Ask Phase:

Business Objective:

Bella Beat is a high-tech manufacturer of health-focused products for women. They have the potential to become a larger player in the global smart device market. The company co-founder and chief creative officer Urska Srsen believes that by analyzing smart device data insights information can be derived from consumers which can lead to marketing strategies that can open the door to new growth opportunities for the company.

Primary stakeholders:

1. Urska Srsen

cofounder and creative officer of Bellabeat

2. Sando Mur

mathematician and executive at Bellabeat

3. Bellabeat Marketing Analytics Team

Prepare Data:

Credibility:

The data is not biased as it is a random sample that is generated from both men and women around a similar time range. The sample also consists of at least 30 individuals which is sufficient enough to produce a normal distribution curve for statistical analysis from a public source so its credibility can be validated. Finally, the data is organized and divided in the form of multiple CSV files making it easy to download and analyze using programs like SQL in cloud platforms like Big Query.

Data source: https://www.kaggle.com/datasets/arashnic/fitbit

Process (in SQL):

Cleaning of the data:

#check to determine if multiple records of participants id's exist which might throw off results

SELECT

id,

COUNT(\*) AS num\_records

FROM temp-424017.Smart\_Health\_Device\_Data.Daily\_Activity\_Table

GROUP BY

id

HAVING num\_records >1;

SELECT DISTINCT(Id)

FROM temp-424017.Smart\_Health\_Device\_Data.Daily\_Activity\_Table;

# check to see if any null values exist within the id column which may throw off the results of the analysis

SELECT \*

FROM temp-424017.Smart\_Health\_Device\_Data.Daily\_Activity\_Table

WHERE Id is NULL;

# check to see if the activity date falls within the expected minimum and maximum values as depicted in the description of the dataset

SELECT min(ActivityDate) as minimum\_date, max(ActivityDate) as maximum\_date

FROM temp-424017.Smart\_Health\_Device\_Data.Daily\_Activity\_Table;

# check to see if any of the columns containing desired measurement parameters contain null values which might throw off the results of the analysis

SELECT \*

FROM temp-424017.Smart\_Health\_Device\_Data.Daily\_Activity\_Table

WHERE ActivityDate IS NULL OR VeryActiveDistance IS NULL OR ModeratelyActiveDistance IS NULL OR LightActiveDistance IS NULL OR VeryActiveMinutes IS NULL OR FairlyActiveMinutes IS NULL OR LightlyActiveMinutes IS NULL;

**Analyze(In SQL):**

Dataset 1:

WITH id\_averages AS (

SELECT id, EXTRACT(MONTH FROM ActivityDate) AS ActivityDate, AVG(VeryActiveDistance) AS Avg\_VeryActiveDistance, AVG(ModeratelyActiveDistance) AS Avg\_ModeratelyActiveDistance, AVG(LightActiveDistance) AS Avg\_LightActiveDistance, AVG(VeryActiveMinutes) AS Avg\_VeryActiveMinutes, AVG(FairlyActiveMinutes) AS Avg\_FairlyActiveMinutes, AVG(LightlyActiveMinutes) AS Avg\_LightlyActiveMinutes

FROM temp-424017.Smart\_Health\_Device\_Data.Daily\_Activity\_Table

GROUP BY

id, ActivityDate

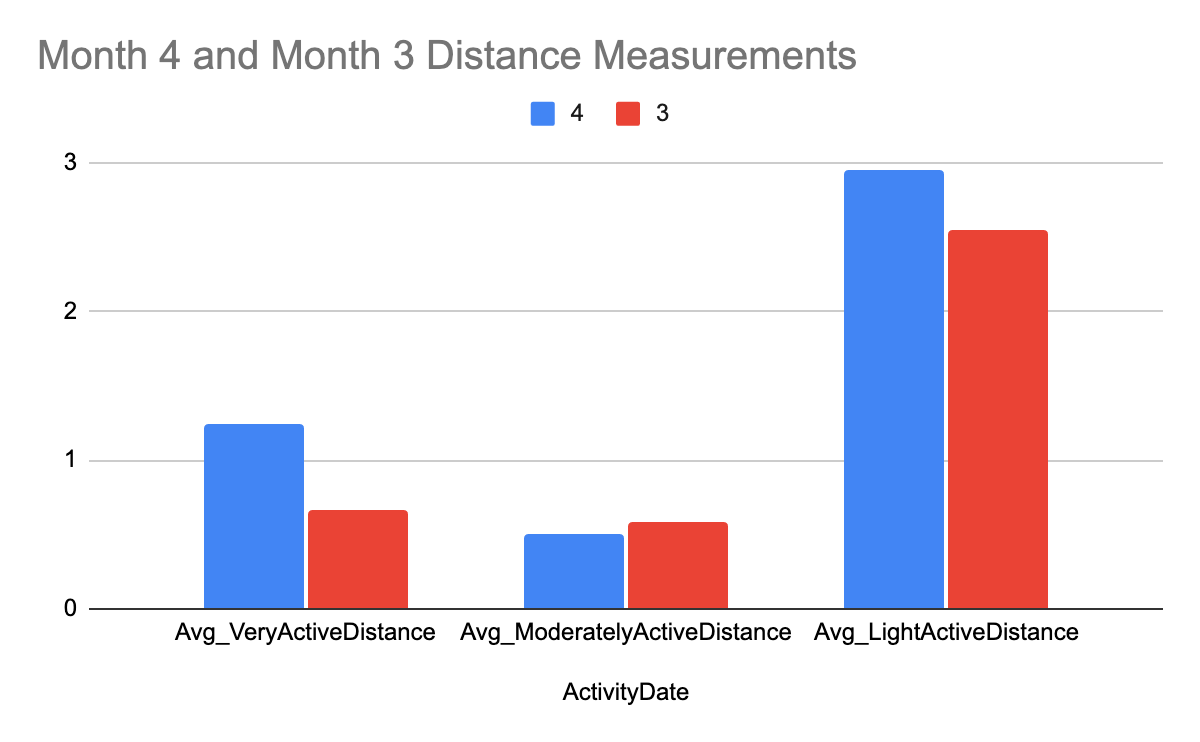
)

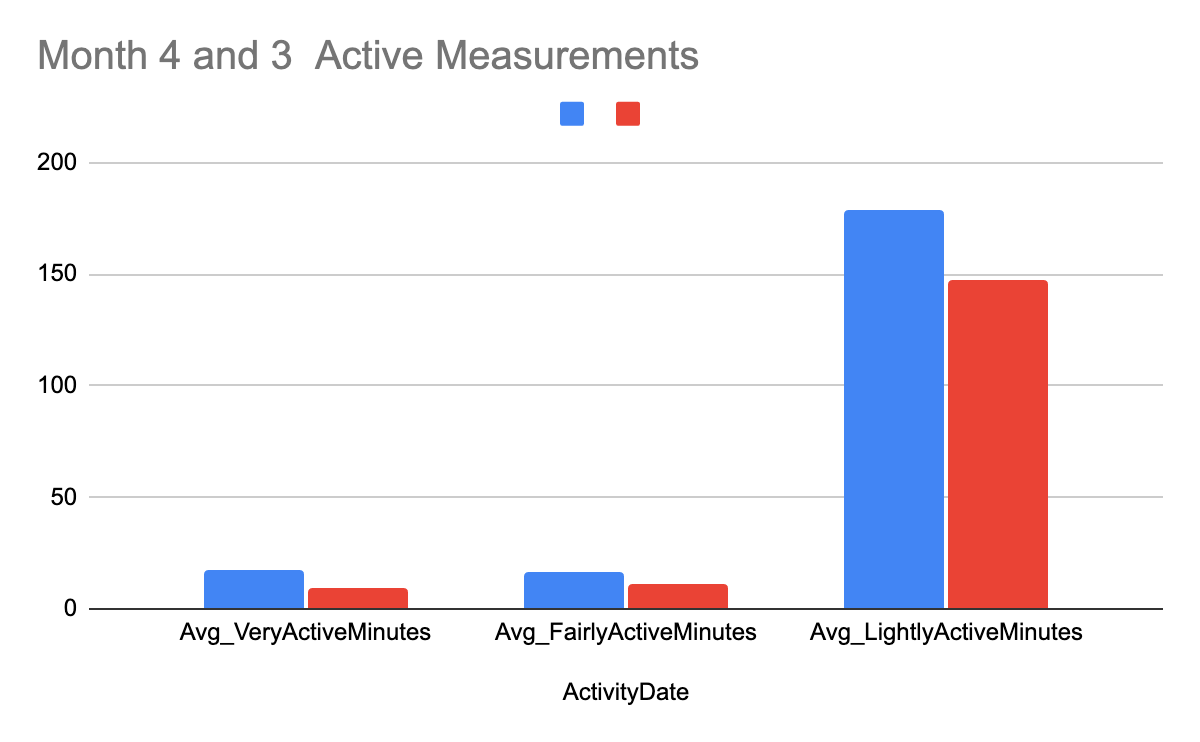
SELECT ActivityDate, AVG(Avg\_VeryActiveDistance) AS Avg\_VeryActiveDistance, AVG(Avg\_ModeratelyActiveDistance) AS Avg\_ModeratelyActiveDistance, AVG(Avg\_LightActiveDistance) AS Avg\_LightActiveDistance, AVG(Avg\_VeryActiveMinutes) AS Avg\_VeryActiveMinutes, AVG(Avg\_FairlyActiveMinutes) AS Avg\_FairlyActiveMinutes, AVG(Avg\_LightlyActiveMinutes) AS Avg\_LightlyActiveMinutes

FROM id\_averages

GROUP BY ActivityDate

ORDER BY Avg\_VeryActiveDistance DESC, Avg\_ModeratelyActiveDistance DESC, Avg\_LightActiveDistance DESC, Avg\_VeryActiveMinutes DESC, Avg\_FairlyActiveMinutes DESC, Avg\_LightlyActiveMinutes DESC





Explanation of analysis: The graphs indicate smart device users are more active in May than they are in March based on the amount of time they spend being active and the amount of distance they travel during those months. Bellabeat should make use of this data by assuming that users are more active in summer months than they are in colder months and should organize their marketing efforts to be more aggressive in colder months than in warmer months in an effort to generate more consumer purchases.

PROCESS( In R):

1. Cleaning and Making sure the Data is intact so it can be analyzed

install.packages('tidyverse')

library(tidyverse)

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

head(daily\_activity)

# checking to make sure that all of the names of the columns included in the data set are in the data table that was uploaded onto the database

colonames(daily\_activity)

# importing the hourlyCalories and hourlySteps data onto the database

hourlyCalories<-read.csv("hourlycalories\_merged.csv")

hourlySteps<-read.csv("hourlySteps\_merged.csv")

#Viewing the first several rows of the data table to get a good understanding of what the data is like

head(daily\_activity)

Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitiesDistance

1 1503960366 3/25/2016 11004 7.11 7.11 0

2 1503960366 3/26/2016 17609 11.55 11.55 0

3 1503960366 3/27/2016 12736 8.53 8.53 0

4 1503960366 3/28/2016 13231 8.93 8.93 0

5 1503960366 3/29/2016 12041 7.85 7.85 0

Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitiesDistance

1 1503960366 3/25/2016 11004 7.11 7.11 0

2 1503960366 3/26/2016 17609 11.55 11.55 0

3 1503960366 3/27/2016 12736 8.53 8.53 0

4 1503960366 3/28/2016 13231 8.93 8.93 0

5 1503960366 3/29/2016 12041 7.85 7.85 0

6 1503960366 3/30/2016 10970 7.16 7.16 0

VeryActiveDistance ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance

1 2.57 0.46 4.07 0

2 6.92 0.73 3.91 0

3 4.66 0.16 3.71 0

4 3.19 0.79 4.95 0

5 2.16 1.09 4.61 0

6 2.36 0.51 4.29 0

VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories

1 33 12 205 804 1819

2 89 17 274 588 2154

3 56 5 268 605 1944

4 39 20 224 1080 1932

5 28 28 243 763 1886

6 30 13 223 1174 1820

> head(hourlyCalories)

Id ActivityHour Calories

1 1503960366 3/12/2016 12:00:00 AM 48

2 1503960366 3/12/2016 1:00:00 AM 48

3 1503960366 3/12/2016 2:00:00 AM 48

4 1503960366 3/12/2016 3:00:00 AM 48

5 1503960366 3/12/2016 4:00:00 AM 48

6 1503960366 3/12/2016 5:00:00 AM 48

colnames(hourlyCalories)

[1] "Id" "ActivityHour" "Calories"

# making sure that the number of distinct rows match up with the number of expected data points to ensure no duplicates exist

> n\_distinct(daily\_activity)

[1] 457

> hourlySteps<-read.csv(hourlySteps\_merged.csv)

> head(hourlySteps)

Id ActivityHour StepTotal

1 1503960366 3/12/2016 12:00:00 AM 0

2 1503960366 3/12/2016 1:00:00 AM 0

3 1503960366 3/12/2016 2:00:00 AM 0

4 1503960366 3/12/2016 3:00:00 AM 0

5 1503960366 3/12/2016 4:00:00 AM 0

6 1503960366 3/12/2016 5:00:00 AM 0

> colnames(hourlySteps)

[1] "Id" "ActivityHour" "StepTotal"

> n\_distinct(daily\_activity$ID)

[1] 0

> n\_distinct(daily\_activity$ID)

[1] 0

> n\_distinct(hourlyCalories$ID)

[1] 0

> n\_distinct(hourlySteps$ID)

[1] 0

> nrow(daily\_activity)

[1] 457

> nrow(hourlyCalories)

[1] 24084

> nrow(hourlySteps)

[1] 24084

#summarize the data to get a pretty good idea of what magnitude of the parameters you will be extracting from it

> daily\_activity %>% select(ActivityDate, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes) %>%

+ summary()

ActivityDate VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes

Length:457 Min. : 0.00 Min. : 0.00 Min. : 0.0

Class :character 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 64.0

Mode :character Median : 0.00 Median : 1.00 Median :181.0

Mean : 16.62 Mean : 13.07 Mean :170.1

3rd Qu.: 25.00 3rd Qu.: 16.00 3rd Qu.:257.0

Max. :202.00 Max. :660.00 Max. :720.0

> daily\_activity %>% select(ActivityDate, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes) %>%

+ summary()

ActivityDate VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes

Length:457 Min. : 0.00 Min. : 0.00 Min. : 0.0

Class :character 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 64.0

Mode :character Median : 0.00 Median : 1.00 Median :181.0

Mean : 16.62 Mean : 13.07 Mean :170.1

3rd Qu.: 25.00 3rd Qu.: 16.00 3rd Qu.:257.0

Max. :202.00 Max. :660.00 Max. :720.0

> daily\_activity %>%

+ select(VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, SedentaryActiveDistance) %>%

+ summary()

VeryActiveDistance ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance

Min. : 0.000 Min. :0.0000 Min. : 0.00 Min. :0.000000

1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.: 0.87 1st Qu.:0.000000

Median : 0.000 Median :0.0200 Median : 2.93 Median :0.000000

Mean : 1.181 Mean :0.4786 Mean : 2.89 Mean :0.001904

3rd Qu.: 1.310 3rd Qu.:0.6700 3rd Qu.: 4.46 3rd Qu.:0.000000

Max. :21.920 Max. :6.4000 Max. :12.51 Max. :0.100000

> daily\_activity%>%

+ select(VeryActiveMinutes, FairlyActiveMinutes, LightActiveMinutes, SedentaryActiveMinutes) %>%

+ summary()

> daily\_activity%>%

+ select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryActiveMinutes) %>%

+ summary()

Error in `select()`:

! Can't subset columns that don't exist.

✖ Column `SedentaryActiveMinutes` doesn't exist.

Run `rlang::last\_trace()` to see where the error occurred.

> daily\_activity%>%

+ select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes)%>%

+ summary()

VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes

Min. : 0.00 Min. : 0.00 Min. : 0.0 Min. : 32.0

1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 64.0 1st Qu.: 728.0

Median : 0.00 Median : 1.00 Median :181.0 Median :1057.0

Mean : 16.62 Mean : 13.07 Mean :170.1 Mean : 995.3

3rd Qu.: 25.00 3rd Qu.: 16.00 3rd Qu.:257.0 3rd Qu.:1285.0

Max. :202.00 Max. :660.00 Max. :720.0 Max. :1440.0

> #people prefer to spend their time being less active than more active

> #people prefer to spend lighter distances traveling too but have a slight tendency to travel wider distances than to spend time being active

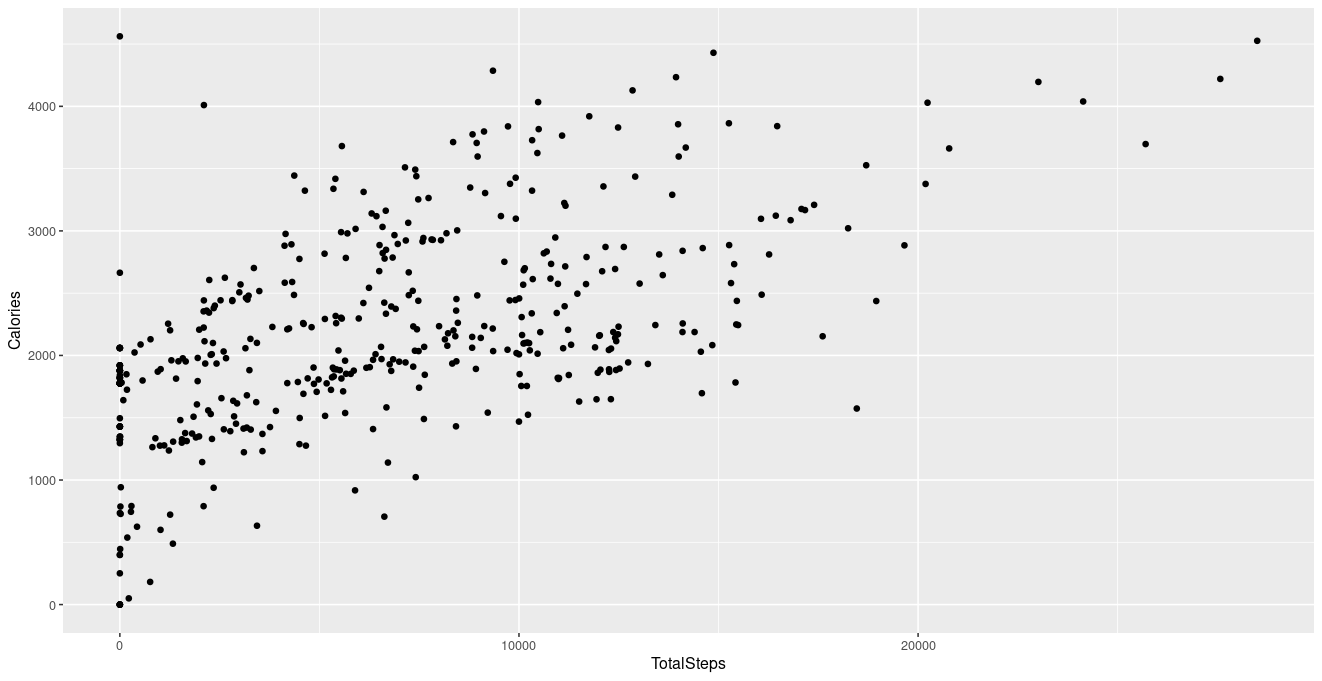
> hourlyCalories %>%

+ select(ActivityHour, Calories)%

Analyze (In R):

# Create a graph of calories vs. total steps

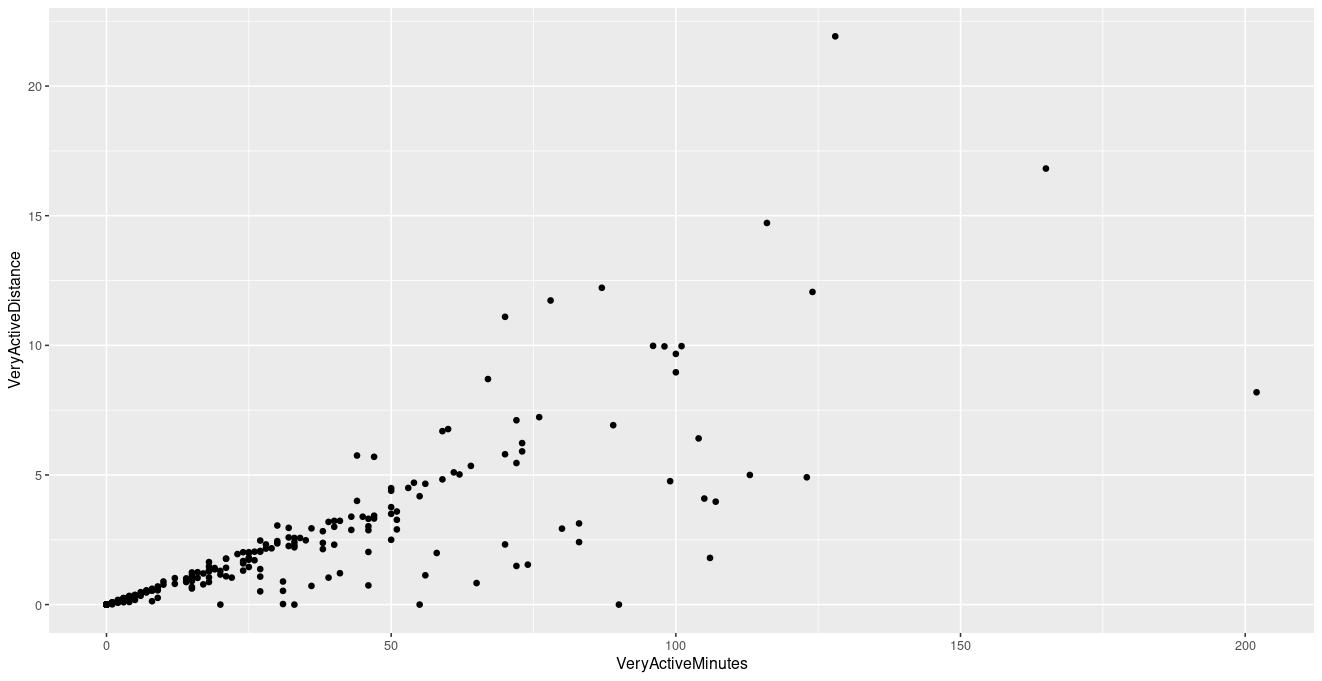
ggplot(data=daily\_activity,aes(x=TotalSteps, y=Calories)) +geom\_point()



Explanation of analysis: Both the scatter plot and line graph demonstrate a strong correlation between the number of steps taken and the number of calories consumed and as a result, if users want to burn calories they should be encouraged to be mindful of the number of steps they are taking for the day. Because the demand for weight loss products is a rapidly growing trend, marketing campaigns about smart devices should both explain the relationship between step-taking and calorie expenditure and how users can make use of the features to get the most out of both activities.

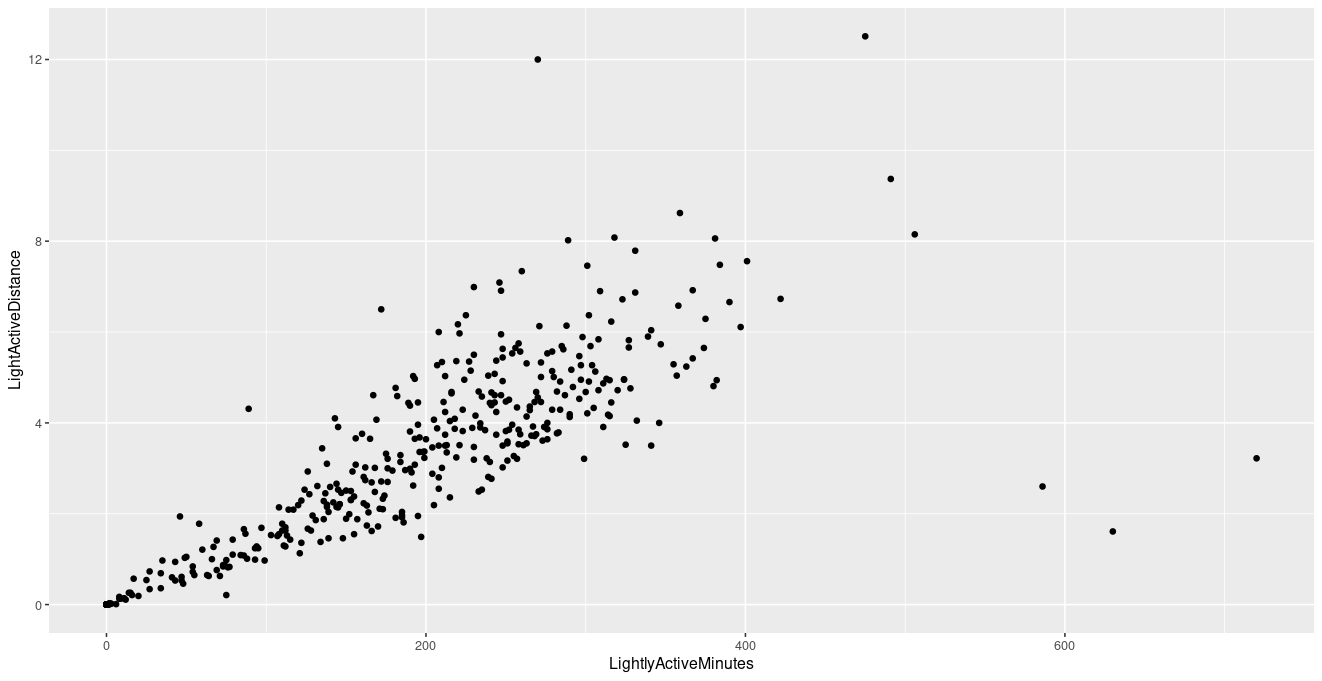
#create a graph of fairly active minutes vs. moderately active distance in the form of a scatter plot

ggplot(data=daily\_activity, aes(x=FairlyActiveMinutes, y=ModeratelyActiveDistance)) +geom\_point



#create a graph of light active minutes vs. light active distance in the form of a scatter plot

ggplot(data=daily\_activity, aes(x=LightlyActiveMinutes, y=LightActiveDistance)) +geom\_point()



Explanation of analysis: Both graphs indicate that users enjoy spending more time engaging in activities that require minimal physical activity. Analysis of both graphs indicates that smart device users like to engage in physical activities that are low in intensity and require movement between short distances. This can be indicated by high concentrations of scatter plots skewing towards the right side of the graph in Figure 2 vs higher concentrations of scatter plots skewed to the left in Figure 1.

The best way to make use of this analysis is to identify the consumer segment that regularly uses smart devices as those that prefer activities and sports that don't require much physical exertion and distance movement (pickleball, water polo, jogging, golf, nature walking, etc). Bellabeat should identify users who demonstrate their interest in these activities and exclusively market towards these types of consumers.