Big Data Analytics Techniques and Applications

Homework 1

Due Date: 2020/10/20 23:59:59

Analyzing NYC Taxi Data

- Dataset

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

Analyze the NYC Taxi Data by using any data analytic tool or package, and answer the following questions:

- Questions
  - Q1: What are the most pickups and drop offs region?
    - hint: You can use some kind of cluster algorithms and count the number of data points of each cluster.
  - Q2: When are the peak hours and off-peak hours for taking taxi?
    - hint: You can count the number of pickups in different hours of day.
  - Q3: What are the differences between short and long distance trips of taking taxi?
    - hint: First, you should define what short and long distance trips are. You may 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 as much data as possible and at least one month of Yellow Taxi Trip Records should be used. The scale of data you used will be counted as an important factor for the score you will get.
  - Submit a report named “Hw1_StudentID.pdf” that describes clearly the following items:
    - Descriptions of the scale of data, tools, and spec of platform you use.

  - Description of how you solve each question in details.
  - 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.