Parachute Steps:

1. **Data Source:**

* **File Path:** 'Input/Data 3 - October, 2024.xlsx'
* **Sheets Loaded:**

1. Monthly Wellness and Engagement'
2. 'Parachute - Cross Section'
3. 'TU - Panel'
4. **Merging Tabs**

* **Monthly Engagement Data:** Loaded from 'Monthly Wellness and Engagement' sheet.
* **Parachute Cross Section Data:** Loaded from 'Parachute - Cross Section' sheet.
* **TU - Panel Data:** Loaded from 'TU - Panel' sheet.
  + **FICO Scores:** Aggregated by 'loan\_id' to compute the mean FICO score per loan.

1. **Missing Values Assessment and Handling**

* **Function:** check\_missing\_values prints the count of missing values per column for each dataset.  
  - Monthly Engagement  
  - Parachute Cross Section  
  - TU – Panel
* **Handling Missing Values:**
  + **Numerical Columns:** 'wellness\_score' filled with the mean value of the column.
  + **Categorical Columns:** Filled with 'Unknown' to maintain category consistency.

1. **Data Type Conversion and Renaming**

* Date Columns Conversion:
  + 'month' in Monthly Engagement
  + 'Disbursement Date' and 'Date of Birth' in Parachute Cross Section
  + Converted to datetime format with error coercion to handle invalid formats.
* ID Columns Renaming: Standardized all ID columns to 'customer\_id' across datasets for consistency.
* **Data Type Adjustment:** Ensured 'customer\_id' is of type string across all datasets.

1. **Duplicate Removal**: Removed duplicate records from all datasets to ensure data integrity.
2. **Merging Datasets**
   1. Monthly Engagement with Parachute Cross Section on 'customer\_id' (Left Join).
   2. Resultant dataset merged with TU - Panel on 'customer\_id' (Inner Join).
   3. **Result:** Consolidated dataset full\_data containing relevant information from all sources.
3. **Feature Engineering**
   1. **Age Calculation:** 'age\_at\_disbursement'
      1. Difference between 'Disbursement Date' and 'Date of Birth' in years.
   2. **Debt-to-Income Ratio:** 'debt\_to\_income\_ratio' = 'Outstanding Balance' / 'Qualified / Verified\nIncome'
   3. **Loan-to-Income Ratio:** 'Loan\_to\_Income\_Ratio' = 'Loan Amount' / 'Qualified / Verified\nIncome'
   4. **Monthly Repayment Burden:** 'Monthly\_Repayment\_Burden' = 'Outstanding Balance' / 'Loan Term (Months)'
   5. **'engagement\_level'** **Categorization:** Based on 'activities\_count'
      1. **Low:** ≤ 5 activities
      2. **Medium:** 6-15 activities
      3. **High:** > 15 activities
4. **Encoding Categorical Variables**One-Hot Encoding using pd.get\_dummies  
   **Columns Encoded:** All categorical columns identified in Monthly Engagement data. For machine learning algorithms.
5. **Handling Missing Values in Engineered Features**
   1. 'dti\_missing': Indicates missing values in 'debt\_to\_income\_ratio'
   2. **Imputation Method:** K-Nearest Neighbors Imputer (KNNImputer) - To predict and fill missing values based on similar data points.
   3. **Features Imputed:**
      1. **‘**debt\_to\_income\_ratio'
      2. 'Outstanding Balance'
      3. 'Qualified / Verified\nIncome'
      4. 'age\_at\_disbursement'
      5. 'Financial Literacy Score'
      6. 'Aptitude for change Score'
6. full\_data\_encoded full\_data\_encoded