Q2. SQL skills assessment:

1. You have the following table at disposal (called *campaign_comms_user_level*):

campaign_id	user_id	date_sent	opened_email	clicked_on_link	visited_landing_page	converted
183946	68324	20180705	1	1	0	0
183946	65715	20180712	1	0	0	0
421564	46546	20180702	1	1	1	1
654654	86451	20180801	0	0	0	0
421564	65158	20180807	1	0	0	0
…	…	…	…	…	…	…

Before starting doing question 2, I have created two empty tables in the base, which we use in Fellowship Project. It is MariaDB base.

I have assumed that there are five campaigns (183946, 421564, 654654, 165715, 283946) sent on five different dates (02.07.2018, 05.072018, 12.07.2018, 01.08.2018, 07.08.2018) to some customer base. I have no information in the assignment as to whether such assumptions are appropriate.

I have filled in 60 fields with random data. Tried to use the information, which were in tables in Question2.

```
INSERT INTO campaign_comms_user_level (id, campaign_id, user_id,
date_sent, opened_email, clicked_on_link, visited_landing_page,
converted)
VALUES

For example:
(1,183946,68324,20180705,1,1,0,0),
(2,183946,65715,20180712,1,0,0,0),
(3,421564,46546,20180702,1,1,1,1),
(4,654654,86451,20180801,0,0,0,0),
(5,421564,65158,20180807,1,0,0,0),
…
```


Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

In my opinion it does mean:
if 0,0,0,0 the event is email_sent,
if 1,0,0,0, the event is opened_email,
if 1,1,0,0, the event is clicked_on_link,
if 1,1,1,0, the event is visited_landing_page,
if 1,1,1,1 the event is converted ?

In the second table were use long name of events:
email_sent, opened_email, clicked_on_link, visited_landing_page, converted

Unfortunately in our base they were too long. They should have till 11 signs (char(11)).

I had to change te names for:
email_sent, opened_em, click_link, visit_page, converted

```
INSERT INTO campaign_comms (id, date, campaign_id, user_id,
event)
VALUES

For example:
(1,20180705, 183946, 68324, 'clic_link'),
(2,20180712, 183946, 65715, 'opened_em'),
(3,20180702, 421564, 46546, 'converted'),
(4,20180801, 654654, 86451, 'email_sent'),
(5,20180807, 421564, 65158, 'opened_em'),
…
```

Write SQL queries to:

a. Calculate a "funnel" per campaign (= ordered milestones with proportion of successes in each milestone based on the successes in the previous milestone).

Ad a.
First I want to count how many **single conversions** are on each milestones.

```
SELECT COUNT(*)
FROM campaign_comms_user_level
WHERE opened_email = '1';
```

Result
57

Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

```sql
SELECT COUNT(*)
FROM campaign_comms_user_level
WHERE clicked_on_link = '1' AND opened_email = '1';
```

```
Result
47
```

```sql
SELECT COUNT(*)
FROM campaign_comms_user_level
WHERE clicked_on_link = '1' AND opened_email = '1' AND
visited_landing_page  = '1';
```

```
Result
34
```

```sql
SELECT COUNT(*)
FROM campaign_comms_user_level
WHERE clicked_on_link = '1' AND opened_email = '1'
AND visited_landing_page  = '1' AND converted = '1';
```

```
Result
20
```

```
Now I want to check the percentage conversion for every
milestones.
```

```sql
SELECT
     (SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE opened_email = '1') AS Opened,
     (SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE clicked_on_link = '1') AS Clicked,
     ROUND((SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE clicked_on_link = '1')*100/ (SELECT COUNT(*)
FROM campaign_comms_user_level
          WHERE opened_email = '1') ,2) AS Percent;
```

Opened	Clicked	Percent
57	47	82.46

Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

```sql
SELECT
     (SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE clicked_on_link = '1') AS Opened,
     (SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE visited_landing_page = '1') AS Clicked,
     ROUND((SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE visited_landing_page = '1')*100/ (SELECT
COUNT(*) FROM campaign_comms_user_level
          WHERE clicked_on_link = '1') ,2) AS Percent;
```

Opened	Clicked	Percent
47	34	72.34

```sql
SELECT
     (SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE visited_landing_page = '1') AS Opened,
     (SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE converted = '1') AS Clicked,
     ROUND((SELECT COUNT(*) FROM campaign_comms_user_level
          WHERE converted = '1')*100.0/ (SELECT COUNT(*) FROM
campaign_comms_user_level
          WHERE visited_landing_page = '1'),2) AS Percent;
```

Opened	Clicked	Percent
34	26	76.47

Next step should be combined calculating. I want to count how many conversions are on each milestones **per campaign.**

I can use SUM() because the value is 1 or 0

Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

```sql
SELECT campaign_id AS Campaign, SUM(opened_email) AS Opened,
SUM(clicked_on_link) AS Clicked, SUM(visited_landing_page) AS
Visited, SUM(converted) AS Converted
FROM campaign_comms_user_level
GROUP BY campaign_id;
```

Campaign	Opened	Clicked	Visited	Converted
165,715	12	9	8	4
183,946	12	7	2	2
283,949	12	11	8	7
421,564	12	11	11	10
654,654	9	9	5	3

```sql
SELECT campaign_id, SUM(opened_email) AS Op,
SUM(clicked_on_link) AS Cl,
SUM(visited_landing_page) AS Vis, SUM(converted) AS Con,
ROUND(SUM(clicked_on_link)*100/NULLIF(SUM(opened_email),0),2) AS
ClPerc,
ROUND(SUM(visited_landing_page)*100/
NULLIF(SUM(clicked_on_link),0),2) AS VisPerc,
ROUND(SUM(converted)*100/NULLIF(SUM(visited_landing_page),0),2)
AS ConPerc
FROM campaign_comms_user_level
GROUP BY campaign_id;
```

campaign_id	Op	Cl	Vis	Con	ClPerc	VisPerc	ConPerc
165,715	12	9	8	4	75	88.89	50
183,946	12	7	2	2	58.33	28.57	100
283,949	12	11	8	7	91.67	72.73	87.5
421,564	12	11	11	10	91.67	100	90.91
654,654	9	9	5	3	100	55.56	60

Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

b. Find the most successful campaign in terms of
i) proportion of opened (among sent) emails
ii) proportion of conversions among emails opened.

Ad b.i)

```
SELECT campaign_id AS Campaign, COUNT(date_sent) AS Send,
SUM(opened_email) AS Opened,
ROUND(SUM(opened_email)*100/COUNT(date_sent),2) AS Percentage
FROM campaign_comms_user_level
GROUP BY campaign_id
ORDER BY Percentage DESC;
```

Campaign	Send	Opened	Percentage
165,715	12	12	100
283,949	12	12	100
183,946	12	12	100
421,564	12	12	100
654,654	12	9	75

Ad b.ii)

```
SELECT campaign_id AS Campaign, SUM(opened_email) AS Opened,
SUM(converted) AS Converted,
ROUND(SUM(converted)*100/COUNT(opened_email),2) AS Percentage
FROM campaign_comms_user_level
GROUP BY campaign_id
ORDER BY Percentage DESC;
```

Campaign	Opened	Converted	Percentage
421,564	12	10	83.33
283,949	12	7	58.33
165,715	12	4	33.33
654,654	9	3	25
183,946	12	2	16.67

Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

c. Find all users who were contacted with at least 5 campaigns and
i) converted in more than 75% of cases
ii) never converted.

Ad c.i)
In this section I don't exactly understand what does it mean the word 'converted' in both sentences.

In my opinion in this section it does it mean every value 1 for whatever milestone.

```
SELECT COUNT(campaign_id) AS NumberCampaign, user_id AS UserId,
(SUM(opened_email)+SUM(clicked_on_link)
+SUM(visited_landing_page)+SUM(converted)) AS OneSum,
COUNT(date_sent)*4 AS Total,
(SUM(opened_email)+SUM(clicked_on_link)
+SUM(visited_landing_page)+SUM(converted))/(COUNT(date_sent)*4)
AS Converted
FROM campaign_comms_user_level
GROUP BY user_id
HAVING COUNT(campaign_id) >= 5 AND Converted > 0.75;
```

NumberCampaign	UserId	OneSum	Total	Converted
5	86,451	16	20	0.8
5	46,465,433	16	20	0.8

Ad c.ii)
I have changed the conditions for campaign_id = 1 and opened = 0.

```
SELECT COUNT(campaign_id) AS NumberCampaign, user_id AS UserId,
SUM(opened_email) AS Opened,
COUNT(date_sent)*4 AS TotalClick, SUM(opened_email)/
(COUNT(date_sent)*4) AS Converted
FROM campaign_comms_user_level
GROUP BY user_id
HAVING COUNT(campaign_id) = 1 AND Opened = 0;
```

NumberCampaign	UserId	Opened	TotalClick	Converted
1	8,645,111	0	4	0

Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

```sql
SELECT COUNT(campaign_id) AS NumberCampaign, user_id AS UserId,
SUM(opened_email) AS Opened,
COUNT(date_sent)*4 AS TotalClick, SUM(opened_email)/
(COUNT(date_sent)*4) AS Converted
FROM campaign_comms_user_level
GROUP BY user_id
HAVING COUNT(campaign_id) >=5 AND Opened = 0;
```

For conditions gave in task there were not results.

2. Regarding the metrics calculated in the previous task, is there any methodological issue / something we need to take care of? How would you assess the quality of the data stored in *campaign_comms_user_level*?

Ad 2.
In my opinion the table ‘campaign_comms_user_level' should has additional columns:

1. dates and time of conversion (morning, lunch, evening...), to optimize each milestone. We would be able to assess the speed of a given path, at each stage of the path. At what time, after the shipping date is the highest conversion. And when customers are likely to open an email or make a purchase.
For example, when a customer stops at a certain step, you can stimulate their interest by sending an additional offer, a discount coupon...

2. The type of campaign (SMS, newsletter, display...) and the source (www, Facebook, Google...) gives us a picture of which campaign performs better in our funnel for specific customers.

3. Demographic data on customers (age, gender, interests, occupation...) and geographical data (city, village, ...) are also very important.
Customer segmentation, i.e. adding information on whether a customer is new or returning.

4. There is a lack of information about what the customer has bought and for how much. And whether they have made a purchase before.

Note: In real world we don’t get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

The combination and analysis of the above data makes it possible to target relevant advertising to a specific group at a specific time.

5. When comparing campaigns to each other, we also need information on how many impressions were generated for each campaign. We get different results when the conversion is from one mailing and from several.

3. Write an SQL script which creates the above table (*campaign_comms_user_level*) using only the information from the table below (called *campaign_comms*):

date	campaign_id	user_id	event
20180722	183946	657464654	opened_email
20180801	421564	6874654651	converted
…	…	…	…

Where event can have the following values:

- 'email_sent', 'email_opened', 'clicked_on_link', 'visited_landing_page', 'converted'

Ad 3.

```
ALTER TABLE campaign_comms
     ADD opened_email varchar(20),
     ADD clicked_on_link varchar(20),
     ADD visited_landing_page varchar(20),
     ADD converted varchar(20);

UPDATE campaign_comms
     SET opened_email = 0,
          clicked_on_link = 0,
          visited_landing_page = 0,
          converted = 0
     WHERE event = 'email_sent';

UPDATE campaign_comms
     SET opened_email = 1,
          clicked_on_link = 0,
          visited_landing_page = 0,
```


Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.

```
          converted = 0
     WHERE event = 'opened_em';

UPDATE campaign_comms
     SET opened_email = 1,
          clicked_on_link = 1,
          visited_landing_page = 0,
          converted = 0
     WHERE event = 'clic_link';

UPDATE campaign_comms
     SET opened_email = 1,
          clicked_on_link = 1,
          visited_landing_page = 1,
          converted = 0
     WHERE event = 'visit_page';

UPDATE campaign_comms
     SET opened_email = 1,
          clicked_on_link = 1,
          visited_landing_page = 1,
          converted = 1
     WHERE event = 'converted';

ALTER TABLE campaign_comms
DROP COLUMN event;

ALTER TABLE campaign_comms
RENAME COLUMN date TO data_sent;
```

In my knowledge, there is no SQL command to define the column ordering.

I can explicitly list the name of columns in the order which is in table above, to be returned in the same order.

```
SELECT campaign_id, user_id, data_sent, opened_email,
clicked_on_link, visited_landing_page, converted
FROM campaign_comms;
```

Kind Regards
Dorota Gawrońska-Popa

Note: In real world we don't get well defined problems and there is no single correct answer. Our intention of this assignment is to see whether the candidates make reasonable assumptions.