**Data Analysis and API Development**

**Objective**:

The objective of this assignment was to assess coding proficiency and data analysis skills. The assignment was divided into three sections, each contributing to a holistic understanding of data engineering processes.

**Section 1: Environment Setup and Data Cleaning:**

In this section, the project began with setting up the environment in Google Colab, where a large dataset was provided. The initial steps involved data cleaning and preprocessing to ensure data integrity. Basic preprocessing steps included dropping unnecessary columns and handling null values. Two separate DataFrames, 'Property Price Details' and 'Property Details,' were created as per the provided instructions.

One of the challenges faced during this section was exporting these DataFrames to a local MySQL Workbench. After multiple unsuccessful attempts, it was discovered that SQLAlchemy and Ngrok could effectively map the local host to the web. This solution proved pivotal in overcoming the challenge.

**Section 2: Data Analysis:**

With the DataFrames successfully exported locally, Section 2 focused on data analysis. Ten specific questions were provided to be answered using the dataset. Two different approaches were adopted for analysis. First, a SQL script was created locally, and the analysis was conducted within MySQL Workbench itself. This approach provided a structured and familiar environment for SQL-based analysis.

The second approach involved running the analysis queries directly within Google Colab. This required establishing a connection between Colab and the local database. The dataset was analysed by executing SQL queries in Colab, which provided a more dynamic and interactive analysis environment.

**Section 3: Exposing Results in API:**

The final section aimed to expose the analysis results through APIs using Flask. Two distinct API methods were implemented. The first method utilised the Get HTTP method and was developed on a local Jupyter notebook. This method successfully created separate API endpoints for each of the ten questions, providing access to the analysis results.

The second method involved using Ngrok to expose the APIs. This process required extensive learning, multiple iterations, and the resolution of various challenges. Ultimately, APIs with ten separate endpoints were created, offering access to the analysis results. Additionally, basic error handling was implemented to gracefully handle common errors.

**Pyspark Exploration**:

In addition to the main assignment, Pyspark was explored to analyse the dataset further. This provided an opportunity to gain insights from a different perspective and enhance the overall data analysis skill set.

**Key Takeaways:**

* This assignment provided invaluable hands-on experience in data engineering and analysis.
* Learning how to establish connections between Colab and local databases was a significant achievement.
* Developing APIs for data exposure showcased practical application and understanding of industry-standard practices.
* Exploring Pyspark extended the toolkit for future data analysis tasks.
* Grateful for the opportunity provided by DataGrokr, this project contributed to a deeper understanding of real-world data engineering challenges and solutions.
* This project not only tested coding proficiency but also provided me a platform to acquire industry-relevant skills and insights, making it a valuable learning experience.