# Analysis into what factors affect the use of fitness app

By Dylan Koordi

## Abstract

As part of an ongoing marketing campaign for virtual bike sharing company, Cyclistic, the aim of this report is to analyze the different factors that are key to converting casual riders to annual riders in the company. This project will place an emphasis on the off bicycle activity. Namely: calories burnt as a KPI as well as average minutes active on app as a key indicator.

## Contents

## Read in the data

Initial observation of data shows that there exists 940 observations consisting of 15 variables. Things to note when initially reading in the dataset, there are too many repeated variables that may not be of use in the analysis. For example, Total Distance and Tracker Distance reflects identical results, which may not be inherently useful for report. Another key observation is that there exists: VeryActiveDistance,ModeratelyActiveDistance,LightActiveDistance. For the scope of this report, the categorization into the 3 may not be effective, as such, part of the transformation and cleaning process would be the remove all repeated variables while also combining and taking subsequent sums or averages of key variables to provide a more accurate analysis.

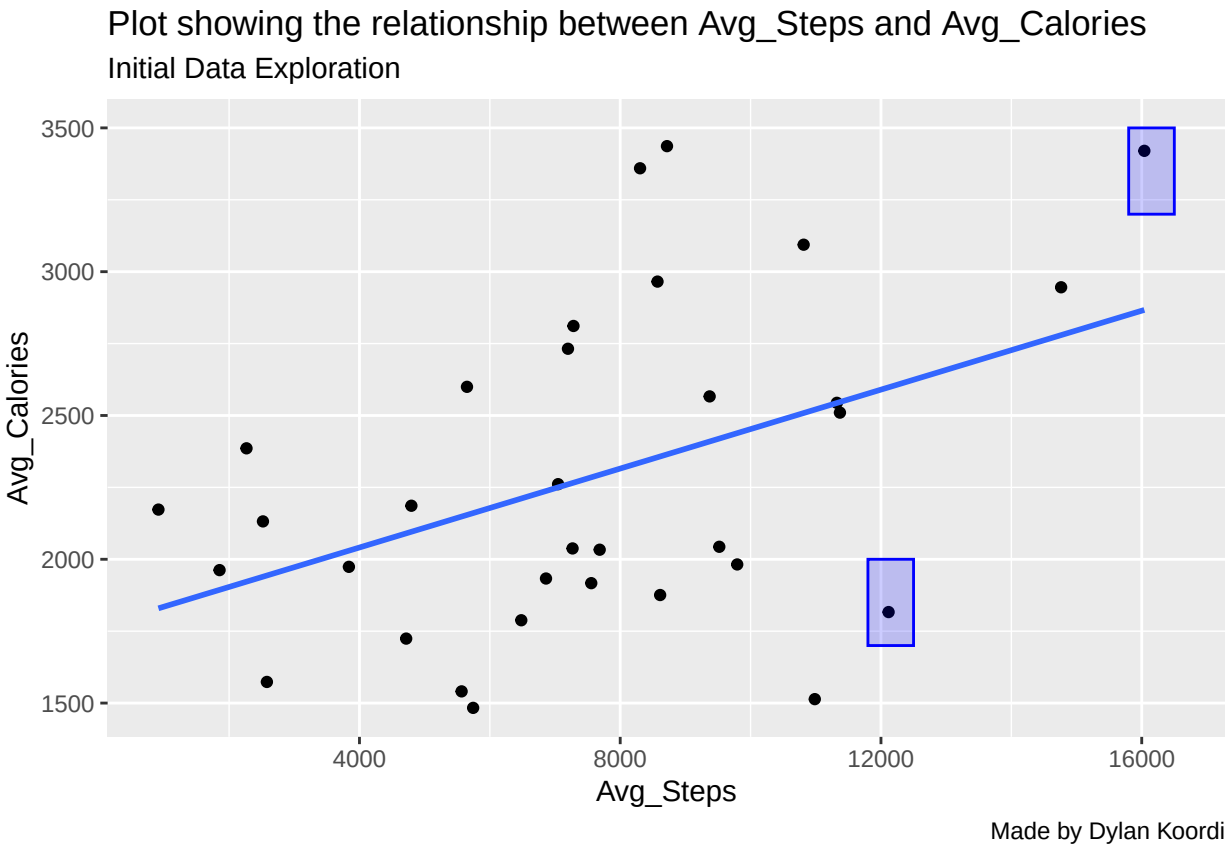
```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    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
```

## Data Cleaning and Transformation

### Data to obtain average steps and average calories: Duplicate Id entries

```
## # A tibble: 6 x 3
##           Id Avg_Steps Avg_Calories
##        <dbl>     <dbl>        <dbl>
## 1 8378563200     8718.        3437.
## 2 8877689391    16040.        3420.
## 3 5577150313     8304.        3360.
## 4 4388161847    10814.        3094.
## 5 4702921684     8572.        2966.
## 6 8053475328    14763.        2946.
```

**Initial data exploration: Is there a signficant relationship between Avg_Calories and Avg_Steps?**




There's a saying that the more steps you take, logically, the more calories you burn because the increased steps taken constitutes a greater amount of time where the body is actively working. But contrary to that theory, our plot shows that there is not a clear relationship between the average steps and average calories. Namely, an individual's combine steps does not directly lead to a higher average calories burnt. The highlighted portion of the diagram reflects points that support the relationship by in true, the other highlighted point on the bottom of the diagram reflects otherwise.

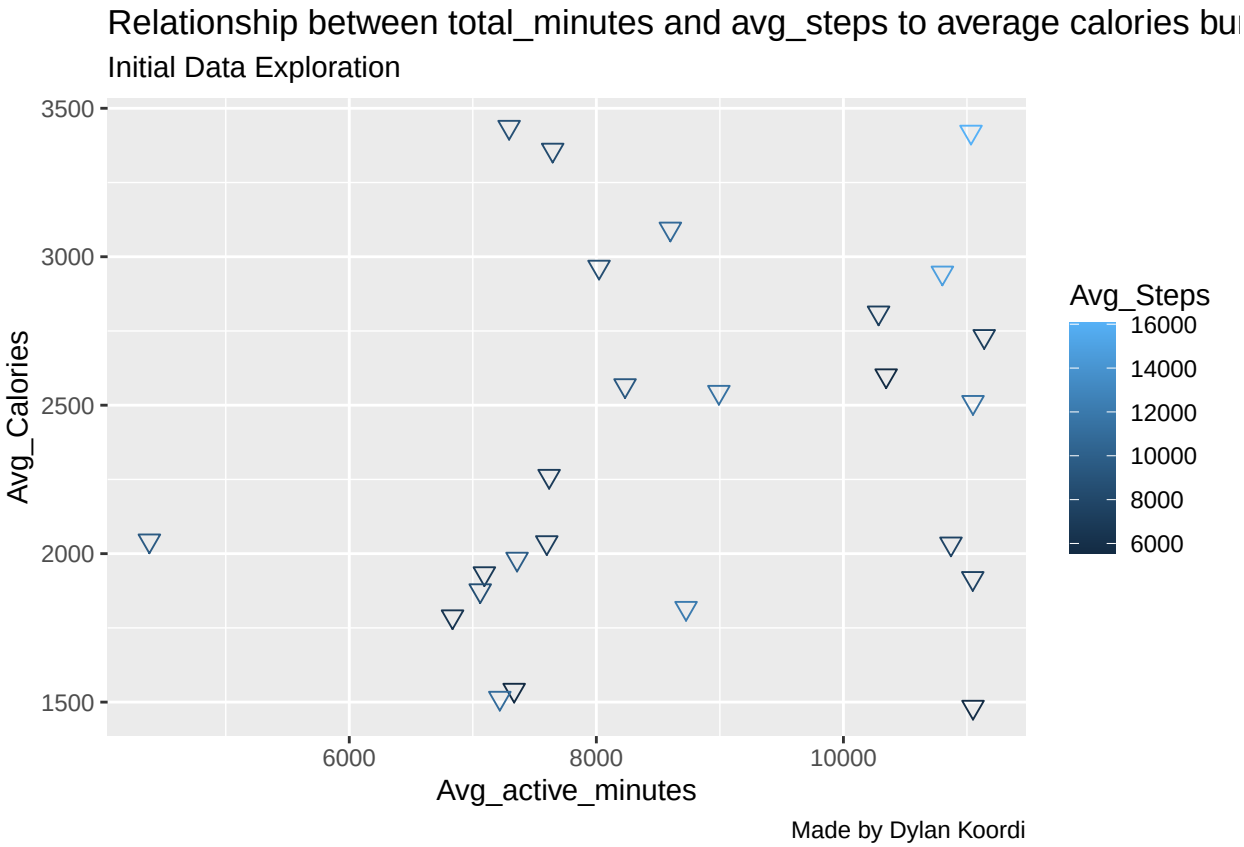
**Another transformation applied to combine the 3 categories of Minutes together.**

```
##           Id Avg_Steps Avg_Calories       Id.1 Total_Active_Minutes
## 1 1503960366  12116.74      1816.42 1503960366                34905
## 2 1624580081   5743.90      1483.35 1624580081                44197
## 3 1644430081   7282.97      2811.30 1644430081                41138
## 4 2022484408  11370.65      2509.97 2022484408                44196
## 5 2026352035   5566.87      1540.65 2026352035                29339
## 6 2347167796   9519.67      2043.44 2347167796                17527
##   Avg_active_minutes
## 1            8726.25
## 2           11049.25
## 3           10284.50
## 4           11049.00
## 5            7334.75
## 6            4381.75
```

This transformation combines the three categories of minutes into one, "Total_Active_Minutes" which essentially sums the time across the 3 categories. Beyond that, the duplicate entries for Id was also further accounted for. Hereby, the table reflects the summary for each user by virtue of the new transformed variables. As an added precaution, I also filtered individuals who have taken more than 5000 steps, the reason being that users who do not accumulate more than 5000 steps to date, are not representative of the sample for analysis.

At this point, the initial data exploration shows that, if average steps and average calories does not have a direct relationship, does average calories and total time spent have any relation?

**A plot highlighting the significance of average time spent against average calories burnt.**




There is some association between active minutes and average calories burned. Namely the plot shows that for every increase in active minutes on the app, the greater the number of average calories burnt. While the trend is generally upward sloping, there are also a lot of outliers, which may indicate that while individuals were active on the app, they may not actually be burning calories (or doing anything).

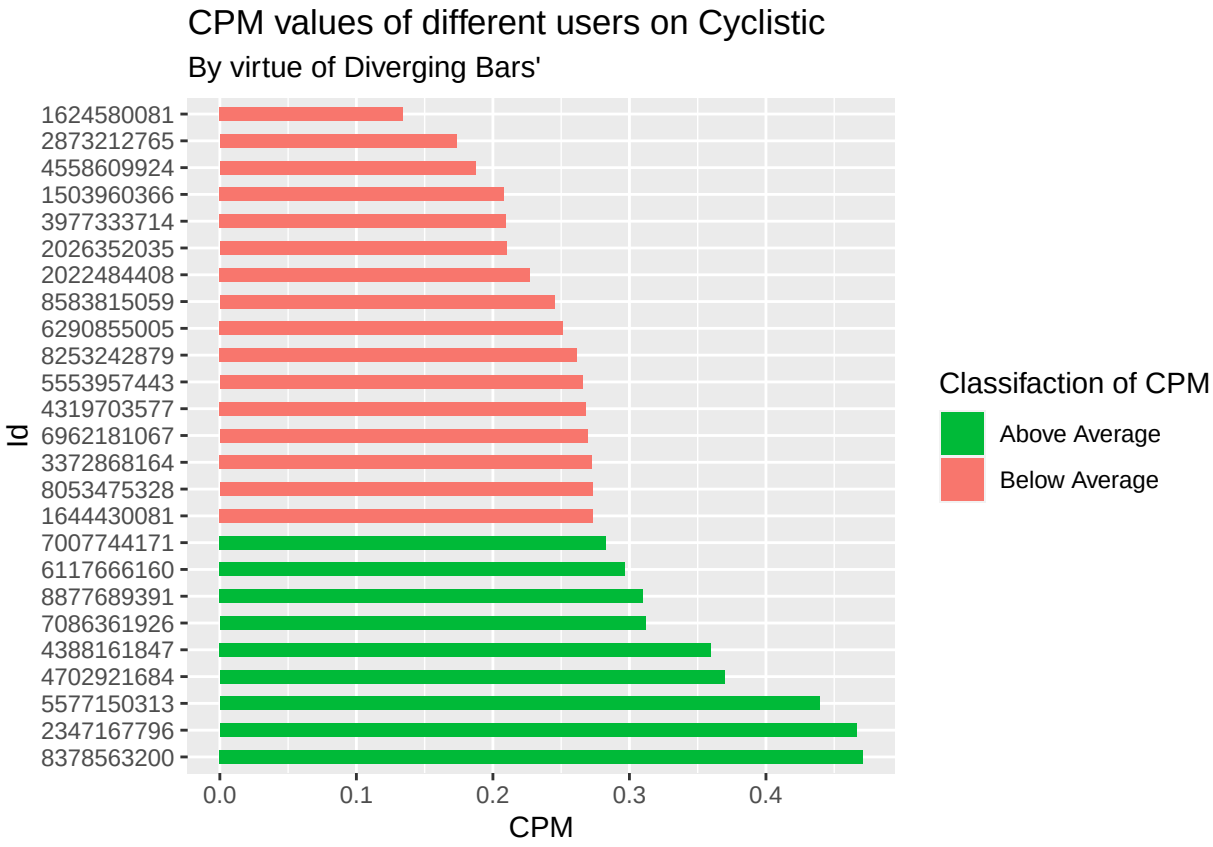
## Calculating a new metric: CPM

CPM refers to the calories per minute. This is essentially the number of calories burnt over the duration of active minutes. Trivially, I incorporated this metric because it measures the efficiency of the individual. For example, an individual is not deemed as efficient if their CPM is low, because for the high amount of active minutes spent walking, they are not burning as many calories. Vise versa

**Plot for Calories Per Minute by Individual**

```
##        Id              Avg_Steps      Avg_Calories       Id.1
##  Min.   :1.504e+09   Min.   : 5567   Min.   :1483   Min.   :1.504e+09
##  1st Qu.:2.873e+09   1st Qu.: 7199   1st Qu.:1917   1st Qu.:2.873e+09
##  Median :4.703e+09   Median : 8572   Median :2261   Median :4.703e+09
##  Mean   :5.044e+09   Mean   : 8986   Mean   :2368   Mean   :5.044e+09
##  3rd Qu.:7.008e+09   3rd Qu.:10814   3rd Qu.:2811   3rd Qu.:7.008e+09
##  Max.   :8.878e+09   Max.   :16040   Max.   :3437   Max.   :8.878e+09
##  Total_Active_Minutes Avg_active_minutes      CPM
##  Min.   :17527        Min.   : 4382      Min.   :0.1342
##  1st Qu.:29339        1st Qu.: 7335      1st Qu.:0.2272
##  Median :32929        Median : 8232      Median :0.2694
##  Mean   :34821        Mean   : 8705      Mean   :0.2815
##  3rd Qu.:43205        3rd Qu.:10801      3rd Qu.:0.3100
##  Max.   :44559        Max.   :11140      Max.   :0.4712
```

Useful statistic to observe from this summary, is that the mean CPM is valued at 0.2815 in the sample. This will be used to value invidiuals that are riding above or below the mean.




Key points of observation for the average user. The time spent on the app deemed as 'active' is not a good reference point for the KPI being calories burnt. We see that individuals who attain the highest level of CPM are usually the ones who spent the least amount of time on the app by virtue of the diagram above.

Now that we have investigated the CPM, what about the different weight classes?

## An investigation into which weight class uses the most?

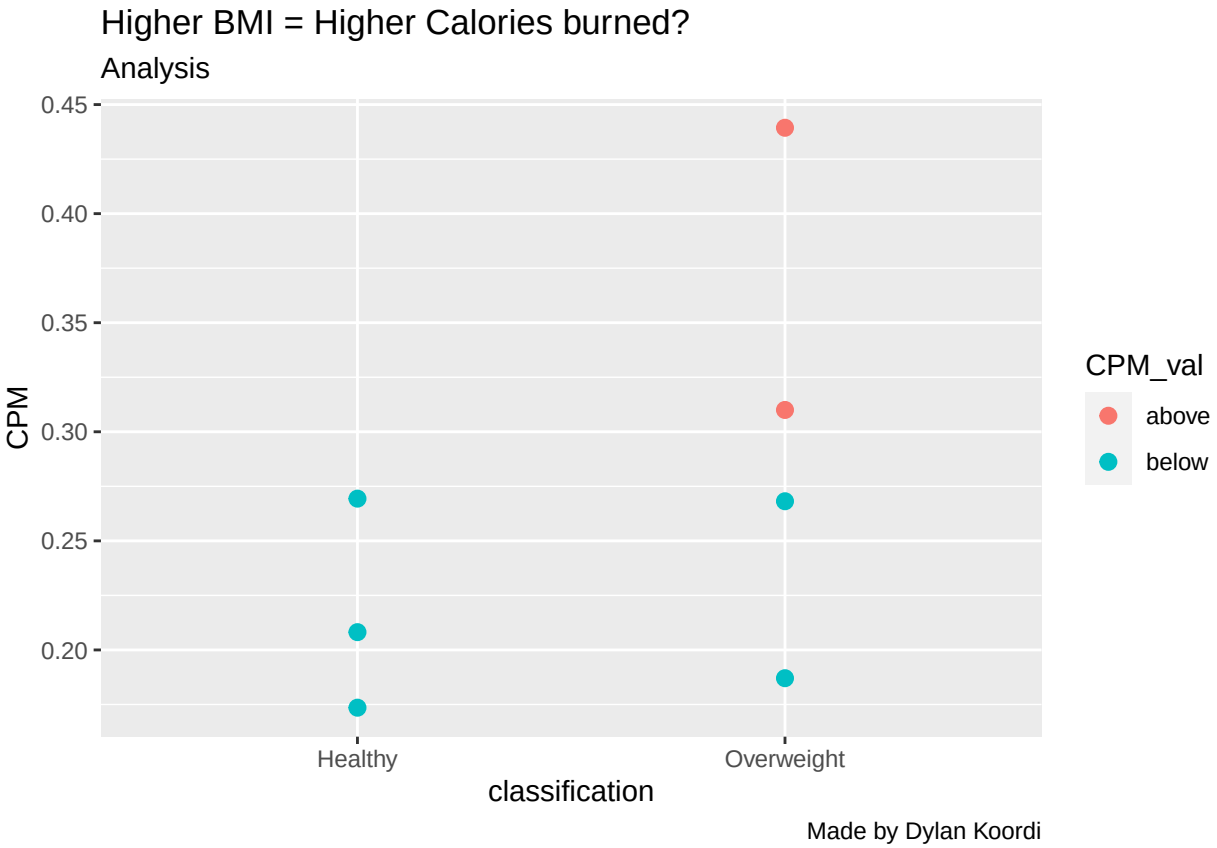
### Classification of riders according to their weight classes

```
## # A tibble: 8 x 4
##           Id Avg_Weight Avg_BMI classification
##        <dbl>      <dbl>   <dbl> <chr>
## 1 1503960366         53    23   Healthy
## 2 1927972279        134    48   Overweight
## 3 2873212765         57    21.5 Healthy
## 4 4319703577         72    27   Overweight
## 5 4558609924         70    27   Overweight
## 6 5577150313         91    28   Overweight
## 7 6962181067         62    24   Healthy
## 8 8877689391         85    26   Overweight
```

As it turns out, a majority of the riders are deemed as overweight. For reference, the healthy BMI is anything that is <24.9, by virtue of that number, we obtain the following classifications above.

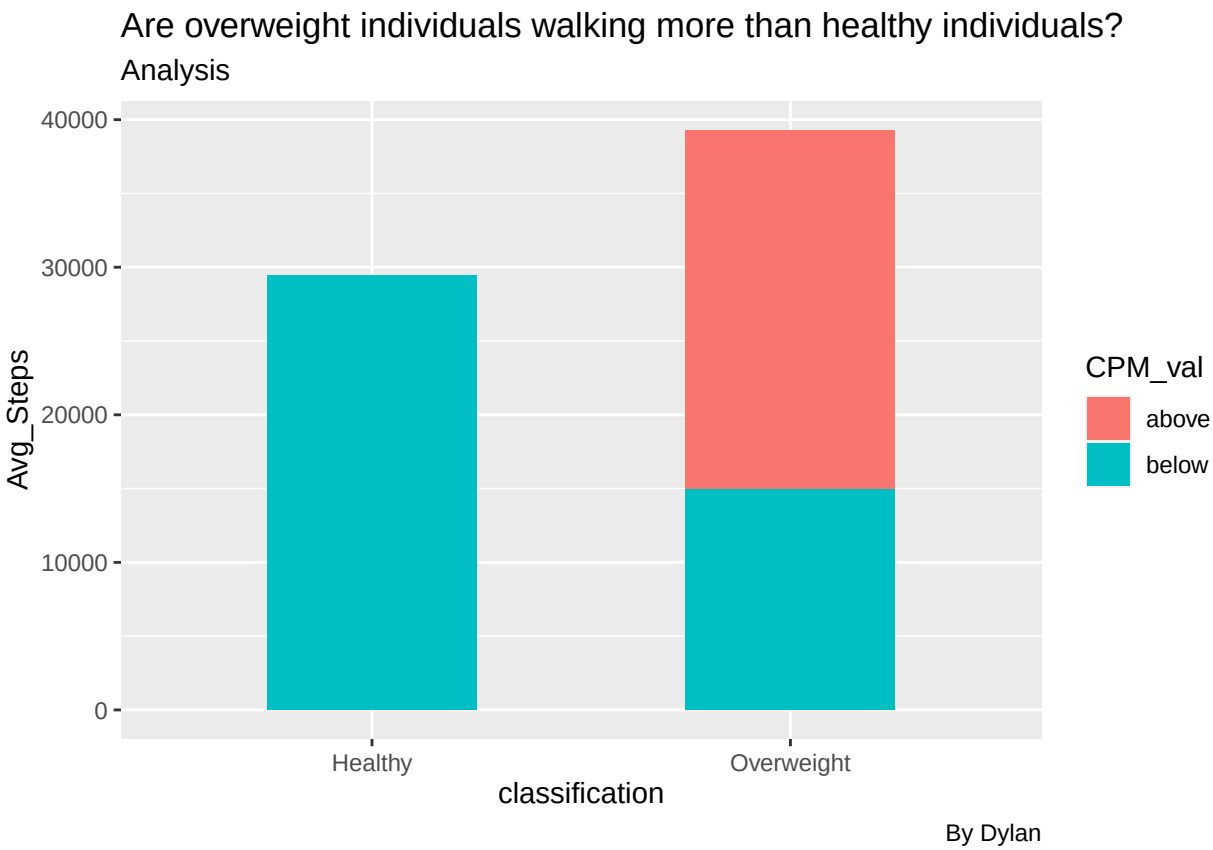
## Merging data for more analysis

### Do overweight individuals tend to burn more fat than a normal weighted individual?



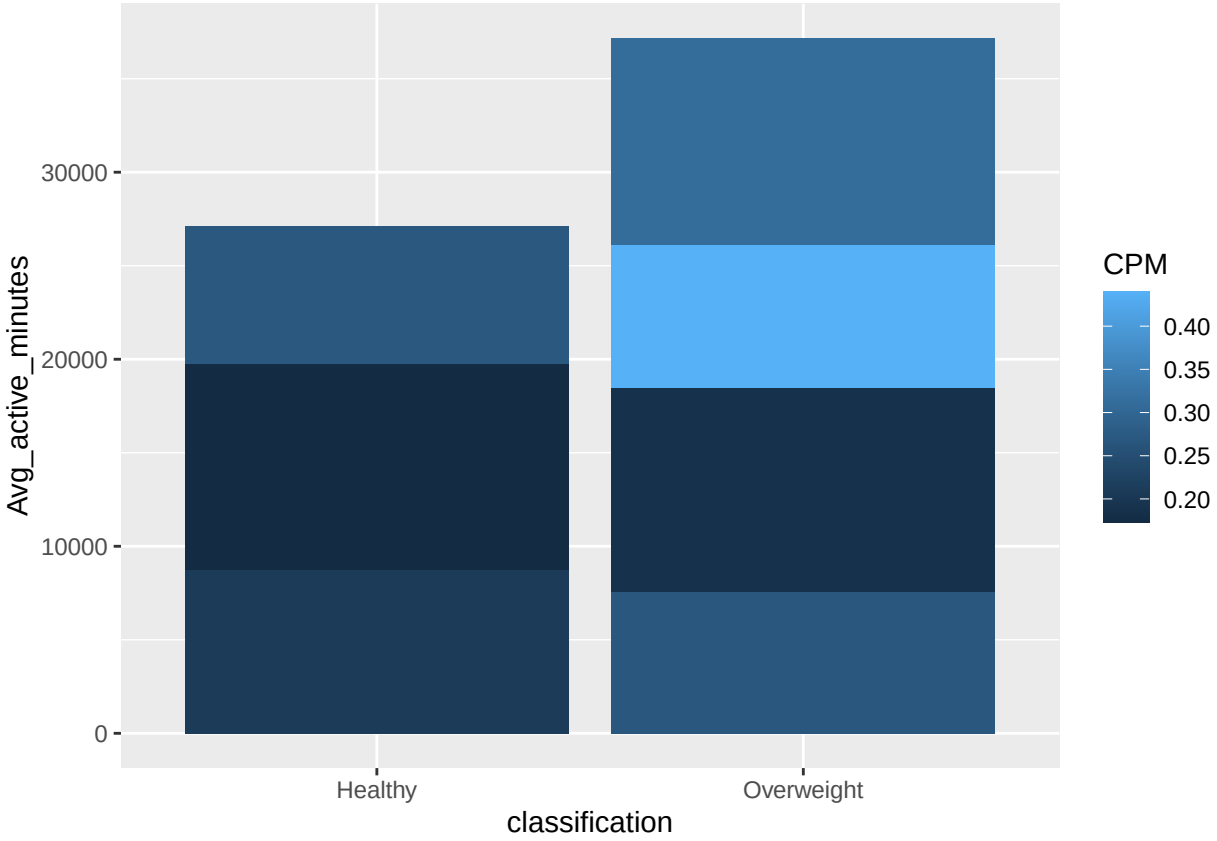

Surprisingly! Riders classified under overweight tend to have a higher CPM than riders that are deemed as healthy. Trivially, half of the individuals classified as overweight, actually have a CPM value that is above the average while in contrary, the CPM for the healthy riders are all below average. This is food for thought.

**Are overweight individuals walking more than normal individuals?**




Coherently, overweight individuals are walking more on average than the healthy individuals. By virtue of the previous plot, the CPM of individuals who are above the average are indeed walking more. This incidentally explains why they are burning more calories over time.

**But are the overweight individuals spending that much more time than the healthy individuals?**




Coherently, there is some evidence suggesting that the overweight individuals are spending more time on the app. But, the big point to point out is that individuals who had the highest CPM actually were active on the app lesser than those that had the most time on the app. This further supports the theory that more minutes does not imply more weight loss/calories burnt.

## Conclusion

The recommendation is for cyclistic to offer more services targeting towards individuals that have a higher BMI or are deemed as 'overweight'. Supporting this claim was that overweight individuals tend to have a higher CPM value compared to individuals deemed as healthy. Another important insight was that the minutes active on the app does not actually correlate to the number of calories burnt. This could point out to the inconsistency of the individual use of the app.