Access the dataset [here](https://docs.google.com/spreadsheets/d/1NGGpjVinFZZIWV7NY95g-A8H2RNoGy5YXIxmet4aKQ4/edit?usp=sharing)

### **Dataset Overview**

You will all be working with the same dataset collected from a previous mentorship cohort.  
 Columns include:

* Timestamp
* ID No.
* Age range
* Gender
* Country
* Where did you hear about Everything Data?
* Years of learning experience in the data field
* Track applied for
* Hours per week available
* Main aim for joining
* Motivation for joining
* Self-assessed skill level in chosen track
* Aptitude test completion status
* Total score
* Graduation status

### 

### **Project Goal**

Analyze, transform, and draw insights from the dataset to:

1. Understand participant demographics, motivations, and performance.
2. Highlight factors that influence graduation rates.
3. Present actionable recommendations for improving future mentorship cohorts.

## **Track-Specific Deliverables**

### **1. Data Analysis Track**

* Data Cleaning: Use Excel or SQL to remove duplicates, handle missing values, and standardize categorical data.
* Exploratory Analysis: Write SQL queries to answer key questions (e.g., average score by track, top countries represented).
* Visualization: Create an interactive Power BI dashboard showing:  
  + Demographics (age, gender, country)
  + Track distribution
  + Relationship between learning hours, experience, and scores
  + Graduation trends
* Insight Report: Write a short summary (max 500 words) of your findings.

### 

### 

### **2. Data Science Track**

* Data Cleaning & Preprocessing: Use Python (Pandas, NumPy) to prepare the dataset.
* Exploratory Data Analysis: Visualize distributions and correlations (Matplotlib/Seaborn).
* Predictive Modeling: Build a simple classification model to predict Graduation status using features like experience, hours/week, and total score. Compare at least two algorithms.
* Model Evaluation: Report accuracy, precision, recall, and F1-score.
* Recommendations: Suggest data-driven ways to improve graduation rates.

### **3. Data Engineering Track**

* Database Design: Create a database schema to store the dataset. Write SQL scripts to insert and query the data.
* ETL Pipeline: Build a Python-based ETL script to:  
  + Extract the dataset (CSV)
  + Transform (clean, standardize formats)
  + Load into your database or a cloud data warehouse (e.g., AWS RDS, Azure SQL).
* Automation: Schedule the ETL job using a cloud service or local scheduler.
* Data Access Layer: Provide SQL queries or an API endpoint for others to access cleaned data.

### 

### 

### **Evaluation Criteria**

* Data Accuracy & Cleaning Quality – 25%
* Completeness of Deliverables – 25%
* Depth of Insights / Model Performance – 25%
* Clarity of Presentation / Dashboard – 25%

### **Submission Guidelines**

* Submit all code, dashboards, and reports in a shared folder. Share with everythndata@gmail.com
* Include a README file explaining your workflow.
* Deadline: Wednesday September 12, 2025 - Presentations Day
* File naming: <Name>\_<Track>\_Project