Report

Group: 30 Students: Fredrik Andresen, Peter Lawrence

Introduction

Git repository

Briefly explain the task and the problems you have solved. How did you work as a group? If you used Git, a link to the repository would be nice.

Our task for this assigment was to clean the Geolife dataset and insert into a database we created. We were then supposed to perform a series of SQL queries on the data. We cleaned and inserted the data by using Python. The data was first extracted from the dataset and made into pandas dataframes in DataInsert.py. The dataframes were then cleaned and inserted into the database with DataInsert.py and SqlQueries.py.

The SQL queries were performed in the database with the help of the SQL queries in SqlQueriespt2.py. We worked well together as a group. We used Git to share code and to keep track of changes and worked together both in person and through Discord.

Results

Add your results from the tasks, both as text and screenshots. Short sentences are sufficient.

Part 1

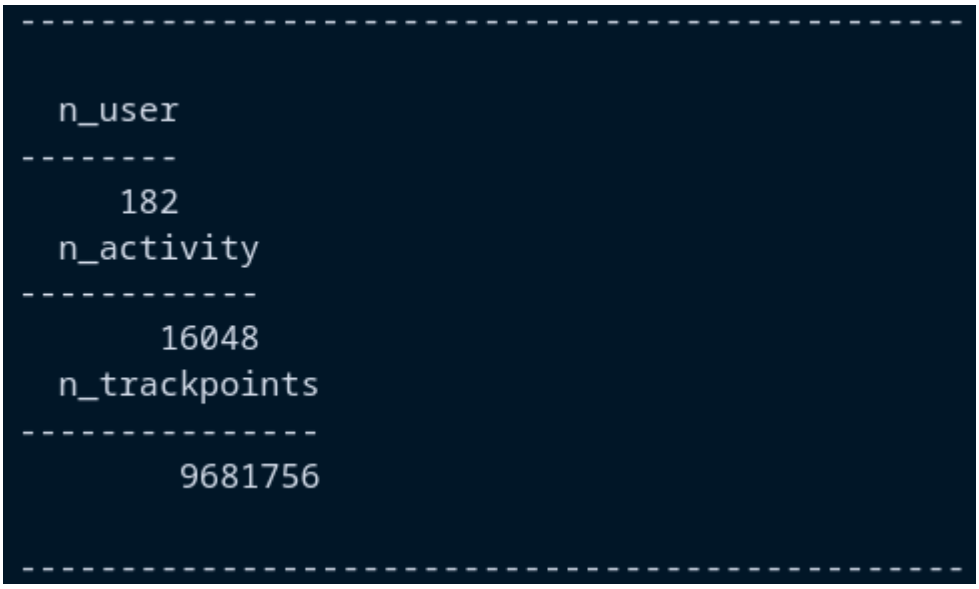
```
mysql> SELECT * FROM User LIMIT 10;
+-----------+-------------+
| user_id   | has_labels  |
+-----------+-------------+
| 000       |        NULL |
| 001       |        NULL |
| 002       |        NULL |
| 003       |        NULL |
| 004       |        NULL |
| 005       |        NULL |
| 006       |        NULL |
| 007       |        NULL |
| 008       |        NULL |
| 009       |        NULL |
+-----------+-------------+
10 rows in set (0.00 sec)
```

```
mysql> SELECT * FROM Activity LIMIT 10;
+-------------+---------+----------------------+---------------------+---------------------+
| activity_id | user_id | transportation_mode  | start_date_time     | end_date_time       |
+-------------+---------+----------------------+---------------------+---------------------+
|           1 | 175     | bus                  | 2007-10-22 01:47:04 | 2007-10-22 01:56:45 |
|           2 | 175     | NULL                 | 2007-10-21 07:21:53 | 2007-10-21 11:56:06 |
|           3 | 175     | NULL                 | 2007-10-19 05:23:15 | 2007-10-19 13:18:10 |
|           4 | 175     | NULL                 | 2007-10-19 22:12:53 | 2007-10-20 02:52:37 |
|           5 | 044     | NULL                 | 2009-03-22 23:20:38 | 2009-03-22 23:21:03 |
|           6 | 044     | NULL                 | 2009-03-26 11:27:26 | 2009-03-26 12:13:26 |
|           7 | 044     | NULL                 | 2009-04-09 12:23:05 | 2009-04-09 13:14:45 |
|           8 | 044     | NULL                 | 2009-03-12 11:48:28 | 2009-03-12 12:01:58 |
|           9 | 044     | NULL                 | 2009-03-22 09:25:07 | 2009-03-22 12:10:02 |
|          10 | 044     | NULL                 | 2009-02-26 09:17:34 | 2009-02-26 14:12:34 |
+-------------+---------+----------------------+---------------------+---------------------+
10 rows in set (0.00 sec)
```

```
mysql> SELECT * FROM TrackPoint LIMIT 10;
+---------------+-------------+-----------------+-----------------+----------+---------------------+
| trackpoint_id | activity_id | lat             | lon             | altitude | date_time           |
+---------------+-------------+-----------------+-----------------+----------+---------------------+
|             1 |           1 |        39.96335 |        116.3307 |      105 | 2007-10-22 01:47:04 |
|             2 |           1 |         39.9659 | 116.314716666667 |      105 | 2007-10-22 01:47:40 |
|             3 |           1 | 39.9659166666667 |       116.31565 |      105 | 2007-10-22 01:48:25 |
|             4 |           1 |         39.9665 | 116.315866666667 |      105 | 2007-10-22 01:49:25 |
|             5 |           1 | 39.9669166666667 | 116.314716666667 |      125 | 2007-10-22 01:49:46 |
|             6 |           1 | 39.9679666666667 | 116.314283333333 |      217 | 2007-10-22 01:51:19 |
|             7 |           1 |         39.9681 | 116.313866666667 |      285 | 2007-10-22 01:56:45 |
|             8 |           2 | 39.1136333333333 |        117.1702 |        0 | 2007-10-21 07:21:53 |
|             9 |           2 | 39.1132333333333 |       117.17065 |       16 | 2007-10-21 07:26:45 |
|            10 |           2 | 39.1132666666667 |       117.17285 |       26 | 2007-10-21 07:30:57 |
+---------------+-------------+-----------------+-----------------+----------+---------------------+
10 rows in set (0.01 sec)
```

Part 2

Question 1



There are 182 users, 16048 activities and 9681756 trackpoints in the dataset.

Question 2

```
-----------------------------------------------

  user_id    AVG(trackpoint_count)
---------  ----------------------
      000                 670.871
      001                 824.965
      002                 924.801
      003                 807.387
      004                 761.858
      005                 587.548
      006                 801.125
      007                 758.275
      008                    1376
      009                 1218.32
      010                 1212.15
      011                 451.756
      012                 629.783
      013                  1388.3
      014                 905.936
      015                 806.583
      016                 690.222
      017                 868.245
      018                 540.795
...
      178                      84
      179                    2419
      180                    1024
      181                     110
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
  user_id    MIN(trackpoint_count)
---------  -----------------------
      000                       5
      001                      33
      002                       4
      003                       3
      004                       4
      005                       5
      006                      14
      007                       6
      008                     165
      009                     134
      010                       3
      011                       3
      012                      64
      013                      13
      014                       3
      015                      22
      016                       7
```

017	4
018	3
019	11
020	5
021	5
022	15

user_id	MAX(trackpoint_count)
000	2359
001	2472
002	2438
003	2485
004	2482
005	2058
006	2478
007	2228
008	2499
009	2396
010	2485
011	2306
012	2277
013	2486
014	2499
015	2411
016	2360
017	2471
018	2245
019	2276
020	2325
021	744
022	2421

Truncated tables of the average, minimum and maximum number of trackpoints per user is seen above.

Question 3

user_id	result
128	2102
153	1793
025	715
163	704
062	691
144	563
041	399
085	364
004	346
140	345
167	320
068	280
017	265
003	261
014	236

The top 15 users with the highest number of activities are seen in the table above.

Question 4

user_id
010
052
062
073
081
084
085
091
092
112
125
128
175

All the users who have taken a bus are in the table above.

Question 5

user_id	mode_count
128	9
062	7
085	4
084	3
058	3
163	3
078	3
081	3
112	3
065	2

The top 10 users with the highest amount of different transportation modes used seen in the table above.

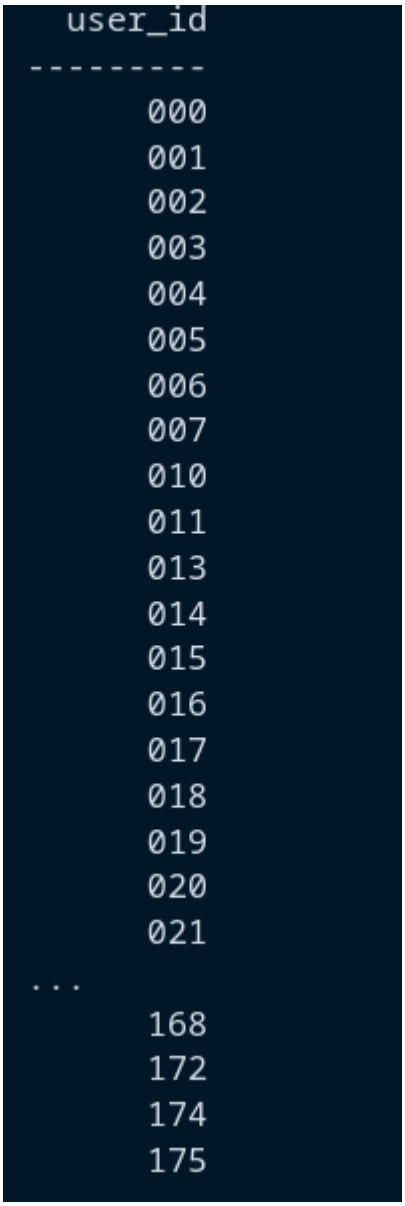
Question 6

activity_id

There are no activities that are registered multiple times.

Question 7

a)



user_id
000
001
002
003
004
005
006
007
010
011
013
014
015
016
017
018
019
020
021
...
168
172
174
175

The number of users who have started an activity in one day and finished it in another day can be seen in the table above.

b)

user_id	transportation_mode	duration
175		0 days 4 hours 39 minutes 44 seconds
044		0 days 2 hours 47 minutes 19 seconds
044		0 days 0 hours 47 minutes 15 seconds
044		0 days 0 hours 58 minutes 45 seconds
044		0 days 12 hours 14 minutes 12 seconds
044		0 days 14 hours 22 minutes 4 seconds
044		0 days 3 hours 0 minutes 45 seconds
044		0 days 6 hours 5 minutes 52 seconds
044		0 days 0 hours 40 minutes 50 seconds
044		0 days 2 hours 33 minutes 30 seconds
044		0 days 0 hours 9 minutes 20 seconds
044		0 days 7 hours 54 minutes 19 seconds
044		0 days 5 hours 43 minutes 45 seconds
044		0 days 1 hours 51 minutes 28 seconds
051		0 days 5 hours 46 minutes 53 seconds
051		1 days 5 hours 1 minutes 3 seconds
011		0 days 0 hours 26 minutes 25 seconds
011		0 days 0 hours 29 minutes 15 seconds
011		0 days 0 hours 34 minutes 25 seconds
...		
039		0 days 0 hours 30 minutes 5 seconds
039		0 days 9 hours 56 minutes 48 seconds
039		0 days 0 hours 32 minutes 20 seconds
039		0 days 0 hours 9 minutes 45 seconds

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

The list of transportation mode, user id and duration for those activities are in the table above.

Question 8

For thistask we tried numerous queries, we've ended up with some queries that we believe have potential but none of them has completed their queries as of now. However we have some suggestions to how such a query can look like:

```
SELECT
    user_id1,
    user_id2,
    T1.tr_activity_id AS activity_id1,
    T2.tr_activity_id AS activity_id2,
    T1.tr_date_time AS date_time1,
    T2.tr_date_time AS date_time2,
    T1.lat AS lat1,
    T1.lon AS lon1,
    T2.lat AS lat2,
    T2.lon AS lon2
```

```
WITH close_time (user_id1, user_id2, activity_id1, activity_id2) AS (
    SELECT DISTINCT A1.user_id AS user1, A2.user_id AS user2,
A1.activity_id AS activity_id1, A2.activity_id AS activity_id2
    FROM Activity AS A1
    INNER JOIN Activity AS A2 ON A1.user_id <> A2.user_id
        AND (ABS(TIMESTAMPDIFF(SECOND, A1.start_date_time,
A2.end_date_time)) <= 30
            OR ABS(TIMESTAMPDIFF(SECOND, A1.end_date_time,
A2.start_date_time)) <= 30
            OR A1.start_date_time BETWEEN A2.start_date_time AND
A2.end_date_time
            OR A1.end_date_time BETWEEN A2.start_date_time AND
A2.end_date_time)
        AND A1.user_id < A2.user_id
        AND A1.user_id < 050
),
trackpoints (tr_activity_id, lat, lon, tr_date_time) AS (
    SELECT activity_id, lat, lon, date_time
    FROM TrackPoint
)
SELECT COUNT(DISTINCT user_id1) AS number_of_users
    FROM close_time
INNER JOIN trackpoints AS T1 ON activity_id1 = T1.tr_activity_id
INNER JOIN trackpoints AS T2 ON activity_id2 = T2.tr_activity_id
AND ST_Distance_Sphere(
        POINT(T1.lon, T1.lat),
        POINT(T2.lon, T2.lat)
        ) <= 50
```

Other possible queries can be found in the folder `src/PART2_individual_SQL`

Question 9

user_id	altitude_gain
128	139415077.9024
144	58499208.3848
153	47142690.7184
111	24241207.1184
163	21519546.3040
041	18172433.7368
004	16746594.0032
085	13334995.5272
002	11986135.4376
039	10640030.8416
030	10227064.8424
003	9684581.5936
167	9124008.9184
084	8343871.6712
013	6994028.2680

15 rows in set (3 min 6.08 sec)

The top 15 users who have gained the most altitude are in the table above.

Question 10

user_id	transportation_mode	total_daily_distance	date_date
108	walk	29273.53825036155	2007-10-03
139	walk	29273.53825036155	2007-10-03
128	bike	73143.37846558148	2008-06-28
062	run	40.082311659663425	2008-09-02
062	train	350357.8121884449	2008-09-02
128	taxi	41157.635837726746	2008-09-30
128	car	520868.34688762604	2008-10-02
128	subway	35760.66025889879	2008-10-31
128	boat	69989.41164522138	2008-11-22
128	bus	221259.96968129047	2009-01-20
128	airplane	44462819.63427694	2009-03-06

11 rows in set (21.15 sec)

The users who have traveled the longest total distance in one day for each transportation mode are in the table above.

Question 11

127	4
128	719
129	6
130	8
131	10
132	3
133	4
134	31
135	5
136	6
138	10
139	12
140	86
141	1
142	52
144	157
145	5
146	7
147	30
150	16
151	1
152	2
153	556
154	14
155	30
157	9
158	9
159	5
161	7
162	9
163	232
164	4

164	6
165	2
166	2
167	134
168	19
169	9
170	2
171	3
172	9
173	5
174	54
175	4
176	8
179	28
180	2
181	14

171 rows in set (3 min 18.69 sec)

user_id	N_invalid_activities
000	101
001	45
002	98
003	179
004	219
005	44
006	17
007	30
008	16
009	31
010	50
011	32
012	43
013	29
014	118
015	46

016	20
017	129
018	27
019	31
020	20
021	7
022	55
023	11
024	27
025	263
026	18
027	2
028	36
029	25
030	112
031	3
032	12
033	2
034	88
035	23
036	34
037	100
038	58
039	147
040	17
041	201
042	54
043	21
044	31
045	7
046	13
047	6
048	1
050	8
051	36
052	44
053	7

054	2
055	15
056	7
057	16
058	13
059	5
060	1
061	12
062	248
063	8

The users who have invalid activities and the number of invalid activities they have are in the tables above.

Question 12

user_id	max_count	transportation_mode
010	3	taxi
020	81	bike
021	1	walk
052	1	bus
056	15	bike
058	2	taxi
060	1	walk
064	1	bike
065	10	bike
067	1	walk
069	1	bike
073	52	walk
075	1	walk
076	3	car
078	37	walk
080	1	taxi
080	1	bike
081	4	bike
082	2	walk
084	9	walk

```
| 085      |          18 | walk                 |
| 086      |           2 | car                  |
| 087      |           5 | walk                 |
| 089      |           7 | car                  |
| 091      |           2 | bus                  |
| 091      |           2 | walk                 |
| 092      |           1 | walk                 |
| 097      |           9 | bike                 |
| 098      |           1 | taxi                 |
| 101      |           3 | car                  |
| 102      |           7 | bike                 |
| 107      |           1 | walk                 |
| 108      |           2 | walk                 |
| 111      |           3 | taxi                 |
| 112      |          69 | walk                 |
| 115      |          80 | car                  |
| 117      |           1 | walk                 |
| 125      |           3 | bike                 |
| 126      |          13 | bike                 |
| 128      |         312 | car                  |
| 136      |           3 | walk                 |
| 138      |           2 | bike                 |
| 139      |           4 | bike                 |
| 144      |           2 | walk                 |
| 153      |           5 | walk                 |
| 161      |           1 | walk                 |
| 163      |          26 | bike                 |
| 167      |          31 | bike                 |
| 175      |           1 | bus                  |
+----------+-------------+----------------------+
49 rows in set (0.04 sec)
```

The users who have registered transportation modes and their most used transportation mode are in the table above.

Discussion

Discuss your solutions. Did you do anything differently than how it was explained in the assignment sheet, in that case why and how did that work? Were there any pain points or problems? What did you learn from this assignment?

We did not do anything differently than how it was explained in the assignment sheet.

For the question queries, we ended up doing all with pure SQL. It would perhaps have been easier to do some of them with Python, as we are more familiar with that. However, we decided to try to do them all with SQL only.

Some pain points were that some of the queries became fairly long and complicated. This made them hard to read and debug when they didn't work. Some of them also took a long time to run, which severly slowed down the process of debugging.

For question 10 we ended up with two users who had the same total distance for the same transportation mode. We decided to include both users in the answer, although we suspect they are duplicates in the database.

> The jupyter notebook Part2.ipynb has some details about the more complicated queries. Also the notebook provides (with the other .py delivered) functions for calling all queries, except Q8. Results from these calls are also in this notebook.

Feedback

The query for question 8 took an excessive amount of time to figure out and to execute (More than 1,5 hours to execute). It would have been helpful if one of the hints gave an idea of how to solve the problem efficiently.