Teamwork: Tiffany Truong, Tianjiao Xu, Sumit Bisawas

CREATING rating_final & userprofile tables

```
MariaDB [(none)]> SHOW DATABASES;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| tttruon5           |
+--------------------+
2 rows in set (0.01 sec)

MariaDB [(none)]> use tttruon5;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MariaDB [tttruon5]> SHOW TABLES;
+--------------------+
| Tables_in_tttruon5 |
+--------------------+
| chefmozcuisine     |
| geoplaces2         |
| regions            |
| warehouse          |
| wb                 |
+--------------------+
5 rows in set (0.00 sec)

MariaDB [tttruon5]> CREATE TABLE rating_final(
    -> userID VARCHAR(20),
    -> placeID BIGINT,
    -> rating BIGINT,
    -> food_rating BIGINT,
    -> service_rating BIGINT
    -> );
Query OK, 0 rows affected (0.09 sec)

MariaDB [tttruon5]> show tables;
+--------------------+
| Tables_in_tttruon5 |
```

```
+--------------------+
| chefmozcuisine     |
| geoplaces2         |
| rating_final       |
| regions            |
| warehouse          |
| wb                 |
+--------------------+
6 rows in set (0.00 sec)

MariaDB [tttruon5]> create table userprofile(
    -> userID VARCHAR(20),
    -> latitude FLOAT,
    -> longitude FLOAT,
    -> smoker VARCHAR(20),
    -> drink_level VARCHAR(60),
    -> dress_preference VARCHAR(100),
    -> ambience VARCHAR(40),
    -> transport VARCHAR(50),
    -> marital_status VARCHAR(30),
    -> hijos VARCHAR(60),
    -> birth_year BIGINT,
    -> interest VARCHAR(70),
    -> personality VARCHAR(100),
    -> religion VARCHAR(60),
    -> activity VARCHAR(40),
    -> color VARCHAR(40),
    -> weight BIGINT,
    -> budget VARCHAR(40),
    -> height FLOAT
    -> );
Query OK, 0 rows affected (0.04 sec)

MariaDB [tttruon5]> show tables;
+--------------------+
| Tables_in_tttruon5 |
+--------------------+
| chefmozcuisine     |
| geoplaces2         |
| rating_final       |
| regions            |
| userprofile        |
| warehouse          |
| wb                 |
+--------------------+
```

```
7 rows in set (0.00 sec)
```

Please perform at least the following steps with MySQL:

• Select and keep the rows containing student users (identified in the "activity" attribute in data file userprofile.csv)

```
SELECT * FROM userprofile WHERE activity='student';
```

```
MariaDB [tttruon5]> SELECT * FROM userprofile WHERE activity='student' LIMIT 10;
+--------+----------+-----------+--------+---------------+------------------+----------+-----------+----------------+-------------+------------+------------+--------------------+---------
| userID | latitude | longitude | smoker | drink_level   | dress_preference | ambience | transport | marital_status | hijos       | birth_year | interest   | personality        | religion
 | activity | color  | weight | budget | height |
+--------+----------+-----------+--------+---------------+------------------+----------+-----------+----------------+-------------+------------+------------+--------------------+---------
+----------+--------+--------+--------+--------+
| U1001  |    22.14 |  -100.979 | false  | abstemious    | informal         | family   | on foot   | single         | independent |       1989 | variety    | thrifty-protector  | none
 | student  | black  |     69 | medium |   1.77 |
| U1002  |  22.1501 |  -100.983 | false  | abstemious    | informal         | family   | public    | single         | independent |       1990 | technology | hunter-ostentatious | Catholic
 | student  | red    |     40 | low    |   1.87 |
| U1003  |  22.1198 |  -100.947 | false  | social drinker | formal           | family   | public    | single         | independent |       1989 | none       | hard-worker        | Catholic
 | student  | blue   |     60 | low    |   1.69 |
| U1005  |  22.1835 |   -100.96 | false  | abstemious    | no preference    | family   | public    | single         | independent |       1992 | none       | thrifty-protector  | Catholic
 | student  | black  |     65 | medium |   1.69 |
| U1006  |    22.15 |  -100.983 | true   | social drinker | no preference    | friends  | car owner | single         | independent |       1989 | variety    | hard-worker        | none
 | student  | blue   |     75 | medium |    1.8 |
| U1007  |  22.1185 |  -100.938 | false  | casual drinker | informal         | solitary | public    | single         | independent |       1989 | variety    | thrifty-protector  | Catholic
 | student  | purple |     60 | low    |   1.59 |
| U1008  |   22.123 |  -100.924 | false  | social drinker | formal           | solitary | public    | single         | independent |       1989 | technology | hard-worker        | Catholic
 | student  | green  |     68 | low    |   1.72 |
| U1009  |  22.1594 |   -100.99 | false  | abstemious    | formal           | family   | on foot   | single         | kids        |       1991 | variety    | thrifty-protector  | Catholic
 | student  | green  |     75 | medium |   1.78 |
| U1010  |  22.1909 |  -100.999 | false  | social drinker | no preference    | friends  | car owner | married        | kids        |       1987 | technology | hard-worker        | none
 | student  | green  |     40 | medium |   1.67 |
| U1011  |   23.725 |  -99.1529 | false  | abstemious    | no preference    | family   | public    | single         | independent |       1989 | variety    | hard-worker        | Catholic
 | student  | purple |     68 | medium |   1.78 |
+--------+----------+-----------+--------+---------------+------------------+----------+-----------+----------------+-------------+------------+------------+--------------------+---------
+----------+--------+--------+--------+--------+
10 rows in set (0.00 sec)

MariaDB [tttruon5]>
```


• Construct a single data file/collection with at least the following attributes:

- **userID**;
- **placeID**;
- **rating** (among the three ratings-related attributes in data file rating_final.csv , only "rating" is relevant to this assignment);
- at least one of the attributes in userprofile.csv that you deem relevant (for instance, "**birth_year**" would qualify if you feel that it might explain some ratings); and
- at least one of the attributes in **geoplaces2.csv** that you deem relevant (for instance, "**price**" would qualify if you feel that it might explain some ratings).

```
INSERT INTO ratingHW4 (userID, placeID, rating, birth_year,
price)

SELECT userprofile.userID, rating_final.placeID,
rating_final.rating, userprofile.birth_year, geoplaces2.price

FROM userprofile

JOIN rating_final

ON userprofile.userID = rating_final.userID

JOIN geoplaces2

ON rating_final.placeID = geoplaces2.placeID;
```

userID	placeID	rating	birth_year	price
U1077	132825	2	1987	low
U1068	132732	0	1988	low
U1067	132732	1	1987	low
U1103	132732	0	1989	low
U1103	132613	2	1989	medium
U1103	132667	1	1989	low
U1070	132608	2	1991	low
U1070	132609	1	1991	low

….

SELECT * FROM ratingHW4

INTO OUTFILE ‘./ratingHW4.csv’

FIELDS TERMINATED BY ‘,‘

ENCLOSED BY ““‘

LINES TERMINATED BY ‘\n’

```
MariaDB [tttruon5]> INSERT INTO ratingHW4 (userID, placeID, rating, birth_year, price)
    -> SELECT userprofile.userID, rating_final.placeID, rating_final.rating, userprofile.birth_year, geopl
aces2.price
    -> FROM userprofile
    -> JOIN rating_final
    -> ON userprofile.userID = rating_final.userID
    -> JOIN geoplaces2
    -> ON rating_final.placeID = geoplaces2.placeID
    -> ;
Query OK, 1161 rows affected (3.77 sec)
Records: 1161  Duplicates: 0  Warnings: 0

MariaDB [tttruon5]> select * from ratingHW4;
+--------+---------+--------+------------+--------+
| userID | placeID | rating | birth_year | price  |
+--------+---------+--------+------------+--------+
| U1077  |  132825 |      2 |       1987 | low    |
| U1068  |  132732 |      0 |       1988 | low    |
| U1067  |  132732 |      1 |       1987 | low    |
| U1103  |  132732 |      0 |       1989 | low    |
| U1103  |  132613 |      2 |       1989 | medium |
| U1103  |  132667 |      1 |       1989 | low    |
| U1070  |  132608 |      2 |       1991 | low    |
| U1070  |  132609 |      1 |       1991 | low    |
| U1070  |  132613 |      1 |       1991 | medium |
| U1031  |  132668 |      0 |       1992 | low    |
| U1082  |  132613 |      0 |       1989 | medium |
| U1082  |  132608 |      0 |       1989 | low    |
| U1082  |  132609 |      0 |       1989 | low    |
| U1060  |  132564 |      2 |       1991 | low    |
| U1123  |  132608 |      1 |       1987 | low    |
| U1123  |  132667 |      1 |       1987 | low    |
| U1123  |  132613 |      1 |       1987 | medium |
| U1021  |  132668 |      2 |       1984 | low    |
| U1083  |  135046 |      2 |       1981 | medium |
| U1108  |  135086 |      1 |       1983 | medium |
| U1108  |  135046 |      1 |       1983 | medium |
| U1012  |  134983 |      2 |       1988 | high   |
```

• Perform at least two operations with missing values. For example, you might drop rows with missing values, impute some of the missing values, etc.

```
DELETE FROM userprofile WHERE activity= ‘?’;

DELETE FROM userprofile WHERE budget= ‘?’;
```

```
MariaDB [tttruon5]> select count(*) from userprofile;
+----------+
| count(*) |
+----------+
|      138 |
+----------+
1 row in set (0.00 sec)

MariaDB [tttruon5]> DELETE FROM userprofile where activity='?';
Query OK, 7 rows affected (0.01 sec)

MariaDB [tttruon5]> select count(*) from userprofile;
+----------+
| count(*) |
+----------+
|      131 |
+----------+
1 row in set (0.00 sec)

MariaDB [tttruon5]>
```

```
MariaDB [tttruon5]> DELETE FROM userprofile where budget='?';
Query OK, 3 rows affected (0.01 sec)

MariaDB [tttruon5]> select count(*) from userprofile;
+----------+
| count(*) |
+----------+
|      128 |
+----------+
1 row in set (0.01 sec)

MariaDB [tttruon5]>
```

• Describe at least two meaningful questions that you could answer with the data set that you have constructed. Discuss what attributes would be relevant to answering the questions. Discuss how generalizable you think your answers would be and discuss why. You do not have to actually perform the analysis to answer the questions.

- We can ask how the price from geoplaces2 will affect the overall rating for the same userID.
    - The attributes we would need to look at would be price from geoplaces2, userID joined from both the userprofile.csv & rating_final.csv datasets, as well as the overall rating from rating_final.

- We can also analyze how the birth year (which tells us the user's age) will affect their ratings and where they would go to eat foods where the prices fall more into their budget.
    - The attributes needed would be birth_year from userprofile, rating from rating_final, and price from geoplaces2.

Between the two questions we've provided above, we think that the first question would be more generalizable as it allows us to analyze how the price will affect the overall rating for a specific place (either positively or negatively) through various trends since a place with a lower price range can either have higher ratings or lower ratings depending on the quality of the food being served. On the other hand, the second question can't be as generalized as the first since it only analyzes the user's age and the assumption that their age will affect their rating or their budget range. The second question doesn't include any other factors that can affect an individual's budget since even though they may be younger, they could come from an affluent family versus an older individual who is in the middle income class.

If we wanted to generalize the conclusions drawn from the second question, we could include how ratings are affected by more than price and age alone. We can include other factors such as activity (as a student or professor), marital status (single or married), or even their dress_preference (informal or formal) as people who are inclined to dress more formally will gravitate to more costly areas.