Homework 0

Data preprocessing and exploring the dataset

Due Date: 23:59, October 13, Friday, 2017

TA: 蔡子豪 stu8978@gmail.com

In this homework, you need to do some data preprocessing,and then get some basic information about the dataset via tools.

Dataset:

New York Citi Bike Trip Histories
https://www.citibikenyc.com/system-data
We'll use "201707-citibike-tripdata.csv.zip" in this homework only.

Schema:

You can see the schema,and descriptions in the previous link.

Tasks:

Preprocess:

1. There might be some noise in the dataset,like strange stations or null values.Please detect it and take proper actions to them.
2. For future use,we need to calculate *in-flow* and *out-flow* for each stations every half hour. The result date set should contains **station_id, time, in_flow_count, out_flow_count**

   in/out flow of a station are define as follow:

   The number of trips move from/to the station within the 30 minutes period.So one day can be splitted into 48 segments.

Query:

1. How many stations are there in this dataset,and what is the average distance between them?

$$average\ distance = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} dist(Si,Sj)}{N^2}$$

2. What are the top 3 frequent stations pairs(start stations,end stations) in weekdays,how about in weekends? (2017/07/01 is Saturday)

   (Si,Sj) is not (Sj,Si)

3. Find the top 3 stations with highest average out-flow ,and top 3 highest average in-flow

4. What is the most popular station(highest average inflow+outflow)?
    a. Draw the in-flow(**A**) and out-flow(**B**) for that station in a line chart
    b. Calculate the distance function between **A** and **B**
    (you can simply use euclidean distance,here is a tool in scikit learn)
    c. Calculate the distance function between **A-mean(A)** and **B-mean(B)**,and draw them both.
    d. Calculate the distance function between **(A-mean(A))/std(A)** and **(B-mean(B))/std(B)**,and draw them both.
    e. Calculate the distance function between $\{A_i - f(i) | Ai \in A\}$ and $\{B_i - f(i) | Bi \in B\}$,and draw them both.

    f *is the linear function that minimize* $\sum (A_i - f(i))^2$ **,**see reference.

    f. Calculate the distance function between **Smooth(A)** and **Smooth(B)**,and draw them both.

    You can choose any smoothing function,just specify it,
    or simply use $Ai = \frac{A_{i-1}+A_i+A_{i+1}}{3}$
    (or take the average of 5,7,9… elements)

5. Please try to find some interesting query or observation in the dataset

Report:

Your report can be in pdf,odt,html but we strongly recommend html format generated from R markdown,or Jupyter notebook(In the following homework we may only accept these two format,so try it as early as you can:))

In the preprocess part,you should mention what you discover and what action you have taken,write down how you calculate in/out flow.

In the query part, if you use some database, please include your SQL and screenshot in the report,if you write a program or use a library, please hand in with code and paste the screenshots(or images) in the report.

Be sure your images are send with report,not a local link
(download from E3 and open the report to check)

If you have any questions or suggestions,fell free to contact me:)
BTW, there is a discussion borad here(hackmd) look around before you ask:)