***Springboard Capstone Blog Post***

Joe Earnshaw

It seems like everybody has some sort of wearable on their wrist these days. From the well-known (and super popular) Apple Watch to the “how many steps have I taken today?” Fitbits, smartwatches and fitness trackers are everywhere and they’re here to stay ([1](https://www.idc.com/getdoc.jsp?containerId=prUS44247418)). This is just the start too. The wearable market is still in its infancy, seeing as the first Fitbit was released around 2009; Samsung released its Galaxy Gear in late 2013; and the Apple Watch wasn’t introduced until 2015 ([2](https://www.businessinsider.com/a-timeline-of-how-the-apple-watch-was-created-2015-3)). Knowing this, there is a lot of potential for what wearables could be in the not-so-distant future. There is one in particular though, called WHOOP, that might be ushering in a new era of wearable technology sooner rather than later.

While WHOOP has flown under the mainstream radar it has nonetheless gained quite the following amongst professional athletes in the NFL, MLB and NBA (including King James himself…). The list doesn’t stop there though; the list of athletes also includes a NASCAR driver, a Tour de France cyclist and top CrossFit Games athletes ([3](https://www.whoop.com/the-locker/)).

Now what makes WHOOP different? Well, the first obvious feature (or lack thereof) is the user interface; there is none. The only way to access the information is via the mobile or web-based app. Admittedly, this is an adjustment at first; after all, how are you going to check your texts?!? This device isn’t trying to be a smart watch though; it is completely devoted to being a health tracker, and nothing else. It is for this reason though that it excels and the data it is able to capture is nothing short of amazing.

Now let’s get into the amazing part, which is the wealth of information WHOOP provides to its user. At the surface, it may seem simplistic in its approach; three primary metrics are provided to the user -- Strain, Recovery, and Sleep Performance -- which show, respectively, how strenuous that particular day has been, how well the body has responded and recovered from the previous day’s stimulus, and how much sleep you got versus how much you needed.

How are these scores calculated though? Well the strap is collecting data 100x/second, 24/hours a day, 7/days a week on the following biological markers:

* average heart rate
* max heart rate achieved that day
* calories burned
* total time spent in bed and asleep
* total time spent in light, REM, and deep (SWS) stages of sleep
* total sleep cycles
* heart-rate variability (HRV)
* resting heart rate (RHR)
* disturbances during sleep
* sleep latency (i.e. how long it took to fall asleep)

As you can see there is a lot more than meets the eye. Quick side note: one of its best features is that the charger is slid over the band, allowing for it to simultaneously continue tracking while also charging the battery.

Personally, this treasure trove of health-related data has significantly contributed to me creating an overall healthier lifestyle. For example, I make a more conscious effort to get to bed early, and ensure I’m getting adequate sleep. Also, I’ve significantly cut back on my drinking (I don’t completely abstain though, I’ll have a brew or glass of wine every now and then).

Yet, it still seemed to be missing something. WHOOP gives feedback on a daily basis, which allows one to adapt their training on a day-to-day basis, but there isn’t a tool that allows for long-term scheduling. Life can be hectic, and there are days where it can be hard to get to the gym. However, what if a day that is going to be particularly busy could land on a ‘rest’ day? Also, what if on a day where your schedule is more open, your body was in peak condition, allowing you to train a little longer and harder?

My Springboard Mentor Jarus Singh and I started wondering: what if, with the information gathered from WHOOP over the past few months, we could be predict high level metrics such as Strain, Sleep Performance and Recovery using lower level variables to assist in creating a longer-term workout plan, say for a week?

Answering this question became the primary goal of the project. We wanted to see using the data available to us, if there were models that could predict each metric as accurately as possible using the variables mentioned above, such as max heart rate, heart rate variability and total sleep.

**Validation**

Before we dive into the models though, I want to address the following question: how do we know these models are reliable predictors of the Strain, Sleep Performance and Recovery metrics? Glad you asked!

The value of models comes not from predicting values on the current data set (which is technically called the ‘training data’) but from making predictions on new, previously unknown, data. If these models’ predictions don’t hold with new data, their essentially worthless ([4](https://www.kaggle.com/dansbecker/model-validation)). So, to check each model’s robustness, we used a method called K-fold cross validation.

The general outline of the procedure works is as follows:

1. Shuffle the dataset randomly
2. Split the dataset into K groups (with K being equal to 10 in our case)
3. For each of these unique groups:
   1. Take the group as a hold out or test data set
   2. Take the remaining groups as the training data set
   3. Fit a model on the training set and evaluate it on the test set
   4. Retain the evaluation score (in our case the RMSE, which I’ll discuss shortly) and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores ([5](https://machinelearningmastery.com/k-fold-cross-validation/))

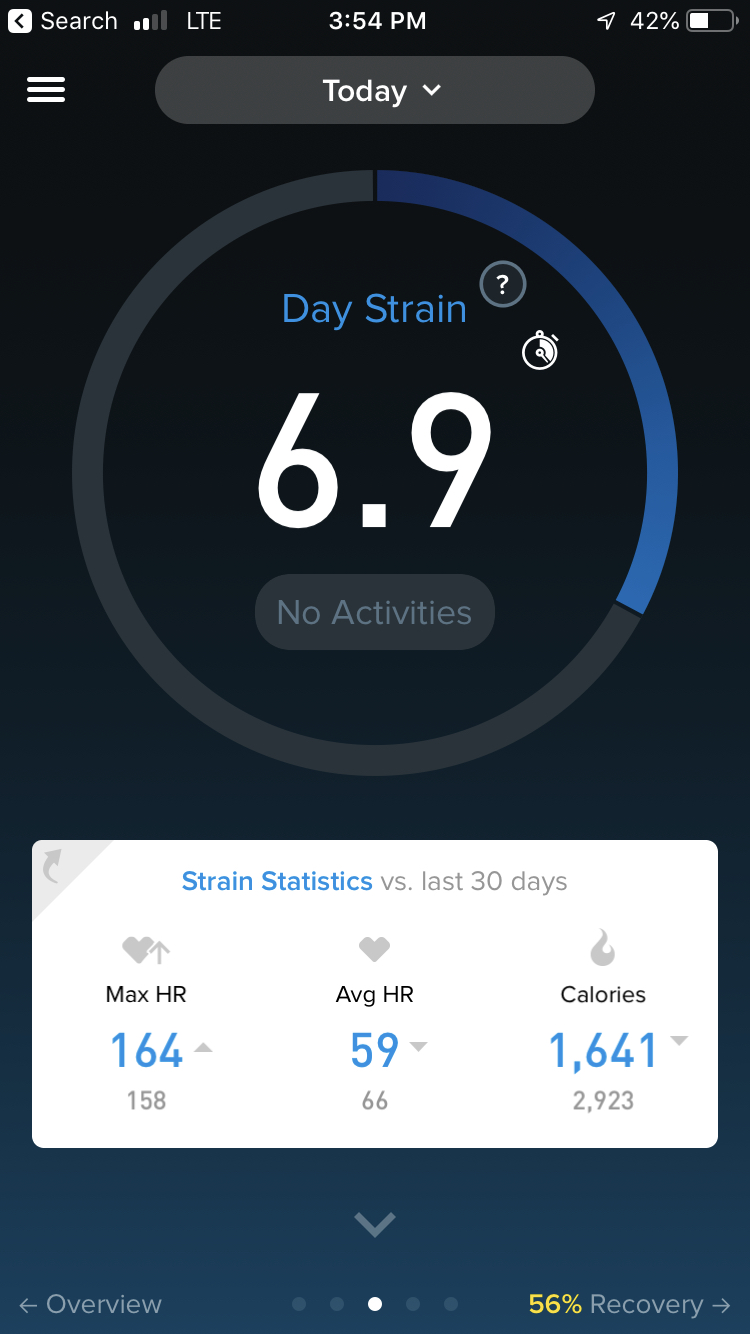
In simpler terms, it means that we’ll randomly split the data into 10 unique groups, after which we’ll use one of these groups as either a hold out or test data set. The rest of the groups are then combined to form the training data set which the model will be trained on first, after which it can then be evaluated on the test set.

The evaluation score in our case is the root-mean square error (RMSE). To avoid going too far down the rabbit hole, RMSE is an indicator of how close the observed data points are from the predicted data points based on the model ([6](https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/)) As an example, a ‘perfect’ model would have a RMSE of 0 because the predicted values are exactly the same as the observed values.

However, perfection is nearly impossible to accomplish but in general the lower the RMSE (i.e. predicted values are closer to the observed values), the better the model is at prediction. After acquiring the mean of the cross-validated RMSEs for the model, we can then compare this number to the model’s RMSE on the full data set. If those numbers are close than our model(s) are valid predictors of that particular metric.

Now with all that out of the way, the next step is to see if we can make some models!

***Strain***



The first metric we will explore is the Strain metric. In the WHOOP app, this is one of the first metrics user will see where it is “reported on a scale from zero to 21, [and] measures the total cardiovascular load experienced over a specified period of time - such as a workout or day - normalized such that a 21.0 represents the maximal cardiovascular load that could be attained in a day” ([1](https://www.whoop.com/the-locker/training-with-whoop-using-recovery-and-strain-to-unlock-your-potential/)).

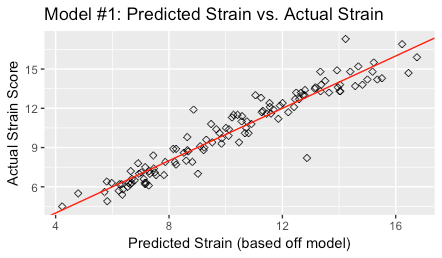
Cardiovascular load is gauged using your heart rate; every second of the day, the device is recording this. If your heart rate stays elevated for longer periods, due to events like working out or going grocery shopping with your kids, you placed more stress on your body and the result will be a higher Strain score.

The benefit of this is that Strain is a holistic metric; it doesn’t just take information from your workouts, but the day-to-day stresses outside the gym as well. Knowing this before going into my exploratory data analysis, I wanted to focus on the three variables -- max heart rate (MHR), average heart rate (AHR) and calories (CALS) -- that WHOOP reported along with the Strain metric.

As expected, all three variables exhibited strong positive relationships with Strain. The next step was to utilize these variables to create a model that could, as accurately as possible, predict the Strain metric.

The first model utilized all three variables -- MHR, AHR and CALS. For the full data set, the model RMSE was approximately 0.9030; what this number tells us is that the predicted Strain from the model would be somewhere between +/- 0.9 of the actual Strain. To check the validity of this model, we then ran it through the 10-fold cross validation and it returned a RMSE value of approximately 0.9381.

This was a very promising first model and set a rather high benchmark for the next models, which as it turned out would not be beat. There were three other models tested; the second model utilized AHR and CALS; the third model utilized MHR and CALS; and the fourth and final model utilized the two heart-related variables, MHR and AHR.

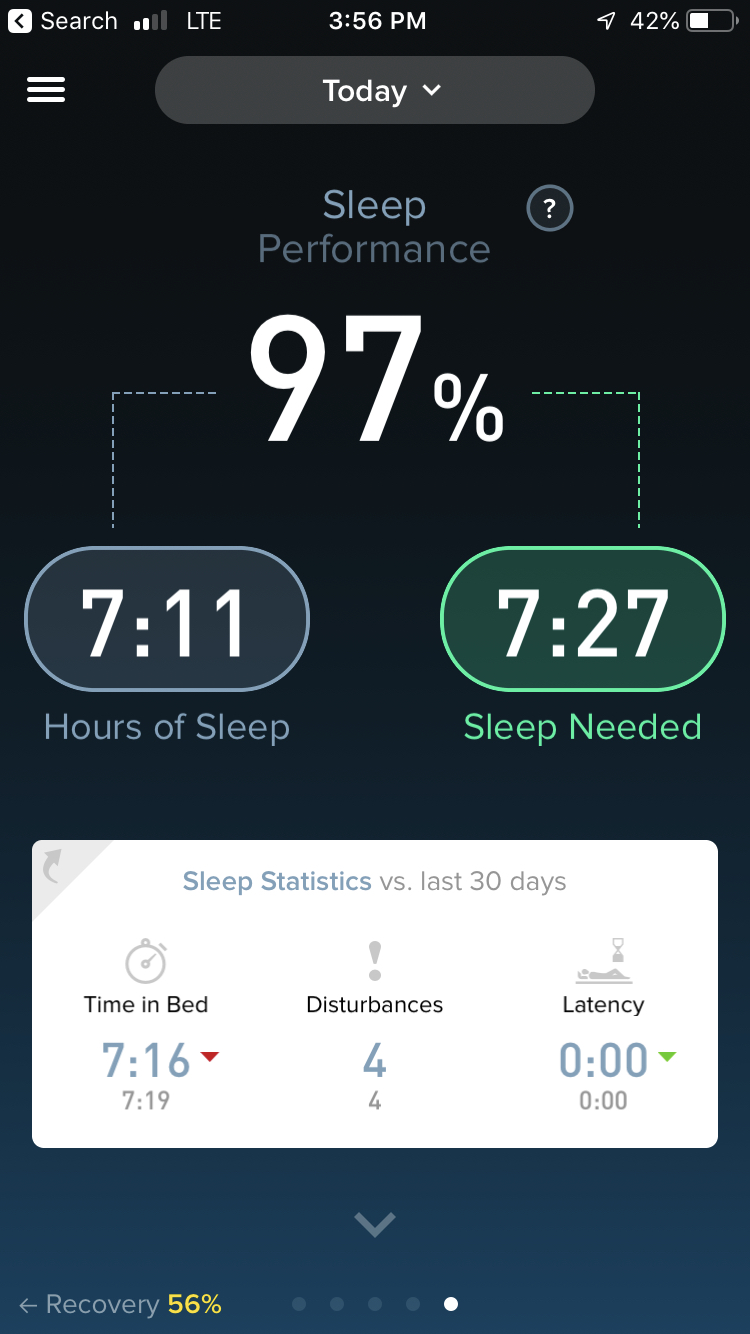


However, the only other model that came close was the third model with a cross-validated RMSE of approximately 0.9454.

As an additional confirmation, the image above is a visual representation of the first model with its predicted Strain scores (the red line) versus the actual observed Strain scores (black dots). As we can see, the vast majority of the points are hugging the line pretty close, which circles back to RMSE. The closer these points are to the line, the smaller the distance between them and as a result, the lower the RMSE is going to be.

***Sleep Performance***

The next metric we focused on was Sleep Performance. As a high-level overview, Sleep Performance is the total amount of sleep you got divided by the amount of sleep your body needed based on that day’s strain.



So, the higher your Strain metric was that day, the more stress you put on your body and as a result your body is going to need more sleep to recover. However, to look at Sleep Performance from this perspective is to miss out on all the interesting sleep-related variables WHOOP tracks while you catch some Z’s at night.

Below are the variables that we had data on for this project:

* Total time spent in bed (awake plus asleep)
* Total time spent asleep
* Time spent in each stage of sleep (light, REM, and deep)
* Sleep cycles (successfully completing each stage of sleep without disruption)

As we can see, sleep is a lot more complicated than just closing your eyes and counting sheep! Due to number of variables available to us, we ended up exploring quite a bit when it came to creating models for Sleep Performance.

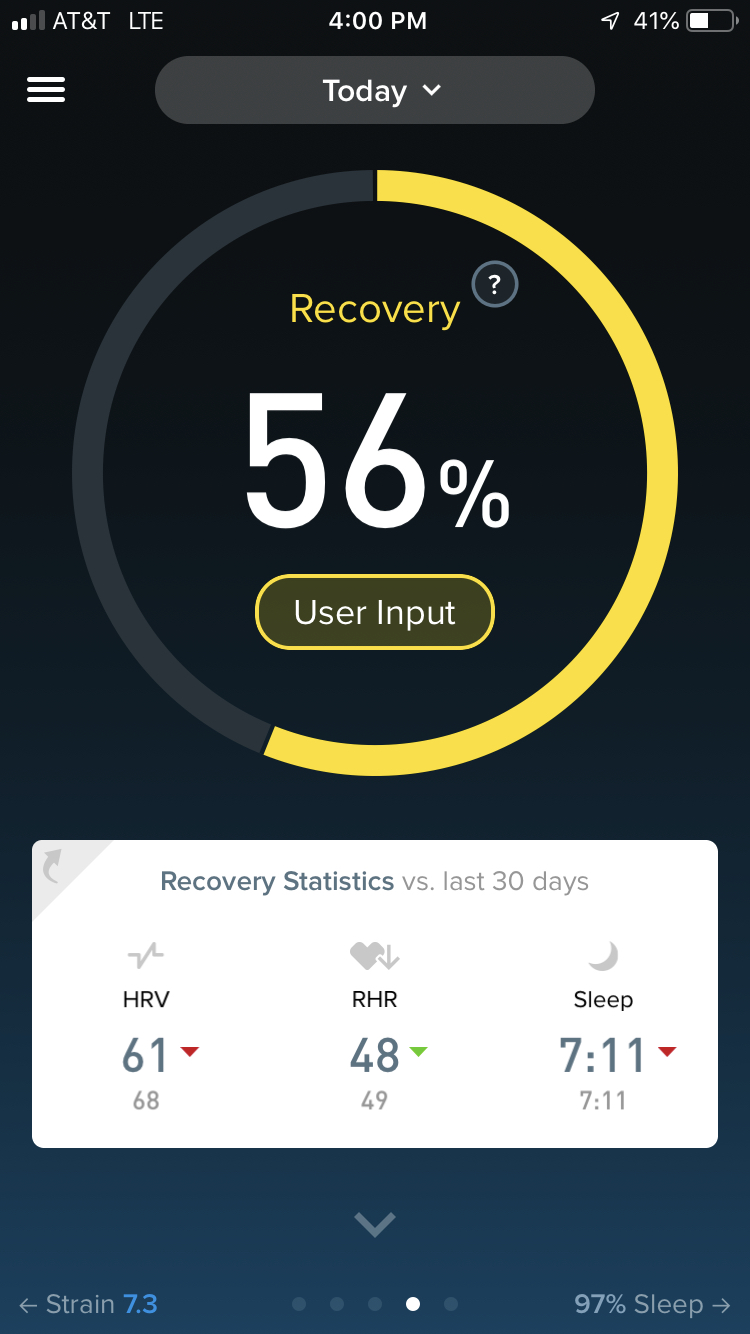
In total there were four models that we decided to test; the first model predicted sleep performance utilizing the time spent in each sleep stage (light, REM, deep) plus sleep cycles; the second model utilized just the sleep stages; the third model used total time in bed, total time asleep and the number of sleep cycles; the fourth and last model was similar to the third but it excluded the sleep cycle variable.

Compared to the Strain models, the RMSE of the cross-validated models were extremely narrow with each being somewhere between approximately +/- 0.07, or 7%, of the actual Sleep Performance value (which is reported as a percentage out of 100 in the app). We had to go out to the ten-thousandths place, to determine the ‘winning’ model which ended up being the third model that used the variables for the total time spent in bed and asleep plus the number of sleep cycles.

The third model’s cross-validated RMSE was 0.07147 (7.147%), which barely beat out the fourth model’s cross-validated RMSE of 0.07179 (7.179%). Despite the third one technically having the lowest RMSE value, the closeness of the four Sleep Performance models was interesting. For further research it would be interesting to dig down a little deeper to determine why and how these different variables were able to produce such similar numbers.

***Recovery***

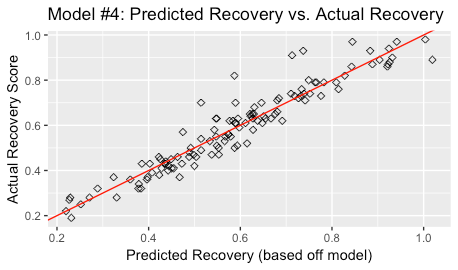
The last metric we’re going to look at is Recovery. According to WHOOP, the “Recovery [metric], reported on a scale from zero to 100 percent, measures the body’s ability to adapt to a training stimulus.” ([1](https://www.whoop.com/the-locker/training-with-whoop-using-recovery-and-strain-to-unlock-your-potential/)) In addition to the Recovery score (presented as a percentage), there are three primary variables that are associated with this particular metric in the app: HRV (heart-rate variability), RHR (resting heart-rate) and sleep (how much time you spent asleep).



During my exploratory data analysis, the relationship Recovery had with the above variables, amongst others, was a little different. In the previous metrics, the lower level variables associated with them tended to have strong positive relationships. And while two out of the three -- HRV and sleep -- had a positive relationship, resting heart rate had a negative relationship with Recovery.

What that meant was that the higher the average resting heart rate, the lower your recovery was going to be. Despite the somewhat contradictory nature of it, I decided to include it in the first model along with the HRV and total sleep variables (since those three variables are shown together with Recovery in the app). The second model was similar to the first in that it utilized HRV and total sleep but excluded resting heart rate.

I also had a hunch from my exploratory data analysis that despite the initial appearance of a somewhat ambiguous relationship, Sleep Performance may be a good explanatory variable to predict Recovery. My intuition was that Sleep Performance is calculated based on how much sleep you needed versus how much sleep you actually got, so theoretically it should be a more reliable indicator of how ‘rested’ you were versus just total sleep. Based off this, I decided to utilize HRV and Sleep Performance for the third model.



Another hypothesis that I wanted to test was that Recovery was an all-encompassing metric; it used information from other metrics in addition to variables such as HRV to calculate its score. Intuitively this makes sense; you cannot determine recovery unless you first know how much stress the body went through, and if the body got an adequate amount of quality sleep to repair any damage done by the stress.

With this in mind, the fourth model for Recovery utilized Sleep Performance, Calories and HRV. The fifth and final model also used Calories and HRV, but total sleep was subbed for Sleep Performance to see if there would be any significant difference between the two.

When cross-validated, the Recovery models’ results were extremely narrow, with the RMSE being approximately +/- 0.06, or 6%, of the actual Recovery value. However, there was one that slide in right under 0.06, and that was the fourth model with a RMSE of approximately 0.05987.

Remember that this model utilized both Calories and Sleep Performance in addition to HRV to calculate the value of Recovery. The cool thing about this result was that opens the door for further investigation into Recovery being an all-encompassing metric.

***Results***

The results for each of these models have been way better than I expected. We were able to find models that predicted Strain within 1 point of its actual value, within roughly 7% for Sleep Performance and within approximately 6% for Recovery.

While not perfect, these models have potential value in optimizing my training. For example, using the model for Strain, in combination with previous workouts, I can now better gauge the number of workouts for each training session while minimizing the risk of overtraining. Additionally, with the Sleep Performance model I can assess the amount of time I should allot for sleeping each night in order to maximize my body’s ability to repair itself. Lastly, with the Recovery model I can better plan out my weekly workout schedule according to how intense I want my training to be on that day, and when I want to take a rest day. With a larger data set, further analysis and some fine-tuning though I believe there is potential to make each of these models more accurate, even to the point of being a viable option as an in-app tool.

Below are the formulas for each of the metrics best model:

Strain = -13.951038 + (0.066266 \* maxHR) + (0.070551 \* averHR) + (0.003163 \* cal)

Sleep Performance = 0.095025 + (0.033334 \* timeInBed) + (0.067111 \* totalSleep) +

(-0.007027 \* sleepCycles)

Recovery = -0.5837 + (0.01299 \* hrv) + (0.2274 \* sleepPerform) + (0.00002720 \* cal)

These formulas represent how each metric – Strain, Sleep Performance and Recovery – is calculated. The first number in each equation (i.e. that isn’t multiplied by a variable) represents the linear models intercept; the numbers in front of the variables are the model’s parameter estimates for that particular variable. For example, let’s focus on the formula for Strain and let’s plug in a 180 for maxHR, 70 for averHR, and 3,500 for cal.

MaxHR: 0.066266 \* (180) = 11.92788

AverHR: 0.070551 \* (70) = 4.93857

Cal: 0.003163 \* (3500) = 11.0705

With these example inputs, the model would predict a Strain score of approximately 13.9.

Strain: -13.951038 + 11.92788 + 4.93857 + 11.0705 = (~)13.9

However, WHOOP rolled out an updated Sleep Coach feature, which I haven’t touched on yet. What this feature used to do was rather simple; it would tell you when to go to bed based on how much sleep you needed to perform at a certain level (i.e. Peak, Perform, Rest) the next day.

With this new update, users can not only optimize sleep with specific sleep/wake times, but it allows you to plan out how recovered you are each day of the week. This essentially means that WHOOP developed a way to plan workouts on a weekly basis!

While it was an awesome coincidence, it highlighted that interpreting data may be just as important as creating models based off it. You may gain some insights by throwing a whole bunch of models at a data set, but are you going to understand the results? And if you don’t understand the results, where is the value in the model?

I think this project highlighted that asking the right questions can lead to some very valuable information. WHOOP was asking a similar question when they came up with their idea for an updated Sleep Coach because they too saw value in giving their users greater flexibility when it comes to planning their workouts.

Overall, learning and becoming more knowledgeable in the statistics and programming skills that go into a project like this have been very rewarding. I’d have to say though that my key takeaway moving forward is this: never stop asking questions, because you never know what you may find.

I want to first thank Jarus Singh, my Springboard mentor for his help with this project. Your guidance throughout this process has been invaluable and without it this process would have taken a lot longer so thank you!

Also, I want to thank Molly Bridges, my Springboard Student Advisor. You helped keep me accountable week to week and I really appreciate your support throughout these past few months.

Lastly, I want to give a shout out to Springboard; what you all are doing is nothing short of amazing. You’re opening up doors to so many people, including me, and will be eternally grateful for the opportunity that you’ve given me. I’m looking forward to learning more and seeing where this journey takes me. Thank you all!

***Sources***

* 1. <https://www.idc.com/getdoc.jsp?containerId=prUS44247418>
  2. <https://www.businessinsider.com/a-timeline-of-how-the-apple-watch-was-created-2015-3>
  3. <https://www.whoop.com/the-locker/>
  4. <https://www.kaggle.com/dansbecker/model-validation>
  5. <https://machinelearningmastery.com/k-fold-cross-validation/>
  6. <https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/>