Data Engineer take home exercise

The following exercise is sent to all Data Engineer applicants, usually after the hiring manager phone screen step. The exercise is pretty simple and straightforward, so the goal is to evaluate the level of attention to detail, and whether candidates put extra effort in their submission, or just submit working code.

Story

ABC, Inc. has started a subscription business and immediately started onboarding users.

Enrollment business rules:

- Users must be 18 years or older upon account creation.
- Users must provide valid identifiers (email, phone number, birthdate) during enrollment.
- Users should never be removed (i.e. rows deleted), they should be marked as cancelled.
- Attached is a DB schema for storing the input, the data should conform completely to the schema.

Task

When the business started there wasn't a data type enforced schema structure for the table to store users info. The business rules existed before but weren't enforced at a database level while storing data to the table. Now the database team wants to properly apply types to the fields. All the new data coming will adhere to these standards but we need to migrate existing data from old to new table. When casually looking through some of that users data, some issues were identified. Now we'd like to have a comprehensive analysis done on the data to find any anomalies.

```
create table users_old (
    id int,
   first_name varchar(255),
    last_name varchar(255),
    email varchar(255),
    phone varchar(100),
    status varchar(255),
    birth_date varchar(100)
    created_at varchar(100)
);
create table users_new (
    id int not null auto_increment,
    first_name varchar(255) not null,
    last_name varchar(255) not null,
    email varchar(255) not null,
    phone int(10) not null,
    status enum('active', 'cancelled') not null,
    birth_date date not null,
    created_at datetime not null
);
```

You are allowed to use any AI tool you have at your disposal. In fact, AI usage will earn you more points.

Your task is to use Python to identify the data that violates any business rule and report the anomalies. You should provide the code used to find the anomalies, and you may report the anomalies in any way that makes sense to you. This is similar to building a change detection system, but hopefully small enough that it isn't overly burdensome. The final solution should, at a minimum, validate data meets reasonable row- and field-level validation for all four business rules above and, more importantly, include your entire AI interaction.

Your code should be able to be compiled (if necessary) and run by the reviewer. To that end, please include a README file or easy to follow instructions on how to resolve dependencies.

Attached is the csv file which is exported from the old users table.

challenge_dataset.csv
14 Apr 2021, 06:22 PM