**Data description:**

The initial dataset is an excel spreadsheet which follows the .xlsx format containing 355 rows and 30 columns. The initial dataset is of size 355 by 30 containing data from the HIGHER EDUCATION LOANS and GRANTS BOARD (HELGB) as well as loan beneficiary responses to loan application and repayment process related questions. The majority of features contain categorical data as marked by more than 25 features containing qualitative values. The dataset to be used for modelling however will contain synthesized data created from the initial dataset using oversampling, generative modelling and random sampling. The dataset will potentially comprise of not less than 1000 rows and it might contain less or more features depending on the results of preprocessing and feature engineering as well as feature selection. In the initial dataset, the feature “repayment\_status” was regarded as the target variable. It represents whether a beneficiary repaid the loan or not. The dataset also contains various predictor variables for example, age\_range and total\_loan which represent various age ranges of beneficiaries as well as the amount of loan applied for respectively.

**Exploratory data analysis:**

1. Univariate analysis:
   1. Frequency distribution tables will be produced for all the features. These tables will be used to check the distributions of all the individual features including missing values.
   2. Bar charts as well as histograms will produced to visualize distributions. These will also be used to detect outliers based on occurrence of values.
   3. Special bar charts visualized in terms of the loan repayment status will be produced to detect outliers.
2. Multivariate analysis:
   1. Cross tabulations. Cross tabulations between the features and the target variables will be created to analyze the relationships between various features and the predictor variable.
   2. Grouping data according to different features to check weighted statistics and distributions. For example, grouping data according to gender, marital status and the loan repayment status.
   3. Scatterplots. Different features will be compared to the target variable using scatterplots.

**NOTE**: more visual analysis will be conducted after data preprocessing since it is easier to visualize numbers than is to get the same level of visual meaning from categorical data.

**Data cleaning and preprocessing:**

1. Converting all column names to lowercase. This will be done to ease further preprocessing and reduce chances of errors due to casing.
2. Converting all the values in the dataset to lowercase. This will be performed to ease further analysis as well as reduce access errors due to mismatching casing.
3. Checking datatypes and converting them where necessary. Datatypes of the values of different features will be checked for uniformity and will be converted to follow a uniform type for the particular feature.
4. Handling outliers. The outliers identified in EDA will be handled using appropriate techniques as follows:
   1. Dropping.
   2. Replacing.
5. Identifying and handling missing values. the dataset will be checked for missing values, and if any are identified they will be handled as follows:
   1. Dropping rows
   2. Dropping columns if missing values exceed a certain threshold.
   3. Imputing using appropriate techniques including using the mode.
6. Feature transformation. Certain features will be changed in order to provide more meaning and relevance for example:
   1. Program of study will be changed to discipline whereby programs conforming to a particular discipline will be labeled as that discipline for instance pharmacy and mbbs will both be registered as medicine as a discipline.
   2. Age range will be converted to age group. the age group will contain the specific intervals of ages of the beneficiaries for instance; ‘38 to 48’ will be converted to ’38 – 40’.
   3. Values under the total loan feature will be converted to sensible intervals and will be stripped of the currency which will be annotated together with the feature itself, thus, loan\_amount (MWK).
7. Dropping features which will be intuitively considered irrelevant. For example, repayment method recommendation which would be irrelevant to the final model.
8. Feature encoding. All the columns will be converted to floating point numbers and integer types using appropriate encoding techniques which are as follows:
   1. Label encoding. For ordinal types.
   2. Mean target encoding. For both ordinal and nominal types of data.
   3. One-hot encoding. For the categories which follow hierarchy.