
No credit history? No problem

Access Loan Default Risk Through Demographic & Financial

--- WHAT FACTORS AFFECT DEFAULT

Ricky Xiong
Chien Tran
Viet Nguyen

HOME CREDIT

Introduction

Source: Kaggle

Objective

Investigating whether an applicant's social demographics and wealth factors are important in predicting whether they can repay a loan or not?

Present Outline

1. **Our Data**
2. **Ethical Considerations & Stakeholders**
3. **Missing Values & Data Cleaning**
4. **Exploratory Data Summary**
5. **Selection Methodology**
6. **Prediction Methodology**

Dataset

Home Credit Default Risk

Can you predict how capable each applicant is of repaying a loan?

Overview Data Code Models Discussion Leaderboard Rules

Dataset Description

- **application_{train|test}.csv**
  - This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
  - Static data for all applications. One row represents one loan in our data sample.
- **bureau.csv**

Files
10 files

Size
2.68 GB

Type
csv

License
Subject to Competition Rules

Data Summary

Metric	Values
Number of rows	307511
Number of columns	122
Character columns	16
Numeric columns	106

Dataset

“TARGET” Variable:
Binary classification target (0 or 1)
Indicates loan payment difficulties
1: The client had a late payment of more than X days on at least one of the first Y installments of the loan
0: All other cases (no significant payment difficulties)

Key Variables Overview:
SK_ID_CURR: Unique loan identifier in the sample
CODE_GENDER: Client's gender
AMT_INCOME_TOTAL: Total income of the client
AMT_CREDIT: Total credit amount of the loan
DAYS_BIRTH: Client's age in days (relative to loan application)
NAME_EDUCATION_TYPE: Highest education level achieved
NAME_FAMILY_STATUS: Marital/family status

Ethical Consideration

Data Ownership, Usage and Privacy

- Belong to Home Credit Group
- Terms of Use and Privacy and Ownership Rights

Community and Individual Welfare

- Discrimination by perpetuating existing biases
- Possible result’s outcomes that led to changes for vulnerable group

Stakeholders

Community/
Consumers

Financial
Institutions

Regulatory
Bodies

Home Credit
Groups

Home Credit's
Clients

Missing Values and Data Cleaning

Missing Values Summary

Column	Missing Count	Total Rows	Missing Percentage (%)
COMMONAREA_AVG	171,839	246009	69.85
COMMONAREA_MODE	171,839	246009	69.85
COMMONAREA_MEDI	171,839	246009	69.85
NONLIVINGAPARTMENTS_AVG	170,786	246009	69.42
NONLIVINGAPARTMENTS_MODE	170,786	246009	69.42
NONLIVINGAPARTMENTS_MEDI	170,786	246009	69.42

Missing Values and Data Cleaning

Numerical Columns with Missing Data and Summary Statistics

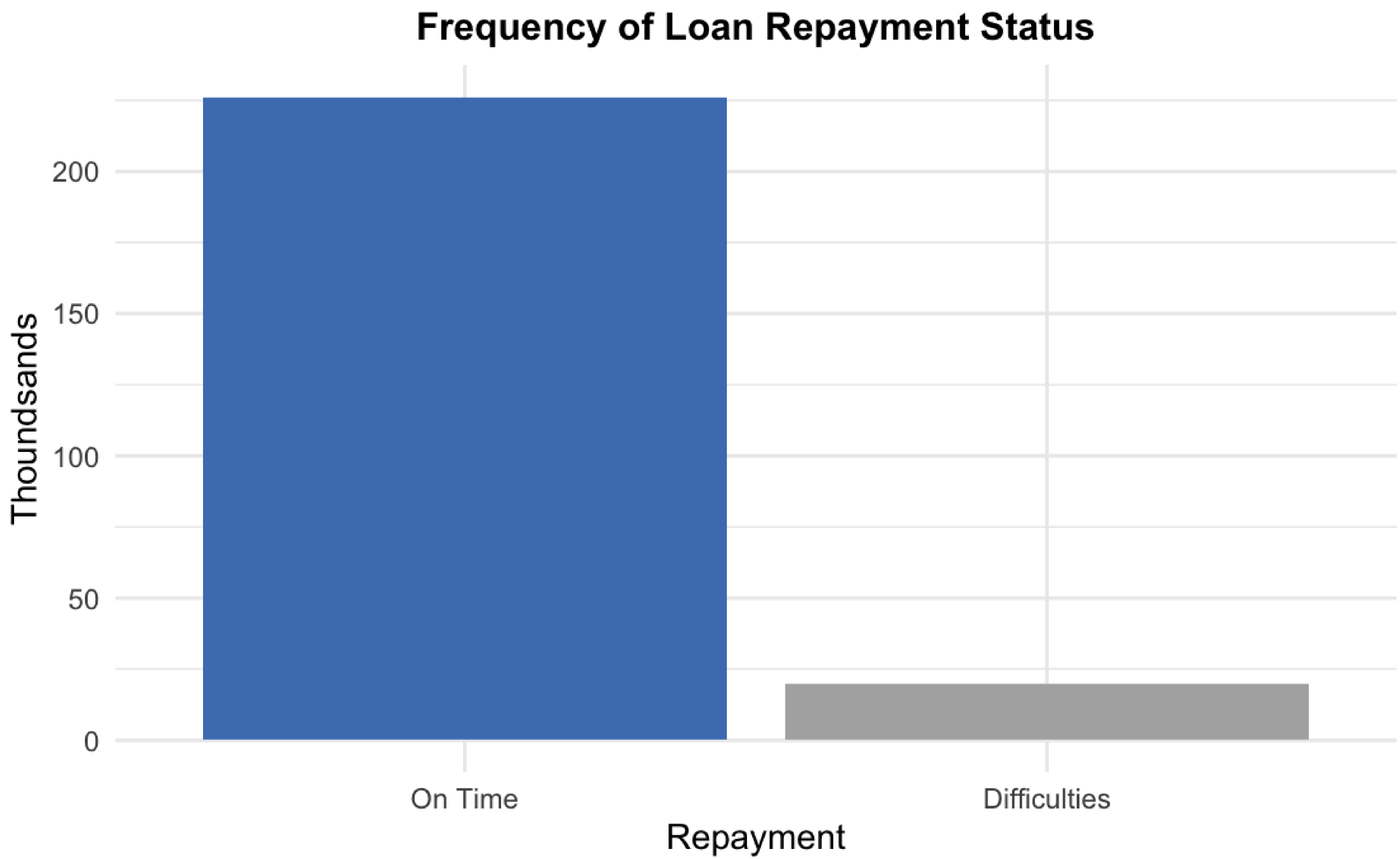
Variable	Percentage (%)	Data Type	Mean	Median	Standard Deviation
EXT_SOURCE_3	19.87	numeric	0.51	0.54	0.20
AMT_REQ_CREDIT_BUREAU_HOUR	13.53	integer	0.01	0.00	0.08
AMT_REQ_CREDIT_BUREAU_DAY	13.53	integer	0.01	0.00	0.11
AMT_REQ_CREDIT_BUREAU_WEEK	13.53	integer	0.03	0.00	0.21
AMT_REQ_CREDIT_BUREAU_MON	13.53	integer	0.27	0.00	0.91
AMT_REQ_CREDIT_BUREAU_QRT	13.53	integer	0.26	0.00	0.61
AMT_REQ_CREDIT_BUREAU_YEAR	13.53	integer	1.90	1.00	1.87
OBS_30_CNT_SOCIAL_CIRCLE	0.33	integer	1.43	0.00	2.43
DEF_30_CNT_SOCIAL_CIRCLE	0.33	integer	0.14	0.00	0.45
OBS_60_CNT_SOCIAL_CIRCLE	0.33	integer	1.41	0.00	2.40
DEF_60_CNT_SOCIAL_CIRCLE	0.33	integer	0.10	0.00	0.36

Missing Values and Data Cleaning

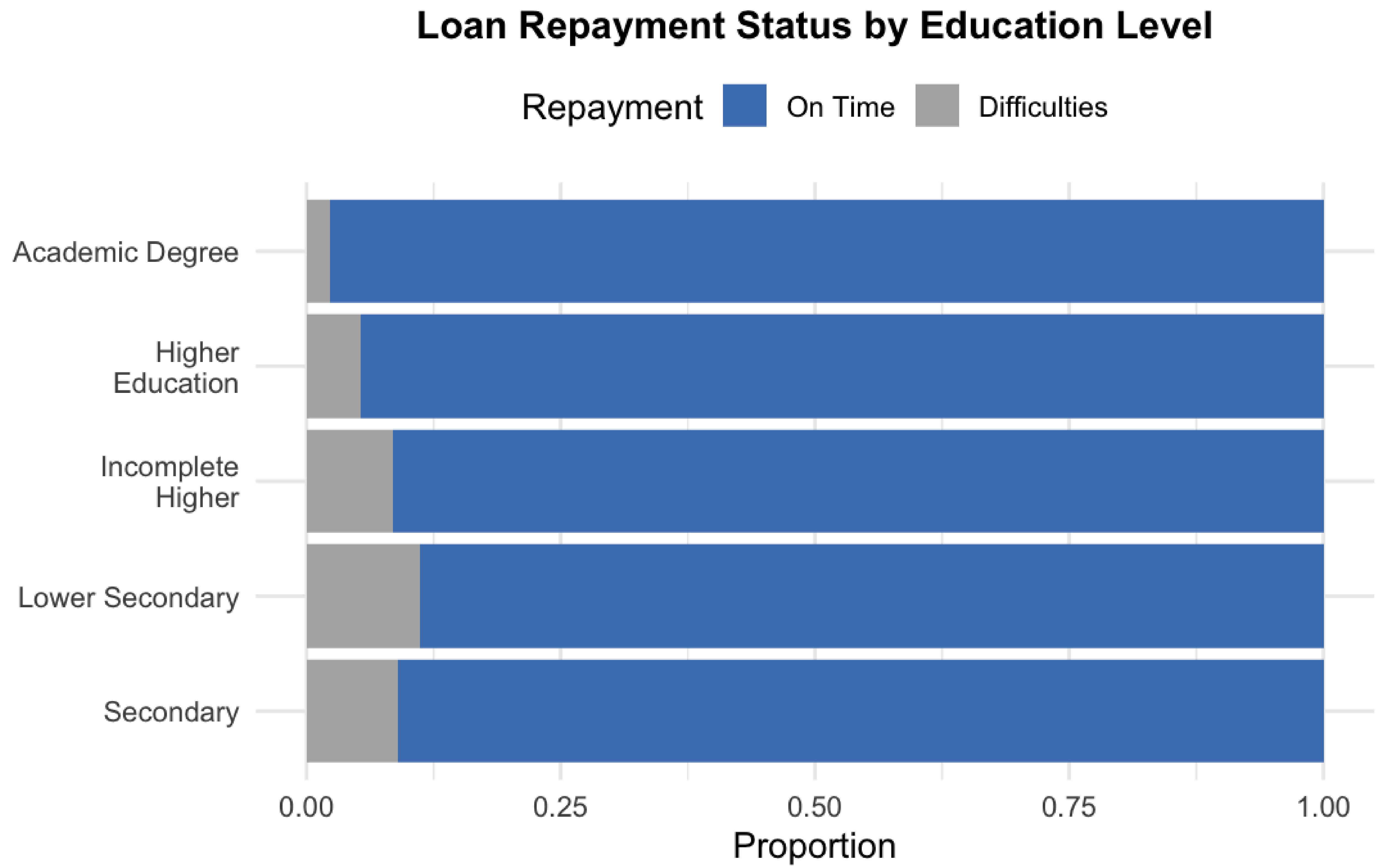
Character Columns Summary

Variable	Missing Count	Percentage (%)	Unique Values	Most Frequent Value
OCCUPATION_TYPE	96,391.00	31.35	19	Laborers
NAME_TYPE_SUITE	1,292.00	0.42	8	Unaccompanied
NAME_CONTRACT_TYPE	0.00	0.00	2	Cash loans
CODE_GENDER	0.00	0.00	3	F
FLAG_OWN_CAR	0.00	0.00	2	N
FLAG_OWN_REALTY	0.00	0.00	2	Y
NAME_INCOME_TYPE	0.00	0.00	8	Working
NAME_EDUCATION_TYPE	0.00	0.00	5	Secondary / secondary special
NAME_FAMILY_STATUS	0.00	0.00	6	Married
NAME_HOUSING_TYPE	0.00	0.00	6	House / apartment
WEEKDAY_APPR_PROCESS_START	0.00	0.00	7	TUESDAY
ORGANIZATION_TYPE	0.00	0.00	58	Business Entity Type 3

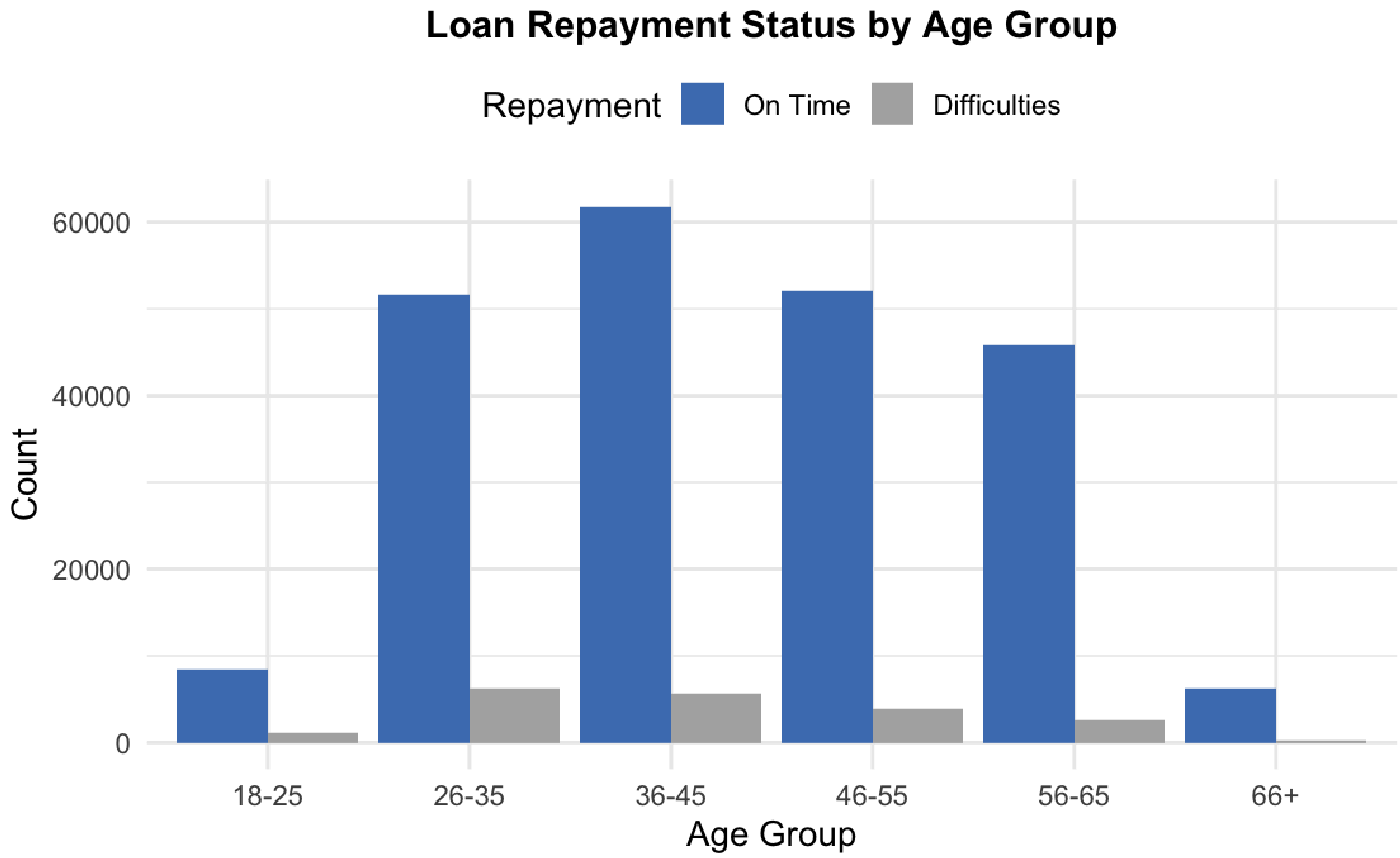
Exploratory Data Analysis



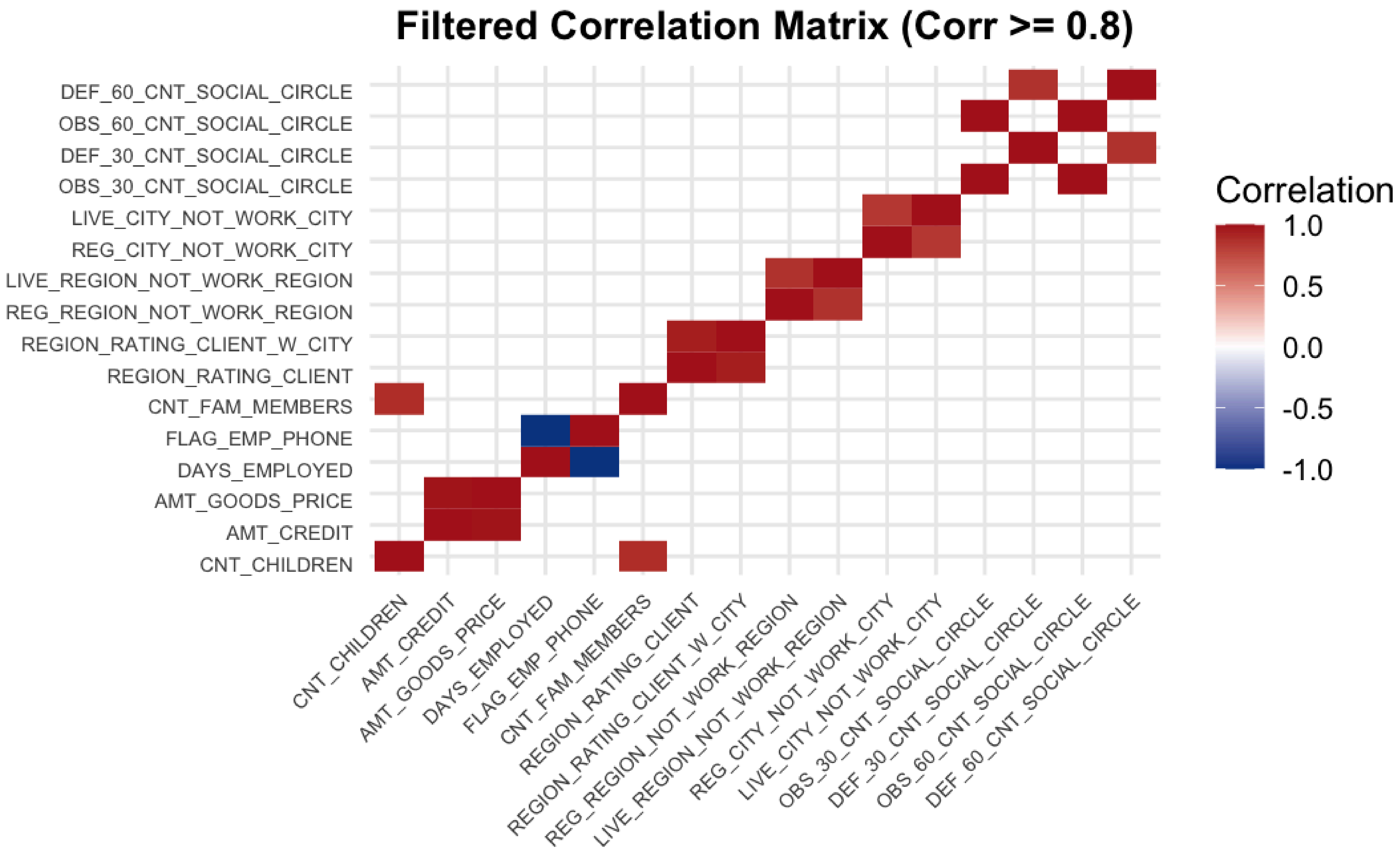
Exploratory Data Analysis



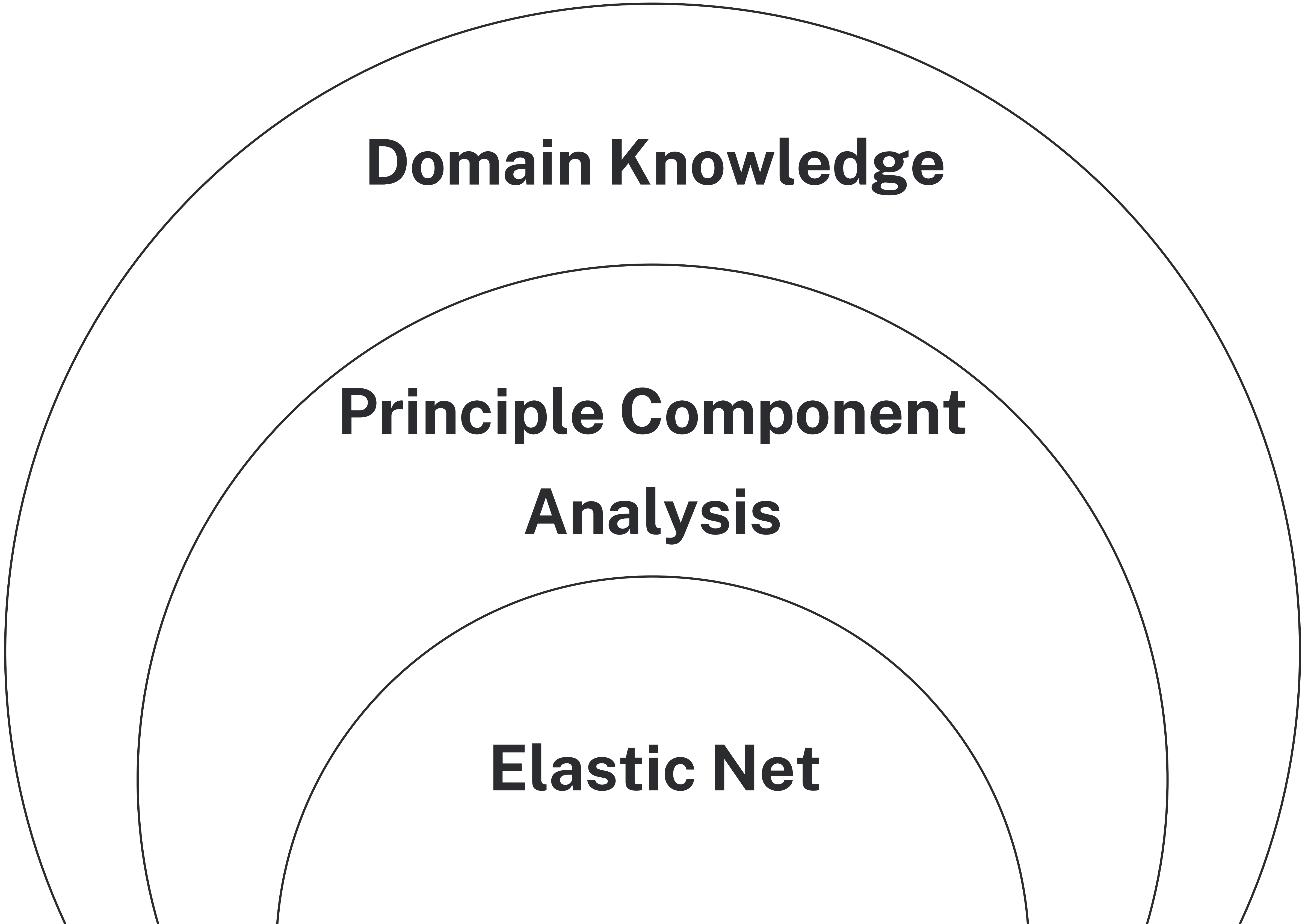
Exploratory Data Analysis



Selection Methodology



Selection Methodology



Prediction Methodology

We will use **Precision** as our metric since False positives (predicting repayment when the loan won’t be repaid) are more costly.

Prediction Methodology

Logistic Regression

Base Line Model for comparision

Random Forest Classification

Can be use for non-linear relationship

Other Model

Considering LightGBM and XGBoosts Model

References

- https://chatgpt.com/
- https://en.wikipedia.org/wiki/Home_Credit
- https://www.homecredit.net/about-us.aspx/#who-we-are
- https://www.kaggle.com/competitions/home-credit-default-risk/overview
- https://www.pewtrusts.org/en/research-and-analysis/articles/2023/01/24/student-loan-borrowers-with-certain-demographic-characteristics-more-likely-to-experience-default
- https://www.urban.org/urban-wire/demographics-income-driven-student-loan-repayment

Thank you!