# Analysis of data from Strava

Yiliang Zhang

This document describes what I did to analyze data from Strava and is divided in to 3 parts: First is how I get these data; second is data analysis and conclusion; last part is summary and discussion about what can be done next for further research.

## 1. DATA SCRAPING

In strava's website, I find most of the data are presented by graphs and maps which is hard to analysis directly. What I want to get are a quite number of data from a wide range of athletes so that these data can be regarded as being randomly chosen from population. After glancing over most of the pages in strava, I found that the useful data are type (mainly run and ride), distance and time of exercise. These three kinds of data are what I am going to scrape next

Unfortunately I did not find any other ready-made data that contains all users' activity information (I mean data containing type, distance and time and can be downloaded directly). Then I find that in some of the users' homepages, there is a link called "training log" in which there are data about everyday exercise in details. These data include type, distance as well as time of exercise that really fit my demand. I want to get data there but it is hard to know which athlete have training log while others do not, and it is impossible to get data from these pages one by one manually. Therefore, I need to write my own code to scrape data.

I find that when you enter the link “training log”, the URL becomes the form of “https://www.strava.com/activities/”+ “id”. Each id number uniquely corresponds to an activity and neighboring ids do not belongs to the same athletes’ activities. Thus I just need to loop ids to get different activity page and use crawler to scrape each page’s data. Actually I haven’t had any similar experience of data scraping before so it took me a lot of time to learn how to use Python to get these data.

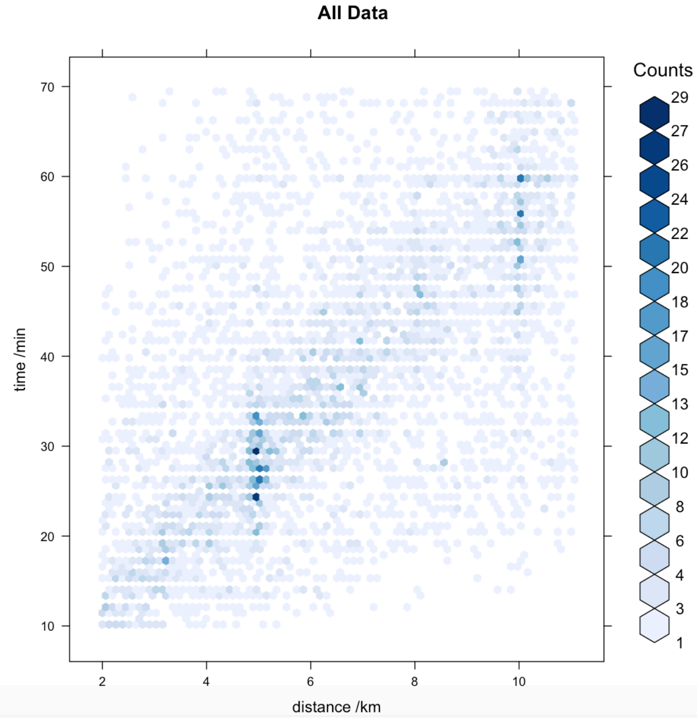
Finally I get 20896 data from strava and the id range is from 746627603 to 746651001, approximately 10 percent of the pages lost some data or have other data forms that my code can not scrape so I delete them and leave 20896 data.

## 2.DATA ANALYSIS & CONCLUSION

I choose R to analyze these data and the main point I focus upon there is any cluster around round numbers in distance and time. The reason why I choose type of exercise is that different types must have different features and might influence each other in analysis. But still I analyze the overall data as a whole to see if there are any other interesting rules. To be concrete, I firstly look at all the data samples and then analyze riding exercise data and finally running exercise data.

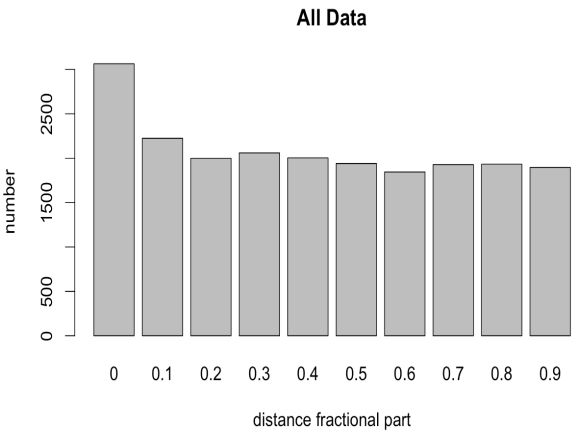
## 2.1 All Data Sample

First of all I draw the graph of distance ~ time distribution of all data sample via “hexbin” package. There is an obvious cluster in line distance=5 and distance=10 in the graph 1:

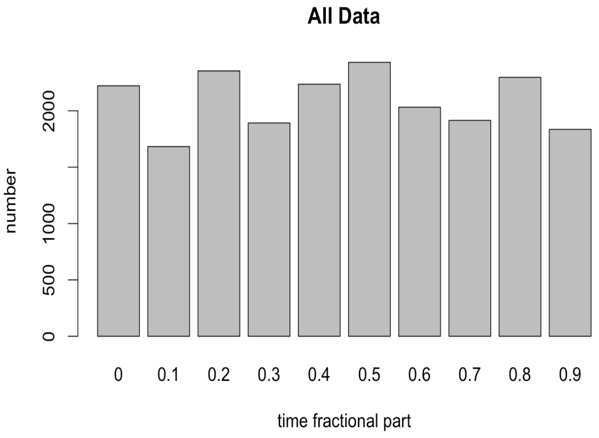


graph 1

To check out whether there is actually more data that near round numbers, I grab the fractional part of both kinds of data (like change 4.8km to 0.8km, change 66.3min to 0.3min) and discover that for all the exercise, distance of exercise clusters in round numbers visibly while time of exercise does not appear this heaping (graph2 ,3):



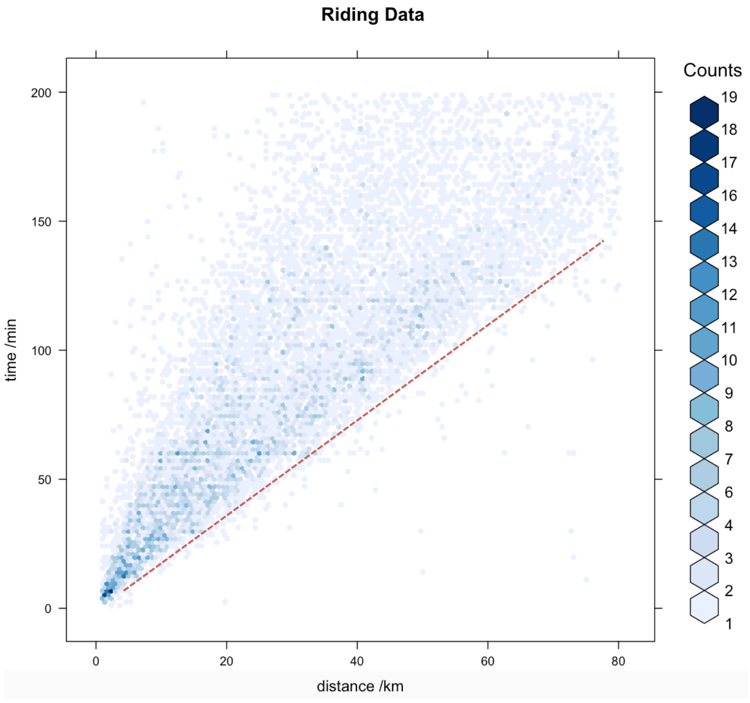
graph 2



graph 3

This appearance relatively fits the theory of "targeting" behavior in distance data but do not have enough evidence in time data. The reason might be that the majority probably believes distance is prior to time in exercise and set their target in distance.

## 2.2 Riding Data

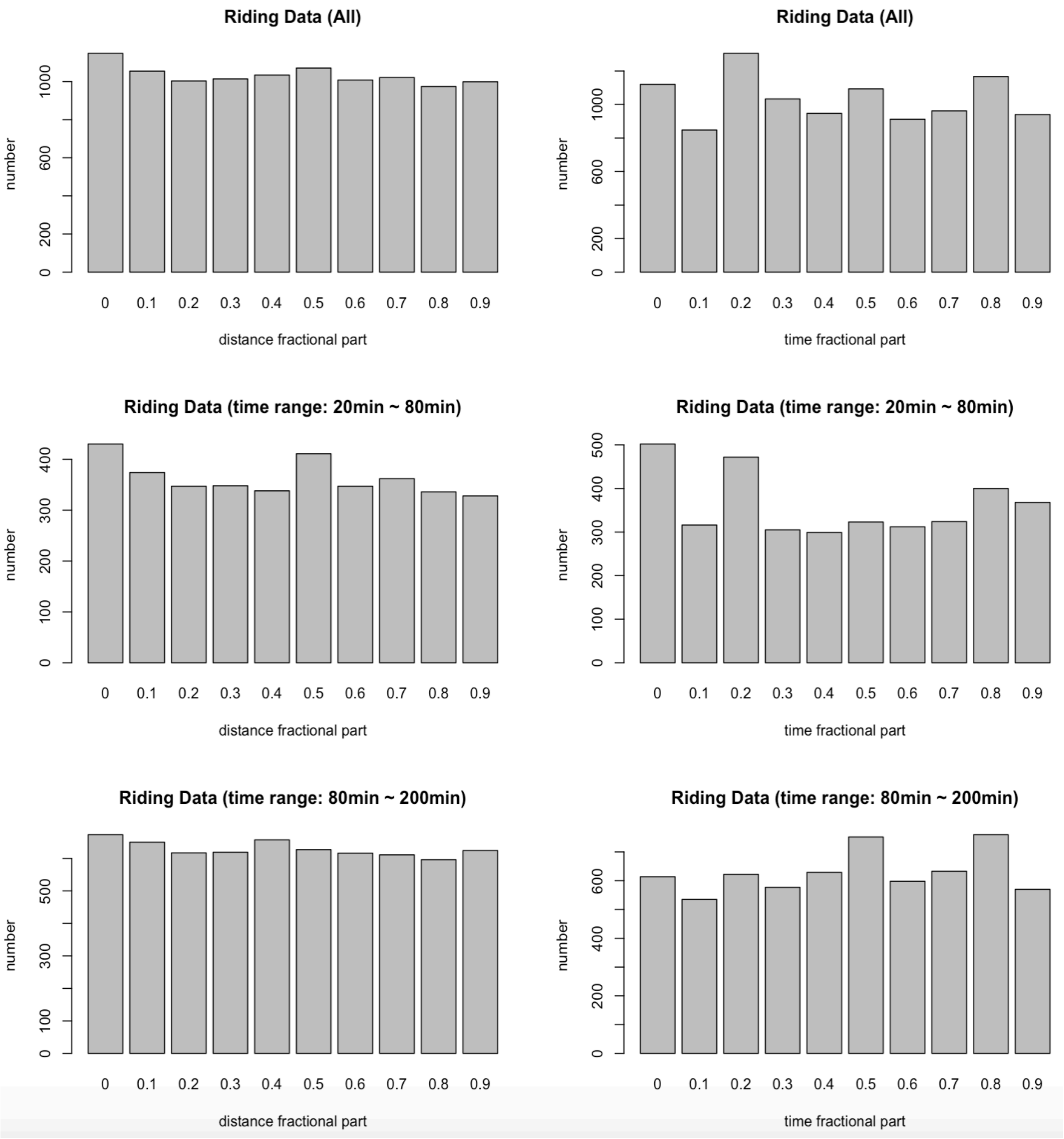


graph 4

I use similar method to see data from riding exercise (with total number of 12663) and get the following graphs. Firstly it is not hard to see that in graph 4 (graph 4 does not show all the riding data as the range is so large, but I do calculate all the data except 0 while drawing graphs showing fractional part of data) there is not an apparent heaping near any distance. (Cluster in bottom left corner only suggests there are more data. When I enlarge that area, there is no cluster). I mark a red dotted line which most of the riding data are above. Tangent of this line can roughly describe the speed of normal people's riding: approximately 30km/h. Those points under the line are probably from professional athletes' exercise data. Via OLS method, I get that the average speed for all the riding data is

$$\overline{v_{ride}} = 20.01 \pm 0.06\ km/h$$

Then I draw graphs about the fractional part of both dimensions of data and find something interesting (graph 5): When I use all the riding data to make graphs (the first row), there is no strong evidence of heaping in round numbers for both distance and time data. But it DO seems to be a little bit more data around round numbers in distance dimension. Then when I choose the data whose time part is in range 20min ~ 80min (second row), cluster around round and half round numbers (0 and 0.5) in distance is clear; there is also an obvious cluster in time dimension around round numbers (there are many data whose time dimension's fractional part is 0.2 but I don't know how to explain it). Therefore, for data in range 20min ~ 80min in time dimension, both time and distance part of data appear a heaping around round numbers. Thirdly I try the range 80min ~ 200min (bottom row) and find that there is no clear cluster anymore for both dimensions (only a little heaping around 0 in distance dimension).



graph 5

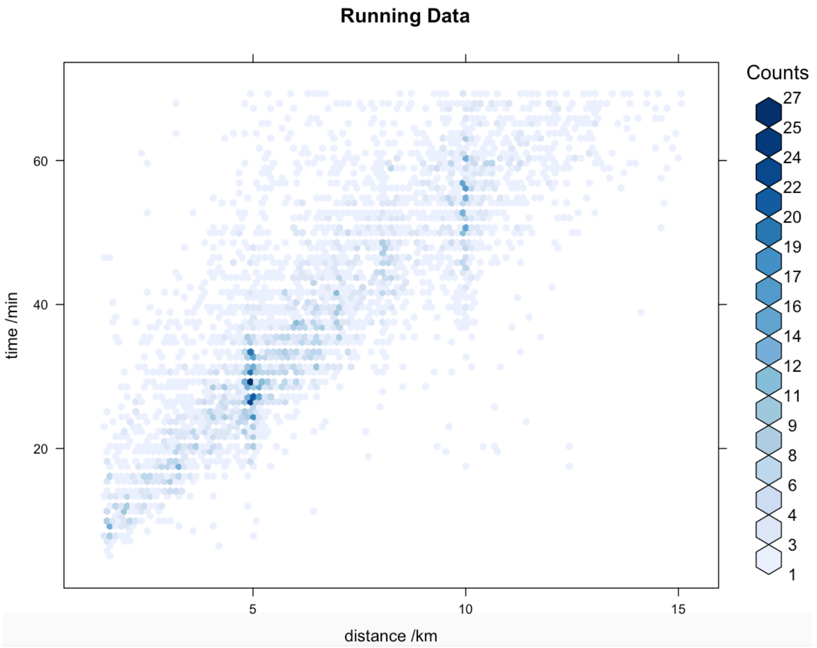
However, when I choose different range of distance for data separation and draw graphs for each group, there is NO such appearance of different distribution as in time dimension.

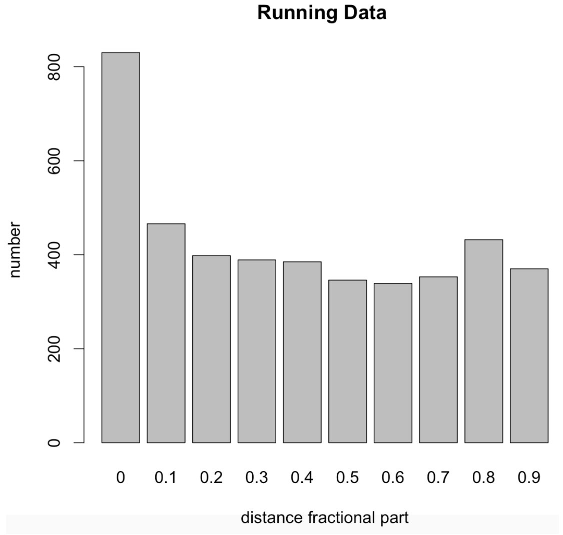
One possible explanation for divergence in time dimension might be the different groups of exercisers. For normal exercisers, riding exercise usually won’t take more than an hour (I choose 80min and the appearance is similar to 60min’s) but the time should not be so short (I choose 20min for bottom line). If normal exercisers have the “target” behavior during exercising, then they would set a time or distance target for every training. Thus clusters in two dimensions can be a strong evidence for the “target” behavior. For athletes that are more professional, they are more likely to exercise longer and further (I choose 80min as a bottom line). When people ride as long as they do, the roads are probably not the normal streets in cities anymore, they might choose some highways that are not crowded in the countryside and start their exercise. But these highways do not have distance that are near round numbers so they can not actually control the time and distance of their training. In another word, the unit of their exercises are not kilometer or minute, but more likely to be a section of highways.

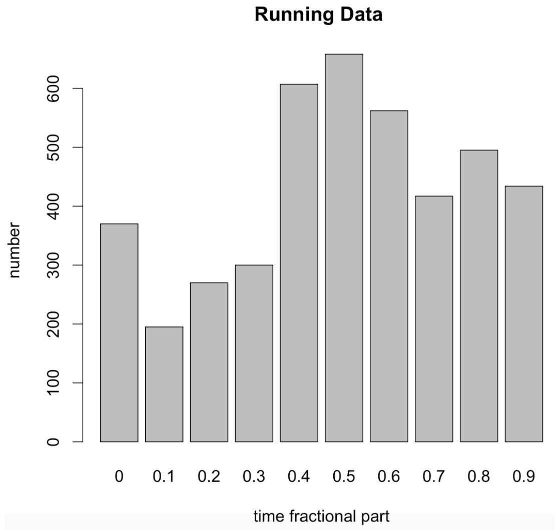
### 2.3 Running Data

I plot graphs using data of running exercise (with total number of 7237) below. From graph 6 there are clear heaping around line distance = 5 and distance = 10. Thus I suppose the cluster of distance in overall data might come from data of running exercise.

graph 6







Via OLS method, I get that the average speed for all the running data is

$$\overline{v_{run}} = 9.23 \pm 0.04\ km/h$$

The cluster in distance dimension is significant for data of running exercise. But time does not show the similar appearance: there are more data near 0.5, which I cannot explain completely. I think it might because that it takes time for exercisers to start and stop the detector in strava app, which is approximately 30 seconds. So when people run for 10 minutes, the detector will write down 10 minutes and a half.

## 3.SUMMARY & DISCUSSION

From all the works above, I believe that, with the existence of cluster of data around round numbers in both time and distance dimensions, indeed there is a “target” behavior in the data I scraped from Strava. But the intensity of this appearance greatly depends on the type of exercise as well as how professional exercisers are.

As for further research, I have several questions that might drive me to understand better about these data:

(1) Is the cluster related to the speed?

(2) Is the scale of cluster in time selected properly? In another word, do people really have “target” minutes in every exercise? How about “target” hours or the “target” is not in one day but in a week or a month? I think maybe the scale might not be how many minutes but more likely to be tens of minutes (For myself, I would not target my exercise time as accurate as 22min or 24min but only 20min or 30min). This is what I think should be done next.

(3) How to distinguish effective time and distance from the raw data? The detector probably writes down data before the beginning of exercise and after the end of exercise. Thus to be more accurate in cluster, being able to distinguish the real effective time and distance from raw data is important.