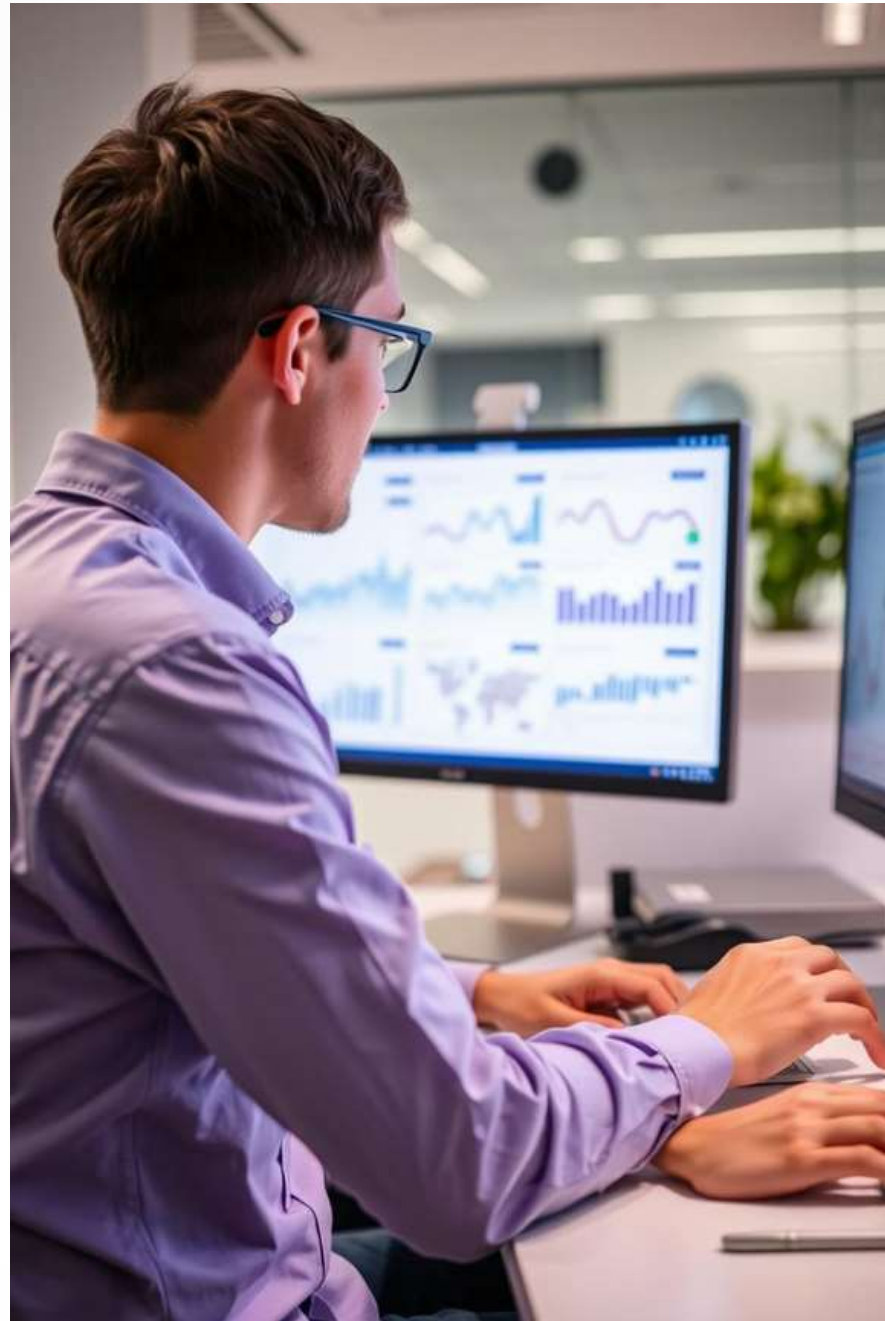


# Exploratory Data Analysis (EDA) in banking

Discover how EDA transforms raw banking data into actionable insights for better loan decisions.

By; **Sohini**  
**Vaibhavi B Raj**

# Introduction to EDA



## What is EDA?

Exploratory Data Analysis helps understand dataset structure, patterns, and relationships before modeling.

## Early Problem Detection

EDA identifies issues in data that could affect analysis quality.

- Maximise the insight in the data set
- Detect outliers and anomalies
- Test underlying assumptions

## Better Decision-Making

It provides insights that improve business and modeling decisions.



John W. Tukey

## EXPLORATORY DATA ANALYSIS



# History of EDA

01

Work of Tukey:1977

02

ABC'S of EDA, Velleman and Hoaglin(1981)

03

More graphical analysis has been developed lately

# How EDA Helps in Jobs



## Spot Insights

Extract valuable patterns from raw data.



## Build Better Models

Create more accurate predictive models.



## Improve Reporting

Enhance quality of data presentations.



## Support Decisions

Provide data-backed business recommendations.

Essential skill for Data Analysts, Business Analysts, and Risk Analysts.



# Where is EDA Applicable

## Banking

Risk profiling and fraud detection



## Healthcare

Patient history and risk prediction



## Insurance

Claims assessment and risk underwriting

## Retail

Customer segmentation and demand forecasting

Conllinets  
Inently moniterd  
pest reathing data  
your d'spcleratient

Eacomts  
Lear redluind wst  
pobliciclers.



agnictent  
Lee reakily yorar  
por appement

Trorofcaturnty  
Eest puoibles, do  
you rdeation lyars

## EDA Step-by-Step Visual



**Business Understanding**



**Data Collection**



**Data Cleaning**



**Univariate Analysis**



**Bivariate Analysis**



**Outlier Detection**



**Correlation Study**



**Insights**

# Problem Statement



## Risk Analytics

Assess credit risk in loan applications.

---



## Customer Types

With and without payment difficulties.

---



## Decision Impact

Rejecting good customers vs. approving risky ones.

---



## Loan Outcomes

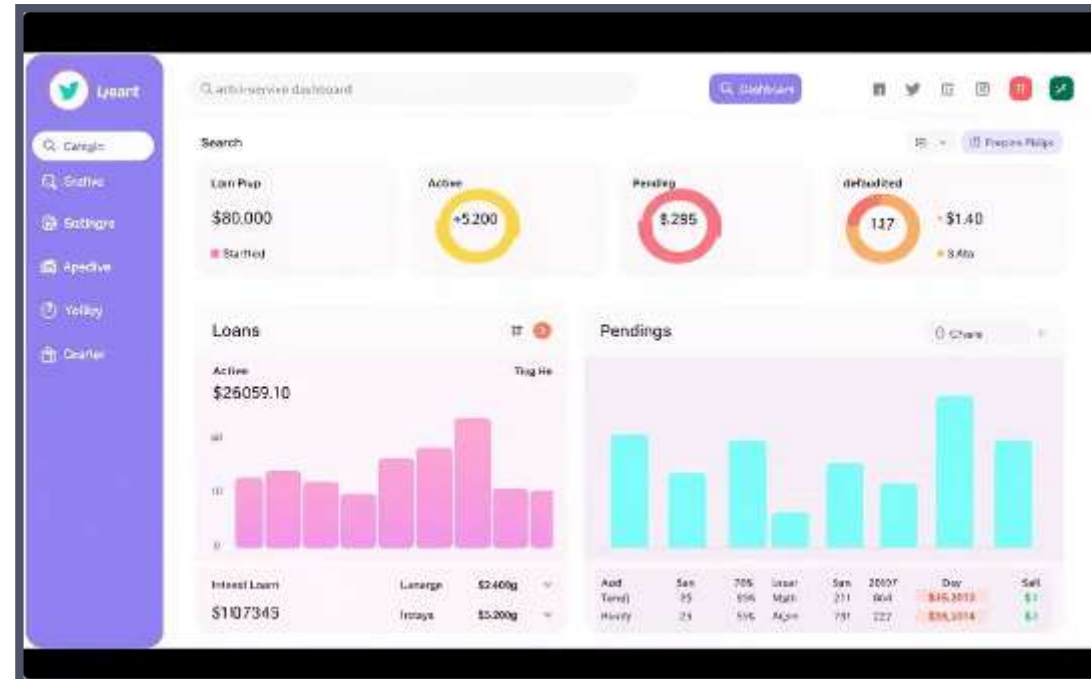
Approved, Refused, Cancelled, Unused Offer.

# Datasets

	Applicant	Applicant	Loan Amount	Interest Amount	Approval Amount	Spouse Interest	Approval Amount	Approval Amount	Approval rate
1	Peter colon	\$21,510.0	150172	779	290	2000	7200	1.05%	270%
2	Men Alory	\$21,510.0	57000	759	779	\$200	4000	2.50%	1587%
3	Loan Amount	\$33,397.70	\$1008	500	500	\$200	3000	3.00%	220%
4	Mat Jorant	457,51.00							
5	Clear Grant	\$67,17.00	\$2500	1579	580	\$800	\$800	3.05%	660%
6	Loan System								
6	Man Maco	161,29.00	15000	579	750	\$300	1900	300%	\$75%
9	Man Lopes	\$17,190.0	\$800	1171	550	300	600	4540	\$70%
18	Man Scors	197,30.30	12000	1771	555	2000	2000	3.00%	\$50%
11	Man Amount	\$77,50.00	11000	800	500	4000	2000	450%	\$25%
11	Loan Amount								
10	Peter Jocus	101,36.00	12800	1172	290	\$300	3018	1250%	\$20%
11	Mat Jorant	173,46.00	300	872	700	2000	2500	1450%	510%
17	Peter Jorant	500							
11	Peter Jorant	157,19.80	\$500	1173	700	4500	2600	550%	550%
19	Peter Loan	151,53.00							
27	Peter Jorant	151,35.60	15250	1132	500	\$600	200%	380%	230%

## Application Data

Current application details with customer information and loan specifics.



## Previous Applications

Customer's past loan application history and outcomes.

```
Column Mext. Uniqu Identifier- buy use)
user,id:
DESCAPPI0, : : INT
username: : laftisr(10/)
email
frails''s- eamait11000
email:
```

## Data Dictionary

Detailed explanation of all variables in the datasets.

<https://www.kaggle.com/datasets/sid9300/credit-eda-case-study-data>

# Approach and Methodology



## Data Cleaning

Handled missing data, dropped high-NaN columns (>19%), fixed negatives.

---



## Categorical Fixes

Replaced invalid entries like 'XNA' with logical values.

---



## Univariate & Bivariate Analysis

Studied variable distributions and relationships.

---



## Segmentation

Separate analysis for defaulters vs non-defaulters.

---

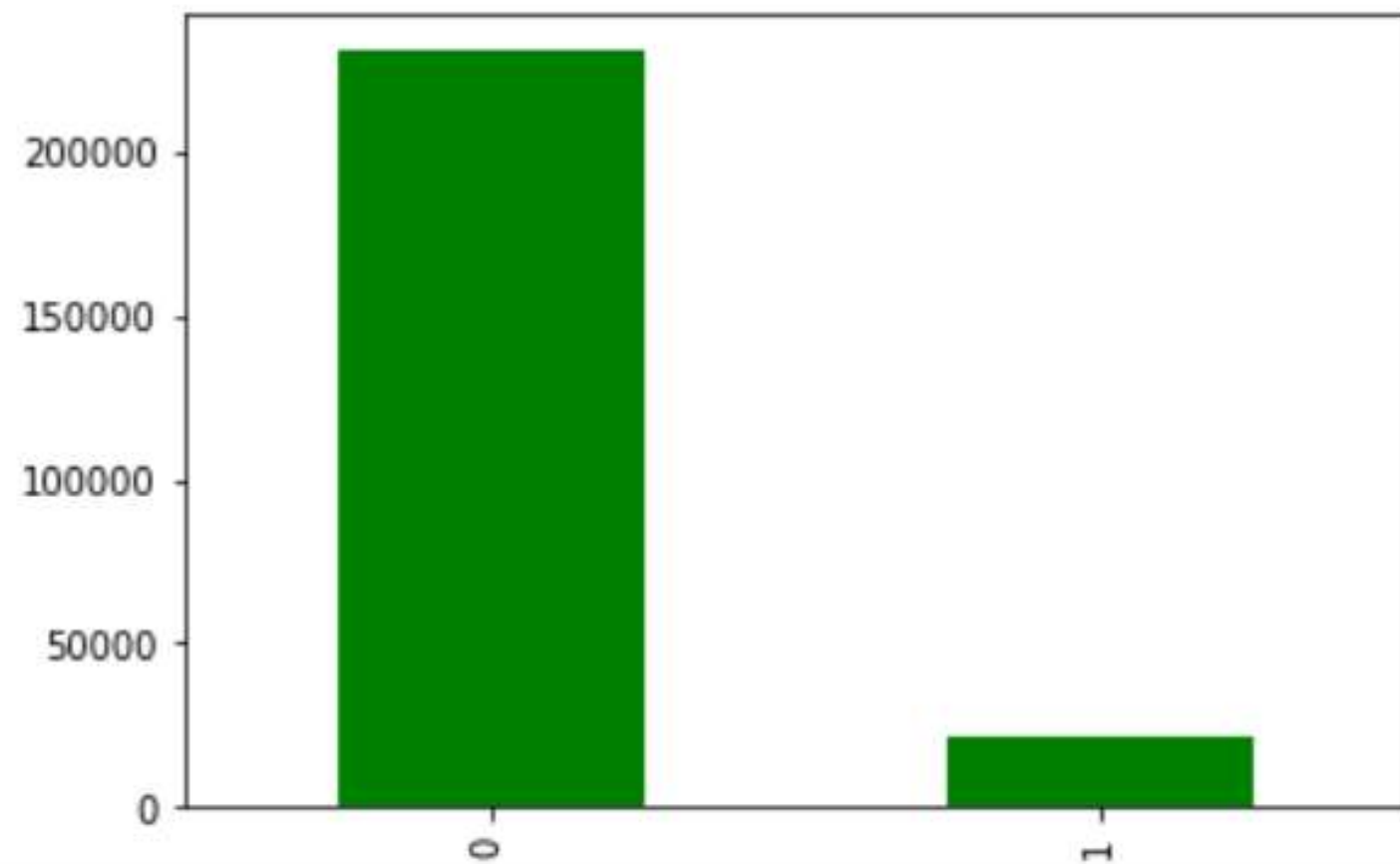


## Correlation & Outliers

Heatmaps to detect key factors; outliers observed via boxplots.

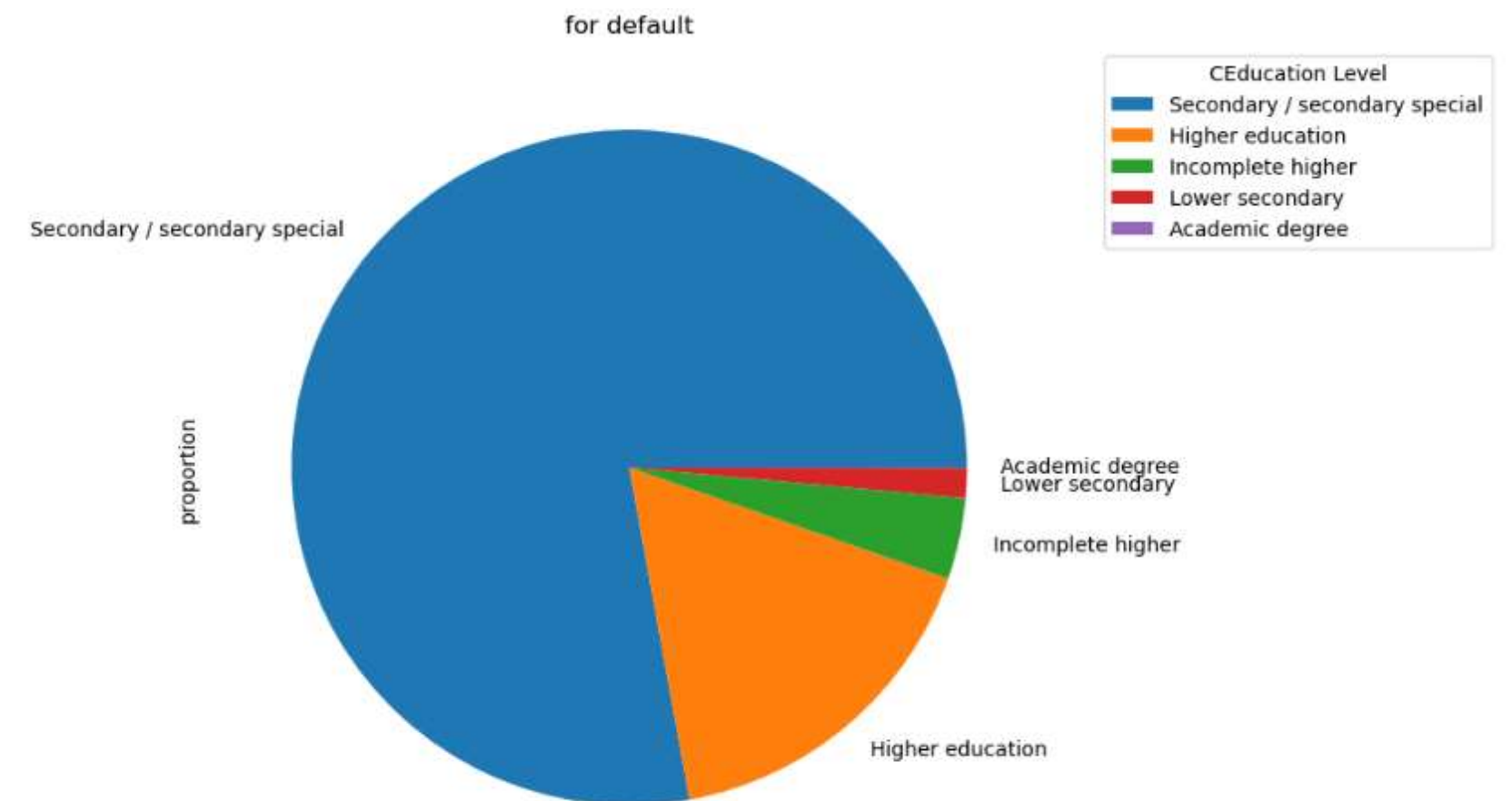
# Target Variable Distribution

Shows non-defaulters dominate



# Education Level Pie Chart

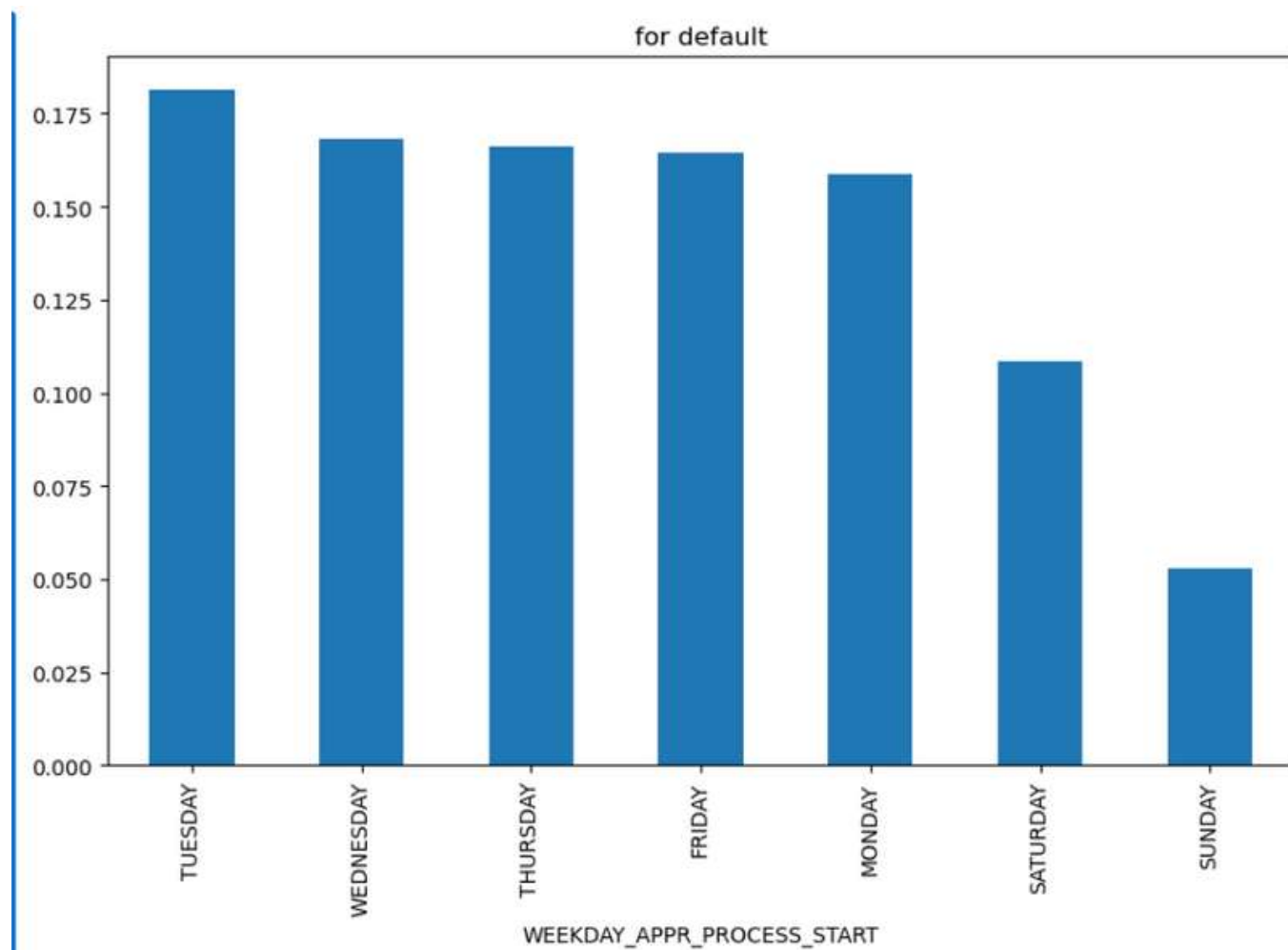
Shows secondary education is most common among defaulters



- 1 = defaulters (customers who failed to repay loans).
- 0 = non defaulter (customers who successfully repaid loans).

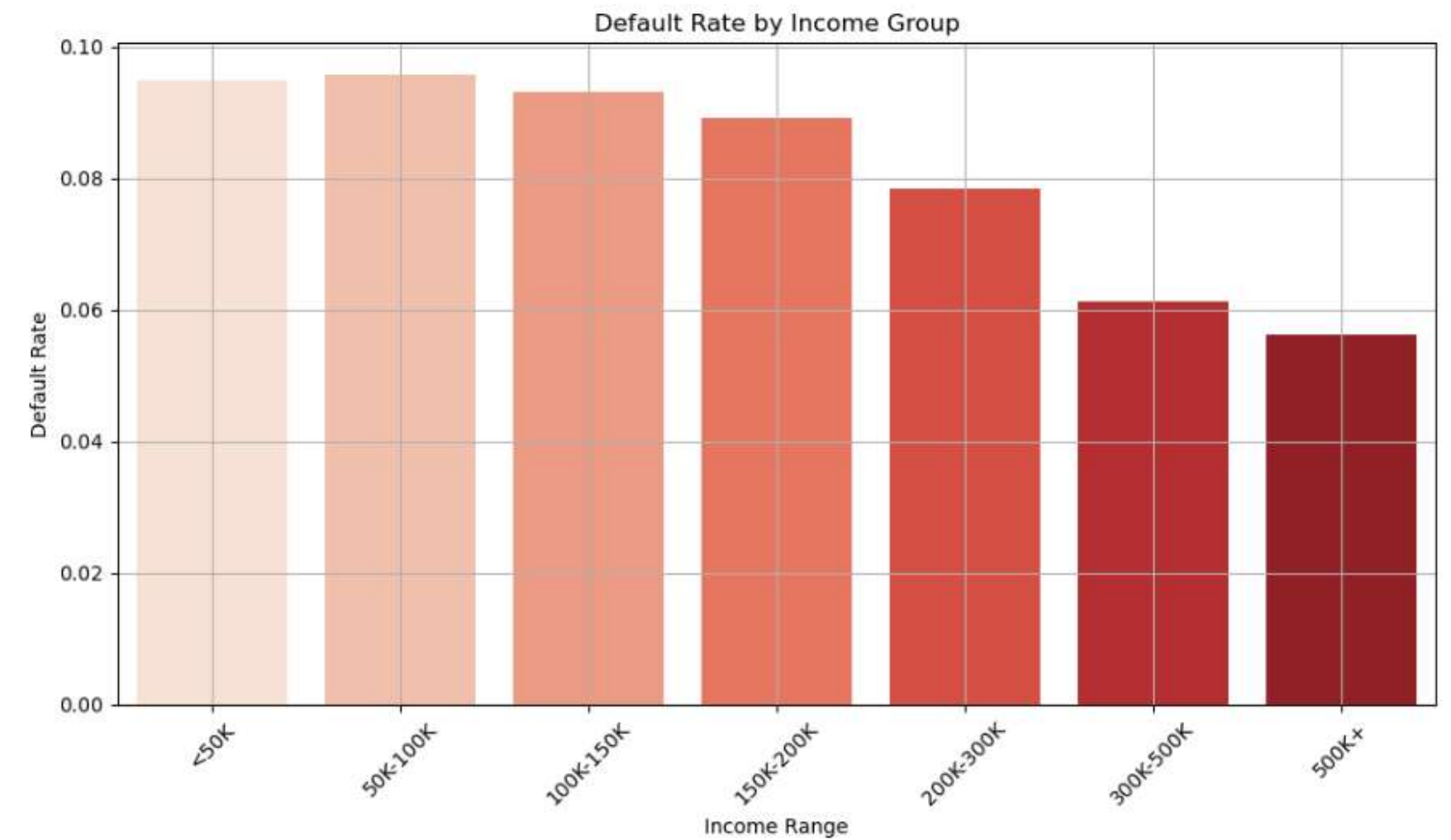
# Application Timing Bar Chart

Shows activity highest on weekdays



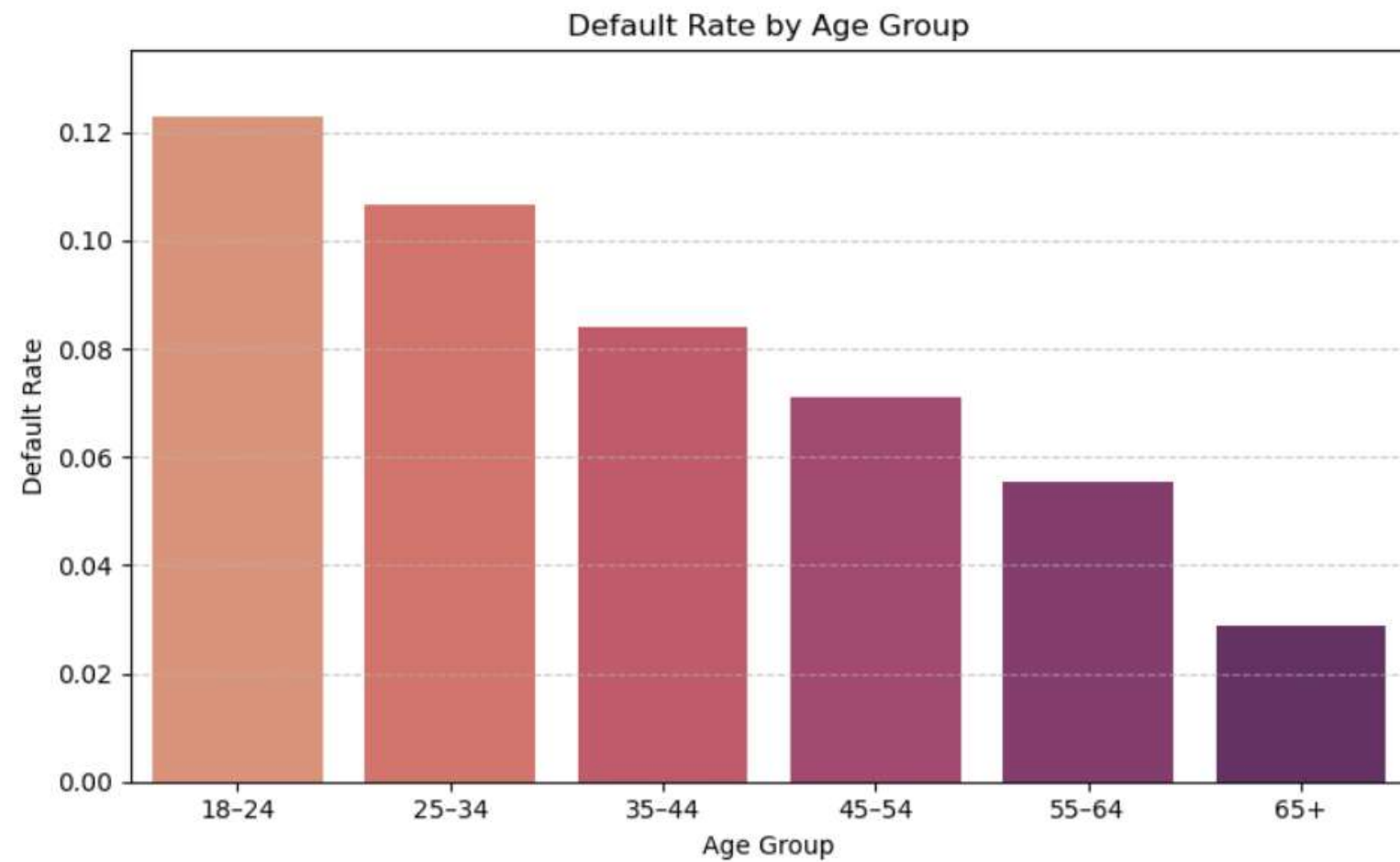
# Income vs Default

Lower income is a significant risk indicator.



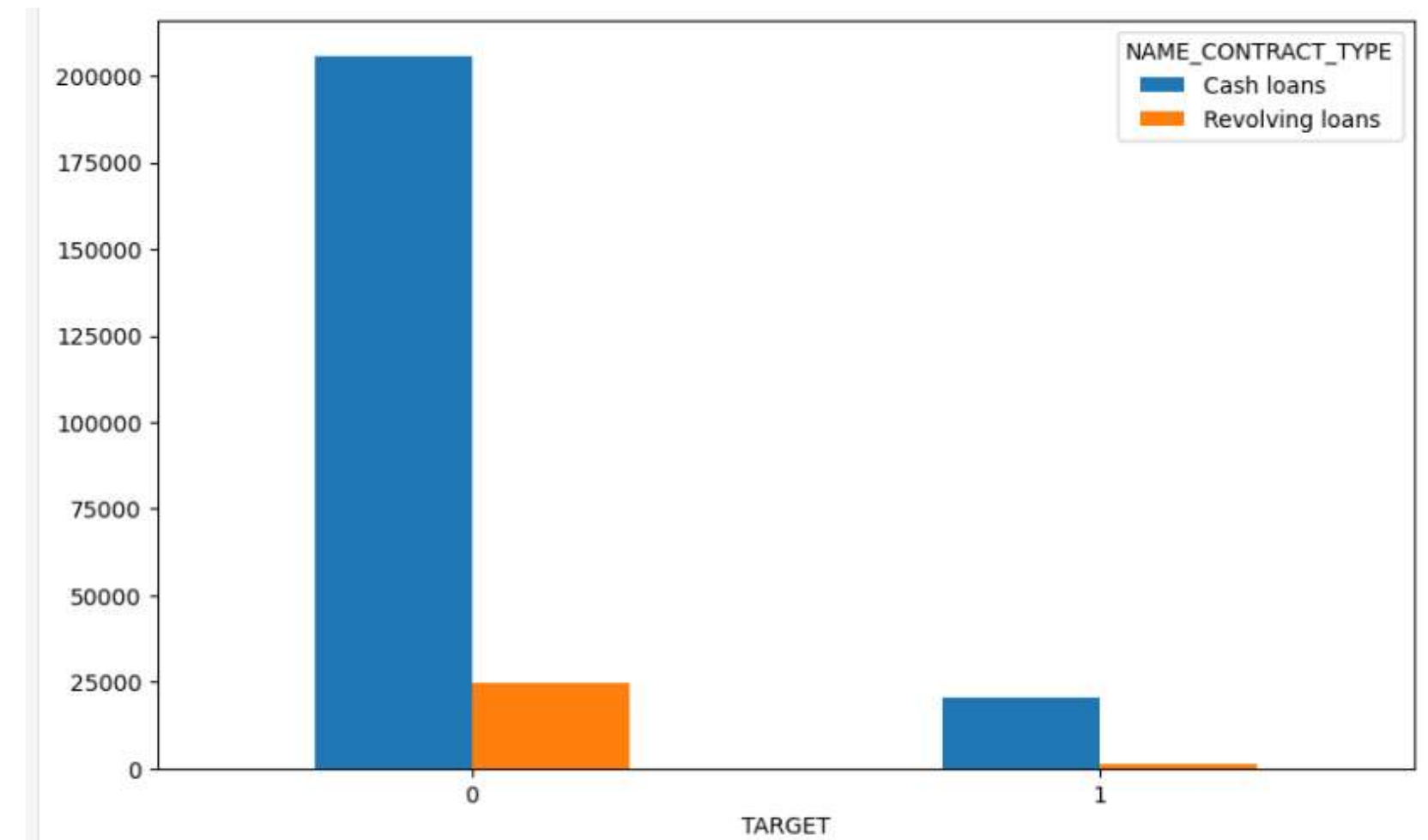
# Age Distribution

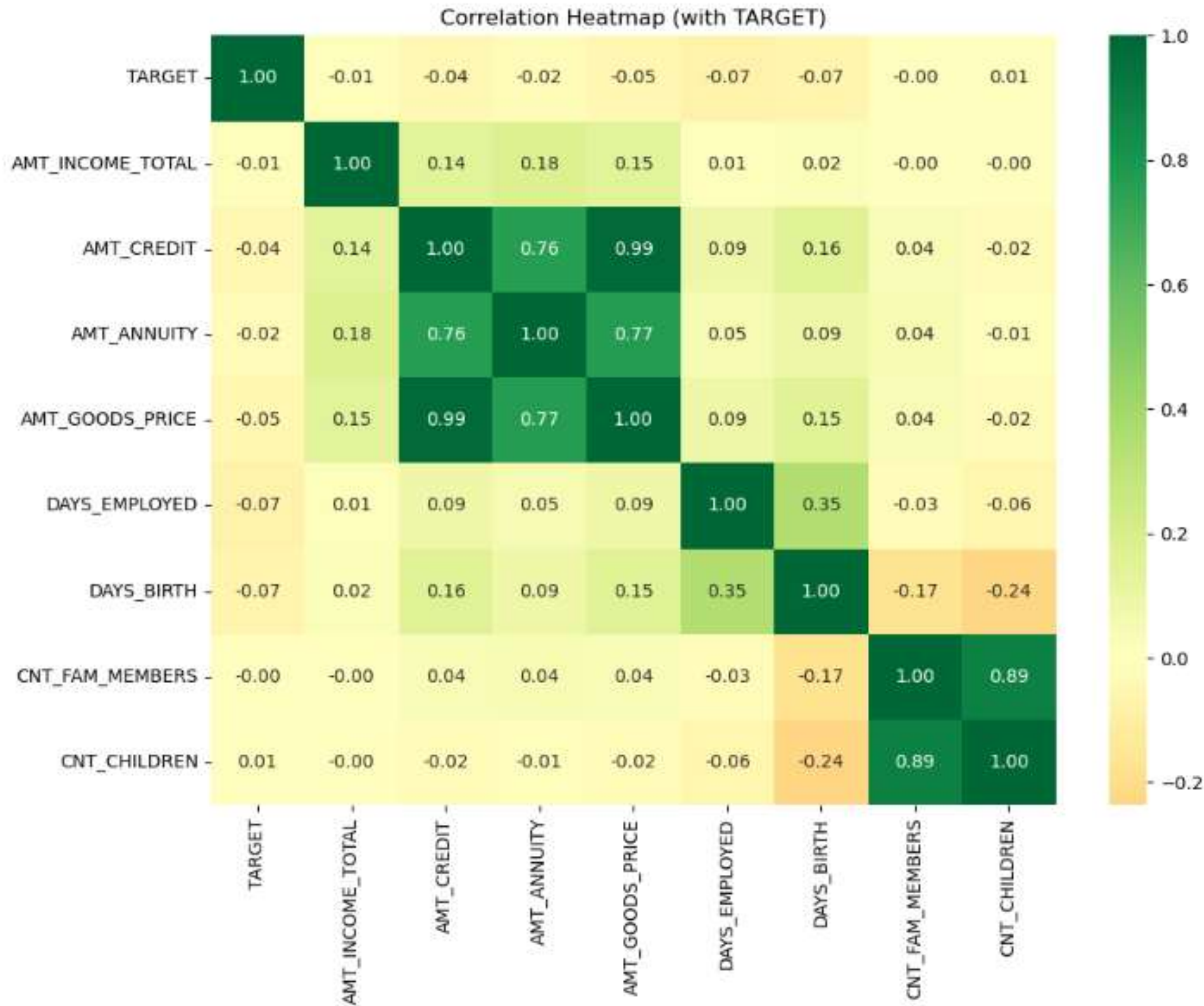
Proves younger customers are riskier



# Loan Type vs Target

Proves cash loans = higher defaults



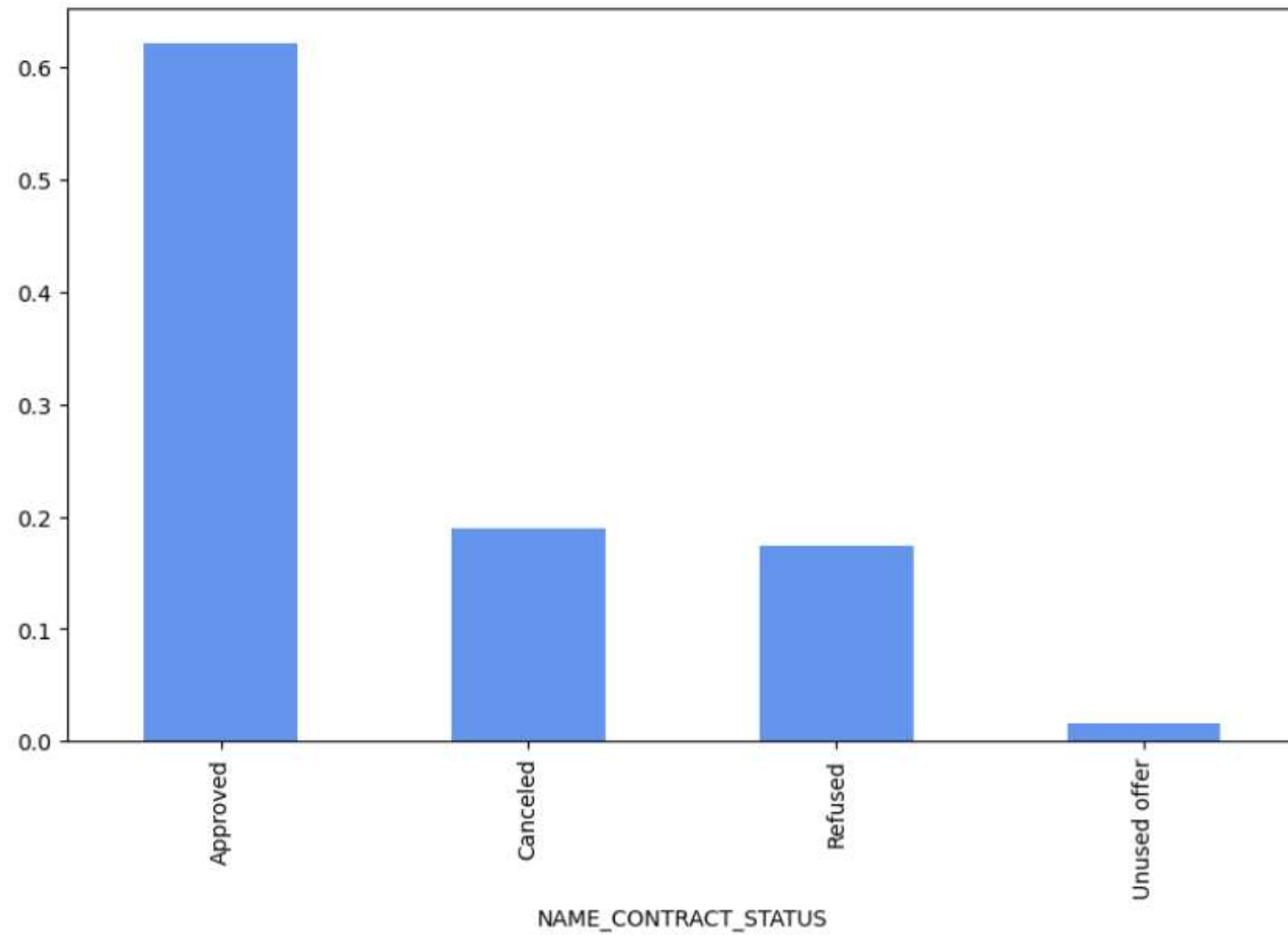


## Correlation Heatmap

Shows strong links: income, credit, employment vs default

# Loan Status Distribution

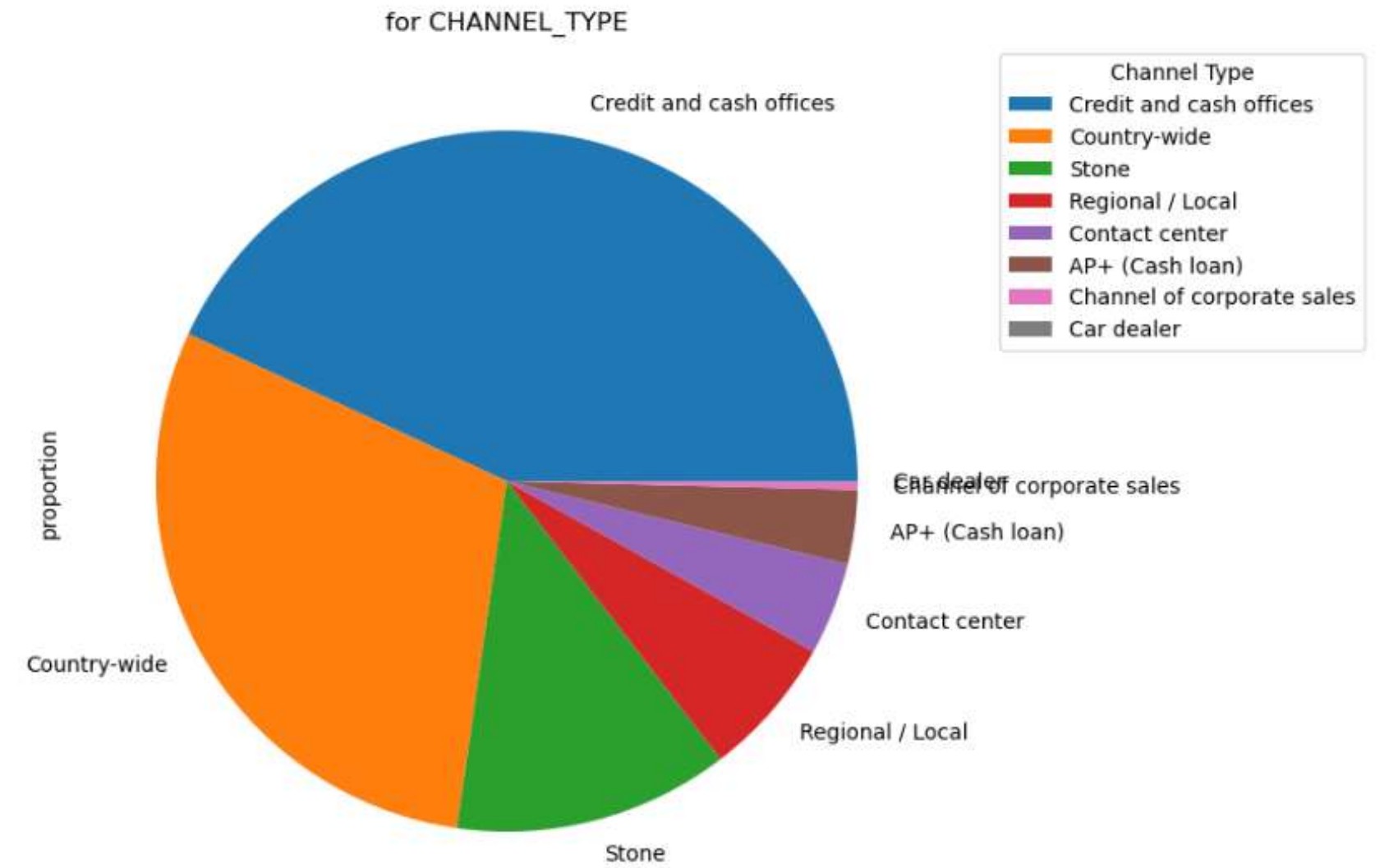
Shows previous refusals are significant



From previous\_application.csv

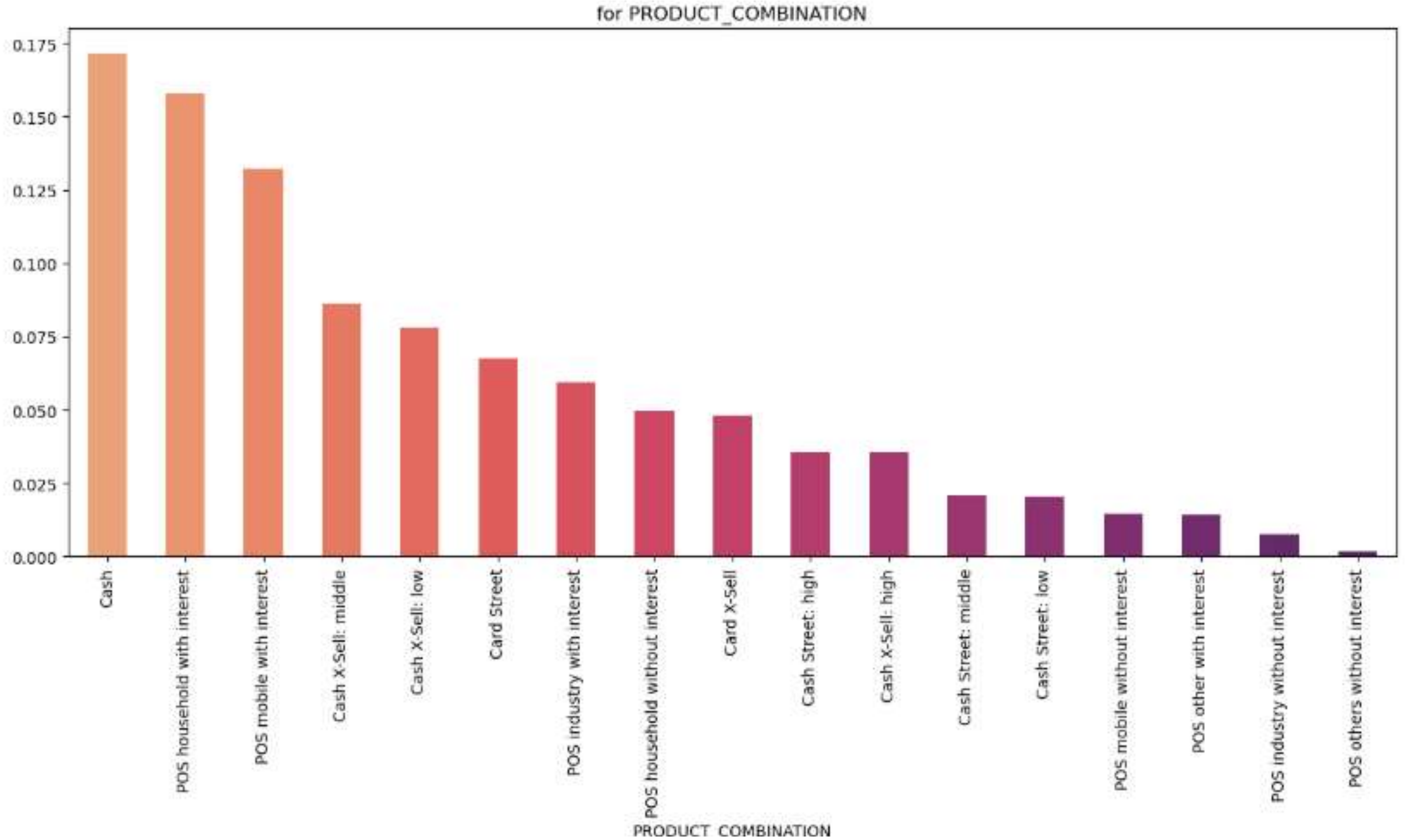
# Channel Type

Supports how loans are processed



# Product Combination

Spot patterns of simple vs bundled loans

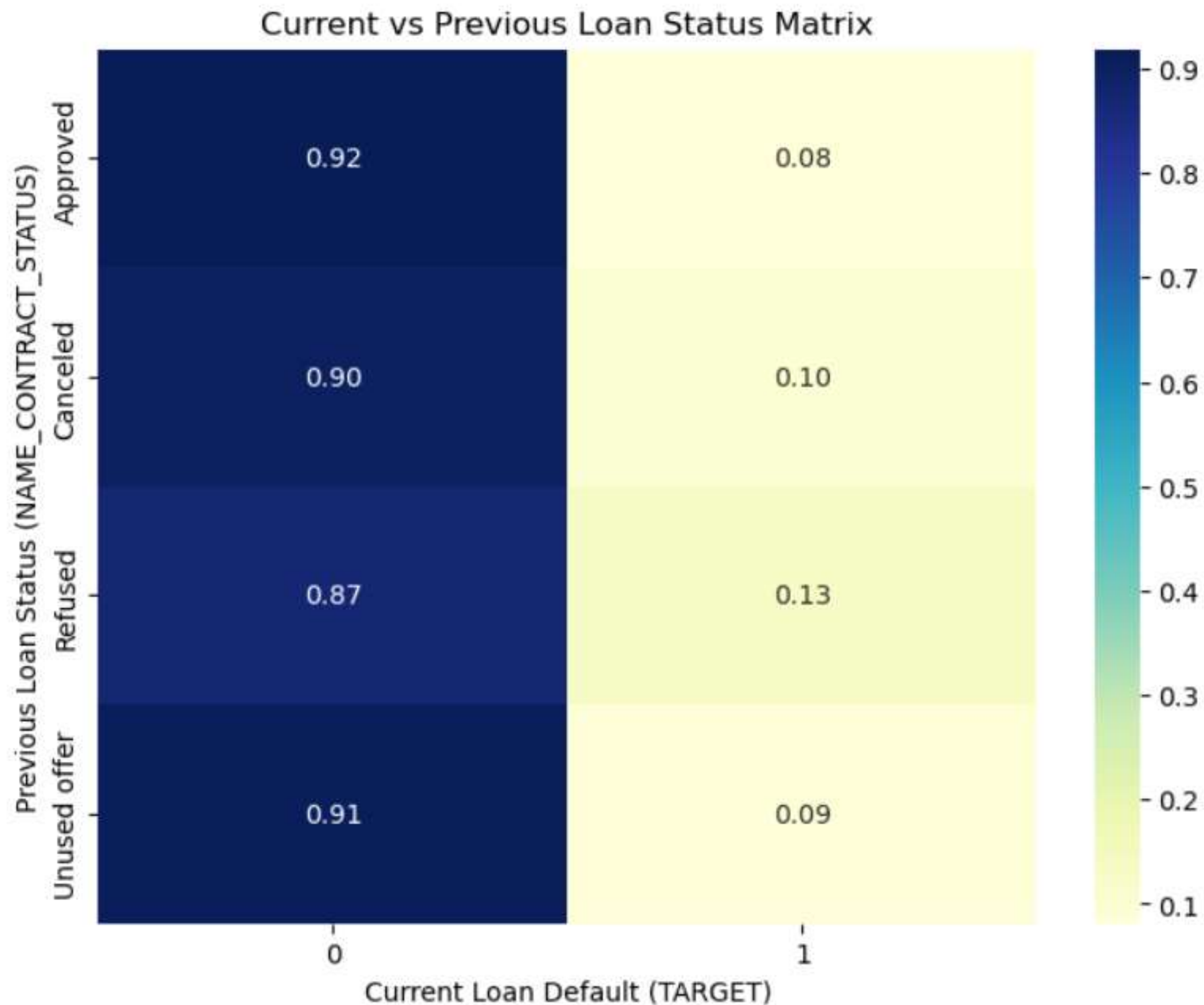


From previous\_application.csv

# Current vs Previous Loan Status

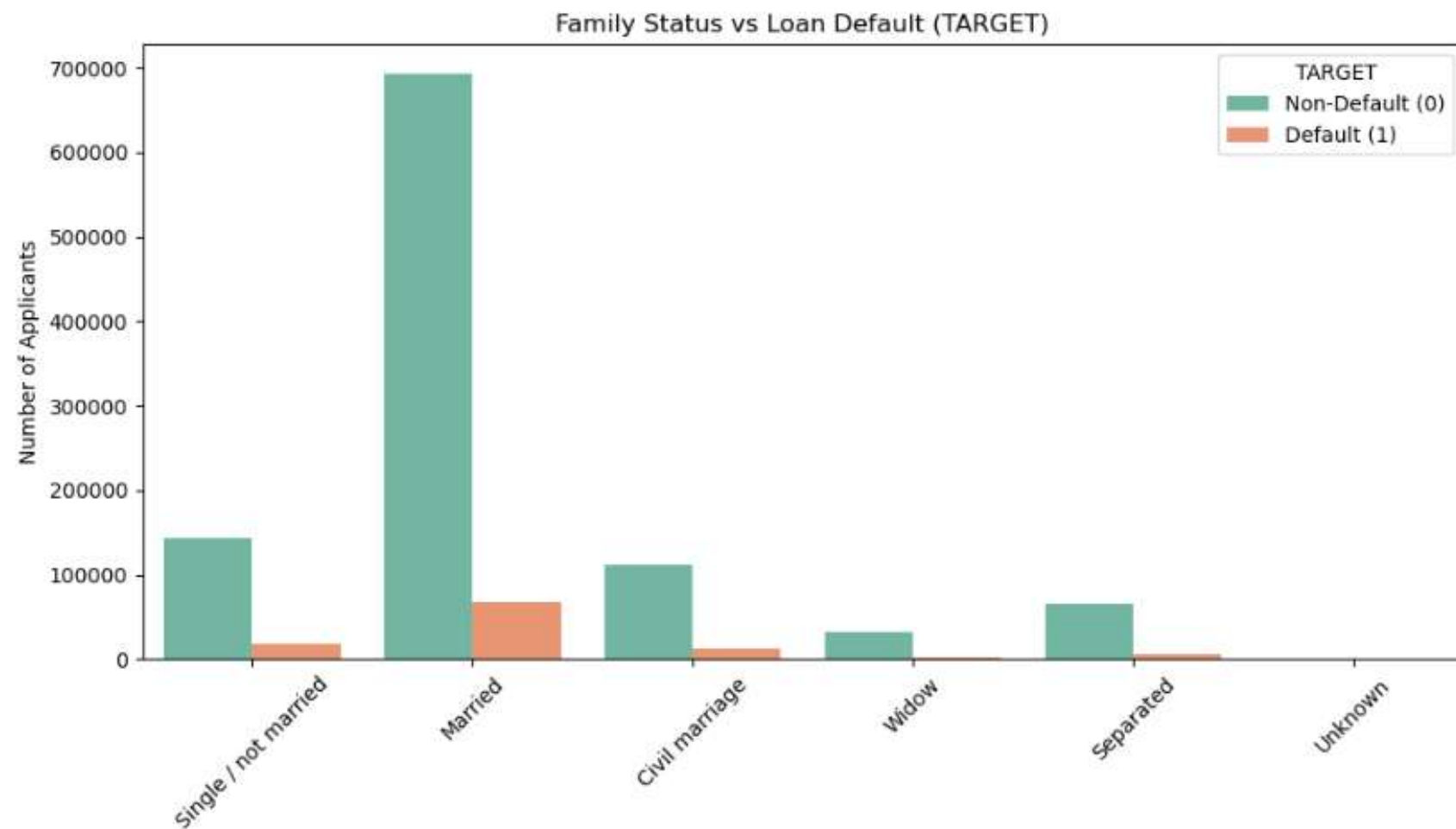
Links past refusals to current defaults

From Merged Dataset



# Family Status vs Loan Status

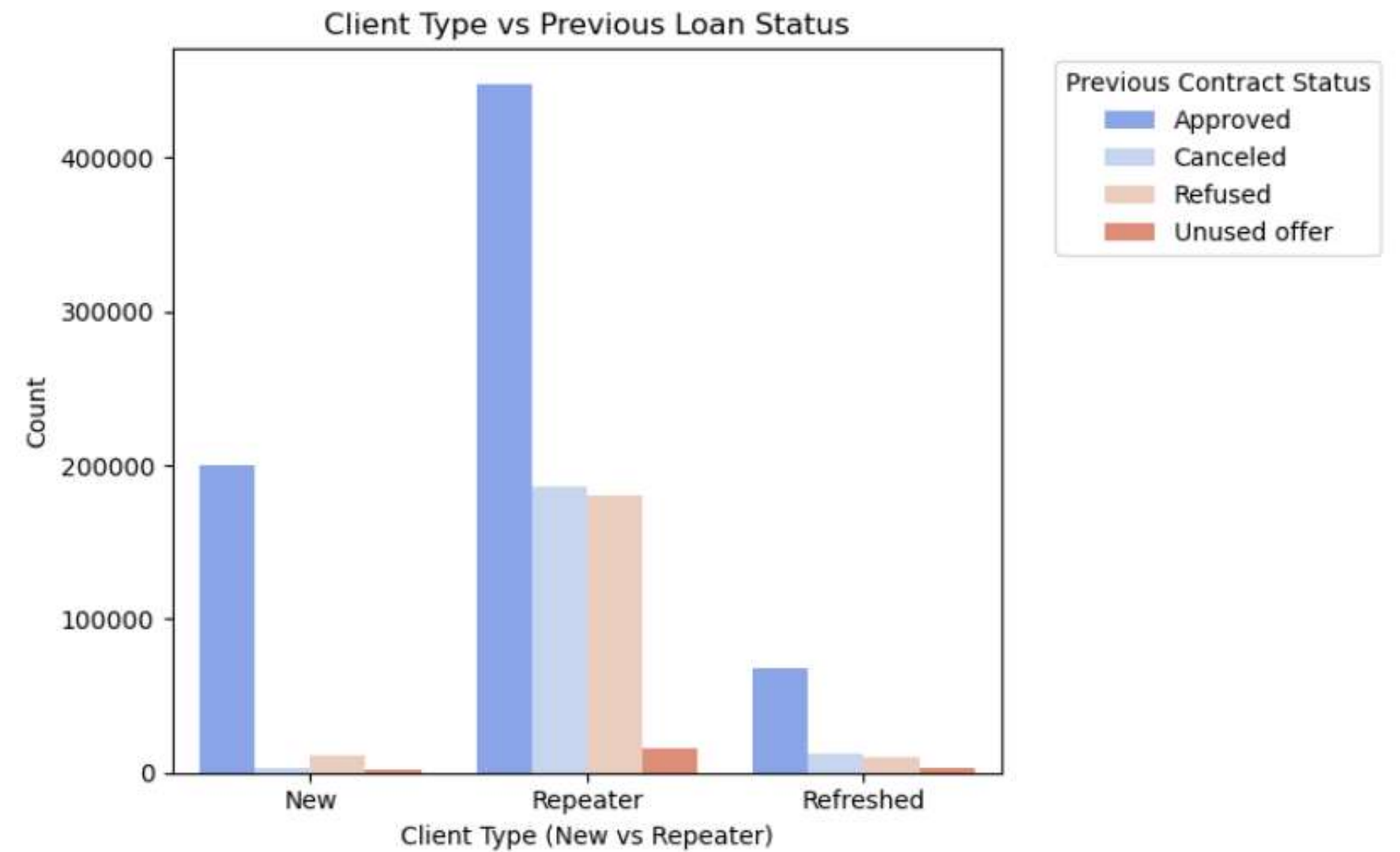
Adds socio-demographic dimension to profile



From Merged Dataset

# Client Type vs Contract Status

Repeating clients more likely to be approved, safer



# Conclusion



## Customer Profile

Most customers are non-defaulters

Secondary education common among defaulters

Application activity highest on weekdays

Strong correlations found between employment, income, and default rates.

## Risk Factors

Lower-income customers show higher risk

Younger customers have higher default rates

Cash loans riskier than revolving loans

