Team:

- Group Name: **TITANS**
- Members:
    - **Name**: Cynthia Onyia
    - **Email**: cyn4jun@gmail.com
    - **Country**: United Kingdom
    - **College/Company**: University of Wolverhampton
    - **Specialisation**: Data Science

    - **Name**: Christopher Avbenake
    - **Email**: kristoville@gmail.com
    - **Country**: United Kingdom
    - **College/Company**: Retail Assets Solution
    - **Specialisation**: Data Science

Problem Description:

Business Problem - Pharmaceutical companies face challenges in understanding if patients will continue their prescribed therapies (which is the persistency_flag in the dataset) or not. ABC Pharma likely wants to automate the identification of this persistence to optimise their marketing strategies, improve patients' adherence, and potentially increase drug sales. They want to identify what factors impact a patient's persistence in continuing their prescribed drug

Business Value: This helps the company reduce drug drop-offs and improve adherence strategies.

Machine Learning Objective - Using a classification model, the prediction of the target variable 'Persistency_Flag' can be achieved, thereby identifying the likelihood of a patient adhering to prescribed medication based on clinical, demographic, and treatment factors through the dataset provided.

In a nutshell, the goal is to use machine learning to build a classification model that predicts whether a patient will be Persistent or Non-Persistent with their treatment.

Data Understanding:

The following are information derived from the dataset provided

- Rows: 3424
- Columns: 69
- Target Variable: Persistency_Flag (Yes/No or similar binary classification)

- **Key sample columns:**
    - Ptid – patient ID
    - Gender, Race, Region, Age_Bucket
    - Ntm_Speciality, Ntm_Specialist_Flag, etc.
    - Multiple risk factor flags (Risk_*) and Count_Of_Risks
    - **Target**: Persistency_Flag

All these are a mix of **categorical**, **ordinal**, and **numerical** features.

- **Types of Variables:**
    - Categorical: 67 columns (e.g., Gender, Race, Comorbidities, Risk Factors)
    - Numeric: 2 columns:
        - Dexa_Freq_During_Rx
        - Count_Of_Risks
- **Data Type Summary:**
    - 67 object (categorical)
    - 2 numeric (int64)

Problems in the Data:

1. Missing Values (NA) – There are no missing values detected in any of the column, showing a sign of a good dataset for modelling.

2. Skewed Data

- Dexa_Freq_During_Rx: Skewness = 6.81 → Highly right-skewed (most patients had very few DEXA scans; few had many).
- Count_Of_Risks: Not heavily skewed.

3. Outliers

- Dexa_Freq_During_Rx: 460 outliers (patients with unusually high scan frequency)
- Count_Of_Risks**: 8 outliers** (patients with very high number of risk factors)

Problems and How to Approach them:

Problem	**Handling Technique**	**Why**
Skewness (Dexa_Freq_During_Rx)	Log/Box-Cox transformation	Reduces skewness and helps models (like logistic regression) perform better
Outliers	Capping (e.g., 95th percentile), or treat based on domain knowledge	Prevents extreme values from disproportionately affecting model training
Categorical Variables	Encoding: One-Hot or Label Encoding	Prepares data for ML models that need numeric input
Feature Scaling **(if needed)**	StandardScaler or MinMaxScaler	Important for models like SVM, KNN, etc., but less so for trees
Imbalanced Data **(if Persistency_Flag is skewed)**	Use SMOTE or class weights	Ensures model doesn't just predict the dominant class
Skewness (Dexa_Freq_During_Rx)	Log/Box-Cox transformation	Reduces skewness and helps models (like logistic regression) perform better
Outliers	Capping (e.g., 95th percentile), or treat based on domain knowledge	Prevents extreme values from disproportionately affecting model training