Ryan Fischbach
Dr. Khuri
CSC373
9/11/2020

Assignment 1: Exploratory Data Analysis

Main Figures and Summary

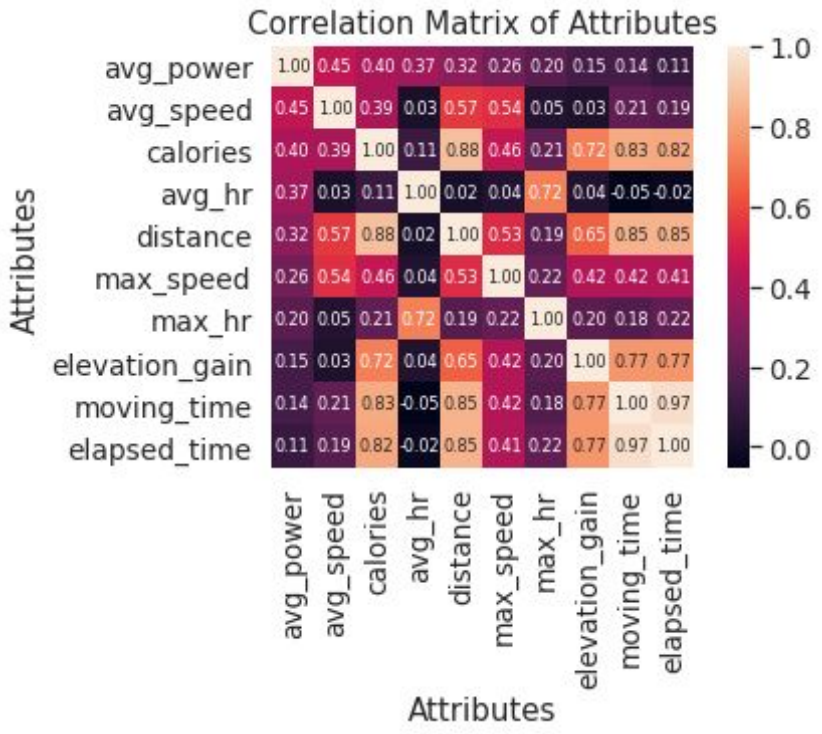

Figure 1: Correlation Matrix of the 10 Attributes with the Highest (Absolute) Correlation to Average Power

A correlation matrix of the attributes highlights some interesting patterns. With the final goal of this data being predicting the average power of a cyclist using other features, attributes with a low correlation to “avg_power” do not provide adequate predictive power. Specifically, both time attributes and the elevation gain add very little information for predicting average power. Other attributes do not provide a high correlation to average power either, suggesting that no attributes featured in this dataset have much predictive power. This would suggest that average power is some sort of calculation or combination of other attributes rather than a single attribute providing predictive power.

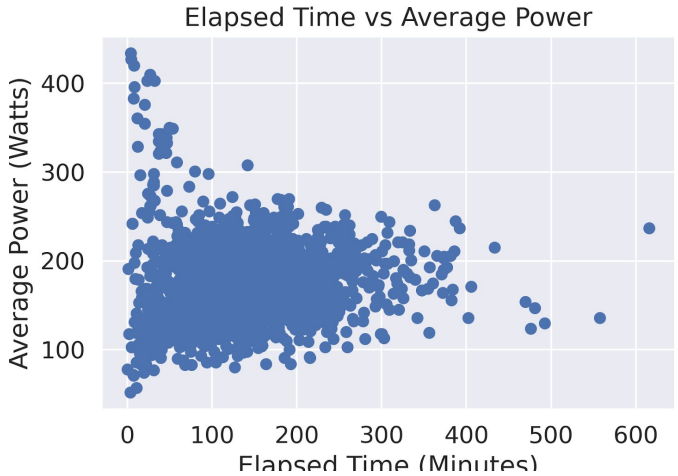

Figure 2: Average Power On Elapsed Time. Average Power is in Watts and Elevation Gain is in Minutes.

This chart suggests that cyclists who ride for longer durations are not “experts” and do not have a higher average power. Rather, it could suggest that there is no difference in the duration of rides between experts who exhibit a higher average power vs beginners who exhibit a lower average power.

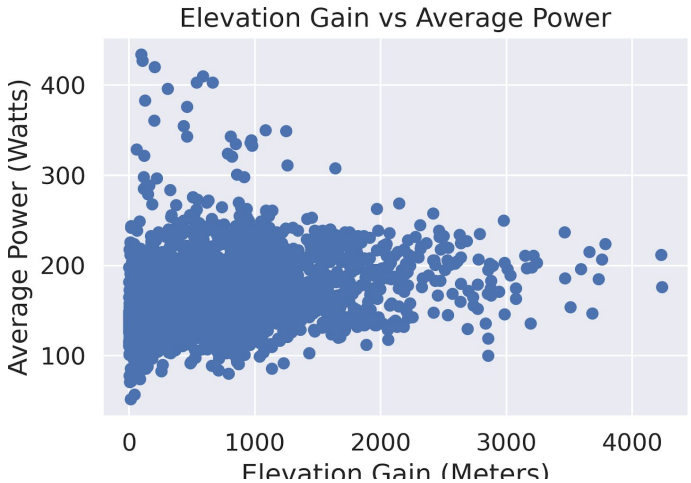

Figure 3: Average Power on Elevation Gain. Average Power is represented in Watts and Elevation Gain is in Meters.

This figure highlights a similar trend, that the amount of elevation gained by a cyclist does not impact their average power output. This could be circumstantial, suggesting that those with access to hills and elevation simply use it while training while others who don’t have access don’t.

Both figures and the correlation of attributes point to an opportunity for clustering to determine who is a “high level” rider and who is a “low level” rider depending on power output. Clustering could potentially identify unseen trends between riders based on their attributes. Based on the resulting clusters, this could provide actionable insights for new riders on what to focus on to increase their “level”.

Data was collected from a limited number of cyclists, so a subset of data from one cyclist might provide more insight into what impacts average power for that rider.

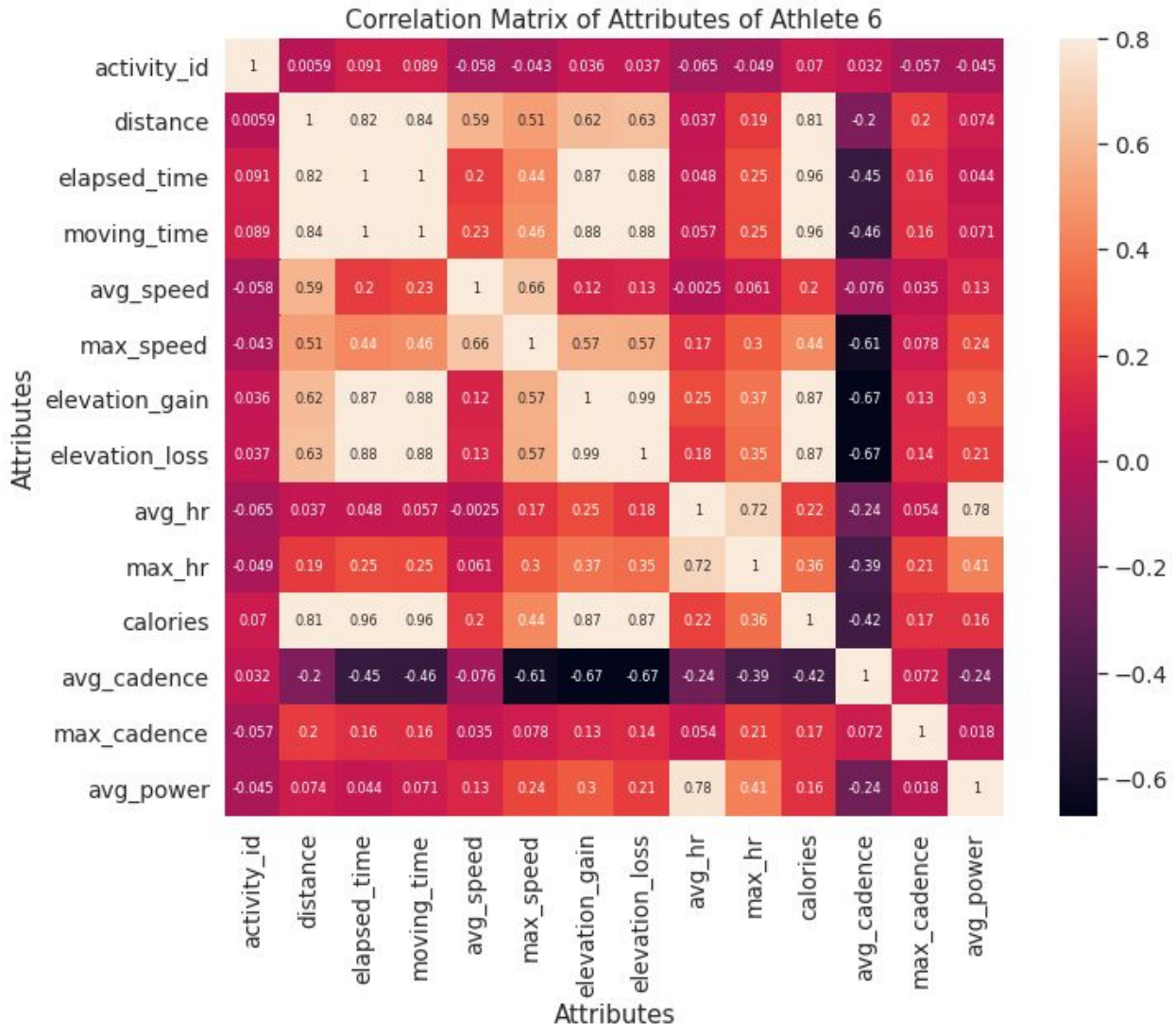

Figure 4: Correlation Matrix of Attributes From Athlete 6

Looking specifically at athlete 6 (the athlete with the lowest average power among rides), there are higher correlations between certain attributes and average power. Specifically, average cadence, average speed, and average heart rate have correlations of 0.62, 0.63, and 0.66

respectively with average power. These numbers would suggest that for this specific rider, these attributes have predictive power towards their power output.
Athlete 6's average power was 140.69 across all rides, putting him below the average for the entire dataset. Choosing a subset of rides from Athlete 2 (with the highest average power of 209.32) might provide differences in correlation between attributes and average power.

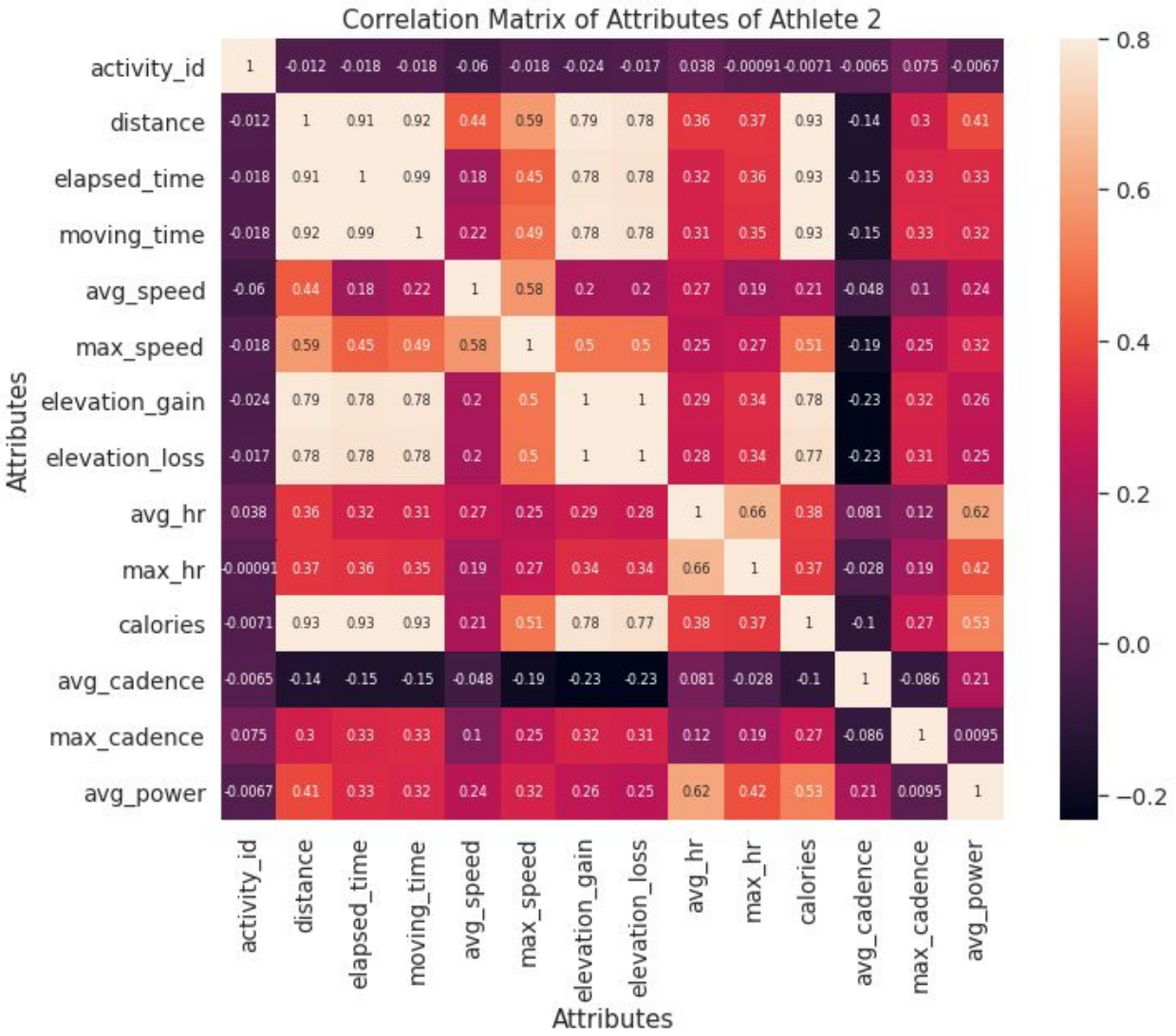

Figure 5: Correlation Matrix of Attributes From Athlete 2

Athlete 2 has different correlations between attributes and average power. Average cadence and average speed are no longer as good of predictors of average power relative to athlete 6 for a ride. However, the average heart rate maintains some predictive power for average power for a ride. It is clear that this more "advanced" rider has different attributes towards their average power. This raises the question, what attributes or combinations of attributes can best predict average power? Additionally, what differences exist in predicting average power between higher-level and lower-level riders (as defined by power)?

Citations

[1] Esteban Murillo Burford, Sarah Parsons, and Natalia Khuri. 2020. Data-Driven Prediction of Cycling Performance. 1, 1 (July 2020), 27 pages. https://doi.org/10.1145/1122445.1122456
[2] Pedro Marcelino. 2017. Comprehensive Data Exploration With Python. https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python/notebook
[3] Numpy. 2020. https://numpy.org
[4] Pandas. 2020. https://pandas.pydata.org
[5] Matplotlib. 2020. https://matplotlib.org
[6] Seaborn. 2020. https://seaborn.pydata.org