# EDA Case Study

Credit Application & Previous Application Analysis

By: Mrunal Paunikar

# Problem Statement

- - High rejection rates in credit applications.
- - Unclear drivers for defaults and refusals.
- - Goal: Identify key patterns and provide actionable insights.

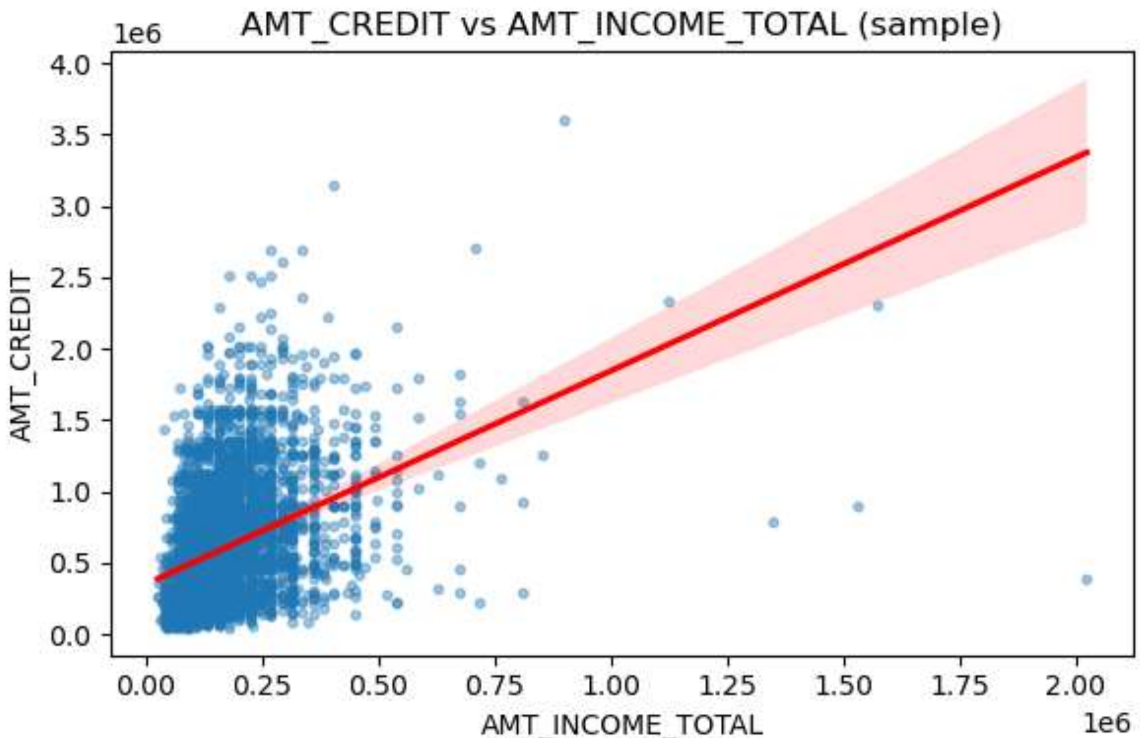
# Steps for Analysis

- 1. Data Cleaning: Handle missing & sentinel values.
- 2. Univariate Analysis: Histograms, boxplots.
- 3. Bivariate Analysis: Scatterplots, jointplots.
- 4. Correlation Analysis: Heatmaps, VIF.
- 5. Previous Applications: Aggregate & analyze.
- 6. Modeling & Interpretability.

# Visualization Techniques

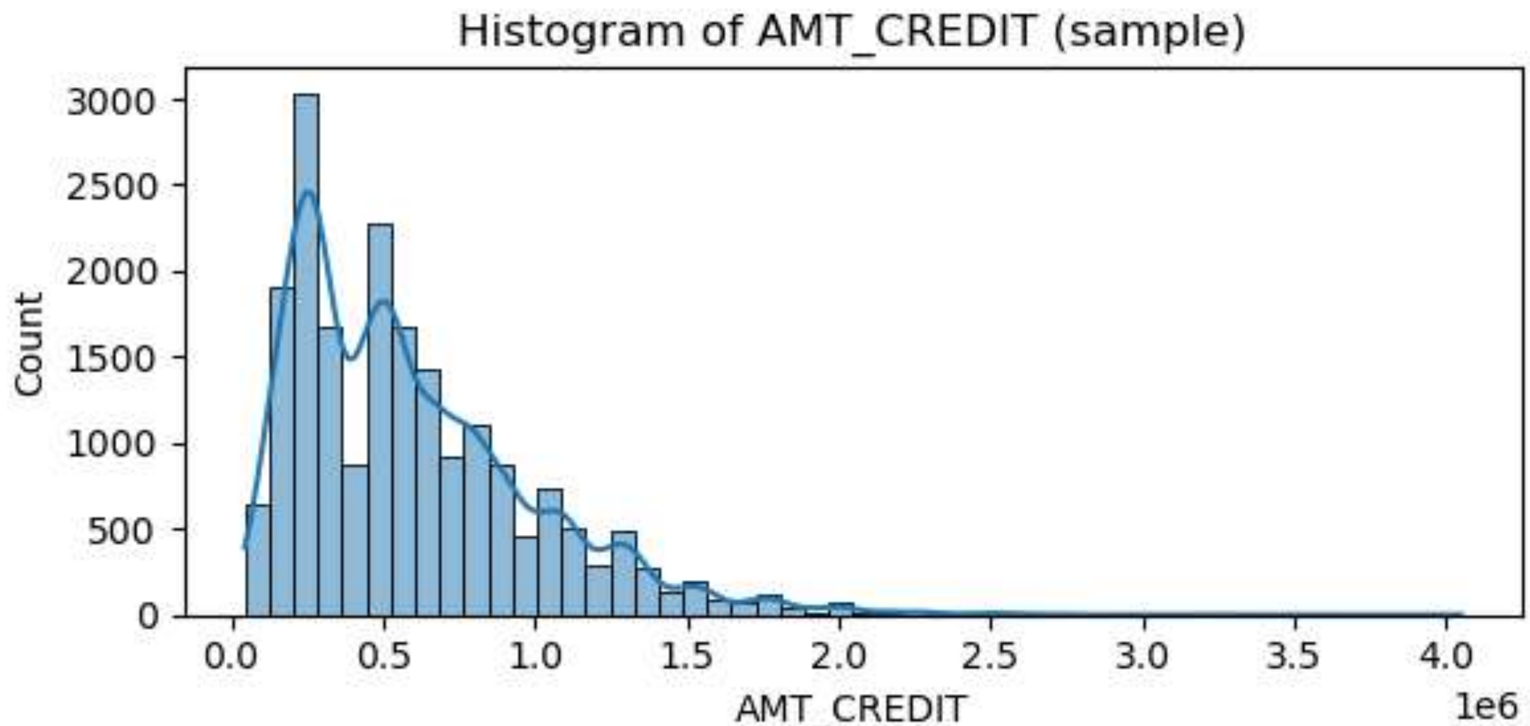
- - Univariate: Histogram, KDE, boxplot.
- - Bivariate: Scatterplot, regression, jointplot.
- - Multivariate: Heatmap, pairplot.
- - Target Analysis: Boxplot AMT_CREDIT by TARGET.
- - Missing Data: Matrix/bar plots.

# Visualization Insight



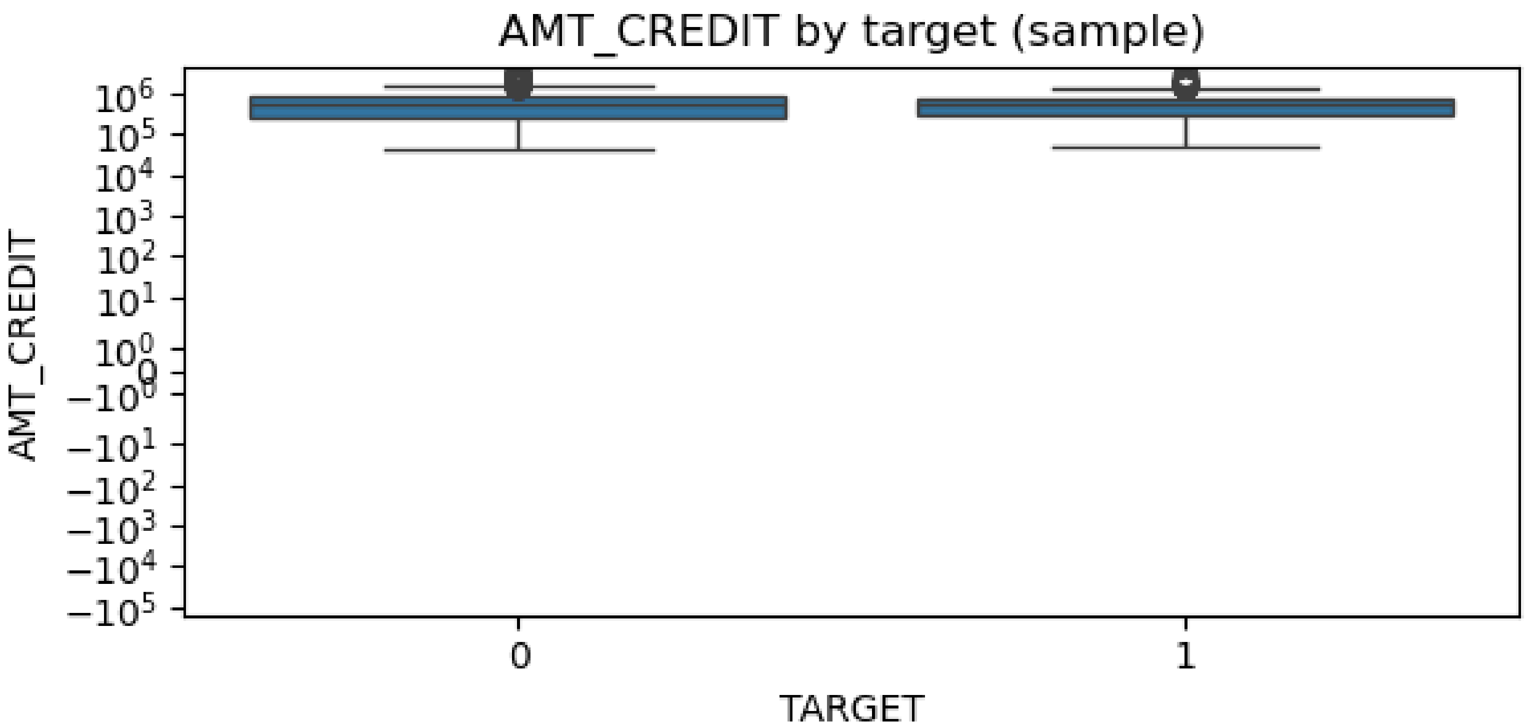

Scatterplot: AMT_CREDIT vs AMT_INCOME_TOTAL
Shows a positive relationship: higher income tends to support higher credit amounts.

# Visualization Insight




Histogram: AMT_CREDIT distribution
Credit amounts are right-skewed; most applicants request lower credit, with few very high outliers.

# Visualization Insight

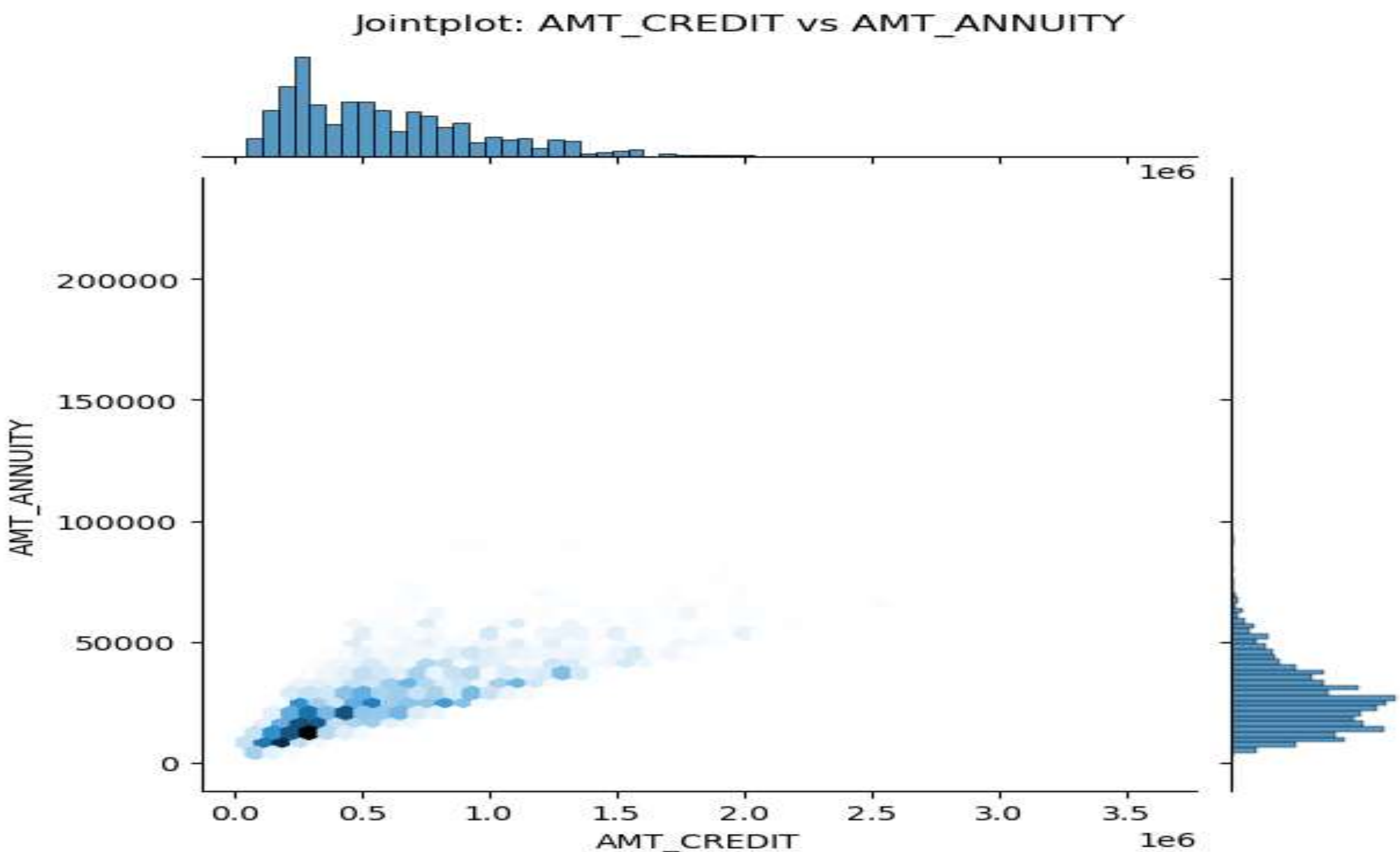



Boxplot: AMT_CREDIT by TARGET
Rejected (1) vs Accepted (0) applicants show overlapping credit levels but rejected tend to cluster at higher amou

# For Understand Purpose

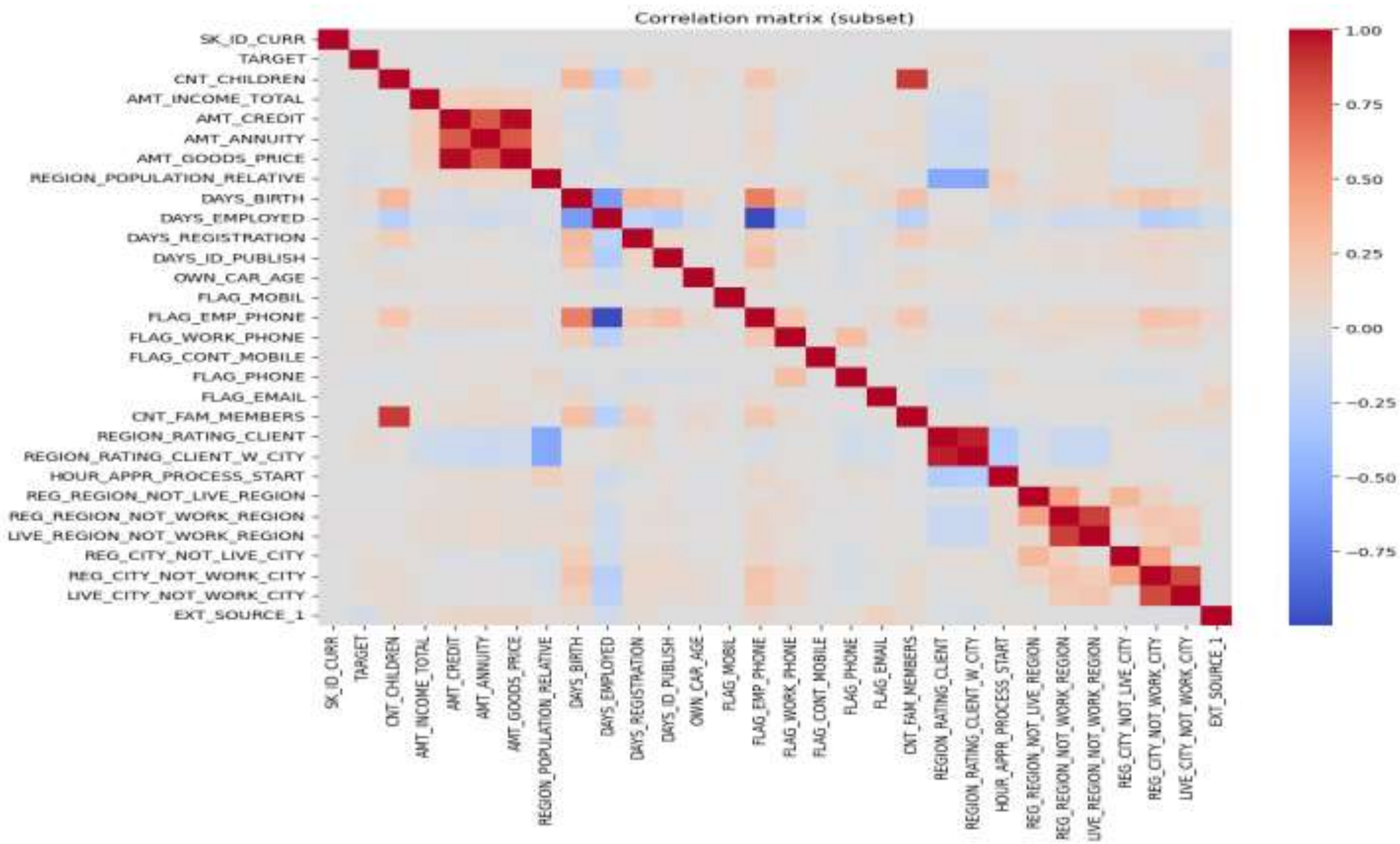

Boxplot: AMT_CREDIT by TARGET
Rejected (1) vs Accepted (0) applicants show overlapping credit levels but rejected tend to cluster at higher amou

# Visualization Insight



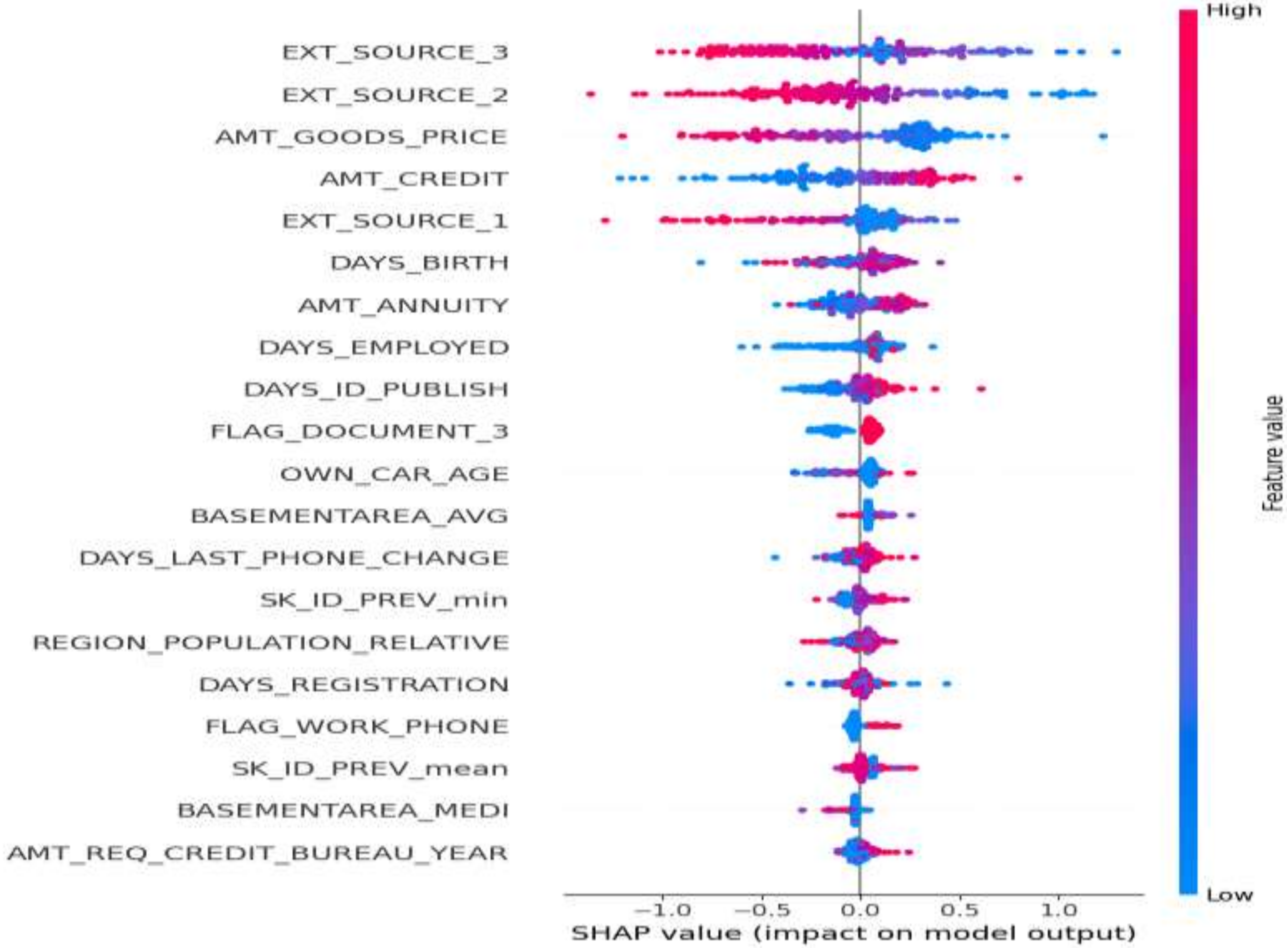

**Jointplot**: AMT_CREDIT vs AMT_ANNUITY
Shows proportionality: higher credit leads to higher annuities, with dense clusters at low to mid levels.

# Visualization Insight




**Correlation Heatmap**
Highlights strong collinearity between income, credit, annuity, and goods price. Useful for feature selection.

# Model Explainability (SHAP)




**Heatmap** Summary Plot
Explains feature impact on model predictions. EXT_SOURCE variables and AMT_GOODS_PRICE are top predictors. Blue = low feature value, Pink = high feature value.

# Root Cause & Findings

- - High loan-to-income ratio drives rejection.
- - Past refusals increase rejection likelihood.
- - Missing demographic info linked to defaults.
- - Employment instability increases risk.
- - DAYS_EMPLOYED extreme values are sentinel placeholders.

# Assumptions

- - TARGET = 1 → rejected/defaulted, 0 → accepted.
- - Sentinel values represent missing entries.
- - Previous status values are accurate.
- - Sampling ~5k rows is representative.

# Solutions & Precautions

- - Engineer features: credit_to_income, annuity_to_income.
- - Cap extreme outliers.
- - Encode categorical variables properly.
- - Apply rejection rules for high-risk ratios.
- - Monitor drift and retrain models.
- - Ensure explainability (SHAP/feature importances).

# Conclusion

- - Credit and income strongly impact rejection.
- - Past refusals and unstable employment predict risk.
- - Combine rules + ML models for fair decisions.
- - Clean, aggregated data supports reliable modeling.
- - Recommend ongoing monitoring & retraining.