CMPT 353 D1: Computational Data Science

Term Project

13/08/2021

|  |  |  |  |
| --- | --- | --- | --- |
| We Love Cycling | | | |
|  | Daniel Wang | 301413271 |  |
|  | Fitzpatrick Laddaran | 301282987 |  |
|  | Hong Cung Quang | 301417603 |  |
|  |  |  |  |
|  | Instructor: | Greg Baker |  |
|  | Teaching Assistant: | Ali Arab |  |
|  | Teaching Assistant: | Ghazal Saheb Jam |  |

# Table of Contents

Project Overview …………………………………………………………………………………3

Gathering the Data ………………………………………………………………………………..4

Cleaning the Data …………………………………………………………………………………4

Analysis and Findings …..………………………………………………………………………...5

Conclusions ……………………………………………………………………………………….8

Limitations and Improvements .....………………………………………………………………..8

Accomplishment Statements ……………………………………………………………………...9

# Project Overview

Inspired by an avid cyclist within our group, the idea behind this project was to determine whether the variation in a cyclist’s power exertion affects one’s performance throughout a cycling ride. A cyclist’s power exertion is measured by the amount of power exerted through the bike’s pedals. This data is recorded via a strain gauge located on the cyclists’ bicycle. Typically, this is located on one or both crank arms, but can also be in the crank spindle, crank spider, or the pedal itself. Power variation is calculated with this formula:

where is the variability index, is the normalized power, and is the average power. is calculated as follows:

1. Calculate rolling averages throughout the entire cycling ride.
2. Calculate the fourth power of the rolling averages.
3. Calculate the average of the values in the previous step.
4. Calculate the fourth root of the value in the previous step.

This calculation is based on Aart Goossens’—a data engineer—description of normalized power, which ultimately was adapted from the *Training and Racing with a Power Meter* by Hunter and Allen[[1]](#footnote-1), [[2]](#footnote-2). Effectively, measures how smooth a cyclist’s power output is throughout a ride.

Additionally, an elevation scoring index was calculated with the following formula:

where is the elevation score, is the total elevation gain over the course of the ride, and is the total distance of the ride. We calculated to find correlations between power variation and different elevation profiles. This is important because we believe that elevation profiles should affect one’s overall performance, that is: a hillier ride is likely to be a harder ride and hence one’s performance should be weaker in comparison to a flat ride of the same distance.

In the following pages, we first discuss the data gathering and cleaning process. We then provide an overview of the type of analyses we have conducted, and then proceed with the conclusions. Additionally, we list some of the limitations that we have encountered while completing this project.

# Gathering the Data

Golden Cheetah is an open-source analytics tool that cyclists can use to analyze the data recorded from their rides. Riders can opt in to share anonymized data to the GoldenCheetah OpenData project. Our data was gathered using their API.

Cleaning the Data

The original uncleaned dataset contains many anomalies, which could have originated from various sources. Such sources include:

* sensor malfunction and inconsistency,
* poor cycling routes that contained many stops, and
* unpredictable cycling behaviours.

Therefore, the data was cleaned under these conditions:

* removed rides with incomplete altitude, distance, or power data,
* removed records where idling periods were 150 seconds or greater,
* removed rides with average speeds less than 10 and over 60 kilometers/hour, and
* removed rides that were less than 5 km and over 300 km.

These conditions eliminated situations where:

* significant fluctuations in measured values,
* extreme measured values,
* a rest stop, and
* prolonged downhill riding was observed.

Evidently, the final dataset captured the entire ride while avoiding segments of data that did not help with our research. The cyclist’s performance is based on segments of data in which the cyclist is consistently pedalling, or small segments of data in which the cyclist is not pedalling but is consistently moving. Therefore, we keep records where the cyclist may be cruising or on a downhill trajectory.

The final dataset contained 28 unique cyclists, totalling to 4116 rides. This averages out to 147 rides per cyclist.

## Analysis and Findings

Given the large dataset, conducting our analyses required us to partition the dataset into an appropriate format. Using Spark, the 28 cyclists were put into small groups of three or four. Additionally, a unique identifier was provided for every record in every cyclist’s ride. This allowed us to keep track of which record belonged to which cyclist. For every group of cyclists, all the data for the cyclists in that group were then put into one data frame. Finally, the calculation for and were done on this data frame. Other statistics of interest were also calculated. These statistics include the total time and distance of the ride, the average power exerted by the cyclist throughout the ride, and the average altitude of the course.

To illustrate, here is an image of a data frame as described above:

Table

Description automatically generated

Figure 1: Data frame after and calculations.

As seen on Figure 1, every data frame will consist of multiple cyclists. This depends on the number of cyclists partitioned into a specific group. To combat any data misinterpretation, a combination of the cyclist’s alias, *cyclist\_id*, and one of their rides, *file\_name*, has been used as a unique identifier. This means that for every cyclist’s ride, we now have some statistics calculated.

At this point, our dataset contained some outliers. This issue originateed from our initial cleaning. The conditions that we have used did not eliminate all the potential anomalies in our data; hence, we eliminated records in which the was above two, and where the average altitude was greater than 2500 meters. We set the first condition because the value of is generally between one and two, where two is on the extreme side for evaluating variance. The second condition was applied because very few rides averaged greater than 2500 meters.

Beginning our analysis, we created a pair plot to determine if any relationships exist between the variables. These are the results:

Diagram

Description automatically generated with medium confidence

Figure 2: Pair plot for all variable pairings.

Figure 2 does not seem to give a good visualization on the relationships between the variables in our data. The huge number of records seem to obstruct the actual relationships if any exist; however, we can still obtain some general information from the plots. As such, there seems to be a linear correlation between the total time it takes to complete a course, *total\_time\_sec*, and the total length of the course, *total\_dist\_km*. Intuitively, this should be true. Looking at the column, it seems that there is a vertical trend between all other variables. This indicates that may not be a factor that affects other variables.

To solidify our general findings, we also performed a parametric and post-hoc test, namely: the ANOVA and Tukey’s Honest Significant Difference (HSD) test. We conducted a parametric test because we have a sufficiently large number of recorded rides, ; however, the lack of normality in the distributions seen in Figure 2 indicates that ANOVA may not be the best statistical tool. Therefore, we considered a non-parametric test, the Kruskal-Wallis test and its corresponding post-hoc test, the Dunn test. The following are the results:

Graphical user interface, text

Description automatically generated

Figure 3: Results from parametric and non-parametric tests.

From Figure 3, the ANOVA test produced some significant results. The first main interest is the p-value obtained from ANOVA, which is . With Tukey’s HSD test, the interesting results indicated that the pairing for and is inconclusive. Same goes for the pairing of with *total\_dist\_km*, and with *total\_dist\_km.*

With the Kruskal-Wallis test, we also have p-value < , so we conducted the Dunn test to determine which pairings were significant. The test tells us that every pairing is significant.

Chart, scatter chart

Description automatically generatedFinally, to answer our original question, we created a density scatter plot of against performance, measured as the average velocity.

Figure 4: Density scatter plot of against performance.

From Figure 4, the density scatter plot shows us that about 60 percent of our rides are on the red-colored surface. This indicates that cyclists who have a of about 1.0 to 1.2 are on the scale for average performance. As we see, for values of that are closer to 2, cyclists also have poorer performance.

Conclusions

Figure 4 partially answers our original prompt. Cyclists with a lower seem to perform better in comparison to those with a higher ; however, it is slightly difficult to determine whether there really does exist a trend due to large number of observations. Due to the difficulties in visualizing as part of the density scatter plot, our group was not able to conclude whether has a correlation with and performance.

We have created Figure 2 to formalize some trends. This way, we could identify patterns that are not intuitive. As discussed, we see very little patterns.

Finally, Figure 3 aided us by providing potential future improvements to this project. We can conclude which pairings of variables are significant; hence, creating models for a variable should somehow depend on these other significant variables.

Conclusively, our group was satisfied with the project that we have completed.

Limitations and Improvements

In general, the limiting factor that hindered our progress was time. Given everyone’s busy schedules, it was hard to coordinate meetings and to motivate progress. If we had more time, we would have done the following:

* research and improve the formulation for ,
* research and deduce other metrics for evaluating performance,
* create a machine learning model that could predict performance based on significant variables,
* improve Figure 2 by creating subsets of the records,
* improve Figure 3 by adding more variables for comparison such as performance,
* improve Figure 4 such that it incorporates the as the colored density,
* increase and pick selective sample sizes to ensure representability, and
* refine our data cleaning methods such that it would mitigate anomalies and improve data quality.

In retrospect however, our group should have been more objective-oriented. Instead of leaving things to the last-minute, we should have started earlier and enforced deadlines for various milestones.

# Accomplishment Statement

*Daniel’s Statement*

Acquired and cleaned dataset from GoldenCheetah OpenData API. Researched and implemented methods in calculating *elevation variability*.

*Fitz’s Statement*

Researched and implemented methods in calculating *power variability* in cycling performance using standard Python libraries such as pandas. Researched and advised on methods for calculating *elevation variability* in different geographical terrains.Conducted non-parametric statistical tests and post-hoc analysis to deliver various conclusions.

*Hong’s Statement*

Developed the data pipeline and cleaned the data. Created and debugged strava\_write.py, strava\_analy.py, strava\_var\_index.py. Implemented statistical analysis.

1. See <https://medium.com/critical-powers/formulas-from-training-and-racing-with-a-power-meter-2a295c661b46>. [↑](#footnote-ref-1)
2. See <https://www.trainingpeaks.com/blog/power-terminology-for-cycling/>. [↑](#footnote-ref-2)