Case Study: How Can a Wellness Technology Company Play It Smart?

Scenario: - You are a junior data analyst working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. You 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. The insights you discover will then help guide marketing strategy for the company. You will present your analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat's marketing strategy.

Characters and products: -

● **Characters**

○ **Urška Sršen**: Bellabeat's cofounder and Chief Creative Officer

○ **Sando Mur**: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team

○ **Bellabeat marketing analytics team**: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy. You joined this team six months ago and have been busy learning about Bellabeat''s mission and business goals — as well as how you, as a junior data analyst, can help Bellabeat achieve them.

● **Products**

○ **Bellabeat app**: The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions. The Bellabeat app connects to their line of smart wellness products.

○ **Leaf**: Bellabeat's classic wellness tracker can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.

○ **Time**: This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. The Time watch connects to the Bellabeat app to provide you with insights into your daily wellness.

○ **Spring**: This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. The Spring bottle connects to the Bellabeat app to track your hydration levels.

○ **Bellabeat membership**: Bellabeat also offers a subscription-based membership program for users. Membership gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

❖ ASK

Business Task: -

To analyze smart device data usage to gain insight into how consumers use Bella beat smart devices. Identify the trends in any one of product & report to marketing team.

Key Stakeholders: -

1) **Urška Sršen**: Bellabeat's cofounder and Chief Creative Officer
2) **Sando Mur**: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive tea

❖ PREPARE

I used Kaggle Public dataset named as 'Fitbit Fitness Tracker Data'. Dataset contains personal fitness tracker data of thirty Bellabeat's users. It includes information about daily activity, steps, and heart rate.

Dataset stored in My computer. Total 18 .csv files available. This data files include weight, sleep data, calories. Data is from 2016 so I can't trust data as latest or real. Data integrity is not fulfilled.

Does my data ROCCC?

ROCCC means Reliable, Original, Comprehensive, Current, Cited.

Reliable- LOW – Not sure about gender.

Original- LOW- It's survey-based data.

Comprehensive- MEDIUM- Data contain most required variable.

Current- LOW- Data is from 2016. Which is 8 years old.

Cited- LOW- As data is collected from survey. So cannot take decision based on that.

❖ PROCESS

Process steps include Data cleaning, sorting & filtering and make sure all data is correct, complete, relevant, and error free.

Steps I followed during Data cleaning: -
# Explore and observe data
# Identify null values if any
# Transform data- format data type
# Perform preliminary statistical analysis.

Tools Used: - R Studio

As data is huge & wide 18 csv files available. So, I can’t use Excel for cleaning.

R used because It’s easy to clean, transform and visualize the data.

Data Selection

Data Frame Creation: I have chosen 5 files to Analyze data and create objects. Files name as: -
1) Daily Activity
2) Daily Calories
3) Weight Log Information
4) Daily Intensities
5) Daily Sleep

After downloading dataset from Kaggle Same has been uploaded in Posit cloud (R studio)

R Packages installed & loaded

Import Packages

```
install.packages("tidyverse")
library(tidyverse)
install.packages("janitor")
library(janitor)
install.packages("skimr")
library(skimr)
install.packages("lubridate")
library(lubridate)
install.packages("sqldf")
library(sqldf)
install.packages("plotrix")
library(plotrix)
install.packages("janitor")
library(janitor)
library(ggplot2)
library(gsubfn)
```

Import 5 selected files for Data cleaning & transformation.

```
## Importing datasets using read_csv function
```

```
dailyactivity <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv")

dailyCalories<- read_csv("Fitabase Data 4.12.16-5.12.16/dailyCalories_merged.csv")

dailyIntensities<- read_csv("Fitabase Data 4.12.16-5.12.16/dailyIntensities_merged.csv")

sleepDay <- read_csv("Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv")

weightLog <- read_csv("Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv")
```

❖ Analyze

Data Exploration

I have explored all 5 datasets & try to find out any common variable/Column name.

```
> colnames(dailyactivity)
 [1] "Id"                         "ActivityDate"
 [3] "TotalSteps"                 "TotalDistance"
 [5] "TrackerDistance"            "LoggedActivitiesDistance"
 [7] "VeryActiveDistance"         "ModeratelyActiveDistance"
 [9] "LightActiveDistance"        "SedentaryActiveDistance"
[11] "VeryActiveMinutes"          "FairlyActiveMinutes"
[13] "LightlyActiveMinutes"       "SedentaryMinutes"
[15] "Calories"
> colnames(dailycalories)
[1] "Id"           "ActivityDay" "Calories"
> colnames(dailyIntensities)
 [1] "Id"                         "ActivityDay"
 [3] "SedentaryMinutes"           "LightlyActiveMinutes"
 [5] "FairlyActiveMinutes"        "VeryActiveMinutes"
 [7] "SedentaryActiveDistance"    "LightActiveDistance"
 [9] "ModeratelyActiveDistance" "VeryActiveDistance"
> colnames(sleepday)
[1] "Id"                   "SleepDay"             "TotalSleepRecords"
[4] "TotalMinutesAsleep" "TotalTimeInBed"
> colnames(WeightLog)
[1] "Id"              "Date"              "WeightKg"        "WeightPounds"
[5] "Fat"             "BMI"               "IsManualReport" "LogId"
```

Output: -

All five datasets have one common variable/column that is “ID”. Now I need to confirm whether all datasets have same observations or different?? Let’s find out.

```
nrow(dailyactivity)
[1] 940
> nrow(dailycalories)
[1] 940
> nrow(dailyIntensities)
[1] 940
> nrow(sleepday)
[1] 413
> nrow(WeightLog)
[1] 67
```

Out of 5 datasets only three datasets have same observations. So, I cannot combine all datasets.
dailyactivity, dailycalories, dailyIntensities datasets have same observations.

Now need to check these three dataset values are same or different? for that I need to create a temporary data frame to see if the two data frames have different values of columns.

```
# Created Temporary Dataframe
dailyact <- dailyactivity %>%select(Id, ActivityDate, Calories)
head(dailyact)
          Id      ActivityDate  Calories
1 1503960366   4/12/2016    1985
2 1503960366   4/13/2016    1797
3 1503960366   4/14/2016    1776
4 1503960366   4/15/2016    1745
5 1503960366   4/16/2016    1863
6 1503960366   4/17/2016    1728
```

Let’s check values in both datasets are equal or not.

```
sqlcheck<-sqldf('SELECT*FROM dailyact INTERSECT SELECT *FROM dailycalories')
head(sqlcheck)
          Id      ActivityDate  Calories
1 1503960366   4/12/2016    1985
2 1503960366   4/13/2016    1797
3 1503960366   4/14/2016    1776
4 1503960366   4/15/2016    1745
5 1503960366   4/16/2016    1863
6 1503960366   4/17/2016    1728
```

From above result we can say that Both DailyActivity & Dailycalories have same values.
Let’s check for DailyIntensities dataset for more confirmation.

```
dailyact1 <- dailyactivity %>%select(Id, ActivityDate, SedentaryActiveDistance, S
edentaryMinutes, FairlyActiveMinutes, ModeratelyActiveDistance, LightlyActiveMinu
tes, VeryActiveMinutes, LightActiveDistance, VeryActiveDistance)
> head(dailyact1)
          Id ActivityDate SedentaryActiveDistance SedentaryMinutes
1 1503960366    4/12/2016                       0              728
2 1503960366    4/13/2016                       0              776
3 1503960366    4/14/2016                       0             1218
4 1503960366    4/15/2016                       0              726
5 1503960366    4/16/2016                       0              773
6 1503960366    4/17/2016                       0              539
  FairlyActiveMinutes ModeratelyActiveDistance LightlyActiveMinutes
1                  13                     0.55                  328
2                  19                     0.69                  217
3                  11                     0.40                  181
4                  34                     1.26                  209
5                  10                     0.41                  221
6                  20                     0.78                  164
  VeryActiveMinutes LightActiveDistance VeryActiveDistance
1                25                6.06               1.88
2                21                4.71               1.57
3                30                3.91               2.44
4                29                2.83               2.14
5                36                5.04               2.71
6                38                2.51               3.19
```

```
sqlcheck1<-sqldf('SELECT*FROM dailyact1 INTERSECT SELECT *FROM dailyIntensities')

> head(sqlcheck1)

        Id ActivityDate SedentaryActiveDistance SedentaryMinutes

1 1503960366   4/12/2016            0          728

2 1503960366   4/13/2016            0          776

3 1503960366   4/14/2016            0         1218

4 1503960366   4/15/2016            0          726

5 1503960366   4/16/2016            0          773

6 1503960366   4/17/2016            0          539

 FairlyActiveMinutes ModeratelyActiveDistance LightlyActiveMinutes

1            13              0.55           328

2            19              0.69           217
```

```
3          11          0.40          181

4          34          1.26          209

5          10          0.41          221

6          20          0.78          164

 VeryActiveMinutes LightActiveDistance VeryActiveDistance

1          25          6.06          1.88

2          21          4.71          1.57

3          30          3.91          2.44

4          29          2.83          2.14

5          36          5.04          2.71

6          38          2.51          3.19
```

Now three datasets like DailyActivity, DataIntensities & DataCalories has same data. Let’s check other datasets to find out unique ID (ID which are not repetitive)

```
n_distinct(WeightLog$Id)
[1] 8
> n_distinct(dailyactivity$Id)
[1] 33
> n_distinct(dailyIntensities$Id)
[1] 33
> n_distinct(sleepday$Id)
[1] 24
```

At start we read 30 users has submitted survey. But from above 33 users are visible. That means duplicate users are registered. Also, only 24 users submitted sleepday data & 8 users submitted weight data.

Let’s examine the summary statistics of some data frame by using SQL.

```
 # Summary statistics of data frame dailyactivity
```

```
dailyactivity %>% 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  
```

```
# Summary statistics of data frame Sleepday

> sleepday %>% select(TotalMinutesAsleep, TotalTimeInBed) %>% summary()
 TotalMinutesAsleep TotalTimeInBed 
 Min.   : 58.0      Min.   : 61.0  
 1st Qu.:361.0      1st Qu.:403.0  
 Median :433.0      Median :463.0  
 Mean   :419.5      Mean   :458.6  
 3rd Qu.:490.0      3rd Qu.:526.0  
 Max.   :796.0      Max.   :961.0  

# Summary statistics of data frame WeightLog
> WeightLog %>% select(WeightKg, BMI) %>% summary()
    WeightKg           BMI       
 Min.   : 52.60   Min.   :21.45  
 1st Qu.: 61.40   1st Qu.:23.96  
 Median : 62.50   Median :24.39  
 Mean   : 72.04   Mean   :25.19  
 3rd Qu.: 85.05   3rd Qu.:25.56  
 Max.   :133.50   Max.   :47.54  
```

From above summary, average Total minutes of sleep is 419.5 i.e., 6.99 Hr. According to CDC (Centers for disease control and prevention) for adult 7 or more hours of sleep necessary. That means half of the woman needed more sleep hours.

Source: - https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html

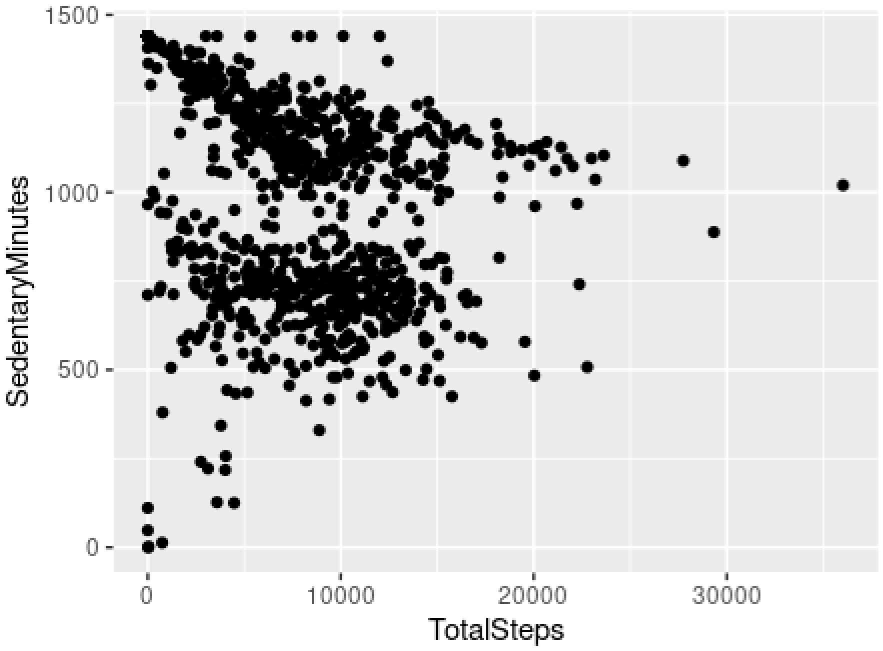
According to DailyActivity dataset, approximately the top 25% of females are walking enough steps per day. As recommended by the CDC, an adult female must aim to walk at least 10,000 steps per day to benefit from general health benefits.

Source: - https://www.cdc.gov/physicalactivity/basics/adults/index.htm

❖ Share

Let’s create visualization to find out relation between Totalsteps vs SedentaryMinutes

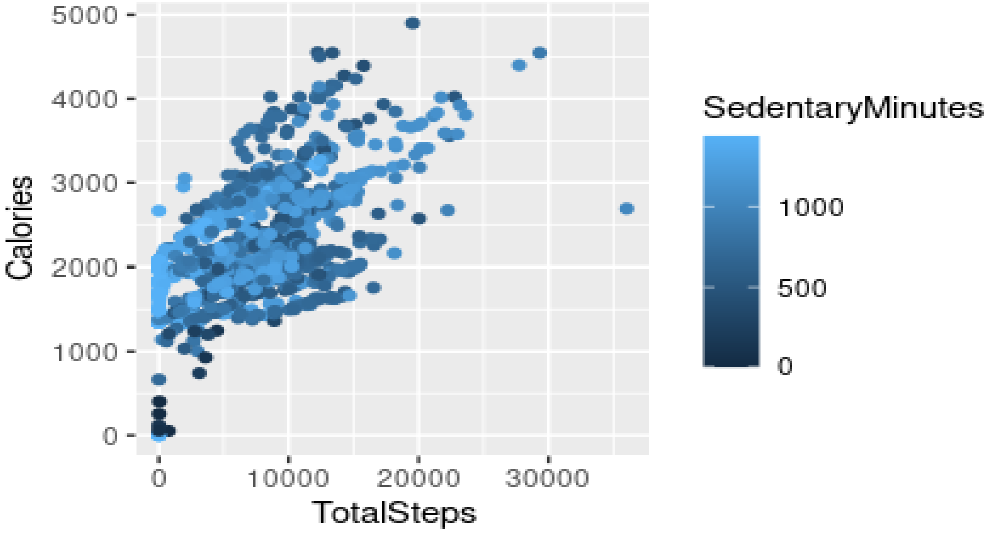
```
install.packages("ggplot2")
library(ggplot2)

ggplot(data = dailyactivity)+ geom_point(mapping = aes(x= TotalSteps, y= SedentaryMinutes))
```



Scatterplot between Sedentary Minutes and Total Steps is a negative relationship. This is an important starting point.

Now I am going to plot the relationship between Total Steps taken and calories burned.

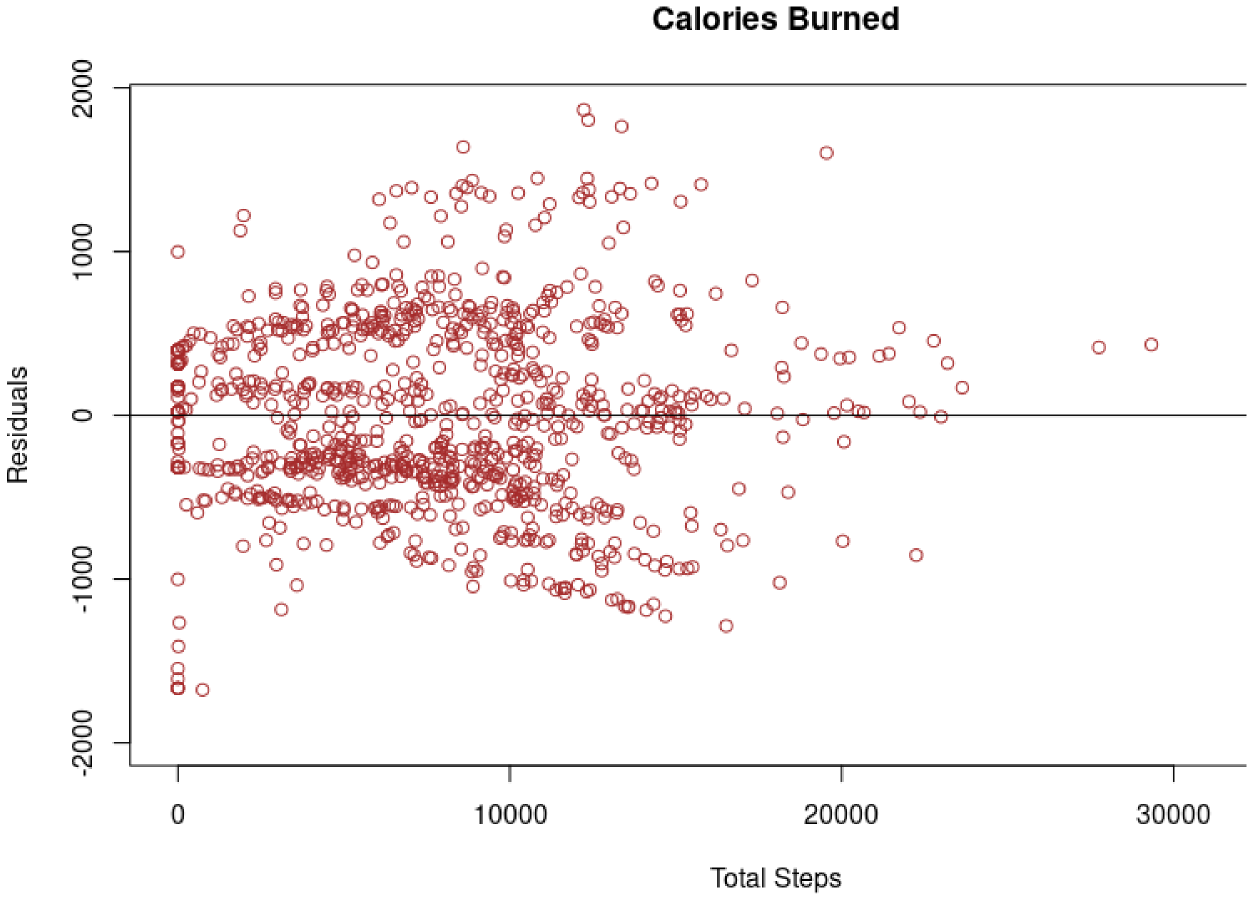
```
ggplot(data= dailyactivity)+ geom_point(mapping = aes(x= TotalSteps, y= Calories, color= SedentaryMinu
tes))
```



From above plot I conclude as steps increases calories burn more. I have use aesthetic for sedentaryMinutes. Positive relationship between TotalSteps & Calories is visible.

Now I am going to check the differences between the observed values and the estimated value.

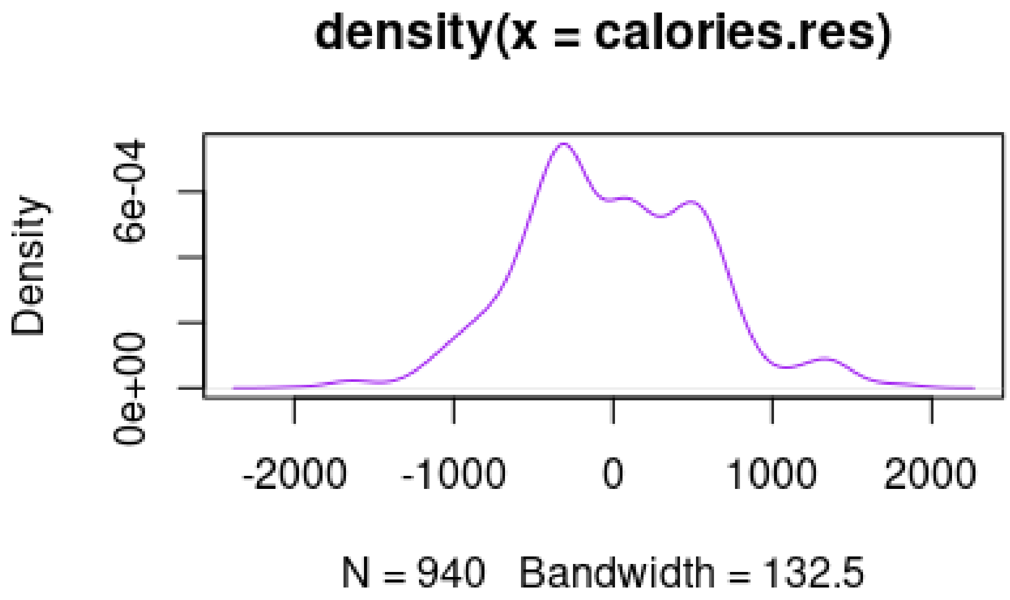
```
calories.lm<-lm(Calories~TotalSteps,data=dailyactivity)
> calories.res<-resid(calories.lm)
> plot(dailyactivity$TotalSteps, calories.res, ylab="Residuals", xlab="Total Steps", main= "Calories Burne
d")
> plot(dailyactivity$TotalSteps, calories.res, ylab="Residuals", xlab="Total Steps", main= "Calories Burne
d", col="brown")
 abline(0,0)
```



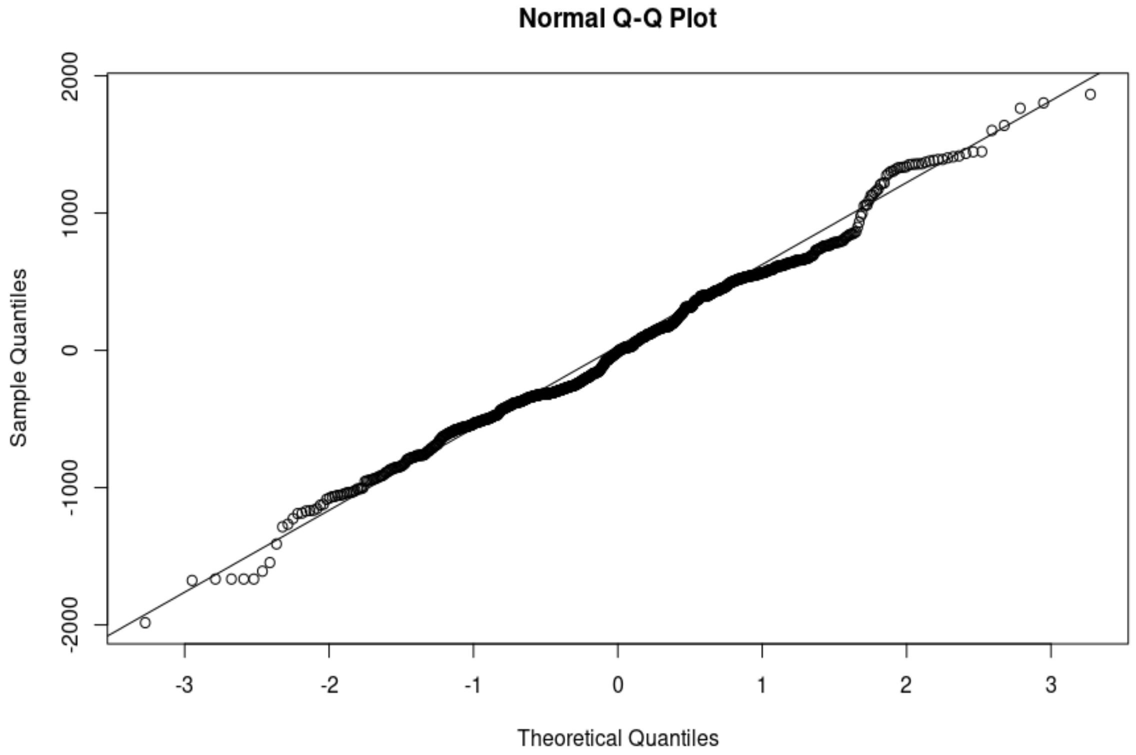
Finally, I am going to graph the density of the residuals and check for normality to determine how statistically far the spread is.

```
#Density plot of the residuals
```

```
plot(density(calories.res), col="Purple")
```



```
# Normality Q-Q plot
qqnorm(calories.res)
qqline(calories.res)
```



Most of the points lie on a straight line then we can say the distribution is normally distributed. As data is normally distributed, we can do further statistical testing such as Pearson correlations to find any statistical significance.

Combine Dataset

As we find out 3 datasets has same data & variable ‘Id’. So, we can combine the dataset to plot a graph to find out relationship between them.

#Merging dailyactivity and sleepday dataset by field 'Id'

```
dailyactivity_Sleepday<-merge(dailyactivity,sleepday, by="Id")
> head(dailyactivity_Sleepday)
```

```
   Id       ActivityDate  TotalSteps  TotalDistance  TrackerDistance
1 1503960366   5/7/2016     11992        7.71           7.71
2 1503960366   5/7/2016     11992        7.71           7.71
3 1503960366   5/7/2016     11992        7.71           7.71
4 1503960366   5/7/2016     11992        7.71           7.71
5 1503960366   5/7/2016     11992        7.71           7.71
6 1503960366   5/7/2016     11992        7.71           7.71
  LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
1                  0           2.46                 2.12
2                  0           2.46                 2.12
3                  0           2.46                 2.12
4                  0           2.46                 2.12
5                  0           2.46                 2.12
6                  0           2.46                 2.12
  LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
1             3.13                  0            37
2             3.13                  0            37
3             3.13                  0            37
4             3.13                  0            37
5             3.13                  0            37
6             3.13                  0            37
  FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
1              46               175           833    1821
2              46               175           833    1821
3              46               175           833    1821
4              46               175           833    1821
5              46               175           833    1821
6              46               175           833    1821
             SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
1 4/12/2016 12:00:00 AM              1            327          346
2 4/13/2016 12:00:00 AM              2            384          407
3 4/15/2016 12:00:00 AM              1            412          442
4 4/16/2016 12:00:00 AM              2            340          367
```

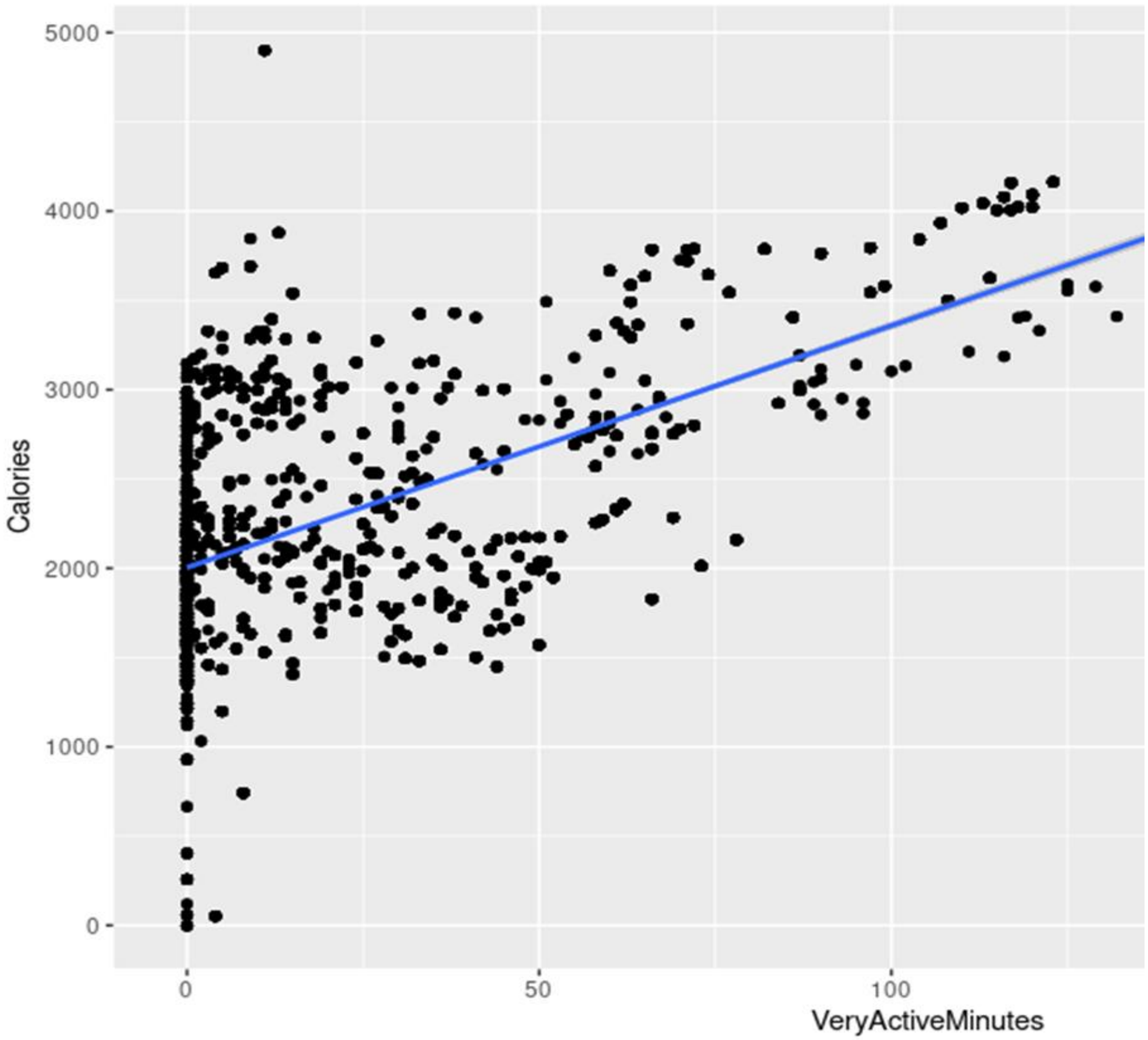
```
5 4/17/2016 12:00:00 AM           1           700          712
6 4/19/2016 12:00:00 AM           1           304          320
```

```
n_distinct(dailyactivity_Sleepday$Id)
[1] 24
```

Since only 24 users registered their sleep date data there are only 24 unique Ids instead of 33.

#Very Active Minutes V Calories burned

```
ggplot(data=dailyactivity_Sleepday, aes(x=VeryActiveMinutes, y=Calories)) + geom_point() +
stat_smooth(method = lm)
```



There is a strong positive correlation between very active minutes and calories burned. Let’s run a correlation to see what the correlation coefficient would be for linear regression.

```
#Linear Regression for Very Active Minutes and Calories Burned
VeryActive.lm<-lm(VeryActiveMinutes~Calories,data=dailyactivity_Sleepday)
> VeryActive.lm
Call:
lm(formula = VeryActiveMinutes ~ Calories, data = dailyactivity_Sleepday)

Coefficients:
(Intercept)     Calories  
  -42.25454      0.02843  

#Pearson Co-relation
cor(dailyactivity_Sleepday$VeryActiveMinutes,dailyactivity_Sleepday$Calories,meth
od="pearson")
[1] 0.6206555
```

Hence 0.62 > 0.05 the co-relation is not statistically correct. Any co-relation is due to random chance or other determining factors. Let’s check final analysis to determine if there is a statistically significant relationship between Sedentary Minutes and Calories burned.

```
#Linear Regression for Sedentary Minutes and Calories Burned
Sedentary.lm<-lm(SedentaryMinutes~Calories,data=dailyactivity_Sleepday)
> Sedentary.lm

Call:
lm(formula = SedentaryMinutes ~ Calories, data = dailyactivity_Sleepday)

Coefficients:
(Intercept)     Calories  
  837.41893     -0.01641  

#Pearson Co-relation
cor(dailyactivity_Sleepday$SedentaryMinutes,dailyactivity_Sleepday$Calories,metho
d="pearson")
[1] -0.04687715
```

Hence, -0.04< 0.05 I can conclude that there is a statistically significant relationship between the number of Sedentary Minutes and the number of calories burned in a day.

❖ ACT

This is the last step of Data Analysis. Here we will answer business questions & making recommendations based on insights discovered in our analysis or according to trends.

Any trends identified?

1) There was a statistically significant negative relationship between Sedentary Minutes and Calories burned
2) Users preferred to track activity data with 33 unique Ids compared to 24 for sleep data and 8 for weight data.

How could these trends help influence Bella beat marketing strategy?

1) Bella beat's marketing team can encourage users to be more active by setting daily goals such as 10,000 steps per day or calories burned. Also, they can set alarm to motivate users to stay active.

2) Encourage users to be more active by giving a notification when you have been sedentary for an allotted time (i.e., prompting users to exercise after being Sedentary for more than 6 hours while awake). This will also encourage users to track other data such as sleep data since most users preferred to track sedentary minutes.

3) Give weekly statistical data about steps, calories burned, weight, sleep hours so user can analyze & work accordingly

4) Daily/Weekly Data sharing on social media handles like Twitter (X), Instagram, Facebook, WhatsApp can encourage users to stay more Active & competitive.