Data 115 Final Project

Jason Cross

12/12/2023

Introduction

For my data research project, I explored my own personal running data, analyzing relationships between cadence, speed, heart rate, and relative effort, hoping I could then use this data to identify an optimal personal cadence to enhance overall running efficiency and economy, maximizing speed while minimizing effort. In this report, I will detail my data gathering process, cleaning steps, exploratory data analysis, and insights gained from linear regression modeling.

Cadence Explained

In running, cadence is measured in steps per minute (SPM). It's possible for runners to adjust their cadence independently of speed (longer and fewer strides vs. a greater number of shorter strides) and for every runner there is a theoretical "sweet spot"–an optimal cadence that strikes the perfect balance between step frequency and stride length. A generalized approximation for optimal running cadence is believed to be somewhere around 180 SPM. However, individual optimal cadence varies depending on body type, running style, body mechanics, and personal comfort.

Motivation and Key Questions

My motivation for this analysis, simply put, is to discover my own personal "sweet spot" as it pertains to cadence. Here are the key questions that drove my research and analysis:

1. What is my average running cadence, and how consistent is it across activities?
2. How close is my average cadence to the generic 180 SPM recommendation?
3. Can I identify an optimal personal running cadence based on relationships between cadence, speed, heart rate, and relative effort data?

Data Gathering

I've maintained a daily running streak for over 3 years, logging activities with a Garmin GPS watch and an app called Strava. I was able to download a comprehensive .csv file from Strava of every activity I'd ever logged, uncovering 89 data categories, including a plethora of extraneous yet intriguing variables like elevation, wind speed, cloud cover, humidity, and even moon phase.

Data Cleaning

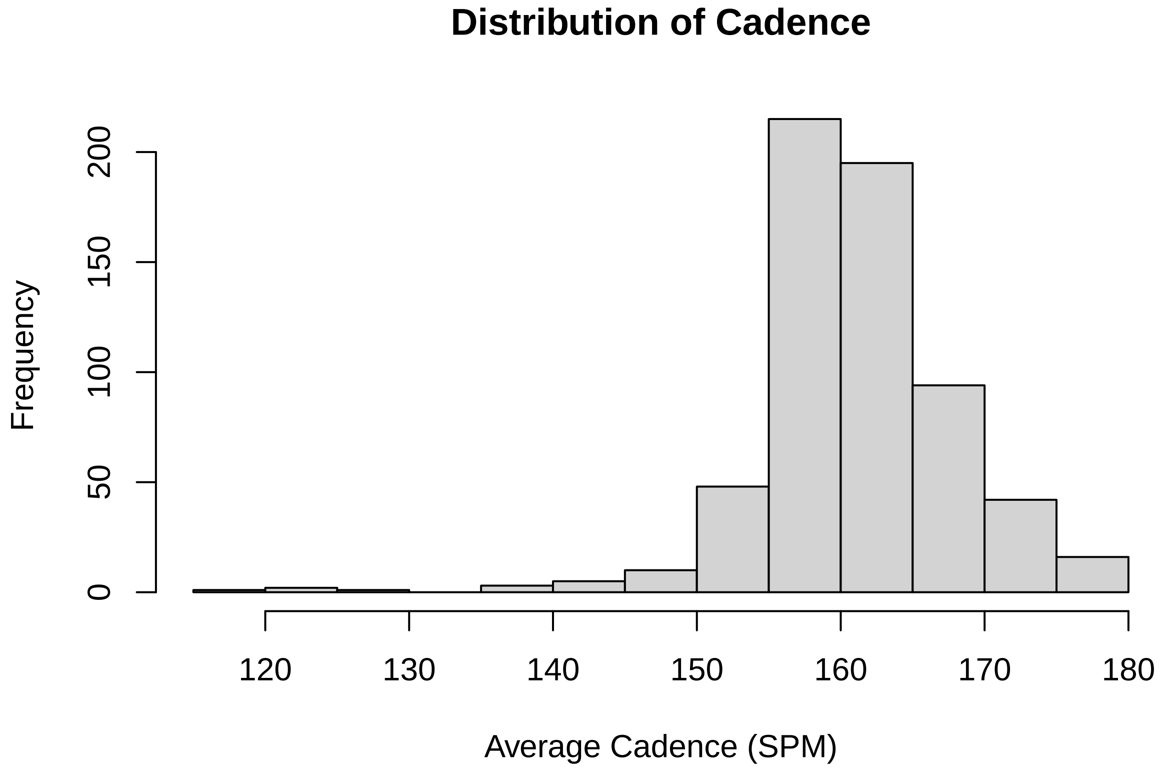
During the data cleaning process, several steps were taken to ensure the integrity and relevance of the dataset. Empty columns were removed to streamline the data structure. Then, formatting uniformity was checked, and although the formatting was consistent, certain cell formats were adjusted for improved readability and handling. Specifically, a single date/time column was split into two separate columns, and cadence values, originally in RPM (revolutions per minute), were converted to SPM (steps per minute).

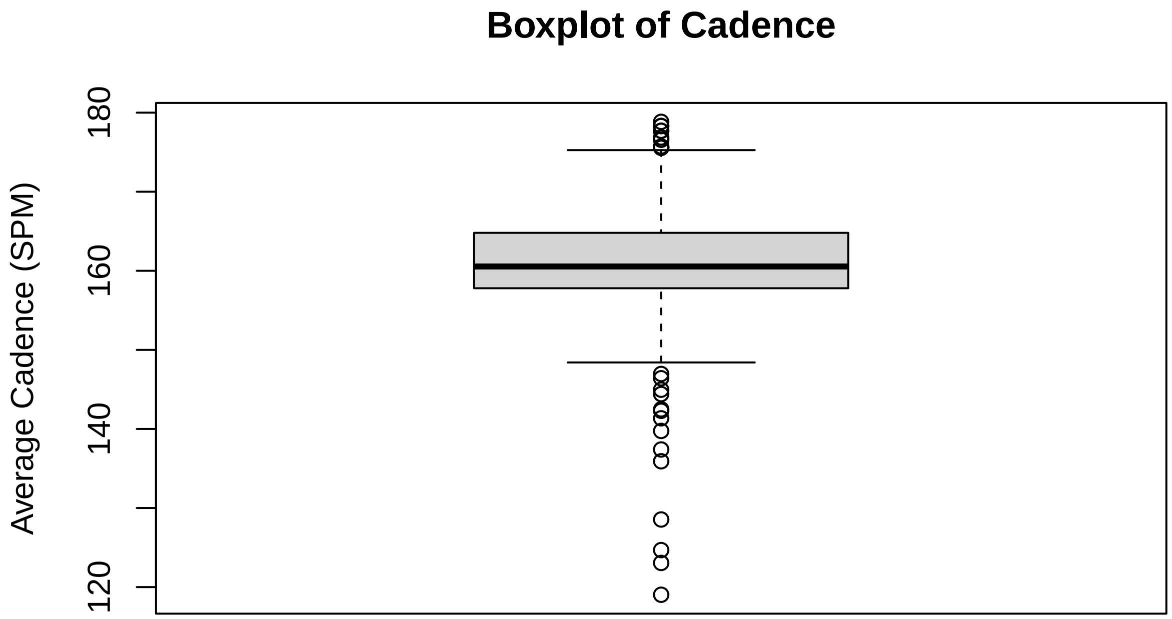
Unwanted data was then systematically excluded. Dates outside the range of interest (prior to 10/01/2020 and after 09/30/2023) were removed. Activities such as walking, cycling, yoga, workouts, and weight training were also eliminated. Duplicate columns, particularly those related to heart rate and relative effort, were identified and removed during this phase, resulting in the removal of 52 columns or categories.

To enhance the organization of the dataset, the column order was rearranged, prioritizing columns of highest interest on the far left. Finally, rows with missing cadence, heart rate, and/or relative effort data were removed from the dataset, and my .csv file was ready for further analysis in RStudio.

Exploratory Data Analysis

In order to answer my first two preliminary questions regarding my average overall cadence, my consistency of cadence from run to run, and how it differs from the general 180 SPM recommendation, I calculated the mean, standard deviation, and variance of my cadence and plotted the following histogram and boxplot for visual analysis.



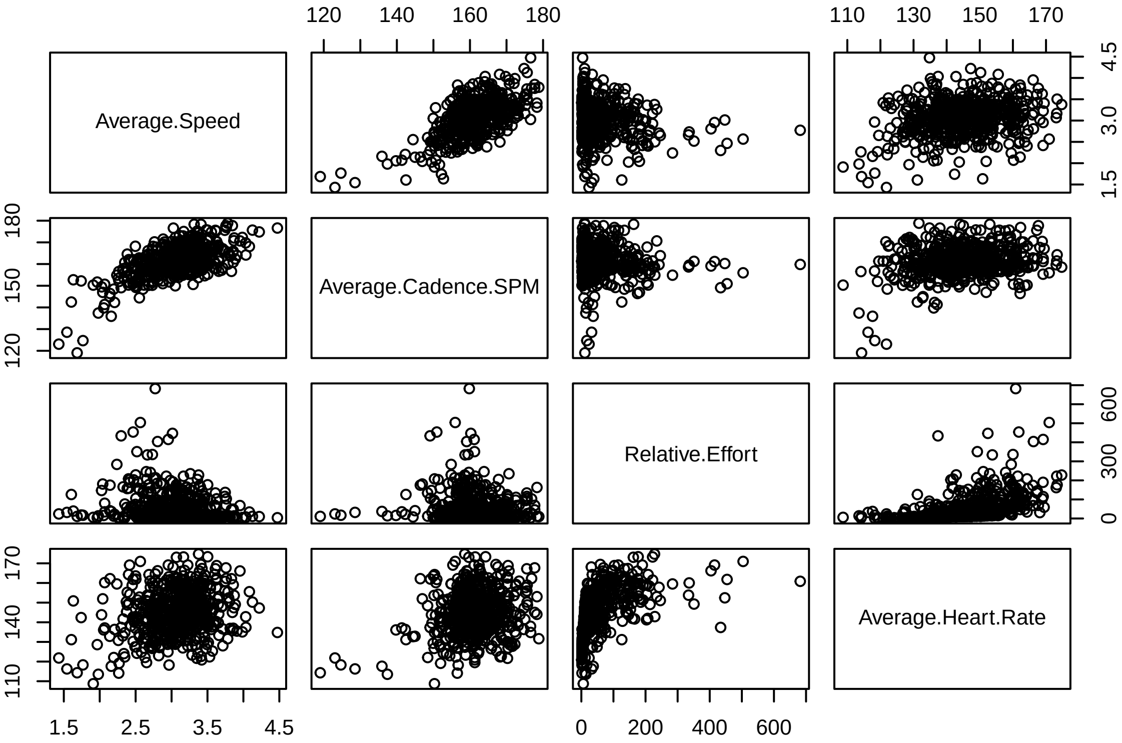


Mean Cadence: 161.0362 ~ Standard Deviation: 6.799002 ~ Variance: 46.22643

Based on these descriptive statistics and accompanying visualizations, my average overall cadence–while fairly consistent from run to run with relatively low spread and variance–is quite a bit lower than the general 180 SPM recommendation at roughly 161 SPM. But is my typical cadence an optimal one for me, personally? How close is 161 SPM to my "sweet spot"? I explored the relationships between cadence, speed, heart rate and relative effort in hopes of finding out.

Scatterplot and Correlation Matrices

Next, I created scatterplot and correlation matrices to explore possible relationships between each of these four variables, respectively. Unfortunately, the correlations between cadence and effort, as well as those between cadence and heart rate, were fairly weak (as you'll see in the summaries below).

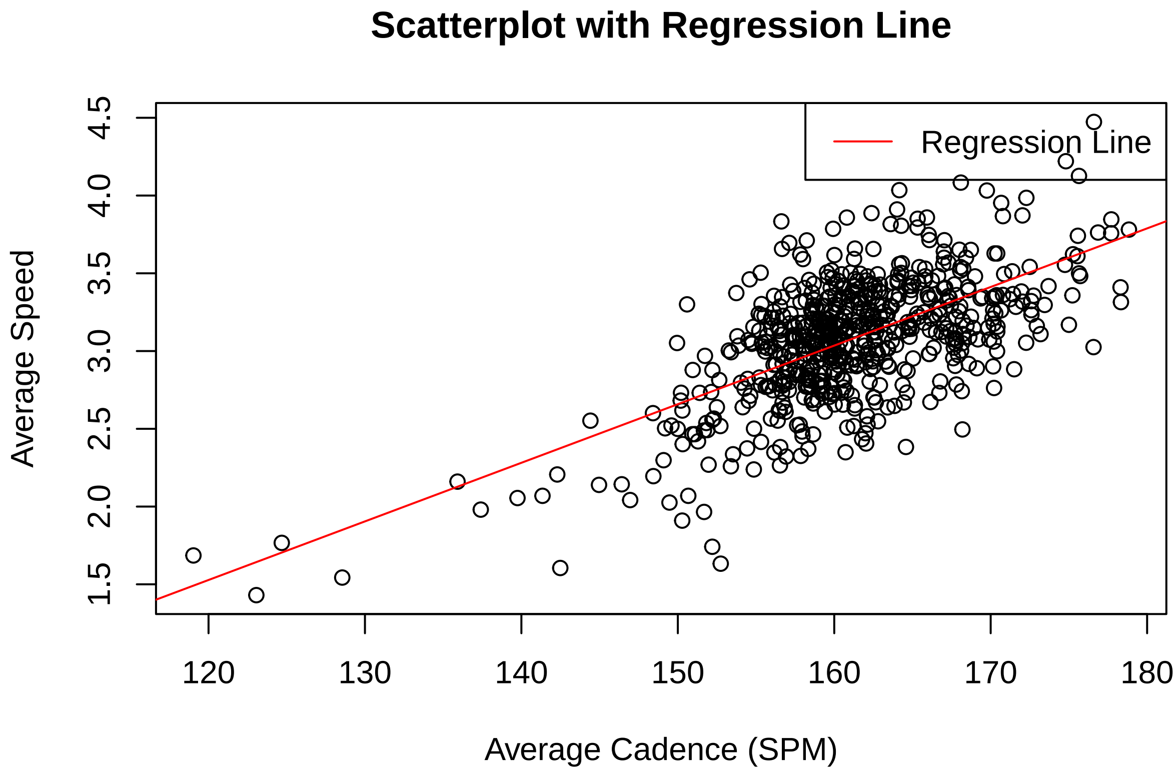


```
##                     Average.Speed Average.Cadence.SPM Relative.Effort
## Average.Speed           1.0000000          0.63891908     -0.23098682
## Average.Cadence.SPM     0.6389191          1.00000000     -0.07768186
## Relative.Effort        -0.2309868         -0.07768186      1.00000000
## Average.Heart.Rate      0.2428202          0.22516477      0.52545003
##                     Average.Heart.Rate
## Average.Speed                0.2428202
## Average.Cadence.SPM          0.2251648
## Relative.Effort              0.5254500
## Average.Heart.Rate           1.0000000
```

Linear Regression Model for Cadence vs Speed

Neither the scatterplot matrix, nor the correlation matrix above, suggest a strong relationship between cadence and effort, unfortunately. Cadence and speed, on the other hand, shows a strong linear trend and a relatively high positive correlation at roughly 0.64. I decided to explore this relationship further with a linear regression model. (Note, the correlation between heart rate and relative effort is also fairly strong, though this isn’t particularly surprising or illuminating given that Strava’s relative effort calculations are based largely on given heart rate data. Therefore, I decided to ignore the heart rate vs. effort correlation.)

```
##
## Call:
## lm(formula = Average.Speed ~ Average.Cadence.SPM, data = selected_columns)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12888 -0.19354  0.01648  0.20519  0.92601
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.994562   0.291373  -10.28   <2e-16 ***
## Average.Cadence.SPM  0.037686   0.001808   20.85   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3087 on 630 degrees of freedom
## Multiple R-squared:  0.4082, Adjusted R-squared:  0.4073
## F-statistic: 434.6 on 1 and 630 DF,  p-value: < 2.2e-16
```



Predicting Speed Percentage Increase for Average Cadence of 180

Using the above linear regression model, I made a prediction about what my average speed would be if I were to increase my 161 SPM average cadence to the general recommendation of 180 SPM. The model was trained on my dataset using the lm() function and, following the model fitting, I predicted an average speed of 3.79 m/s–a 23.25% increase from my current overall average speed of 3.07 m/s.

Conclusions, Lessons, and Next Steps

The data revealed a strong positive correlation between cadence and speed, suggesting that an increase in one generally corresponds to an increase in the other, and vice versa–an outcome that was more or less expected. Sadly, the results of my intended analysis regarding optimal cadence were mostly inconclusive. However, with my cadence vs. speed linear regression model and predictive analysis, it will be interesting to test the prediction by running at exactly 180 SPM (which can be achieved by running to the beat of a song with a tempo of 180 BPM) and seeing how close my actual speed is to the predicted 3.79 m/s.

In hindsight, I believe the lack of variance in my dataset with regards to average cadence made it difficult to accurately or meaningfully interpret possible relationships between cadence and effort. In order to explore this further, I believe I would need to collect running data with more variance and even distribution across a range of cadences, so that the effects of each SPM value (with respect to effort) are more adequately represented.

I am strongly considering conducting this experiment over the coming year, aiming to collect running data with a more diverse and evenly distributed range of cadences