

PS4-Onishi

Saryu Onishi

February 2023

1 part I

I am interested in scraping data from a site called strava.com. It is a social media platform commonly used by athletes to upload GPS data of their activities. I am not sure how exactly this would work, but I assume it is no different than what we have done in class. And there is an R package called rStrava, but I am yet to look into it.

2 part II

2.1 5d

date in mydf is a character object.

2.2 6-7

"class(df1)" showed that df1 was a tibble, a table-like data structure, and a traditional data frame. "class(df)" showed that df was a Spark DataFrame, but also one that could be queried using SQL syntax and a regular R data frame. It also showed that the operations on the Spark DataFrame are evaluated lazily. This means that any operation on the data frame is only executed when needed to optimize performance in spark.

2.3 6-8

Yes, there is a difference in names. df1 uses periods, where df uses underscore. Since df is the copy of df1 in spark environment, the column name difference suggests a syntax difference between spark and R. Specifically, about the usage of periods ".". This is done in the spark environment to follow the conventions of the SQL language.