Robert Swanson

Dr. Maxwell

MAT 311

18 May 2020

Data Science Final Project Writeup

My goal was to better understand what factors influence my quality of sleep, specifically my sleeping heart rate. I while at school the some nights my heart rate averaged up above 100 and other nights it averaged around the low 40s. I was curios to see what caused those numbers, as well as visualize when I experienced those high heart rate nights. As I suspected, those nights mostly took place during stressful periods of life, so I wondered what real world data might correlate, such as my fitness data that I already had access to. I was also curios to check on the reliability of the data my devices were giving me. As it turned out, one metric, the resting calories burned turned out to be a historically problematic one, as it was not agreeing with other sensors. Even now, people aren't really sure how it's calculated since it shows variation, yet such a metric is typically based off of static measures such as weight, and height.

One of the best parts of this exercise was the fact that this data was my own, describing my own history. This both allowed me to verify the reasonability of the results given my background knowledge of the data itself, and allowed me to remember/discover new things about myself and my habits (such as Monday being my most exercised day on average). This New York Times article discusses another example of analyzing personal data, though in this case that data is derived from email.

My solution was to use an app to export this data as a .csv file and clean it up in RStudio. from there I grouped them by datetime, by date, by time, by weekday, and the combination of time and weekday. I then had a lot of fun visualizing the data in different ways, including using google charts by way of the googleVis library. I then ran a bunch of different correlation checks between different variables and chose variables to be a part of the end formula. I ultimately used both a linear regression model and a decision tree, testing each on a separate set of testing data.