PROBLEM STATEMENT

1. Overview:
- Apply Exploratory Data Analysis (EDA) in a real business scenario.
- Understand risk analytics in banking and financial services.
- Use data to minimize the risk of losing money while lending.

2. Business Understanding:
- Loan providers struggle with insufficient or non-existent credit history.
- Some consumers exploit this by becoming defaulters.
- Analyze patterns to ensure capable applicants are not rejected.

3. Data Description:
- Contains loan application information at the time of applying.
- Scenarios: payment difficulties and on-time payments.

4. Loan Application Outcomes:
- Approved: Loan application approved.
- Cancelled: Client canceled during approval due to various reasons.
- Refused: Loan application rejected.
- Unused offer: Loan canceled by the client at different stages.

5. Business Objectives:
- Identify patterns for denying, reducing loan amounts, or higher interest rates for risky applicants.
- Ensure capable consumers are not rejected.
- Utilize insights for portfolio and risk assessment.

PROBLEM STATEMENT

6. Analysis Approach:
- Identify and handle missing data.
- Identify and understand outliers.
- Analyze data imbalance.
- Perform univariate, bivariate analysis, and multivariate analysis.

7. Correlation Analysis:
- Find top correlations for clients with payment difficulties and others.
- Segment data based on the target variable for correlation analysis.

METHODOLOGY

1. Drop Features: Remove features with more than 40% missing values to maintain data integrity and ensure robust analysis.

2. Imputation:
- Numerical Features: Use the median to impute missing values, providing a robust measure against outliers.
- Categorical Features: Impute missing values with the mode to preserve the most frequent category.
- Selective Non-Imputation: In some cases, avoid imputing missing values to maintain the natural data distribution and characteristics.

3. Univariate analysis:
- Examine the distribution, central tendency, and variability of individual features.
- Identify patterns, trends, and anomalies within each feature.

METHODOLOGY

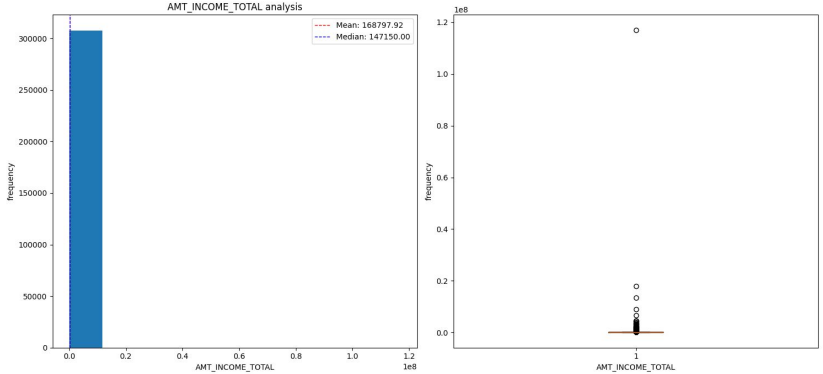
4. Bivariate analysis:
- Weight of Evidence (WOE): Calculate WOE to evaluate the predictive power of features relative to the target variable.
- Information Value (IV): Use IV to select the top 5 risk indicators. Features with high IV values are strong predictors of default risk.

5. Multivariate analysis:
- Identify the top 10 pairs of features with the highest correlation to understand the linear relationships between them.

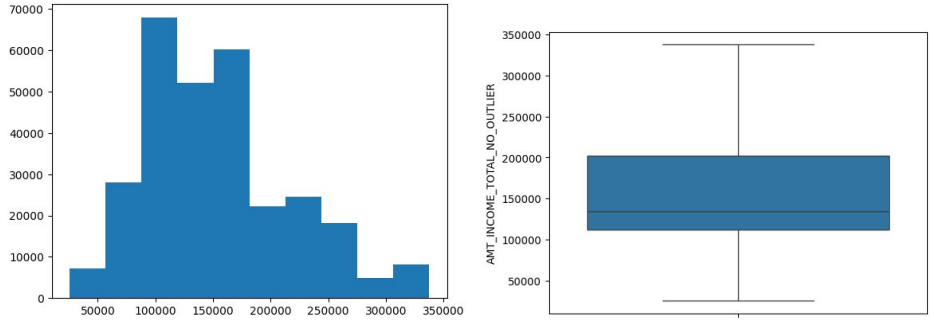
6. Correlation analysis: Identify the top 10 pairs of features with the highest correlation to understand the linear relationships between them.

INSIGHT (Univariate analysis)

- Drop all features with more than 40% missing data, approximately 41 features from the `APPLICATION` dataset and 13 features from the `PREVIOUS_APPLICATION` dataset.
- Exclude the detected outliers during analysis to improve clarity (using IQR), e.g.AMT_INCOME_TOTAL
- With outliers removed, a clear pattern emerges, showing that most applicants have an income ranging from 110k to 200k.



Before outliers

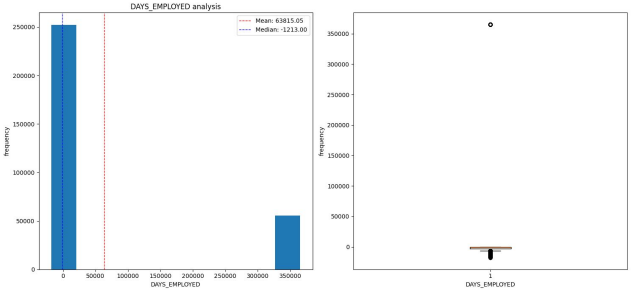


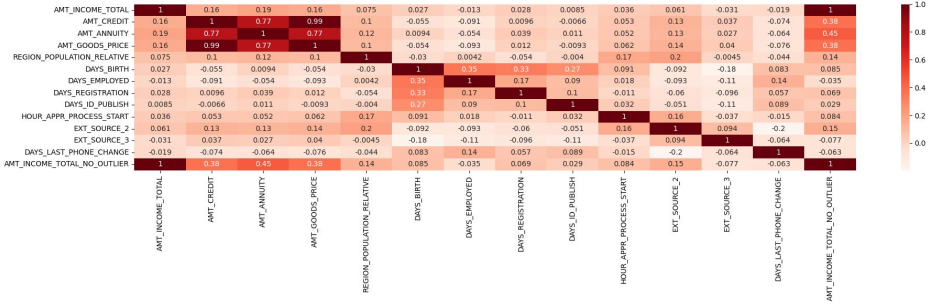
After outliers removal

INSIGHT (Univariate analysis)

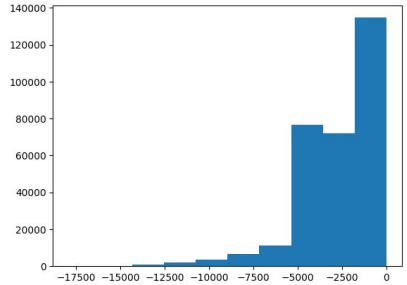
- Certain features exhibit abnormalities in their distribution.
- For example, 18% of the values in DAYS_EMPLOYED are abnormally recorded as 365243, which introduces noise and obscures data patterns.
- To address this, impute the abnormal values using linear regression.
- The feature DAYS_EMPLOYED has a moderate correlation (0.35) with DAYS_BIRTH, making it a suitable candidate for this imputation method.
- The graph below highlights the result of this imputation.
- Orange points represent the imputed values, showing a clearer pattern and reducing noise in the data.

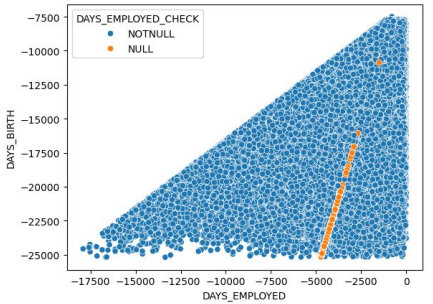
Before imputation





After imputation

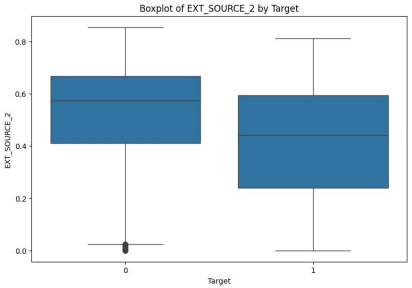


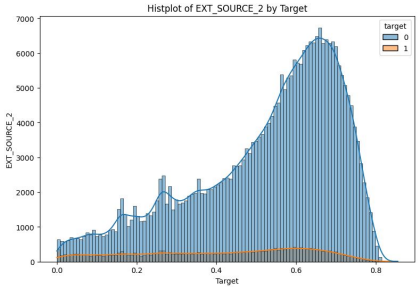


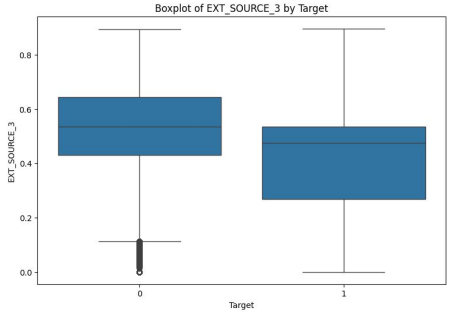
INSIGHT (Bivariate analysis)

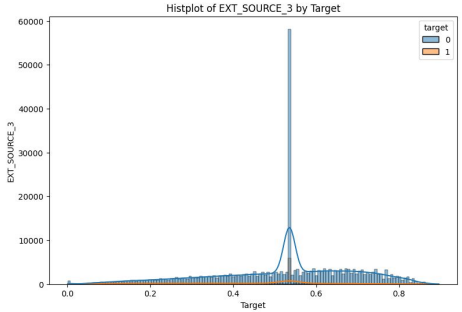
- Calculate the Information Value (IV) of each features and then find top 5 the most predictive power.
- **EXT_SOURCE_2 (credit score from source 2)** and **EXT_SOURCE_3 (credit score from source 3)** are the most significant predictors of default risk with **predictive power around 0.3.**
- Higher scores are associated with a lower likelihood of default (target = 0), while lower values correlate with a higher likelihood of default (target = 1).

Feature	IV	Predictive Power
EXT_SOURCE_3	0.32629	Strong
EXT_SOURCE_2	0.31933	Strong
DAYS_EMPLOYED	0.11650	Medium
DAYS_BIRTH	0.08926	Weak
AMT_GOODS_PRICE	0.08868	Weak



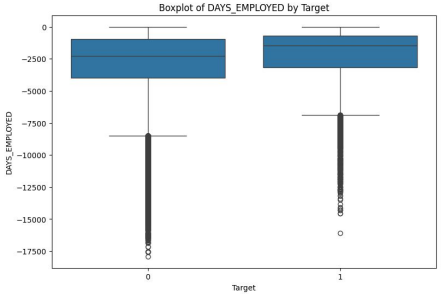


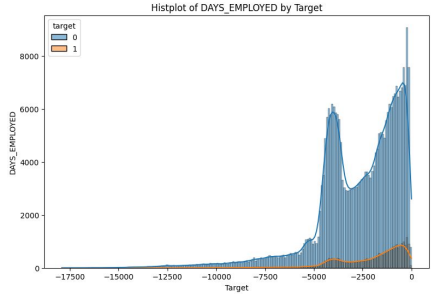


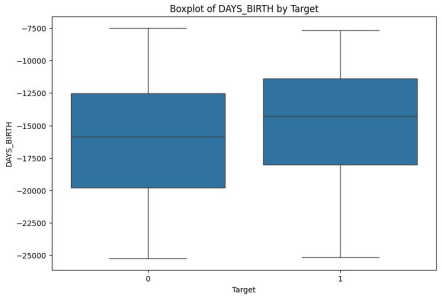


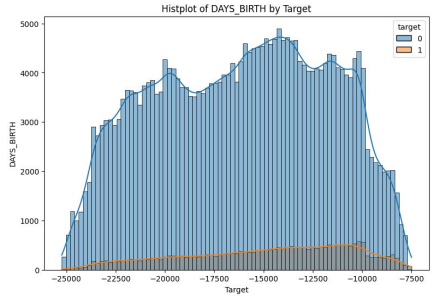
INSIGHT (Bivariate analysis)

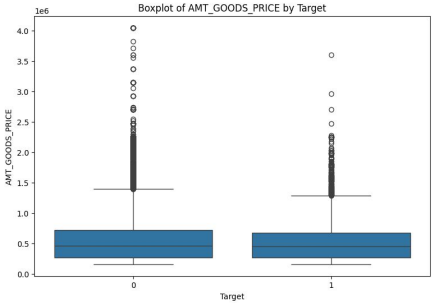
- **DAYS_EMPLOYED (IV=0.1)**: Longer employment duration is generally associated with lower default risk, but variability exists.
- **DAYS_BIRTH (IV=0.08)**: Age is not a strong predictor of default risk, with similar distributions for both defaulted and non-defaulted groups.
- **AMT_GOODS_PRICE (IV=0.08)**: Higher loan amounts for goods tend to correlate slightly with lower default risk, but many outliers indicate complexity in this relationship.

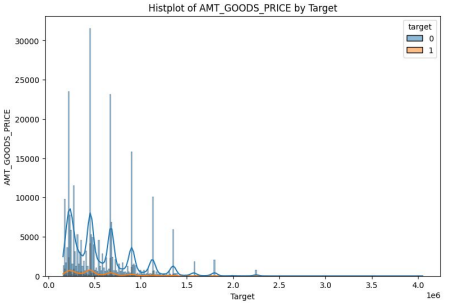










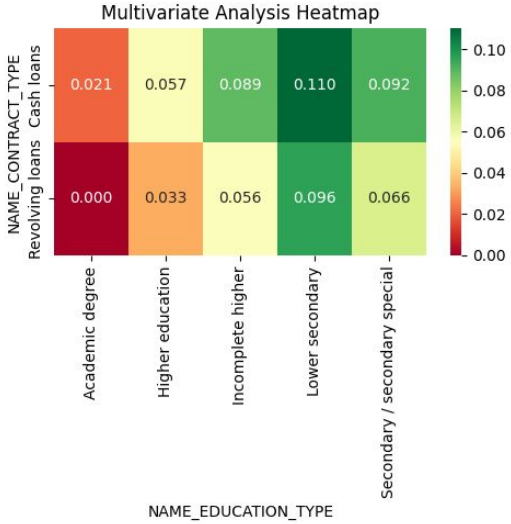


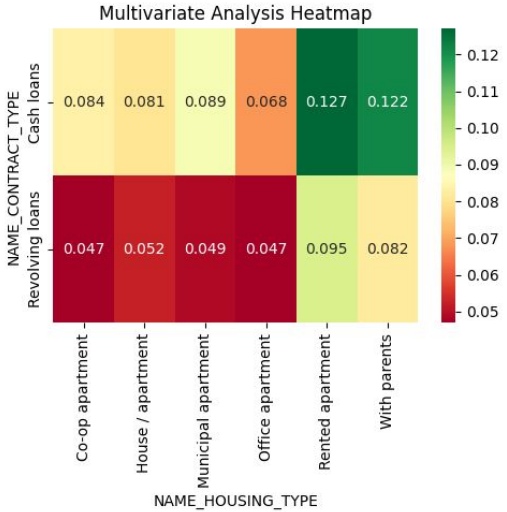
INSIGHT (Multivariate analysis)

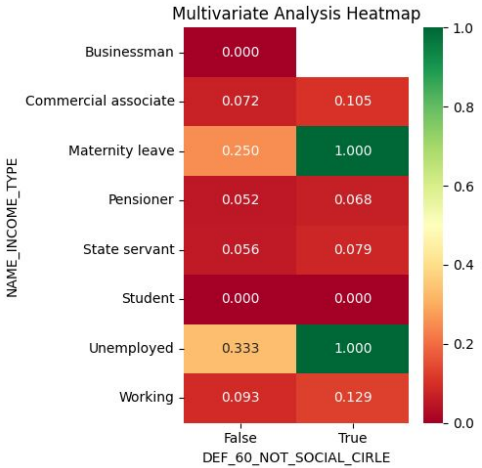
Educational Influence:
- Applicants with academic degrees are less likely to default, particularly with revolving loans.
- Lower secondary and incomplete higher education levels show higher default rates across both loan types.

Housing Influence:
- Applicants living with parents or rented apartment show higher default rates for both loan types.

Income Influence:
- Certain income types, particularly those on maternity leave or unemployed, show extremely high default rates, especially when involved in social circles.
- The "Working" group also shows increased risk when part of social circles.







INSIGHT (Correlation analysis)

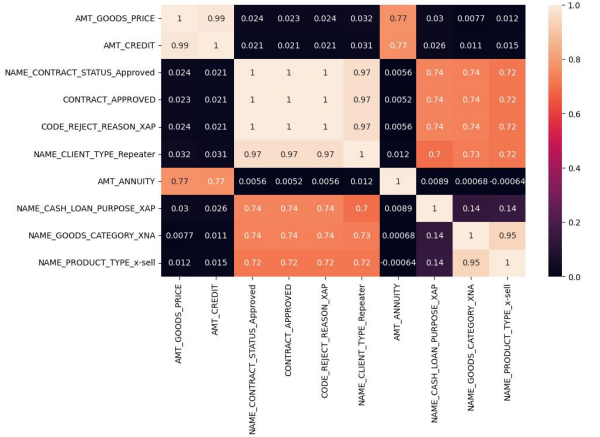
AMT_GOODS_PRICE vs. **AMT_ANNUITY**: Strong positive correlation (0.774), showing that as the goods price increases, the annuity amount also tends to increase, the scatterplot reveals a clear upward trend with some spread, indicating that while the relationship is strong, other factors might influence annuity amounts as well.

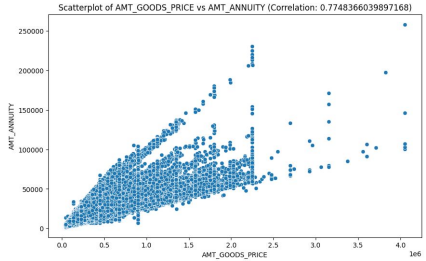
AMT_CREDIT vs. **AMT_ANNUITY**: High positive correlation (0.771), similar to the relationship between goods price and annuity. This scatterplot also shows a clear upward trend, reinforcing the idea that higher credit amounts result in higher annuity payments.

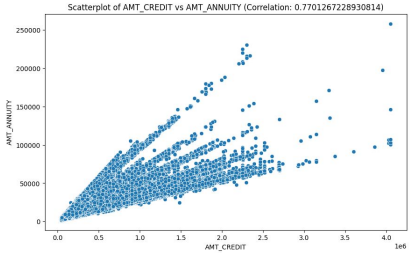
NAME_CONTRACT_STATUS_Approved vs. **NAME_CASH_LOAN_PURPOSE_XAP**:Positive correlation (0.744), suggesting a strong association between the number of approved contracts and the purposes of cash loans.

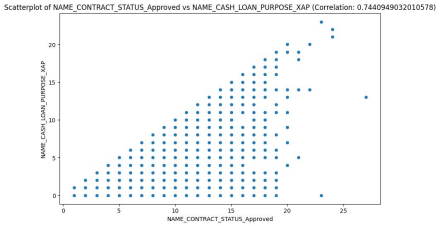
- The scatterplot indicates a clustered distribution, showing a direct relationship where an increase in one variable is associated with an increase in the other.

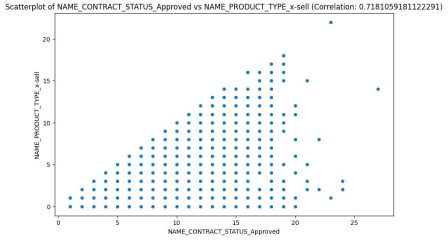
NAME_CONTRACT_STATUS_Approved vs. **NAME_PRODUCT_TYPE_x-sell**: Strong positive correlation (0.718), indicating that the number of approved contracts is closely linked to the product type for cross-selling. The scatterplot shows a linear trend with some spread, implying a consistent relationship but also highlighting variability due to other influencing factors.











RECOMMENDATION

1. **Focus on EXT_SOURCE_2 and EXT_SOURCE_3** for risk assessment due to their strong predictive power.
2. **Leverage employment duration** as a key variable, but account for its variability.
3. **Refine loan policies** for applicants based on education and housing types to minimize default risk.
4. **Enhance monitoring** of high-risk income groups, particularly those on maternity leave or unemployed.
5. **Utilize correlations** to refine loan offerings, ensuring appropriate annuity structures based on credit and goods prices.
6. **Promote targeted cross-selling** strategies for approved contracts to optimize product offerings.