Part 1 - SQL questions

1. Write a query that returns the account_id, number of orders, and total amount of all orders for each account with more than 3 orders.

Answer:

```
SELECT a.account_id, o.order_id, SUM(amount)
FROM account a, order o
WHERE a.account_id = o.account_id
AND o.order_id > 3
GROUP BY a.account_id, o.order_id, SUM(amount)
```

2. Write a query that returns the account_id, number of orders, total amount of all orders, number of loans, and total amount of all loans for each account

Answer:

```
SELECT a.account_id, o.order_id, l.loan_id, COUNT(l.payments), SUM(l.amount)
FROM account a, order o, loan l
WHERE a.account_id = o.account_id
AND l.account_id = a.account_id
GROUP BY a.account_id, o.order_id, l.loan_id, COUNT(l.payments), SUM(l.amount)
```

3. Use the same query as #2 and add the logic for only keeping the accounts that the total loan amount is lower than the total orders amount (including accounts without any loans).

Answer:

```
SELECT a.account_id, o.order_id, l.loan_id, COUNT(l.payments), SUM(l.amount), SUM(o.amount)
FROM account a, order o, loan l
WHERE a.account_id = o.account_id
AND l.account_id = a.account_id
AND SUM(l.amount) < SUM(o.amount)
GROUP BY a.account_id, o.order_id, l.loan_id, COUNT(l.payments), SUM(l.amount), SUM(o.amount)
```

4. Write a query that returns the account_id, order_id, order amount, and the share out of the total order amount for that account. For example, for account_id = 2 the total amount of all orders is 10638.7, that account has 2 orders: order_id=29402 and order_id=29403. The amount of order_id=29402 is 3372.7. The share of this order is 31.7% out of all the orders of that account. We want to calculate that share for every order.

Answer:

```
SELECT a.account_id, o.order_id, SUM(t.amount) * 100 / o.amount

FROM account a, order o, trans t

WHERE a.account_id = o.account_id

AND o.account_id = t.account_id

ORDER BY o.order_id
```

5. Write a query that returns a list of all “k_symbol” unique values from both order and trans tables.

Answer:

```
SELECT DISTINCT o.k_symbol, t.k_symbol

FROM order o, trans t

WHERE o.account_id  = t.account_id

ORDER BY 1,2
```

Part 2 - Data question

Task:

She has asked you to help in this task by calculating which customers are most similar to each other based on the countries they have factories in. Specifically, she wants to know the similarity between any two customers.

Given two customers, how would you calculate their similarity? In what situations might your approach not work well?

Answer:

Assuming that the similarity between customers based on countries' factories would be evaluated using alphanumeric, continuous and higher dimensional data, I would choose to use clustering techniques and similarity distance measures. To measure the similarity across customer data points in each country, I would choose to cluster the text data using k-means and reduce dimensionality in the data through Principal Component Analysis (PCA). To visualize the data clustering I would t-SNE, and evaluate the clusters formed across customers within and across compared countries. For similarity distance measures, I would choose to create a supervised similarity measure using Euclidean Distance. Dimensions/features are unrelated across countries, and the features need to be normalized to avoid over emphasis on one feature. In order to normalize the unrelated country features, I would use Euclidean Distance as the formula is rotation invariant.

In the situation of higher dimensional data, one might argue that Euclidean Distance might not work as effectively. In higher dimensional data, some variables tend to be correlated, thus obfuscating the process as Euclidean Distance treats all dimensions equally. However, a counter-argument to that would be that of the existence of nonuniformity in a real-world, higher dimensional data distribution. Nonuniformity contends that real data is probably not going to be distributed evenly in the higher dimensional space, but will rather hold a small, clustered subset of that dimensional space.