Final Project - Proposal

1) Project Description:

- This project focuses on building an end-to-end data engineering solution for predicting and analyzing customer churn. The dataset from Kaggle contains comprehensive information on bank customers, including demographic, financial, and behavioural data (e.g., credit score, account balance, product usage), with a target variable indicating whether a customer has churned. The primary goal is to engineer a robust, automated data pipeline that ingests the raw dataset, performs rigorous cleaning and transformation, and loads the processed data into a structured SQL database. The final output will be an interactive Power BI dashboard that provides actionable insights and key performance indicators (KPIs) on churn drivers, enabling data-driven decision-making to reduce customer attrition. The pipeline will also perform feature engineering to prepare the data for future predictive modeling.

2) Group Members & Roles:

Name	ID	Email	Assigned roles
Jana Mohamed El-Sayed Mohamed	21032500	Janamohamed7178@gmail.com	M1, M3
Steven Tamer Soliman	21081767	Steven.tamer.s@gmail.com	M1, M2
Mohab Sherif Mohamed	21039536	Mohab.sherif04@gmail.com	M1, M2
Arwa Mohamed	21098099	arwa3169@gmail.com	M1, M3
Lamis Abdallah Essmat Abdelhamid	21044350	lamisabdallah2811@gmail.com	M1, M2
Malak Mahmoud Shehata Mahmoud	21086742	malaksoliman73@gmail.com	M1, M3

3) Team Leader: Jana Mohamed

4) Tools & Technologies:

- **Programming Languages:** Python (Pandas, NumPy, Matplotlib)
- **Development Environment:** Google Colab
- **Version Control:** GitHub

5) Objectives:

- **Data Pipeline Automation:** To design and implement a robust, automated ETL pipeline that ingests the raw CSV data, performs cleansing and transformations, and loads the processed data into a structured data warehouse.
- **Data Quality & Integrity:** To establish and execute data validation rules and quality checks throughout the pipeline to ensure the accuracy, consistency, and reliability of the output data.
- **Database Engineering & Optimization:** To design a normalized relational database schema in SQL and optimize it for efficient data storage, complex joins, and fast analytical query performance.
- **Feature Engineering for Analytics:** To create new derived features (e.g., customer tenure groups, total product count) that enhance the dataset for downstream churn analysis and reporting.
- **Analytical Dashboard Development:** To build an interactive Power BI dashboard that provides actionable insights into customer churn metrics, trends, and key business drivers.
- **End-to-End Documentation:** To thoroughly document the entire data pipeline architecture, database schema, ETL processes, and dashboard definitions for maintainability and knowledge transfer.

6) KPIs (Key Performance Indicators):

A. Data Preprocessing (Python script, cleaned CSV):

- **% of missing/duplicate data correctly handled: Target → 100%**
- **Script efficiency: Data cleaning script runs within expected time (<10 seconds) for the dataset.**
- **Data consistency and validity checks passed: Target → ≥95%**

B. Feature Engineering & Analysis:

- **Number of meaningful derived features created: Target → ≥3 (e.g., tenure groups, total products).**
- **Feature relevance: New features improve interpretability or reveal clearer churn patterns (qualitative evaluation).**

C. Visualization (Python - Matplotlib/Seaborn):

- **% of main insights/metrics visualized: Target → ≥90% (e.g., churn distribution by age, balance, credit score, etc.)**
- **Visualization clarity and accuracy: Target → ≥95% (plots are properly labeled, easy to interpret).**
- **Execution performance: All visualization scripts run smoothly without errors.**

D. Presentation (Report, slide deck):

- **Report completeness: Target → 100% (includes data overview, preprocessing, analysis, and visuals).**
- **Presentation clarity and engagement: Target → ≥4/5 average feedback score from peers or instructor.**

7) Milestones & Deadlines:

Milestone 1: Data Collection, Exploration, & Preprocessing		**1-Week 31/Auc – 5/Sep**
Data Collection:	Jana Mohamed El-Sayed Mohamed	**2-days 31/Auc–1/Sep**
	Malak Mahmoud Shehata Mahmoud	
Data Exploration	Steven Tamer Soliman	**2-days 1/Sep – 3/Sep**
	Mohab Sherif Mohamed	
Data Preprocessing	Arwa Mohamed	**3-days 3/Sep – 5/Sep**
	Lamis Abdallah Essmat Abdelhamid	
Milestone 2: Predictive Model Development		**12-Days 27/Sep – 9/Oct**
Model Selection	Lamis Abdallah Essmat Abdelhamid	**5-days 27/Sep – 2/Oct**
Training & Evaluation	Steven Tamer Soliman	**7-days 2/Oct – 9/Oct**
Tuning & Interpretation	Mohab Sherif Mohamed	
Milestone 3: Deployment & Retention Strategy		**2-Weeks 9/Oct – 23/Oct**
Deployment	Arwa Mohamed	**1-week 9/Oct – 16/Oct**
Retention Strategy	Jana Mohamed El-Sayed Mohamed	**4-days 16/Oct–20/Oct**
	Malak Mahmoud Shehata Mahmoud	**3-days 20/Oct–23/Oct**