# Bank Loan Case Study

## Project Description:

As a Data Analyst at a finance company specializing in lending to urban customers, I undertook a critical case study aimed at addressing a core business challenge: identifying reliable loan applicants in the absence of complete credit history. The company often encounters cases where customers with insufficient or no credit history are either wrongly rejected or, worse, accepted and then default on their loans – leading to significant financial losses.

To tackle this issue, I conducted **Exploratory Data Analysis (EDA)** on a real – world loan dataset to uncover hidden patterns and trends in customer behaviour and financial profiles. The objective was twofold:

**Minimized False Negative** – Ensure that potentially trustworthy borrowers without a full credit history are not unjustly denied loans.

**Detect Risky Profiles Early –** Identify common traits among defaulters to reduce loan default rates.

Key steps in the project included:

- Cleaning and pre-processing data to handle missing and inconsistent entries.
- Segmenting applicants based on credit history, income levels, loan amounts, and demographic features.
- Identifying strong correlations between applicant attributes and loan default risk.
- Visualizing patterns using charts, heat maps, and distributions to support actionable insights.

Through this analysis, the project aims to empower the risk assessment and underwriting teams to make more data-driven decisions, leading to better credit risk management and increased loan approval accuracy.

## Task A: Identify Missing Data and Deal with it Appropriately

### Objective:

To **identify variables with missing data** in the dataset and **apply appropriate data-cleaning techniques** (like imputation, deletion, or substitution) using

Excel's built-in functions (such as IF, ISBLANK, AVERAGE, MEDIAN, MODE, etc.). The goal is to ensure the dataset is clean, consistent, and reliable for further analysis or modelling.

**Key Insight:**

1. **High Missingness in Specific Variables:**
    - COMMONAREA_AVG, COMMONAREA_MODE, NONLIVINGAREA_MODE, and similar variables have **over 200% blank values,** indicating multiple records may be missing across different related fields.
    - Such high levels of missingness suggest poor data capture or irrelevance of these variables to many observations.
2. **Moderate Missingness (50% - 150%):**
    - Features like YEARS_BUILD_MODE, OWN_CAR_AGE, LANDAREA_MODE, BASEMENTAREA_MEDI, etc., fall in this range. These may still be recoverable via statistical imputations (e.g., mean/median/mode imputation).
3. **Low to No Missingness (<10%):**
    - Many important demographic and financial features (e.g., TARGET, CNT_FAM_MEMBERS, AMT_INCOME_TOTAL, CODE_GENDER) have very few or no missing values. These are **reliable for analysis.**
4. **Cumulative Line Insight:**
    - The orange line shows that a small set of variables contributes the majority of the missing data. This implies that **handling a few variables** can significantly improve overall data quality.

**Business Impact:**

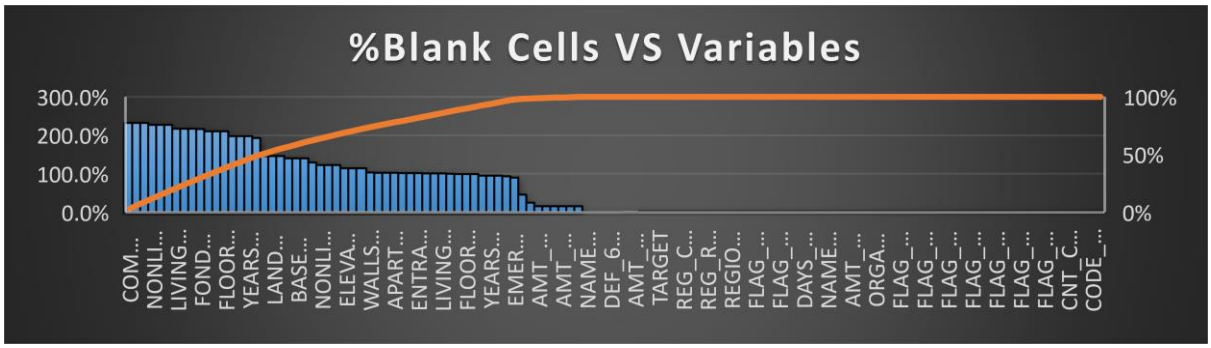
1. **Improved Model Accuracy:**
    - Cleaning high-missing-value variables prevents skewed models and **increases predictive accuracy** in risk scoring, customer profiling, or loan default prediction.
2. **Efficient Feature Selection:**
    - Helps identify variables that may be **dropped without losing much information,** especially if they're irrelevant or redundant, improving computational performance and interpretability.

3. **Strategic Data Collection:**
    - Repeated missingness in certain fields highlights **areas for improving data collection processes,** possibly leading to cost savings and better customer insights in the future.
4. **Maintaining Regulatory and Analytical Integrity:**
    - For industries like finance or healthcare, proper handling of missing data ensures compliance with audit and governance standards, avoiding incorrect conclusions that could impact business decisions.

**Recommendation:**

- Use =ISBLANK() to detect missing values.
- Apply conditional formatting to highlight them.
- Use = IF(ISBLANK(cell), AVERAGE(range), cell) or similar logic to impute values.
- Consider removing columns with over 35% missing if they are not critical.s

**Graph of Percentage of Blank Cells VS Variables:**




**Task B: Identify Outliers in the Dataset**

**Objective:** The objective of this analysis is to **detect and interpret outliers** in key numerical variables of the dataset using **box plots and statistical functions in Excel.** Outliers, which are data points that deviate significantly from the rest of the dataset, can distort summary statistics and reduce the accuracy of predictive models. Identifying these outliers is crucial for **ensuring data quality,** improving **model reliability,** and supporting **robust business decision-making.**

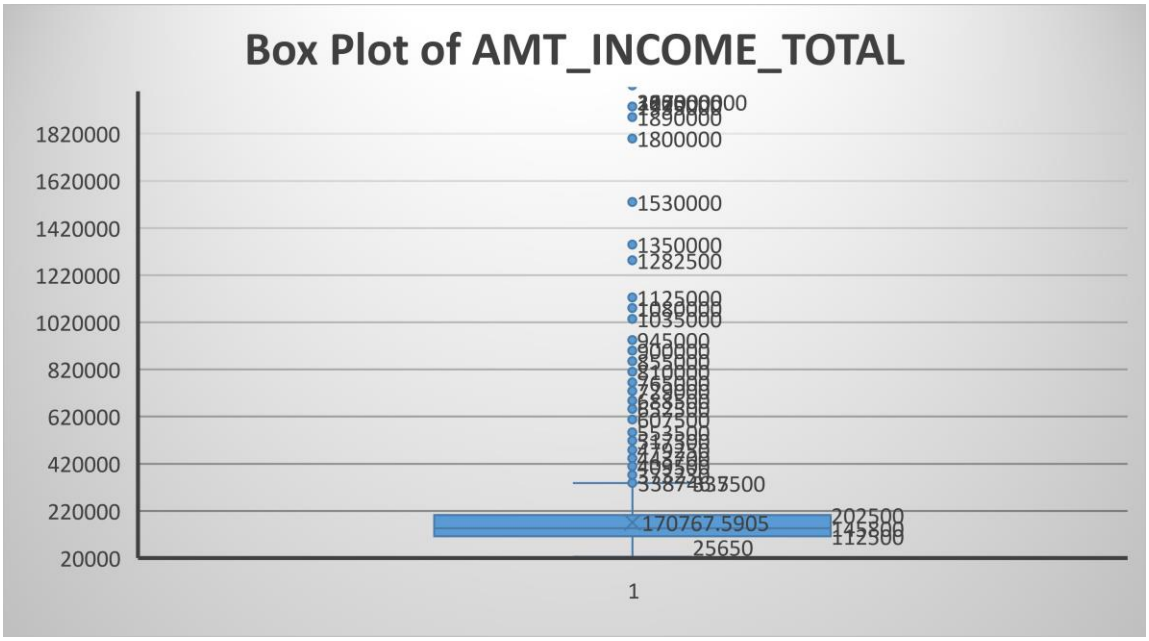
**Key Insights:**

- **AMT_INCOME_TOTAL:** Outliers observed above 10,00,000; unusually high income values detected.
- **Credit Amount:** Several outliers detected above 20,00,000; a few values exceed 3,00,000.
- **Annuity Amount:** Outliers found above 1,00,000, which are significantly higher than the typical range.
- **Children Count:** Most values range from 0 to 2, but outliers are seen up to 11 children.
- **Goods Price:** Clear outliers found above 20,00,000.
- **Family Members:** Typical family size ranges from 1 to 4; outliers like 10, 13 are present.
- **Days Employed:** Some extremely high values (positive and negative) identified as anomalies.
- **Registration Duration (Years):** Outliers found above 50 years, which is unlikely.
- **Last Phone Change (Years):** Most values are within 1-5 years; some outliers exceed 10 years.
- **EXT_SOURCE_3:** Few mild outliers below 0.1; otherwise distribution is largely normal.

**Business Impact:**

- **Model Performance:** Outliers can skew algorithms like Linear Regression, Decision Trees, or Clustering models.
- **Risk Scoring Accuracy:** Extremely high or low income/credit data may mislead loan approval or rejection models.
- **Customer Profiling:** Outliers can distort segmentation, leading to poor targeting in marketing or product design.
- **Operational Decision – making:** Inaccurate insights from outlier-heavy data may result in flawed business strategies.
- **Fraud and Data Errors:** Unusual values may indicate **data entry errors** or **potential fraud,** which need correction.

## Box Plot of AMT_INCOME_TOTAL




## Task C: Analyse Data Imbalance

**Objective:** To assess the distribution of the target variable (Defaulter VS. Non-defaulter) in the loan application dataset and determine the extent of data imbalance using Excel functions and visual representation.

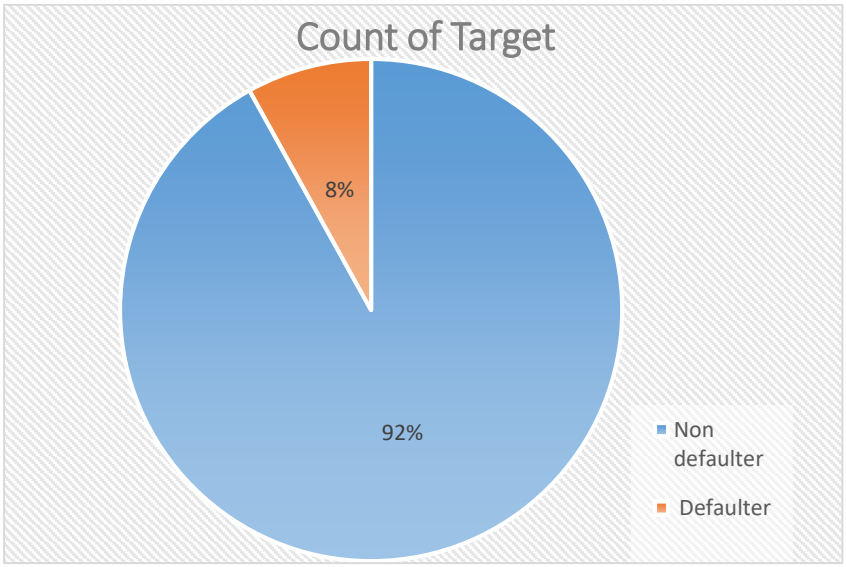
**Key Insight:**

- The target variable has **two categories:**
  - **Non-defaulters (0):** 45,973 records - > 92%
  - **Defaulters (1):** 4,026 records - > 8%
- The dataset is **highly imbalanced,** with a majority class (non-defaulters) significantly outnumbering the minority class (defaulters).
- This **8:92 ratio** indicates a **class imbalance problem,** which can potentially lead to biased predictive models that favour the majority class.

**Business Impact:**

- **Risk of Misclassification:** In credit risk modelling, under-representation of defaulters can lead to **poor detection of risky applicants,** increasing the chances of loan defaults and financial losses.
- **Model Performance Degradation:** Machine learning models trained on imbalanced data may show **high accuracy** but **low recall/precision for defaulters,** reducing their real-world effectiveness.
- **Strategic Actions Required:**
    - Apply **data balancing techniques** (e.g., SMOTE, under-sampling, class weighing).
    - Monitor performance using metrics like **ROC-AUC, Precision-Recall Curve, F1 Score,** rather tha just accuracy.
- Ensuring balanced modelling leads to **improved decision-making, better credit risk management,** and **reduced non-performing assests (NPAs).**

## Count of Target




## Task D: Analyse Key Drivers of Loan Default

**Objective:**

To analyse the loan applicant dataset using:

- **Univariate Analysis** for understanding the distribution and summary statistics of individual features.
- **Segmented Univariate Analysis** to compare how variables differ between defaulters and non-defaulters.
- **Bivariate Analysis** to explore relationships between categorical variables and the target (default status), identifying key risk indicators.

**Key Insights:**

**Univariate Analysis:**

- **Income and Credit:** The average income is approximately 1,70,677 with a high standard deviation, indicating **wide income variability.** The median credit amount is also substantial (~5,13,775), showing most applicants request large loans.
- **Education Level:** Most applicants fall under **'Secondary/Special Secondary'** and **'Higher Education'** categories.
- **Income Type:** Majority of applicants are **'Working'** class, followed by **'Commercial Associate'** and **'Pensioner'.**
- **Occupation Type:** Most common occupations are **'Labourers'** and **'Sales staff'**, which may be relevant in risk modelling.

**Segmented Univariate Analysis:**

- **Income and Credit Patterns:** Non-defaulters show higher average **income** and **credit** compared to defaulters.
- **Children Count & Age:** Defaulters tend to be younger with slightly more children, suggesting a **potential correlation between age, family size, and credit risk.**

- **Education vs Default:** Higher default rates are seen in applicants with **Secondary or Lower Secondary** education, whereas **Academic Degree holders** show very low default rates.

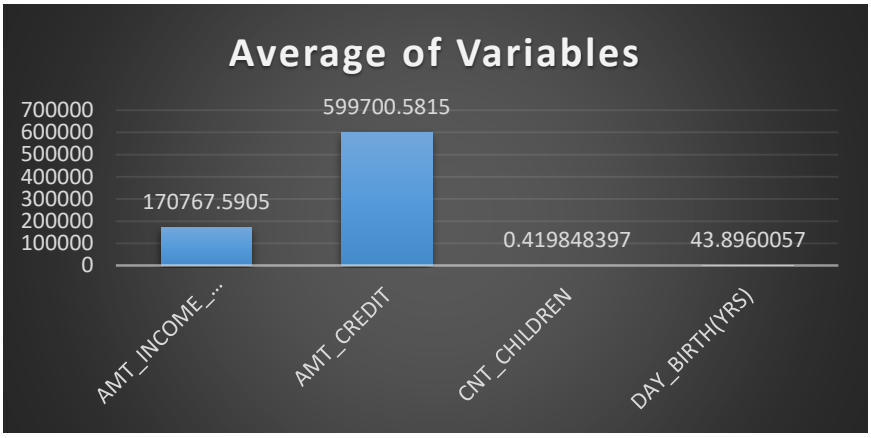
**Bivariate Analysis:**

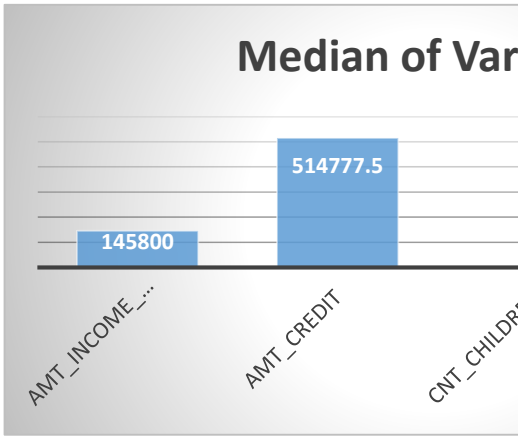
- **Income Type vs Default:**
  - **Unemployed** applicants have the **highest default rate** (~35.38%).
  - **Students** show **0% default** rate.
  - **Businessmen and State Servants** show **no defaults,** indicating low risk.
  - **Working class,** despite being the largest group, has a **default rate of ~9.47%,** which is close to the average but relevant due to volume.

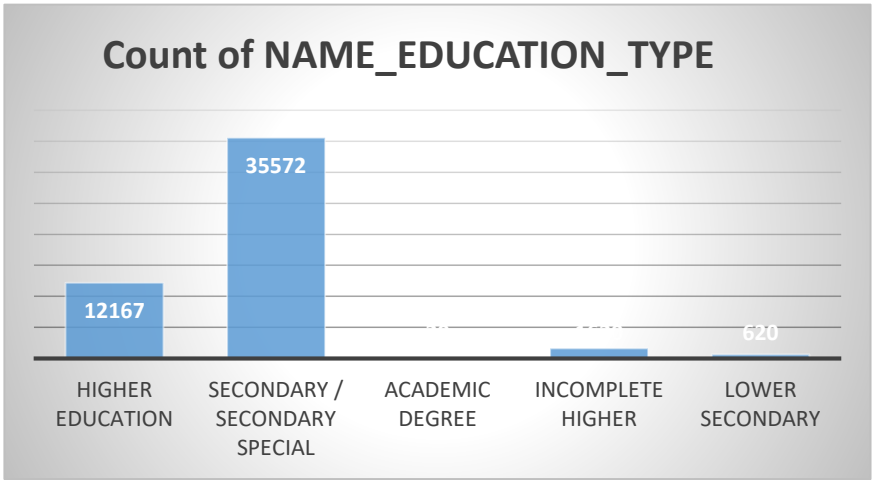
**Business Impact:**

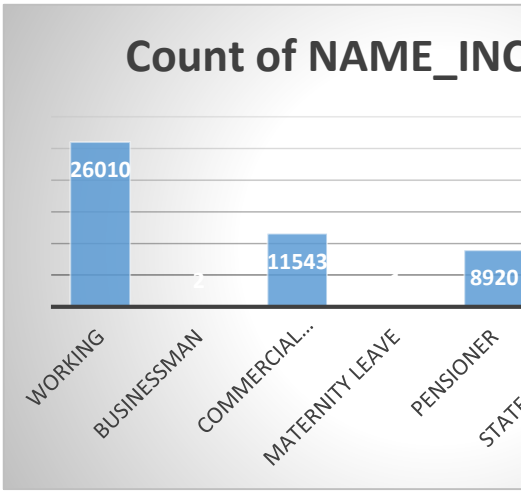
- **Risk-Based Segmentation:** The analysis helps identify **high-risk applicant segments** (e.g., unemployed, low education level) that need **stricter loan approval policies** or **additional verification checks.**
- **Improved Credit Strategy:** Lenders can **customize credit policies** for safer segments like students, pensioners, and state servants, while being cautious with groups showing high risk.
- **Data-Driven Underwriting:** Insights can directly support **better underwriting rules,** enhancing loan portfolio health and reducing default rates.
- **Personalized Loan Products:** Based on income and education profiles, financial institutions can design **custom loan products** suited to specific demographics.

# Univariate An

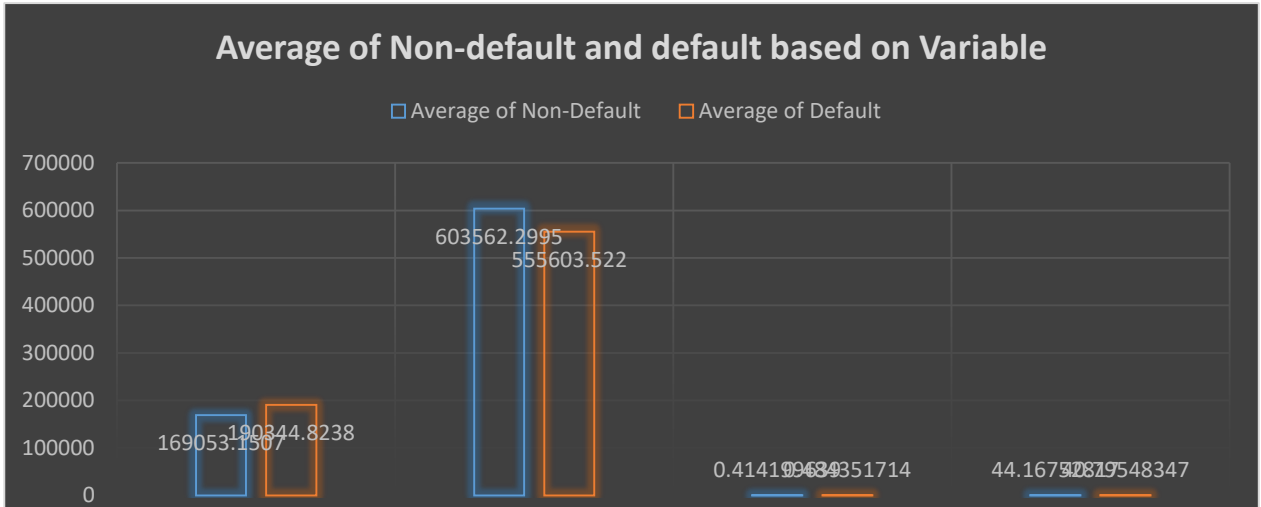


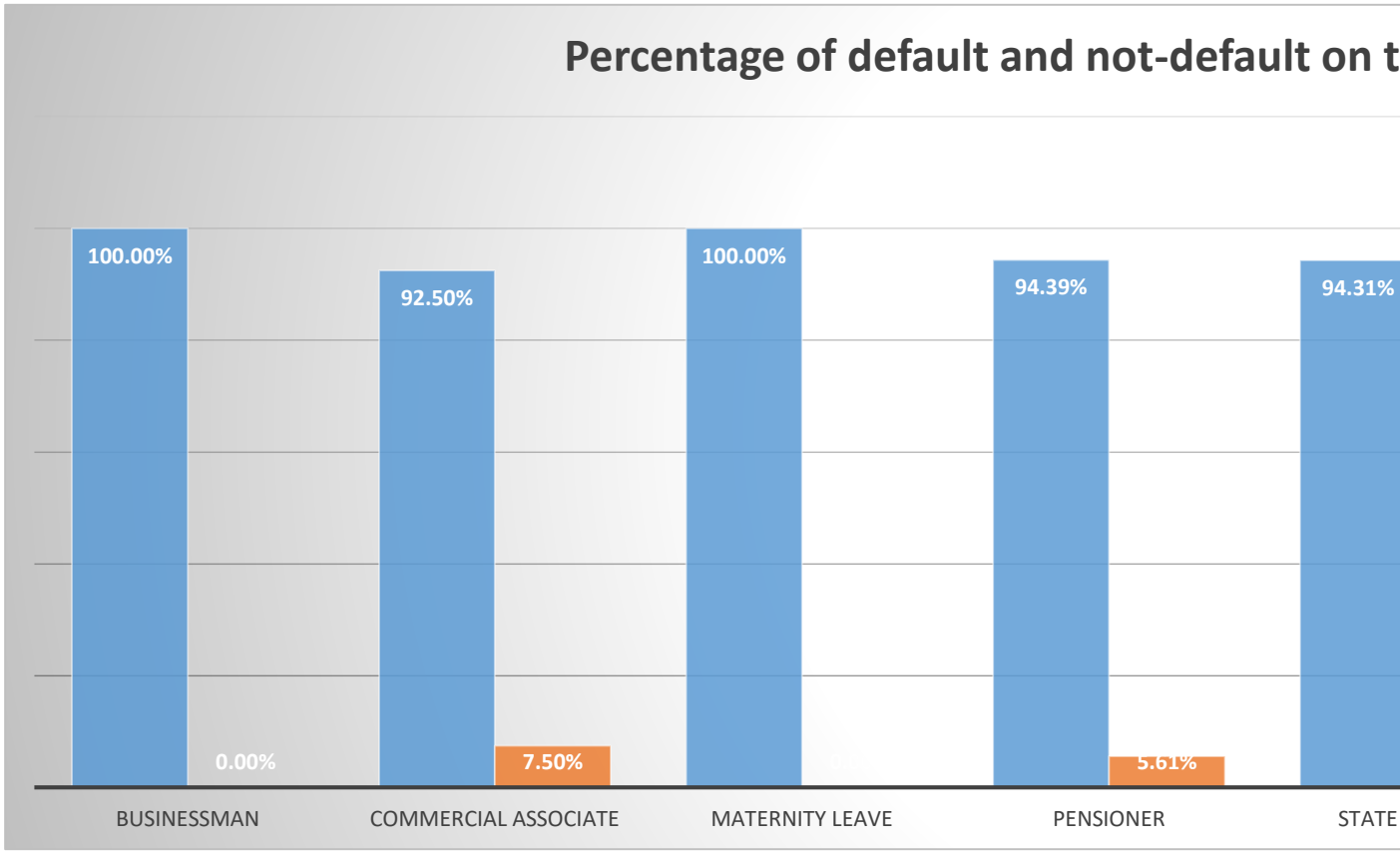




# Segmented Univ



## Bivariate Ana




**<u>Task E:</u> Identify Top Correlations for Different Variables**

**Objective:**

To analyse relationships between key variables in the loan application dataset by calculating correlation coefficients. This helps understand multi collinearity, uncover strong associations, and refine feature selection for predictive modelling.

**Key Insight:**

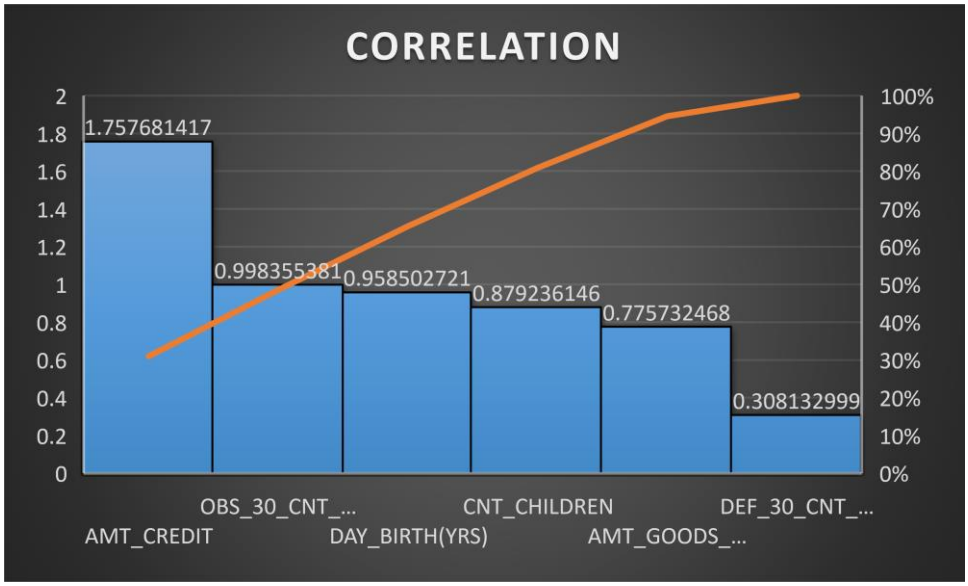
1. **Strong Correlations:**
    - AMT_CREDIT vs. AMT_GOODS_PRICE: High correlation of **0.987** indicates loan amount closely tracks the price of goods/services being financed.
    - OBS_30_CNT_SOCIAL_CIRCLE vs. OBS_60_CNT_SOCIAL_CIRCLE: Nearly perfect correlation of **0.998** implies redundancy; both variables capture very similar social behaviour metrics.
2. **Moderate to Strong Correlations:**
    - CNT_CHILDREN vs. CNT_FAM_MEMBERS: Correlation of **0.879** suggests family size is heavily influenced by the number of children, reinforcing the predictive importance of household demographics.
    - AMT_GOODS_PRICE vs. AMT_ANNUITY and AMT_CREDIT vs. AMT_ANNUITY: Both ~0.77, showing repayment structure aligns with credit and purchase amounts.
3. **Demographic Time-based Correlations:**
    - DAY_BIRTH vs. DAY_EMPLOYED: Moderate correlation of **0.623**, indicating some age-employment duration relationship.
    - DAY_BIRTH vs. DAY_REGISTRATION: Weaker correlation of **0.335,** implying more variation in account registration timing.
4. **Social Circle Defauts:**

- DEF_30_CNT_SOCIAL_CIRCLE vs. DEF_60_CNT_SOCIAL_CIRCLE: Low correlation of **0.308**, indicating default behaviour within a borrower's circle varies across 30 vs. 60-day timeframes.

**Business Impact:**

- **Improved Feature Selection:** Identifying and eliminating highly correlated variables (e.g., one from the pair OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE) can reduce multi collinearity and improve model accuracy.
- **Credit Risk Modelling:** Strong correlations between credit amounts and annuity/goods price help build more reliable credit scoring models.
- **Operational Efficiency:** Understanding that social behaviour metrics or demographic indicators are tightly linked can guide streamlined data collection and model design.
- **Better Customer Profiling:** Insight into demographic and behavioural linkages (e.g., family size, employment duration) aids in targeted loan offerings or risk segmentation.

## Correlations to Different Variables



## Overall Business Outcome:

1. **Cleaned and Reliable Dataset:** Enhanced data quality via imputation and cleaning, forming the base for further modelling or reporting.
2. **Customer Segmentation and Profiling:** Detailed analysis enabled profiling of high-risk vs. low-risk applicants based on education, employment, and demographic indicators.
3. **Risk Identification and Mitigation:** Detected data imbalance and high-risk segments (e.g., unemployed, working class) to improve credit policies and reduce non-performing loans.
4. **Data-Driven Decision Support:** Insights from correlation and bivariate analysis empowered data-backed decisions in marketing, loan approval, and credit modelling.