AN ANALYSIS ON DATA FROM A WEARABLE TECHNOLOGY

Bernard Wong, Giorgio Quer

INTRODUCTION

Wearable technology has become a dominant industry in the past decade and as a result has changed both consumers and the health industry. With CCS Insight's prediction of 411 million smart wearable devices or almost $34 billion worth of technology being sold in 2020 there's no doubt that wearables will continue to grow into t he foreseeable future. This growth of technology is not only exciting and innovative, but it also bridges the gap between the health industry and real time data. With wearable technologies becoming more and more mainstream, medical professionals now have access to collect data from daily users on a massive scale. This data can then be used to analyze how different populations are doing, discover patterns hidden within diseases and illnesses, and assist health professionals in creating solutions for complex medical issues.

I had been given data collected by a new innovative technology filled with sensors that took comprehensive measurements involving heart conditions, sleep, and motion. With this data, I hoped to analyze the data and create methods so that this new wearable would have a good way to analyze these lifestyle factors to a greater scale.

UNDERSTANDING AND CLEANING THE DATA

The data given to me was one sample selected from a pool of 223 people. There were 3 datasets with varying amounts of information, but each had well over 40,000 measurements. The goal was simple; I wanted to take the data and create methods that prepared the datasets for future manipulation such as visualization and machine learning, visualize the data to better understand it and discover patterns, and prepare methods that could be applied to future larger datasets in a similar format. However, I first needed to understand and clean the data before doing any analysis, and because this was new data that the company collected, I explored and cleaned the data while also creating a dictionary explaining what each of the datasets measured.

IBI Dataset

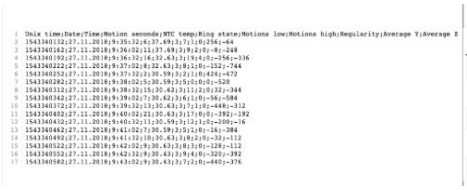
```
Date;Time;Validity;Padded IBI;IBI;UTC time
27.11.2018;12:37:43;0;1312;1312;1543351057
27.11.2018;12:37:43;1;980;980;1543351058
27.11.2018;12:37:43;1;956;956;1543351059
27.11.2018;12:37:43;1;1096;1096;1543351060
27.11.2018;12:37:43;1;1172;1172;1543351061
27.11.2018;12:37:43;1;1080;1080;1543351062
27.11.2018;12:37:50;1;1044;1044;1543351064
27.11.2018;12:37:50;1;1068;1068;1543351065
27.11.2018;12:37:50;1;1048;1048;1543351066
27.11.2018;12:37:50;1;996;996;1543351067
27.11.2018;12:37:50;1;984;984;1543351068
27.11.2018;12:37:50;1;1032;1032;1543351069
```

	Date	Time	Validity	Padded IBI	IBI	UTC time	UTC time (converted)	time frame	time according to UTC time
1	27.11.2018	12:37:43	1	980	980.0	1543351058	2018-11-27 12:37:38	afternoon	12:37:38
2	27.11.2018	12:37:43	1	956	956.0	1543351059	2018-11-27 12:37:39	afternoon	12:37:39
3	27.11.2018	12:37:43	1	1096	1096.0	1543351060	2018-11-27 12:37:40	afternoon	12:37:40
4	27.11.2018	12:37:43	1	1172	1172.0	1543351061	2018-11-27 12:37:41	afternoon	12:37:41
5	27.11.2018	12:37:43	1	1080	1080.0	1543351062	2018-11-27 12:37:42	afternoon	12:37:42
6	27.11.2018	12:37:50	1	1044	1044.0	1543351064	2018-11-27 12:37:44	afternoon	12:37:44
7	27.11.2018	12:37:50	1	1068	1068.0	1543351065	2018-11-27 12:37:45	afternoon	12:37:45
8	27.11.2018	12:37:50	1	1048	1048.0	1543351066	2018-11-27 12:37:46	afternoon	12:37:46
9	27.11.2018	12:37:50	1	996	996.0	1543351067	2018-11-27 12:37:47	afternoon	12:37:47
10	27.11.2018	12:37:50	1	984	984.0	1543351068	2018-11-27 12:37:48	afternoon	12:37:48

Interbeat interval, or IBI, is the time interval between individual beats of the heart. It's a useful measurement because it reflects how effective the brain's communication with the heart is. A healthy body will have many varying IBIs due to the constant changes between the sympathetic and parasympathetic nervous system, so when IBI doesn't vary this can be a warning sign to slowing neural communication or other neural issues.

The dataset that we had contained almost 500,000 measurements of IBIs that occurred from 11.27.2018 to 12.12.2018, with a measurement occurring nearly every second (non inclusive of the time the wearable device was off or wasn't running). The dataset contained information about when the IBI measurements were taken along with how valid the IBI measurement was. Overall, the dataset was relatively easy to clean and had no missing data. Because more than 95% of the data was clean, we only selected the most valid data and did an analysis on those.

Motion Dataset



	Unix time	Date	Time	Motion seconds	NTC temp	Ring state	Motions low	Motions high	Regularity	Average Y	Average Z
0	1543340132	27.11.2018	9:35:32	6	37.69	3	7	1	0	256.0	-64.0
1	1543340162	27.11.2018	9:36:02	11	37.69	3	9	2	0	-8.0	-248.0
2	1543340192	27.11.2018	9:36:32	16	32.63	3	19	4	0	-256.0	-336.0
3	1543340222	27.11.2018	9:37:02	8	32.63	3	8	1	0	-152.0	-744.0
4	1543340252	27.11.2018	9:37:32	2	30.59	3	2	1	0	424.0	-472.0

This dataset was the simplest of the three datasets. It contained information about the date, time, how long motion occurred in seconds, the NTC temperature (the temperature of the internal components), ring state, number of low and high motions per minute, and average y and z motion during the duration of measurement. Due to the interests of the project being primarily related to sleep and heart rate, not much further analysis was done with the motion dataset. However, a dictionary explaining the columns of this dataset should make it easier for analysis in the future.

Sleep Dataset

	Date	Bedtime start Unix	Bedtime end Unix	Bedtime start	Bedtime end	TimeZone	Debug info	Battery consumption	Is longest	Time in bed	...	1731	1732	1733	1734	1735	1736	1737
0	27.11.2018	1.543351e+09	1.543352e+09	12:35:53	12:49:53	-8.0	NaN	NaN	1.0	14.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	27.11.2018	1.543368e+09	1.543369e+09	17:17:55	17:37:55	-8.0	NaN	NaN	1.0	20.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	28.11.2018	1.543386e+09	1.543419e+09	22:16:15	7:32:15	-8.0	NaN	1.2%	1.0	556.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	28.11.2018	1.543438e+09	1.543440e+09	12:52:26	13:12:26	-8.0	NaN	NaN	0.0	20.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	29.11.2018	1.543463e+09	1.543464e+09	19:44:43	19:54:43	-8.0	NaN	NaN	1.0	10.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

The sleep dataset was by far the most complicated due to a number of factors. One reason was that it had the most information and as a result took some deeper exploring to understand. For example, there were columns such as 'is_longest' and 'Lowest HR time minutes' which could mean a variety of things. There were more than 500 different measurements about sleep, but after some communication with the company we were able to make sense of what most of the columns meant. Along with that, not each data point had the same amount of information; due to the organization of the columns of the dataset, there were a few regarding the sleep cycles that occurred during sleep. Because not each sleep measurement lasted the same duration and therefore didn't have the same number of sleep cycles, certain data points had measurements of sleep cycles while others didn't. The data had naively just recorded no data whenever there weren't sleep cycles and as a result formatting with the data was extremely poor. To resolve the issue, I replaced missing data with

placeholders representing nonexistent data, which enabled me to see full sleep cycles and make the data more readable.

Being such a big dataset, there were bound to be plenty of errors within the data. Time measurements were often slightly off, conditions were often erroneous, and sometimes software issues such as the wearable coming off or losing power would lead to errors in measuring sleep cycles and breathing rate. By sweeping the data and removing as many of these inaccurate data points, I was able to make the data a little more readable and easier to process later on. Even after the cleaning there were 1780 columns of data, making this dataset extremely informative.

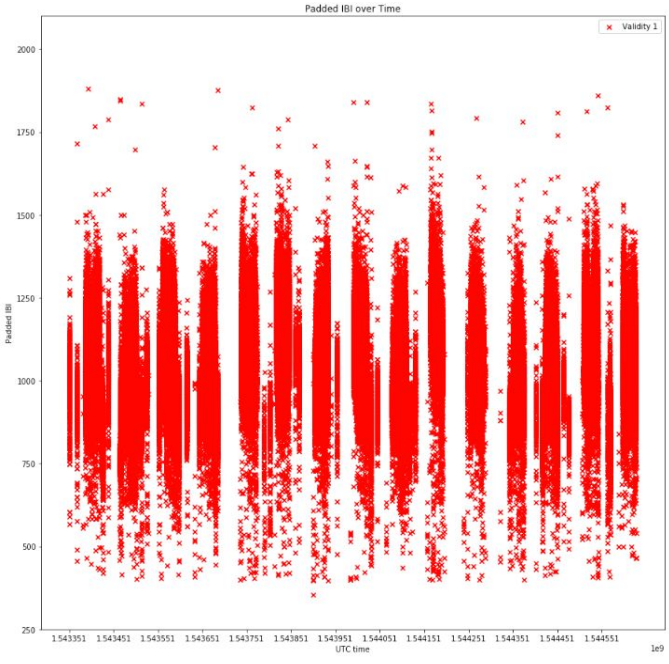
Overall, the process of understanding and cleaning the datasets was a long and tedious process. However, its importance cannot be stressed enough; through this organization, I was able to understand the data and get a better understanding on what information I was working with. Along with that, I was able to determine what data was truly useful and what wasn't, which made the analytics and visualizations a much easier process in the future. Although I gave a quick summary in this paper, much more work went into understanding the datasets, including writing a description and finding meaningful statistics for each section. If you're interested in seeing what additional work was done or to understand the dataset a little bit more, the dictionary can be accessed here.

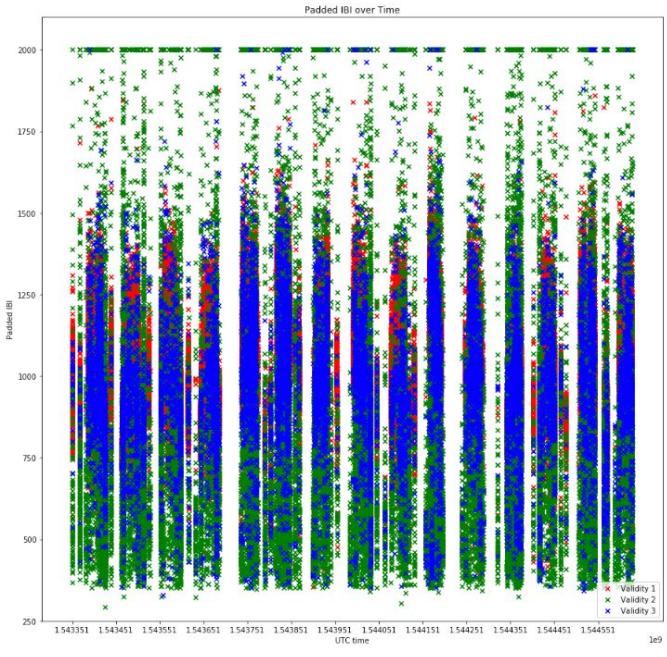
DATA ANALYSIS

Now that a majority of the data cleaning and organizing had been completed, we could now continue forward and begin analyzing and visualizing the data. The following are some of the visuals and statistics done for each dataset:

IBI Dataset

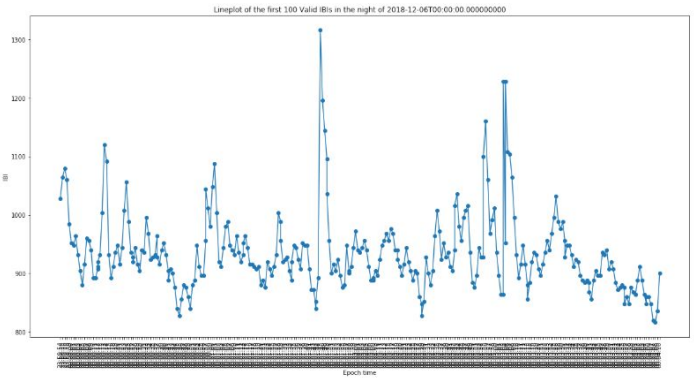
Because of my unfamiliarity with IBI, I wanted to get a basic understanding using a few visuals to help me. Further research and explanations from my mentor taught me that IBI tends to vary a lot in a healthy human, and can also vary depending on time of day, especially comparing waking and sleeping hours. To make this more apparent, I plotted IBIs of various validities over time.

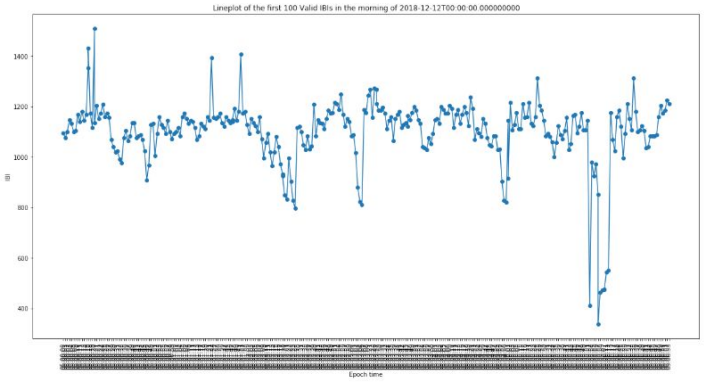




There is a lot that can be learned from these visuals. There are noticeable white gaps between certain time periods which helps us see when data isn't being recorded. Along with that, we can also start to see the differences between valid data. While IBI measurements with validity 1 stay in the range of 800 to 1700 milliseconds, IBI measurements of validity 2 and 3 are often at the extremes like 2000 milliseconds and are a lot more sparse. What's also interesting to see is how much IBI varies daily. Because the IBIs are greatly spread apart and create 'smears' of different IBIs, this signifies that the sample has greatly varying IBIs and as a result have a generally healthy neural communication network.

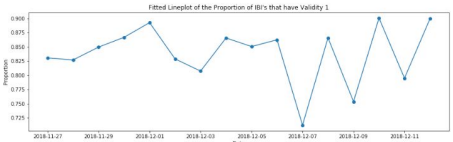
To explore the differences between waking and sleeping hours, I decided to plot IBIs during different times of the day. Using a method, I was able to plot the first 100 valid IBIs of a random time period during the morning, afternoon, evening, or night time.

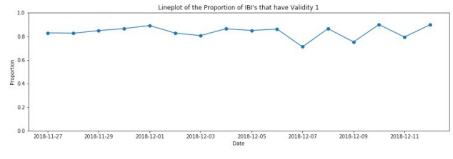




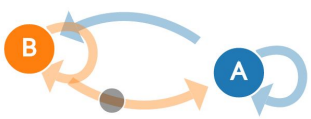
This is where the jumps in IBI become extremely apparent. As you can see in these line plots, there are consistently peaks that occur over time, most likely signifying jumps from the sympathetic and parasympathetic nervous system. You can also begin to see a slight difference between IBI variance at night and in the morning. While the night line plot has plenty of small spikes, the morning IBI stays relatively calm and stable. This can possibly be explained the lack of consciousness during the late night; during sleep, the brain can fully focus on moderating and controlling the body, while during waking hours the body has to focus on interpreting a lot more signals and controlling a lot more things.

Another part of the data that we spent a lot of time exploring was the validity of the data. We were interested in a few questions; 1) was the validity of IBI affected by the day?





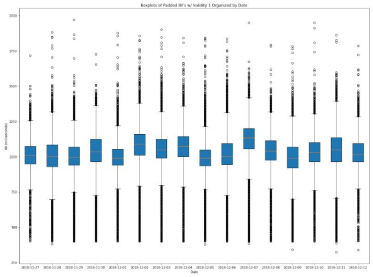
next initial	False	True
False	0.758826	0.241174
True	0.046645	0.953355

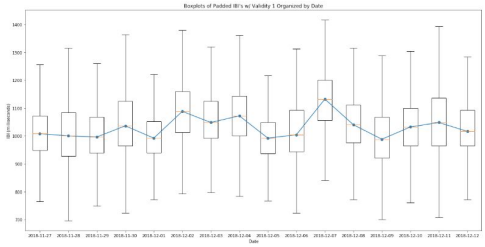


And 2) did valid IBI occur in chunks, or were they sporadic? Using both a fitted and regular line plot, we were able to plot the proportion of data that was the most valid (or had a validity of one) and discovered that almost every day more than 80% of the IBI measurements were valid. While the amount of data recorded varied a lot more during the month of December, a majority of the data was valid for each of the days recorded, meaning that it was fair to only do statistical work on just the most valid IBI data. To help us answer the second question, we

implemented a stochastic model known as Markov's chain. The markov chain is a system that models transitions from one 'state' to another according to certain probability rules; our task was to discover what these probability rules are. By finding out the probability of a valid IBI being followed by an invalid or valid IBI, and by finding out the probability of an invalid IBI being followed by an invalid or valid IBI, we can see if the valid IBIs are found together or sporadically spread apart. With our Markov's chain, we were able to discover that valid IBIs are typically found together; when a valid IBI is recorded, there is a 95% chance of the next IBI measurement being valid. This makes sense, especially considering the fact that invalid data is typically caused by sensor or technological error, which doesn't happen too often.

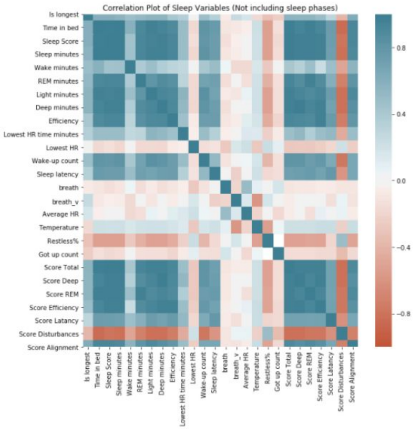
As a final overview of the IBI data, we decided to take a statistical approach and do some research on the outliers each day while also calculating some medical numbers involving heart rate variance (or HRV). By creating a few boxplots of the IBI each day, we can easily see the median of each day, see the outliers that occur, and also see how each day compares to the other. While outliers are normally not statistically great, they have a different meaning for IBIs. Because a greater variance in IBI signifies a stronger neural communication system, the great number of outliers is actually a great signifier for our sample's health. Along with that, we can start to see general trends of IBI depending on the date. These plots will be even more informative as they are applied to bigger datasets, as we can see if there are any seasonal patterns or any other potential factors that might affect IBI. The last analysis we did with the IBI data was come up with a few methods that helped calculate different numbers regarding HRV, which included RMSSD (the root mean square of successive RR interval differences), SDNN (the standard deviation of the IBI), NN50 (the number of NN intervals where the the interval difference is more than 50 milliseconds), and PNN50 (the proportion of the NN intervals where the interval difference is more than 50 milliseconds). Overall, we were able to create lots of useful visuals, and it was a unique experience learning about IBIs using the data collected by these new and unique sensors in the wearable.





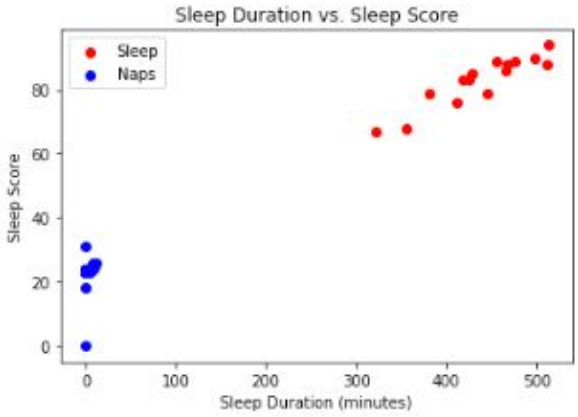
Sleep Dataset

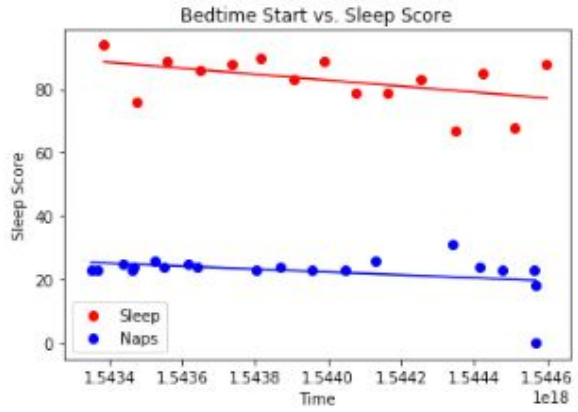
Because of the large amount of variables within the sleep dataset, I needed to determine which of the columns would give the most information or had the biggest impact on sleep. In order to narrow down my options, I created a correlation heat map that shows the positive and negative correlation between variables in the dataset. As can be seen by this correlation heatmap, there are a few items that have little to no correlation with none of the other variables such as the breath measurements. Overall, this heatmap is a great way to quickly glance and see what the relationship between two sleep measurements have on each other. I was primarily interested in what affected the sleep score,



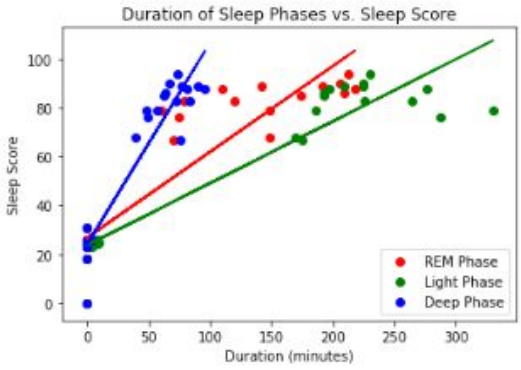
and to no surprise it seemed to be that variables regarding duration of sleep, whether it be sleep cycles or sleep as a whole, seemed to have the largest positive correlation with the sleep score.

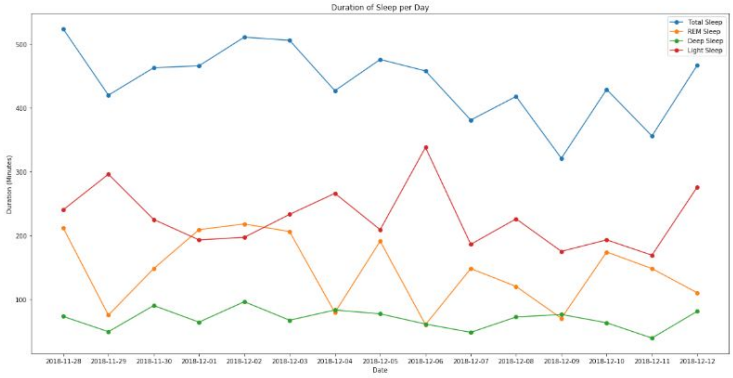
Because the duration of sleep played such a large role in the sleep score, I decided to create a few plots comparing the sleep duration and the bedtime start with the sleep score. By dividing up the sleep data points into two groups, a 'nap' group where sleep was less than 30 minutes and a 'sleep' group which was any sleep that was more than 30 minutes, I hoped to discover some interesting patterns. What was interesting to see was that the plots for Sleep Duration vs. Sleep Score and Bedtime Start vs. Sleep Score both showed stark differences for the group. Overall, the nap group had significantly smaller scores which was honestly expected. Along with that, naps that lasted longer had a significantly greater increase in score compared to other shorter naps, while the score change was less significant for sleeps. An interesting thing to note was the similarities between the groups. No matter the sleep or nap group, the later the bedtime start was, the worse the score tended to be, which supports previous studies regarding sleep time.





Another interesting thing that we had decided to plot was sleep score and its effect on sleep in general. We were able to plot two different aspects regarding sleep score: the duration of sleep phases and its effect on sleep score and the amount of each sleep phase our subject had per day.





There are a lot of great takeaways from these graphs. The Duration of Sleep Phases vs. Sleep Score helped determine what were some of the more important sleep phases, and it's quite clear that more deep sleep and REM sleep leads to a more significant increase in sleep score, especially when compared to light sleep. The durations of each sleep cycle for each day help us

find days where certain sleep phases are lacking and give us a better idea on how much of sleep phase is happening throughout our subject's life. This information might help future exploration, as you might be able to find outlier dates where certain sleep cycles are different and use it to see its effects on IBI, motion, and daily life.

CONCLUSION

Working with this data was truly a great and unique experience, but there are still many more things that can be done with it. While I had begun to string IBI and sleep together and work on determining the effects each had on one another, I had been limited in time and had not been able to explore the two datasets together as much as they could be. Along with that, little was done with the motion dataset, leading to much more exploration. It would also be great to use these cleaning methods that we've created to help implement certain machine learning algorithms that could potentially predict swings in IBI, IBIs effects on sleep, and other potential events. Most importantly, however, is trying to use these methods on a larger population to find more generalized trends. Because this exploration only involved one sample, we were able to learn a lot about the organization of the data and create visuals describing this individual's two week period. However, the true potential comes when these methods are used on a large sample, which can reveal a plethora of different trends that occur with heartbeats and sleep.

While there is still much more exploration that can be done, there are a few things that are certain. Big data and wearable technology is changing society on a very large scale, and is not only allowing medical professionals to analyze and solve problems that they couldn't solve before, but is also educating the general public about personal health and assisting in personal growth. With such a short amount of time, we were able to create beautiful visuals, get a better understanding about IBI and sleep, and create methods that can be used for wide scale data to better understand a larger population. This is simply the tip of the iceberg and as sensor prices decrease, big data plays a more prominent role in the health industry, and wearables become more and more frequent in our everyday lives, the positive impact and medical research will continue to grow with it.