Global Health Informatics Institute

Data Analyst and Health Data Fellow Screening Questions

Instructions:

1. **Please attempt to answer the following questions in your preferred database language.**
2. **Save the answers in a PDF format and upload with your full name.**
3. **Prohibition of AI: The use of any form of artificial intelligence, including but not limited to chatbots, language models, or automated tools to attempt questions on the online test, is strictly prohibited. Any violation of this rule will result in disqualification from the job application process.**

Exercise One: From the table below, please write an SQL query that will display the duplicate records.

Disclaimer: the names are not real names

User Table					
user_id	**first_name**	**last_name**	**gender**	**accepted_at**	**applied_at**
1	Dziwe	Maliketi	male	2020-04-17	2020-04-11 06:27:55.000 +0200
2	Tsoka	Mpamba	male	2011-01-30	2011-01-01 18:44:05.000 +0200
3	Pamoto	Mtengo	female	2021-01-03	2021-01-03 12:47:52.000 +0200
4	Dziwe	Maliketi	male	2020-04-17	2020-04-11 06:27:55.000 +0200
5	Kuunika	Malo	female	2019-03-25	2019-12-25 15:45:38.000 +0200
6	Kaduka	Bamusi	male	2020-01-12	2021-01-03 12:37:52.000 +0200
7	Kameza	Kondani	male	2021-09-23	2021-11-01 18:24:05.000 +0200
8	Pamoto	Mtengo	female	2021-01-03	2021-01-03 12:47:52.000 +0200
9	Ngende	Chizungu	female	2019-09-23	2020-11-01 11:24:05.000 +0200
10	Kameza	Kondani	male	2021-09-23	2021-11-01 18:24:05.000 +0200

Exercise Two: The SQL query below returns the maximum visit date of all patients which tested positive for Malaria using the following tables:

Diagnosis_Stage Table	
person_id	**malaria_stage**
123	Plus One
1234	Plus Two
12345	Plus One

Patient_Visit Table			
visit_id	**person_id**	**visit_date**	**clinician_id**
1	111	2023-01-01	8
2	1234	2023-01-01	4
3	12345	2023-01-04	3
4	222	2023-01-12	8
5	456	2023-01-12	8
6	123	2023-01-13	3
7	111	2023-01-14	2
8	1234	2023-01-15	8
9	12345	2023-01-18	8

Patient_Demographic Table		
person_id	**first_name**	**last_name**
123	Nyamalikiti	Mtengo
111	Duwa	Mtedza
1234	Pepala	Mwala
12345	Basikolo	Phiko
222	Fumwe	Mphambano
456	Sipokosi	Chokha

Query:

```
SELECT pd.person_id, pd.first_name, pd.last_name, max_visits.visit_date
FROM Patient_Demographic pd
INNER JOIN
(
        SELECT pv.person_id, max(pv.visit_date) as visit_date
        FROM Patient_Visit pv
        group by pv.person_id
) as max_visits ON max_visits.person_id = p.person_id
WHERE pd.person_id in (SELECT DISTINCT ds.person_id FROM Diagnosis_Stage ds)
```

a) The query takes quite some time to return results. Mention any two ways in which the query can be restructured (optimized) to run efficiently

b) Write a query that applies the solutions mentioned in question a above

Exercise three: Data Deduplication

Using a python script do the following:

a) Read the provided CSV file "**client_purchases.csv**".

b) Find and Identify duplicate records.

c) Remove the duplicate records

d) Export the cleaned CSV file into a file named "**client_purchases_deduplicated.csv**".

e) Identify unique clients, assign them a unique ID and export them into a file named "**clients_unique.csv**"

Exercise four: Data Privacy and Security:

As a Data Analyst or Health Data Fellow you will be dealing with highly sensitive data such as PII (Personal Identifiable Information) which is defined as: Any representation of information that can be used to identify an individual whom the information applies to be reasonably inferred by either direct or indirect means.

a) From the given deduplicated CSV file mention all the variables which qualify as PII

b) Write a python script to anonymize sensitive data (PII) to ensure privacy and security compliance.

c) Export the anonymize data set in a csv file named "**clients_deidentified.csv**"

Good luck