PEAK PERFORMANCE

A Fitness Tracker Data Exploration

Payton Wiscovitch
C00122694
Informatics 512
Dr. Tozal
Spring 2023

Table of Contents

The Data

Description:

The following data is a collection of my own personal health and fitness information captured from a Whoop fitness and recovery device. The data was captured over approximately two years spanning from February 28, 2020, to May 2, 2022. However, there are some periods of missing data for dates which the device was not worn. Also, according to Whoop user information, it takes approximately one month for the device to properly calibrate to your physiology. For those reasons I will focus this analysis on a subset of one year of the data from April 1, 2020, to March 31, 2021. The data was acquired by requesting a full export of the data collected by Whoop during the time for which I was a subscriber. The data is distributed over four datasets: physiological, workouts, sleeps, and journal entries. The physiological dataset contains information from both the sleep and workouts dataset but is not comprehensive as it does not include all variables from the respective datasets.

Background Information:

The Whoop strap is a fitness and health wearable that is a subscription-based service. The device itself has no interactive display and solely functions as a monitoring and data collection device. The data captured is collected by Whoop and processed using proprietary algorithms to produce personalized performance data for the user via the Whoop App. The purpose is to help the user to use these insights to help advance their personal performance. At the time that I subscribed to this service the main purpose was to provide the user with a Recovery Score. The Recovery Score was a numerical value that corresponded to a color-coded system. The user could be in the red, yellow, or green. Green indicated a high level of recovery, and that the user was in prime condition to perform athletically. [1] The device also tracked sleep metrics offering a Sleep Performance %, tracked workouts which resulted in an Activity Strain Score, and offered the user the ability to track habits in a daily journal to see how these factors contributed to recovery, sleep, and strain. Whoop tracks data on a cycle basis instead of by the day. A cycle runs from the user's wake time to the onset of sleep. Each row of the data corresponds to a cycle and the data collected as a result of that cycle.

Analysis Plan and Goals:

In the following analysis I want to explore the data to gain insights and discover patterns about behaviors as they relate to my recovery, sleep, and strain/workouts. I will analyze each data set separately to evaluate the key metrics that contribute to recovery, sleep, and strain/workouts. I will then compare this data with the journal entry data to see what effect if any daily habits had on my scores. I will also explore which factors lead to having a high recovery of >= 90 and low recovery days which is defined based on recovery of <= 33. I will take a closer look at the highest and lowest recovery days compared with the variables and journal entries for those days to see if I can determine which factors contribute to the highest and lowest recovery. For example, how do different sleep cycles impact recovery? What daily habits impact recovery the most and which should I try to use to increase my recovery scores? How do different types of exercise impact recovery? The goal of this analysis is to gain insight into the factors and behaviors that drive my own personal performance so that I may use them to achieve peak performance.

Data Dictionary

Variable Name	Short Name	Data Type	Description
Cycle Start Time	cstart	Timestamp	Starting date and time for data collection
Cycle End Time	cend	Timestamp	End date and time for data collection
Recovery Score %	recovery	Integer	Calculation of HRV, RHR, sleep performance and respiratory rate
Recovery Class	Score	Categorical	Green: sufficient, recovery is 67% or above body is primed to take on heavy training load Yellow: adequate, recovery is between 34-66% body can adapt to high training load, but performance could be compromised on the lower end Red: low, recovery is 33% or less, consider a low load day or day off.
Resting Heart Rate	rhr	Integer	Heart rate when the body is at rest
Heart Rate Variability	hrv	Integer	The variance in time between the beats of your heart
Day Strain	daystrain	Integer	Strain put on cardiovascular system and body
Energy Burned	cals	Integer	Calories burned
Max HR	maxhr	Integer	Maximum heart rate of the day
Average HR	avghr	Integer	Average heart rate throughout the day
Sleep Onset	sleepstart	Timestamp	Time sleep state was detected as starting
Wake Onset	wake	Timestamp	Time waking was detected
Sleep Performance %	sleepscore	Integer	= time asleep/sleep need
Respiratory Rate	resprate	Integer	Number of breaths per minute
Asleep Duration	asleep	Integer	Number of minutes slept
In Bed Duration	inbed	Integer	Number of minutes in bed
Light Sleep Duration	lightsleep	Integer	Minutes spent in light sleep
Deep (SWS) Sleep	deepsleep	Integer	Minutes spent in deep sleep
REM duration	rem	Integer	Minutes spent in REM sleep
Awake Duration	awake	Integer	Minutes during sleep cycle spent awake

Sleep Need	sleepneed	Integer	Minutes of sleep needed based on baseline, sleep debt, strain, and naps.
Sleep Dept	sleepdebt	Integer	Carryover of leftover unslept minutes from previous sleep cycle
Sleep Efficiency %	sleepeff	Integer	Time in bed vs. time slept and quality of sleep
Workout Start time	wostart	Timestamp	Time workout began
Workout end time	woend	Timestamp	Time workout ended
Duration(workout)	duration	Integer	Length of workout in minutes
Activity Name	activity	Categorical	Name or type of workout
Activity Strain	actstrain	Float	A calculation of time spent in each HR zone
HR Zone 1	hrz1	Integer	Minutes spent in HR zone 1
HR Zone 2	hrz2	Integer	Minutes spent in HR zone 2
HR Zone 3	hrz3	Integer	Minutes spent in HR zone 3
HR Zone 4	hrz4	Integer	Minutes spent in HR zone 4
HR Zone 5	hrz5	Integer	Minutes spent in HR zone 5

Journal Entry Questions:

Journal Entry answers are Boolean where TRUE indicates a YES response and FALSE indicates a NO response. These entries will be explored by how they relate to green/high recovery days with recovery scores of >= 67 and all red recovery days.

Have any alcoholic drinks?
Take an anti-inflammatory NSAID?
Wear blue-light blocking glasses before bed?
Have any caffeine?
View a screened device while in bed?
Consume meat?
Take a melatonin supplement?
Sleep in the same bed as usual?
Experience any stress?
Spend time stretching?
Experiencing COVID-19 symptoms?
Injured / Not Injured
Sick / Not Sick
Single / Not Single

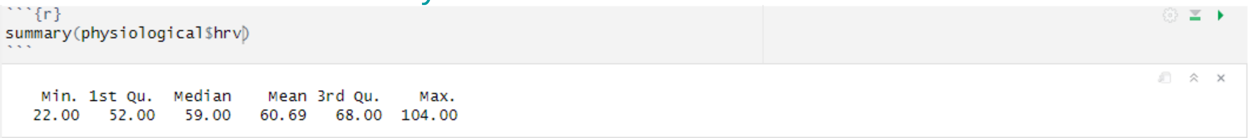
Baseline Metrics

According to documentation on the Whoop website, the user's baseline metrics are used in the calculation of recovery, sleep need, and strain. To better understand trends in the data, it is important to understand what the baseline metrics are. We will look at the summary statistics for key variables to compare deviations from baseline to better understand how they impact recovery. Research on the Whoop device explains that recovery score is a calculation of HRV, RHR, sleep performance, and respiratory rate.[2] Sleep performance is the difference in the amount of sleep achieved vs. the amount of sleep needed. Sleep efficiency is the difference between the amount of time slept vs. the amount of time spent in bed. Sleep need is a calculation of baseline sleep need, sleep debt, strain, and naps. Activity strain is a measure of cardiovascular exertion and quantifies the amount of time spent in each heart rate zone during a recorded activity. To better understand my baseline metrics, we will look at summary statistic for each of the mentioned metrics. The averages should serve as a baseline.

Recovery Metrics

I will use the following summary statistics to explore and compare data related to high and low recovery scores. The mean/average will serve as a baseline for each metric. Exploring deviations from these baselines will help determine factors contributing to high and low recovery.

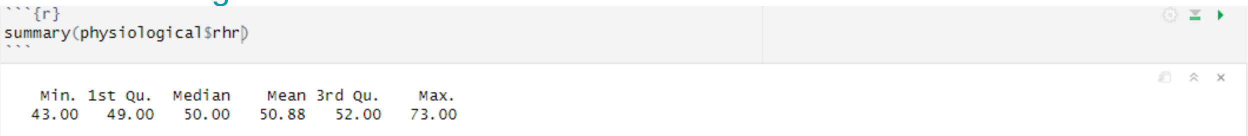
HRV – Heart Rate Variability



~~~
```{r}
summary(physiological$hrv)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.00   52.00   59.00   60.69   68.00  104.00 
```

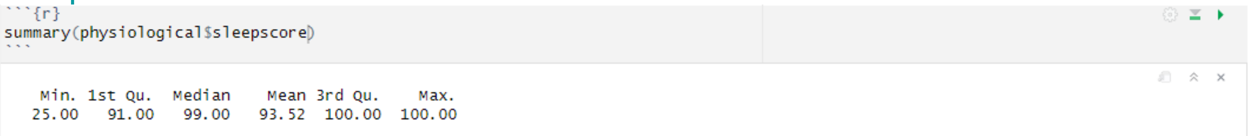
RHR – Resting Heart Rate



~~~
```{r}
summary(physiological$rhr)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  43.00   49.00   50.00   50.88   52.00   73.00 
```

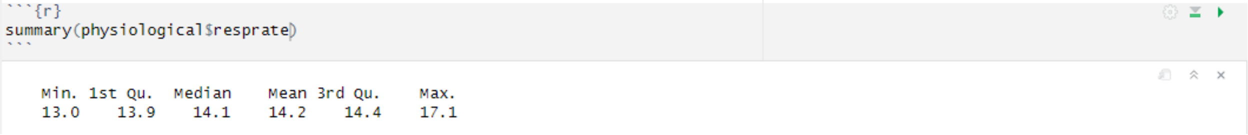
Sleep Performance



~~~
```{r}
summary(physiological$sleepscore)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  25.00   91.00   99.00   93.52  100.00  100.00 
```

Respiratory Rate



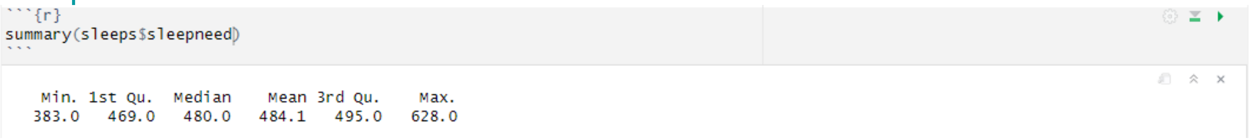
~~~
```{r}
summary(physiological$resprate)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   13.0    13.9    14.1    14.2    14.4    17.1 
```

Sleep Metrics

The following sleep metrics will be used in understanding how deviation from the average impacts sleep need, sleep performance and ultimately high and low recovery days.

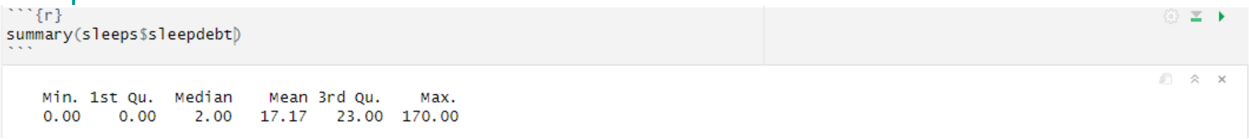
Sleep Need



~~~
```{r}
summary(sleeps$sleepneed)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  383.0   469.0   480.0   484.1   495.0   628.0 
```

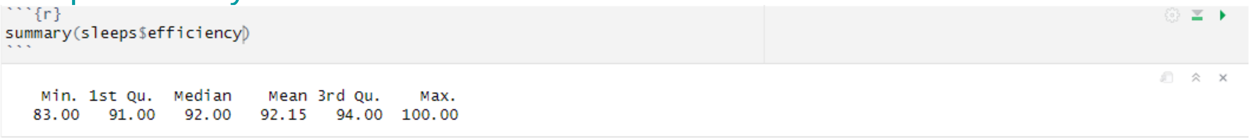
Sleep Debt



~~~
```{r}
summary(sleeps$sleepdebt)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.00    2.00   17.17   23.00  170.00 
```

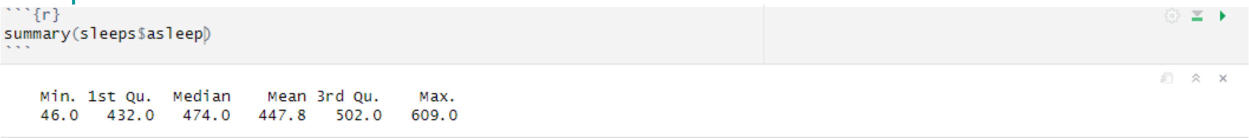
Sleep Efficiency



~~~
```{r}
summary(sleeps$efficiency)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  83.00   91.00   92.00   92.15   94.00  100.00 
```

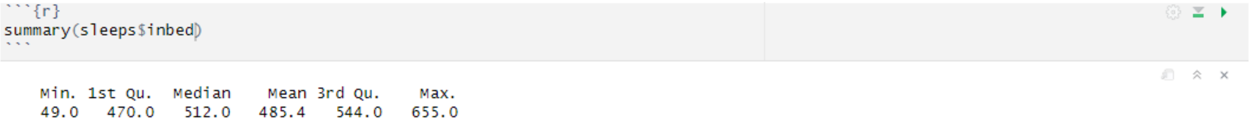
Sleep in Minutes



~~~
```{r}
summary(sleeps$asleep)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   46.0   432.0   474.0   447.8   502.0   609.0 
```

Minutes in Bed



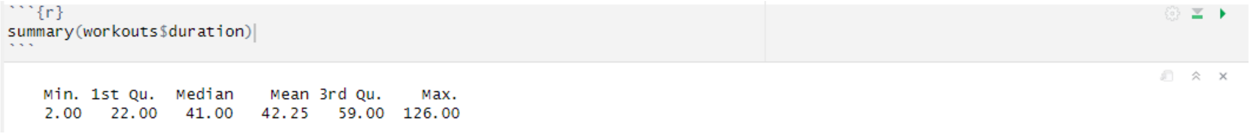
~~~
```{r}
summary(sleeps$inbed)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   49.0   470.0   512.0   485.4   544.0   655.0 
```

Workout Metrics

Workout metrics will be used in evaluating how deviations from average contribute to activity strain. Strain plays a role in sleep need which ultimately impacts recovery.

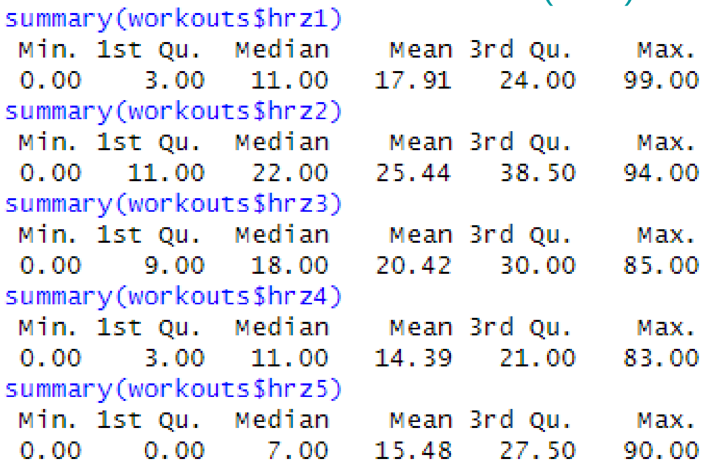
Workout Duration in Minutes



~~~
```{r}
summary(workouts$duration)
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00   22.00   41.00   42.25   59.00  126.00 
```

% of time in each Heart Rate Zone (HRZ)

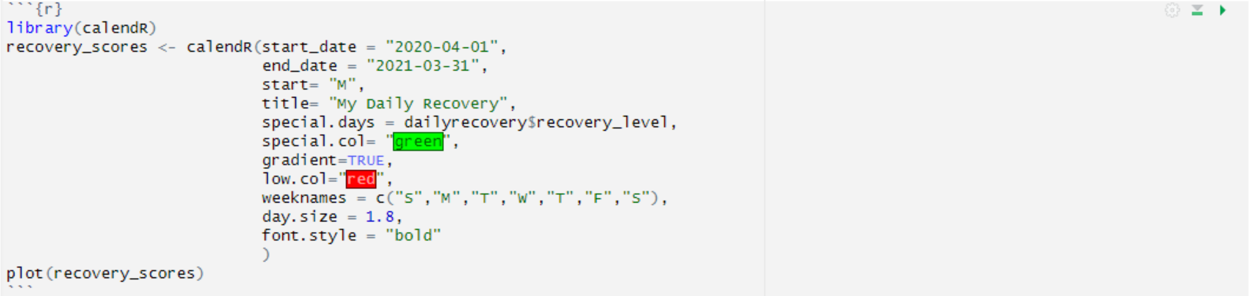


```
summary(workouts$hrz1)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    3.00   11.00   17.91   24.00   99.00
summary(workouts$hrz2)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00   11.00   22.00   25.44   38.50   94.00
summary(workouts$hrz3)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    9.00   18.00   20.42   30.00   85.00
summary(workouts$hrz4)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    3.00   11.00   14.39   21.00   83.00
summary(workouts$hrz5)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    0.00    7.00   15.48   27.50   90.00
```

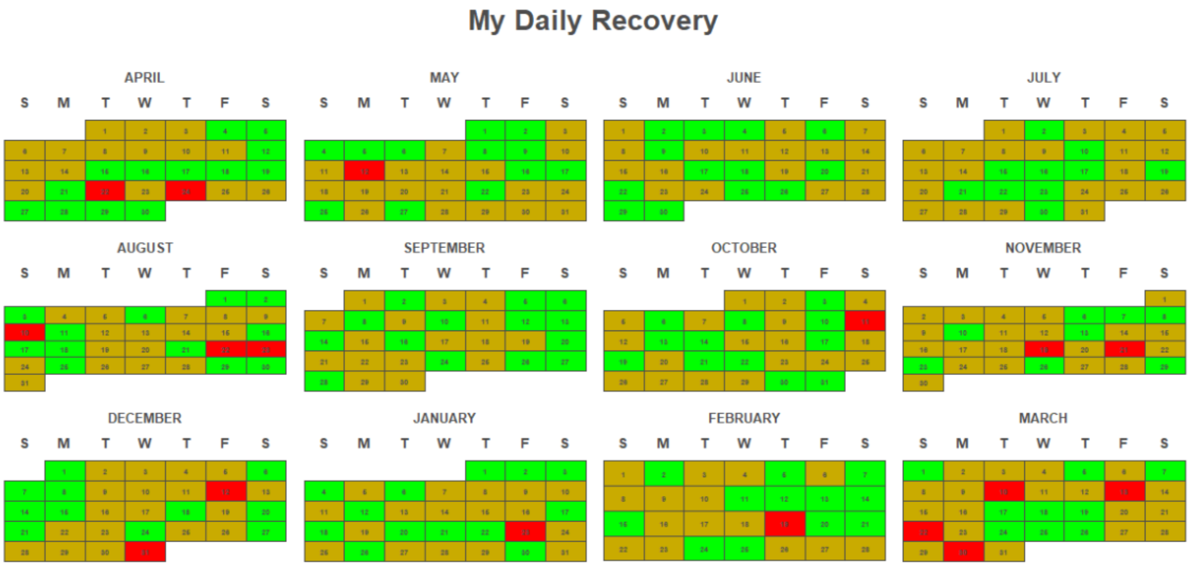
Recovery Exploration and Analysis

Recovery is key in athletic performance. Afterall, you can’t drive a car that doesn’t have gas. This is how we can think about recovery. Recovery is the fuel that drives the body and provides the ability to perform, expend energy, and take on cardiovascular exertion.

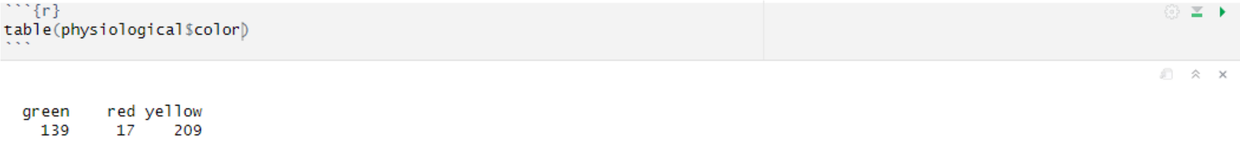
Visualizing a Year of Recovery



~~~
```{r}
library(calendR)
recovery_scores <- calendR(start_date = "2020-04-01",
                           end_date = "2021-03-31",
                           start= "M",
                           title= "My Daily Recovery",
                           special.days = dailyrecovery$recovery_level,
                           special.col= "green",
                           gradient=TRUE,
                           low.col="red",
                           weeknames = c("S","M","T","W","T","F","S"),
                           day.size = 1.8,
                           font.style = "bold"
                           )
plot(recovery_scores)
```
~~~



Whoop recovery scores are categorized as green, yellow, and red. Green recovery is high recovery and indicates the body’s readiness to take on cardiovascular strain, stress, and exertion. Yellow indicates moderate recovery. While red indicates low recovery and the need for rest and recovery procedures to recuperate.



~~~
```{r}
table(physiological$color)
```
~~~

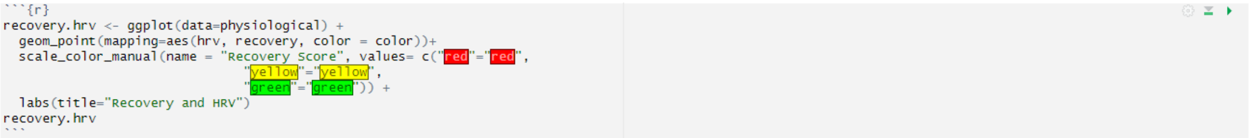
```
 green    red yellow 
   139     17    209 
```

Over the course of the year, I spent more days with moderate/yellow recovery than high or low recovery. I only spent 17/365 days in the red with low recovery. So, how can I turn some of these moderate recovery days into high recovery days and what factors contributed to the 17 low recovery days? I will specifically explore the 17 lowest recovery days and the highest recovery days of >= 90 later in this analysis to see if I can develop criteria for increasing my recovery and avoiding low recovery days.

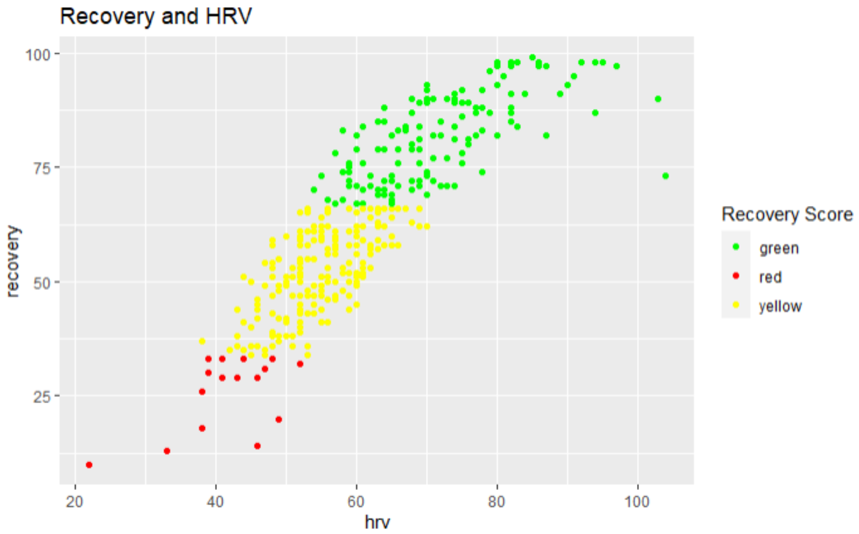
Recovery and Key Metrics

According to documentation from Whoop, recovery is a calculation of HRV, RHR, sleep performance, and respiratory rate. So, I will look at some visualizations of these metrics alongside recovery.

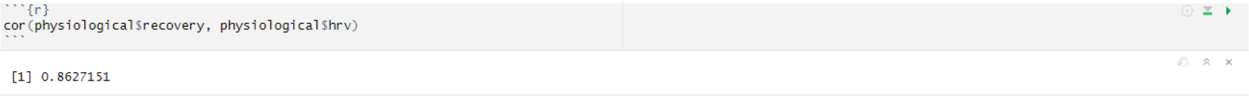
Recovery and HRV



~~~
```{r}
recovery.hrv <- ggplot(data=physiological) +
  geom_point(mapping=aes(hrv, recovery, color = color))+
  scale_color_manual(name = "Recovery Score", values= c("red"="red",
                                                 "yellow"="yellow",
                                                 "green"="green")) +
  labs(title="Recovery and HRV")
recovery.hrv
```
~~~



Based on this plot of recovery and HRV, we can see that there is a clear positive linear relationship between the two variables. As HRV increases so does recovery. We can confirm this by calculating the Pearson Correlation Coefficient to determine if these variables are in fact correlated, how strongly they are correlated and if the variables do in fact have a positive relationship.

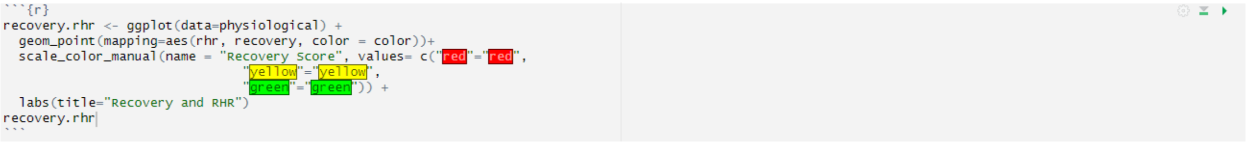


~~~
```{r}
cor(physiological$recovery, physiological$hrv)
```
~~~

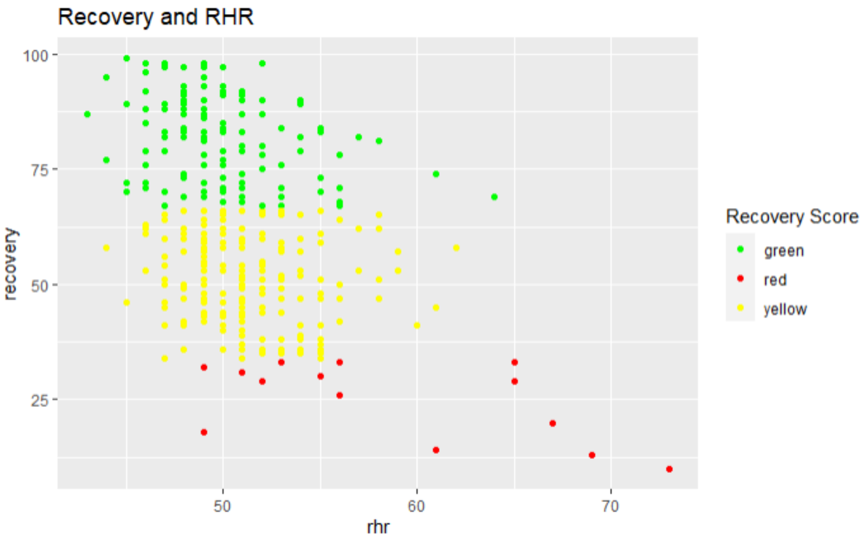
```
[1] 0.8627151
```

The closer the correlation coefficient is to -1 or 1 the stronger the negative or positive relationship between the variables. With a correlation coefficient of 0.86, I can say with confidence that recovery and HRV are highly correlated with a strong positive relationship. Thus, HRV should be a good prediction variable for recovery.

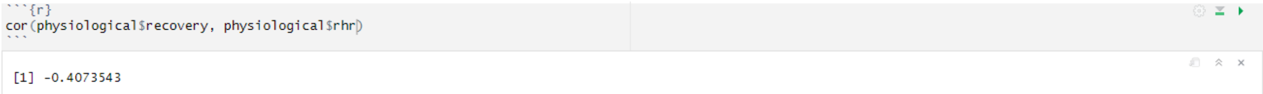
Recovery and RHR



~~~
```{r}
recovery.rhr <- ggplot(data=physiological) +
  geom_point(mapping=aes(rhr, recovery, color = color))+
  scale_color_manual(name = "Recovery Score", values= c("red"="red",
                                                        "yellow"="yellow",
                                                        "green"="green")) +
  labs(title="Recovery and RHR")
recovery.rhr
```
~~~



It appears that recovery and resting heart rate have a weak negative linear relationship. As resting heart rate increases, recovery seems to decrease. While I know that RHR is a variable used to calculate recovery, it appears that the relationship is weak and that RHR does not play as big a factor in the calculation as other variables.

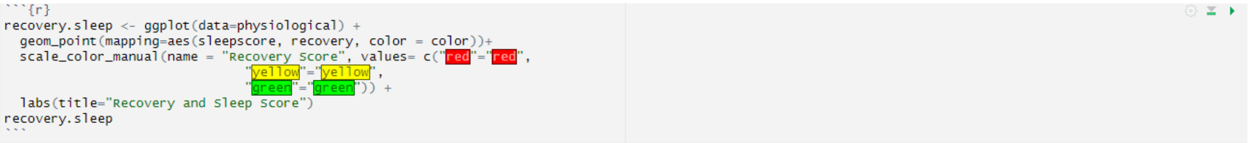


~~~
```{r}
cor(physiological$recovery, physiological$rhr)
```
~~~

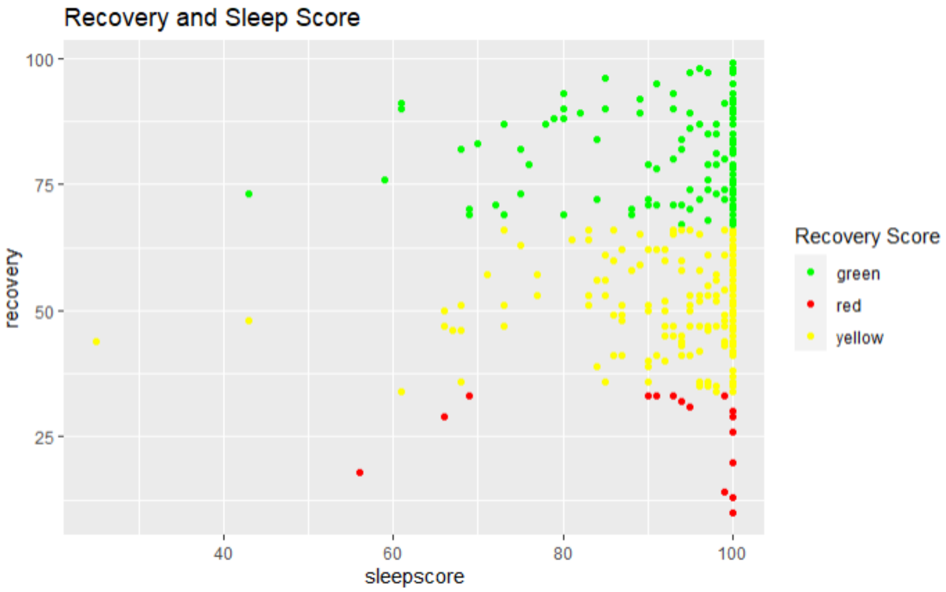
```
[1] -0.4073543
```

After computing the correlation coefficient, I can confirm that there is a weak negative relationship between the two variables.

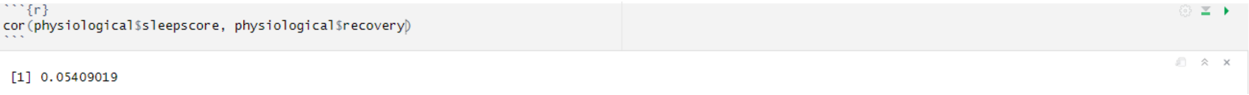
Recovery and Sleep Performance



~~~
```{r}
recovery.sleep <- ggplot(data=physiological) +
  geom_point(mapping=aes(sleepscore, recovery, color = color))+
  scale_color_manual(name = "Recovery Score", values= c("red"="red",
                                                       "yellow"="yellow",
                                                       "green"="green")) +
  labs(title="Recovery and Sleep Score")
recovery.sleep
```
~~~

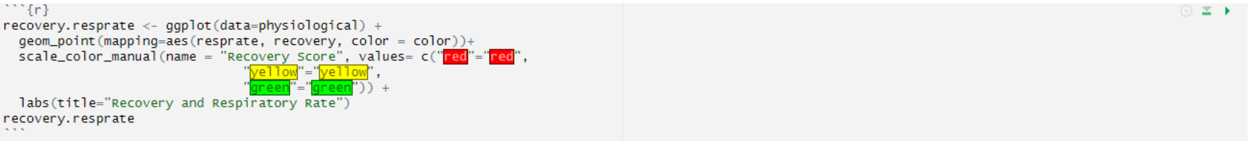


There is no linear relationship between recovery and sleep performance. This is confirmed with a correlation coefficient of 0.05.

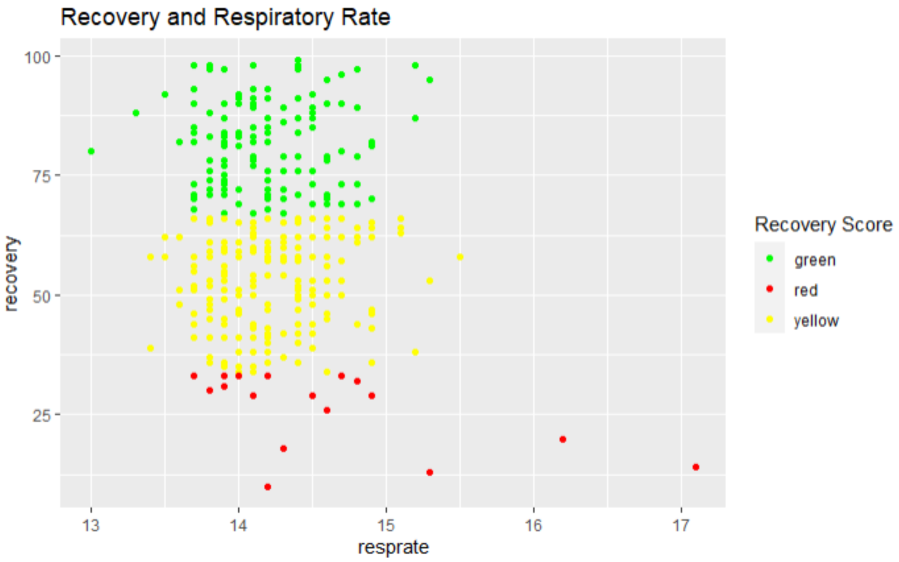


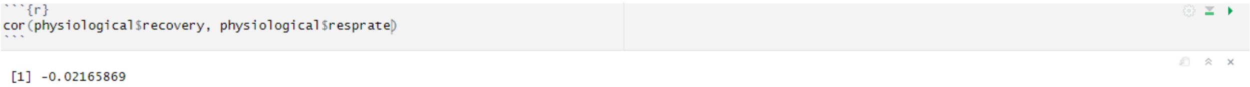
~~~
```{r}
cor(physiological$sleepscore, physiological$recovery)
```

[1] 0.05409019
~~~

Recovery and Respiratory Rate



~~~
```{r}
recovery.resprate <- ggplot(data=physiological) +
  geom_point(mapping=aes(resprate, recovery, color = color))+
  scale_color_manual(name = "Recovery Score", values= c("red"="red",
                                                       "yellow"="yellow",
                                                       "green"="green")) +
  labs(title="Recovery and Respiratory Rate")
recovery.resprate
```
~~~





~~~
```{r}
cor(physiological$recovery, physiological$resprate)
```
~~~

```
[1] -0.02165869
```

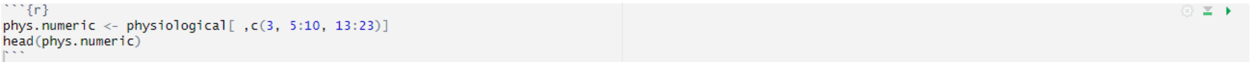
There is no linear relationship between recovery and respiratory rate.

Recovery and Key Metrics Conclusion

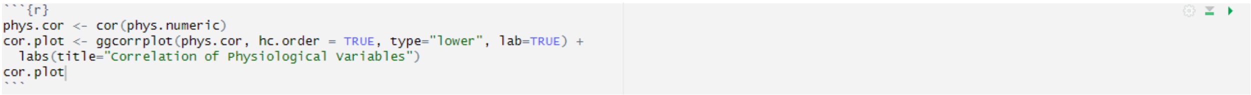
At this point I know based on research and documentation from Whoop that they use HRV, RHR, sleep performance, and respiratory rate in their calculation of recovery. However, the plots and correlation coefficients make it seem as though HRV is the main determinant of recovery. This will be evaluated further during the prediction modeling process. While several of the variables had no linear relationship with recovery that does not necessarily mean that there is no relationship between the two variables.

Correlation of all Numerical Variables

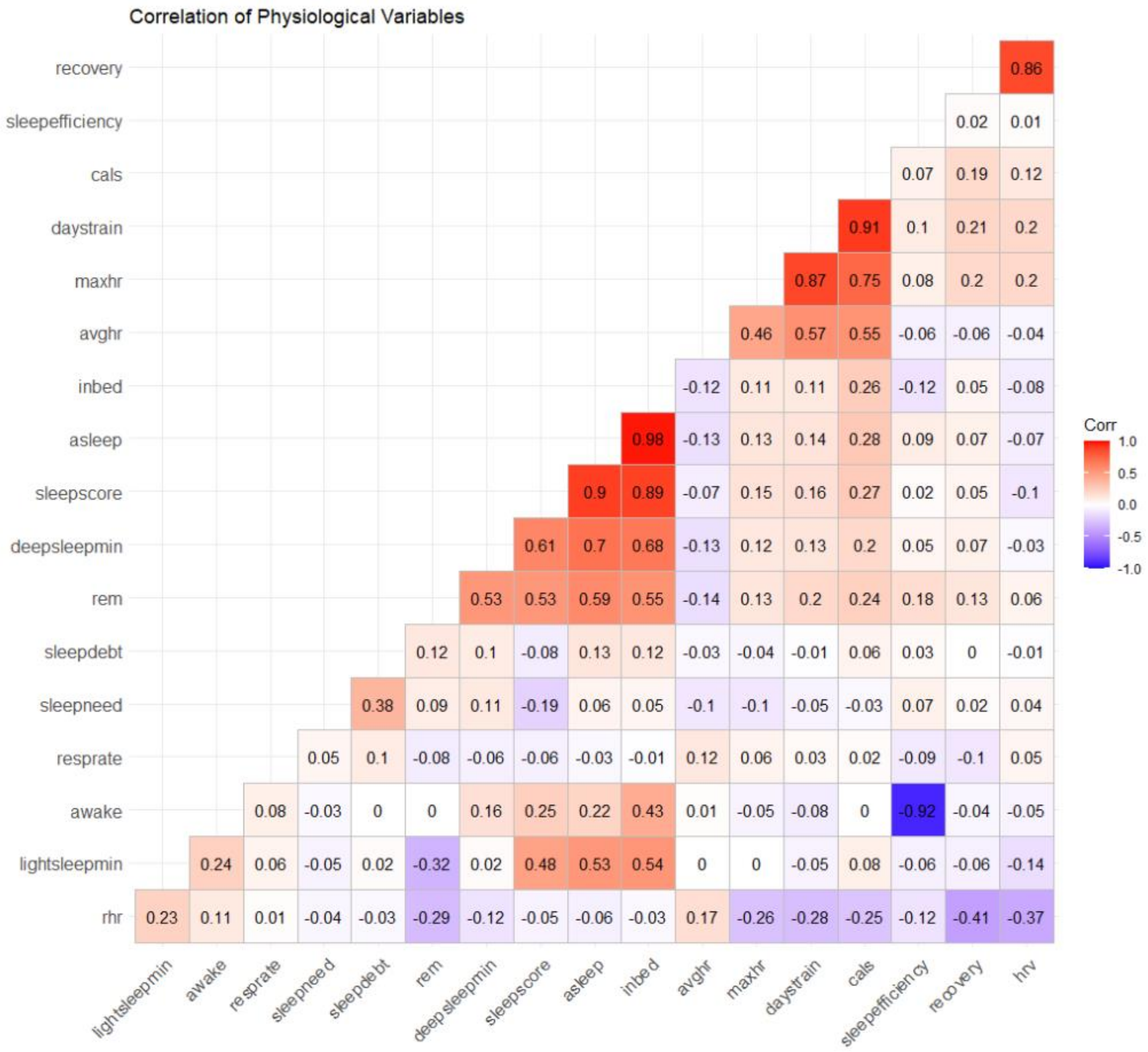
Since the above visualization and analysis on recovery and key metrics returned weak correlations, I decided to run a full correlation matrix on the numeric variables in the physiological dataset. I started by separating this information into a new data frame called phys.numeric. I then created a visualization of the correlation matrix.



~~~
```{r}
phys.numeric <- physiological[ ,c(3, 5:10, 13:23)]
head(phys.numeric)
```
~~~



~~~
```{r}
phys.cor <- cor(phys.numeric)
cor.plot <- ggcorrplot(phys.cor, hc.order = TRUE, type="lower", lab=TRUE) +
  labs(title="Correlation of Physiological Variables")
cor.plot
```
~~~



Based on the correlation matrix, my assumption that HRV was the most important contributing factor in relation to recovery was correct. There are also some other highly correlated variable pairs. The following variable pairs present with a strong positive linear relationship: day strain and calories burned, max heart rate and day strain, time asleep and time in bed, sleep performance and time asleep. There is a strong negative linear relationship between time spent awake and sleep efficiency. Positive relationships indicate that both variables increase together. While negative relationships indicate that as one variable increases the other decreases.

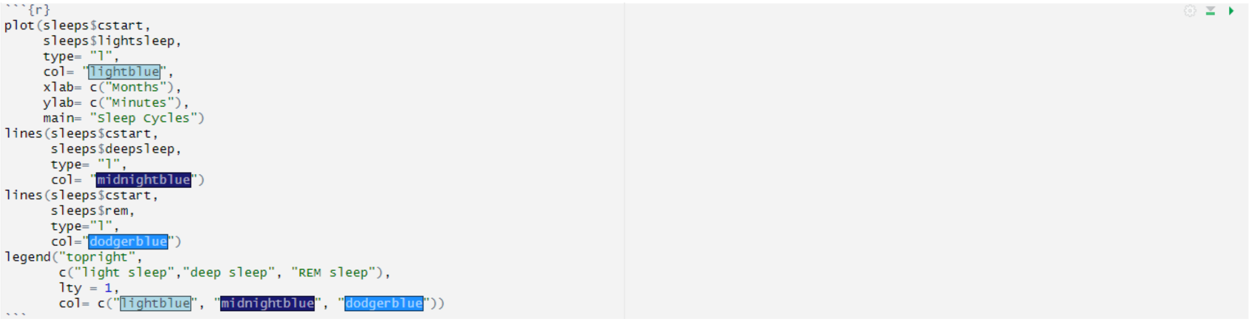
Sleep Exploration and Analysis

Sleep is a restorative exercise. It gives the body a chance to recuperate, allow muscles to heal and rest, and our body to produce hormones that promote healing and well-being. According to Whoop, “deep and REM sleep are the two restorative stages of sleep where your body and mind heal and repair themselves.” [3] The time we spend in each sleep cycle as well as overall time spent asleep contributes to a feeling of rest and well-being. It is also important in the calculation of recovery score.

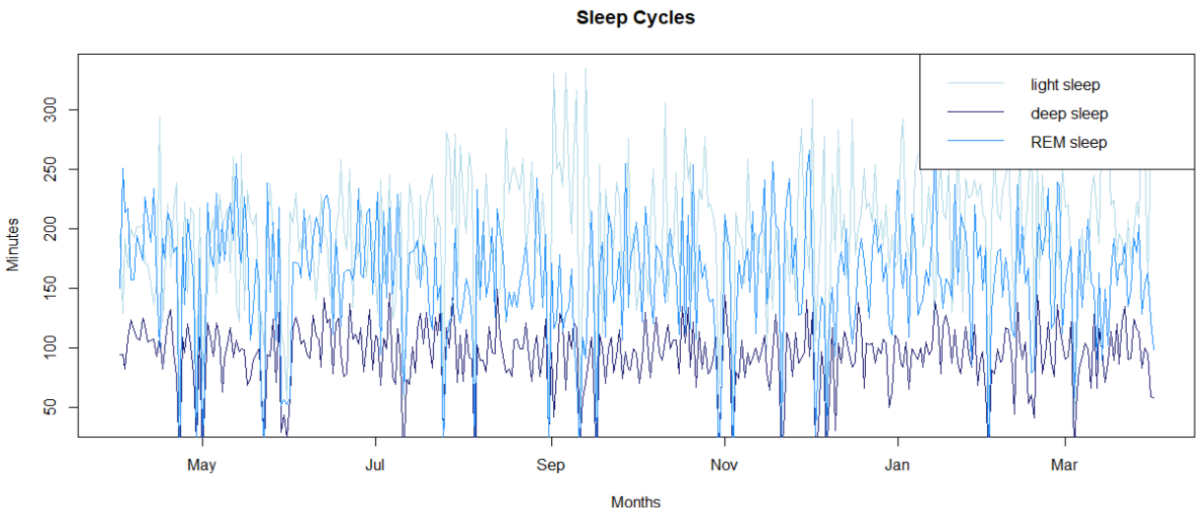
Visualizing a Year of Sleep

Sleep Cycles

Since deep sleep and REM sleep are the most important restorative stages of sleep, I wanted to look at how much time I was spending in each sleep cycle.

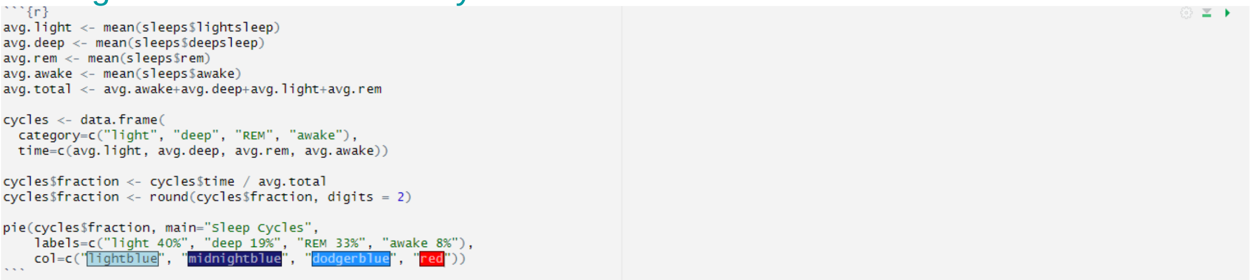


~~~
```{r}
plot(sleeps$cstart,
     sleeps$lightsleep,
     type= "l",
     col= "lightblue",
     xlab= c("Months"),
     ylab= c("Minutes"),
     main= "Sleep Cycles")
lines(sleeps$cstart,
      sleeps$deepsleep,
      type= "l",
      col= "midnightblue")
lines(sleeps$cstart,
      sleeps$rem,
      type="l",
      col="dodgerblue")
legend("topright",
       c("light sleep","deep sleep", "REM sleep"),
       lty = 1,
       col= c("lightblue", "midnightblue", "dodgerblue"))
```
~~~

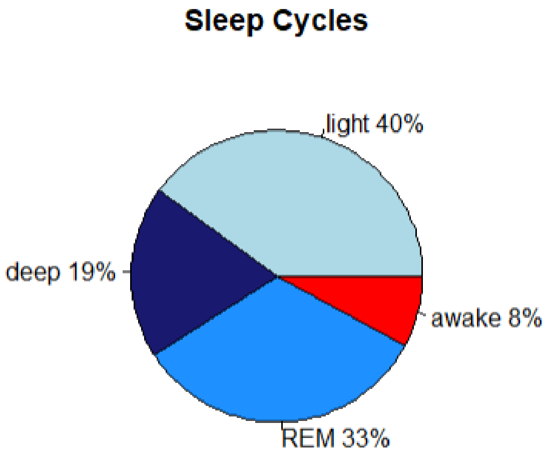


This chart indicates that I spent more time in light sleep than any other sleep cycle. I spent the least amount of time in a deep sleep.

Average % of Time in Each Cycle



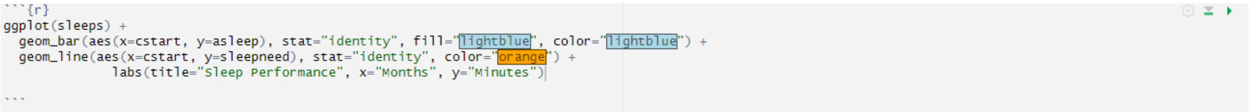
~~~
```{r}
avg.light <- mean(sleeps$lightsleep)
avg.deep <- mean(sleeps$deepsleep)
avg.rem <- mean(sleeps$rem)
avg.awake <- mean(sleeps$awake)
avg.total <- avg.awake+avg.deep+avg.light+avg.rem

cycles <- data.frame(
  category=c("light", "deep", "REM", "awake"),
  time=c(avg.light, avg.deep, avg.rem, avg.awake))

cycles$fraction <- cycles$time / avg.total
cycles$fraction <- round(cycles$fraction, digits = 2)

pie(cycles$fraction, main="Sleep Cycles",
    labels=c("light 40%", "deep 19%", "REM 33%", "awake 8%"),
    col=c("lightblue", "midnightblue", "dodgerblue", "red"))
```
~~~



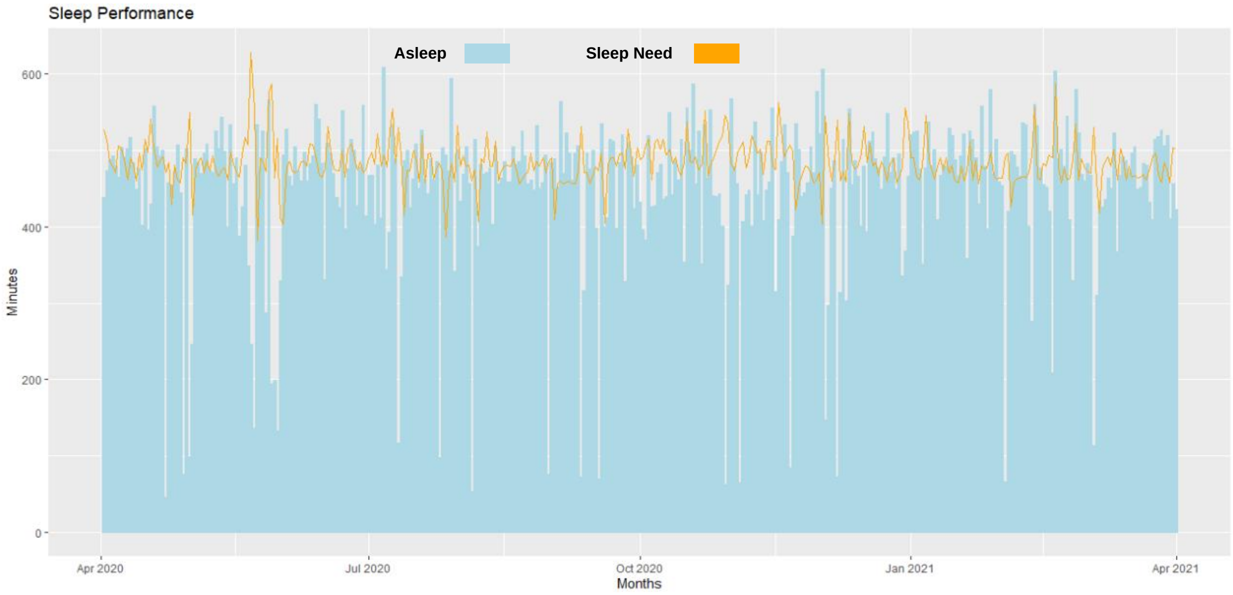
As you can see, on average I spend 52% of my night in restorative sleep cycles (deep and rem). This is likely why I have such a high number of moderate and high recovery days over low recovery.

Sleep Performance

Sleep performance is calculated as the difference between time spent asleep and sleep need. Sleep performance = asleep/sleep need.

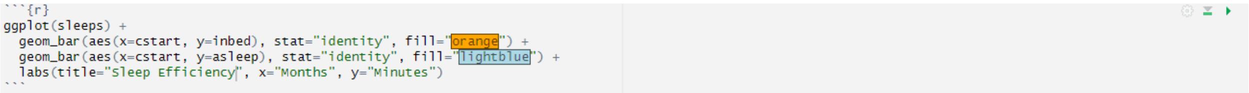


~~~
```{r}
ggplot(sleeps) +
  geom_bar(aes(x=cstart, y=asleep), stat="identity", fill="lightblue", color="lightblue") +
  geom_line(aes(x=cstart, y=sleepneed), stat="identity", color="orange") +
            labs(title="Sleep Performance", x="Months", y="Minutes")

```
~~~

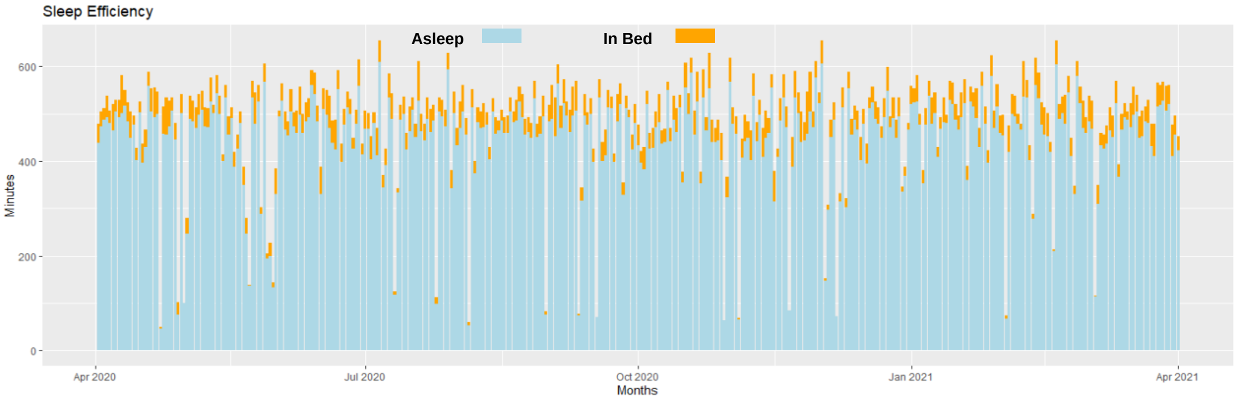


Sleep Efficiency

Sleep efficiency is the difference between time spent in bed and time actually slept. Sleep efficiency = asleep/in bed. The image below shows time slept vs. time in bed. The orange section represents the amount of time it took to fall asleep plus the sleep disturbances or time spent awake during sleep cycles. Overall, I have a relatively consistent pattern of sleeping approximately 92% of the minutes I spend in bed.

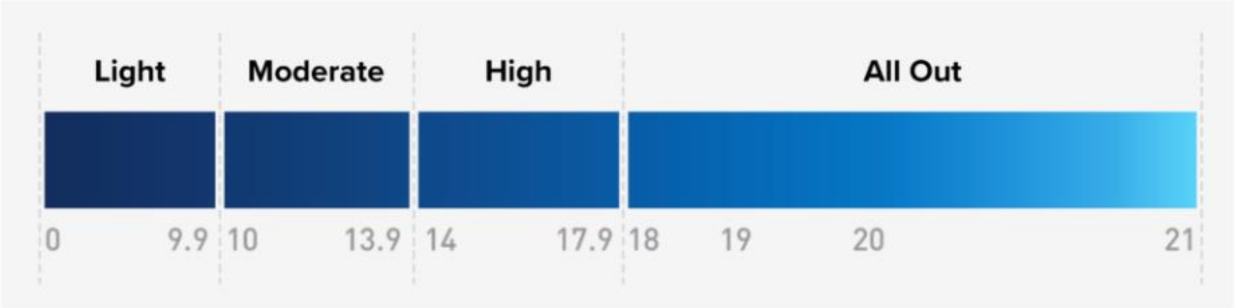


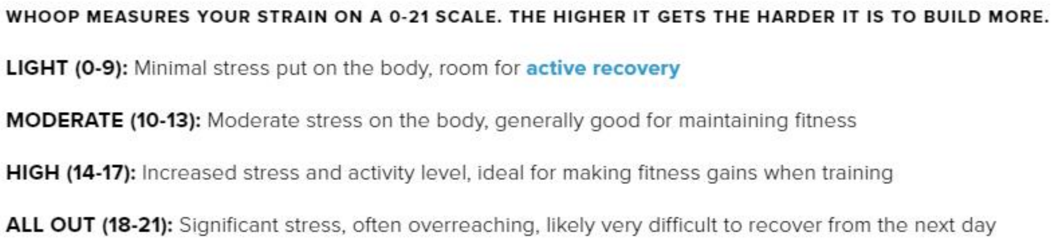
~~~
```{r}
ggplot(sleeps) +
  geom_bar(aes(x=cstart, y=inbed), stat="identity", fill="orange") +
  geom_bar(aes(x=cstart, y=asleep), stat="identity", fill="lightblue") +
  labs(title="Sleep Efficiency", x="Months", y="Minutes")
```
~~~



Workouts and Activity Strain Exploration and Analysis

Workouts and activities that put strain on the body and cardiovascular system result in an activity strain score and day strain. Whoop bases strain on Borg's Rating of Perceived Exertion. [4] Activity strain is a calculation of the percentage of time spent in each heart rate zone (hrz). Strain impacts recovery, sleep need, and HRV. Cardiovascular activity also increases endurance, the body's ability to adapt to stress, and increases physical fitness. Understanding metrics related to workouts is important in understanding what habits lead to high recovery and increased performance as well as trying to avoid low recovery days.





WHOOP MEASURES YOUR STRAIN ON A 0-21 SCALE. THE HIGHER IT GETS THE HARDER IT IS TO BUILD MORE.

LIGHT (0-9): Minimal stress put on the body, room for **active recovery**

MODERATE (10-13): Moderate stress on the body, generally good for maintaining fitness

HIGH (14-17): Increased stress and activity level, ideal for making fitness gains when training

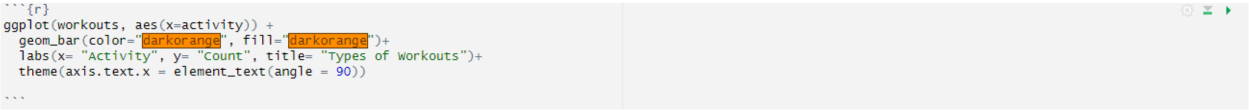
ALL OUT (18-21): Significant stress, often overreaching, likely very difficult to recover from the next day

*Information and Image from Whoop documentation regarding strain.[5]

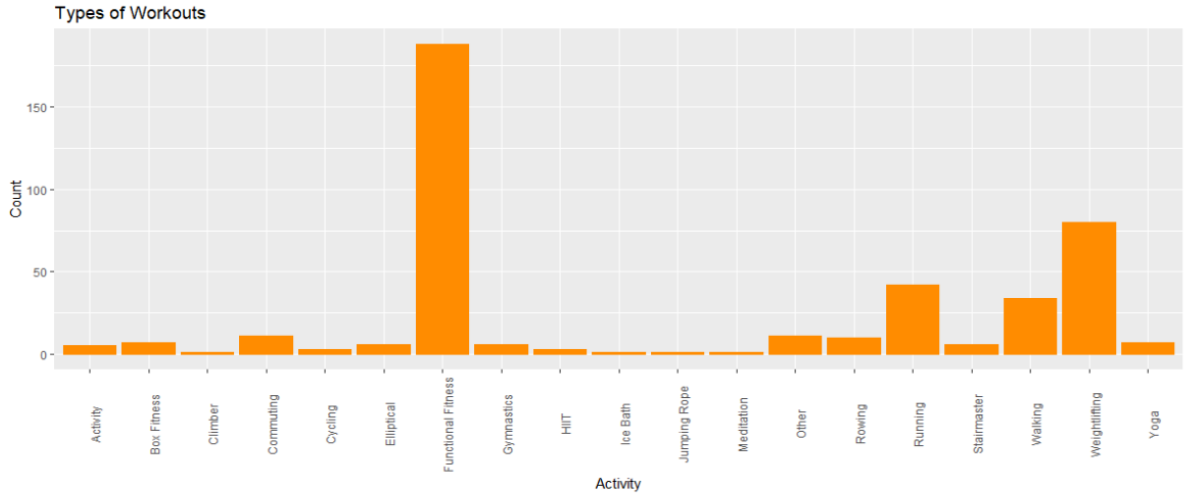
Visualizing a Year of Workouts

Activity Types and Frequency

I performed many different types of activities and workouts over the course of this year. I was a member at a CrossFit gym and most of the workouts tracked were considered Functional Fitness or Box Fitness which are synonyms for CrossFit.

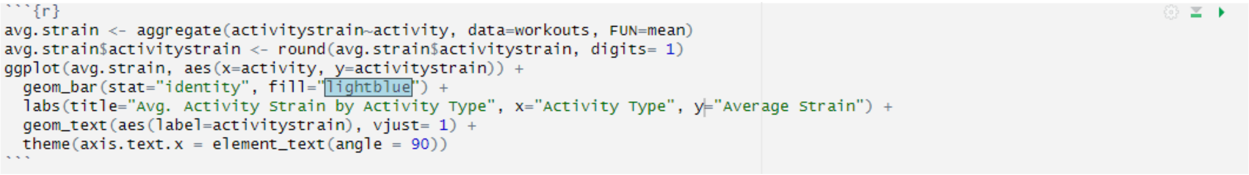


~~~
```{r}
ggplot(workouts, aes(x=activity)) +
  geom_bar(color="darkorange", fill="darkorange")+
  labs(x= "Activity", y= "Count", title= "Types of Workouts")+
  theme(axis.text.x = element_text(angle = 90))

```
~~~

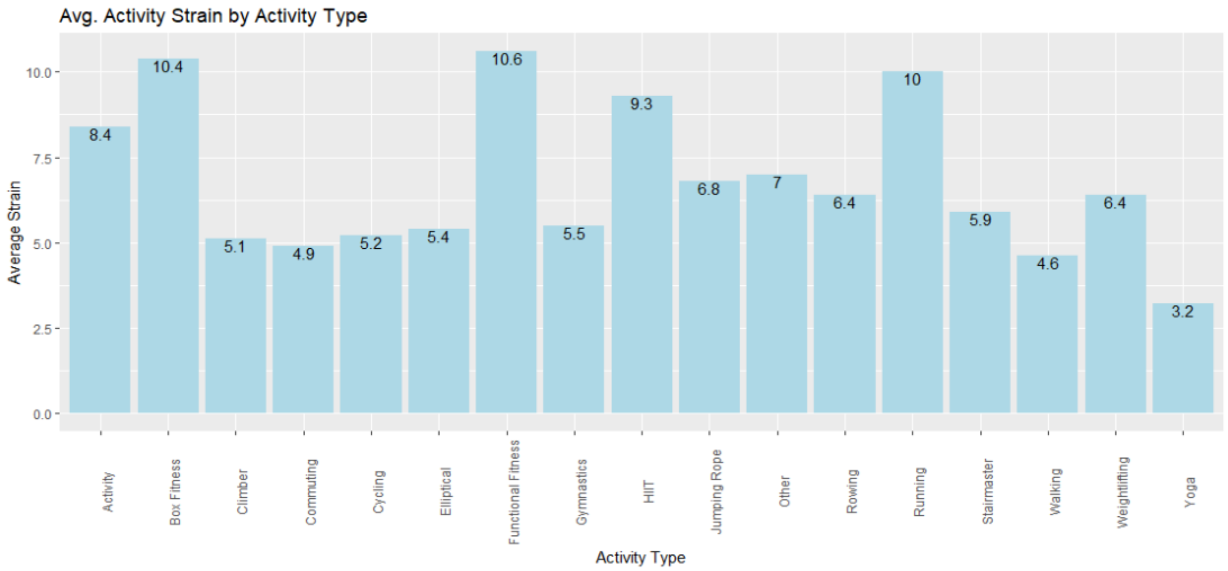


Average Activity Strain

Each activity type has different cardiovascular needs and the amount of time spent in the various heart rate zones will vary. Thus, the average strain incurred by workout/activity type should vary. The following chart is a visual representation of the difference between each activity by average activity strain.



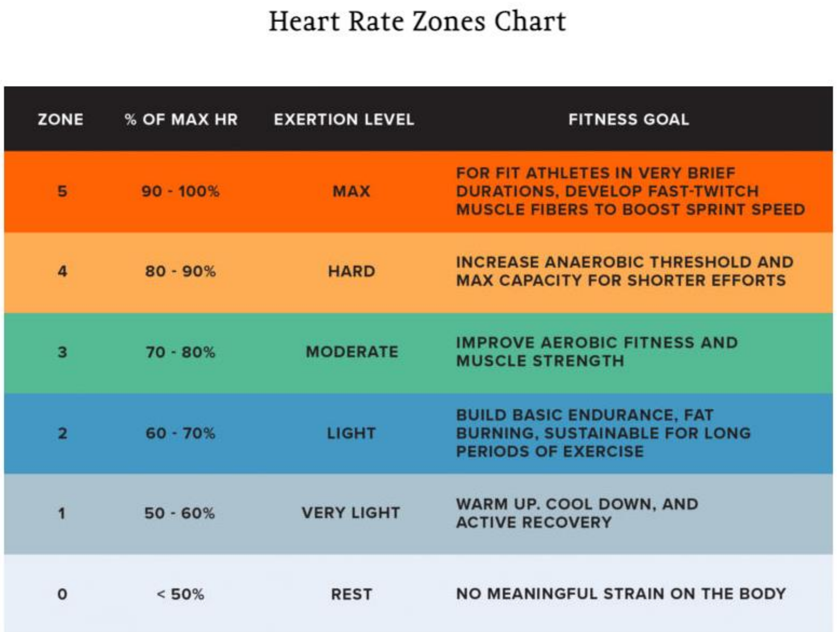
~~~
```{r}
avg.strain <- aggregate(activitystrain~activity, data=workouts, FUN=mean)
avg.strain$activitystrain <- round(avg.strain$activitystrain, digits= 1)
ggplot(avg.strain, aes(x=activity, y=activitystrain)) +
  geom_bar(stat="identity", fill="lightblue") +
  labs(title="Avg. Activity Strain by Activity Type", x="Activity Type", y="Average Strain") +
  geom_text(aes(label=activitystrain), vjust= 1) +
  theme(axis.text.x = element_text(angle = 90))
```
~~~



Based on this chart, box fitness, functional fitness, HIIT, and running result in the highest activity strain and thus require the most cardiovascular exertion. I am interested to see the breakdown of heart rate zones for each activity. I will do this by taking the average percentage of time for each heart rate zone and activity and create a color-coded bar chart.

Heart Rate Zones by Activity

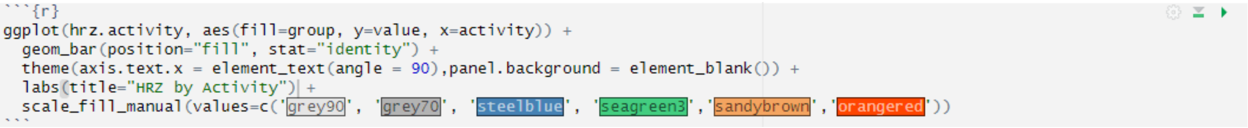
Heart rate zones are based on your own max heart rate. The amount of time spent in each heart rate zone represents a level of exertion and helps to achieve different fitness goals. The heart rate zones used by Whoop are represented by the following image sourced from the Whoop website.



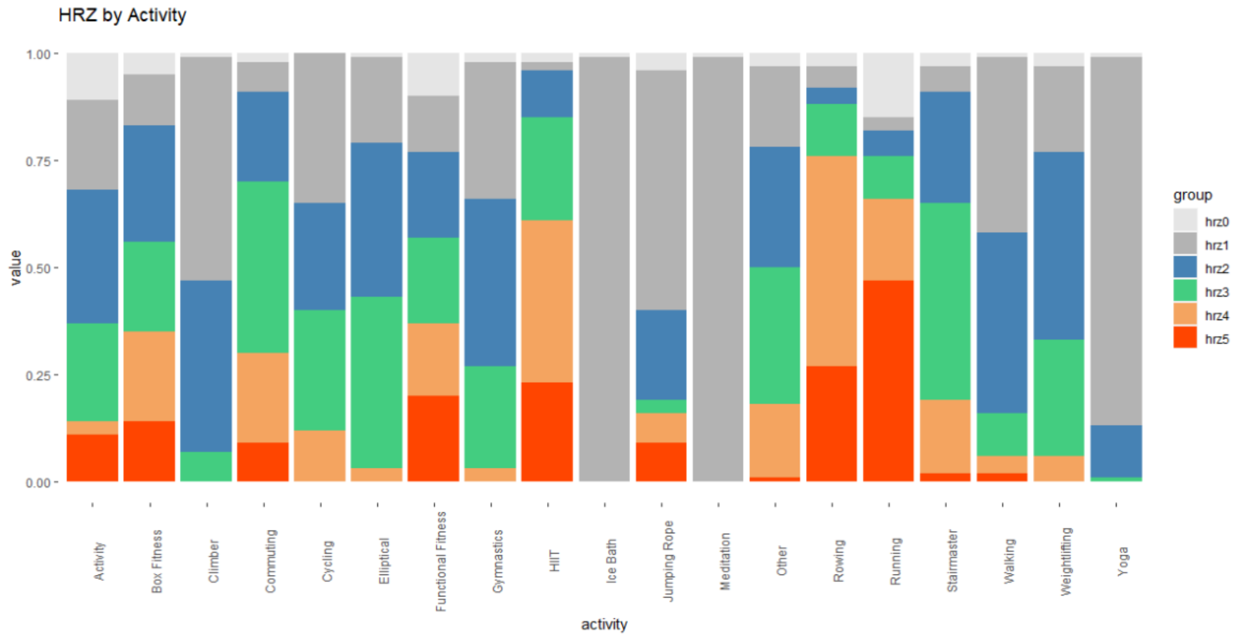
Heart Rate Zones Chart

ZONE	% OF MAX HR	EXERTION LEVEL	FITNESS GOAL
5	90 - 100%	MAX	FOR FIT ATHLETES IN VERY BRIEF DURATIONS, DEVELOP FAST-TWITCH MUSCLE FIBERS TO BOOST SPRINT SPEED
4	80 - 90%	HARD	INCREASE ANAEROBIC THRESHOLD AND MAX CAPACITY FOR SHORTER EFFORTS
3	70 - 80%	MODERATE	IMPROVE AEROBIC FITNESS AND MUSCLE STRENGTH
2	60 - 70%	LIGHT	BUILD BASIC ENDURANCE, FAT BURNING, SUSTAINABLE FOR LONG PERIODS OF EXERCISE
1	50 - 60%	VERY LIGHT	WARM UP. COOL DOWN, AND ACTIVE RECOVERY
0	< 50%	REST	NO MEANINGFUL STRAIN ON THE BODY

*Image retrieved from Whoop documentation on heart rate zones.[6]



~~~
```{r}
ggplot(hrz.activity, aes(fill=group, y=value, x=activity)) +
  geom_bar(position="fill", stat="identity") +
  theme(axis.text.x = element_text(angle = 90),panel.background = element_blank()) +
  labs(title="HRZ by Activity") +
  scale_fill_manual(values=c('grey90', 'grey70', 'steelblue', 'seagreen3','sandybrown','orangered'))
```
~~~



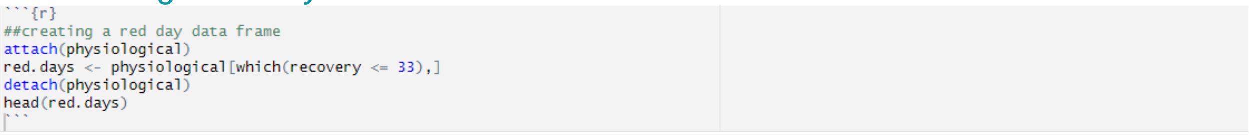
Based on the above chart and information provided by Whoop about heart rate zones, I can conclude that taking an ice bath, meditating, and doing yoga are all restorative. Activities that have more time spent in hrz4 and hrz5 will increase strain on the body and likely contribute to the need for rest and recovery. They also, however, increase the body’s cardiovascular threshold and ability to break through training plateaus which results in a lower RHR and higher HRV. So, even though they may require more recovery time and activities, they may not necessarily contribute to low recovery scores. As we have already discovered, as HRV increases so does Recovery Score.

Highest Highs and Lowest Lows

I know from the previous exploration that I had 17 red/low recovery days. I had 139 high/green recovery days. I am interested in exploring if there are any obvious contributing factors or trends related to the lowest recovery days and highest recovery days. I will create two subsets. One with all red recovery days and the other focusing in on green recovery >= 90. I will then merge all data based on these dates to see if there are any deviations from my baseline that can be pinpointed to understand how that translates to the recovery score. If there are obvious trends for these highest highs and lowest lows, that means I can use this information to change behaviors that will help me increase my recovery more often to achieve peak performance.

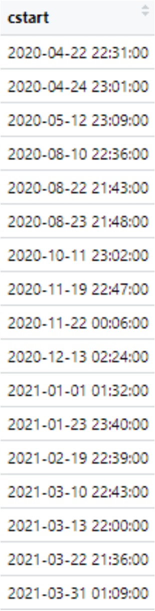
Exploring Differences in Physiological Metrics

Sub-setting Red Days



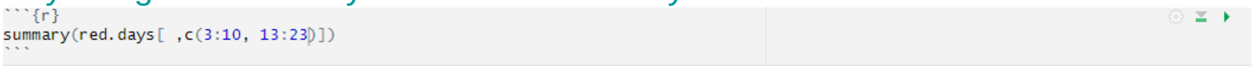
~~~
```{r}
##creating a red day data frame
attach(physiological)
red.days <- physiological[which(recovery <= 33),]
detach(physiological)
head(red.days)
```
~~~

Red Dates

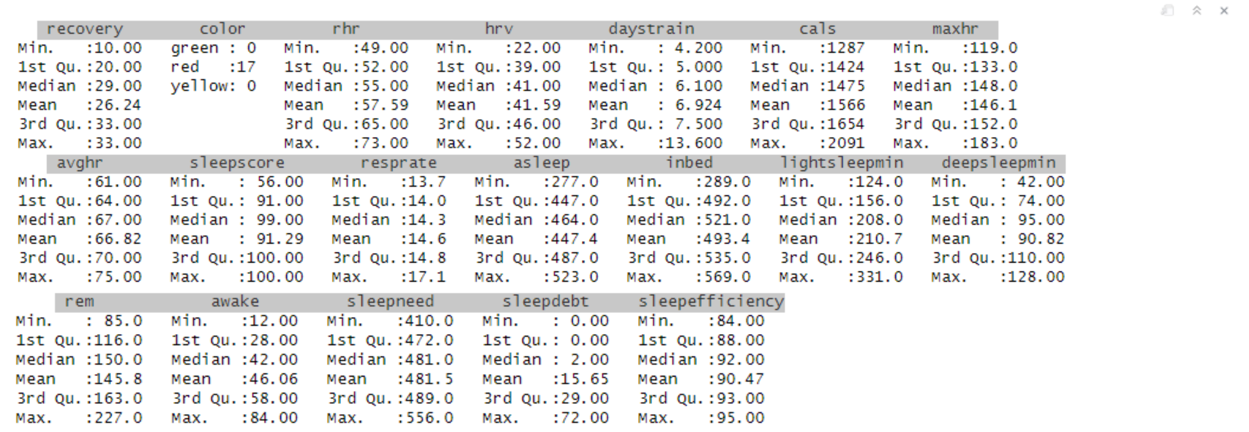


cstart
2020-04-22 22:31:00
2020-04-24 23:01:00
2020-05-12 23:09:00
2020-08-10 22:36:00
2020-08-22 21:43:00
2020-08-23 21:48:00
2020-10-11 23:02:00
2020-11-19 22:47:00
2020-11-22 00:06:00
2020-12-13 02:24:00
2021-01-01 01:32:00
2021-01-23 23:40:00
2021-02-19 22:39:00
2021-03-10 22:43:00
2021-03-13 22:00:00
2021-03-22 21:36:00
2021-03-31 01:09:00

Physiological Summary Statistics – Red Days

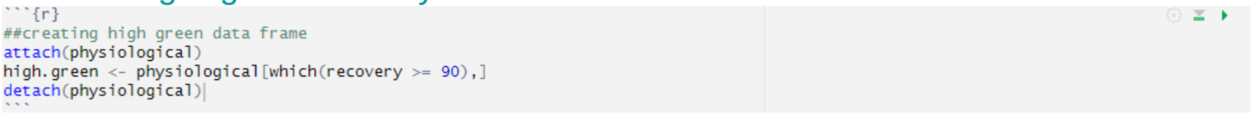


~~~
```{r}
summary(red.days[ ,c(3:10, 13:23)])
```
~~~



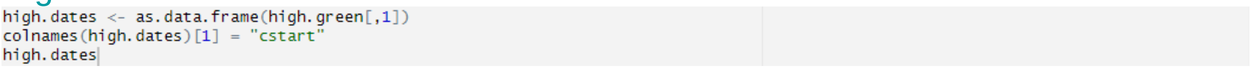
```
    recovery         color          rhr             hrv           daystrain          cals          maxhr      
 Min.   :10.00   green : 0   Min.   :49.00   Min.   :22.00   Min.   : 4.200   Min.   :1287   Min.   :119.0  
 1st Qu.:20.00   red   :17   1st Qu.:52.00   1st Qu.:39.00   1st Qu.: 5.000   1st Qu.:1424   1st Qu.:133.0  
 Median :29.00   yellow: 0   Median :55.00   Median :41.00   Median : 6.100   Median :1475   Median :148.0  
 Mean   :26.24               Mean   :57.59   Mean   :41.59   Mean   : 6.924   Mean   :1566   Mean   :146.1  
 3rd Qu.:33.00               3rd Qu.:65.00   3rd Qu.:46.00   3rd Qu.: 7.500   3rd Qu.:1654   3rd Qu.:152.0  
 Max.   :33.00               Max.   :73.00   Max.   :52.00   Max.   :13.600   Max.   :2091   Max.   :183.0  
     avghr         sleepscore        resprate         asleep          inbed        lightsleepmin    deepsleepmin   
 Min.   :61.00   Min.   : 56.00   Min.   :13.7   Min.   :277.0   Min.   :289.0   Min.   :124.0   Min.   : 42.00  
 1st Qu.:64.00   1st Qu.: 91.00   1st Qu.:14.0   1st Qu.:447.0   1st Qu.:492.0   1st Qu.:156.0   1st Qu.: 74.00  
 Median :67.00   Median : 99.00   Median :14.3   Median :464.0   Median :521.0   Median :208.0   Median : 95.00  
 Mean   :66.82   Mean   : 91.29   Mean   :14.6   Mean   :447.4   Mean   :493.4   Mean   :210.7   Mean   : 90.82  
 3rd Qu.:70.00   3rd Qu.:100.00   3rd Qu.:14.8   3rd Qu.:487.0   3rd Qu.:535.0   3rd Qu.:246.0   3rd Qu.:110.00  
 Max.   :75.00   Max.   :100.00   Max.   :17.1   Max.   :523.0   Max.   :569.0   Max.   :331.0   Max.   :128.00  
      rem            awake          sleepneed       sleepdebt     sleepefficiency
 Min.   : 85.0   Min.   :12.00   Min.   :410.0   Min.   : 0.00   Min.   :84.00  
 1st Qu.:116.0   1st Qu.:28.00   1st Qu.:472.0   1st Qu.: 0.00   1st Qu.:88.00  
 Median :150.0   Median :42.00   Median :481.0   Median : 2.00   Median :92.00  
 Mean   :145.8   Mean   :46.06   Mean   :481.5   Mean   :15.65   Mean   :90.47  
 3rd Qu.:163.0   3rd Qu.:58.00   3rd Qu.:489.0   3rd Qu.:29.00   3rd Qu.:93.00  
 Max.   :227.0   Max.   :84.00   Max.   :556.0   Max.   :72.00   Max.   :95.00  
```

Sub-setting High Green Days

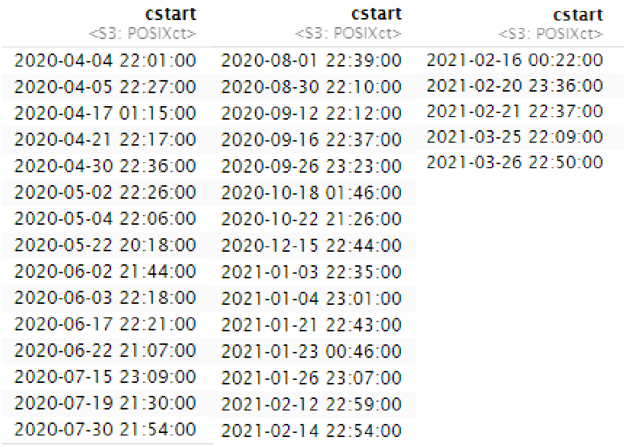


~~~
```{r}
##creating high green data frame
attach(physiological)
high.green <- physiological[which(recovery >= 90),]
detach(physiological)
```
~~~

High Green Dates

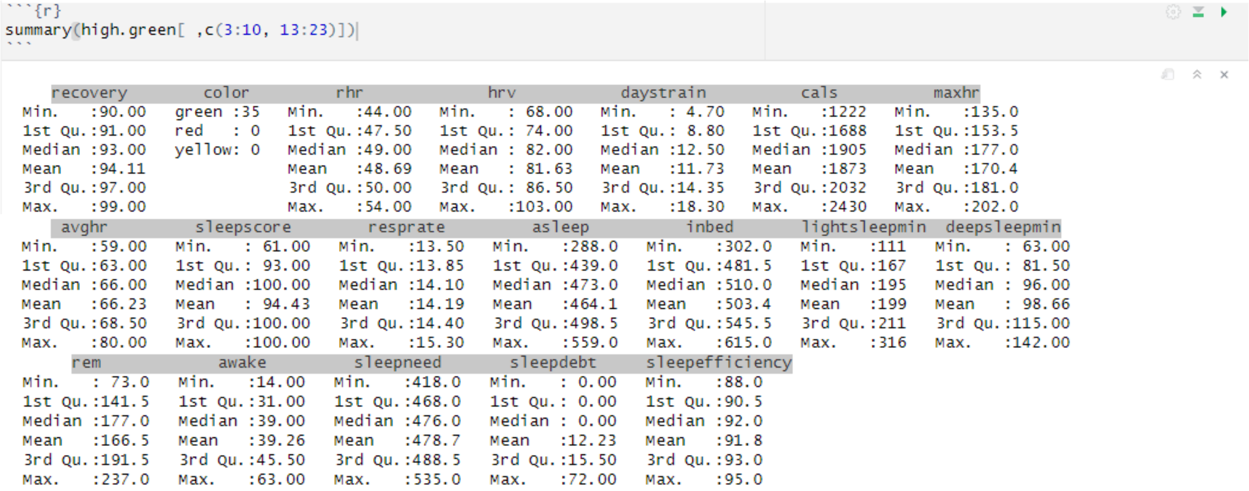


```
high.dates <- as.data.frame(high.green[,1])
colnames(high.dates)[1] = "cstart"
high.dates
```



cstart <S3: POSIXct>	cstart <S3: POSIXct>	cstart <S3: POSIXct>
2020-04-04 22:01:00	2020-08-01 22:39:00	2021-02-16 00:22:00
2020-04-05 22:27:00	2020-08-30 22:10:00	2021-02-20 23:36:00
2020-04-17 01:15:00	2020-09-12 22:12:00	2021-02-21 22:37:00
2020-04-21 22:17:00	2020-09-16 22:37:00	2021-03-25 22:09:00
2020-04-30 22:36:00	2020-09-26 23:23:00	2021-03-26 22:50:00
2020-05-02 22:26:00	2020-10-18 01:46:00	
2020-05-04 22:06:00	2020-10-22 21:26:00	
2020-05-22 20:18:00	2020-12-15 22:44:00	
2020-06-02 21:44:00	2021-01-03 22:35:00	
2020-06-03 22:18:00	2021-01-04 23:01:00	
2020-06-17 22:21:00	2021-01-21 22:43:00	
2020-06-22 21:07:00	2021-01-23 00:46:00	
2020-07-15 23:09:00	2021-01-26 23:07:00	
2020-07-19 21:30:00	2021-02-12 22:59:00	
2020-07-30 21:54:00	2021-02-14 22:54:00	

Physiological Summary Statistics – High Green Days



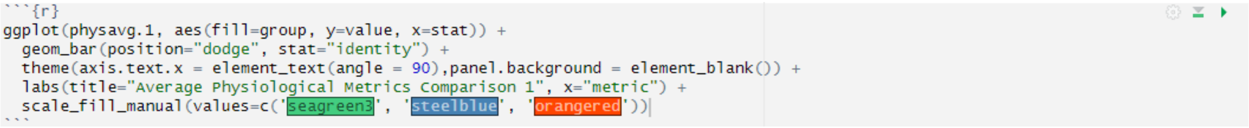
~~~
```{r}
summary(high.green[ ,c(3:10, 13:23)])
```

~~~

```
    recovery         color          rhr              hrv              daystrain           cals             maxhr      
 Min.    :90.00   green :35   Min.    :44.00   Min.    : 68.00   Min.    : 4.70   Min.    :1222   Min.    :135.0  
 1st Qu.:91.00   red    : 0   1st Qu.:47.50   1st Qu.: 74.00   1st Qu.: 8.80   1st Qu.:1688   1st Qu.:153.5  
 Median :93.00   yellow: 0   Median :49.00   Median : 82.00   Median :12.50   Median :1905   Median :177.0  
 Mean   :94.11               Mean   :48.69   Mean   : 81.63   Mean   :11.73   Mean   :1873   Mean   :170.4  
 3rd Qu.:97.00               3rd Qu.:50.00   3rd Qu.: 86.50   3rd Qu.:14.35   3rd Qu.:2032   3rd Qu.:181.0  
 Max.   :99.00               Max.   :54.00   Max.   :103.00   Max.   :18.30   Max.   :2430   Max.   :202.0  
     avghr          sleepscore         resprate          asleep           inbed        lightsleepmin  deepsleepmin  
 Min.   :59.00   Min.   : 61.00   Min.   :13.50   Min.   :288.0   Min.   :302.0   Min.   :111   Min.   : 63.00  
 1st Qu.:63.00   1st Qu.: 93.00   1st Qu.:13.85   1st Qu.:439.0   1st Qu.:481.5   1st Qu.:167   1st Qu.: 81.50  
 Median :66.00   Median :100.00   Median :14.10   Median :473.0   Median :510.0   Median :195   Median : 96.00  
 Mean   :66.23   Mean   : 94.43   Mean   :14.19   Mean   :464.1   Mean   :503.4   Mean   :199   Mean   : 98.66  
 3rd Qu.:68.50   3rd Qu.:100.00   3rd Qu.:14.40   3rd Qu.:498.5   3rd Qu.:545.5   3rd Qu.:211   3rd Qu.:115.00  
 Max.   :80.00   Max.   :100.00   Max.   :15.30   Max.   :559.0   Max.   :615.0   Max.   :316   Max.   :142.00  
      rem             awake          sleepneed        sleepdebt      sleepefficiency
 Min.   : 73.0   Min.   :14.00   Min.   :418.0   Min.   : 0.00   Min.   :88.0   
 1st Qu.:141.5   1st Qu.:31.00   1st Qu.:468.0   1st Qu.: 0.00   1st Qu.:90.5   
 Median :177.0   Median :39.00   Median :476.0   Median : 0.00   Median :92.0   
 Mean   :166.5   Mean   :39.26   Mean   :478.7   Mean   :12.23   Mean   :91.8   
 3rd Qu.:191.5   3rd Qu.:45.50   3rd Qu.:488.5   3rd Qu.:15.50   3rd Qu.:93.0   
 Max.   :237.0   Max.   :63.00   Max.   :535.0   Max.   :72.00   Max.   :95.0   
```

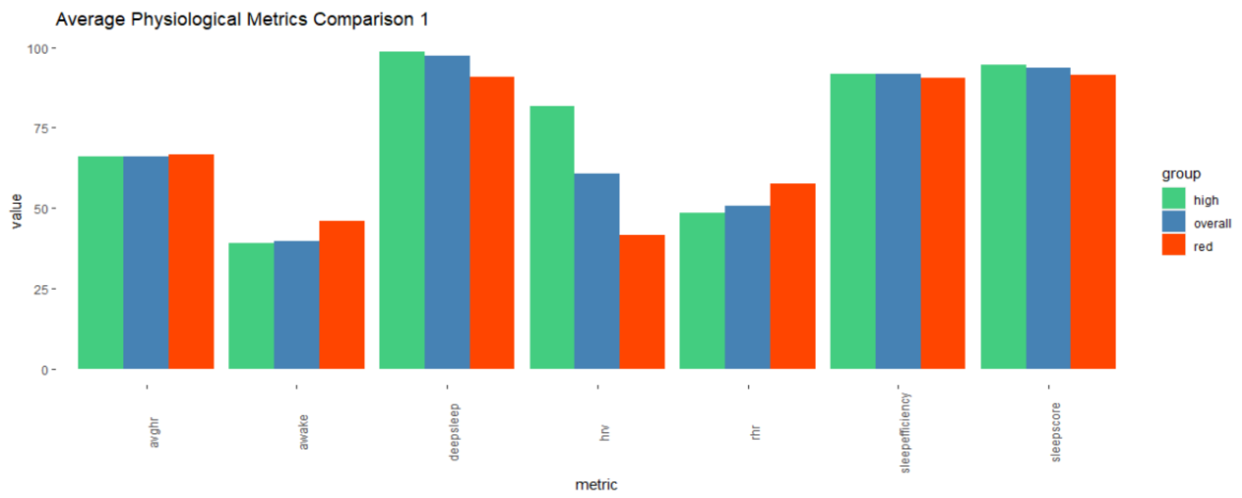
Visualizing the Differences in Physiological Metrics

To see differences in physiological metrics I calculated the averages for each physiological metric for red days, high green days, and for the entire year. I then separated the results into 5 groups based on the numerical values of the metrics. Group 1 holds the metrics presenting values between 39-100. Group 2 holds the metrics presenting values between 6 and 20. Group 3 holds the metrics presenting values between 145 and 215. Group 4 holds the metrics presenting values between 445 and 510. Group 5 holds the metrics presenting values between 1500 and 2000. I then visualized red, high green, and overall physiological metrics in a side-by-side comparison bar chart to see if I could spot any obvious trends.

Group 1 Comparison



~~~
```{r}
ggplot(physavg.1, aes(fill=group, y=value, x=stat)) +
  geom_bar(position="dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 90),panel.background = element_blank()) +
  labs(title="Average Physiological Metrics Comparison 1", x="metric") +
  scale_fill_manual(values=c('seagreen3', 'steelblue', 'orangered'))
```
~~~



Looking at the side-by-side comparisons of each metric makes it easy to spot differences.

Average HR

There is only a slight difference between high green days, overall, and red days for average heart rate. However, I do see that for red days my average heart rate was higher than average.

Time Spent Awake (During sleep cycles)

Here there is an obvious trend of less time spent awake during sleep cycles for high green days vs overall average and red days. Red days present with more time spent awake during sleep cycles, also known as sleep disturbances.

Deep Sleep

Deep sleep shows that the more time spent in the deep sleep cycle results in higher recovery where less deep sleep results in lower recovery.

HRV

HRV has already been identified as a main contributor to overall recovery score. This is again validated by the chart showing the significance of HRV in resulting recovery scores. The higher the HRV the higher the average recovery scores for high green, overall, and red days.

RHR

Resting heart rate works opposite HRV. The lower my average resting heart rate was, the higher the recovery score.

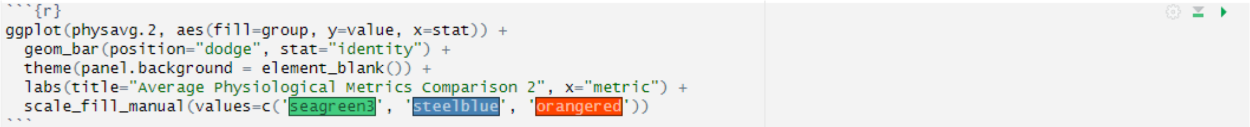
Sleep Efficiency

While the differences in sleep efficiency are not extreme, they still show that there is a difference between high, low, and average overall recovery. The lower sleep efficiency results in lower recovery.

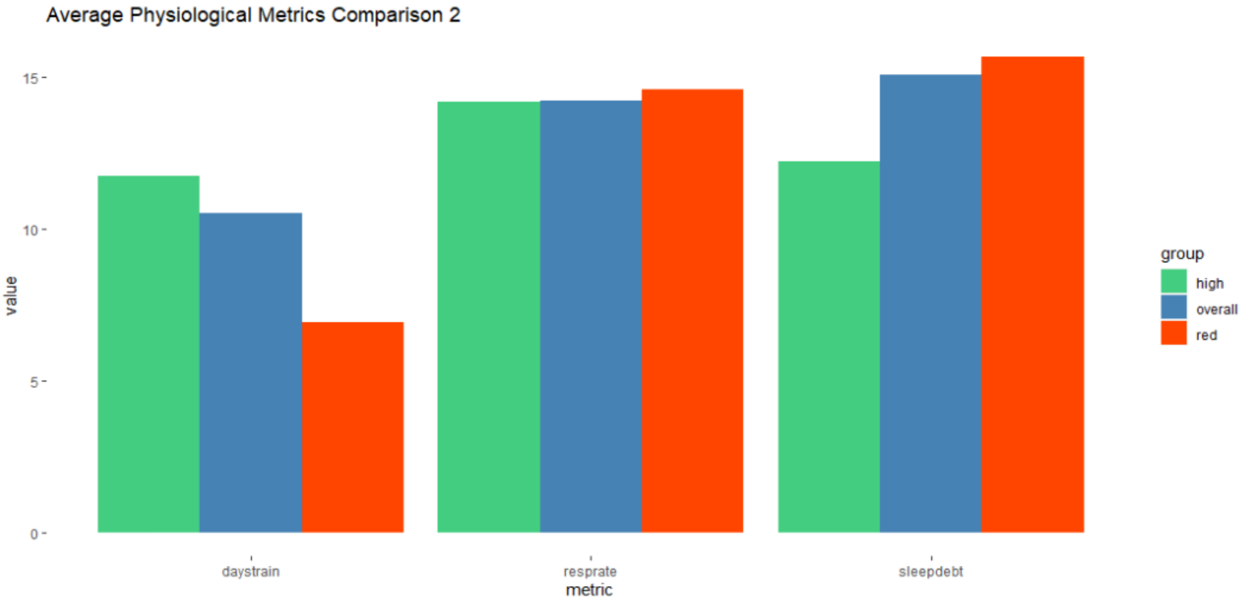
Sleep Performance

Sleep performance also shows that as sleep performance decreases so does recovery.

Group 2 Comparison



~~~
```{r}
ggplot(physavg.2, aes(fill=group, y=value, x=stat)) +
  geom_bar(position="dodge", stat="identity") +
  theme(panel.background = element_blank()) +
  labs(title="Average Physiological Metrics Comparison 2", x="metric") +
  scale_fill_manual(values=c('seagreen3', 'steelblue', 'orangered'))
```
~~~



Day Strain

As day strain decreases, so does recovery. This makes sense because strain is a result of cardiovascular exertion. Without the increase of heart rate, HRV remains low and thus recovery is also low.

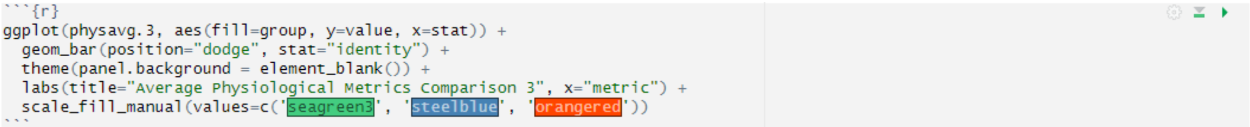
Respiratory Rate

Higher respiratory rates lead to lower recovery. Respiratory rate is the number of breaths we take per minute. This metric remains relatively constant and increases in respiratory rate may implicate illness. Whoop was even able to use this information to detect the early onset of Covid-19. [7]

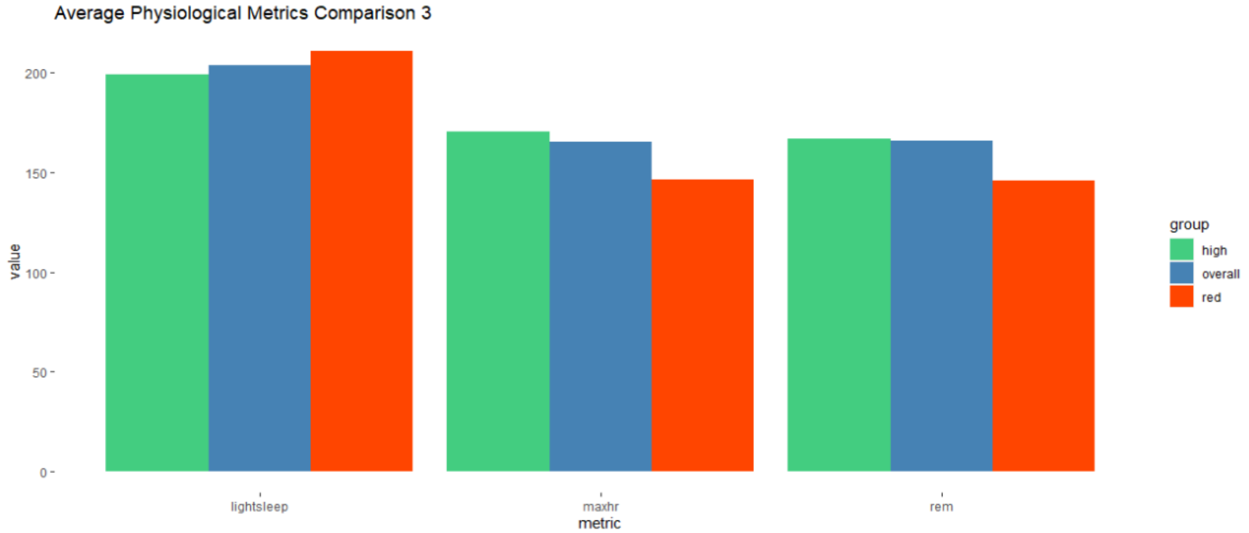
Sleep Debt

As sleep debt decreases, recovery increases. This shows the significance of getting adequate sleep.

Group 3 Comparison



~~~
```{r}
ggplot(physavg.3, aes(fill=group, y=value, x=stat)) +
  geom_bar(position="dodge", stat="identity") +
  theme(panel.background = element_blank()) +
  labs(title="Average Physiological Metrics Comparison 3", x="metric") +
  scale_fill_manual(values=c('seagreen3', 'steelblue', 'orangered'))
```
~~~



Light Sleep

As light sleep decreases, recovery increases. This indicates the importance of quality sleep over quantity. Deep sleep and REM sleep are restorative sleep cycles. This is why more time spent in light sleep equates to lower recovery scores.

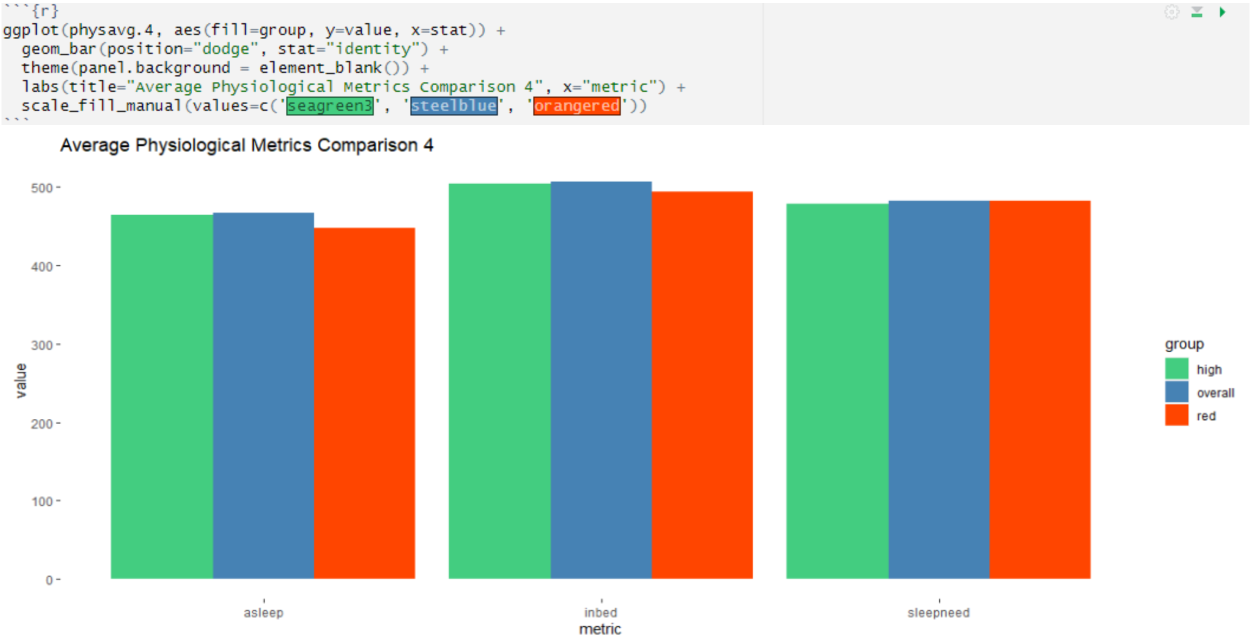
Max HR

As max HR increases so does recovery. Increasing and decreasing heart rate contributes to differences in heart rate variability (HRV). So, increasing your max heart rate contributes to a higher HRV and ultimately a higher recovery.

REM

As previously mentioned, REM is a restorative sleep cycle. I can see that here because as time spent in REM increases so does my recovery.

Group 4 Comparison



~~~~
```{r}
ggplot(physavg.4, aes(fill=group, y=value, x=stat)) +
  geom_bar(position="dodge", stat="identity") +
  theme(panel.background = element_blank()) +
  labs(title="Average Physiological Metrics Comparison 4", x="metric") +
  scale_fill_manual(values=c('seagreen3', 'steelblue', 'orangered'))
```
~~~~

Time Spent Asleep

On average, when I spent less time asleep, I had lower recovery scores than when I spent more time asleep.

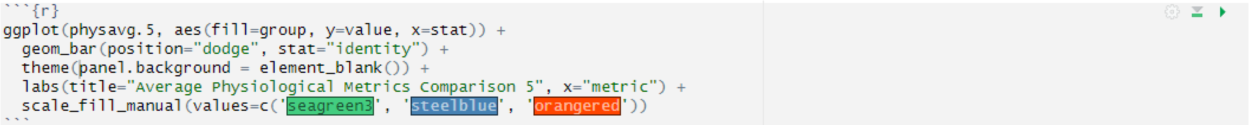
Time Spent In Bed

While the difference in recovery scores is not significant for time spent in bed, it is true that as I spent less time in bed my recovery score decreased. Less time in bed results in less time asleep which then results in lower recovery scores.

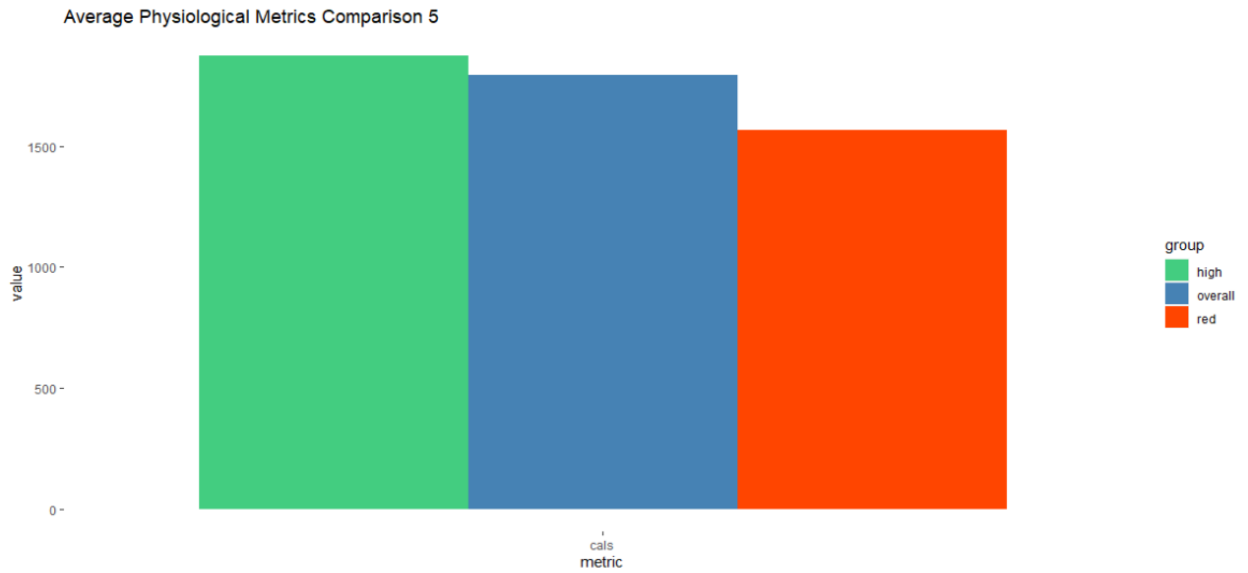
Sleep Need

There is no significant difference in recovery based on sleep need.

Group 5 Comparison



~~~
```{r}
ggplot(physavg.5, aes(fill=group, y=value, x=stat)) +
  geom_bar(position="dodge", stat="identity") +
  theme(panel.background = element_blank()) +
  labs(title="Average Physiological Metrics Comparison 5", x="metric") +
  scale_fill_manual(values=c('seagreen3', 'steelblue', 'orangered'))
```
~~~



Calories Burned

The final physiological metric is calories burned. The chart shows that the more calories burned the higher the resulting recovery score. Exercising increases heart rate, heart rate variability, day strain, and ultimately calories burned. This indicates that often, doing something is better than doing nothing when it comes to recovery. Restorative exercise is better than no exercise at all.

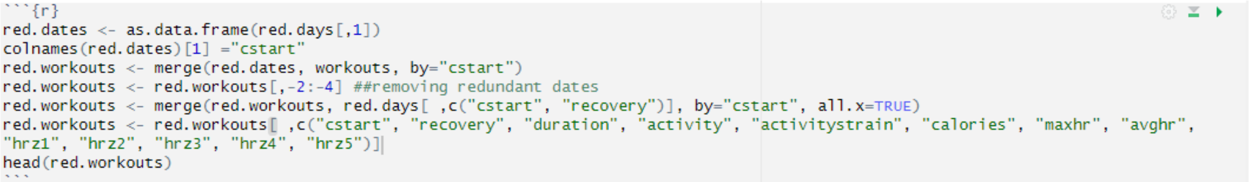
Differences in Physiological Metrics Conclusion

By comparing the average physiological metrics between low/red recovery days, high green recovery days, and my average baseline, I was able to understand the importance of quality sleep and exercise in preparing my body for peak performance. Getting enough sleep is important, but more importantly spending enough time in deep sleep and REM sleep helps to restore the body and prepare for exertion. Furthermore, putting strain on the cardiovascular system through exercise and increasing my heart rate helped in achieving higher HRV and higher recovery scores.

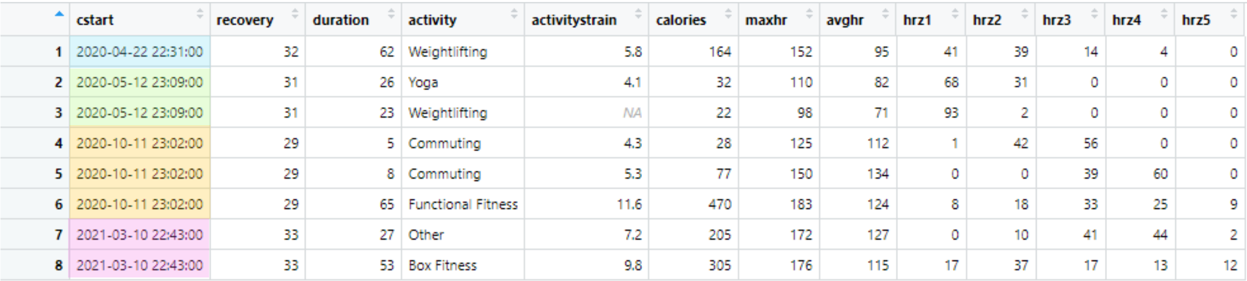
Exploring Workouts for High and Low Days

Red Day Workouts

To get the necessary data I merged the red dates with the workout dataset and then added recovery score for reference.



~~~
```{r}
red.dates <- as.data.frame(red.days[,1])
colnames(red.dates)[1] ="cstart"
red.workouts <- merge(red.dates, workouts, by="cstart")
red.workouts <- red.workouts[,-2:-4] ##removing redundant dates
red.workouts <- merge(red.workouts, red.days[ ,c("cstart", "recovery")], by="cstart", all.x=TRUE)
red.workouts <- red.workouts[ ,c("cstart", "recovery", "duration", "activity", "activitystrain", "calories", "maxhr", "avghr",
"hrz1", "hrz2", "hrz3", "hrz4", "hrz5")]
head(red.workouts)
```
~~~

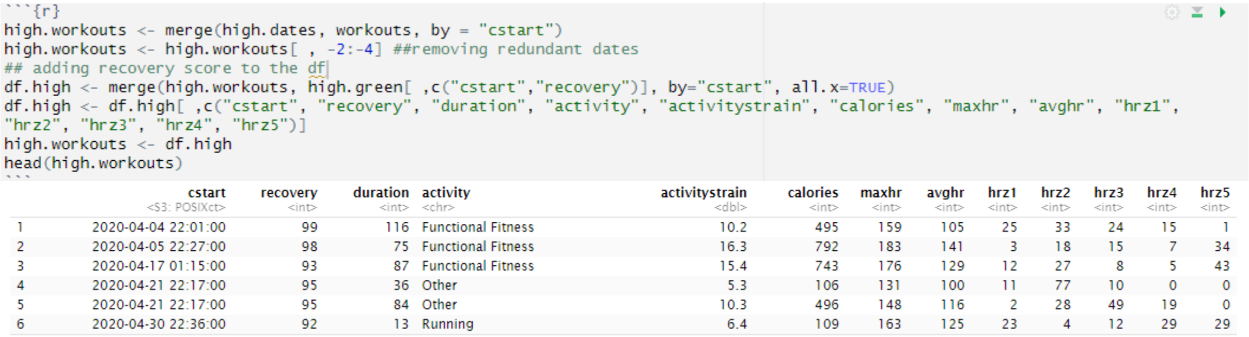


	cstart	recovery	duration	activity	activitystrain	calories	maxhr	avghr	hrz1	hrz2	hrz3	hrz4	hrz5
1	2020-04-22 22:31:00	32	62	Weightlifting	5.8	164	152	95	41	39	14	4	0
2	2020-05-12 23:09:00	31	26	Yoga	4.1	32	110	82	68	31	0	0	0
3	2020-05-12 23:09:00	31	23	Weightlifting	NA	22	98	71	93	2	0	0	0
4	2020-10-11 23:02:00	29	5	Commuting	4.3	28	125	112	1	42	56	0	0
5	2020-10-11 23:02:00	29	8	Commuting	5.3	77	150	134	0	0	39	60	0
6	2020-10-11 23:02:00	29	65	Functional Fitness	11.6	470	183	124	8	18	33	25	9
7	2021-03-10 22:43:00	33	27	Other	7.2	205	172	127	0	10	41	44	2
8	2021-03-10 22:43:00	33	53	Box Fitness	9.8	305	176	115	17	37	17	13	12

The first thing I noticed was that I only exercised on 4 of the red recovery days. Based on everything I have already explored; I know that exercise is an important part of recovery. So, not exercising is a contributing factor to lower recovery days. Additionally, I notice here that the activity strains for these red days are relatively low aside from the functional fitness workout on 10-11-2020 and the box fitness workout on 3-10-2021. This makes me wonder if perhaps I overextended myself on these days and the result was a low recovery score and for the other days perhaps, I didn't exercise enough.

High Green Workouts

To get the necessary data I merged the green dates with the workout dataset and added recovery score for reference.

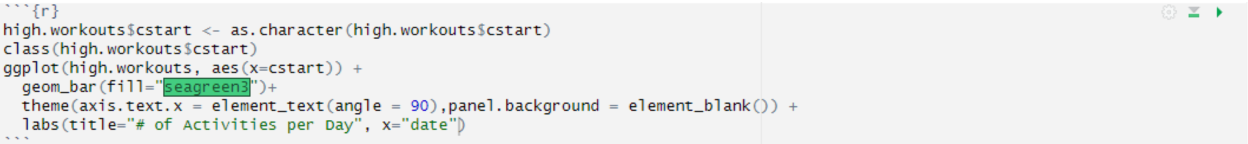


~~~
```{r}
high.workouts <- merge(high.dates, workouts, by = "cstart")
high.workouts <- high.workouts[ , -2:-4] ##removing redundant dates
## adding recovery score to the df
df.high <- merge(high.workouts, high.green[ ,c("cstart","recovery")], by="cstart", all.x=TRUE)
df.high <- df.high[ ,c("cstart", "recovery", "duration", "activity", "activitystrain", "calories", "maxhr", "avghr", "hrz1",
"hrz2", "hrz3", "hrz4", "hrz5")]
high.workouts <- df.high
head(high.workouts)
```
~~~

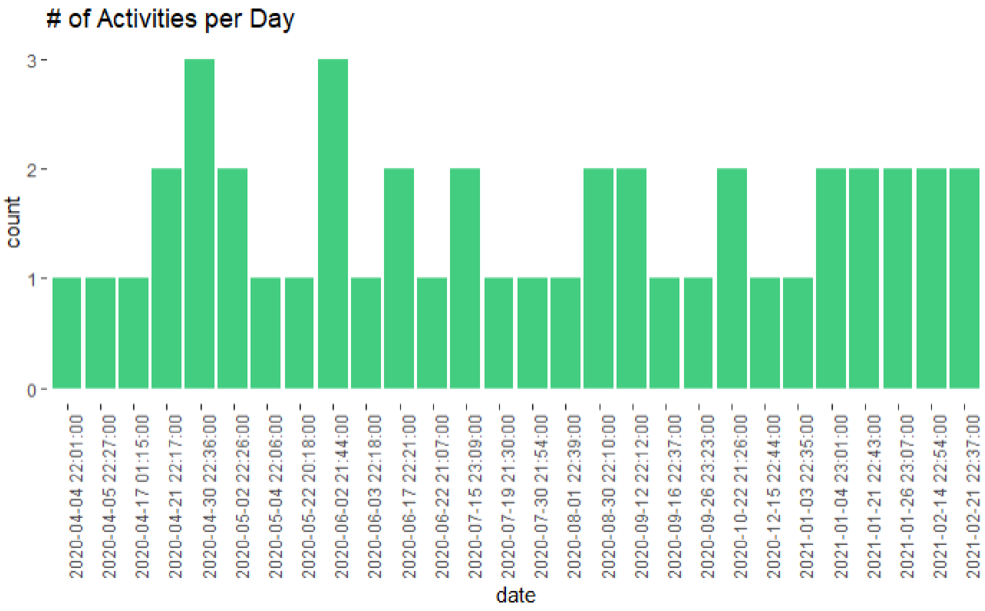
```
                 cstart  recovery  duration activity                          activitystrain  calories  maxhr  avghr  hrz1  hrz2  hrz3  hrz4  hrz5
           <S3: POSIXct>     <int>     <int> <chr>                                     <dbl>     <int>  <int>  <int> <int> <int> <int> <int> <int>
1   2020-04-04 22:01:00        99       116 Functional Fitness                         10.2       495    159    105    25    33    24    15     1
2   2020-04-05 22:27:00        98        75 Functional Fitness                         16.3       792    183    141     3    18    15     7    34
3   2020-04-17 01:15:00        93        87 Functional Fitness                         15.4       743    176    129    12    27     8     5    43
4   2020-04-21 22:17:00        95        36 Other                                       5.3       106    131    100    11    77    10     0     0
5   2020-04-21 22:17:00        95        84 Other                                      10.3       496    148    116     2    28    49    19     0
6   2020-04-30 22:36:00        92        13 Running                                     6.4       109    163    125    23     4    12    29    29
```

Workouts per Day

When looking at my high green recovery day workouts, I noticed that I often performed multiple workouts per day.

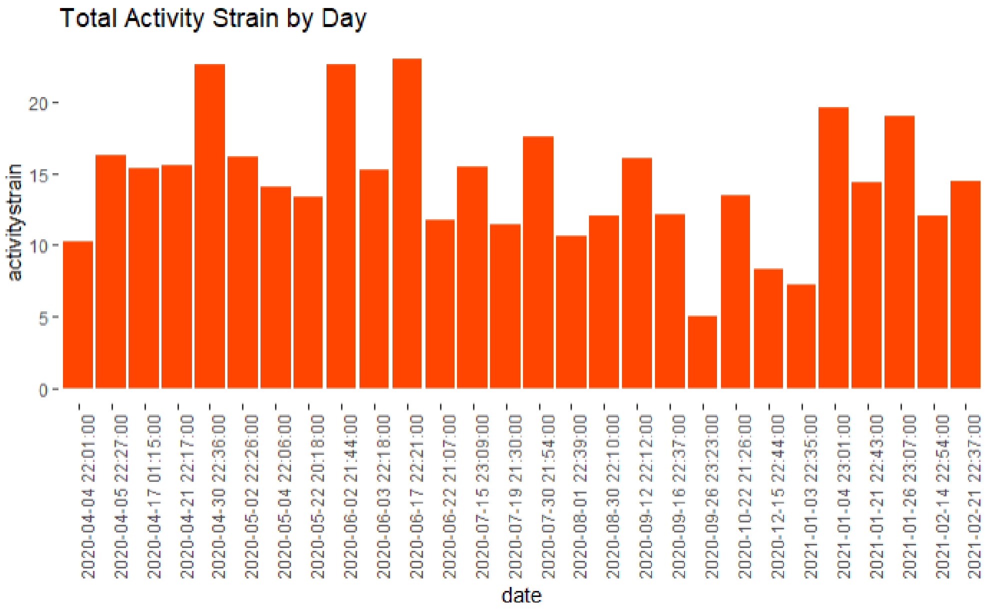


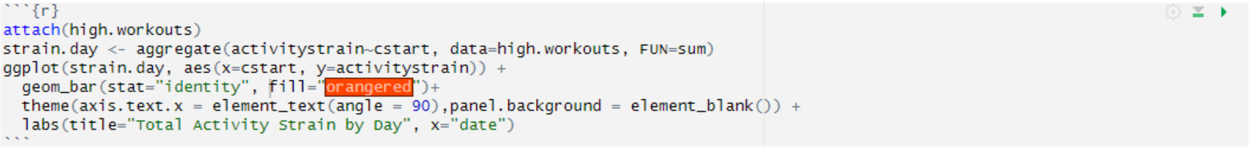
~~~
```{r}
high.workouts$cstart <- as.character(high.workouts$cstart)
class(high.workouts$cstart)
ggplot(high.workouts, aes(x=cstart)) +
  geom_bar(fill="seagreen3")+
  theme(axis.text.x = element_text(angle = 90),panel.background = element_blank()) +
  labs(title="# of Activities per Day", x="date")
```
~~~



Activity Strain by Day

I noticed that in the high green day workout data that most days I had activity strains >= 10. So, I took the sum of activity strain per day and visualized it by day.





~~~
```{r}
attach(high.workouts)
strain.day <- aggregate(activitystrain~cstart, data=high.workouts, FUN=sum)
ggplot(strain.day, aes(x=cstart, y=activitystrain)) +
  geom_bar(stat="identity", fill="orangered")+
  theme(axis.text.x = element_text(angle = 90),panel.background = element_blank()) +
  labs(title="Total Activity Strain by Day", x="date")
```
~~~

Higher activity strains lead to more calories burned, higher max heart rate, higher average heart rate and more time spent in the higher heart rate zones.

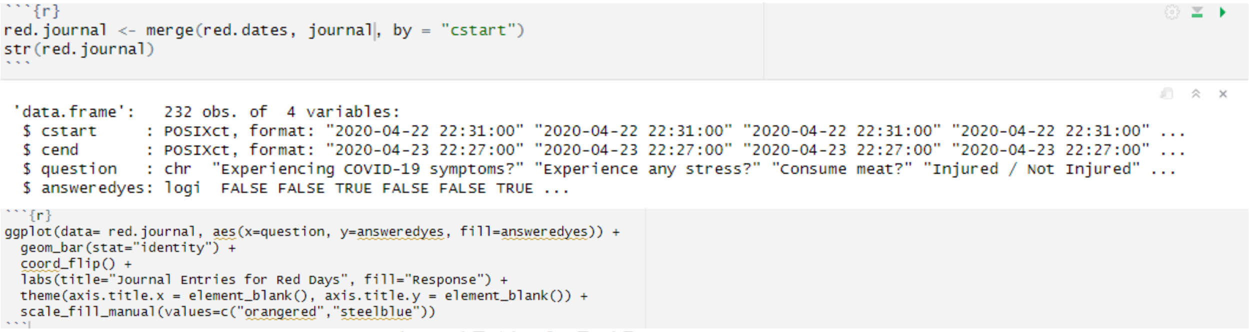
High vs. Low Recovery Workouts Conclusion

On days resulting in low recovery, I tend to not workout or not exert myself hard enough. Whereas on high recovery days I often perform more workouts that are higher in strain, burn more calories, and have more time spent in higher heart rate zones. What I learned is that doing something for exercise is better than doing nothing in relation to recovery, but that sometimes you can overexert yourself. So, it is important to know when to push your limits and when to reserve energy.

Exploring Journal Entries for High and Low Recovery Days

One of the unique features offered by Whoop is the daily journal. Whoop allows users to personalize their journal selecting questions most pertaining to their daily habits. I will admit to not having used this feature enough, but I am interested to see if there are noticeable journal entry differences between high and low recovery days.

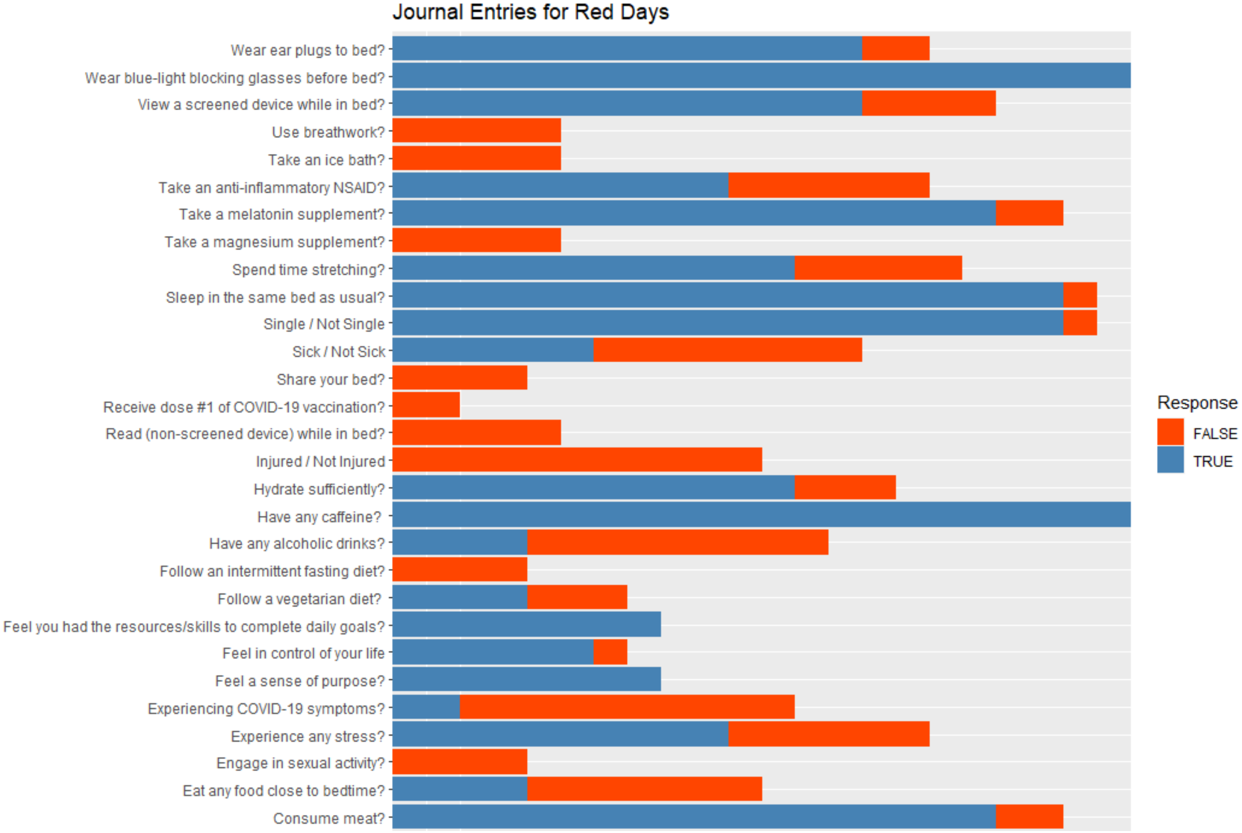
Red Day Journal Entries



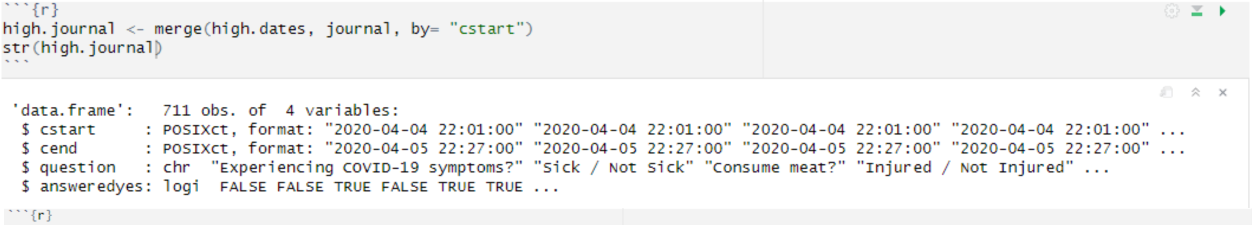
~~~
```{r}
red.journal <- merge(red.dates, journal, by = "cstart")
str(red.journal)
```
~~~

```
 'data.frame':   232 obs. of  4 variables:
  $ cstart     : POSIXct, format: "2020-04-22 22:31:00" "2020-04-22 22:31:00" "2020-04-22 22:31:00" "2020-04-22 22:31:00" ...
  $ cend       : POSIXct, format: "2020-04-23 22:27:00" "2020-04-23 22:27:00" "2020-04-23 22:27:00" "2020-04-23 22:27:00" ...
  $ question   : chr  "Experiencing COVID-19 symptoms?" "Experience any stress?" "Consume meat?" "Injured / Not Injured" ...
  $ answeredyes: logi  FALSE FALSE TRUE FALSE FALSE TRUE ...
```

~~~
```{r}
ggplot(data= red.journal, aes(x=question, y=answeredyes, fill=answeredyes)) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(title="Journal Entries for Red Days", fill="Response") +
  theme(axis.title.x = element_blank(), axis.title.y = element_blank()) +
  scale_fill_manual(values=c("orangered","steelblue"))
```
~~~



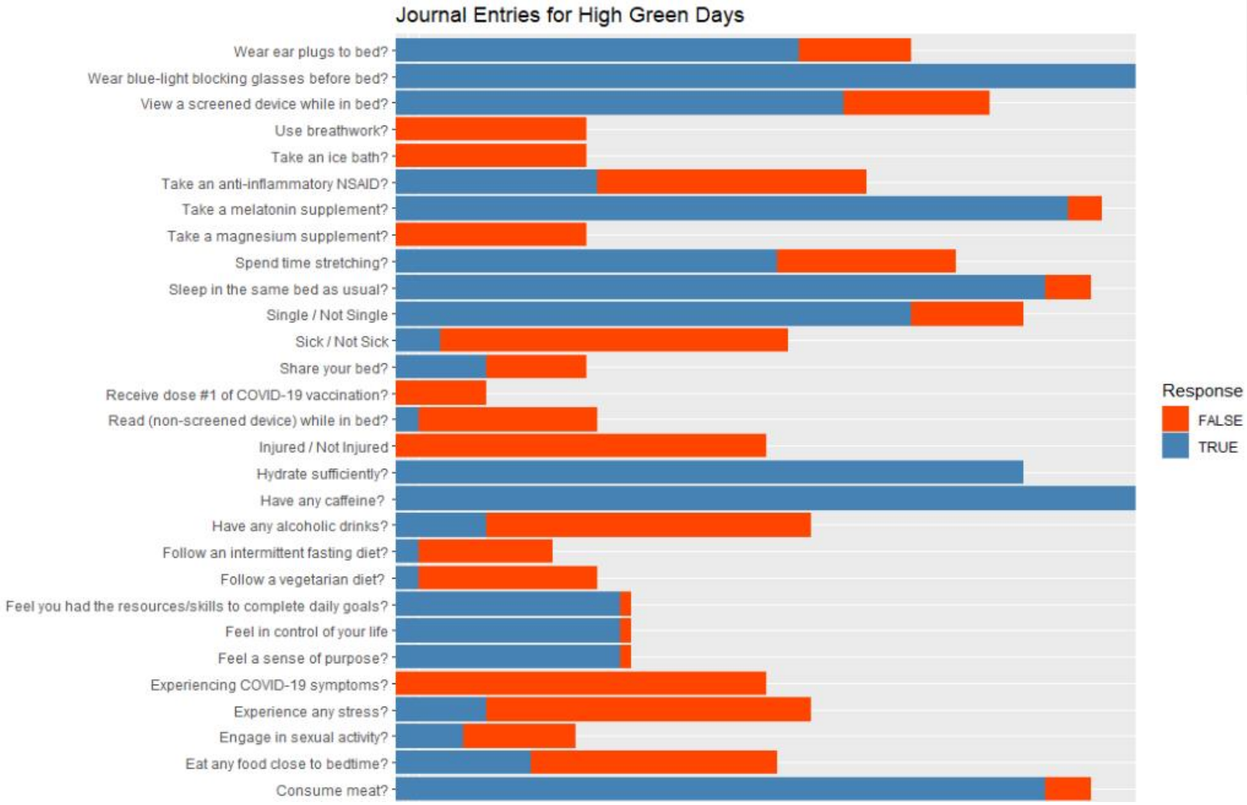
High Green Journal Entries



~~~
```{r}
high.journal <- merge(high.dates, journal, by= "cstart")
str(high.journal)
```
~~~

```
'data.frame':   711 obs. of  4 variables:
 $ cstart      : POSIXct, format: "2020-04-04 22:01:00" "2020-04-04 22:01:00" "2020-04-04 22:01:00" "2020-04-04 22:01:00" ...
 $ cend        : POSIXct, format: "2020-04-05 22:27:00" "2020-04-05 22:27:00" "2020-04-05 22:27:00" "2020-04-05 22:27:00" ...
 $ question    : chr  "Experiencing COVID-19 symptoms?" "Sick / Not Sick" "Consume meat?" "Injured / Not Injured" ...
 $ answeredyes: logi  FALSE FALSE TRUE FALSE TRUE TRUE ...
```

~~~
```{r}
~~~



Journal Entry Insights

Some daily habits exhibited no notable differences when comparing low recovery days vs. high recovery days. However, some notable insights are as follows:

- I tended to take less NSAIDs (ibuprofen) on days I had high recovery.
- I had more low recovery days where I didn't take melatonin than when I did.
- All red recovery days I reported being single. Where some of the highest recovery days I reported being in a relationship, sharing my bed, and engaging in sexual activity.
- I reported being sick more on red days.
- I reported reading a non-screened device more on high recovery days.
- I reported sufficient hydration for all high recovery days and insufficient hydration more on low recovery days.
- I reported consuming alcohol more on low recovery days.

- I had more low recovery days after reporting a vegetarian diet vs. high recovery days when reporting consuming meat.
- On low recovery days I reported experiencing more stress and feeling less in control of my life.

What is interesting about the journal entries is that it seems happiness or emotional well-being plays a significant role in recovery. Alcohol, hydration, and diet also play a role in recovery. Physical well-being seems to be in direct relation to recovery like how when I reported sick was most often on low recovery days. Based on journal entries, some habitual factors that I can control to increase my recovery scores would be socializing more, staying hydrated, minimizing stress, and consuming meat.

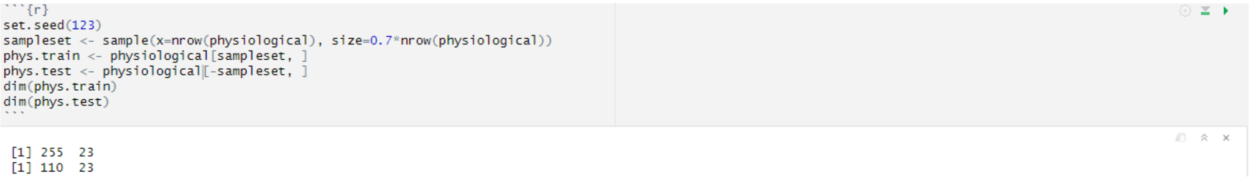
Predicting Recovery

In the following section I will explore various prediction modeling techniques to evaluate which variables are most influential in predicting recovery. Each prediction model will be tested for performance. It is important to test a variety of methods and evaluate performance to avoid overfitting or selecting a biased model.

Simple Linear Regression

In the exploration of recovery, I discovered a high correlation and strong positive linear relationship between HRV and Recovery. So, I will begin my predictive modeling methods with a simple linear regression between the two variables. I will split the data into a training set and a test set for validation using 70% of the data for training the simple linear regression model and 30% for testing the model.

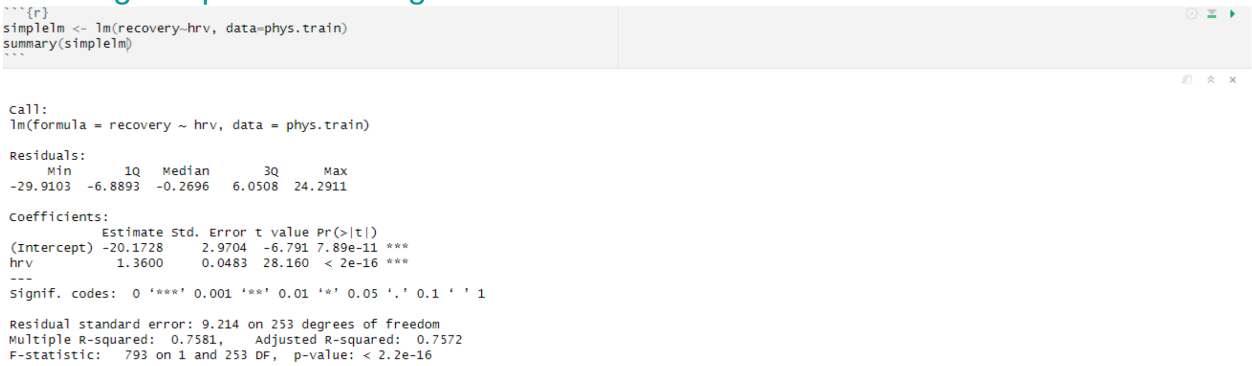
Creating Train and Test Sets



~~~
```{r}
set.seed(123)
sampleset <- sample(x=nrow(physiological), size=0.7*nrow(physiological))
phys.train <- physiological[sampleset, ]
phys.test <- physiological[-sampleset, ]
dim(phys.train)
dim(phys.test)
```
~~~

```
 [1] 255  23
 [1] 110  23
```

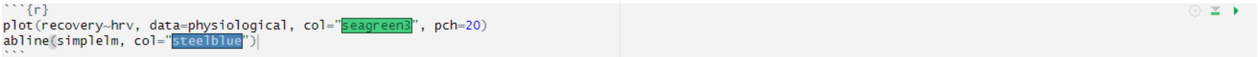
Running Simple Linear Regression



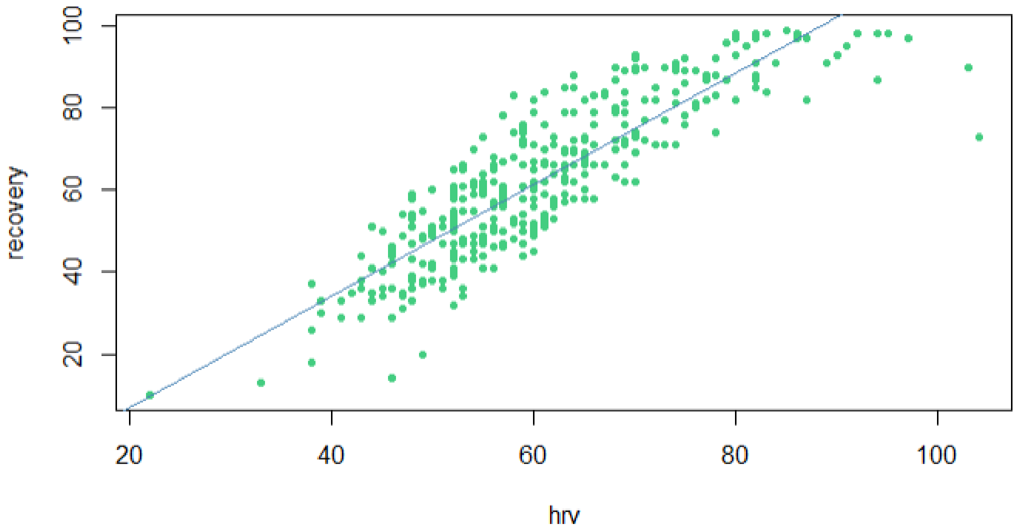
~~~
```{r}
simplelm <- lm(recovery~hrv, data=phys.train)
summary(simplelm)
```
~~~

```
Call:
lm(formula = recovery ~ hrv, data = phys.train)

Residuals:
     Min       1Q   Median       3Q      Max 
-29.9103  -6.8893  -0.2696   6.0508  24.2911 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -20.1728     2.9704  -6.791 7.89e-11 ***
hrv           1.3600     0.0483  28.160  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.214 on 253 degrees of freedom
Multiple R-squared:  0.7581,	Adjusted R-squared:  0.7572 
F-statistic:   793 on 1 and 253 DF,  p-value: < 2.2e-16
```

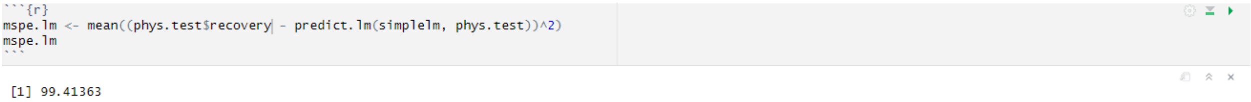
After running a simple linear regression on recovery by HRV, I can see that the p-value is less than 0.05 and can conclude that HRV does have a significant linear relationship with recovery. We can look at the linear regression line in comparison with the data to understand how well it matches.



```{r}
plot(recovery~hrv, data=physiological, col="seagreen3", pch=20)
abline(simplelm, col="steelblue")
```



I can see that the simple linear regression model fits the data relatively well. However, I would like to test this model on the remaining test data that was reserved for validation and calculate the test error of this model.



```{r}
mspe.lm <- mean((phys.test$recovery - predict.lm(simplelm, phys.test))^2)
mspe.lm
```

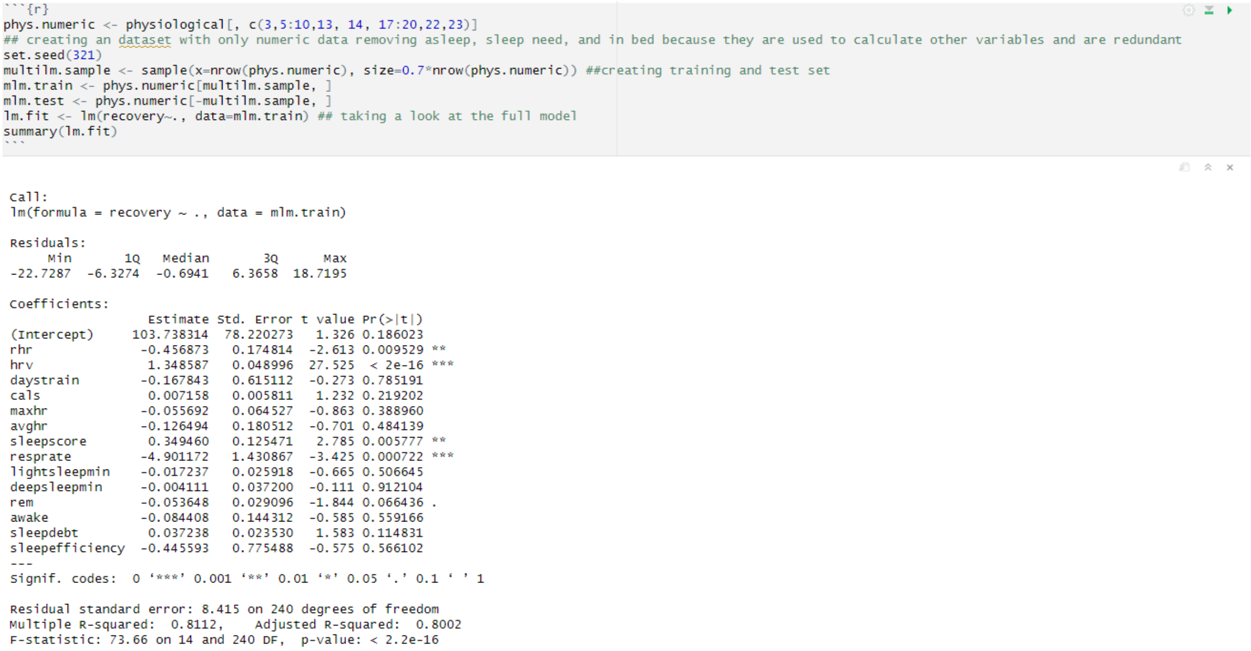
```
[1] 99.41363
```

The test MSE (mean squared error) for this simple linear model is 99.41. I will see if I can improve the test MSE with other models.

Multiple Linear Regression

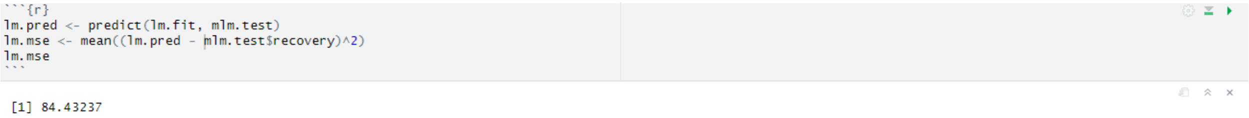
Next, I want to run a multiple linear regression model on all numerical variables in the physiological dataset.

Basic Multiple Linear Regression



~~~
```{r}
phys.numeric <- physiological[, c(3,5:10,13, 14, 17:20,22,23)]
## creating an dataset with only numeric data removing asleep, sleep need, and in bed because they are used to calculate other variables and are redundant
set.seed(321)
multilm.sample <- sample(x=nrow(phys.numeric), size=0.7*nrow(phys.numeric)) ##creating training and test set
mlm.train <- phys.numeric[multilm.sample, ]
mlm.test <- phys.numeric[-multilm.sample, ]
lm.fit <- lm(recovery~., data=mlm.train) ## taking a look at the full model
summary(lm.fit)
```
~~~

```
Call:
lm(formula = recovery ~ ., data = mlm.train)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.7287  -6.3274  -0.6941   6.3658  18.7195 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     103.738314  78.220273   1.326 0.186023    
rhr              -0.456873   0.174814  -2.613 0.009529 ** 
hrv               1.348587   0.048996  27.525  < 2e-16 ***
daystrain        -0.167843   0.615112  -0.273 0.785191    
cals              0.007158   0.005811   1.232 0.219202    
maxhr            -0.055692   0.064527  -0.863 0.388960    
avghr            -0.126494   0.180512  -0.701 0.484139    
sleepscore        0.349460   0.125471   2.785 0.005777 ** 
resprate         -4.901172   1.430867  -3.425 0.000722 ***
lightsleepmin    -0.017237   0.025918  -0.665 0.506645    
deepsleepmin     -0.004111   0.037200  -0.111 0.912104    
rem              -0.053648   0.029096  -1.844 0.066436 .  
awake            -0.084408   0.144312  -0.585 0.559166    
sleepdebt         0.037238   0.023530   1.583 0.114831    
sleepefficiency  -0.445593   0.775488  -0.575 0.566102    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.415 on 240 degrees of freedom
Multiple R-squared:  0.8112,	Adjusted R-squared:  0.8002 
F-statistic: 73.66 on 14 and 240 DF,  p-value: < 2.2e-16
```



~~~
```{r}
lm.pred <- predict(lm.fit, mlm.test)
lm.mse <- mean((lm.pred - mlm.test$recovery)^2)
lm.mse
```
~~~

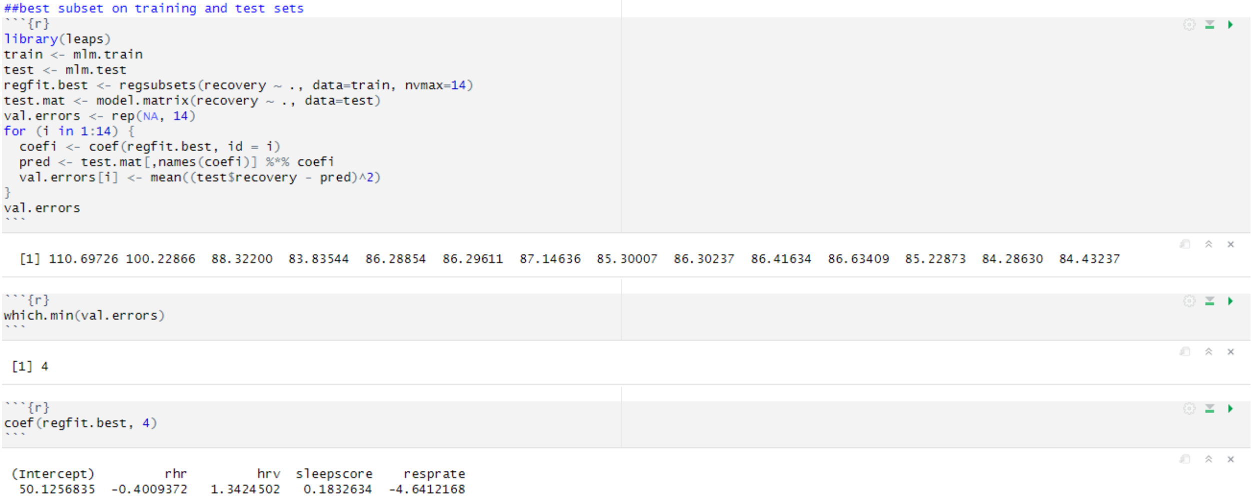
```
[1] 84.43237
```

Based on the output of the multiple linear regression model including all variables, the variables having p-values < .05 should be significant in predicting recovery. That means that rhr, hrv, sleepscore, and respiratory rate should be good predictor variables. However, I will use best subset selection to see if the subset of variables offering the best outcome for predicting recovery change. The test error (MSE) for the full multiple linear regression model is 84.43. I will see if this is reduced after performing best subset selection with cross validation.

Best Subset Selection

Best subset selection runs all possible models for the data. In other words, every combination of variables will be run on the data to determine which subset of variables offers the best variables for predicting recovery.

Best Subset Validation Set Approach



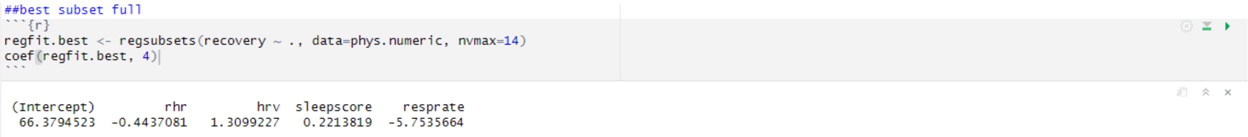
~~~
##best subset on training and test sets
```{r}
library(leaps)
train <- mlm.train
test <- mlm.test
regfit.best <- regsubsets(recovery ~ ., data=train, nvmax=14)
test.mat <- model.matrix(recovery ~ ., data=test)
val.errors <- rep(NA, 14)
for (i in 1:14) {
  coefi <- coef(regfit.best, id = i)
  pred <- test.mat[,names(coefi)] %*% coefi
  val.errors[i] <- mean((test$recovery - pred)^2)
}
val.errors
~~~

 [1] 110.69726 100.22866 88.32200 83.83544 86.28854 86.29611 87.14636 85.30007 86.30237 86.41634 86.63409 85.22873 84.28630 84.43237

```{r}
which.min(val.errors)
```

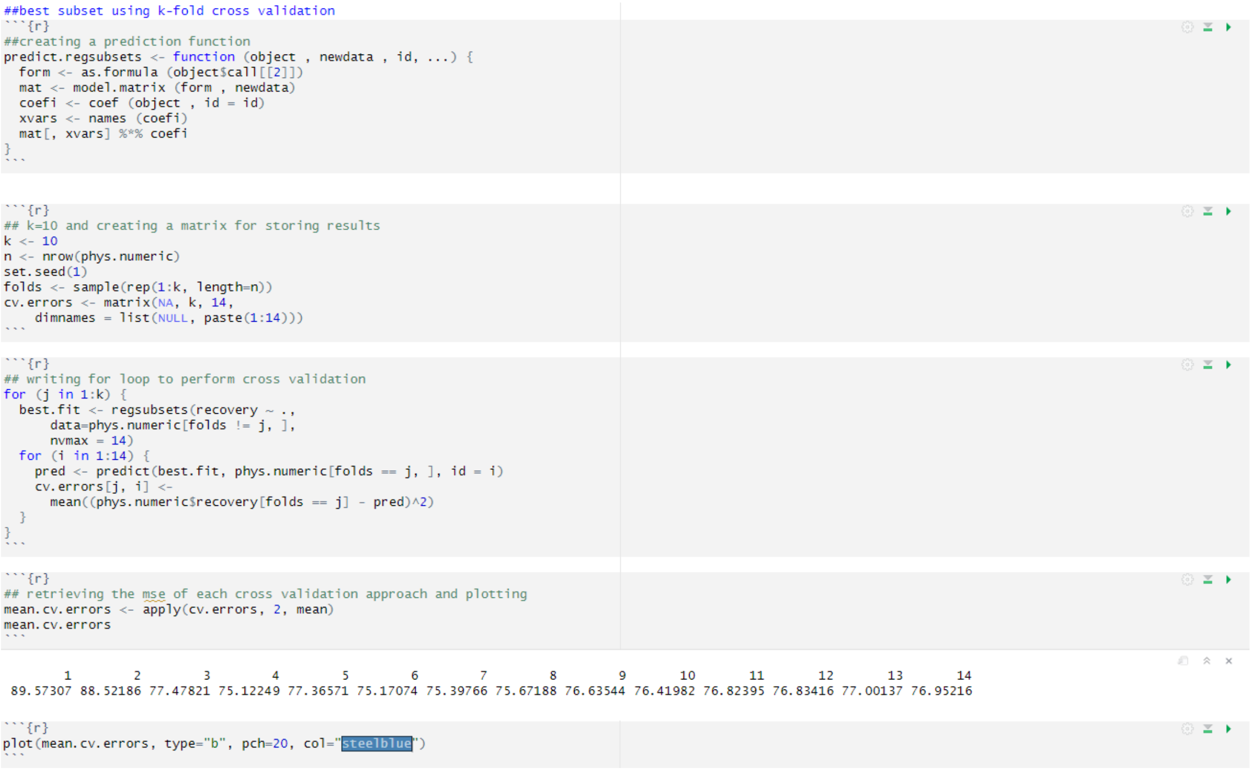
[1] 4

```{r}
coef(regfit.best, 4)
```

(Intercept)   rhr   hrv sleepscore   resprate
 50.1256835 -0.4009372   1.3424502   0.1832634 -4.6412168
```

After performing best subset selection on the training and test set, I can say that the best subset is a model containing 4 variables: rhr, hrv, sleepscore, and respiratory rate. However, it is important to run the best subset on the full dataset to get the most accurate coefficients. So, I will run best subset again to find the best 4 variables from the full dataset.

### Best Subset Full



```
##best subset full
```{r}
regfit.best <- regsubsets(recovery ~ ., data=phys.numeric, nvmax=14)
coef(regfit.best, 4)
```

(Intercept)   rhr   hrv sleepscore   resprate
 66.3794523 -0.4437081   1.3099227   0.2213819 -5.7535664
```

The variables remained the same, however the intercept and coefficients changed to offer a more accurate prediction of recovery. The last step for best subset selection is running the best subset with k-folds cross validation.
```

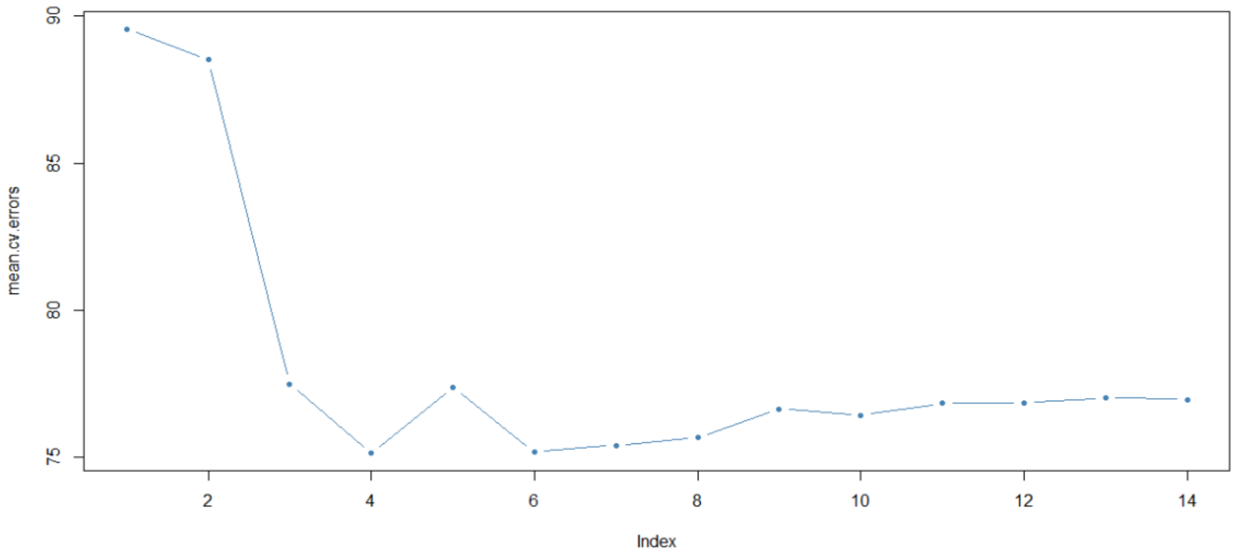
Best Subset using K-Folds Cross Validation



~~~
##best subset using k-fold cross validation
```{r}
##creating a prediction function
predict.regsubsets <- function (object , newdata , id, ...) {
  form <- as.formula (object$call[[2]])
  mat <- model.matrix (form , newdata)
  coefi <- coef (object , id = id)
  xvars <- names (coefi)
  mat[, xvars] %*% coefi
}
```

```{r}
## k=10 and creating a matrix for storing results
k <- 10
n <- nrow(phys.numeric)
set.seed(1)
folds <- sample(rep(1:k, length=n))
cv.errors <- matrix(NA, k, 14,
    dimnames = list(NULL, paste(1:14)))
```

```{r}
## writing for loop to perform cross validation
for (j in 1:k) {
  best.fit <- regsubsets(recovery ~ .,
      data=phys.numeric[folds != j, ],
      nvmax = 14)
  for (i in 1:14) {
    pred <- predict(best.fit, phys.numeric[folds == j, ], id = i)
    cv.errors[j, i] <-
      mean((phys.numeric$recovery[folds == j] - pred)^2)
  }
}
```

```{r}
## retrieving the mse of each cross validation approach and plotting
mean.cv.errors <- apply(cv.errors, 2, mean)
mean.cv.errors
```
~~~

```
        1        2        3        4        5        6        7        8        9       10       11       12       13       14 
 89.57307 88.52186 77.47821 75.12249 77.36571 75.17074 75.39766 75.67188 76.63544 76.41982 76.82395 76.83416 77.00137 76.95216 
```

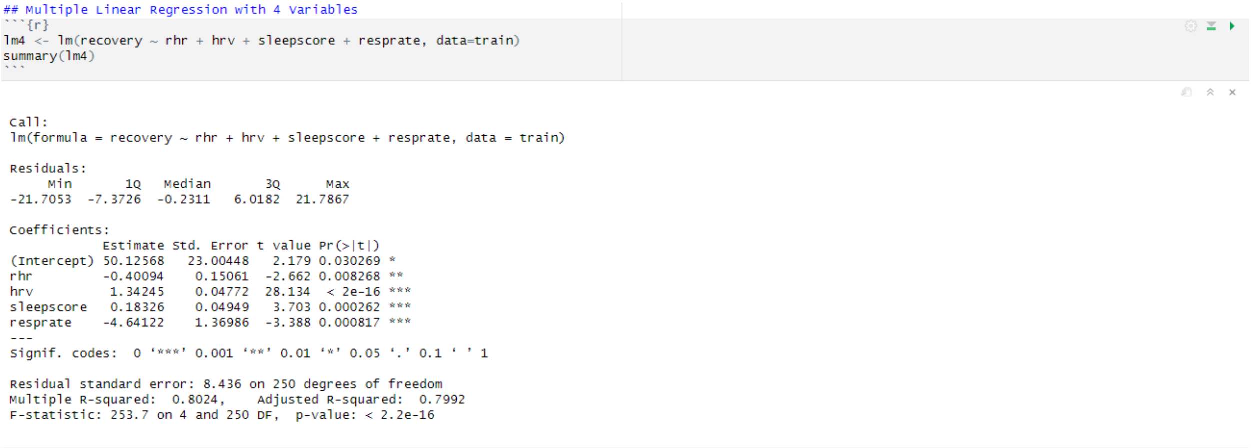
~~~
```{r}
plot(mean.cv.errors, type="b", pch=20, col="steelblue")
```
~~~



Based on the returned vector of MSEs and the plot visualizing these errors the 4 variable model is still the best prediction model for recovery. The 4 variable model has an MSE of 75.12. Again, our best prediction variables are rhr, hrv, sleepscore, and respiratory rate.

Re-running Multiple Linear Regression using Best Subset

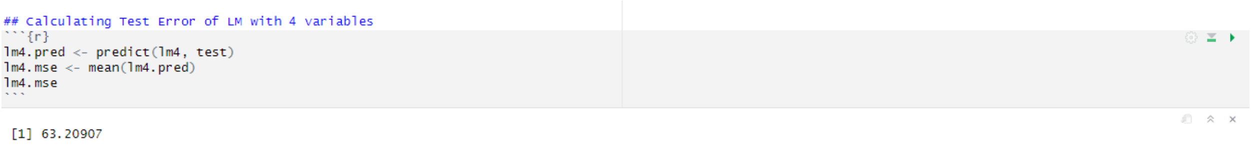
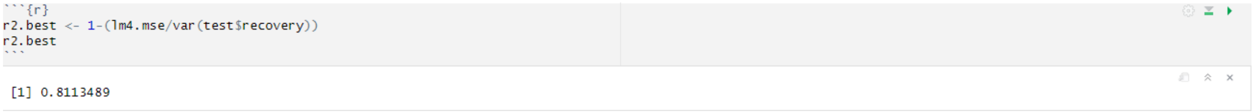
After running the best subset selection using 10-fold cross validation, I can confirm that the 4 most influential variables for predicting recovery are rhr, hrv, sleepscore, and resprate. I will now run a multiple linear regression model on these variables and then calculate the MSE.



~~~
## Multiple Linear Regression with 4 Variables
```{r}
lm4 <- lm(recovery ~ rhr + hrv + sleepscore + resprate, data=train)
summary(lm4)
~~~
```

```
Call:
lm(formula = recovery ~ rhr + hrv + sleepscore + resprate, data = train)

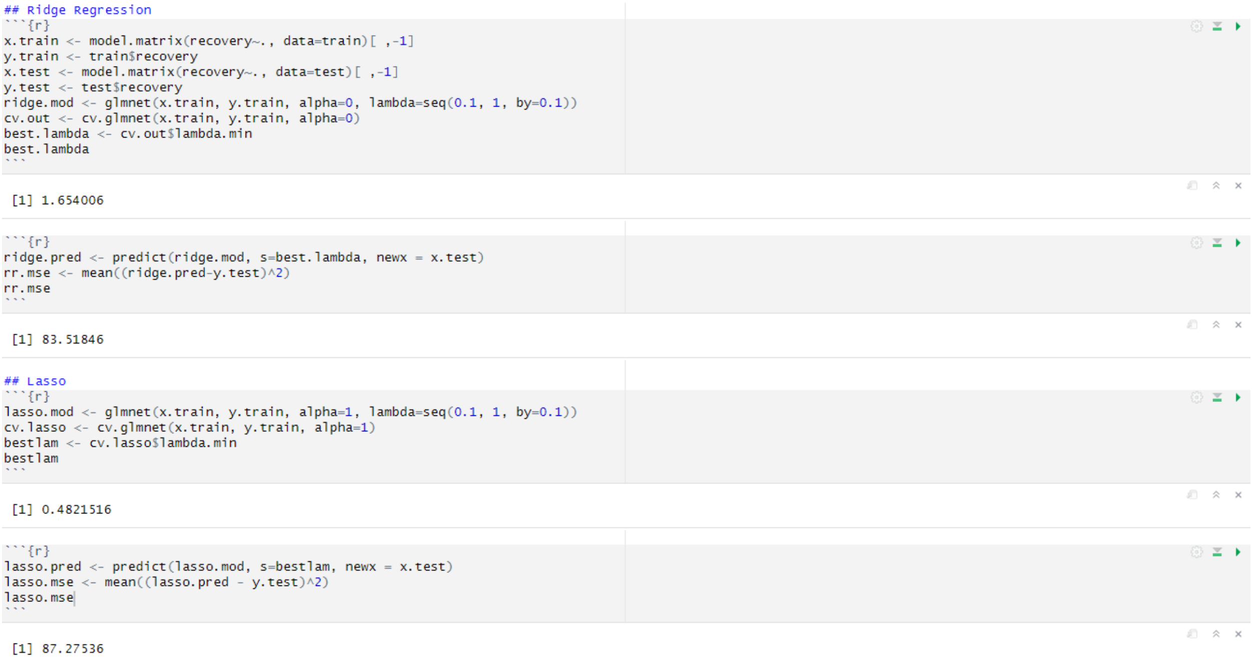
Residuals:
     Min       1Q   Median       3Q      Max 
-21.7053  -7.3726  -0.2311   6.0182  21.7867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.12568   23.00448   2.179 0.030269 *  
rhr         -0.40094    0.15061  -2.662 0.008268 ** 
hrv          1.34245    0.04772  28.134  < 2e-16 ***
sleepscore   0.18326    0.04949   3.703 0.000262 ***
resprate    -4.64122    1.36986  -3.388 0.000817 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.436 on 250 degrees of freedom
Multiple R-squared:  0.8024,	Adjusted R-squared:  0.7992 
F-statistic: 253.7 on 4 and 250 DF,  p-value: < 2.2e-16
```



```
## Calculating Test Error of LM with 4 variables
```{r}
lm4.pred <- predict(lm4, test)
lm4.mse <- mean(lm4.pred)
lm4.mse
```
```

```
[1] 63.20907
```

This new multiple linear regression model utilizing the 4 variables selected using best subset selection and k-fold cross validation has provided the best model yet. With an MSE of 63.21, the lowest MSE of all models run this far. The prediction formula using this model would be:

Recovery = intercept +/- rhr coefficient +/- hrv coefficient +/- sleepscore coefficient +/- resprate coefficient +/- ε. Using the output of the lm() function the equation is as follows.

Recovery = 50.13 – 0.40(rhr) + 1.34(hrv) + 0.18(sleepscore) – 4.64(resprate) + ε

The R^2 for this model is 0.81 which indicates that this model has a high probability of accurately predicting recovery.



```
```{r}
r2.best <- 1-(lm4.mse/var(test$recovery))
r2.best
```
```

```
[1] 0.8113489
```
```

## Ridge Regression and Lasso

Next, I'll run a Ridge Regression model and Lasso model to see if either will improve upon the MSE. The lower the MSE the better the model. As of right now the best model is the best subset model with 4 variables using k-folds cross validation. With an MSE of 75.12.

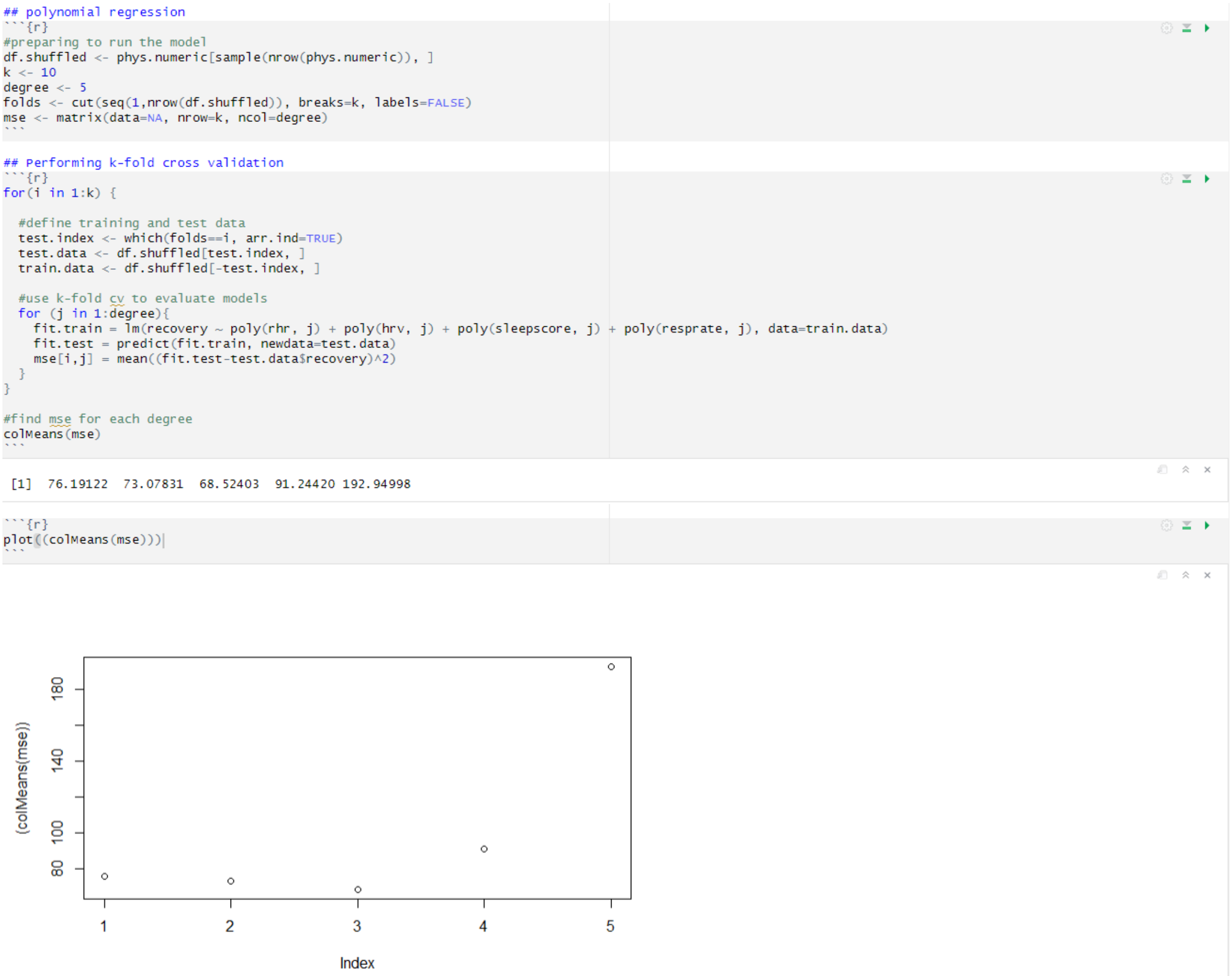


~~~
## Ridge Regression
```{r}
x.train <- model.matrix(recovery~., data=train)[ ,-1]
y.train <- train$recovery
x.test <- model.matrix(recovery~., data=test)[ ,-1]
y.test <- test$recovery
ridge.mod <- glmnet(x.train, y.train, alpha=0, lambda=seq(0.1, 1, by=0.1))
cv.out <- cv.glmnet(x.train, y.train, alpha=0)
best.lambda <- cv.out$lambda.min
best.lambda
```

 [1] 1.654006

```{r}
ridge.pred <- predict(ridge.mod, s=best.lambda, newx = x.test)
rr.mse <- mean((ridge.pred-y.test)^2)
rr.mse
```

 [1] 83.51846

Lasso
```{r}
lasso.mod <- glmnet(x.train, y.train, alpha=1, lambda=seq(0.1, 1, by=0.1))
cv.lasso <- cv.glmnet(x.train, y.train, alpha=1)
bestlam <- cv.lasso$lambda.min
bestlam
```

 [1] 0.4821516

```{r}
lasso.pred <- predict(lasso.mod, s=bestlam, newx = x.test)
lasso.mse <- mean((lasso.pred - y.test)^2)
lasso.mse
```

 [1] 87.27536
~~~

Neither Ridge Regression nor Lasso model improves upon the MSE. So, I will reject these models.

Polynomial Regression using K-Folds Cross Validation

The last method I used to predict recovery was polynomial regression. While the relationship between recovery and hrv was linear, the relationship between recovery and rhr, sleepscore, and resprate were not. Polynomial regression is used when relationships of variables are nonlinear.



~~~
## polynomial regression
```{r}
#preparing to run the model
df.shuffled <- phys.numeric[sample(nrow(phys.numeric)), ]
k <- 10
degree <- 5
folds <- cut(seq(1,nrow(df.shuffled)), breaks=k, labels=FALSE)
mse <- matrix(data=NA, nrow=k, ncol=degree)
```

## Performing k-fold cross validation
```{r}
for(i in 1:k) {

  #define training and test data
  test.index <- which(folds==i, arr.ind=TRUE)
  test.data <- df.shuffled[test.index, ]
  train.data <- df.shuffled[-test.index, ]

  #use k-fold cv to evaluate models
  for (j in 1:degree){
    fit.train = lm(recovery ~ poly(rhr, j) + poly(hrv, j) + poly(sleepscore, j) + poly(resprate, j), data=train.data)
    fit.test = predict(fit.train, newdata=test.data)
    mse[i,j] = mean((fit.test-test.data$recovery)^2)
  }
}

#find mse for each degree
colMeans(mse)
```
~~~

```
 [1]  76.19122  73.07831  68.52403  91.24420 192.94998
```

~~~
```{r}
plot((colMeans(mse)))
```
~~~

The polynomial regression using degree of 3 provides the lowest MSE for these models with an MSE of 68.52. However, the multiple linear regression model utilizing the best subset still provides the best outcome.

Prediction Models Conclusion

The best model for predicting recovery is the multiple linear regression model using 4 prediction variables rhr, hrv, sleep score, and respiratory rate. With this model, I was able to predict recovery within +/- 8 points. A variety of prediction methods were used to evaluate which model provided the best outcome or lowest MSE. I considered running logistic models to determine whether recovery score would be high or low. However, I decided that for the purposes of this project and analysis it was more important to be able to predict overall recovery. By analyzing all the variables and examining the trends I was able to identify habits and controllable factors which would allow me to improve my recovery score in the future and ultimately increase my athletic performance.

Peak Performance - Analysis Conclusion

After exploring the data, I have gained insights into some variables that I can use to increase my recovery and in return peak performance. I learned that my recovery is highly dependent on my heart rate variability. Sleep is also important in recovery and not just the amount of time slept, but the quality of sleep. The more time spent in restorative sleep cycles of deep sleep and REM sleep are associated with overall higher recovery scores. While sleep disturbances and not getting enough sleep contribute to lower recovery scores.

Exercise and cardiovascular strain are key contributors to increasing HRV and in turn recovery scores. I discovered that doing something, in terms of exercise, is better than doing nothing. Being able to control my cardiovascular output to spend more time in restorative heart rate zone could help me to increase my recovery scores in the future. Pushing myself to higher thresholds on high green recovery days will ultimately lead to higher HRV and my ability to withstand more intense training activities without compromising my recovery.

Based on journal entry information, I learned that psychosocial factors such as happiness, social activities, and being in a relationship lead to better recovery days. I also learned that days where I experienced higher than normal stress or was feeling depressed led to lower recovery days. My diet also played a factor in high and low recovery. I often had lower recovery on days where I did not consume meat and lower recovery on days where I reported consuming a vegetarian diet.

After developing and testing several predictive models for predicting recovery, I can confirm that heart rate variability (HRV), resting heart rate (RHR), sleep performance, and respiratory rate are the best prediction variables for recovery. The multiple linear regression model utilizing these four variables provided the most accurate model for predicting recovery score.

Overall, my key take aways are that I should be more social and do things to minimize my stress. Keeping my diet consistent and consuming meat more often may help me maintain or increase my recovery. I should explore pacing my workouts to exert the right amount of energy for optimizing recovery and pay close attention to the time I am spending in each heart rate zone. Implementing these habits and changes will contribute to achieving peak performance.

References

[1] https://www.whoop.com/experience/recovery/

[2] https://www.whoop.com/experience/

[3] https://www.whoop.com/thelocker/everything-to-know-about-sleep/

[4] https://www.whoop.com/thelocker/borg-scale-perceived-exertion-rpe/

[5] https://www.whoop.com/thelocker/how-does-whoop-strain-work-101/

[6] https://www.whoop.com/thelocker/max-heart-rate-training-zones/

[7] https://www.whoop.com/thelocker/what-is-respiratory-rate-normal/