Report

Assignment 2 - MySQL

Group: 10
Students: Anna Sofie Lunde, Astrid Kristine Ragnhildsdatter Bakken and Maria Lundin Brenna

Introduction

The task was to create tables, clean and insert data into the tables, and then query the database using Python and MySQL. The data we worked with was from Microsoft's Geolife GPS Trajectory dataset. The dataset contains data on users' outdoor movements and our task was inspired by the workout application Strava.

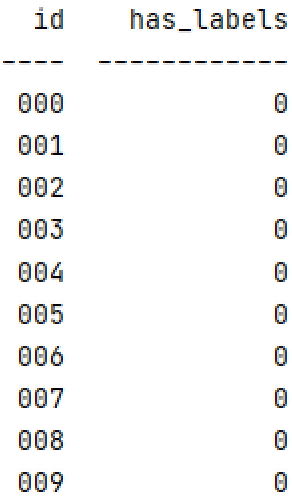
We mostly worked together during the creation of the tables and the data cleaning as we considered this part as crucial for the project. After the data was inserted into the database, each group member got responsibility for some queries. When each query had a draft, we discussed the different drafts and agreed on the final version of the query.

Link to our Git repository: https://github.com/annalunde/TDT4225-Assignment-2

Results

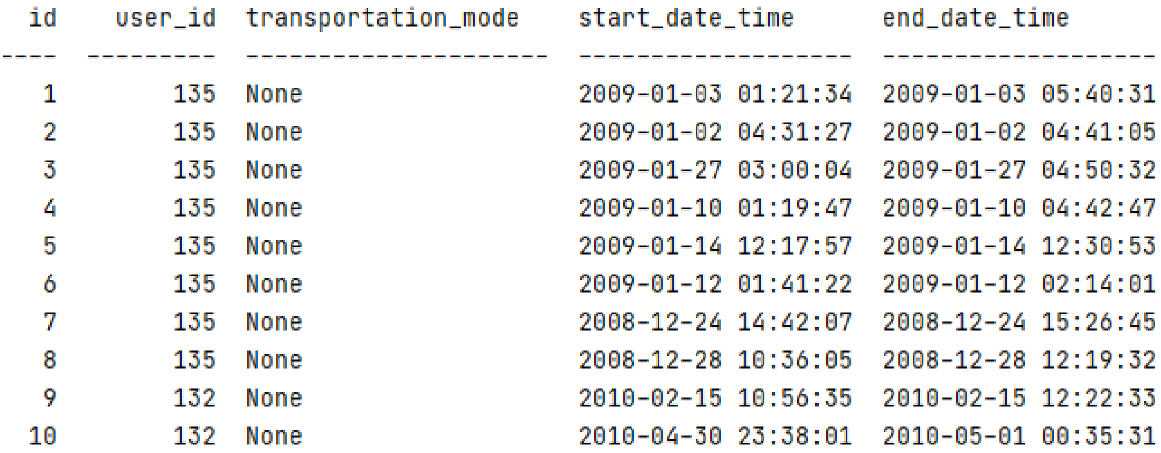
Part 1

User – top 10 rows



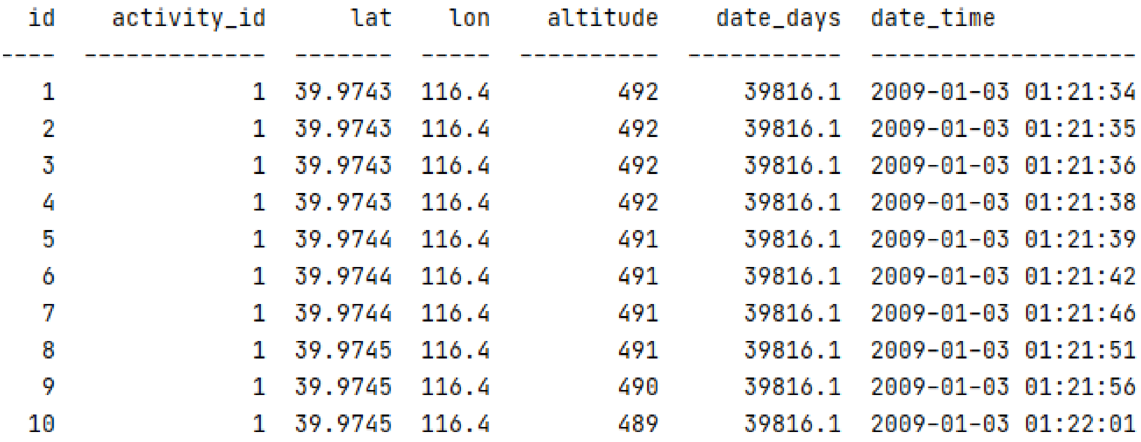
```
  id    has_labels
----  ------------
 000             0
 001             0
 002             0
 003             0
 004             0
 005             0
 006             0
 007             0
 008             0
 009             0
```

Activity – top 10 rows



```
  id    user_id  transportation_mode    start_date_time      end_date_time
----  ---------  ---------------------  -------------------  -------------------
   1        135  None                   2009-01-03 01:21:34  2009-01-03 05:40:31
   2        135  None                   2009-01-02 04:31:27  2009-01-02 04:41:05
   3        135  None                   2009-01-27 03:00:04  2009-01-27 04:50:32
   4        135  None                   2009-01-10 01:19:47  2009-01-10 04:42:47
   5        135  None                   2009-01-14 12:17:57  2009-01-14 12:30:53
   6        135  None                   2009-01-12 01:41:22  2009-01-12 02:14:01
   7        135  None                   2008-12-24 14:42:07  2008-12-24 15:26:45
   8        135  None                   2008-12-28 10:36:05  2008-12-28 12:19:32
   9        132  None                   2010-02-15 10:56:35  2010-02-15 12:22:33
  10        132  None                   2010-04-30 23:38:01  2010-05-01 00:35:31
```

TrackPoint – top 10 rows



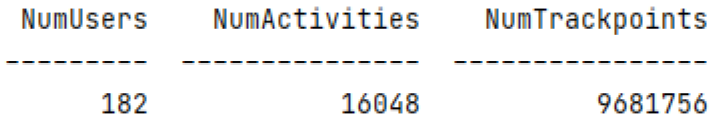
```
  id    activity_id      lat    lon    altitude    date_days  date_time
----  -------------  -------  -----  ----------  -----------  -------------------
   1              1  39.9743  116.4         492      39816.1  2009-01-03 01:21:34
   2              1  39.9743  116.4         492      39816.1  2009-01-03 01:21:35
   3              1  39.9743  116.4         492      39816.1  2009-01-03 01:21:36
   4              1  39.9743  116.4         492      39816.1  2009-01-03 01:21:38
   5              1  39.9744  116.4         491      39816.1  2009-01-03 01:21:39
   6              1  39.9744  116.4         491      39816.1  2009-01-03 01:21:42
   7              1  39.9744  116.4         491      39816.1  2009-01-03 01:21:46
   8              1  39.9745  116.4         491      39816.1  2009-01-03 01:21:51
   9              1  39.9745  116.4         490      39816.1  2009-01-03 01:21:56
  10              1  39.9745  116.4         489      39816.1  2009-01-03 01:22:01
```

Part 2

Note: The assumptions for the different queries are listed under *Discussion*

Query 1:

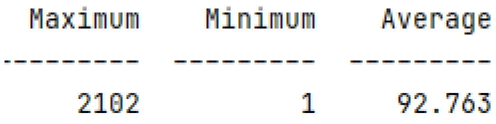
Number of users: 182
Number of activities: 16048
Number of trackpoints: 9681756



```
 NumUsers     NumActivities     NumTrackpoints
---------   ---------------   ----------------
      182             16048            9681756
```

Query 2:

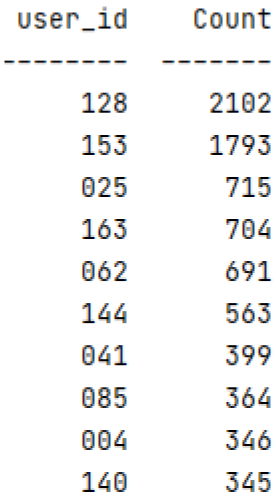
Maximum number of activities per user: 2102
Minimum number of activities per user: 1
Average number of activities per user: 92.763



```
  Maximum     Minimum     Average
---------   ---------   ---------
     2102           1      92.763
```

Query 3:

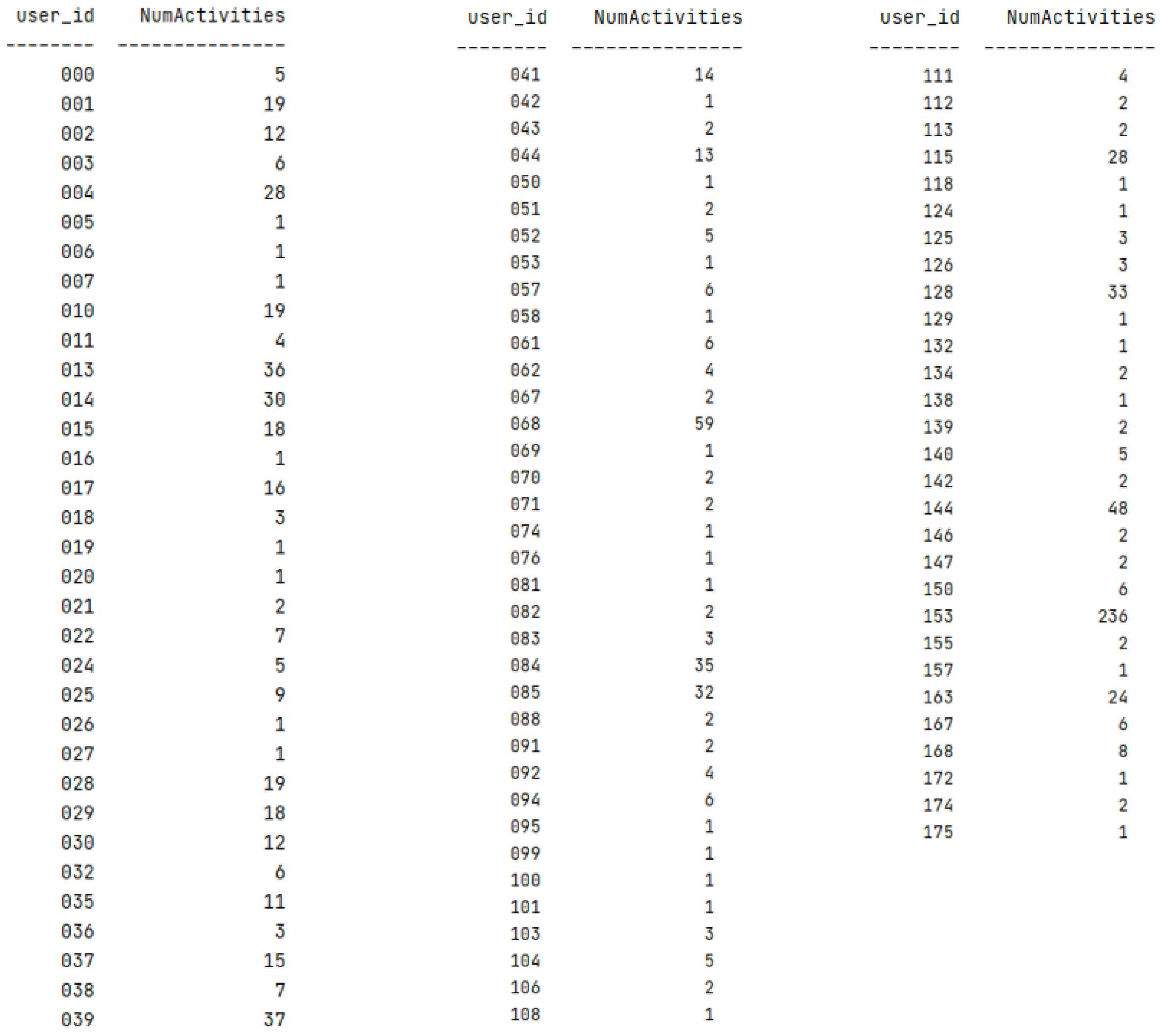
Top 10 users with the highest number of activities:



```
 user_id     Count
--------   -------
     128      2102
     153      1793
     025       715
     163       704
     062       691
     144       563
     041       399
     085       364
     004       346
     140       345
```

Query 4:

The number of activities for users, where the activity is started in one day and ended in the next day:



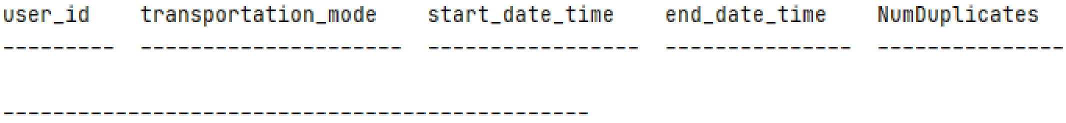
```
user_id    NumActivities
--------  ---------------
     000                5
     001               19
     002               12
     003                6
     004               28
     005                1
     006                1
     007                1
     010               19
     011                4
     013               36
     014               30
     015               18
     016                1
     017               16
     018                3
     019                1
     020                1
     021                2
     022                7
     024                5
     025                9
     026                1
     027                1
     028               19
     029               18
     030               12
     032                6
     035               11
     036                3
     037               15
     038                7
     039               37
```

```
user_id    NumActivities
--------  ---------------
     041               14
     042                1
     043                2
     044               13
     050                1
     051                2
     052                5
     053                1
     057                6
     058                1
     061                6
     062                4
     067                2
     068               59
     069                1
     070                2
     071                2
     074                1
     076                1
     081                1
     082                2
     083                3
     084               35
     085               32
     088                2
     091                2
     092                4
     094                6
     095                1
     099                1
     100                1
     101                1
     103                3
     104                5
     106                2
     108                1
```

```
user_id    NumActivities
--------  ---------------
     111                4
     112                2
     113                2
     115               28
     118                1
     124                1
     125                3
     126                3
     128               33
     129                1
     132                1
     134                2
     138                1
     139                2
     140                5
     142                2
     144               48
     146                2
     147                2
     150                6
     153              236
     155                2
     157                1
     163               24
     167                6
     168                8
     172                1
     174                2
     175                1
```

Query 5:

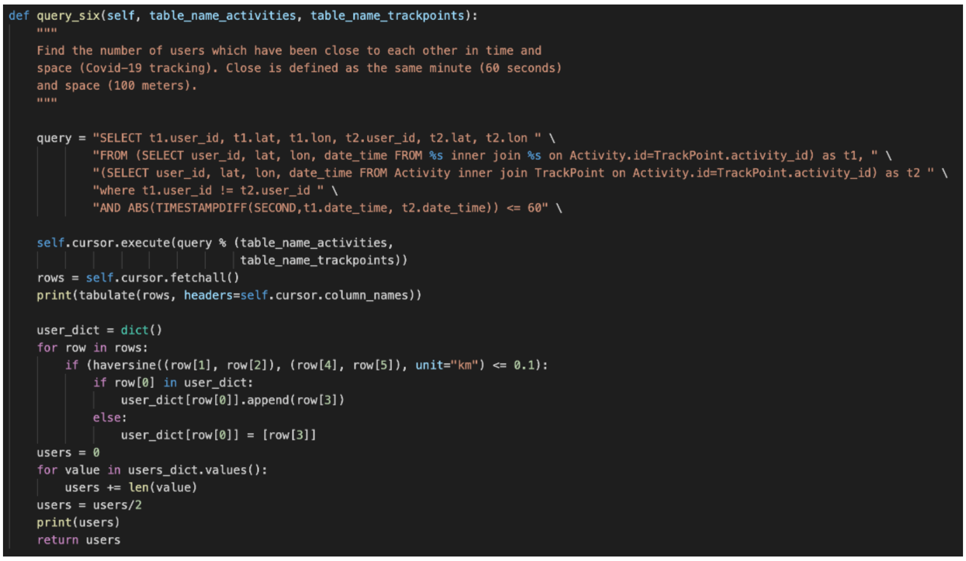
Activities that are registered multiple times: Zero results



```
user_id    transportation_mode    start_date_time    end_date_time    NumDuplicates
---------  ---------------------  -----------------  ---------------  ---------------

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

Query 6:

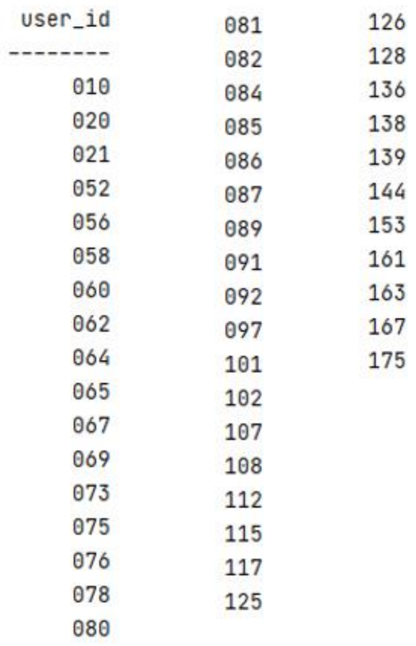
Have tried to run query 6, but the query is too heavy, so no results are reported.



```python
def query_six(self, table_name_activities, table_name_trackpoints):
    """
    Find the number of users which have been close to each other in time and
    space (Covid-19 tracking). Close is defined as the same minute (60 seconds)
    and space (100 meters).
    """

    query = "SELECT t1.user_id, t1.lat, t1.lon, t2.user_id, t2.lat, t2.lon " \
            "FROM (SELECT user_id, lat, lon, date_time FROM %s inner join %s on Activity.id=TrackPoint.activity_id) as t1, " \
            "(SELECT user_id, lat, lon, date_time FROM Activity inner join TrackPoint on Activity.id=TrackPoint.activity_id) as t2 " \
            "where t1.user_id != t2.user_id " \
            "AND ABS(TIMESTAMPDIFF(SECOND,t1.date_time, t2.date_time)) <= 60" \

    self.cursor.execute(query % (table_name_activities,
                                 table_name_trackpoints))
    rows = self.cursor.fetchall()
    print(tabulate(rows, headers=self.cursor.column_names))

    user_dict = dict()
    for row in rows:
        if (haversine((row[1], row[2]), (row[4], row[5]), unit="km") <= 0.1):
            if row[0] in user_dict:
                user_dict[row[0]].append(row[3])
            else:
                user_dict[row[0]] = [row[3]]
    users = 0
    for value in users_dict.values():
        users += len(value)
    users = users/2
    print(users)
    return users
```

Query 7:

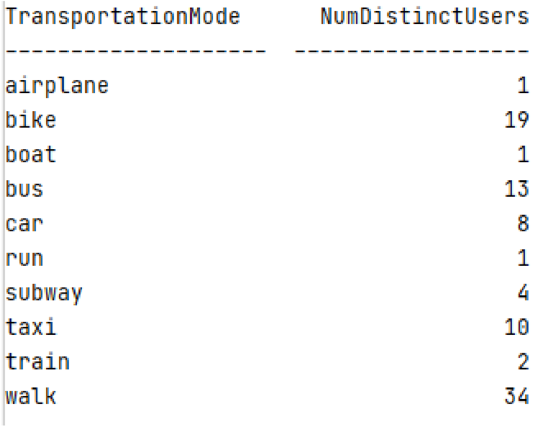
All users that have never taken a taxi:



```
  user_id     081    126
---------     082    128
      010     084    136
      020     085    138
      021     086    139
      052     087    144
      056     089    153
      058     091    161
      060     092    163
      062     097    167
      064     101    175
      065     102
      067     107
      069     108
      073     112
      075     115
      076     117
      078     125
      080
```

Query 8:

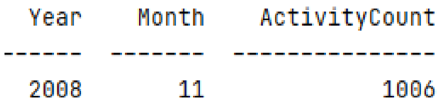
All types of transportation modes and how many distinct users have used the different transportation modes:



```
TransportationMode      NumDistinctUsers
--------------------  ------------------
airplane                               1
bike                                  19
boat                                   1
bus                                   13
car                                    8
run                                    1
subway                                 4
taxi                                  10
train                                  2
walk                                  34
```

Query 9:

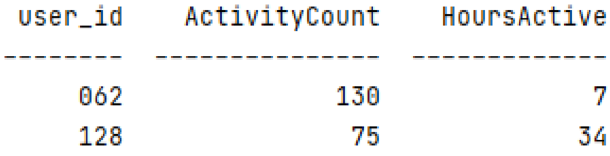
a) Year and month with the most activities: November 2008



```
  Year    Month    ActivityCount
------  -------  ---------------
  2008       11             1006
```

b) The user that had the most activities in November 2008, and the number of recorded hours: User 062 with 130 activities and 7 hours recorded

The user with the most activities in November 2008 (User 062 – 7 hours) does not have more hours recorded than the user with the second most activities (User 128 – 34 hours).



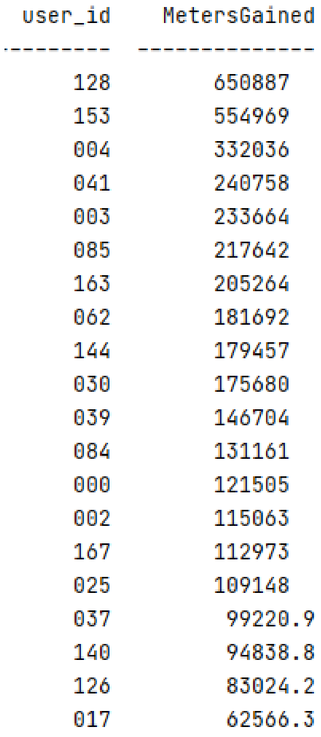
```
 user_id    ActivityCount    HoursActive
---------  ---------------  -------------
      062              130              7
      128               75             34
```

Query 10:

Total distance (in km) walked in 2008 by user with id=112: 115.47465961507991 km

Query 11:

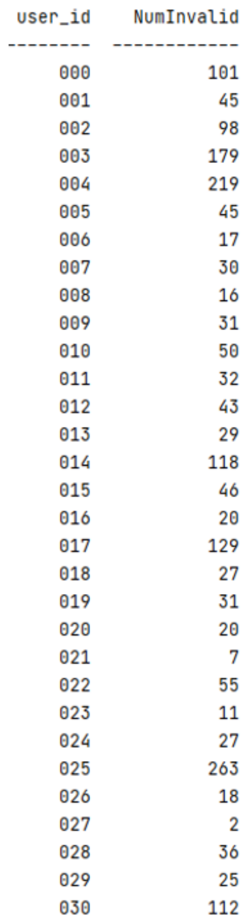
Find the top 20 users who have gained the most altitude:



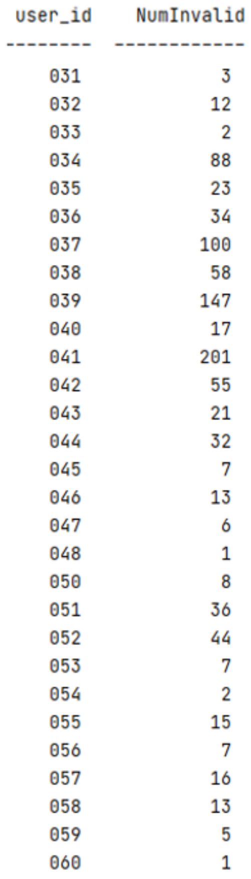
```
 user_id    MetersGained
---------  --------------
      128        650887
      153        554969
      004        332036
      041        240758
      003        233664
      085        217642
      163        205264
      062        181692
      144        179457
      030        175680
      039        146704
      084        131161
      000        121505
      002        115063
      167        112973
      025        109148
      037         99220.9
      140         94838.8
      126         83024.2
      017         62566.3
```

Query 12:

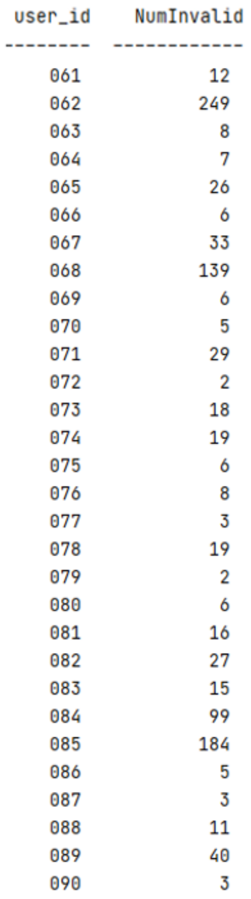
All users who have invalid activities, and the number of invalid activities per user:



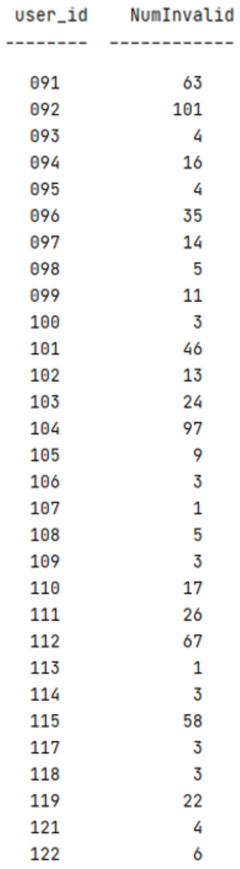
```
user_id   NumInvalid
--------  ------------
     000           101
     001            45
     002            98
     003           179
     004           219
     005            45
     006            17
     007            30
     008            16
     009            31
     010            50
     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
```



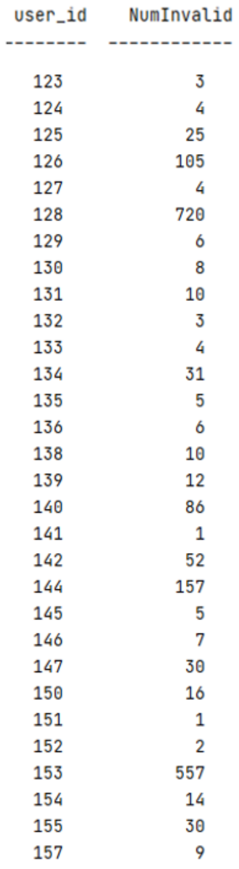
```
user_id   NumInvalid
--------  ------------
     031             3
     032            12
     033             2
     034            88
     035            23
     036            34
     037           100
     038            58
     039           147
     040            17
     041           201
     042            55
     043            21
     044            32
     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
```



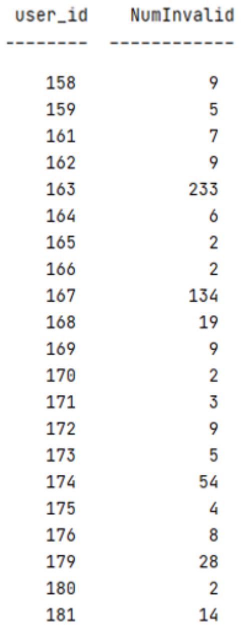
```
user_id   NumInvalid
--------  ------------
     061            12
     062           249
     063             8
     064             7
     065            26
     066             6
     067            33
     068           139
     069             6
     070             5
     071            29
     072             2
     073            18
     074            19
     075             6
     076             8
     077             3
     078            19
     079             2
     080             6
     081            16
     082            27
     083            15
     084            99
     085           184
     086             5
     087             3
     088            11
     089            40
     090             3
```



```
user_id   NumInvalid
--------  ------------

     091            63
     092           101
     093             4
     094            16
     095             4
     096            35
     097            14
     098             5
     099            11
     100             3
     101            46
     102            13
     103            24
     104            97
     105             9
     106             3
     107             1
     108             5
     109             3
     110            17
     111            26
     112            67
     113             1
     114             3
     115            58
     117             3
     118             3
     119            22
     121             4
     122             6
```



```
user_id   NumInvalid
--------  ------------

     123             3
     124             4
     125            25
     126           105
     127             4
     128           720
     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           557
     154            14
     155            30
     157             9
```



```
user_id   NumInvalid
--------  ------------

     158             9
     159             5
     161             7
     162             9
     163           233
     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
```

Discussion

Part 1

- To keep track of the activityID given in the database, the file name of the activity along with activityID were saved in a separate .txt-file. This .txt-file was used when data was inserted into the TrackPoint table, so that a trackpoint would be linked to the correct activity with the correct activityID.

Part 2

- In Query 4, we found the number of activities for users, where the activity is started in one day and ended in the next day (as mentioned on Piazza).
- Query 6 was very heavy, which led to it not returning any results.
- For Query 7, we assumed that we should only consider labelled activities, but that not all activities were to be labelled for that user to be considered as never have taken a taxi.
- For Query 9a, we assumed that the activities belonged to the year and month according to their start_date_time. It was deemed reasonable due to that it would be very complicated to keep track of the "border activities" - that started in one year/month and ended in another year/month. We also assumed that this would apply to few activities.
- A tip that was given in the assignment sheet was that variables in SQL might come in handy. For Query 9b, we decided to use the output for month and year from Query 9a instead, since this was an easier solution that we knew would provide the correct answer. However, it might have been a more elegant solution to use variables.

Learning points

- During this exercise we refreshed our knowledge within SQL and got practice in setting up and populating a database.
- We also learned the importance of knowing the structure of your dataset. We had to really get familiar with the data before being able to create tables, populate them, and write queries to extract the desired data.

Feedback

The exercise was interesting and fun to execute. However, Query 6 was time consuming, and it would have been nice with more tips on how to test queries that are too heavy to be executed within a reasonable time frame.