# Fitbase_Analysis (Final Project for Google Data Analytics Certificate

Ranjani Ramanathan

2023-11-28

## INTRODUCTION

Fitbase in a pioneering fitness application that will track different metrics of a person's health and activity so they can keep track of their habits and find insights to improve their health. Data was collected from the Fitbit application (similar type of application) to see trends with historical data of fitness habits. The data was analysed and the unique observations were used to make recommendations on how to enhance user experiences for the Fitbase application.

## ASK

How do intensity, distance and steps affect the outcome of calorie burn?

With these insights, I can highlight what healthy habits can burn calories and lose weight. Also, identifying the most effective healthy habits can market products specifically catered to those. Finally, these statistics can give recommendations to people on how to optimize their fitness.

## PREPARE

The data is stored in multiple files in the same directory as the R markdown. The Activity data that I will use for most of my analysis is wide, because there are lots of columns for each of the entry numbers. However, the Sleep and Heartbeat data seem narrow because they only track one variable. I plan to mostly study the daily data, and hourly data in specific circumstances. Bias can happen if the process is self selecting, people who choose to report their data do so. The key word is consent, and the people who consent are a self selecting group. However, this bias is inevitable because it is a violation of right to report someone's usage without their consent. The data seems to pass the ROCC except it is from 2016, so it is considered historical data since we do not have current data on hand. Although it would be a good idea to collect new data in the future.

Possible problems: No units for distance

## PROCESS

The steps used to clean data and make sure it is in the best format for the analysis I need.

```
install.packages("here")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library("here")
```

```
## here() starts at /cloud/project
```

```r
install.packages("skimr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
library("skimr")
install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
library("janitor")
```

```
##
## Attaching package: 'janitor'
```

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

```r
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library("tidyverse")
```

Steps to process so the data is ready to analyze: 1) Add a total active minutes in activities 2) The tracked distance and the total distance is different for 15 rows (hopefully clean by filter)

```
#Filters out the rows where tracked activities are different from actual activities
activities = read.csv("dailyActivity_merged.csv")
nrow(activities)
```

```
## [1] 940
```

```
activities <- filter(activities, activities$TrackerDistance - activities$TotalDistance == 0)
nrow(activities)
```

```
## [1] 925
```

```
head(activities)
```

```
##           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
```

```
#Mutate to get a new column with all active minutes combined

activities_new <-activities%>%mutate(TotalActiveMinutes = LightlyActiveMinutes+ FairlyActiveMinutes + V
activities<-activities_new
ncol(activities)
```

```
## [1] 16
```

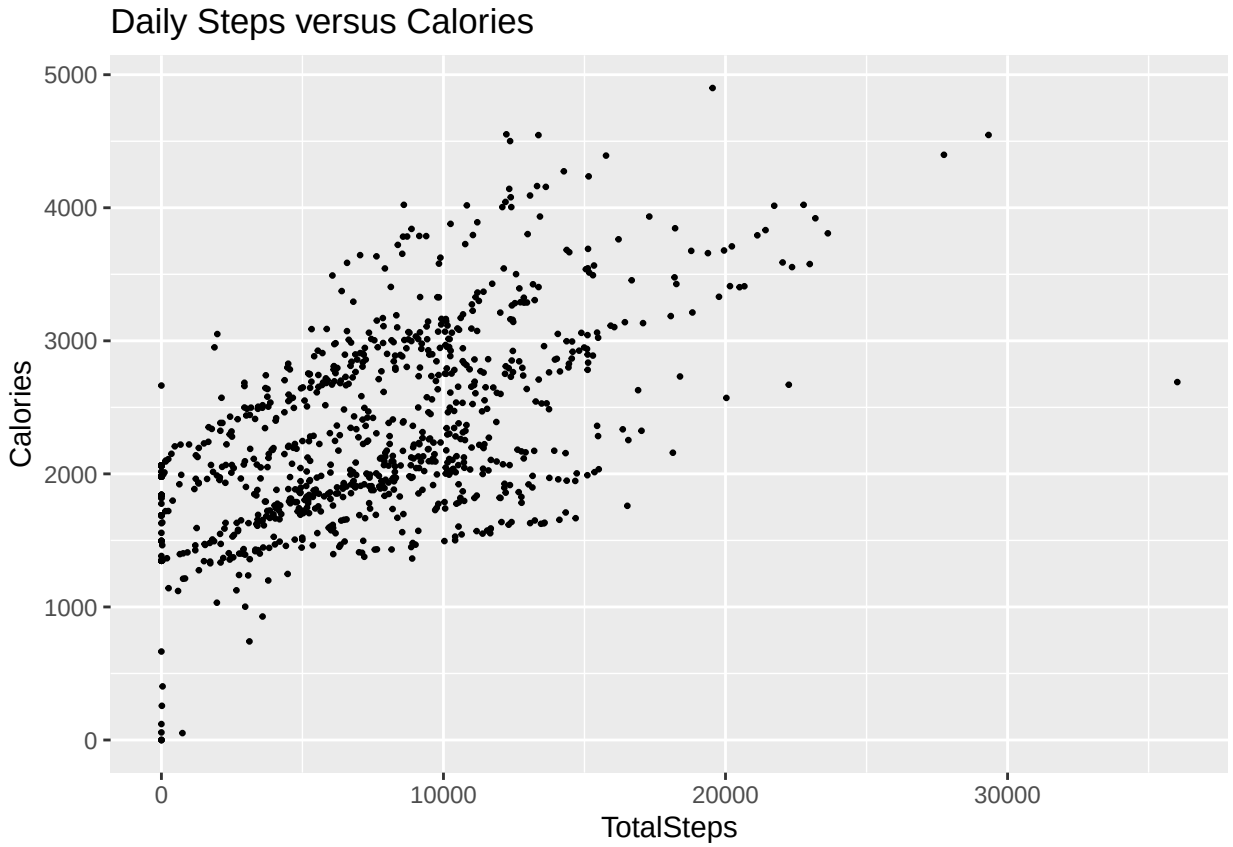
```
head(activities)
```

```
##           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
##   TotalActiveMinutes
## 1                366
## 2                257
## 3                222
## 4                272
## 5                267
## 6                222
```

## ANALYZE PART 1

This part of analysis deals with how intensity, steps and distance correlate to calorie burn

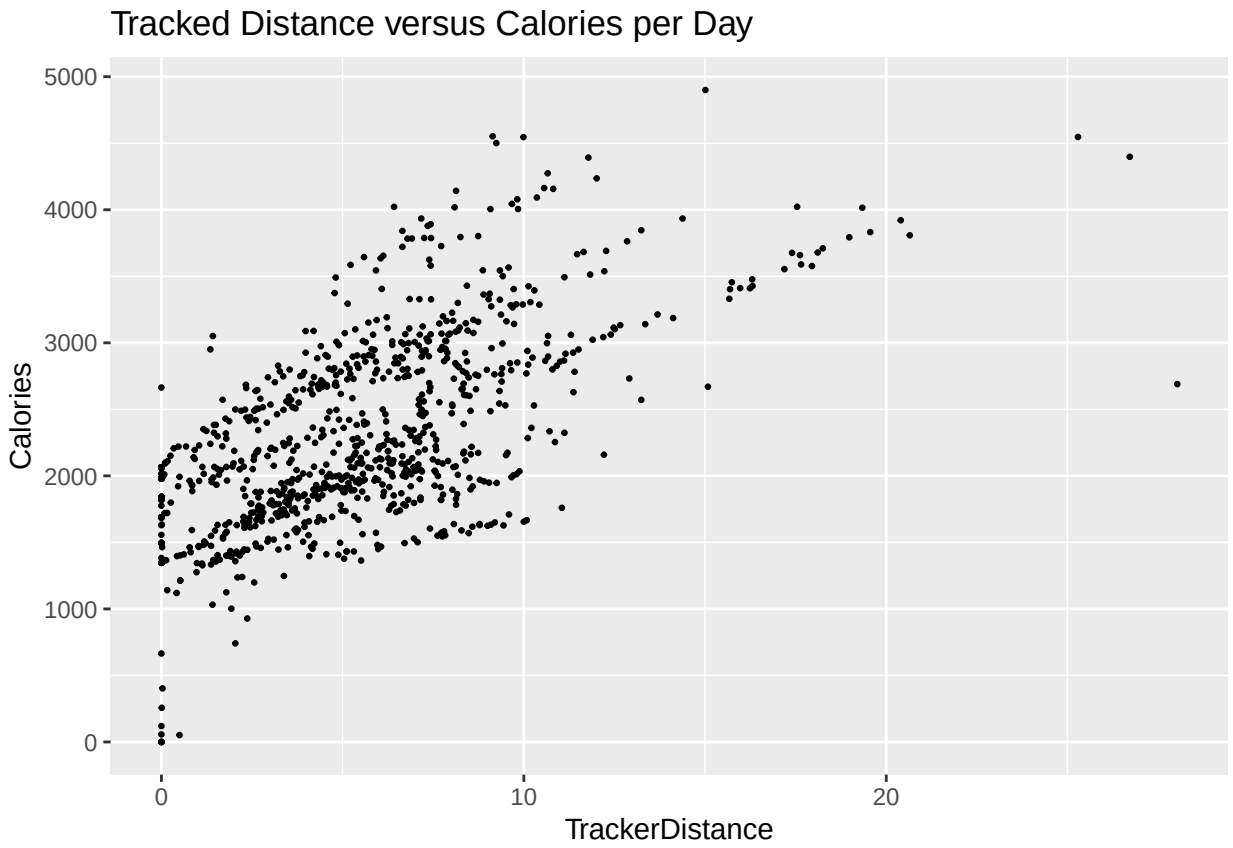
```
#Do a scatter plot with steps on x axis and calorie bun on y
ggplot(data = activities) +
geom_point(mapping = aes(x = TotalSteps, y = Calories), size=0.5) + labs(title = "Daily Steps versus Ca
```




```
cor(activities$TotalSteps, activities$Calories)
```

```
## [1] 0.5881113
```

```
#Do a scatter plot with distance on x axis and calorie bun on y
ggplot(data = activities) +
geom_point(mapping = aes(x = TrackerDistance, y = Calories), size=0.5) + labs(title = "Tracked Distance
```

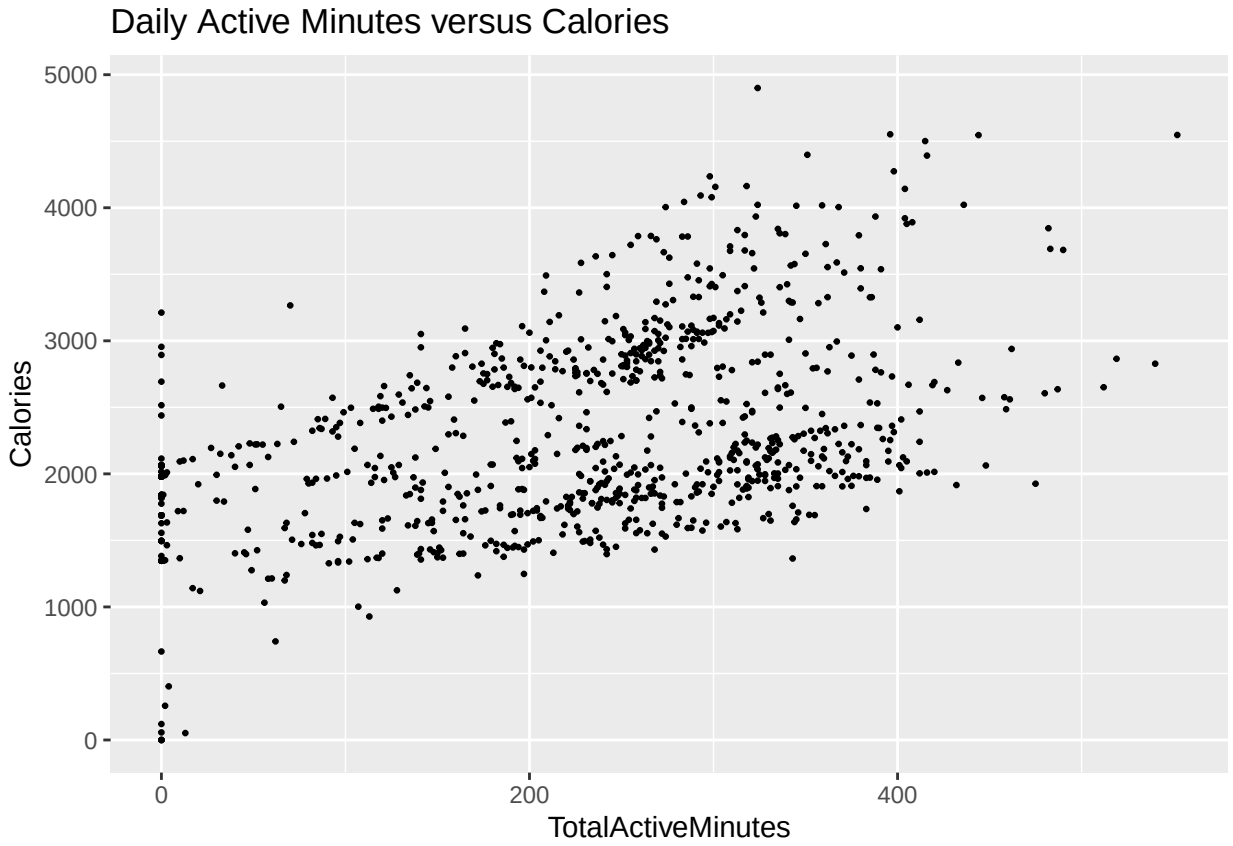



```
cor(activities$TrackerDistance, activities$Calories)
```

```
## [1] 0.6425057
```

When comparing steps to distance, the distance is more an indicator of how many calories are burned, because its correlation coefficient, 0.64, is greater than steps at 0.58

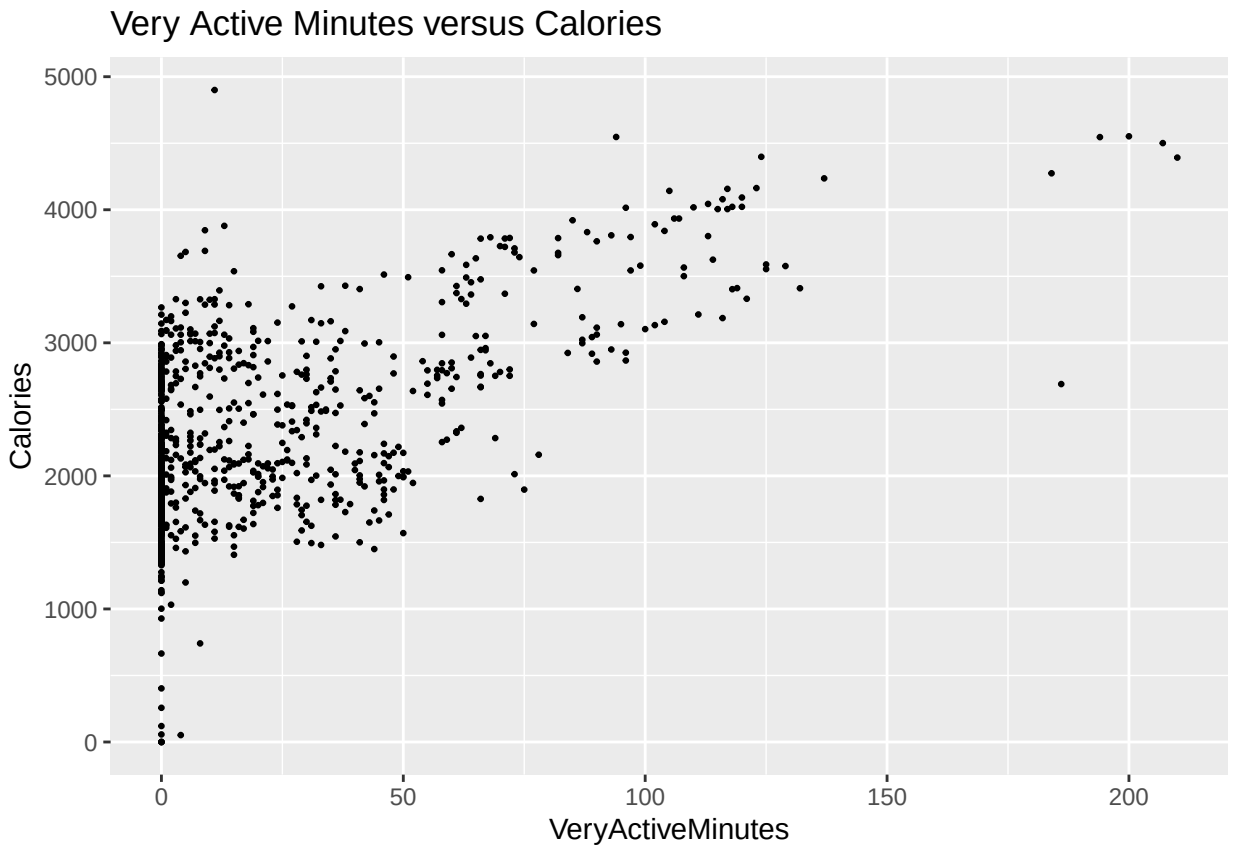
```
#Do a scatter plot with total active on x axis and calorie burn on y
ggplot(data = activities) +
geom_point(mapping = aes(x = TotalActiveMinutes, y = Calories), size=0.5)  + labs(title = "Daily Active
```




```
cor(activities$TotalActiveMinutes, activities$Calories)
```

```
## [1] 0.4655258
```

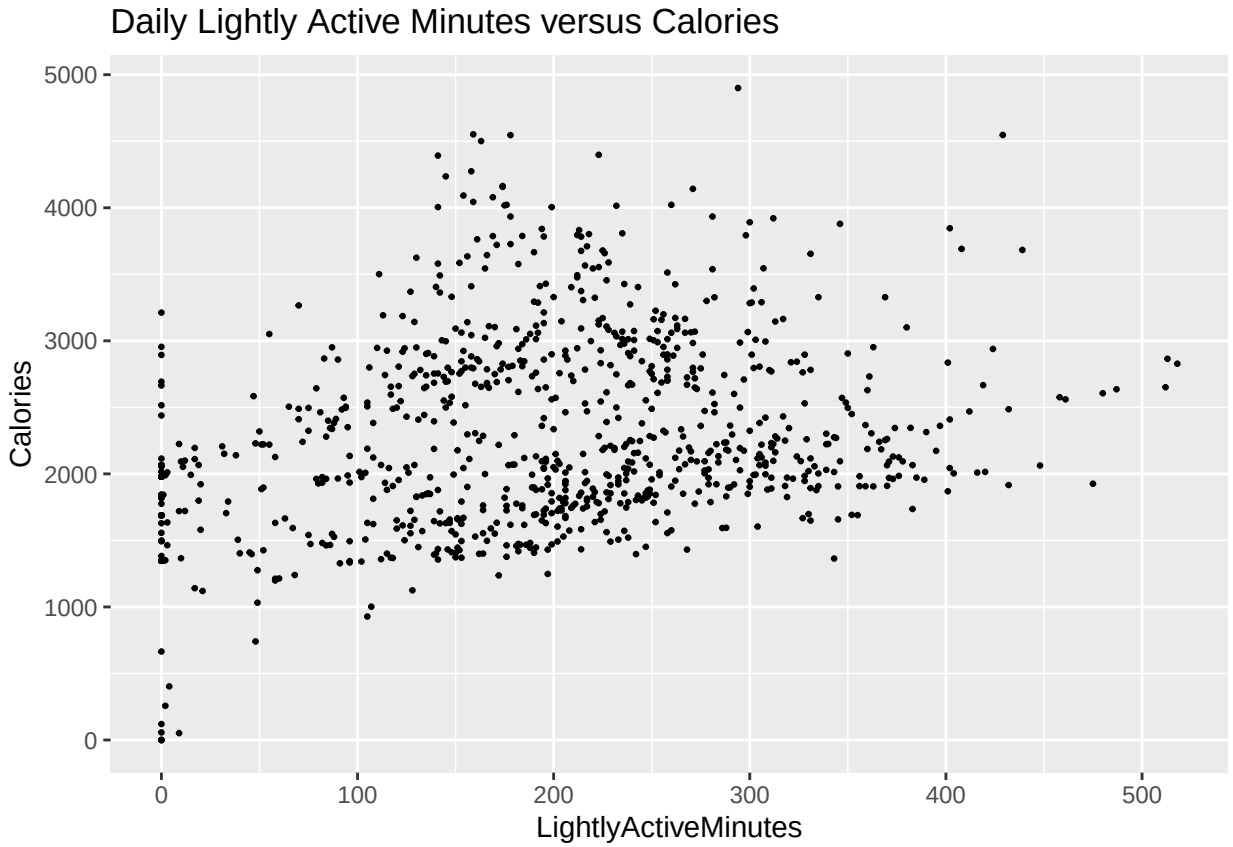
```
#Do a scatter plot with very active on x axis and calorie bun on y
ggplot(data = activities) +
geom_point(mapping = aes(x = VeryActiveMinutes, y = Calories), size=0.5) + labs(title = "Very Active Mi
```




```
cor(activities$VeryActiveMinutes, activities$Calories)
```

```
## [1] 0.6131723
```

```
#Do a scatter plot with lightly active on x axis and calorie bun on y
ggplot(data = activities) +
geom_point(mapping = aes(x = LightlyActiveMinutes, y = Calories), size=0.5) + labs(title = "Daily Light
```




```
cor(activities$LightlyActiveMinutes, activities$Calories)
```

```
## [1] 0.2777964
```

An interesting observation to note is that when comparing the two graphs, sample point is that on average takes 200 light minutes to burn 2500 calories, but 100 very active minutes to burn 3500 calories (which is half the time burning more)

Very active minutes have a much stronger correlation with calories burned than total active minutes, meaning that the majority of calories are burned when a person is most active. Maximizing the moments a person is very active is more effective in burning calories than maximizing the total amount of active minutes because fair and light activity is only weakly associated with calorie burn (like the correlation coefficients shown below 1) fairly, 2) lightly). Similarly, the correlation between activity and calories only drastically increases when it is very active (fairly and lightly active have a negligible difference).

```
cor(activities$FairlyActiveMinutes, activities$Calories)
```

```
## [1] 0.2949369
```

```
cor(activities$LightlyActiveMinutes, activities$Calories)
```

```
## [1] 0.2777964
```

In conclusion for all of the graphs, distance is the best indicator of calorie burn compared to steps and total active minutes

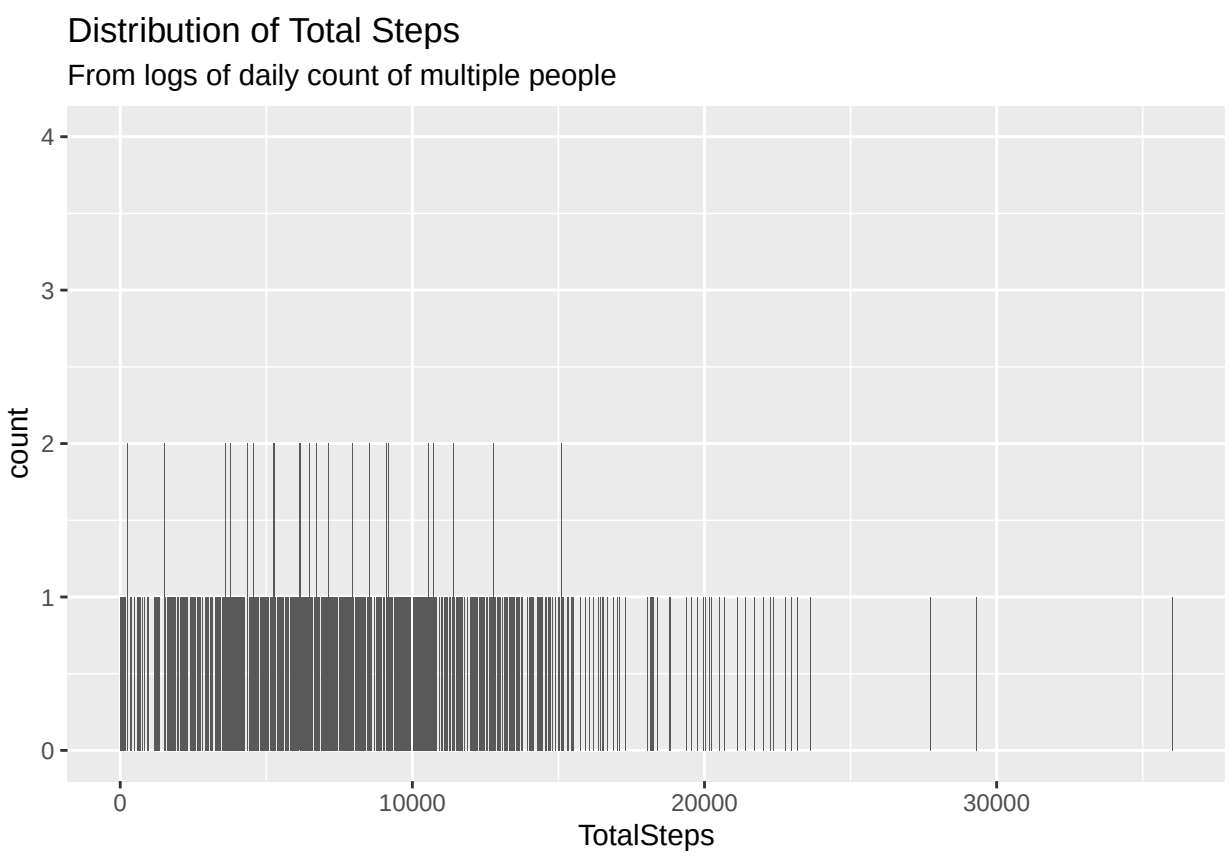
0.59 = steps 0.64 = distance 0.47 = total active 0.61 = very active 0.27 = light 0.29 = fairly

*NOTE: There could be other factors that contribute to calorie burn that are not measured here (weightlifting, height, weight, more)

Finding distribution, mean, standard deviation for distance, steps, active minutes and calories burned

```
#Do a histogram for steps, distance, active minutes, calories
ggplot(data = activities) +
geom_bar(mapping = aes(x=TotalSteps)) + scale_y_continuous(limits = c(0,4))  + labs(title = "Distributi
```

```
## Warning: Removed 1 rows containing missing values (`geom_bar()`).
```




```
summarize(activities, mean(TotalSteps,na.rm=TRUE))
```

```
##   mean(TotalSteps, na.rm = TRUE)
## 1                       7529.765
```

```
summarize(activities, median(TotalSteps,na.rm=TRUE))
```

```
##   median(TotalSteps, na.rm = TRUE)
## 1                             7328
```

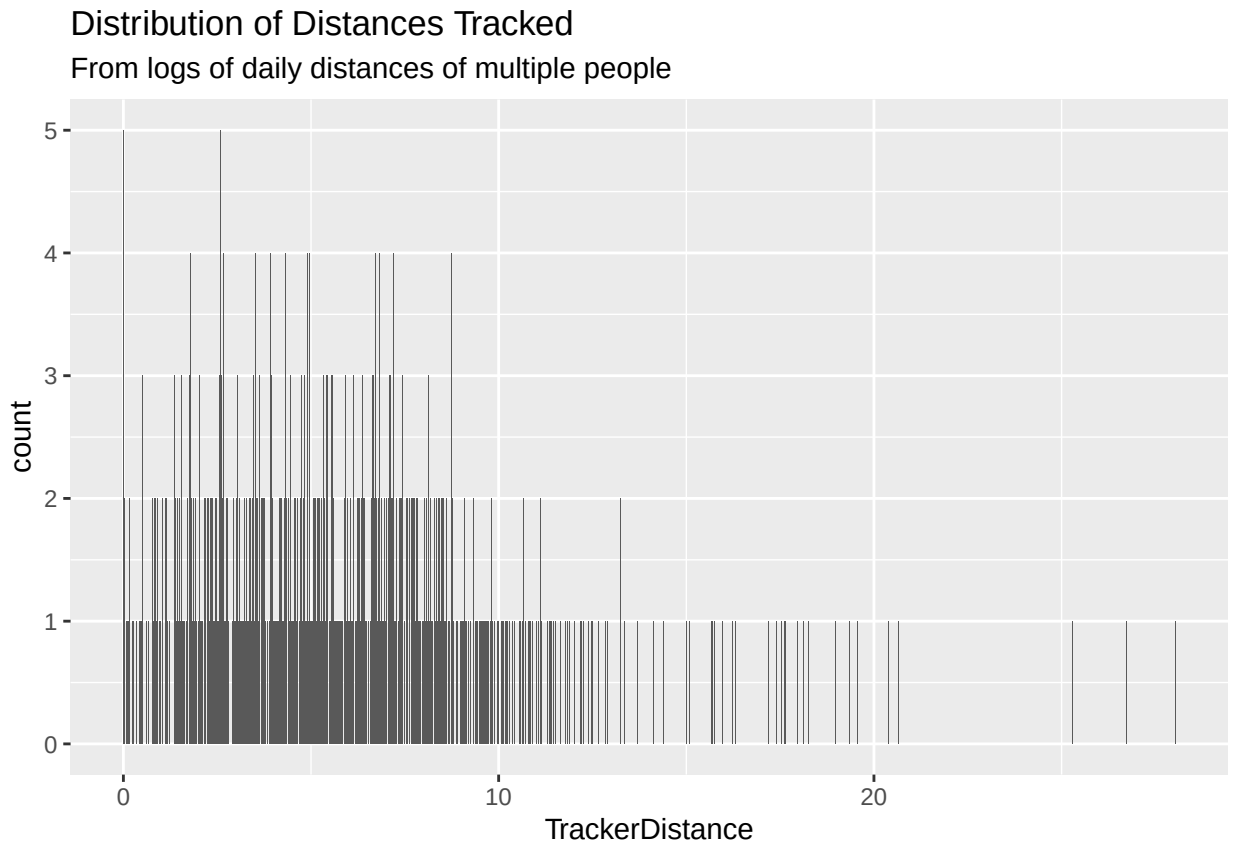
```
summarize(activities, sd(TotalSteps,na.rm=TRUE))
```

```
##   sd(TotalSteps, na.rm = TRUE)
## 1                     5048.079
```

Mean and median are similar which means the outliers are negligible, however standard deviation is 5048 so data is spread far apart.

```
#Do a histogram for steps, distance, active minutes, calories
ggplot(data = activities) +
geom_bar(mapping = aes(x=TrackerDistance)) + scale_y_continuous(limits = c(0,5)) + labs(title = "Distril
```

```
## Warning: Removed 1 rows containing missing values (`geom_bar()`).
```




```
summarize(activities, mean(TrackerDistance,na.rm=TRUE))
```

```
##   mean(TrackerDistance, na.rm = TRUE)
## 1                            5.408281
```

```
summarize(activities, median(TrackerDistance,na.rm=TRUE))
```

```
##   median(TrackerDistance, na.rm = TRUE)
## 1                                  5.18
```

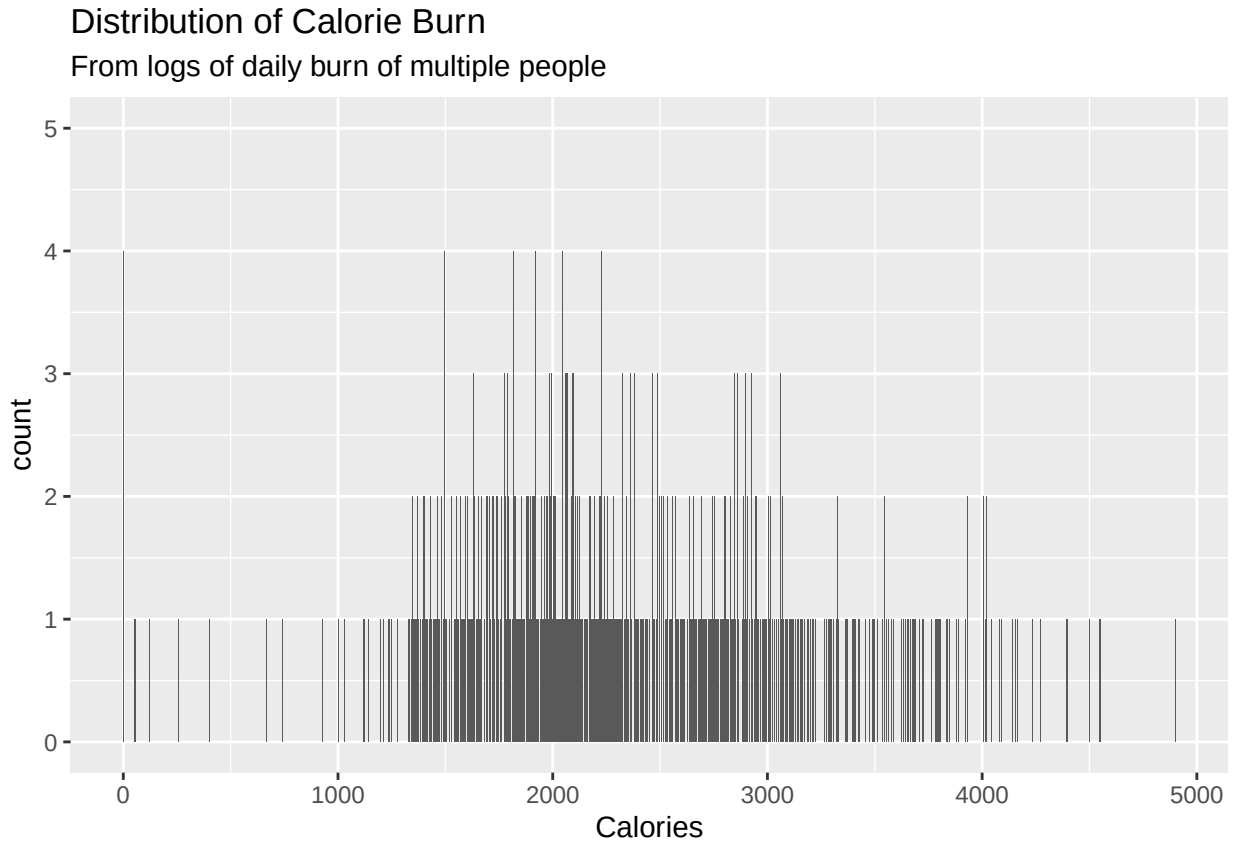
```
summarize(activities, sd(TrackerDistance,na.rm=TRUE))
```

```
##   sd(TrackerDistance, na.rm = TRUE)
## 1                          3.898302
```

For the distance and calorie burn, there is a good majority at the lower end, but there are more distance and more active people skewed to the right. A next step is to analyze the habits for people trending towards the right and encourage the majority to pursue the same habits.

```
#Do a histogram for steps, distance, active minutes, calories
ggplot(data = activities) +
geom_bar(mapping = aes(x=Calories)) + scale_y_continuous(limits = c(0,5)) + labs(title = "Distribution
```

```
## Warning: Removed 5 rows containing missing values (`geom_bar()`).
```




```
summarize(activities, mean(Calories,na.rm=TRUE))
```

```
##   mean(Calories, na.rm = TRUE)
## 1                     2296.445
```

```
summarize(activities, median(Calories,na.rm=TRUE))
```

```
##   median(Calories, na.rm = TRUE)
## 1                           2124
```

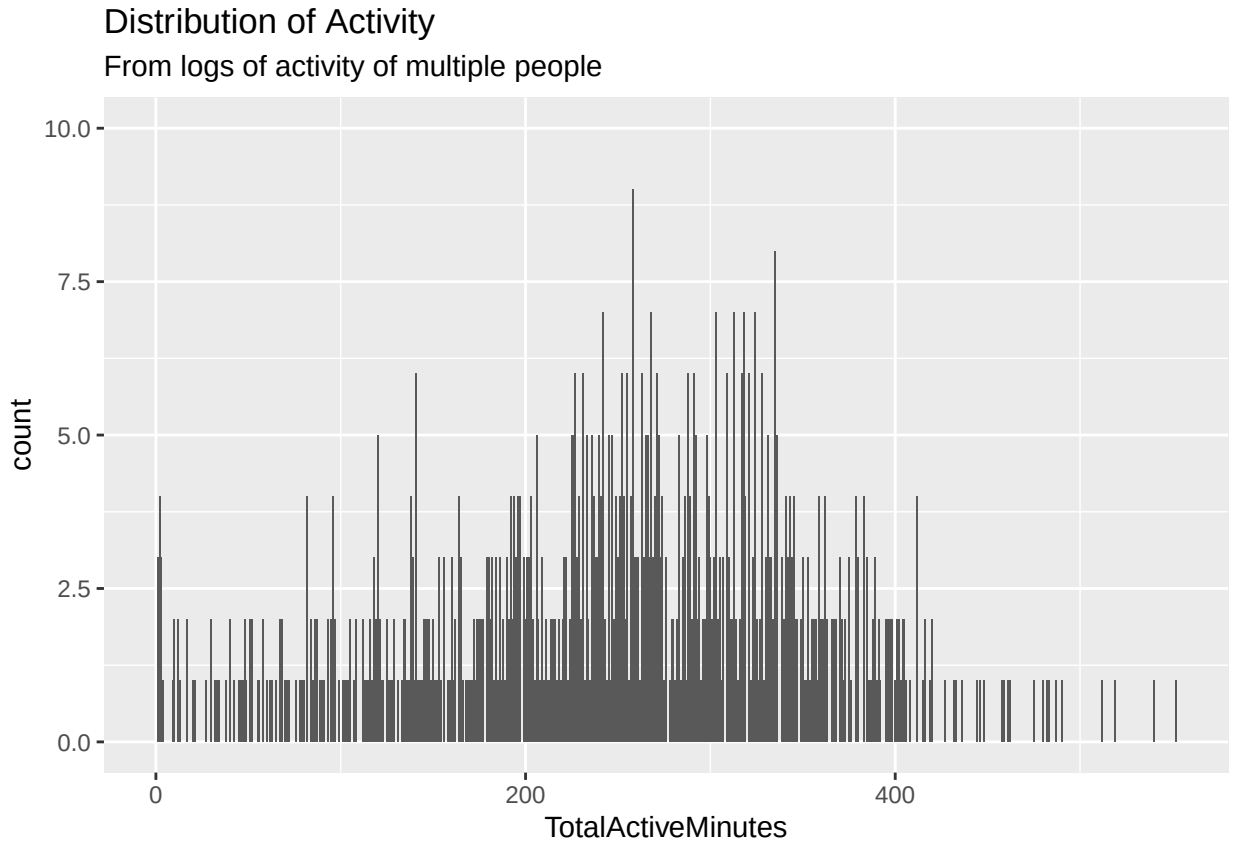
```
summarize(activities, sd(Calories,na.rm=TRUE))
```

```
##   sd(Calories, na.rm = TRUE)
## 1                   720.5423
```

Mean is 2296 and median is 2124 so the data is skewed to the right, the standard deviation is only 720.

```
#Do a histogram for steps, distance, active minutes, calories
ggplot(data = activities) +
geom_bar(mapping = aes(x=TotalActiveMinutes)) + scale_y_continuous(limits = c(0,10)) + labs(title = "Dis
```

```
## Warning: Removed 1 rows containing missing values (`geom_bar()`).
```




```
summarize(activities, mean(TotalActiveMinutes,na.rm=TRUE))
```

```
##   mean(TotalActiveMinutes, na.rm = TRUE)
## 1                               225.0865
```

```
summarize(activities, median(TotalActiveMinutes,na.rm=TRUE))
```

```
##   median(TotalActiveMinutes, na.rm = TRUE)
## 1                                      245
```

```
summarize(activities, sd(TotalActiveMinutes,na.rm=TRUE))
```

```
##   sd(TotalActiveMinutes, na.rm = TRUE)
## 1                             121.0137
```

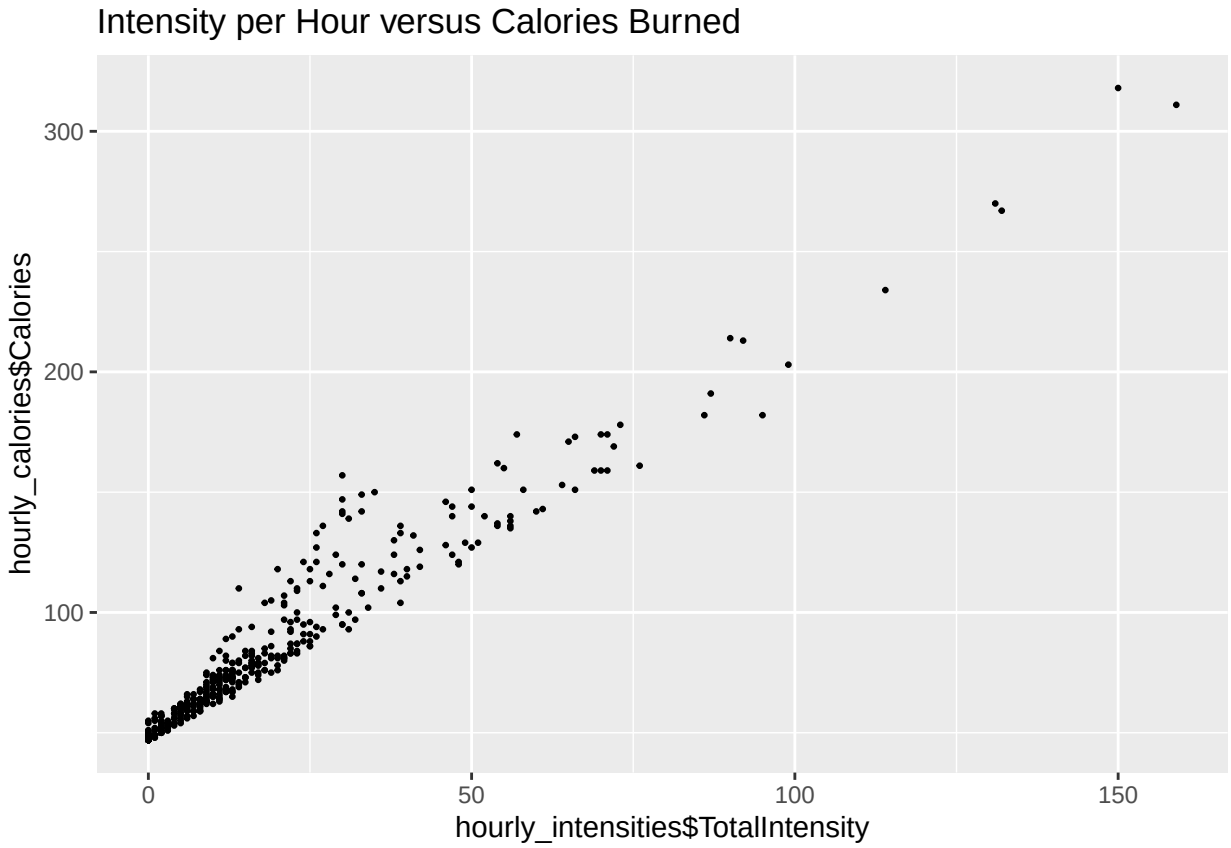
Mean is 225 and median is 245, so the data is centered around the mean/median with little/no skew (only graph with close to normal distribution). Standard deviation of 121 is over half the mean

```
#Analyze heartbeat
hourly_calories <-read.csv("hourlyCalories_merged.csv")
hourly_intensities <-read.csv("hourlyIntensities_merged.csv")
hourly_calories <- head(hourly_calories, 500)
hourly_intensities <-head(hourly_intensities, 500)
```

Find out if the TotalIntensity per hour has a linear relationship with calorie burn. This shows a most linear (slope nearly 1) correlation coefficient when intensity and calories are looked at closely.

```
#Analyze heartbeat

ggplot(data = NULL) +
geom_point(mapping = aes(x = hourly_intensities$TotalIntensity, y = hourly_calories$Calories), size=0.5)
```




```
cor(hourly_intensities$TotalIntensity, hourly_calories$Calories)
```

```
## [1] 0.9718523
```

When the intensity of activity and the calories burned are looked closely by the hour, you can see more of a direct correlation (nearly linear). Smaller timeframes clearly demonstrate the intensity's effect on calorie burn.

‘‘‘

## ACT

The main insight from the line graphs comparing each of the predictors (distance, steps, and types of active minutes) to calorie burn, the distance each person traveled is the strongest predictor and the amount of *very active* minutes is the second best predictor. A good way to market the app is to make features that document those statistics as highlights in a person's fitness journey, while also documenting other statistics (steps, light active) as supplements.

As for the bar graphs, the main action items are to study the habits of the higher skew distance and calorie burners via survey, and encourage users to practice those habits. The median distances and calorie burns are not in the middle, but in the lower end so a good improvement would be to see the mean and median closer to each other.