DATA SCIENCE ENGINEER INTERN ASSESSMENT(TG TEAM)

For this assessment, the candidate is expected to **ATTEMPT** question 1 and answer any other question from the second section.

SECTION A (Data Ingestion and API Communication) - Python

1. We want to ingest github data of techies or github users in the Machine learning Field in the following locations on github. (**Lagos, Nigeria, Rwanda**). It is expected that you use the **github search for users API** for this task. The final result for this task should be a dataframe that contains details of Github Users who do **Machine Learning** from the three location tags indicated above. A sample search Keyword for users would be like **"Machine Learning location: Lagos**"
    a. The expected data points to be ingested or seen in the data frame include name, login/username, Bio, Blog, Company, email, followers(count of followers), following(count of following), id, url, and location.)

HINTS - Check out using **Pygithub** or **Github3.py** Python Packages for this task which can be installed using pip. As these packages can be used to interact with Github APIs directly.

SECTION B (Data Analytics - ML & Business Analytics) - Python or SQL

Answer any one of these questions. (You can choose only one out of questions 2 and 3)

The data to be used for these questions can be found here .

2. With the data above, perform clustering analysis or customer segmentation. Create a Report of your approach and explain the different segments or clusters you have business wise.

3. Perform Credit Scoring using the data above, it can be rule based where you award points based on certain conditions to different users

or you can take a machine learning approach but note there is no column indication loan default so you are free to come up with your own criterion on choosing who is likely to default or not.
Hint: You can also go from unsupervised to supervised by first creating clusters and getting the cluster or segment of users likely to default. From which you can now take a Supervised classification approach to generate credit scores.

Data Dictionary -

- User_id - ID of a customer or user
- Channel - Channel used for last transaction by the customer
- TotalTransactionAmount - Total amount transacted by the customer in the last few months
- Tx_Count - Count of transactions done by the customer on the system
- DaysSinceLastTrans - Number of days since the customer did his last transaction with reference to a specific date
- Reseller_id - Id of the reseller that the customer belongs to.
- Superdealer_id - ID of the SuperDealer that the customer belongs to.
- Product_ServiceProvider - Service Provider of the Last Product purchased by the customer.
- Product_Type - The Product type of the last product purchased by the customer
- Product_Category - Category for the last Product purchased by the Customer.