'dplyr'

##   
## The downloaded binary packages are in  
## C:\Users\caleb\AppData\Local\Temp\Rtmp6zMBMO\downloaded\_packages

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

## package 'skimr' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\caleb\AppData\Local\Temp\Rtmp6zMBMO\downloaded\_packages

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

## package 'janitor' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\caleb\AppData\Local\Temp\Rtmp6zMBMO\downloaded\_packages

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

## package 'here' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\caleb\AppData\Local\Temp\Rtmp6zMBMO\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\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(lubridate)

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

library(ggplot2)  
library(rmarkdown)  
library(readr)  
library(dplyr)  
library(skimr)  
library(janitor)

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

library(here)

## here() starts at C:/Users/caleb/Documents

#Importing and cleaning datasets ##Loading datasets

daily\_activity\_df <- read\_csv("dailyActivity\_merged.csv")

## Rows: 940 Columns: 15  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityDate  
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDi...  
##   
## 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.

sleep\_day <- read\_csv("sleepDay\_merged.csv")

## Rows: 413 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): SleepDay  
## 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.

##Sorting the Data

daily\_activity\_df\_sorted <- daily\_activity\_df %>% arrange(Id, ActivityDate)

sleep\_day\_sorted <- sleep\_day %>% arrange(Id, SleepDay)

##removing original datasets

remove(daily\_activity\_df)

remove(sleep\_day)

##cleaning names

daily\_activity\_df\_sorted <- clean\_names(daily\_activity\_df\_sorted)

sleep\_day\_sorted <- clean\_names(sleep\_day\_sorted)

##duplicates

daily\_activity\_df\_sorted <- unique(daily\_activity\_df\_sorted)

sleep\_day\_sorted <- unique(sleep\_day\_sorted)

##removing rows with suspiciously low activity (suspected device not used or died early)

daily\_activity\_df\_sorted <- daily\_activity\_df\_sorted %>% filter(total\_steps >= 400, total\_distance > 0, tracker\_distance > 0)

##exporting to fix dates in google sheets

write.csv(sleep\_day\_sorted, "sleep\_day\_sorted.csv")

write.csv(daily\_activity\_df\_sorted, "daily\_activity\_sorted.csv")

#Importing the datasets from google sheets

#importing fixed datasets prepped for merging  
  
activity\_day\_fixed <- read.csv("activity\_day\_fixed.csv")

sleep\_day\_fixed <- read.csv("sleep\_day\_fixed.csv")

#joining datasets

#joining sleep\_day\_fixed and activity\_day\_fixed  
  
activity\_joined <- merge(x = activity\_day\_fixed, y = sleep\_day\_fixed, by = c("date", "id"))

##removing old dataframes

remove(daily\_activity\_df\_sorted)

remove(sleep\_day\_sorted)

##renaming datasets

daily\_activity <- activity\_day\_fixed

remove(activity\_day\_fixed)

sleep\_day <- sleep\_day\_fixed

remove(sleep\_day\_fixed)

#Checking distinct IDs since I forgot earlier

n\_distinct(sleep\_day$id)

## [1] 24

n\_distinct(daily\_activity$id)

## [1] 33

#Making averages datsets for pie charts #making an averages dataframe for each ID in daily\_activity

daily\_averages <- daily\_activity %>%  
 group\_by(id) %>%  
 summarize(steps = mean(total\_steps), distance = mean(total\_distance), very\_active\_distance = mean(very\_active\_distance), moderately\_active\_distance = mean(moderately\_active\_distance),   
 light\_active\_distance = mean(light\_active\_distance), very\_active\_minutes = mean(very\_active\_minutes), fairly\_active\_minutes = mean(fairly\_active\_minutes),   
 lightly\_active\_minutes = mean(lightly\_active\_minutes), sedentary\_minutes = mean(sedentary\_minutes), average\_calories = mean(calories))

#making an averages dataframe for the activity\_joined to include sleep analysis

sleep\_averages <- activity\_joined %>%  
 group\_by(id) %>%  
 summarize(steps = mean(total\_steps), distance = mean(total\_distance), very\_active\_distance = mean(very\_active\_distance), moderately\_active\_distance = mean(moderately\_active\_distance),   
 light\_active\_distance = mean(light\_active\_distance), very\_active\_minutes = mean(very\_active\_minutes), fairly\_active\_minutes = mean(fairly\_active\_minutes),   
 lightly\_active\_minutes = mean(lightly\_active\_minutes), sedentary\_minutes = mean(sedentary\_minutes), average\_calories = mean(calories),  
 average\_min\_asleep = mean(total\_minutes\_asleep), average\_time\_in\_bed = mean(total\_time\_in\_bed))

#Making user classifications based averages

##classifying users based on steps according to <https://www.10000steps.org.au/articles/healthy-lifestyles/counting-steps/> ##this allows me to see how many users are active or not to get an idea of the users

activity\_status <- daily\_averages %>%  
 mutate(activity\_status = case\_when(steps < 5000 ~ "Sedentary", steps >= 5000 & steps <= 7499 ~ "Low Activity", steps >= 7500 & steps <= 9999 ~ "Medium Activity",   
 steps >= 10000 & steps < 12500 ~ "Active", steps >= 12500 ~ "High Activity"))

##daily step analysis - not used in final analysis

daily\_steps <- daily\_activity %>%  
 mutate(activity\_status = case\_when(total\_steps < 5000 ~ "Sedentary", total\_steps >= 5000 & total\_steps <= 7499 ~ "Low Activity",  
 total\_steps >= 7500 & total\_steps <= 9999 ~ "Medium Activity", total\_steps >= 10000 & total\_steps < 12500 ~ "Active",   
 total\_steps >= 12500 ~ "High Activity"))

##Calculating users sleep based on mayo clinic recommendations for adults. 7-9 hours is the recommended.

sleep\_status <- sleep\_averages %>%  
 mutate(sleep\_duration = case\_when(average\_min\_asleep < 420 ~ "Too Little", average\_min\_asleep >= 420 & average\_min\_asleep <= 540 ~ "Good", average\_min\_asleep > 540 ~ "Too Much"))

##daily sleep analysis - not used in final analysis

daily\_sleep <- sleep\_day %>%  
 mutate(sleep\_duration = case\_when(total\_minutes\_asleep < 420 ~ "Too Little", total\_minutes\_asleep >= 420 & total\_minutes\_asleep <= 540 ~ "Good", total\_minutes\_asleep >540 ~ "Too Much"))

#creating another dataset including activity, sleep, and all ratings

#activity joined + sleep duration rating  
  
activity\_join\_plus <- activity\_joined %>%  
 mutate(sleep\_duration = case\_when(total\_minutes\_asleep < 420 ~ "Too Little", total\_minutes\_asleep >= 420 & total\_minutes\_asleep <= 540 ~ "Good", total\_minutes\_asleep > 540 ~ "Too Much")) %>%  
 mutate(activity\_status = case\_when(total\_steps < 5000 ~ "Sedentary", total\_steps >= 5000 & total\_steps <= 7499 ~ "Low Activity",  
 total\_steps >= 7500 & total\_steps <= 9999 ~ "Medium Activity", total\_steps >= 10000 & total\_steps < 12500 ~ "Active",   
 total\_steps >= 12500 ~ "High Activity"))

#creating datasets based on activity and sleep status

##analyzing percentages of sleep and activity to get an idea of who is using these products

##activity\_status using all daily inputs - not used

daily\_activity\_percents <- daily\_steps %>%  
 group\_by(activity\_status) %>%  
 summarize(daily\_activity\_percent = 100 \* n() / nrow(daily\_steps))

##average activity status per user

avg\_activity\_percents <- activity\_status %>%  
 group\_by(activity\_status) %>%  
 summarize(average\_activity\_percent = 100 \* n() / nrow(activity\_status))

## `summarise()` has grouped output by 'activity\_status'. You can override using  
## the `.groups` argument.

#getting percents for sleep trends

##avg sleep % for each user

avg\_sleep\_percent <- sleep\_status %>%  
 group\_by(sleep\_duration) %>%  
 summarize(average\_sleep\_percent = 100 \* n() / nrow(sleep\_status))

##daily sleep % based on all daily inputs - not used

daily\_sleep\_percent <- daily\_sleep %>%  
 group\_by(sleep\_duration) %>%  
 summarize(daily\_sleep\_percent = 100 \* n() / nrow(daily\_sleep))

#rounding the numbers form those percentage sets

#rounding the numbers  
  
avg\_activity\_percents <- avg\_activity\_percents %>% mutate\_if(is.numeric, round, digits = 2)

## `mutate\_if()` ignored the following grouping variables:  
## • Column `activity\_status`

daily\_activity\_percents <- daily\_activity\_percents %>% mutate\_if(is.numeric, round, digits = 2)  
avg\_sleep\_percent <- avg\_sleep\_percent %>% mutate\_if(is.numeric, round, digits = 2)  
daily\_sleep\_percent <- daily\_sleep\_percent %>% mutate\_if(is.numeric, round, digits = 2)

#exporting those datasets to make pie charts in google sheets

write\_csv(avg\_activity\_percents, "avg\_activity\_percents.csv")

write\_csv(avg\_sleep\_percent, "avg\_sleep\_percent.csv")

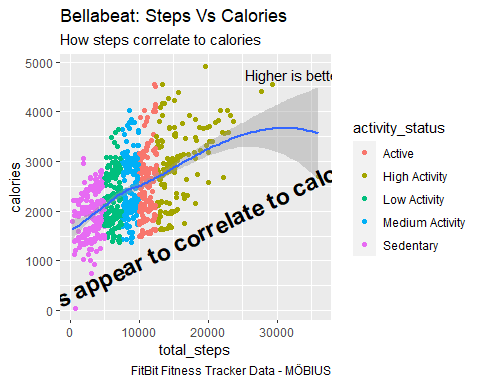
write\_csv(daily\_activity\_percents, "daily\_activity\_percents.csv")

write\_csv(daily\_sleep\_percent, "daily\_sleep\_percent.csv")

#graphing daily steps compared to calories

#graphing daily\_steps  
  
ggplot(data = daily\_steps, mapping = aes(x = total\_steps, y = calories)) +  
 geom\_point(aes(color = activity\_status))+  
 geom\_smooth() +  
 labs(title = "Bellabeat: Steps Vs Calories", subtitle = "How steps correlate to calories", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 17500, y = 1500, label = "Steps appear to correlate to calories", fontface = "bold", angle = 25, size = 6) +  
 annotate("text", x = 32500, y = 4750, label = "Higher is better")

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



#creating long data to graph the different activity classifications

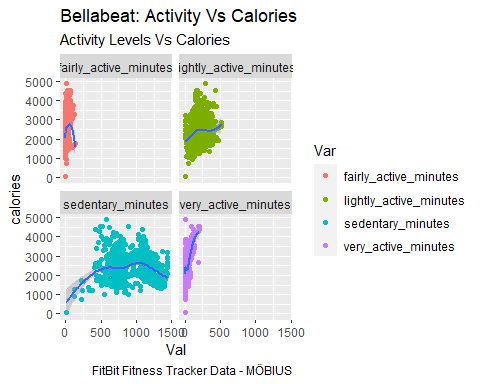
#creating longer data for daily\_steps to graph activity minutes  
  
daily\_steps\_long <- daily\_steps %>%  
 select(id, date, lightly\_active\_minutes, sedentary\_minutes, very\_active\_minutes, fairly\_active\_minutes, calories, activity\_status, total\_steps) %>%  
 pivot\_longer(cols = c(lightly\_active\_minutes, sedentary\_minutes, very\_active\_minutes, fairly\_active\_minutes), names\_to = "Var", values\_to = "Val")

activity\_joined\_plus\_long <- activity\_join\_plus %>%  
 select(id, date, lightly\_active\_minutes, sedentary\_minutes, very\_active\_minutes, fairly\_active\_minutes, calories, total\_minutes\_asleep, sleep\_duration, activity\_status) %>%  
 pivot\_longer(cols = c(lightly\_active\_minutes, sedentary\_minutes, very\_active\_minutes, fairly\_active\_minutes), names\_to = "Var", values\_to = "Val")

#graphing the long data

ggplot(data = daily\_steps\_long, mapping = aes(x = Val, y = calories)) +  
 geom\_point(aes(color = Var)) +  
 geom\_smooth() +  
 facet\_wrap(~Var) +  
 labs(title = "Bellabeat: Activity Vs Calories", subtitle = "Activity Levels Vs Calories", caption = "FitBit Fitness Tracker Data - MÖBIUS")

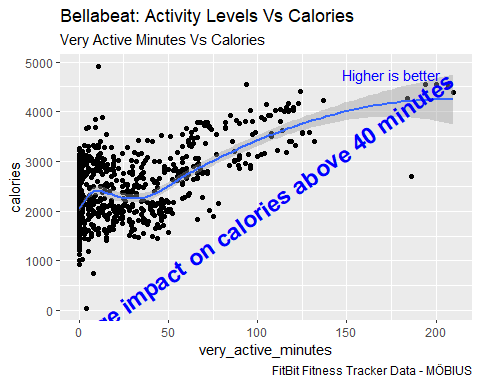
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



#graphing the above charts seperately to zone in ##graphing very active minutes

ggplot(data = daily\_steps, mapping = aes(x = very\_active\_minutes, y = calories)) +  
 geom\_point() +  
 geom\_smooth() +  
 labs(title = "Bellabeat: Activity Levels Vs Calories", subtitle = "Very Active Minutes Vs Calories", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 100, y = 2000, label = "Large impact on calories above 40 minutes", angle = 35, fontface = "bold", size = 6, color = "blue") +  
 annotate("text", x = 175, y = 4750, label = "Higher is better", color = "blue")

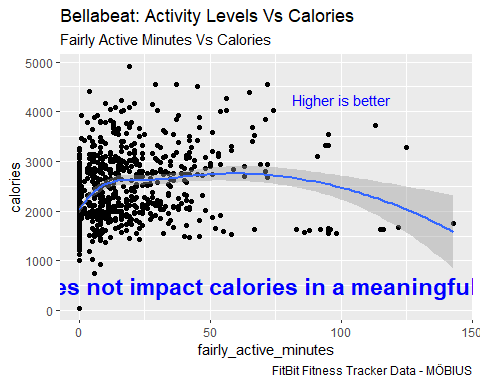
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



##fairly active minutes graph

ggplot(data = daily\_steps, mapping = aes(x = fairly\_active\_minutes, y = calories)) +  
 geom\_point() +  
 geom\_smooth() +  
 labs(title = "Bellabeat: Activity Levels Vs Calories", subtitle = "Fairly Active Minutes Vs Calories", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 75, y = 500, label = "Does not impact calories in a meaningful way", fontface = "bold", size = 6, color = "blue") +  
 annotate("text", x = 100, y = 4250, label = "Higher is better", color = "blue")

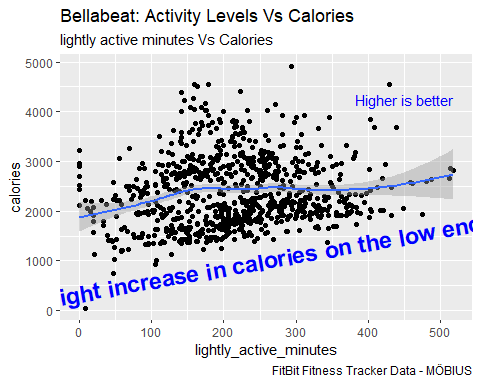
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



##graphing lightly active minutes

ggplot(data = daily\_steps, mapping = aes(x = lightly\_active\_minutes, y = calories)) +  
 geom\_point() +  
 geom\_smooth() +  
 labs(title = "Bellabeat: Activity Levels Vs Calories", subtitle = "lightly active minutes Vs Calories", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 250, y = 1000, label = "Slight increase in calories on the low end", color = "blue", fontface = "bold", angle = 10, size = 6) +  
 annotate("text", x = 450, y = 4250, label = "Higher is better", color = "blue")

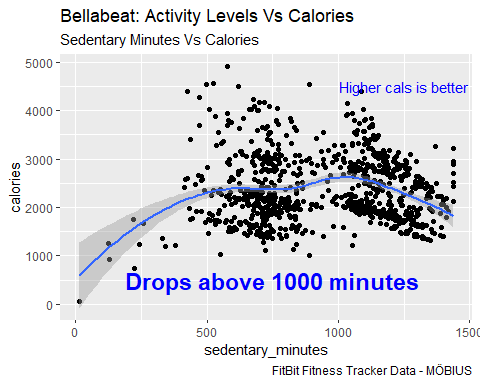
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



##graphing sedentary minutes

ggplot(data = daily\_steps, mapping = aes(x = sedentary\_minutes, y = calories)) +  
 geom\_point() +  
 geom\_smooth() +  
 labs(title = "Bellabeat: Activity Levels Vs Calories", subtitle = "Sedentary Minutes Vs Calories", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 750, y = 500, label = "Drops above 1000 minutes", color = "blue", fontface = "bold", size = 6) +  
 annotate("text", x = 1250, y = 4500, label = "Higher cals is better", color = "blue")

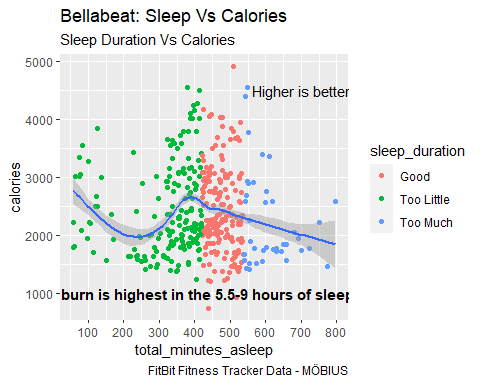
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



#graphing sleep related data ##graphing sleep and calories

ggplot(data = activity\_joined\_plus\_long, mapping = aes(x = total\_minutes\_asleep, y = calories)) +  
 geom\_point(aes(color = sleep\_duration)) +   
 geom\_smooth() +  
 scale\_x\_continuous(breaks = seq(0, 800, by = 100)) +  
 labs(title = "Bellabeat: Sleep Vs Calories", subtitle = "Sleep Duration Vs Calories", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 425, y = 1000, label = "Calorie burn is highest in the 5.5-9 hours of sleep range", size = 4, fontface = "bold") +  
 annotate("text", x = 700, y = 4500, label = "Higher is better")

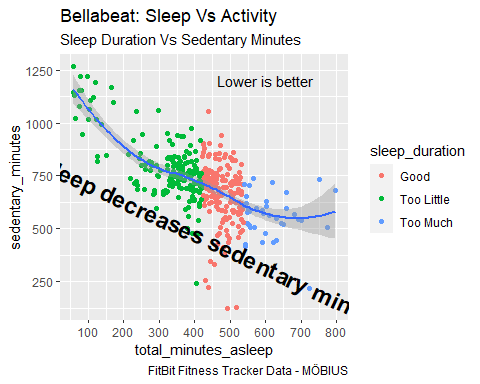
## `geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'



##sleep to sedentary minutes

ggplot(data = activity\_join\_plus, mapping = aes(x = total\_minutes\_asleep, y = sedentary\_minutes)) +  
 geom\_point(aes(color = sleep\_duration)) +  
 geom\_smooth() +  
 scale\_x\_continuous(breaks = seq(0, 800, by = 100)) +  
 labs(title = "Bellabeat: Sleep Vs Activity", subtitle = "Sleep Duration Vs Sedentary Minutes", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 375, y = 500, label = "Better sleep decreases sedentary minutes", size = 6, angle = 335, fontface = "bold") +  
 annotate("text", x = 600, y = 1200, label = "Lower is better")

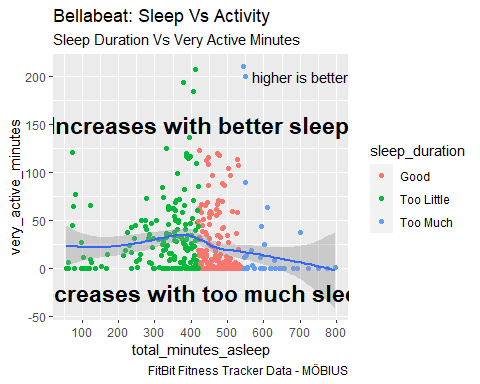
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



##sleep to very active minutes

ggplot(data = activity\_join\_plus, mapping = aes(x = total\_minutes\_asleep, y = very\_active\_minutes)) +  
 geom\_point(aes(color = sleep\_duration)) +  
 geom\_smooth() +  
 scale\_x\_continuous(breaks = seq(0, 800, by = 100)) +  
 labs(title = "Bellabeat: Sleep Vs Activity", subtitle = "Sleep Duration Vs Very Active Minutes", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 425, y = 150, label = "Increases with better sleep", size = 6, fontface = "bold") +  
 annotate("text", x = 425, y = -25, label = "decreases with too much sleep", size = 6, fontface = "bold") +  
 annotate("text", x = 700, y = 200, label = "higher is better")

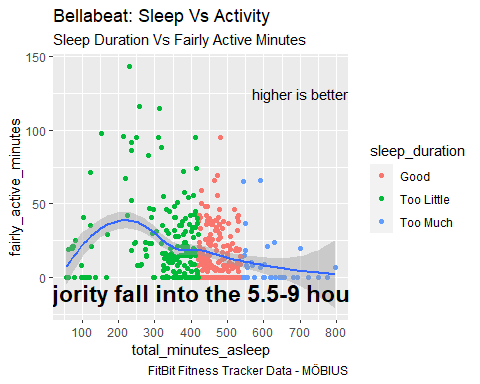
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



##sleep to fairly active minutes

ggplot(data = activity\_join\_plus, mapping = aes(x = total\_minutes\_asleep, y = fairly\_active\_minutes)) +  
 geom\_point(aes(color = sleep\_duration)) +  
 geom\_smooth() +  
 scale\_x\_continuous(breaks = seq(0, 800, by = 100)) +  
 labs(title = "Bellabeat: Sleep Vs Activity", subtitle = "Sleep Duration Vs Fairly Active Minutes", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 425, y = -10, label = "The majority fall into the 5.5-9 hour range", fontface = "bold", size = 6) +  
 annotate("text", x = 700, y = 125, label = "higher is better")

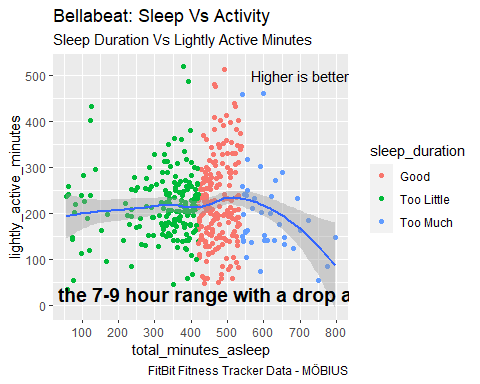
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



##sleep to lightly active minutes

ggplot(data = activity\_join\_plus, mapping = aes(x = total\_minutes\_asleep, y = lightly\_active\_minutes)) +  
 geom\_point(aes(color = sleep\_duration)) +  
 geom\_smooth() +  
 scale\_x\_continuous(breaks = seq(0, 800, by = 100)) +  
 labs(title = "Bellabeat: Sleep Vs Activity", subtitle = "Sleep Duration Vs Lightly Active Minutes", caption = "FitBit Fitness Tracker Data - MÖBIUS") +  
 annotate("text", x = 425, y = 25, label = "Increases in the 7-9 hour range with a drop after 9 hours", fontface = "bold", size = 5) +  
 annotate("text", x = 700, y = 500, label = "Higher is better")

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



#From this point graphs were exported and used in google slides to make a presentation