Big Data Analytics Techniques and Applications

Homework 1
Due Date: 2022/03/22 23:59:59

Analyzing NYC Taxi Data

- Dataset

  NYC Taxi Data: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

You need to analyze the NYC Taxi Data by using any data analytic tool or package, and answer the following questions.

*Note that in this homework, we are going to use the **Yellow Taxi Trip Records** from January, 2009 to March, 2009 (totally three months). Make sure that you use the correct dataset.

- Questions:
  - Q1: What regions have the most pickups? What are the top-5 regions with the most pickups and drop-offs (pickups and drop-offs should be counted separately)?
  - Q2: When are the peak hours and off-peak hours for taking a taxi?
    - hint: You can count the number of pickups in different hours of day.
  - Q3: What are the differences between big and small total amounts when taking a taxi?
    - hint: First, you should define what big and small total amounts are. And then, you should point out the difference between them. You should at least observe the results of Q1 and Q2.

- Requirements
  - You might encounter "Big Data" issues in analyzing the NYC dataset (e.g., the data is too large for you to come out the analysis results by your tools/machines). In this case, try your best to incorporate big data and the datasets are Yellow Taxi Trip Records from January, 2009 to March, 2009 (totally three months).
  - Submit a report named "HW1_StudentID.pdf" to E3 and describe clearly the following items:

        - Descriptions of the scale of data, analytical tools, and spec of the platform you use. (You may use any platform/analytical tools you like.)
        - Source code or manipulation steps of data analysis tools
        - Descriptions of how you solve each question in detail.
        - Some figures or tables to illustrate your analyzed answers to each question.
        - Anything else worth mentioning (e.g. other valuable observations, or difficulties encountered in this work and how you resolve them).

- Penalty for late submission
    - If your work is submitted within one day after the deadline, a penalty of 20 percentage marks will be applied.
    - If your work is submitted within two days after the deadline, a penalty of 50 percentage marks will be applied.
    - If your work is submitted over two days after the deadline, you will get 0 in this homework.