My Health: Data Analysis

2022-10-05

Research Questions

Throughout this presentation I will dive into many questions regarding trends and patterns within my apple watch sensory data over the past 90 days. Some of the big questions are included below.

How has my resting heart rate changed over time?

What days of the week do I tend to burn the most calories?

Which health metric is most correlated with calories burned?

What workouts did I do the most often?

How does the duration of my workouts correlate with active calories?

What is the most efficient workout for calorie expenditure?

Health and Workout Data

Health Data

This health csv file is derived from the Apple Watch sensory data. I organized my health data into a total of 12 variables, all being numerical, in order to get a wide range of analysis on my day to day health metrics.

```
#reads in my health and workout csv files
healthData <- read.csv("/Users/jonpaino/Library/Mobile
Documents/com~apple~CloudDocs/Stats 10/HealthData.csv")
head(healthData)

##                  Date  actCal exerTime hoursStand minStand restCal minHR
maxHR
## 1 2022-07-14 0:00:00  964.29       98         14      173    1829    40
152
## 2 2022-07-15 0:00:00  568.82       59         12      141    1838    40
126
## 3 2022-07-16 0:00:00 1503.00      109         15      448    1866    40
171
## 4 2022-07-17 0:00:00 1325.00      127         13      227    1826    40
150
## 5 2022-07-18 0:00:00 1108.00       88         12      169    1818    44
```

```
185
## 6 2022-07-19 0:00:00 1483.00     140        13      232    1855    41
165
##   avgHR    hrv restHR totalDist
## 1 80.80 105.88     43      4.53
## 2 67.88 148.22     42      3.22
## 3 96.74  99.09     50     17.06
## 4 93.62  96.19     46     13.74
## 5 94.25 142.60     48      9.62
## 6 99.08 111.22     47     13.82
```

Workout Data

This workout csv file is also derived from the Apple Watch sensory data. Within this data there are a total of 10 variables, with one being categorical that describes what type of workout I completed.

```
workoutData <- read.csv("/Users/jonpaino/Library/Mobile
Documents/com~apple~CloudDocs/Stats 10/WorkoutData.csv")
head(workoutData)

##                             type            start              end duration
## 1 Traditional Strength Training  2022-07-14 6:56  2022-07-14 7:38  0:41:53
## 2                Stair Climbing  2022-07-14 7:39  2022-07-14 8:25  0:45:31
## 3 Traditional Strength Training  2022-07-15 7:28  2022-07-15 8:24  0:55:26
## 4                      Swimming 2022-07-15 15:06 2022-07-15 15:40  0:33:57
## 5                       Running  2022-07-16 6:37  2022-07-16 7:44  1:06:56
## 6                       Running  2022-07-17 7:01  2022-07-17 8:12  1:11:01
##     tCal actCal maxHR  avgHR distance speed
## 1 207.14 147.48   112  79.84       NA    NA
## 2 503.26 419.85   152 136.92       NA    NA
## 3 259.72 165.55   126  89.34       NA    NA
## 4 122.20  54.49   124  71.77       NA 0.181
## 5 867.52 752.35   171 154.59     8.01 7.180
## 6 856.13 735.49   150 137.55     8.01 6.770

library(tidyverse)
library(lubridate)
library(modelr)
```

Discussion and data analysis

Basic statistical measurements for cardiovascular health data

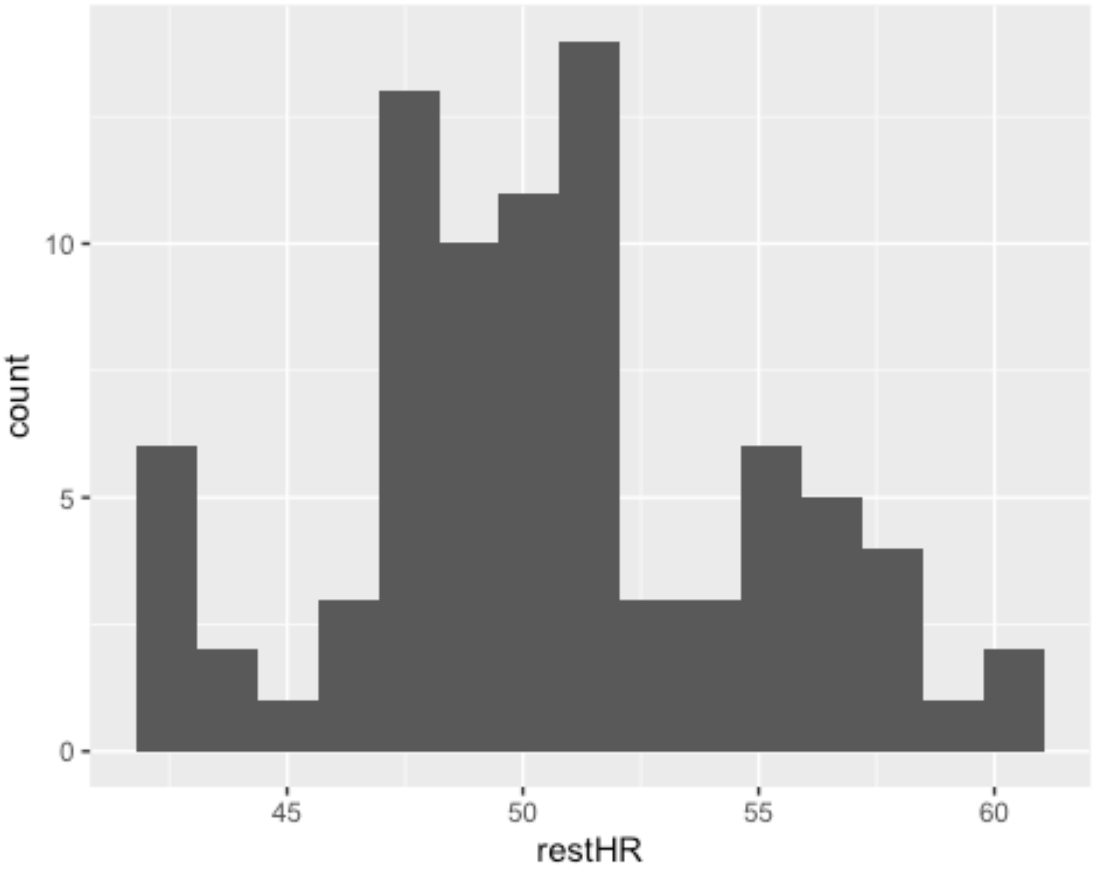
Resting Heart Rate

```
#average resting heart rate
mean(healthData$restHR)
```

```
## [1] 50.57143

#standard deviation of resting heart rate
sd(healthData$restHR)

## [1] 4.333355

#finds the most common occurence (heart rate on top, occurences on bottom)
mode <- sort(table(healthData$restHR),decreasing=TRUE)[1]
mode

## 50
## 11
```

The average of my resting heart rate is about 50.7 bpm. My resting heart rate typically varies by around 4.3 bpm. The most common resting heart rate was 50 bpm with a total of 11 occurrences over the past 90 days.

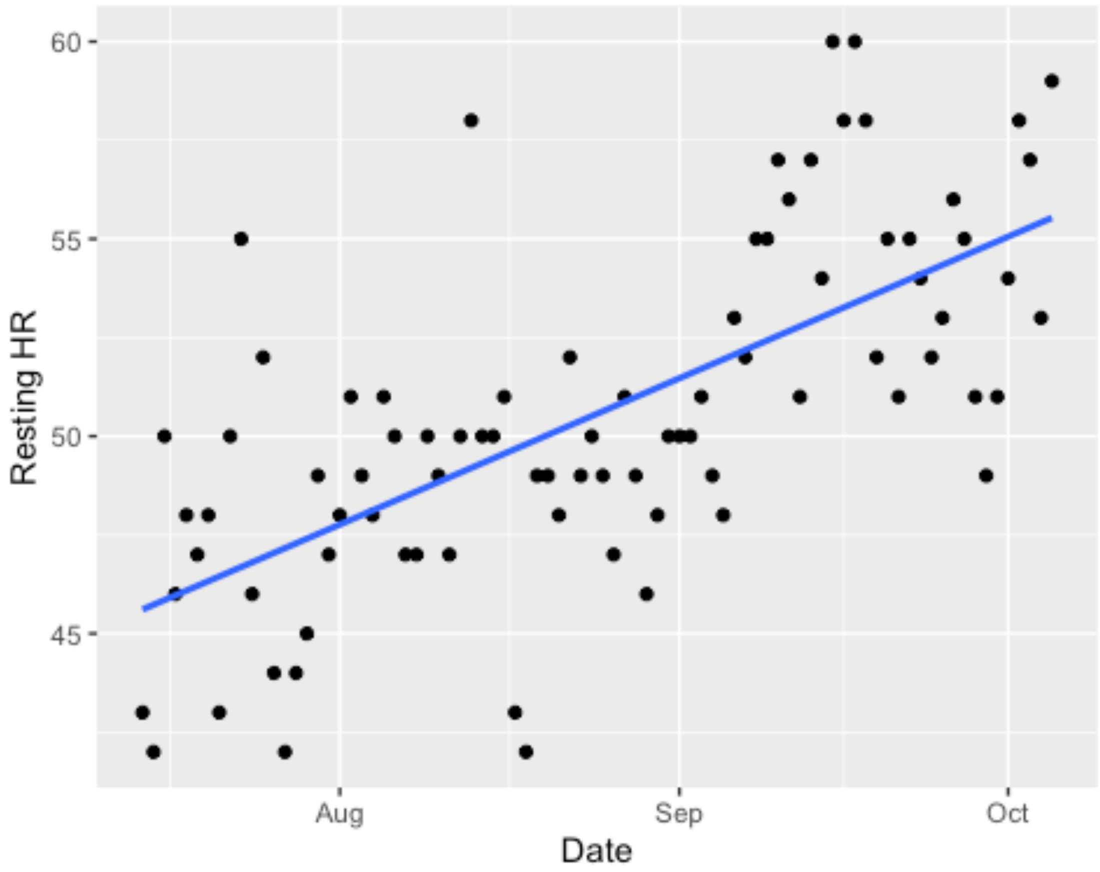
```
#plots resting heart rate occurences in a histogram
ggplot(healthData, aes(restHR)) + geom_histogram(bins = 15)
```



```
range(healthData$restHR)

## [1] 42 60
```

You can see from the distribution of my resting heart rate that I most frequently sit at around 50 with a range of +- 10 beats per minute. My lowest resting HR was 42 bpm while my highest was at 60.

```r
#Creates a date-time function from the Date variable that is currently a
string
newDate <- ymd_hms(healthData$Date)
#Creates a plot for resting heart rate over the 90 day period
ggplot(healthData, aes(newDate,restHR)) +
geom_point() + geom_smooth(method = "lm", se = F) + labs(x = "Date", y =
"Resting HR")

## `geom_smooth()` using formula 'y ~ x'
```



You can see an increasing trend in my resting heart rate over the last 90 days which is very understandable from my knowledge of the circumstances behind the data. I would likely attribute it to the intentional weight gain that I have been working on during this time period and the move to college which involved lots of stress and uncertainty. Both of these scenarios are likely culprits of increasing heart rate.

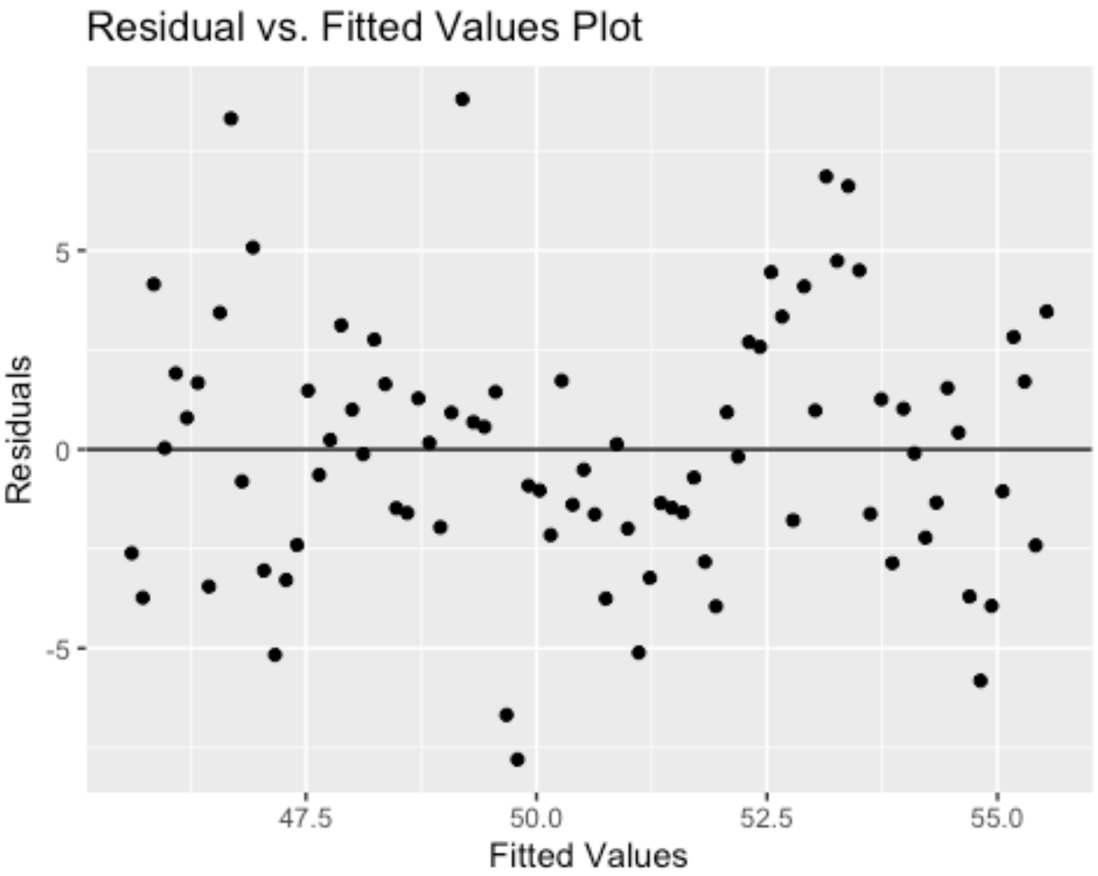
The relationship between my resting HR and time does clearly have an increasing trend, however the association is fairly weak with the data points not being too close to the regression line.

```
#Creates a linear model of all the summary statistics betweeen resting heart
rate and time
 model <- lm(restHR ~ newDate, data = healthData)
summary(model)
```

```
##
## Call:
## lm(formula = restHR ~ newDate, data = healthData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7942 -2.0307 -0.1095  1.6814  8.8036
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.249e+03  2.790e+02  -8.060 5.32e-12 ***
## newDate      1.384e-06  1.679e-07   8.241 2.32e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.224 on 82 degrees of freedom
## Multiple R-squared:  0.453,  Adjusted R-squared:  0.4464
## F-statistic: 67.91 on 1 and 82 DF,  p-value: 2.325e-12
```

You can see from this output that the R^2 value is .45 which indicates that time explains 45% of the variation in resting heart rate.

```
#Creates a model of the regression points
ggplot(model, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title='Residual vs. Fitted Values Plot', x='Fitted Values',
y='Residuals')
```



Given the residual plot above you can see how there seems to be a random scattering of the points around the regression line, indicating a good correlation. However, given the distance of these points away from the regression line, the correlation appears to be fairly weak.

Max Heart Rate

```
#finds highest heart rate over past 90 days
max(healthData$maxHR)

## [1] 185
```

My Highest heart rate in the past 90 days was 185 beats per minute.

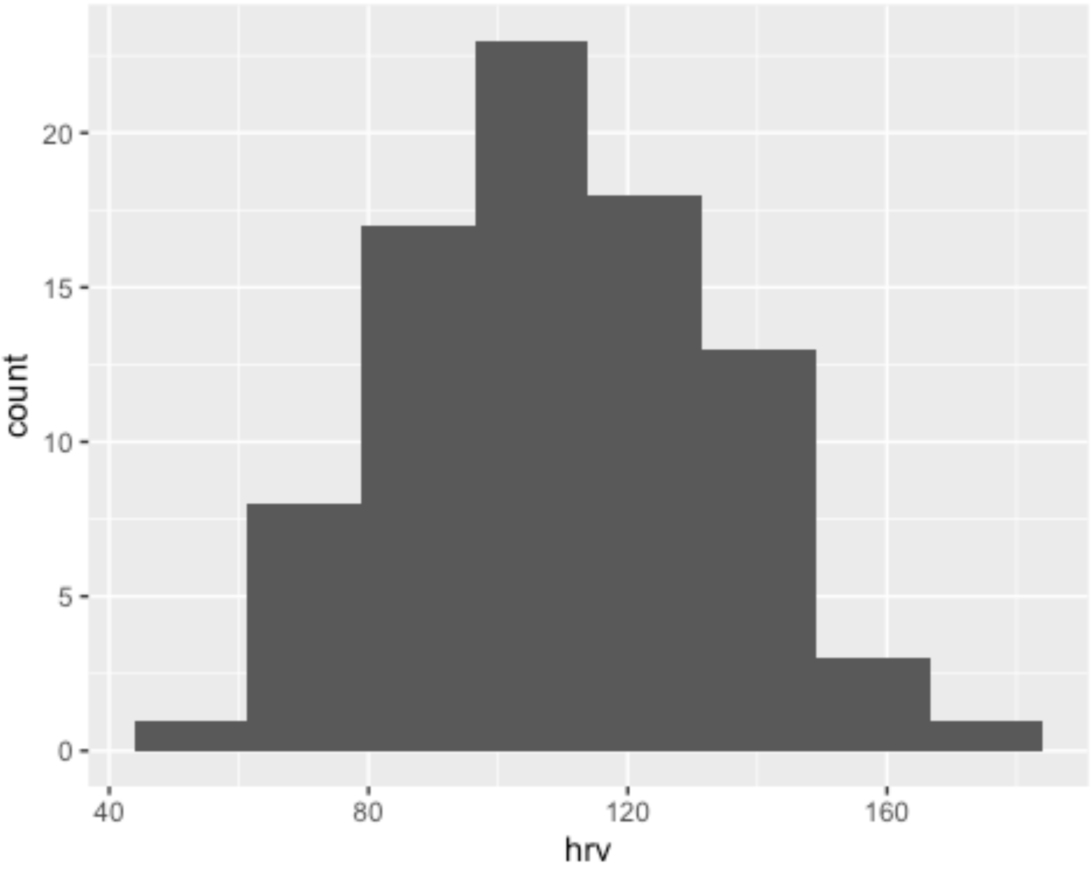
How many calories did I burn on my highest heart rate day?

```
#creates a tibble with actCal, restCal and a new totalCal variabel
cals <- healthData %>%
  group_by(actCal, restCal) %>%
  mutate(totalCal= actCal + restCal)
#searches for the value of total calories for the data points where max heart
rate is equal to 185
cals$totalCal[which(healthData$maxHR == 185)]

## [1] 2926 2834
```

You can see from this output that I hit a heart rate of 185 on two different occasions. I hit a total calorie expenditure of 2926 and 2834.

Heart Rate Variability

```
ggplot(data = healthData, aes(hrv)) + geom_histogram(bins = 8)
```



```
mean(healthData$hrv)

## [1] 109.6165

sd(healthData$hrv)

## [1] 24.49833
```

My heart rate variability forms roughly a normal distribution with the mean being around 110 milliseconds.

With most sources claiming a healthy heart rate variability being >60 ms, this data is quite surprising as mine is consistently almost double this baseline.

What percent of the time am I above a healthy HRV?

```
#Finds proportion of healthy HRV occurences
healthyhrv <- length(which(healthData$hrv >= 60))/length(healthData$hrv) *
```

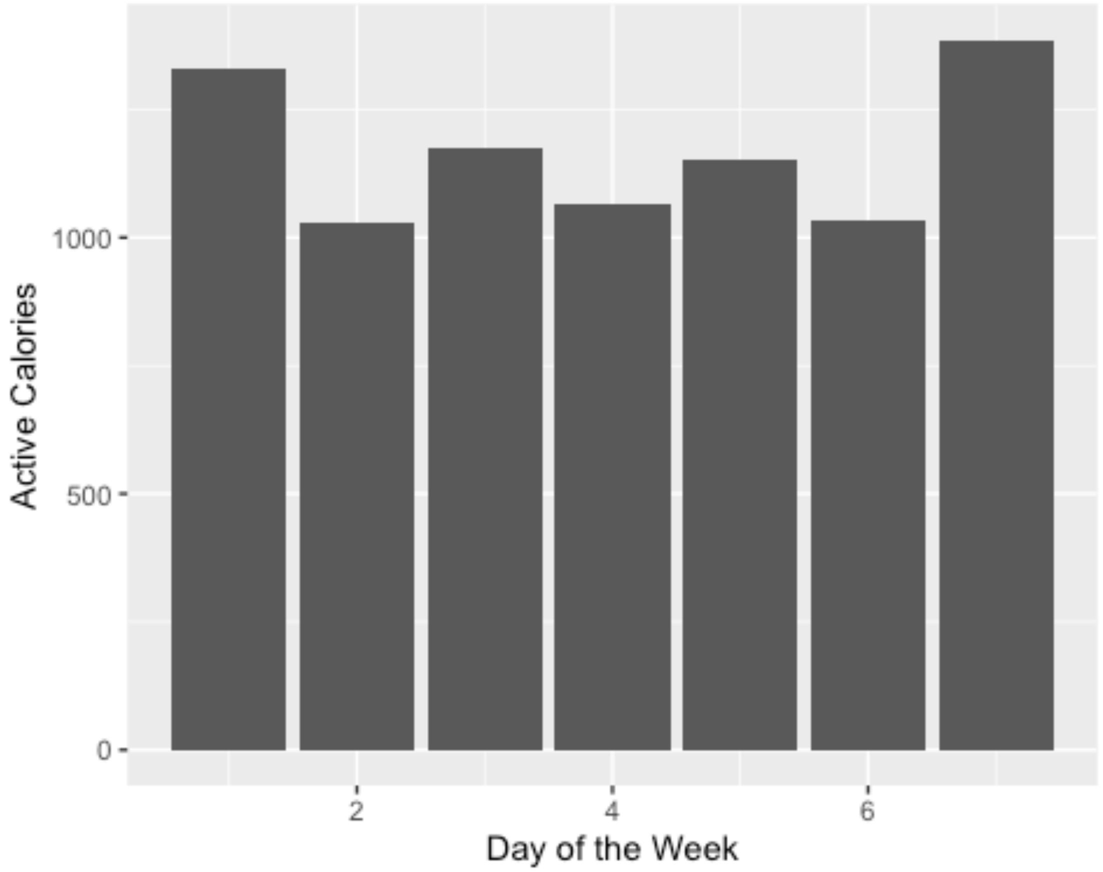
```
100
round(healthyhrv, 0)

## [1] 99

#Finds proportion of high HRV occurences
highhrv <- length(which(healthData$hrv >= 100))/length(healthData$hrv) * 100
round(highhrv,0)

## [1] 63
```

You can see that about 99% of the time I fall in the appropriate range.

I am also above the range 63% of the time given that most sites tend to cap HRV at 100ms. I would assume this is a fault of the apple watch sensor metrics as most sources say HRV is an intricate measure that takes lots of precision.

Which day of the week did I tend to burn the most calories?

```
healthData %>%
#Converts the string of Date into a Date-Time class
  mutate(Date = ymd_hms(Date)) %>%
#Converts the Date-Time class into just a date class
  mutate(Date = as.Date(Date)) %>%
#creates a variable that extracts the days of the week
  mutate(dow = wday(Date)) %>%
  group_by(dow) %>%
#Finds the average active calories for the days of the week
  summarize(act_cal = mean(actCal)) %>%
#Plots the active calories for days of the week in a bar graph
  ggplot(aes(dow, act_cal)) + geom_bar(stat="identity") + labs(x = "Day of
the Week", y = "Active Calories")
```



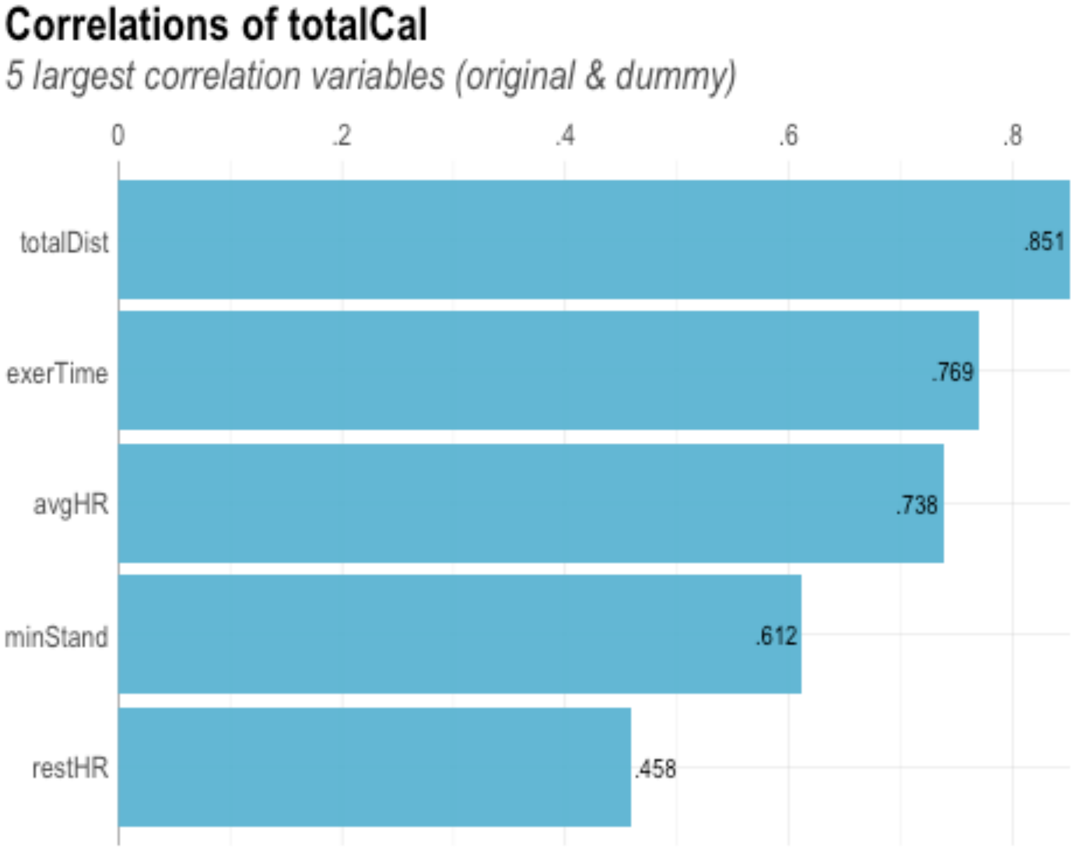
It appears that I burn a significant amount more calories on the weekends. This is most likely due to the fact that I like to do my longer runs on the weekend when I have more time available.

Which health metric is most correlated with the amount of calories I burn in a day?

```
library(extrafont)

## Registering fonts with R

extrafont::font_import()

## Importing fonts may take a few minutes, depending on the number of fonts
and the speed of the system.
## Continue? [y/n]

## Exiting.

#uses the lares library for cleaner data visualization
library(lares)
```

```
##
## Attaching package: 'lares'

## The following objects are masked from 'package:modelr':
##
##     mae, mape, mse, rmse
```

```
healthData %>%
  #adds an additional column to my data sat for total calories
  mutate(totalCal = actCal, restCal) %>%
  #find the top 5 biggest correlations with total calories burned
  corr_var(totalCal, top = 5)
```



The data displayed here demonstrates which health metrics are the biggest indicators of the total amount of calories I burn throughout a day using a pearson correlation coefficient. The pool of variables chosen only came from ones with a p-value less than .05 to demonstrate significance. As you can see, the total amount of distance I travel throughout the day, including both walking and running, is the most important metric when determining my caloric expenditure.

All of the other metrics in this top 5 list were fairly intuitive, however I am surprised to see resting heart rate at a .458 correlation coefficient. Even though it appears resting heart rate increases my caloric expenditure, I would assume that more exercise simply increases resting heart rate and it is not the other way around.

Multiple variable regression analysis to determine which health metric is most correlated with total calories all else being equal

```
#creates a variable for the total amount of calories throughout the day
  totalCal = healthData$actCal + healthData$restCal
#creates an independent linear regression model for each of the top 5
correlating variables to total calories
  model <- lm(totalCal ~ totalDist + exerTime + avgHR + minStand + restHR,
data = healthData)
  summary(model)
```

```
##
## Call:
## lm(formula = totalCal ~ totalDist + exerTime + avgHR + minStand +
##     restHR, data = healthData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -739.10  -35.36    7.56   54.76  140.21
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1642.6721   158.5454  10.361 2.57e-16 ***
## totalDist     26.3612     5.5280   4.769 8.46e-06 ***
## exerTime       3.9799     0.5418   7.346 1.72e-10 ***
## avgHR          9.1646     2.1278   4.307 4.77e-05 ***
## minStand       0.7110     0.2503   2.840  0.00575 **
## restHR        -5.7774     3.7811  -1.528  0.13056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 107.8 on 78 degrees of freedom
## Multiple R-squared:   0.86,  Adjusted R-squared:  0.851
## F-statistic: 95.83 on 5 and 78 DF,  p-value: < 2.2e-16
```

```
#Finds the standard error rate
  sigma(model)/mean(totalCal)
```

```
## [1] 0.03598834
```

Now for interpreting these results. Overall the p-value of all variables is less than 2.2 x 10^-16 which implies these variables are very significantly correlated with total calories. It appears that the lowest p-value within the set comes from exercise time at 1.72 x 10^-10 instead of the predicted total distance as from the previous display. The given t-statistics, which measure the significance of each variable in the outcome of total calories, reveals that exercise time is the highest predictor with average HR, total distance, minutes standing, and resting HR following in descending order.

The estimated correlation column explains how the total calories would change per unit increase in the variable specified. For example a 1 unit increase in total distance, which corresponds to 1 mile, would increase total calories by around 26. A one unit increase in exercise time, which is 1 minute, corresponds to an increase in 4 total calories.

The variance is measured under the multiple R-squared which is the overall correlation coefficient squared. The data shows that approximately 86% of the variance within total calories could be explained by these variables alone.

The residual standard error can tell us the error rate of this model being able to accurately predict the total calories throughout the day. Calculating error rate through dividing RSE by the average total cals gives me an error rate of about 3.5% for this model.

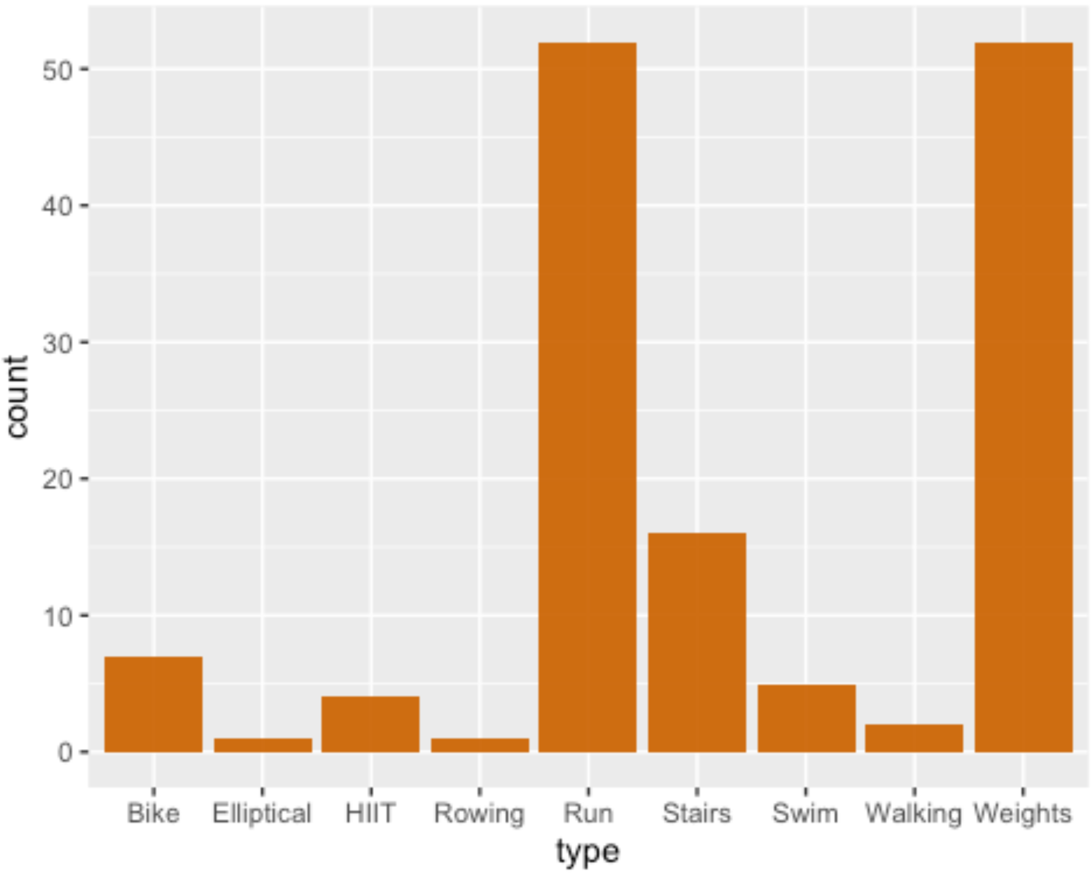
This multiple variable analysis proves my original thought that resting HR should not be that correlated with total calories as was shown by the previous top 5 ranking display. When controlling for all other factors, it in fact has a slight negative correlation with total calories and an insignificant p-value.

Which workouts did I tend to favor the most?

```
# makes all the default Apple workout types shorter to fit them in a bar
graph
workoutData$type <- str_replace(workoutData$type, "Cycling", "Bike")
workoutData$type <- str_replace(workoutData$type, "Running", "Run")
workoutData$type <- str_replace(workoutData$type, "Traditional Strength
Training", "Weights")
workoutData$type <- str_replace(workoutData$type, "Swimming", "Swim")
workoutData$type <- str_replace(workoutData$type, "Stair Climbing", "Stairs")
workoutData$type <- str_replace(workoutData$type, "High Intensity Interval
Training", "HIIT")
```

As you can see from the data below, I heavily favored running and weights over other forms of exercise

```
#makes a bar graph by workout type
ggplot(data = workoutData, aes(type)) + geom_bar()
```



Here is a display of the frequency and proportions of my workout type data.

```
# creates a summary of just my workout type data with frequency and
proportion
workoutData %>%
  count(type) %>%
  #applies the percentages into the round function for 1 decimal because it
displays 5 decimal places otherwise
  mutate(prop = (n/ sum(n) *100) %>% round(1))

##          type  n prop
## 1        Bike  7  5.0
## 2 Elliptical  1  0.7
## 3        HIIT  4  2.9
## 4     Rowing  1  0.7
## 5         Run 52 37.1
## 6     Stairs 16 11.4
## 7        Swim  5  3.6
## 8    Walking  2  1.4
## 9    Weights 52 37.1
```

Which workout tended to burn the most calories?

```
workoutData %>%
  group_by(type) %>%
#takes all the types and summarizes their median active calories
  summarize(medianActCal = median(actCal)) %>%
#adds a column for ranking the types based on median active calories
  mutate(rank = dense_rank(desc(medianActCal)))
```

```
## # A tibble: 9 × 3
##   type       medianActCal  rank
##   <chr>             <dbl> <int>
## 1 Bike              341.      3
## 2 Elliptical          8.7     9
## 3 HIIT              227.      4
## 4 Rowing             67.9     7
## 5 Run               573.      1
## 6 Stairs            387.      2
## 7 Swim               65.1     8
## 8 Walking           106.      6
## 9 Weights           135.      5
```

Based on the median amount of active calories burned during my workouts you can see that running tends to be the highest at 573 calories while the elliptical is the least at only 8.7 calories.

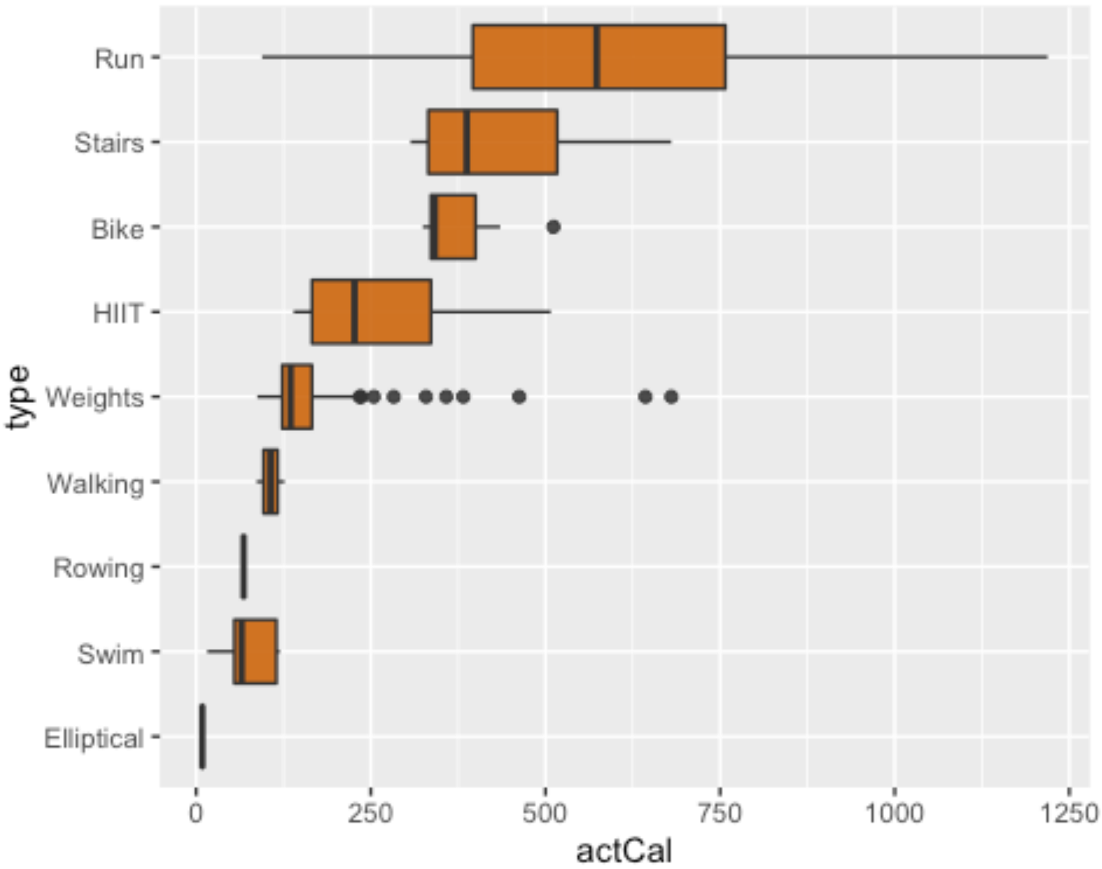
This elliptical data is a form of a data entry error and should not be interpreted as such. I do not use the elliptical and must have accidentally clicked on it temporarily when trying to start an alternative workout.

```
workoutData %>%
#determines how many occurences of elliptical workouts there were
  count(type == "Elliptical")
```

```
##   type == "Elliptical"   n
## 1                FALSE 139
## 2                 TRUE   1
```

As you can see I only had 1 elliptical workout and it was a mistake on my apple watch.

```
#organizes a boxplot display in increasing order based on median active
calories
workoutData %>%
  ggplot(aes(type)) + geom_boxplot(aes(reorder(type, actCal, FUN = median),
actCal)) + coord_flip()
```



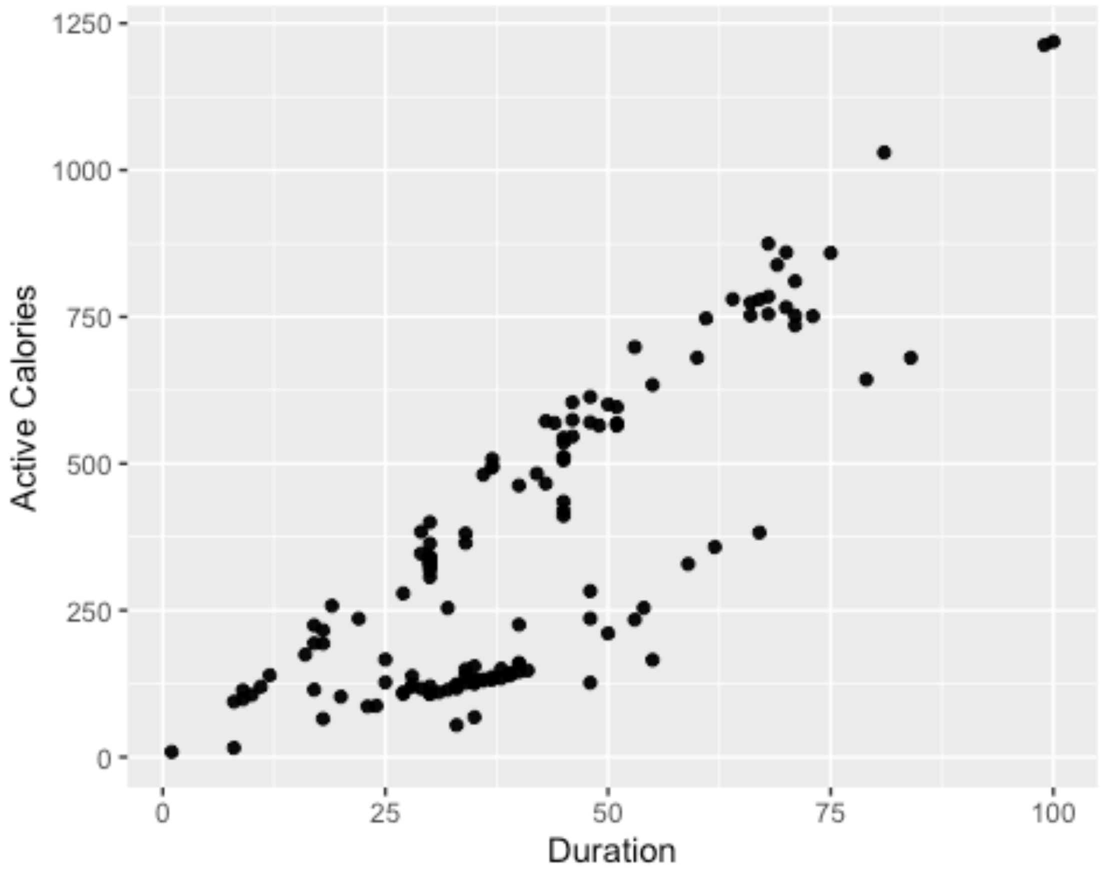
```
#I used the coordinate flip technique here to allow for a cleaner display of
workout types instead of individually changing all of their names
```

You can see more clearly which workouts tended to have the highest active calories in an increasing order within the box plots.

There are some interesting observations to be made regarding some of the workout types. For example, the weight lifting category has a lot of potential outliers, meaning there were several occasions in which a workout was over 1.5 times the 3rd quartile. I am sure this observation comes about due to the times where I have forgotten to switch my workout type after a weight lifting session as I tend to do stairs directly afterwards.

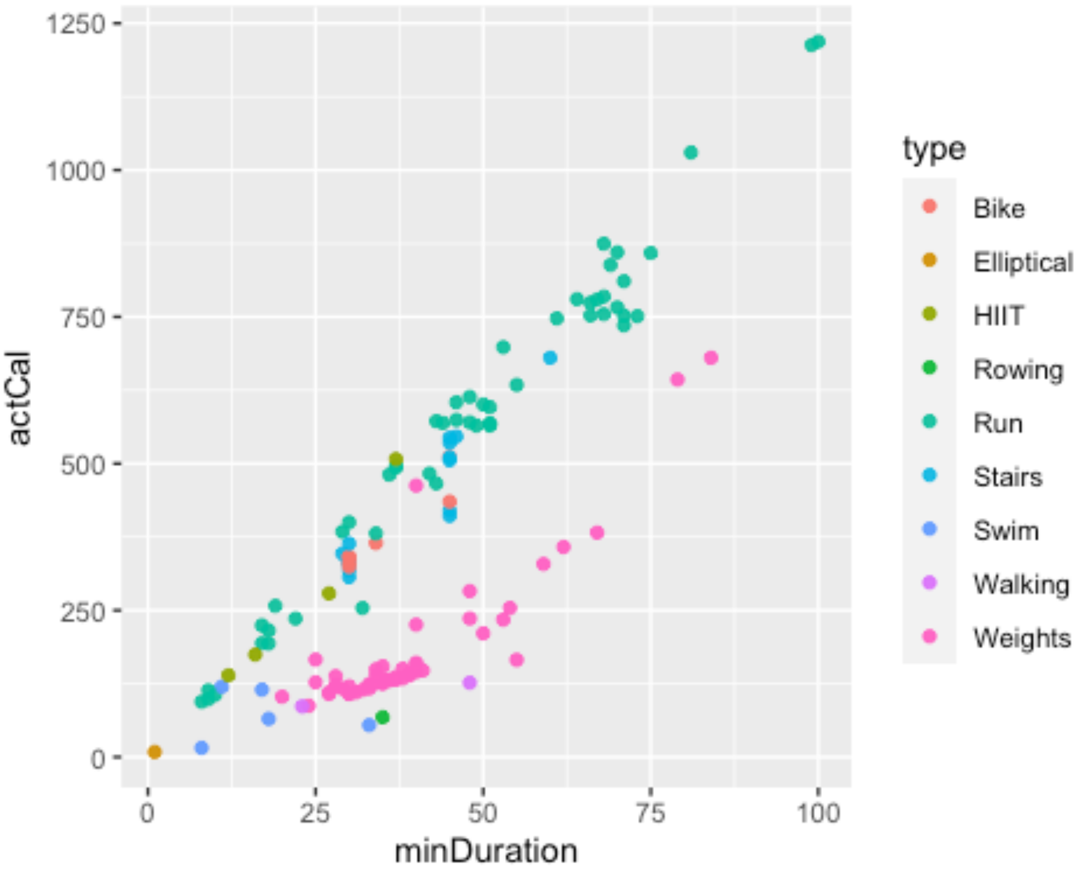
How does the duration of my workouts correlate with active calories?

```
#Creates a new data fram that has the duration column in minutes
minutes <- workoutData %>%
#Turns the duration from a string into a date-time class
  mutate(duration = hms(duration)) %>%
#Extracts the total minutes in each workout
  mutate(minDuration = hour(duration)*60 + minute(duration))
ggplot(minutes, aes(minDuration, actCal)) + geom_point() + labs(x =
"Duration", y = "Active Calories")
```



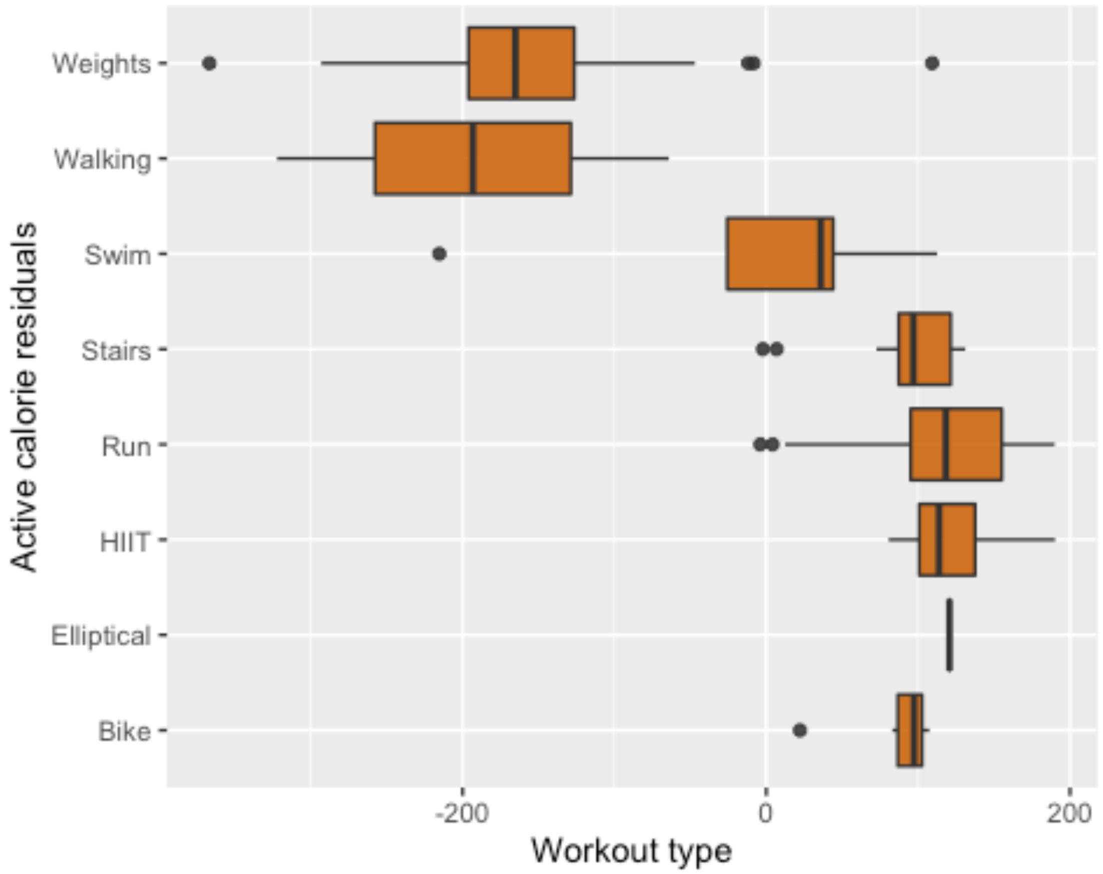
This display shows a very clear upward trend in the active calories as duration increases. However there is a small cluster that seems to have a weaker trend for some reason.

```
#Creates a plot for the duration of my workouts and the amount of calories
burned within them. This time there is a color at each point for a specific
workout type
ggplot(data = minutes, aes(minDuration, actCal, col = type)) + geom_point()
```



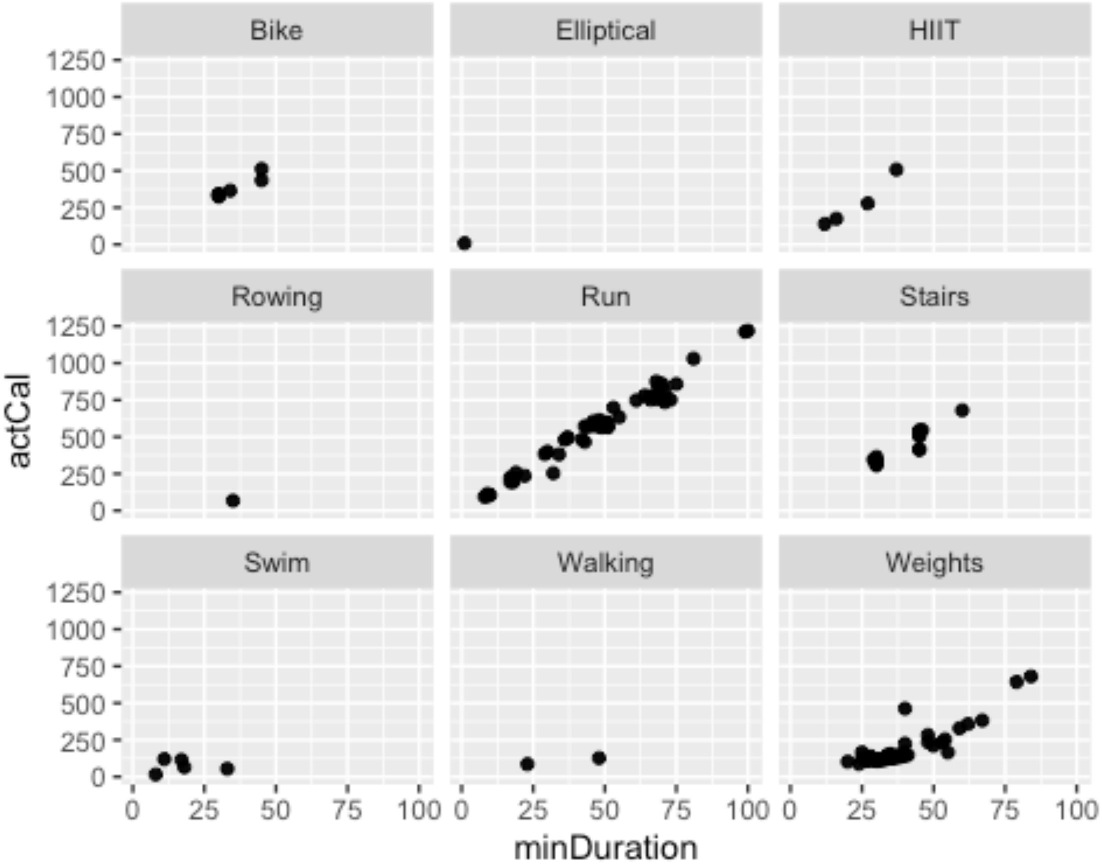
This time you can see that this small cluster all happens to fall under the same workout type of weightlifting. This makes sense as it is not as cardio intensive as the other workout types listed here.

```r
#Creates a linear model for the active calories against duration
mod_exer <- lm(actCal ~ minDuration, data = minutes)

#Creates a new data frame without the least performed workouts
minutes2 <- minutes %>%
  filter(type != "Rowing") %>%
  filter(type != "Swimming") %>%
  filter(type != "Eliptical") %>%
  add_residuals(mod_exer, "actcalresid")

#Displays the residuals of the workout types against the regression line
ggplot(minutes2, aes(type, actcalresid)) + geom_boxplot() + coord_flip() +
labs(x = "Active calorie residuals", y = "Workout type")
```



From the display above it becomes apparent that weight lifting and walking tend to be close to 200 active calories below the regression line.

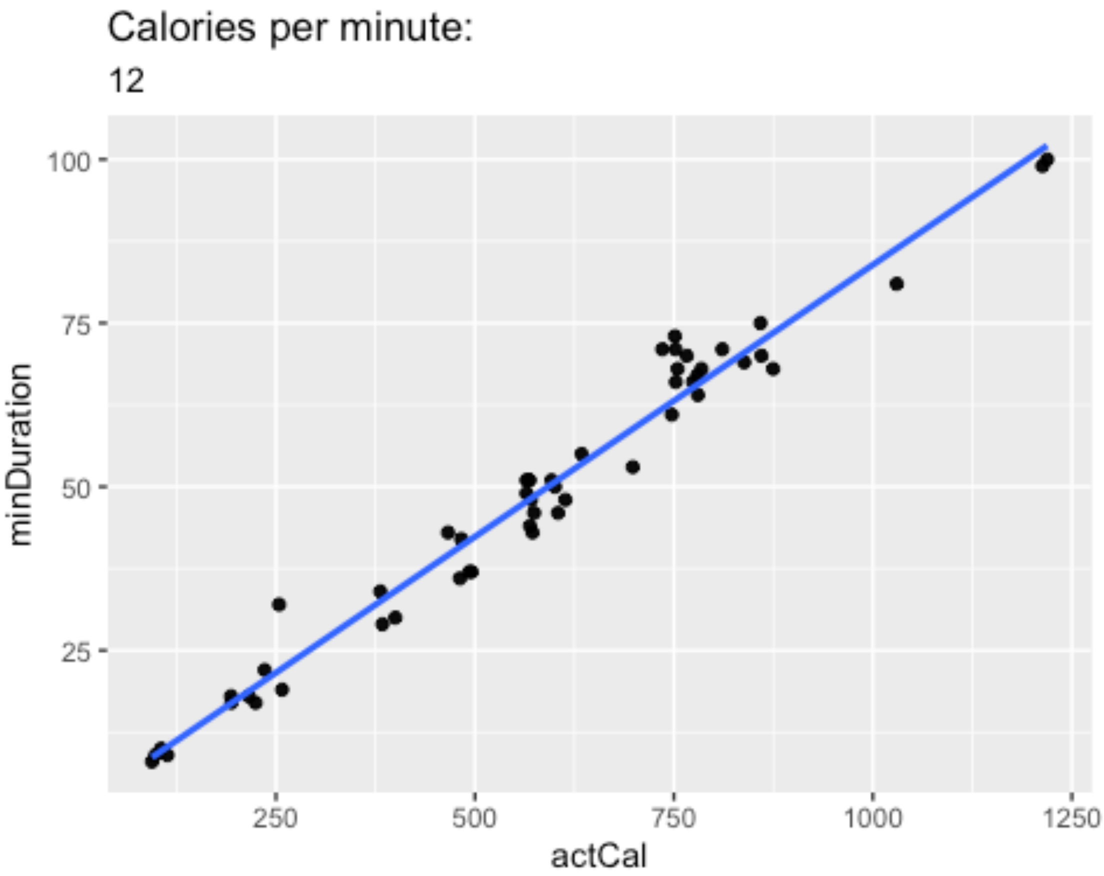
```
#Creates a plot for the duration of my workouts and the amount of calories
burned within them
ggplot(data = minutes, aes(minDuration, actCal)) + geom_point() +
facet_wrap(~ type, nrow = 3)
```



From this display you can see each type of workouts trend with duration. Running and stair climbing have large enough data points to see a very strong linear relationship. On the other hand it appears that the relationship of weightlifting with time is potentially exponential. However I believe this goes back to the point I made about me forgetting to turn off the weight lifting session when I went to do cardio afterwards.

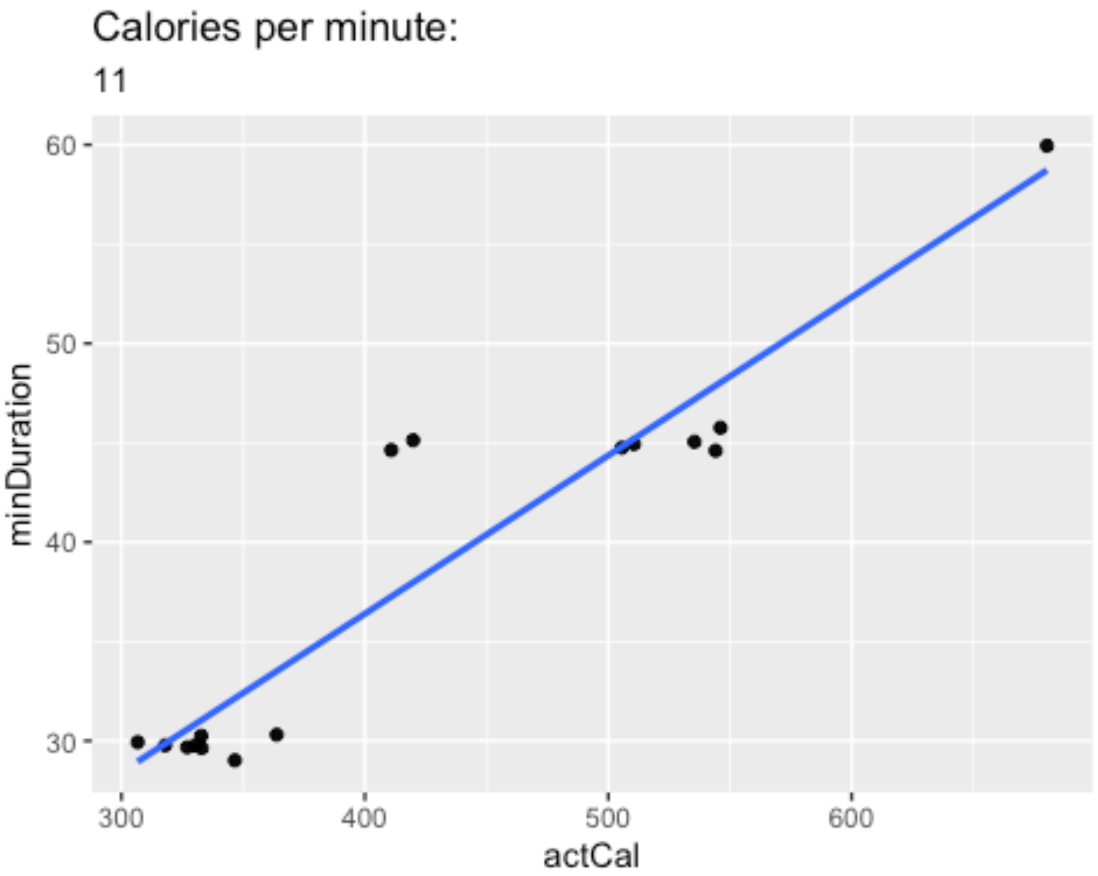
What is the most efficient type of workout if calorie expenditure is my priority?

```
#creates a tibble with just the running workouts
run <- minutes %>%
    filter(type == "Run") %>%
  select(!c(distance,speed))

#Extracts the slope from the list of coefficients from a linear model and
rounds it to 2 digits
slope=format(signif(coef(lm(actCal~minDuration, data = run))[2],2))
#Plots the minutes against active calories for all my run data
ggplot(run, aes(actCal,minDuration)) + geom_point() + geom_smooth(method =
"lm", se = F) + ggtitle("Calories per minute:", slope)

## `geom_smooth()` using formula 'y ~ x'
```



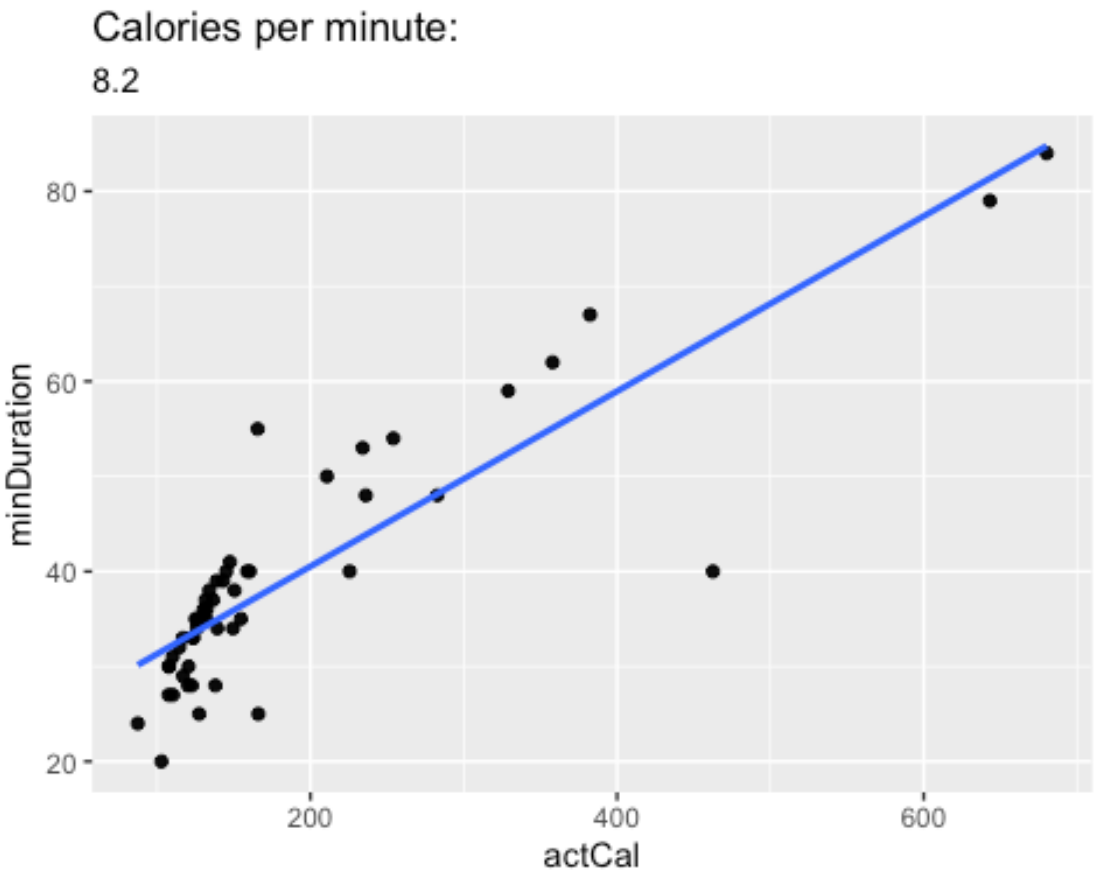
From all of my runs in the past 90 days I burned around 12 calories per minute.

```r
stair <- minutes %>%
    filter(type == "Stairs")

slope=format(signif(coef(lm(actCal~minDuration, data = stair))[2],2))
#Plots the minutes against active calories for all my run data
ggplot(stair, aes(actCal,minDuration)) + geom_jitter() + geom_smooth(method =
"lm", se = F) + ggtitle("Calories per minute:", slope)

## `geom_smooth()` using formula 'y ~ x'
```



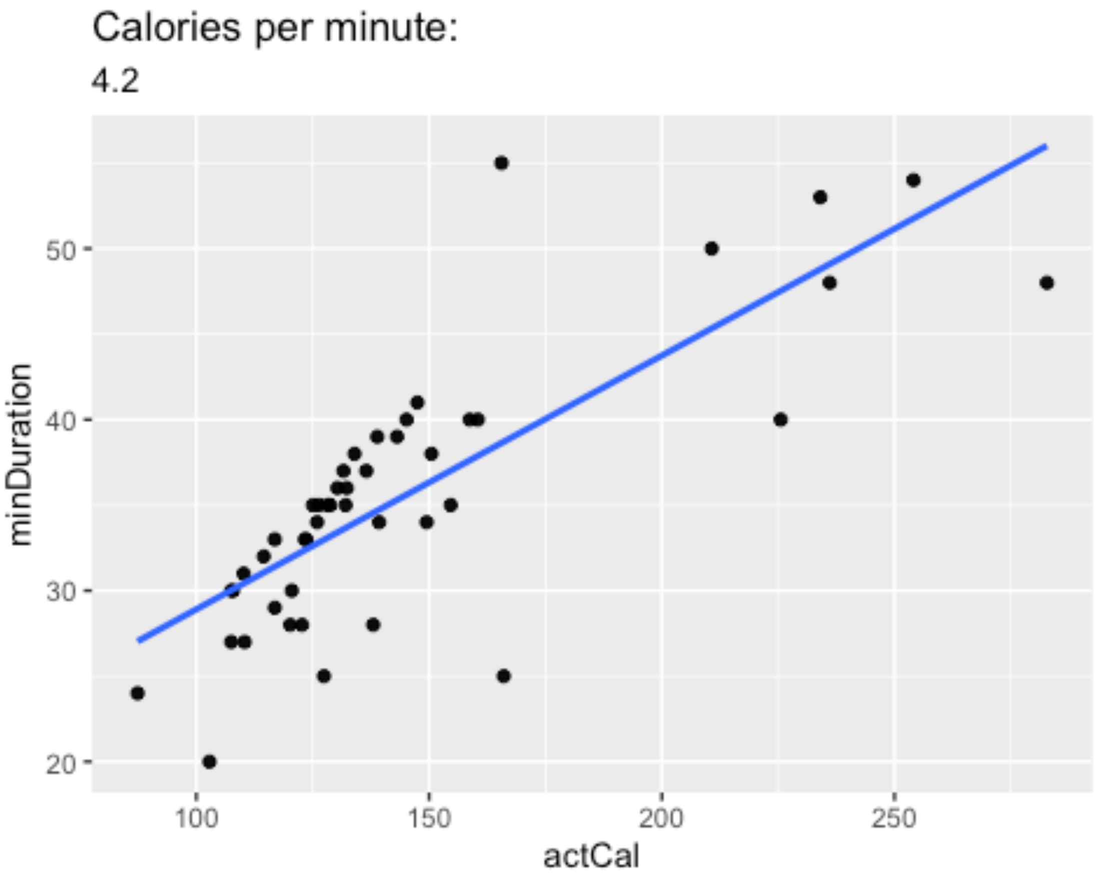
From all of my stair master workouts I burned around 11 calories per minute

```
lift <- minutes %>%
    filter(type == "Weights")
slope=format(signif(coef(lm(actCal~minDuration, data = lift))[2],2))
#Plots the minutes against active calories for all my run data
ggplot(lift, aes(actCal,minDuration)) + geom_point() + geom_smooth(method =
"lm", se = F) + ggtitle("Calories per minute:", slope)

## `geom_smooth()` using formula 'y ~ x'
```



From all my weight lifting workouts I burned around 8.2 calories per minute. However, I believe this is an overestimate due to the times where I continued on to some form of cardio after weights and never changed the workout type.

```
lift2 <- minutes %>%
    filter(type == "Weights", minDuration < 60, actCal < 300)
slope=format(signif(coef(lm(actCal~minDuration, data = lift2))[2],2))
#Plots the minutes against active calories for all my run data
ggplot(lift2, aes(actCal,minDuration)) + geom_point() + geom_smooth(method =
"lm", se = F) + ggtitle("Calories per minute:", slope)

## `geom_smooth()` using formula 'y ~ x'
```



After controlling the data points being plotted for those that are below 60 minutes in duration and lower than 300 calories it appears that lifting weights burns around 4.2 calories per minute which makes much more sense given my observations in real time. I used these parameters because my lifts never extend past 1 hour and they never reach 300 calories which would indicate I was likely doing some form of cardio instead of lifting.

Overall, running appears to be the most efficient form of calorie expenditure for my time at a total of 12 calories per minute.

Real world applications

This data analysis consistently reflects assumptions I would have had about the outcomes given real world observations. For example, the increasing trend in resting heart rate over time reflecting scenarios in my life that would lead to such an outcome.

Most of my workouts coming from running and weightlifting also make sense as those are my top priorities. The weekends being my most calorie intensive days is reflected by the fact that my longer runs take place on those days.

The metrics most correlated with the total calories burned throughout the day did not totally reflect my assumptions about how the data would turn out. However after multiple variable regression, all of the top correlating variables made complete sense. Lastly, my most efficient workout for time being running is also reflective of real world application as it tends to be the most difficult task to physically perform.

Limitations

There were no real big limitations within the data sets that I had. The only thing that would have made this analysis much better would have been a larger number of data points. Given the fact that I have only had the apple watch for 90 days, this was not possible.

Additional variables such as sleep metrics would be interesting to add in as I have lots of questions regarding physical performance and heart rate data and how they relate to sleep.