Credit EDA Assignment

By – Rachit Jha

Problem Statements

Problem Statement -1

Analyze consumer finance loan data using Exploratory Data Analysis (EDA) to identify key factors influencing loan repayment and default. The aim is to develop insights that will help improve loan approval decisions, minimizing financial risk while ensuring that creditworthy applicants are not unjustly denied loans.

Problem Statement -2

Conduct an Exploratory Data Analysis (EDA) on loan application data to identify patterns and key factors influencing loan repayment and default. Address data quality issues such as missing values and outliers, analyze data imbalance, and perform univariate and bivariate analyses. The goal is to derive insights that can inform better loan approval decisions and minimize financial risk. Present findings through visualizations and a comprehensive report, highlighting the most significant variables affecting loan defaults.

Abstract

Analysis Design

- The analysis is structured into three main steps:
- **Exploratory Analysis**: Gain a comprehensive understanding of the dataset, clean it, and identify key columns crucial for analysis.
- **Analytical Techniques**: Conduct univariate, bivariate, and multivariate analyses on both categorical and numerical variables related to the target.
- **Identification of High-Risk Variables**: Identify variables that contribute significantly to predicting high-risk customers.

Domain Variables Considered:

- **TARGET**: Identifies customers who have delayed at least one payment, serving as a primary variable for analysis.
- **NAME_CONTRACT_STATUS**: Indicates the status of previous loan applications, crucial for assessing risk.
- **Assumption**: All customers with at least one missed payment are treated equally, regardless of the number of missed payments.

Understanding the Data Sets

This is the first step of analysis and focus was preparing the data set to perform predictive analysis

Understand the

- *data type*
- *Column types*
- *Identify the columns with missing / null values and outliers*

- *Removing columns with missing value*
- *Finding different approach to treat outliers*
- *Fixing the column with incorrect data types*

In this step, objective was

- *Creating required columns(binning)*
- *Removing irrelevant columns*
- *Creating required data frame*

Exploring Data Sets – Categorical Variable

Univariate Exploration

We examine the dataset to understand the distribution of defaulters (target = 1) and non-defaulters (target = 2) across various variables.

Observations:

- **Fig 1: Gender**
  - Defaulters have a higher percentage of males.
- **Fig 2: Age Group**
  - There is a higher proportion of customers in their 30s among defaulters.
- **Fig 3: Education**
  - Defaulters show a higher percentage of individuals with secondary education.
- **Fig 4: Income Type**
  - Defaulters exhibit a higher proportion of individuals with working-type income.

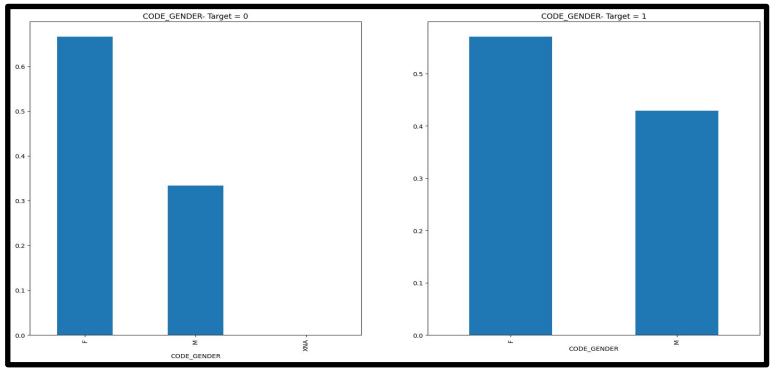


Fig 1

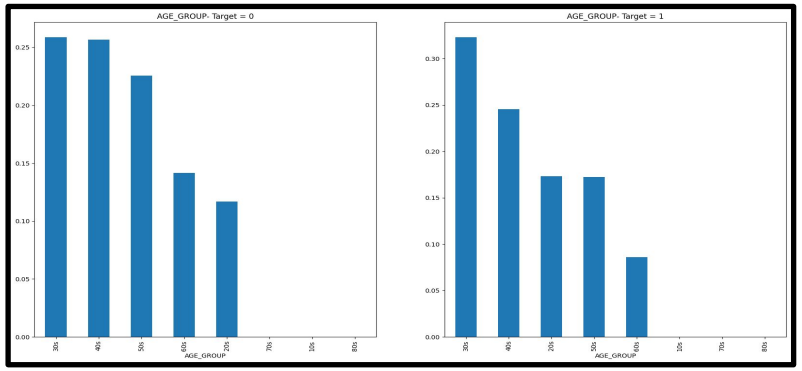


Fig 2

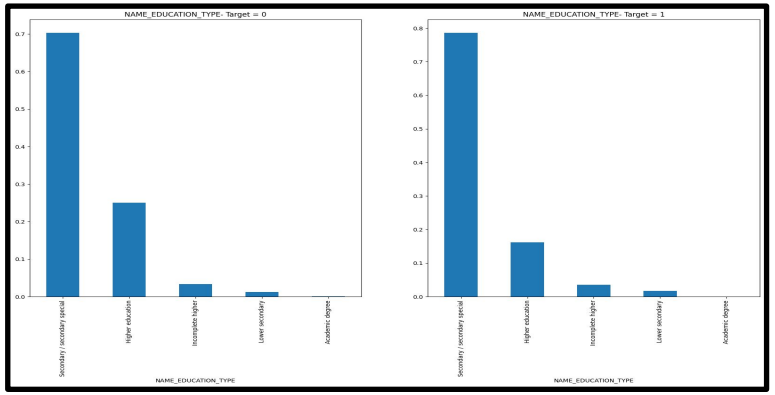


Fig 3

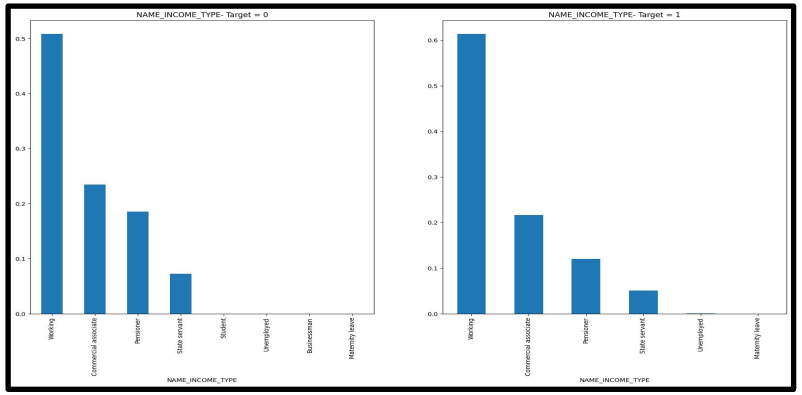
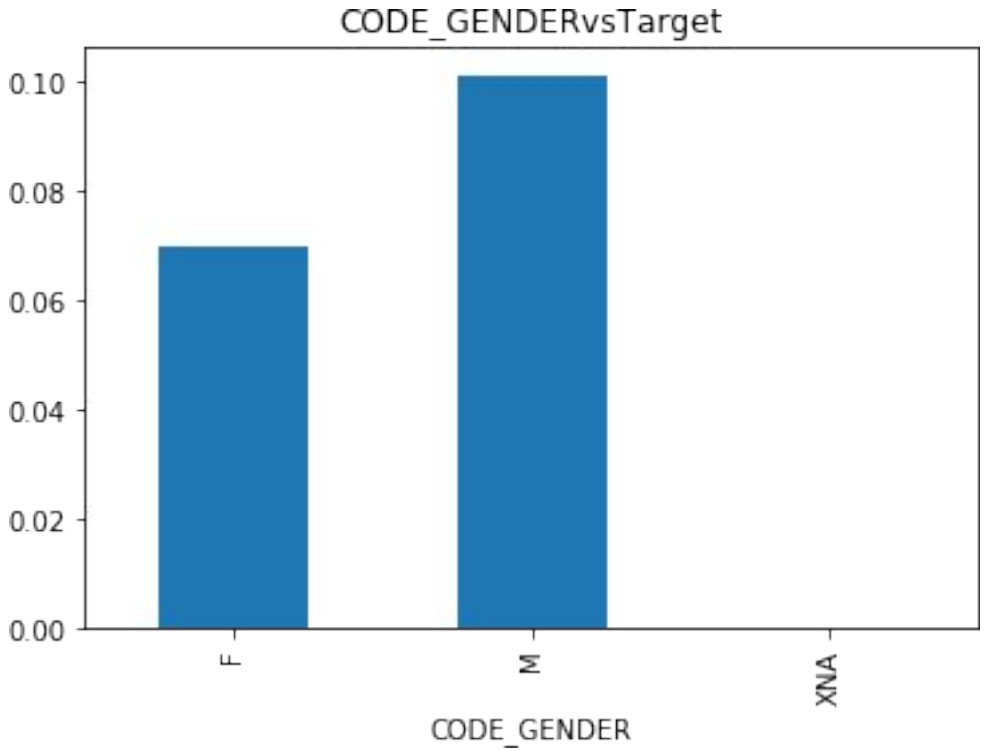


Fig 4

Exploring Data Sets – Categorical Variable

Bivariate Exploration
We explore the dataset to understand the relationship between loan default and gender.

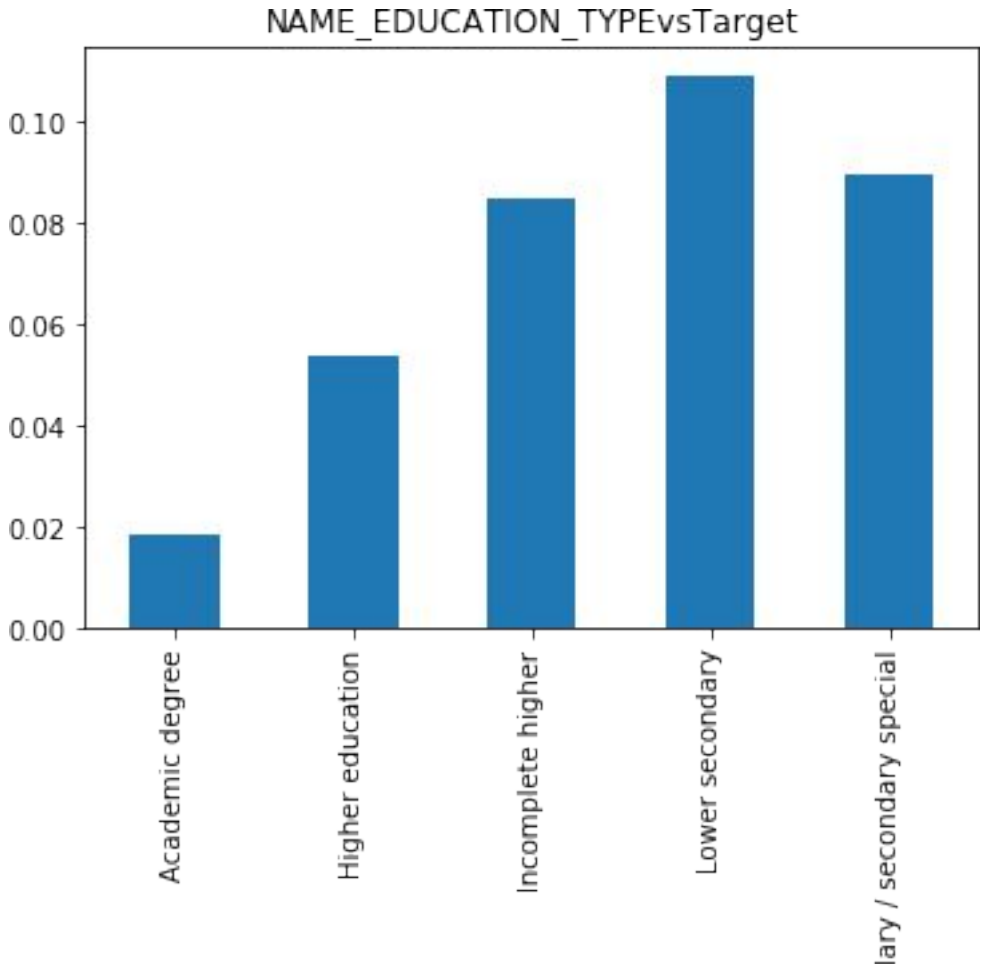
Observation:
- Males have a higher probability of defaulting (~10%) compared to females (~7%).



Exploring Data Sets – Categorical Variable

Bivariate Exploration
We explore the dataset to understand the relationship between loan default and education type.

Observation:
- Individuals with lower secondary education have a higher probability of defaulting (~11%) compared to any other education type.



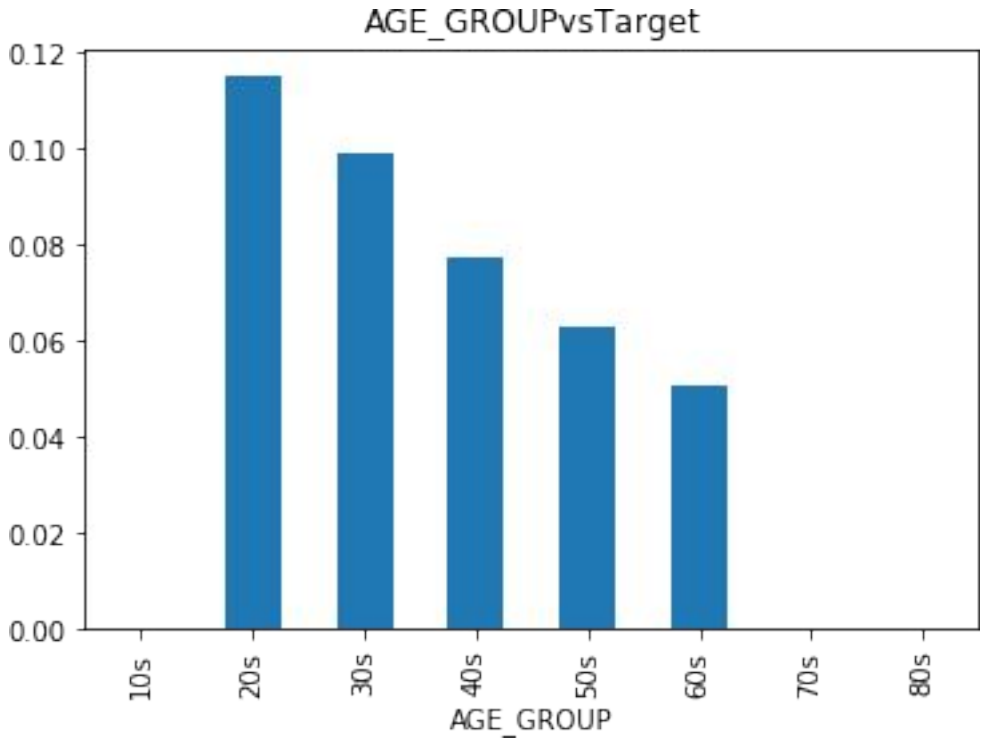
Exploring Data Sets – Categorical Variable

Bivariate Exploration

We explore the dataset to understand the relationship between loan default and age group.

Observation:

- Customers in their 20s and 30s have a higher risk of defaulting (>10%).



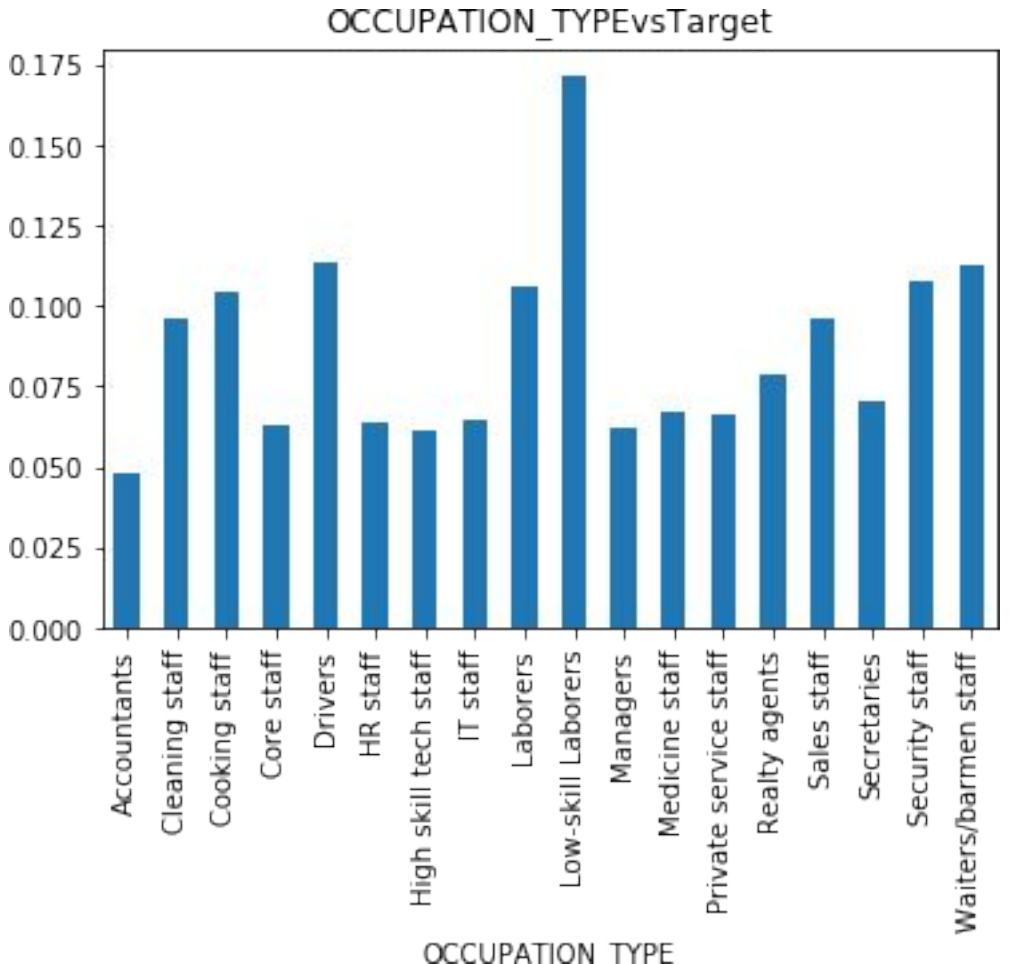
Exploring Data Sets – Categorical Variable

Bivariate Exploration

We examine the dataset to understand the relationship between loan default and occupation type.

Observation:

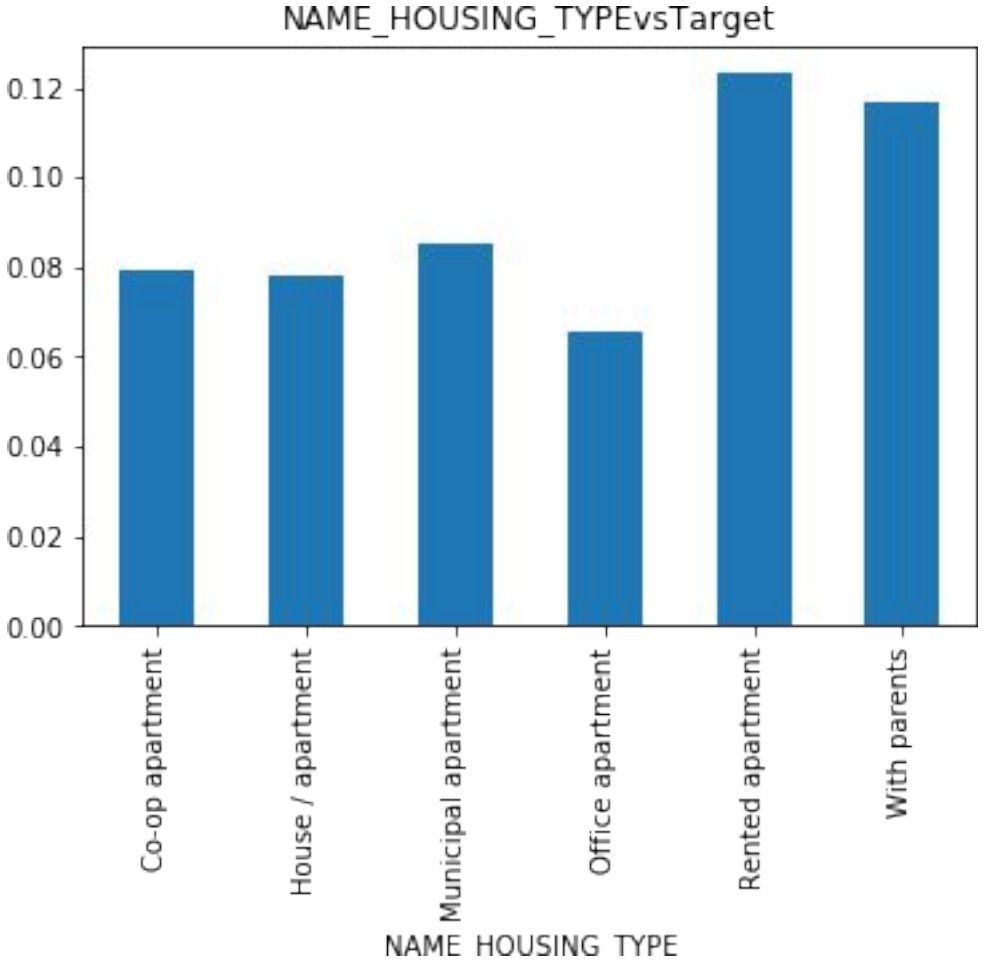
- Customers employed as low-skill laborers exhibit a higher risk of defaulting (~17%).



Exploring Data Sets – Categorical Variable

Bivariate Exploration
We investigate the dataset to understand the relationship between loan default and housing type.

Observation:
- Customers living with parents or in rented apartments show a higher risk of defaulting (>10%).



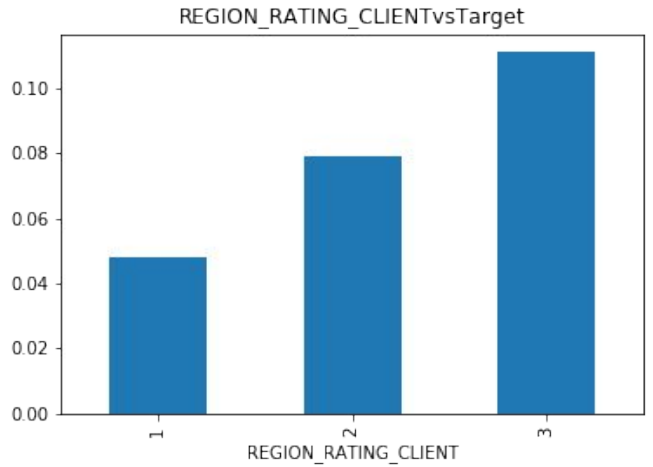
Exploring Data Sets – Categorical Variable

Bivariate Exploration

We analyze the dataset to understand the relationship between loan default and region rating.

Observation:

- Customers residing in region rating 3 exhibit a higher risk of defaulting (~10%).



Exploring Data Sets – Numerical Variable

Univariate Exploration

We analyze the dataset to understand how defaulters (target = 1) and non-defaulters (target = 2) are distributed across various numerical variables.

Observations:

- **Fig 1: DAYS_LAST_PHONE_CHANGE**
  - Defaulters (Target = 1) have a lower median and 75th percentile value compared to non-defaulters (Target = 2). This suggests that defaulters tend to change their phone numbers more frequently before loan application.

- **Fig 2: DAYS_ID_PUBLISH**
  - Defaulters (Target = 1) appear to change their IDs more frequently than non-defaulters (Target = 2).

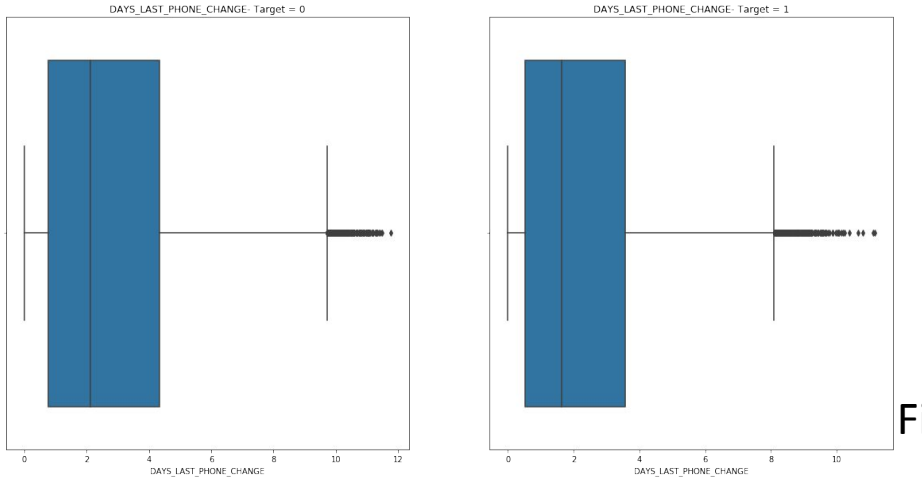


Fig-1

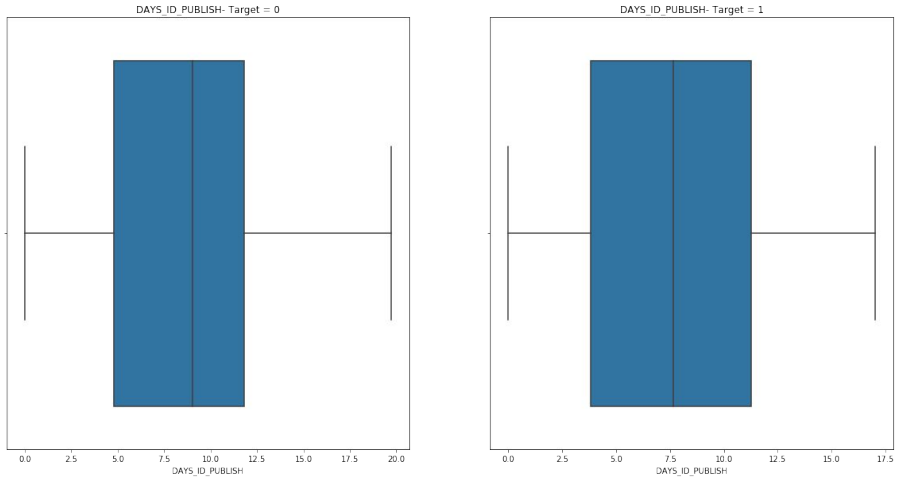


Fig-2

Exploring Data Sets – Numerical Variable

Bivariate Exploration

We explore the dataset to understand the relationship between defaulters (target = 1) and non-defaulters (target = 2) across various numerical variables.

Observations:

- **Fig 1: DAYS_LAST_PHONE_CHANGE**
  - Defaulters tend to change their phone numbers closer to the submission of the loan application.

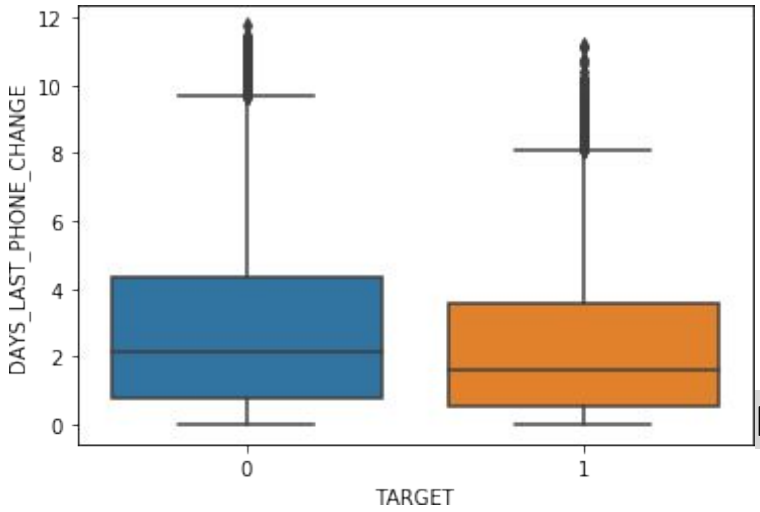


Fig-1

- **Fig 2: DAYS_ID_PUBLISH**
  - Defaulters tend to change their IDs closer to the submission of the loan application.

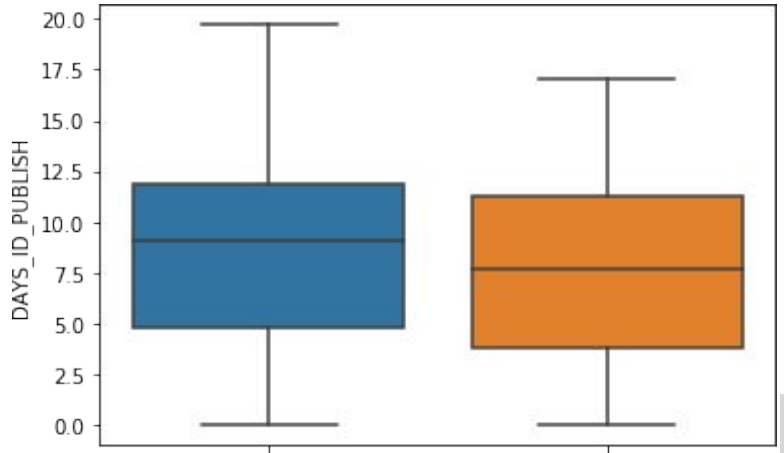


Fig-2

Exploring Data Sets – Numerical Variable

Correlation Analysis

We examine datasets for target = 0 and target = 1 to compare the top 10 correlated pairs of variables.

Observation:

- **Fig 1: Correlation for variables in target = 0**
- **Fig 2: Correlation for variables in target = 1**

While correlation values appear similar between both datasets, a tabular format will be utilized to facilitate direct comparison.

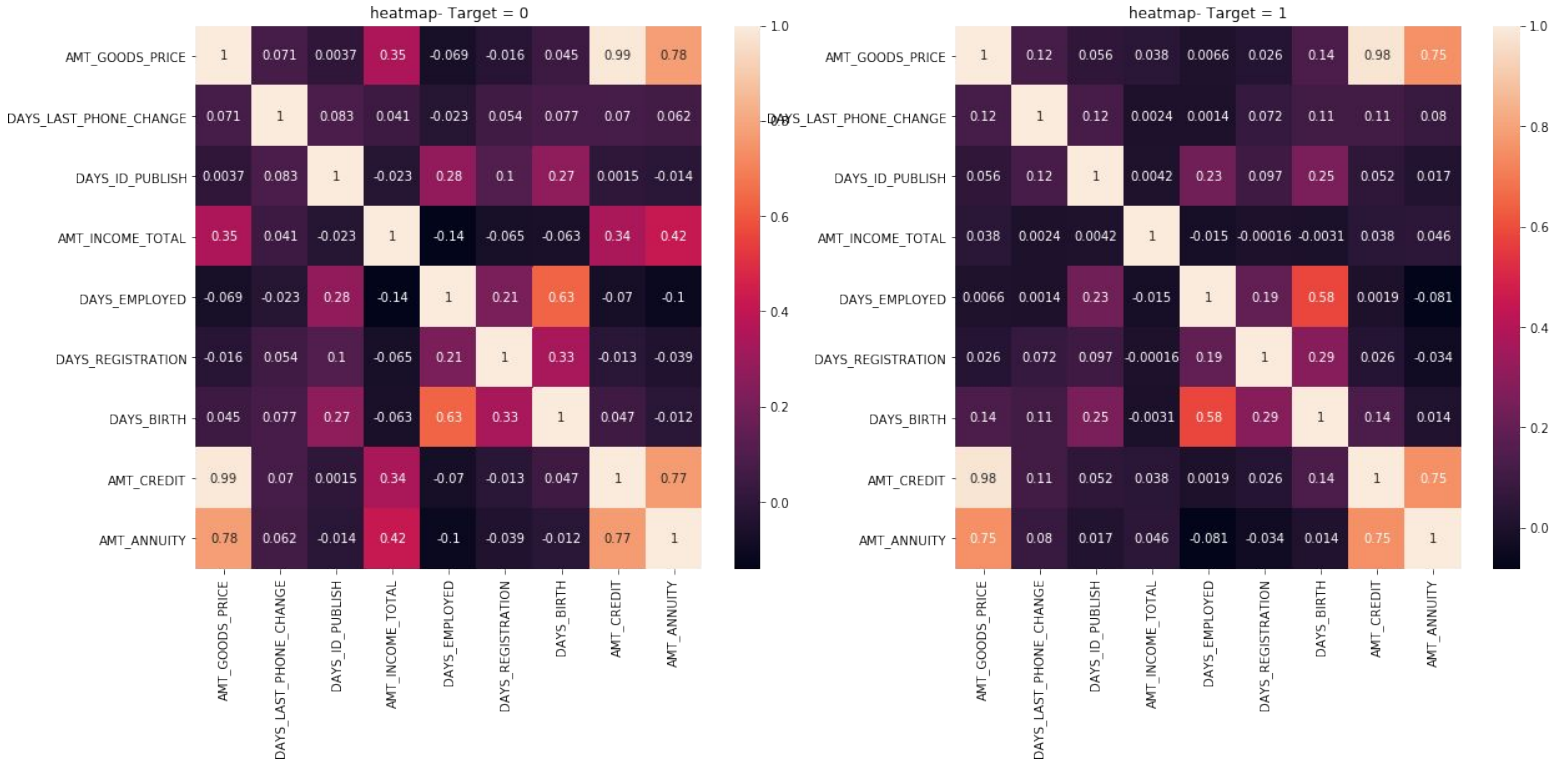


Fig-1

Fig-2

Exploring Data Sets – Numerical Variable

Correlation Analysis
We analyze datasets for target = 0 and target = 1 to compare the top 10 correlated pairs of variables.

Observation:
- **Fig 1: Top 10 correlated variables for target = 0**
- **Fig 2: Top 10 correlated variables for target = 1**

It is noteworthy that 8 out of the top 10 correlated variables are common across both datasets.

Fig-1

	VAR1	VAR2	Correlation	Correlation_abs
414	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	1.00
154	AMT_GOODS_PRICE	AMT_CREDIT	0.98	0.98
337	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96	0.96
277	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89	0.89
440	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87	0.87
129	AMT_ANNUITY	AMT_CREDIT	0.75	0.75
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.75	0.75
207	DAYS_EMPLOYED	DAYS_BIRTH	0.58	0.58
415	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34	0.34
389	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33	0.33

Fig-2

	VAR1	VAR2	Correlation	Correlation_abs
414	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	1.00
154	AMT_GOODS_PRICE	AMT_CREDIT	0.98	0.98
337	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96	0.96
277	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89	0.89
440	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87	0.87
129	AMT_ANNUITY	AMT_CREDIT	0.75	0.75
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.75	0.75
207	DAYS_EMPLOYED	DAYS_BIRTH	0.58	0.58
415	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34	0.34
389	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33	0.33

Summary

As per analysis made below is the summary, which can help in predicting which customer can account to default.

S.No.	Variable / Column	Variable Type
1	CODE_GENDER	Categorical
2.	NAME_EDUCATION_TYPE	Categorical
3.	AGE_GROUP	Categorical
4.	NAME_HOUSING_TYPE	Categorical
5.	NAME_INCOME_TYPE	Categorical
6.	OCCUPATION_TYPE	Categorical
7.	REGION_RATING_CLIENT	Categorical
8.	REGION_RATING_CLIENT_W_CITY	Categorical
9.	DAYS_LAST_PHONE_CHANGE	Numerical
10.	DAYS_ID_PUBLISH	Numerical