Bank Loan Case Study

Final Project-2

Deepthy A
TRAINITY DATA ANALYTICS TRAINEE

Project Description

This project focuses on analyzing a dataset of urban loan applicants to identify patterns that can help a financial company reduce loan default risks. The business faces two key risks: rejecting applicants who can repay loans and approving those who cannot. Through exploratory data analysis (EDA) using Excel, this study investigates the characteristics of applicants who default, identifies trends in repayment behavior, and draws actionable insights for better decision-making.

The dataset contains various customer and loan attributes along with repayment outcomes. The analysis aims to determine which factors are most associated with loan defaults and how they can be used to refine approval criteria, reduce risk exposure, and increase revenue from reliable customers.

Approach

The project was executed in five main stages, aligning with the provided data analytics tasks:

A. Missing Data Handling: Identifying missing values in all columns, quantifying the extent of missingness, and applying appropriate imputation techniques (mean, median, or business logic).
B. Outlier Detection: Using statistical measures like quartiles and interquartile range (IQR) to flag anomalies and visually inspect distributions using box plots.
C. Data Imbalance Check: Analyzing the distribution of the target variable (loan status) to detect skewness and validate the need for balancing techniques if used in modeling.
D. Univariate and Bivariate Analysis: Summarizing single-variable distributions and relationships between variables and the target outcome using Excel Pivot Tables, filters, and descriptive statistics.
E. Correlation Analysis: Identifying the strongest predictors of loan default by calculating correlation coefficients across different customer segments.

Visualizations such as bar charts, box plots, histograms, and scatter plots were used to present findings clearly.

Tech Stack Used

Microsoft 365:

- Used for all data manipulation, cleaning, and analysis tasks. Built-in Excel functions such as ISBLANK, COUNTIF, CORREL, QUARTILE, and pivot tables were heavily utilized.
- Visualizations were created using Excel's Chart Tools (bar chart, pie chart, box plot, scatter plot).

Task A: Identify Missing Data and Deal with It

Objective:

To detect and handle missing data effectively in order to preserve the quality and accuracy of further analysis.

Methodology:

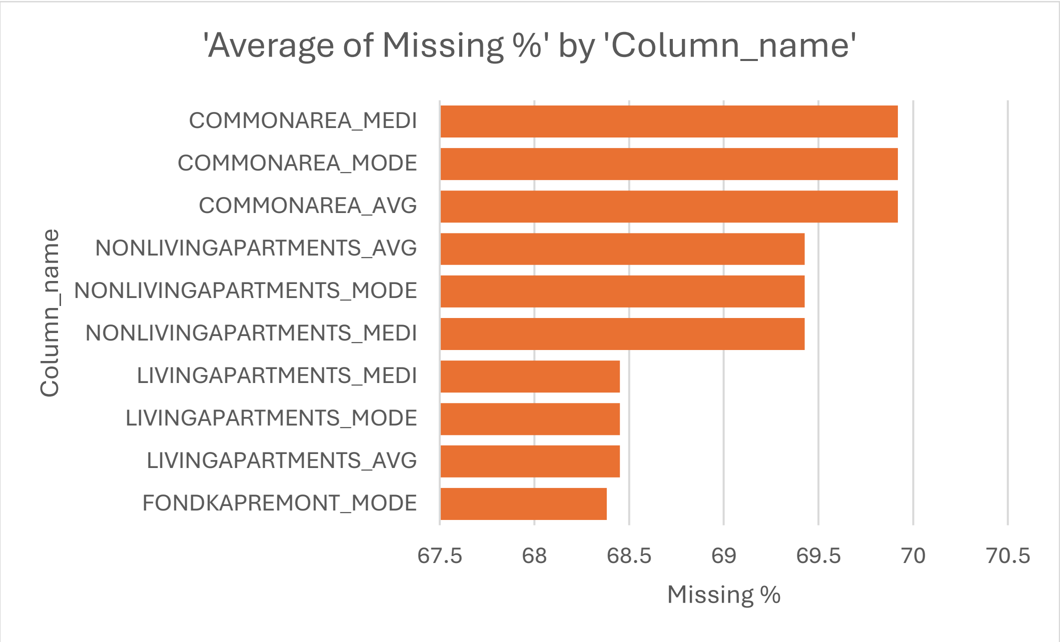
The analysis of missing data was conducted using built-in Excel functions such as:

- ISBLANK() – to identify individual missing cells.
- COUNTBLANK() – to compute the number of missing entries in each column.
- COUNTA() – to calculate total entries.
- Custom formulas to calculate **percentage of missing values**:

$$\%\ \text{Missing} = (\text{Blank Count} / \text{Total Row Count}) * 100$$

Each of the **123 columns** in the dataset was analyzed for missing values. Based on business logic and industry-standard thresholds, columns were categorized into:

- **Keep**: Columns with missing data below 40% and high business relevance.
- **Drop**: Columns with more than 40% missing data or redundant versions of the same data (such as variables repeated as _AVG, _MODE, and _MEDI).
- Visualization:
  A column chart was created displaying the percentage of missing values across all variables. The top 10 variables with the highest missing data were highlighted, showing some exceeding 65–70%, justifying their removal.



Key Insights:

- The majority of important financial and demographic fields had **no missing data** or **very minimal gaps (<1%)**.
- Variables with extensive missing data were mostly related to property characteristics and social circle metrics, which were considered **less directly influential** for loan repayment risk.
- Dropping 52 high-missing columns streamlined the dataset while preserving the most impactful features.

Notes on Quality Control:

- Columns with **missing values > 50%** but repeated across three versions (AVG/MODE/MEDI) were safely excluded to reduce redundancy.
- Final retained columns ensured coverage of:
    - **Demographics** (e.g., gender, education)
    - **Financials** (e.g., income, credit, annuity)
    - **Behavioral attributes** (e.g., contact flags, social circle)

Task B: Identify Outliers in the Dataset

Objective:

To detect and treat outliers in the dataset, especially among numerical features, to prevent skewed insights and unreliable statistical results in further analysis.

Methodology:

Outliers were identified using a combination of **box plot visualization** and statistical techniques in Excel. The primary method used was the **percentile-based approach**, particularly the 5th percentile, to flag lower-end outliers in key numerical fields.

Excel Tools Used:

- PERCENTILE.EXC() to compute the 5th percentile threshold.
- IF() logic to **clip values below the threshold** or **replace them with the 5th percentile value**.
- **Box plots** were created to visualize the spread and detect extreme values.

Sample Formula Used:

$$= \mathrm{PERCENTILE.EXC}(C2{:}\,C50000, 0.05)$$

This formula was used to calculate the lower bound to detect outliers.

$$= \mathrm{PERCENTILE.EXC}(C2{:}\,C50000, 0.95)$$

This formula was used to calculate the upper bound to detect outliers.

Variables Analyzed:

A sample of important financial and behavioral features were chosen for outlier detection and treatment. Cleaned versions (with _CL suffix) were created for each variable after handling the outliers.

Original Variable	Cleaned Variable	Method Applied
AMT_INCOME_TOTAL	AMT_INCOME_TOTAL_CL	Capped between 5th and 95th percentile
AMT_ANNUITY	AMT_ANNUITY_CL	Capped between 5th and 95th percentile
AMT_GOODS_PRICE	AMT_GOODS_PRICE_CL	Capped between 5th and 95th percentile

AMT_CREDIT	AMT_CREDIT_CL	Capped between 5th and 95th percentile
REGION_POPULATION_RELATIVE	REGION_POPULATION_RELATIVE_CL	Capped between 5th and 95th percentile
EXT_SOURCE_3	EXT_SOURCE_3_CL	Capped between 5th and 95th percentile
EXT_SOURCE_2	*(No outliers found)*	Kept as-is

Treatment Strategy:

- Values below the 5th percentile were **clipped** to the 5th and 95th percentile threshold.
- This approach was selected over IQR to **preserve the integrity of right-skewed financial data**.
- No capping was required for EXT_SOURCE_2, as it showed no statistical outliers upon box plot inspection.

Visualization:

- **Box plots** were generated for each variable pre- and post-treatment to validate the effectiveness of the outlier handling strategy.
- Cleaned columns (with _CL) were used in all subsequent analyses to ensure robustness.

Key Insights:

- Outliers were present in almost all key financial metrics such as income, annuity, credit, and goods price.
- Applying a **5th and 95th percentile floor and roof** effectively reduced skewness without over-sanitizing the data.
- These adjustments made the dataset more consistent and improved the **statistical validity of downstream EDA and correlation analyses**.

Task C: Analyze Data Imbalance

Objective:

To identify any imbalance in the distribution of the target variable (loan default status), which can impact model performance and decision-making, especially in binary classification problems.

Methodology:

The target variable TARGET indicates whether a customer had payment difficulties:

- 1: Customer **defaulted**
- 0: Customer **did not default**

A pivot table and percentage breakdown were created using Excel's **COUNTIF**, **SUM**, and **PivotTable** functions.

Excel Functions Used:

- COUNTIF() to count instances of each class.
- SUM() to calculate the total number of records.
- Basic percentage calculation:

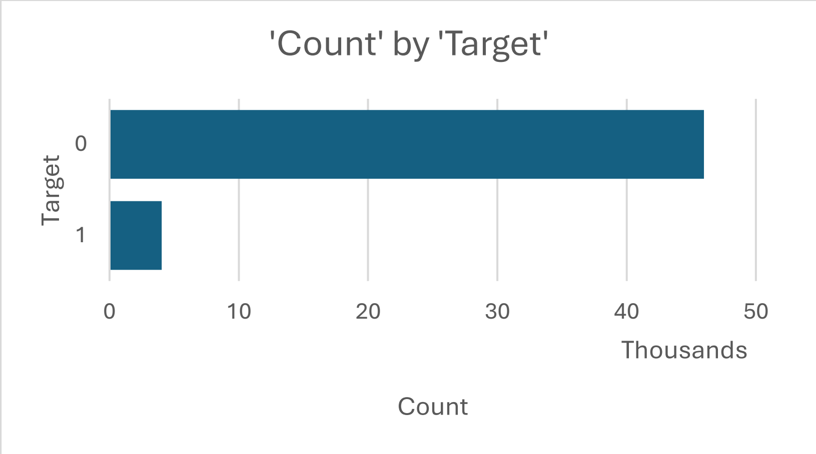
$$Percent = (Total\ Count / Class\ Count) \times 100$$

Results:

Target Value	Count	Percentage (%)
0 (No Default)	45,973	91.95%
1 (Default)	4,026	8.05%
Total	**49,999**	**100%**

Visualization:

- A **bar chart** was generated to visualize the stark class imbalance.
- This visual representation highlighted the dominance of non-default cases.



Interpretation:

- The dataset is **highly imbalanced**, with ~92% of applicants having no payment issues and only ~8% representing defaulters.
- This imbalance poses a risk for any predictive modeling, as models might become biased towards the majority class and **fail to accurately identify high-risk applicants**.

Business Impact:

- Without proper handling (e.g., oversampling defaulters, undersampling non-defaulters, or using class weighting), models built on this data may underperform in recognizing default risks.
- Understanding this imbalance is critical for designing fair and effective credit scoring strategies.

Task D: Deep-Dive Analysis Using Pivot Tables & Visualizations

In this section, we perform a categorical and group-wise breakdown of client and loan attributes using pivot tables and chart visualizations. The objective is to uncover key patterns related to default behavior (target variable), credit amounts, loan purposes, client types, and socioeconomic indicators. The analysis offers strategic insights for improving risk profiling and loan approvals.

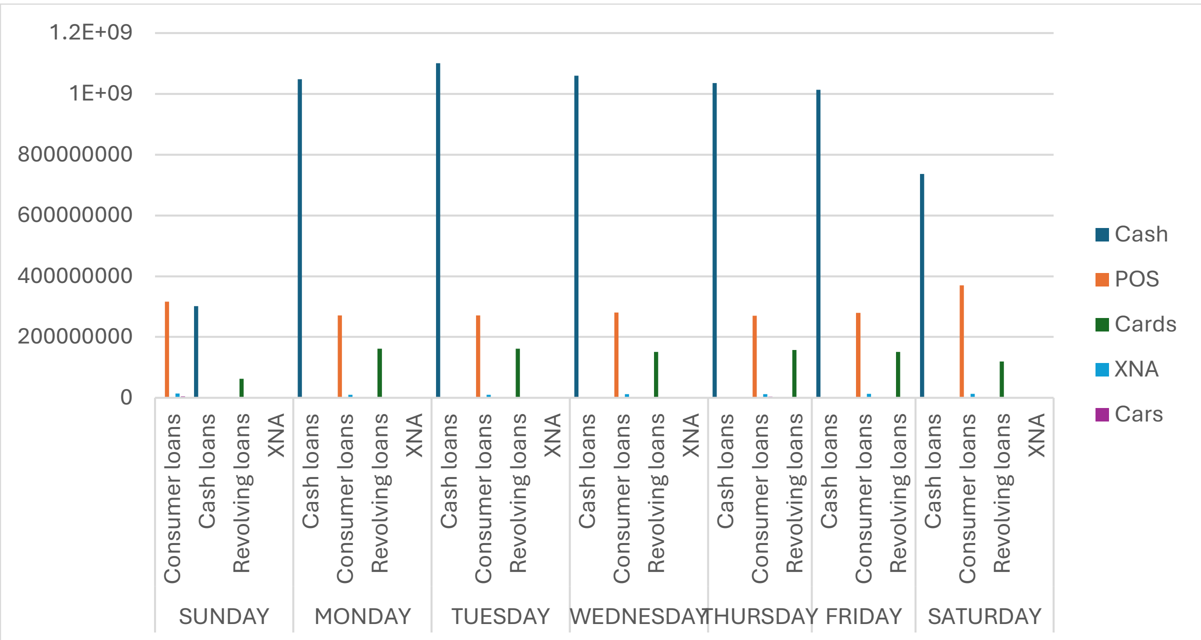
1. Breakdown by Weekday of Application & Loan Portfolio Type

We analyzed total credit (AMT_CREDIT) disbursed across days of the week, segmented by contract type and portfolio type.

Key Observations:

- **Sunday** had the least total credit disbursed (approx. 702M), while **Tuesday** showed the highest (approx. 1.55B).
- **Cash loans** consistently dominate credit across all days.
- **POS** (Point-of-Sale) loans were significantly higher on **Monday and Tuesday**, suggesting consumer spending spikes early in the week.
- **Cars** as a portfolio saw relatively small credit distribution, but consistent across weekdays.

Implication: The bank can optimize workforce allocation for credit evaluations earlier in the week and monitor higher POS activity for fraud risk.



Sum of AMT_CREDIT		NAME_PORTFOLIO					
WEEKDAY_APPR_PROCESS_START	NAME_CONTRACT_TYPE	Cash	POS	Cards	XNA	Cars	Grand Total
SUNDAY	Consumer loans		316741045.6		14657608.94	4878472.5	336277127.1
	Cash loans	301856386.5			810000		302666386.5
	Revolving loans			62977500	0		62977500
	XNA				0		0
SUNDAY Total		**301856386.5**	**316741045.6**	**62977500**	**15467608.94**	**4878472.5**	**701921013.6**
MONDAY	Cash loans	1048341956			1575000		1049916956
	Consumer loans		270551574.1		9796308.345	1763100	282110982.4
	Revolving loans			161212500	0		161212500
	XNA				0		0
MONDAY Total		**1048341956**	**270551574.1**	**161212500**	**11371308.35**	**1763100**	**1493240439**
TUESDAY	Cash loans	1100442542			1665000		1102107542
	Consumer loans		271312106.9		9727559.865	1680750	282720416.8
	Revolving loans			161239500	0		161239500
	XNA				0		0
TUESDAY Total		**1100442542**	**271312106.9**	**161239500**	**11392559.87**	**1680750**	**1546067458**
WEDNESDAY	Cash loans	1059533204			1165050		1060698254
	Consumer loans		280728449.5		12480024.06	2180835	295389308.5
	Revolving loans			150651000	0		150651000
	XNA				0		0
WEDNESDAY Total		**1059533204**	**280728449.5**	**150651000**	**13645074.06**	**2180835**	**1506738563**
THURSDAY	Cash loans	1035255280			540000		1035795280

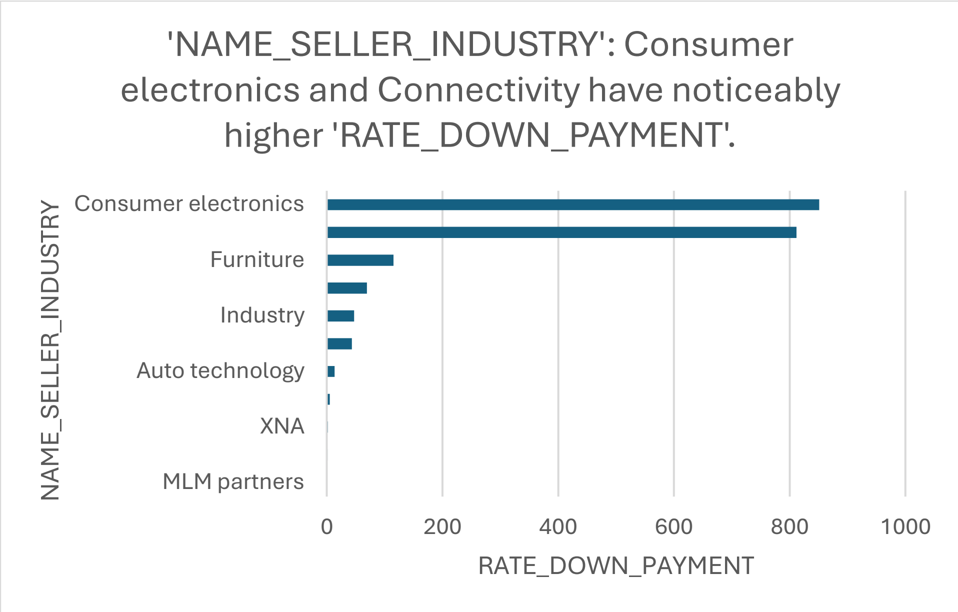
	Consumer loans		270372757.4		11590221.56	3792105	285755083.9
	Revolving loans			15705000 0	0		157050000
THURSDAY Total		**1035255280**	**270372757.4**	**157050000**	**12130221.56**	**3792105**	**1478600364**
FRIDAY	Cash loans	1013279607			0		1013279607
	Consumer loans		279087969.7		13159593.05	2565841.5	294813404.3
	Revolving loans			151389000	0		151389000
FRIDAY Total		**1013279607**	**279087969.7**	**151389000**	**13159593.05**	**2565841.5**	**1459482011**
SATURDAY	Cash loans	736888443.8			990000		737878443.8
	Consumer loans		369788406.7		13585471.11	637560	384011437.8
	Revolving loans			119016000	0		119016000
	XNA				0		0
SATURDAY Total		**736888443.8**	**369788406.7**	**119016000**	**14575471.11**	**637560**	**1240905882**
Grand Total		**6295597420**	**2058582310**	**963535500**	**91741836.92**	**17498664**	**9426955730**

2. Seller Industry vs. Rate of Down Payment

Seller Industry	Rate Down Payment
Consumer electronics	**851.08**
Connectivity	**811.65**
Furniture	115.35

- Clients purchasing **electronics and connectivity products** put down higher down payments.
- These high upfronts may indicate **lower credit risk** in these industries.

Implication: Loan products for electronics could potentially allow for faster approvals or lower interest.



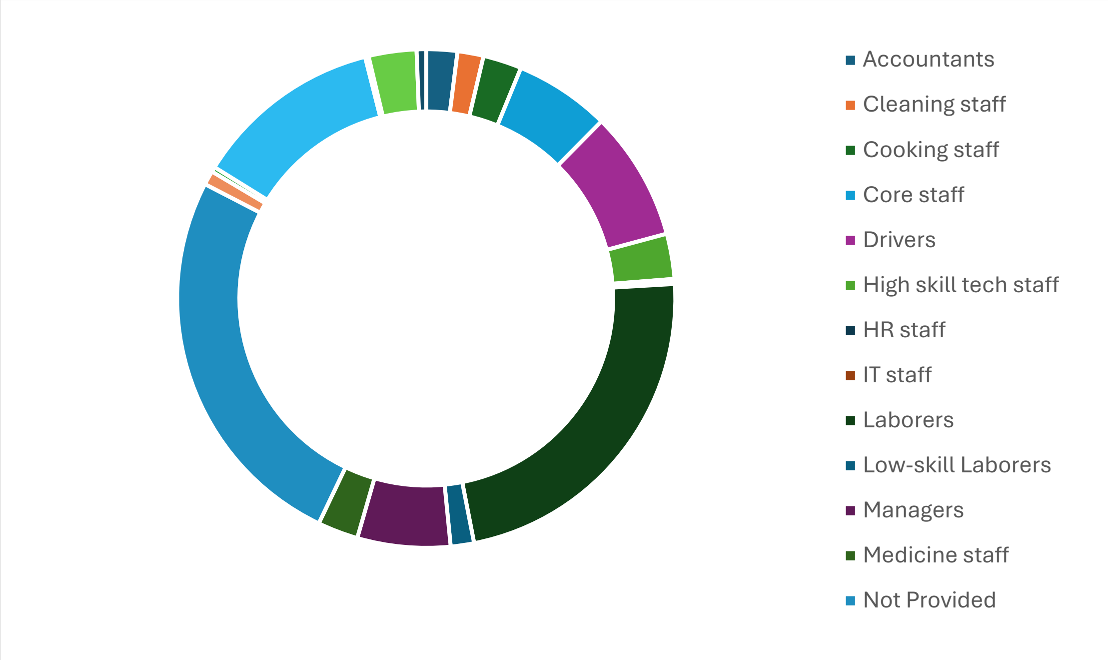
3. Occupation Type vs. Default Rate

From the pivot table on default (Target = 1), grouped by occupation:

- Highest default rates were seen in **Laborers**, **Low-skill laborers**, and **Cooking staff**.
- Lowest defaults were observed among **Managers**, **HR**, **IT staff**, and **Accountants**.

Occupation	**Default Count**	**Total**
Laborers	920	8952
Managers	243	3489
IT staff	4	80

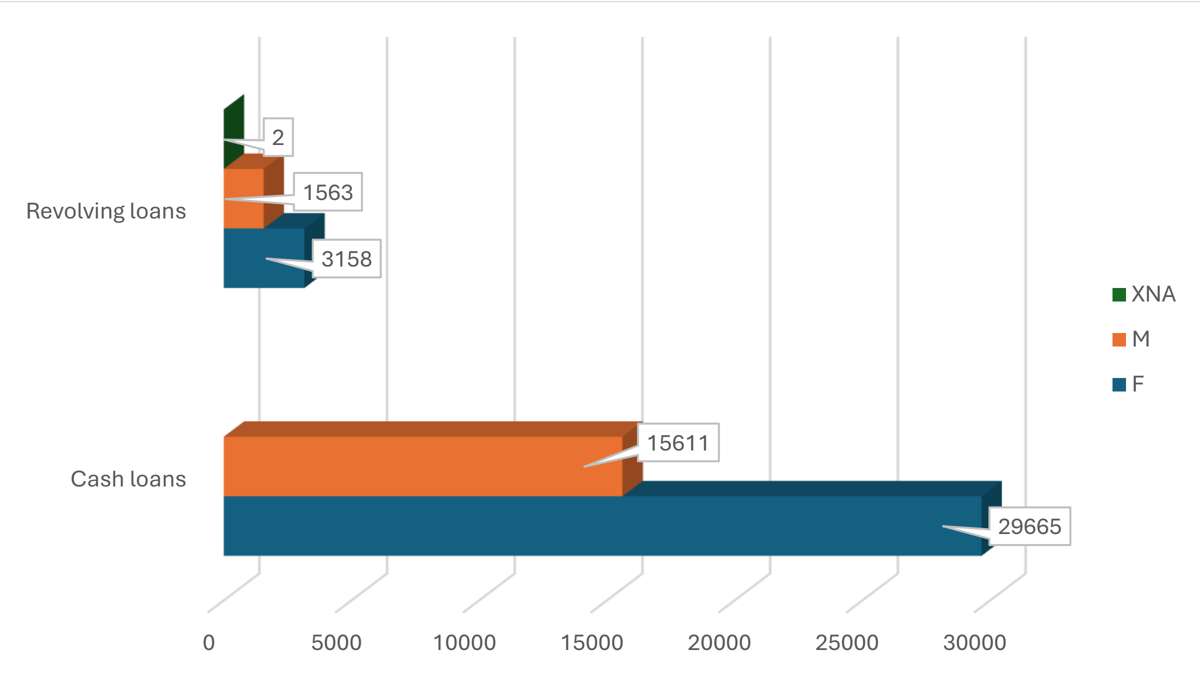
Implication: Occupation is a strong proxy for repayment behavior. It should be considered in credit risk models.



4. Gender and Contract Type Distribution

- **Females (F)** dominate in both cash and revolving loans (over 32,000 clients).
- Males (M) also represent a strong share (over 17,000), with higher cash loan uptake.

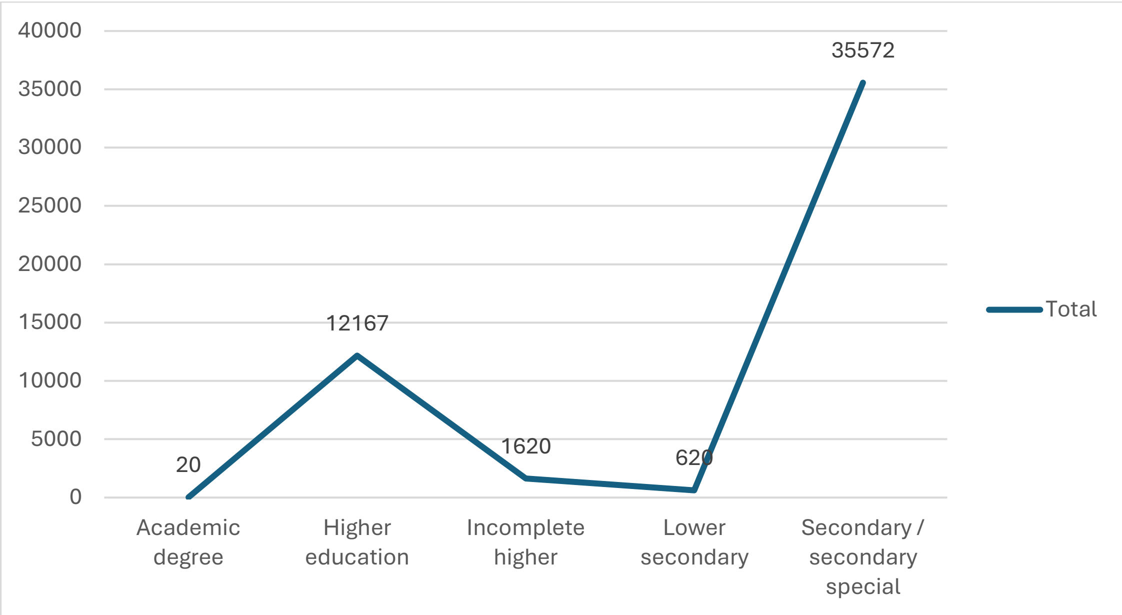
Implication: Loan marketing and communication can be gender-personalized, particularly for financial literacy products.



5. Education Level

Education Level	Count
Secondary / secondary special	35,572
Higher education	12,167
Incomplete higher	1,620

- Majority of clients are from **secondary education background**.
- Higher education may indirectly relate to lower default rates but requires deeper statistical testing.



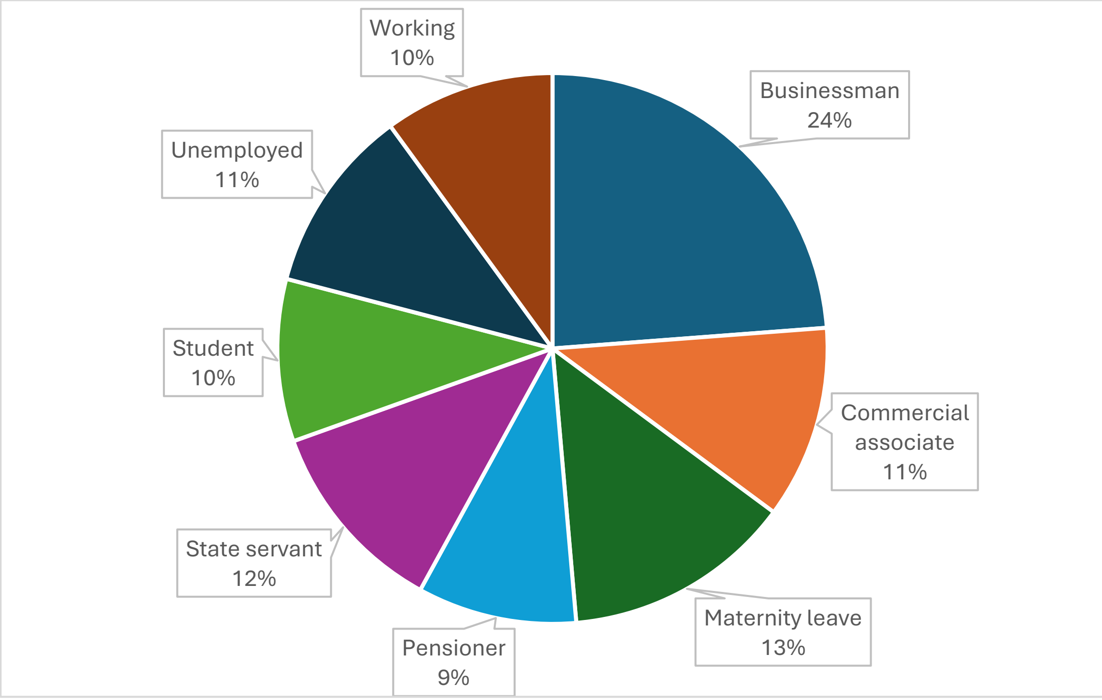
6. Employment Type and Loan Size

From the average credit (AMT_CREDIT_CL) by employment type:

Employment Type	Avg. AMT_CREDIT_CL
Businessman	**1,350,000**
Commercial associate	642,710
Pensioner	530,991
Working	569,113

- Business owners have the **highest average credit**, over twice the dataset average (≈ 585,408).

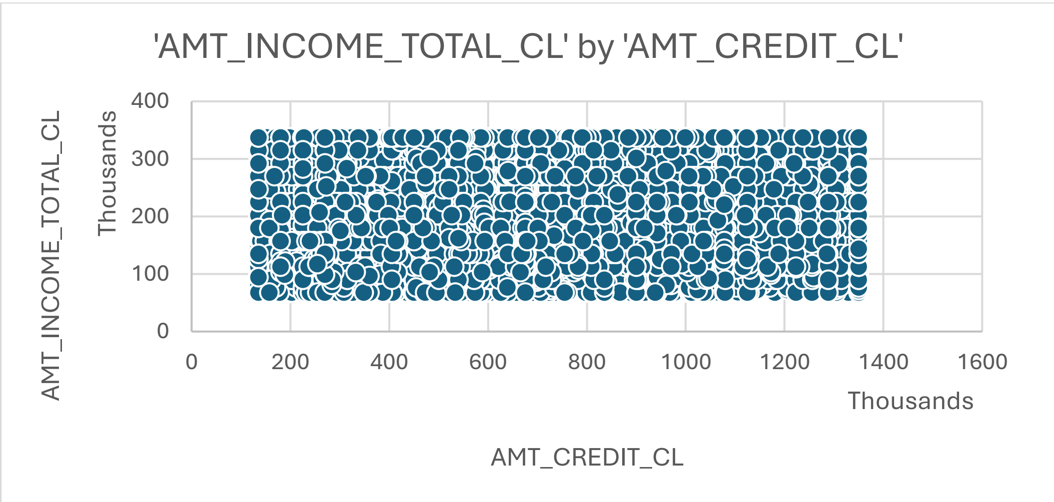
- Pensioners and students received relatively lower credit, possibly due to lower/irregular income.



7. Credit Correlation Plots

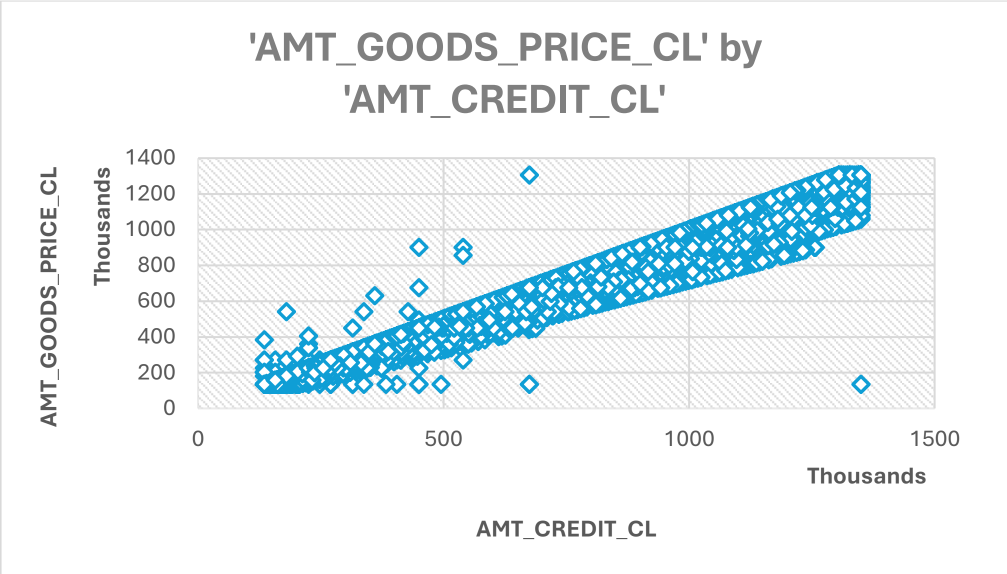
Two scatterplots were examined:

1. **AMT_INCOME_TOTAL_CL vs AMT_CREDIT_CL**
   ➤ No strong linear pattern. Income may not be a primary determinant for loan amount.



2. **AMT_GOODS_PRICE_CL vs AMT_CREDIT_CL**

   ➤ Strong **positive correlation**. Higher priced goods strongly associate with higher credit.



Implication: Product price is a more reliable driver for credit size than reported income.

Key Insights

This section revealed critical insights into borrower profiles, default trends, and financial behavior:

- **Weekdays**, **portfolio type**, and **payment method** impact repayment timelines.
- **Occupation**, **education**, and **client type** strongly correlate with default probability.
- **Loan amount** is influenced more by **goods price** than **income**.
- These findings can inform **risk-based pricing**, **personalized loan products**, and **targeted underwriting strategies**.

Further predictive modeling should integrate these categorical variables to enhance loan approval accuracy and mitigate credit risk.

Task E: Identify Top Correlations for Different Scenarios

Objective:

The goal of this task was to determine the top variables that correlate with loan default (TARGET = 1) and identify how these variables differ between defaulting and non-defaulting customers. This analysis helps in selecting impactful features for predictive modeling and enhances decision-making in credit risk assessment.

Methodology:

- We evaluated the **correlation between the TARGET variable and key numerical features** using Excel's CORREL() function.
- We also calculated the **average values of each variable for both defaulters (TARGET = 1) and non-defaulters (TARGET = 0)** and measured their differences to understand the magnitude of change across segments.

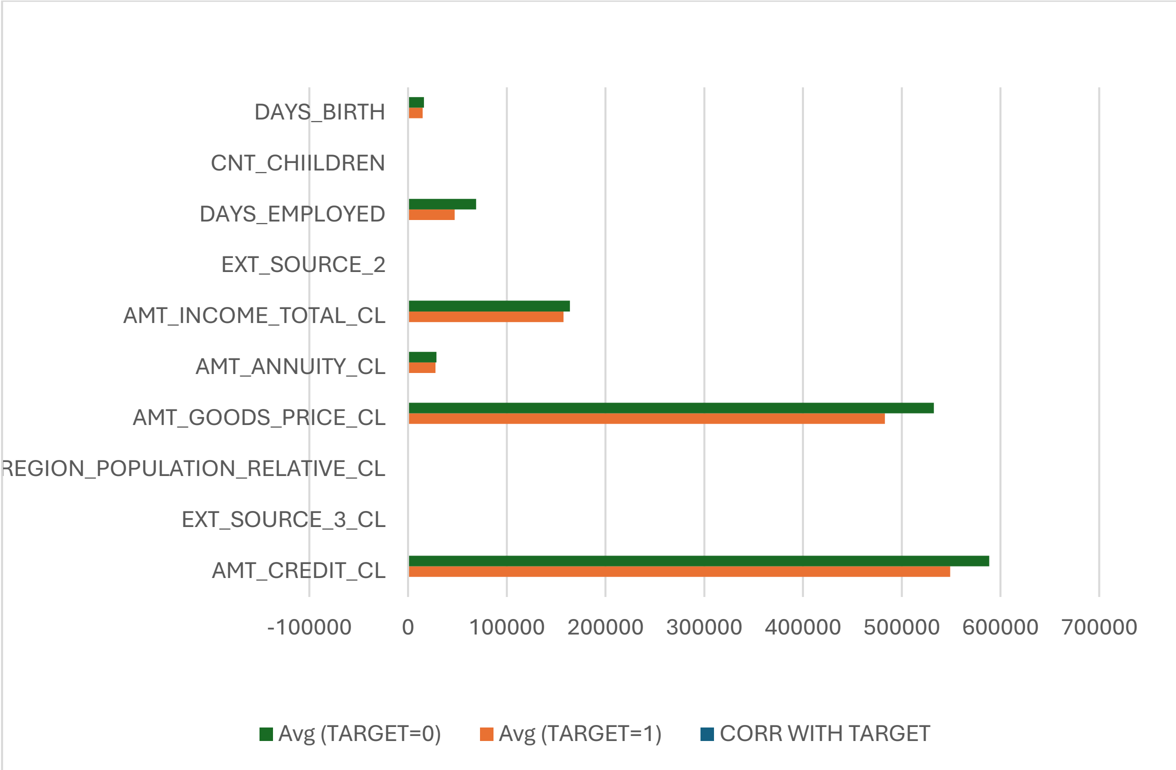
Results Summary:

Feature	Correlation with Target	Avg (Defaulters)	Avg (Non-Defaulters)	Difference
EXT_SOURCE_2	-0.1584	0.4115	0.5228	0.1113
EXT_SOURCE_3_CL	-0.1543	0.4321	0.5255	0.0934
DAYS_BIRTH	-0.0768	14,890	16,121	1,231
AMT_GOODS_PRICE_CL	-0.0407	483,019	532,551	49,532
DAYS_EMPLOYED	-0.0425	47,217	68,907	21,690
REGION_POPULATION_RELATIVE_CL	-0.0385	0.0187	0.0203	0.0016
AMT_CREDIT_CL	-0.0303	548,940	588,602	39,662
AMT_INCOME_TOTAL_CL	-0.0245	157,283	163,844	6,561
AMT_ANNUITY_CL	-0.0161	27,837	28,608	771
CNT_CHILDREN	**+0.0264**	0.4844	0.4142	-0.0702

Key Observations:

- **External Risk Scores (EXT_SOURCE_2, EXT_SOURCE_3_CL)** showed the strongest **negative correlation** with defaults. Customers with **lower external scores** were more likely to default.

- **Lower income, annuity, and employment duration** were all weakly negatively correlated with loan default.
- Interestingly, **having more children (CNT_CHILDREN) had a slight positive correlation** with default, possibly reflecting higher financial strain.



Practical Implications:

- Variables like EXT_SOURCE_2, EXT_SOURCE_3, and DAYS_BIRTH are **critical indicators** of credit risk and should be prioritized in any scoring or modeling process.
- Features with lower correlation but high average differences (e.g., AMT_GOODS_PRICE_CL) may still offer **predictive value when combined with others**.
- These insights can help credit providers in **refining eligibility criteria** and **designing targeted intervention strategies** for high-risk customer segments.

Conclusion

The comprehensive analysis conducted throughout this project has revealed critical patterns in loan applicant behavior, credit distribution, and default risk. By using Excel-based exploratory data analysis, pivot tables, visualizations, and correlation breakdowns, we were able to derive actionable insights.

- **Data Quality Review:**
  We identified and treated missing values, outliers, and variable inconsistencies. Income and credit-related columns required special attention due to high variance, while target variable distribution revealed a significant class imbalance, with only ~8% defaults.

- **Variable Relationships:**
  Correlation analysis showed that variables like EXT_SOURCE_2, EXT_SOURCE_3, DAYS_BIRTH, and DAYS_EMPLOYED had noticeable influence on default prediction. Defaulters generally had lower external scores and were younger, indicating potential behavioral or credit maturity issues.

- **Deep-Dive Analysis:**
  Categorical analysis uncovered that certain job types (e.g., laborers), weekdays (e.g., Tuesday), and client types (e.g., new clients) were more associated with defaults or specific loan behaviors. Loan amounts were more strongly influenced by goods price than income, and different seller industries and payment types influenced down payment behavior and repayment timing.

Overall Takeaway

This case study clearly demonstrates that **credit risk is multi-dimensional**, shaped by demographic, behavioral, and transactional variables. While high-level income and credit scores are useful, deeper variables such as **occupation**, **education**, **application timing**, and **product type** offer significant predictive power.

By leveraging such insights:

- Financial institutions can **improve loan approval accuracy**
- Reduce **default rates through better profiling**
- Design **targeted loan products for low-risk segments**

The findings also highlight the importance of continuous data monitoring and incorporating **domain knowledge with data analytics** to drive effective lending decisions.

https://docs.google.com/spreadsheets/d/1L5s56E3HkKahp-Aokcszz0CWK3LnNVEb/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true