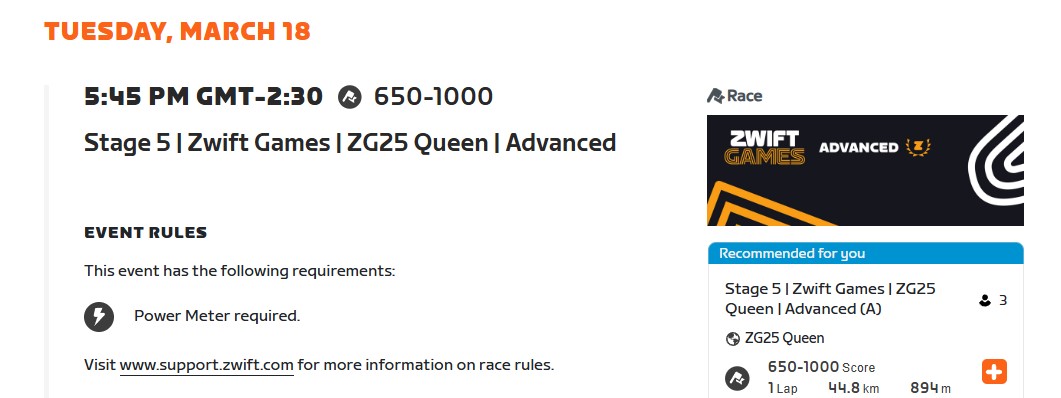
CP1370 Final Project ( 20% final mark )

For this project you are encouraged to work in groups of 2. You may also work individually

You are responsible for advising Eric on his cycling training. You will load all of Eric’s training and race data into spark dataframes, perform transformations and actions to output summary information which Eric can use to help guide his training.

Eric uses a program called strava to record every workout and race which he does. Since October 2023, Eric has completed hundreds of workouts and hundreds of virtual e-sport races.

To kick off the project, Eric <https://www.twitch.tv/coachstock> will be live streaming a race today, Tuesday March 18th at 5:45.



Strava provides the ability to allow each user to download an archive of their data. Amongst this data are files which represent data from workouts and races. The format for these files is a proprietary binary format from a company named Garmin. The extension for the file is “.fit”.

Eric’s fit files are in the folder “activities.zip”. Each of the fit files will need to be unzipped and then the .fit file will need to be converted to a .csv file before it can be uploaded to databricks.

You are required to write a script to perform both the unzip of all files and the conversion from .fit to csv. Please note that the conversion from .fit to .csv can be facilitated using the garmin fit sdk. <https://developer.garmin.com/fit/download/>

The .fit files contain various pieces of data related to cycling. We will be focus on the recorded workout data in the .fit files. These contain the following

-timestamp ( the current time in seconds, this time is in garmin epoch time )

-distance ( the distance covered in the workout )

-heart rate (probably not that useful)

-power ( the amount of power Eric is generating at a given time. This is the best indicator of performance during a workout or race )

-speed (probably not that useful)

Overall we we are mostly interested in the timestamp so we can determine how much training Eric does. We are also interested in how much power Eric can generate during races. For example a rider’s average power over various time intervals can be used to measure their cycling ability. Popular time intervals are as follows:

5 sec

15 sec

30 sec

1 minute

2 minute

5 minute

8 minute

20 minute

Project Submission:

Part 1: Project Prep 5% final mark. ( Due Monday March 31st )

For this part of the project, you will unzip and convert the .fit files into .csv files. You must then load the .csv files into a spark dataframe.

You will need to prepare the dataframe so it can be analyzed. This will involve the following:

-adding a unique identifier to each workout so the data can be grouped by workout

-removing rows from the data which do not contain workout data

-removing columns from the data which will not be used (ex// position\_lat)

-ajusting the schema of the data with the correct data types and names for each column value

As you are completing this part of the project, you are not to perform any manual steps. For example instead of unzipping each .fit file manually, you should write a python script to perform this.

For submission, you must include the following:

-all scripts used to convert the .fit files with

-exported python notebook showing transformations to setup dataframe with correct schema and data

-folder containing all .fit files which have been converted to .csv

Part 2: Final Project Submission 15% final mark. ( Due Thurs April 17th)

Your group will first meet with Eric shortly after submitting the Project Prep part of the project.

Any issues with the data preparation from Part 1 will be highlighted and will need to be fixed before completing part 2.

You will discuss the data and come up with a plan to analyze the data and provide meaningful insights into the data. Specifically you will be looking to analyze where Eric has improved his performance and where Eric has not imporoved his performance. Based on your findings you will then suggest activities/workouts which can target Eric’s weaknesses.

More details and a rubric will be provided for this section