Big Data Analytics Techniques and Applications

Homework 1

Due Date: 2023/03/22 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 will use the **Yellow Taxi Trip Records** in **February, June**, and **October 2009** (a total of three months). Please download all datasets on E3 and 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)?
    - Notice: Before analyzing, you should state the definition of the region first.
  - Q2: When are the peak hours and off-peak hours for taking a taxi?
    - Hint: You can count the number of pickups at different hours of the 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 with the analysis results by your tools/machines). In this case, try your best to deal with big data, and then **write down the problem you encountered** (**with some observation**) in the report.
  - Submit a report named "HW1_{StudentID}.pdf" (e.g., HW1_310456099.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.)
    - Manipulation steps of data analysis tools (e.g., providing the screenshot of source code or software GUI with explanation).
    - 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).
  - If you use programming-based data analytic tool or package (e.g., Python, R), you need to submit source code files to E3.
    - You should **zip** all source code files in a file named "HW1_{StudentID}_Code.**zip**" (e.g., HW1_310456099_Code.zip).

- **Penalty for late submission:**.
  - If your work is submitted within one day after the deadline, you will get only **20%** of original score.
  - If your work is submitted within two day after the deadline, you will get only **50%** of original score.
  - If your work is submitted over two days after the deadline, you will get **zero score** on this homework.

- **Penalty for format error:**
  - The report or source code file name has any format error. (-5%)
  - The report is not in pdf. (-5%)
  - The source code file is not in zip. (-5%)