CREDIT EDA CASE STUDY USING PYTHON:

(By ~ VIVEK RAWAT)

PROBLEM STATEMENT:

1. **CREDIT EDA OBJECTIVES**:

★ Challenges Faced: Lending companies encounter difficulties when dealing with individuals lacking a robust credit history, potentially leading to intentional defaults.

★ Company Focus: Working in a consumer finance firm specializing in urban lending, the aim is to scrutinize loan applications through data analysis techniques.

★ Objective Clarification: Beyond numerical data, the primary goal is to ensure deserving applicants aren't rejected solely due to credit limitations while comprehending associated risks.

2. **CHALLENGES IN LOAN APPROVALS**:

★ Critical Decisions: Approving or rejecting loans involves tough choices; declining a capable applicant risks losing business, while approving a risky one may lead to financial setbacks.

★ Client Categorization: The available dataset categorizes clients into distinct groups: those facing payment difficulties and those consistently meeting payment schedules.

3. **DATA ANALYSIS FOR RISK ASSESSMENT**:

★ The data analysis aims to comprehend how client and loan details impact default probabilities, enabling fair lending decisions beyond credit history and supporting intelligent risk management in loan approvals.

ANALYSIS STRUCTURE:

1. **UNDERSTANDING AND CLEANING THE DATA:**
   First, we explore the dataset to grasp its nature and clean it for analysis, identifying crucial columns.

2. **PERFORMING DETAILED ANALYSIS:**
   We conduct univariate and bivariate analyses on specific categorical and numerical variables.

3. **IDENTIFYING PREDICTIVE VARIABLES:**
   Our focus lies in finding variables that aid in predicting high-risk customers.

CERTAIN VITAL VARIABLES FROM OUR DOMAIN STAND OUT:

★ **TARGET:** It highlights customers who've experienced payment delays, serving as a primary variable for our analysis.

★ **NAME_CONTRACT_STATUS:** This variable denotes the status of prior loan applications, holding significance in our analysis.

An important assumption guiding our analysis is that all customers with at least one missed payment are considered equally, irrespective of the number of missed payments.

EXPLORING THE DATA:

1. **DATA PREPARATION:**
   This marks the initial stage of our analysis, concentrating on readying the dataset for predictive analysis purposes.

★ **UNDERSTANDING DATA:**

- Data dimensions
- Column type
- Identify columns with missing values
- Identify outliers

2. **DATA CLEANING:**
   At this stage, we emphasize developing approaches to:

- Drop columns with data missing more than 50%
- Handle null values
- Treat outliers
- Fix incorrect datatype of columns.

3. **DATA TRANSFORMATION:**
   At this stage, we focus on:

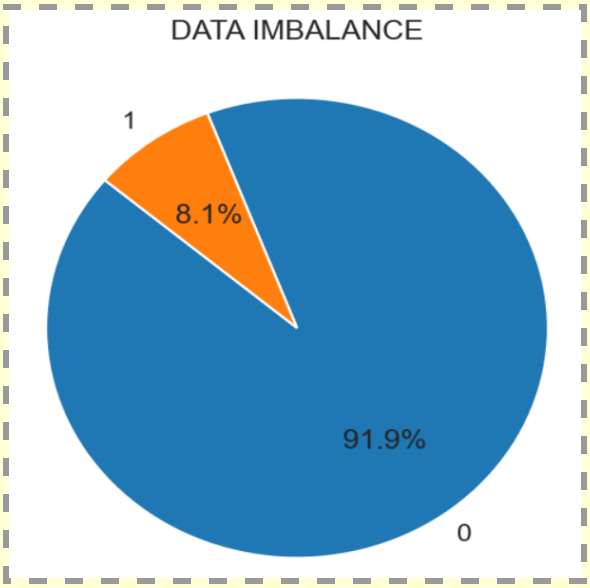
- Creation of necessary columns through binning
- Removal irrelevant columns
- Creation of segmented data frames based on the TARGET variable

DATA IMBALANCE:

DATA IS HIGHLY IMBALANCED:

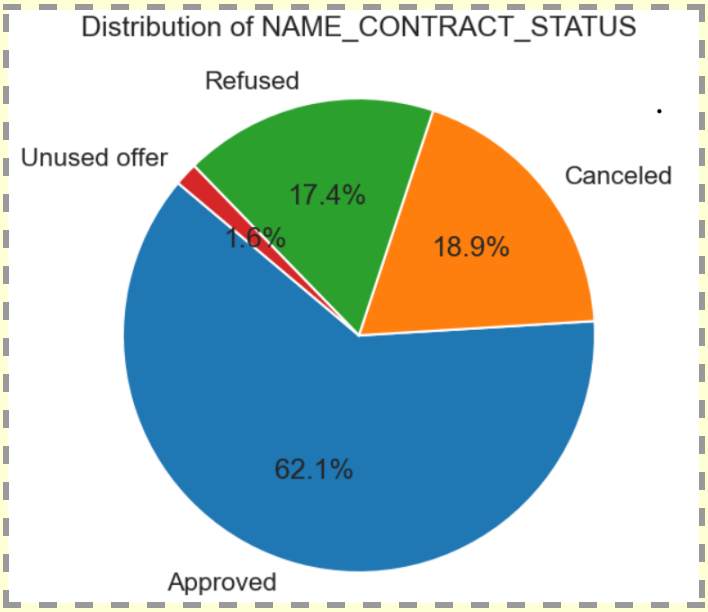
1.In application_data.csv

- Only 8.1% of data is present for defaulters.
- Rest 91.9% of the data is present for non-defaulters.



2.In previous_application.csv

- 62.1% data is available for customers with loan status: Approved.



ANALYSING CATEGORICAL COLUMNS:

I have analyzed the dataset to understand how individuals who face difficulties with payments (TARGET = ‘1’) compare to those who don't (TARGET = ‘0') across different aspects.
Here's a breakdown of what I observed:

- Fig 1: Gender - There's a higher ratio of males among those who struggle with payments.
- Fig 2: Age Group - Interestingly, there's a notable cluster of individuals in their 30s who seem to have trouble keeping up with payments.
- Fig 3: Education - Among those facing payment challenges, there's a larger percentage of individuals with secondary education backgrounds.
- Fig 4: Income Type - found that a greater proportion of those facing payment issues are in occupations characterized by earning a living through work.

Fig 1:

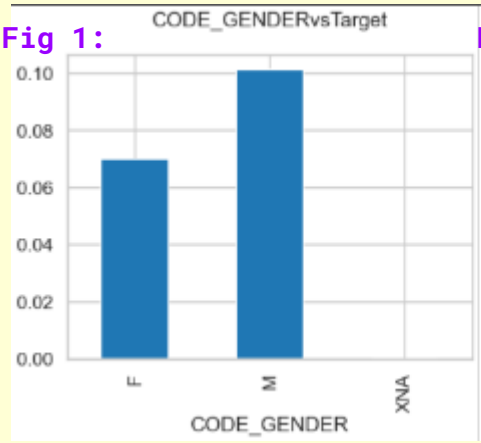


Fig 2:

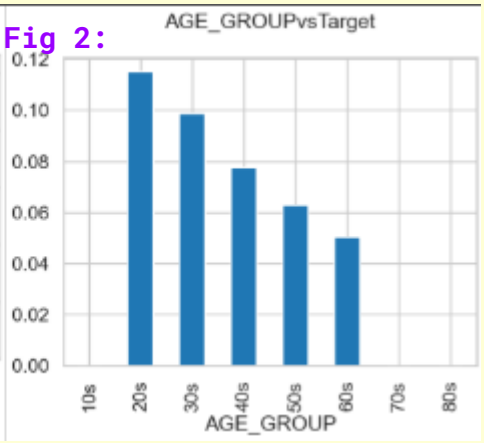


Fig 3:

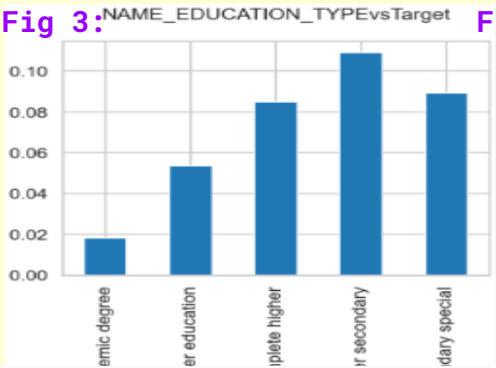
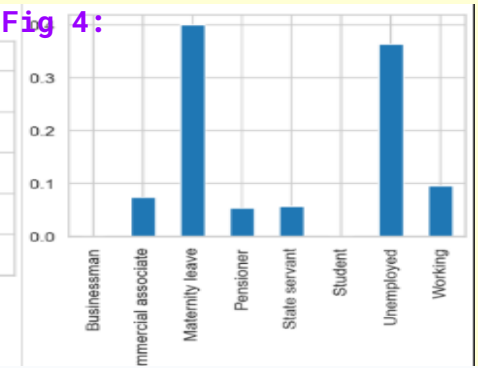
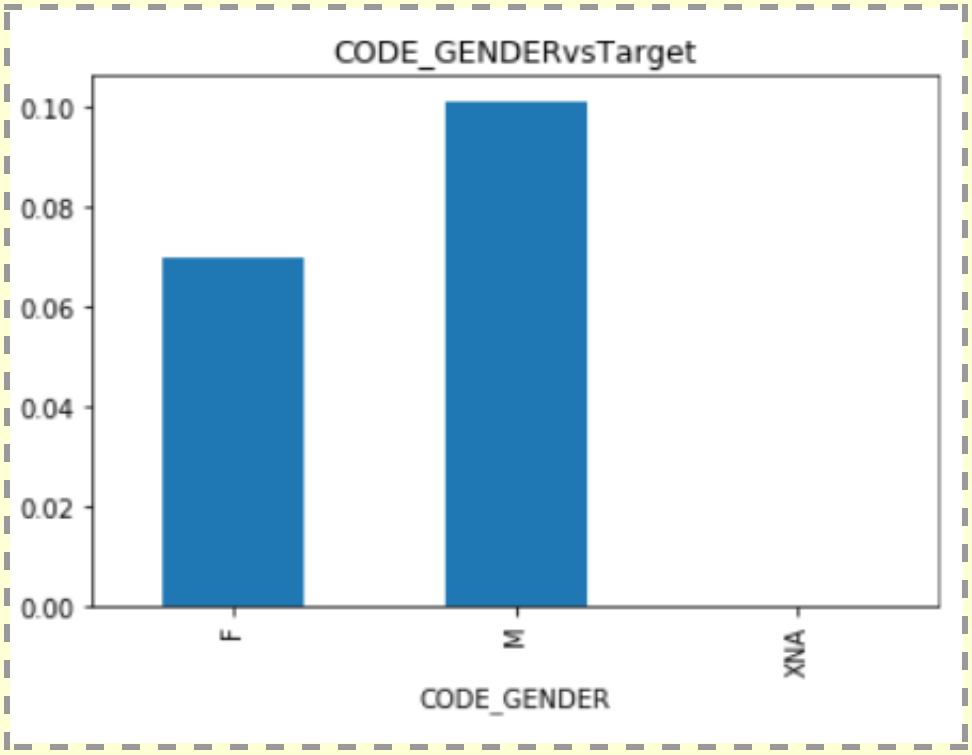


Fig 4:

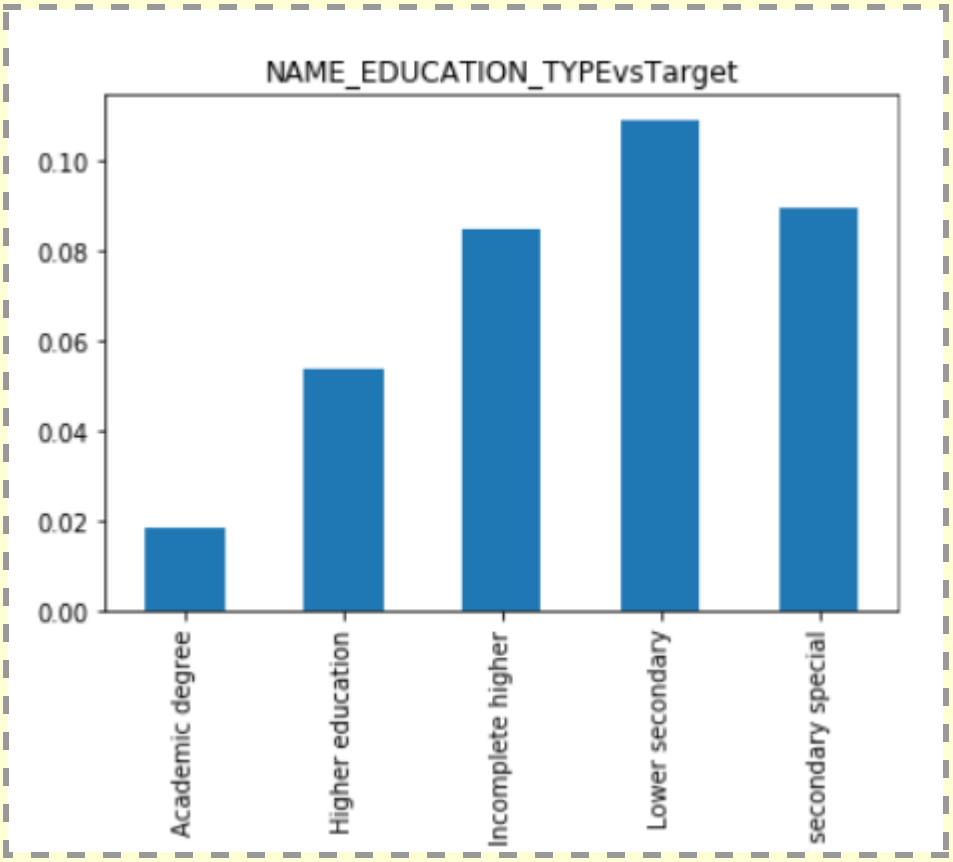


BIVARIATE CATEGORICAL ANALYSIS:

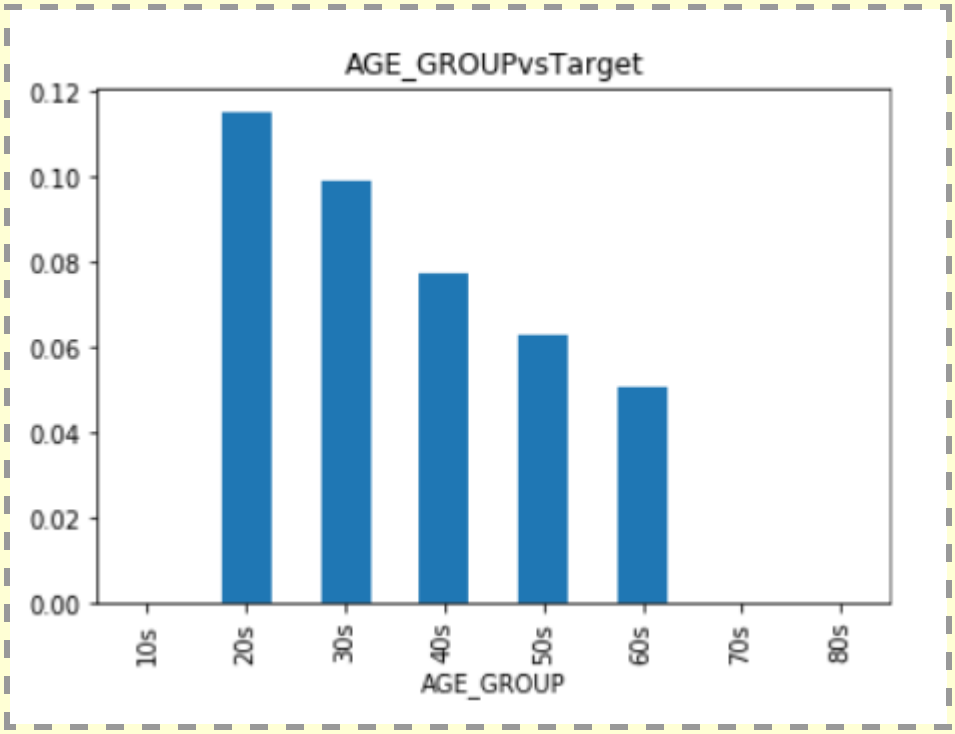
- During the Analysis of the dataset, I explored the correlation between **loan defaulters** and **gender**. The results underscored that **males** tend to have a higher likelihood of defaulting, hovering around **10%**, whereas **females** exhibit a lower **default rate**, approximately **7%**.



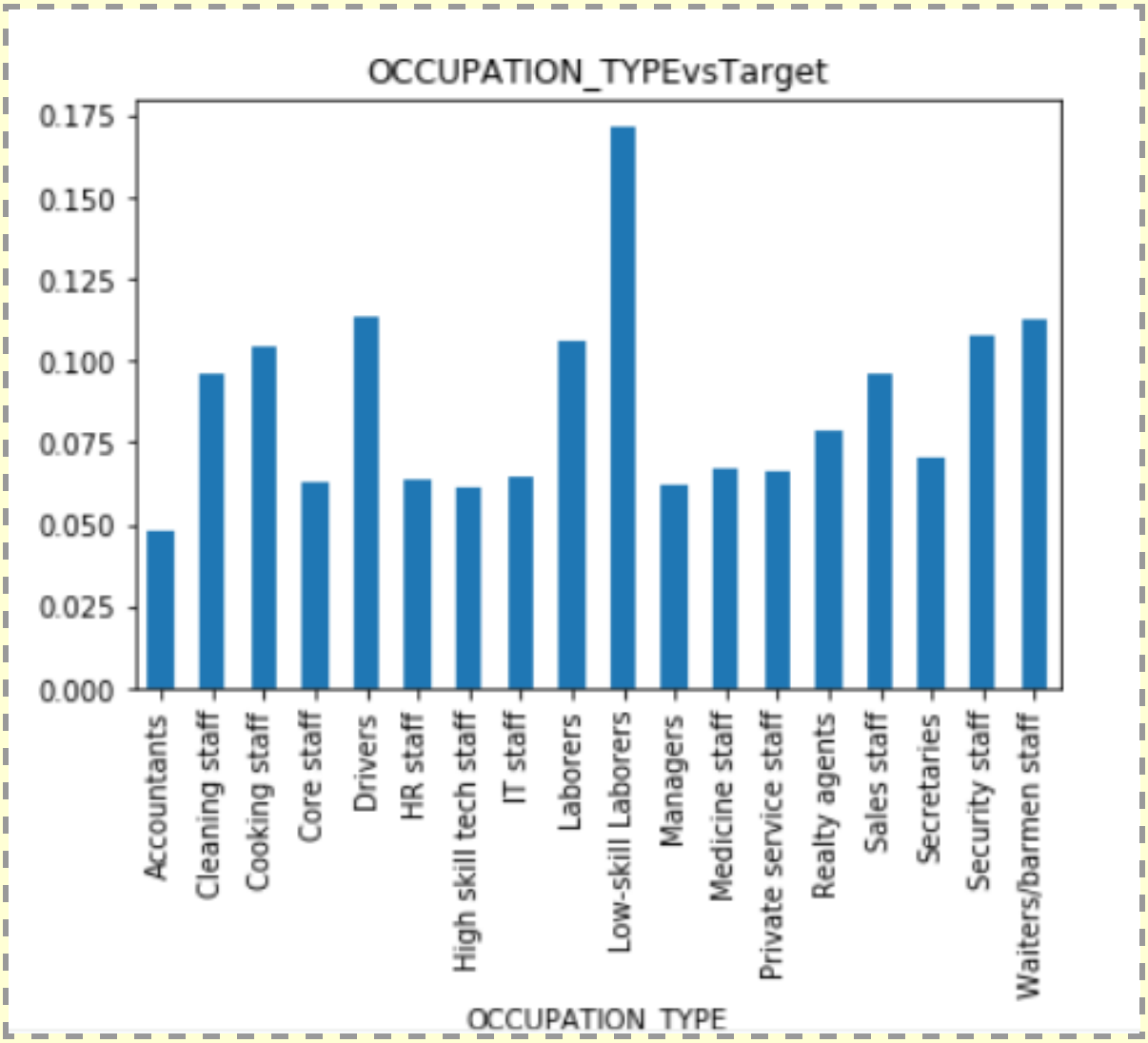
- During the analysis of the dataset, I explored the correlation between **loan defaulters** and **education type**. My findings revealed that individuals with a **lower secondary education** level show a higher probability of defaulting, approximately **11%**, compared to any other **education type**.



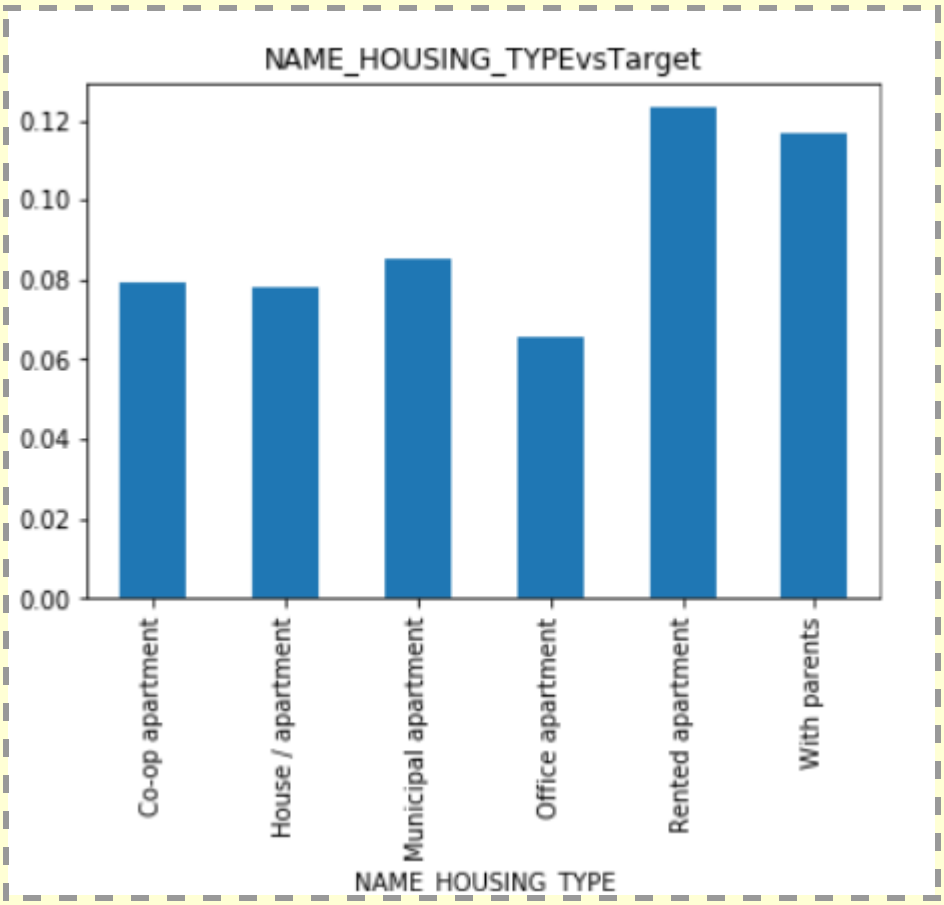
- During the analysis of the dataset, I explored the correlation between **loan defaulters** and the **Age group**. My findings revealed that customers in their **20s and 30s** exhibit a higher risk of defaulting, with rates exceeding **10%**.



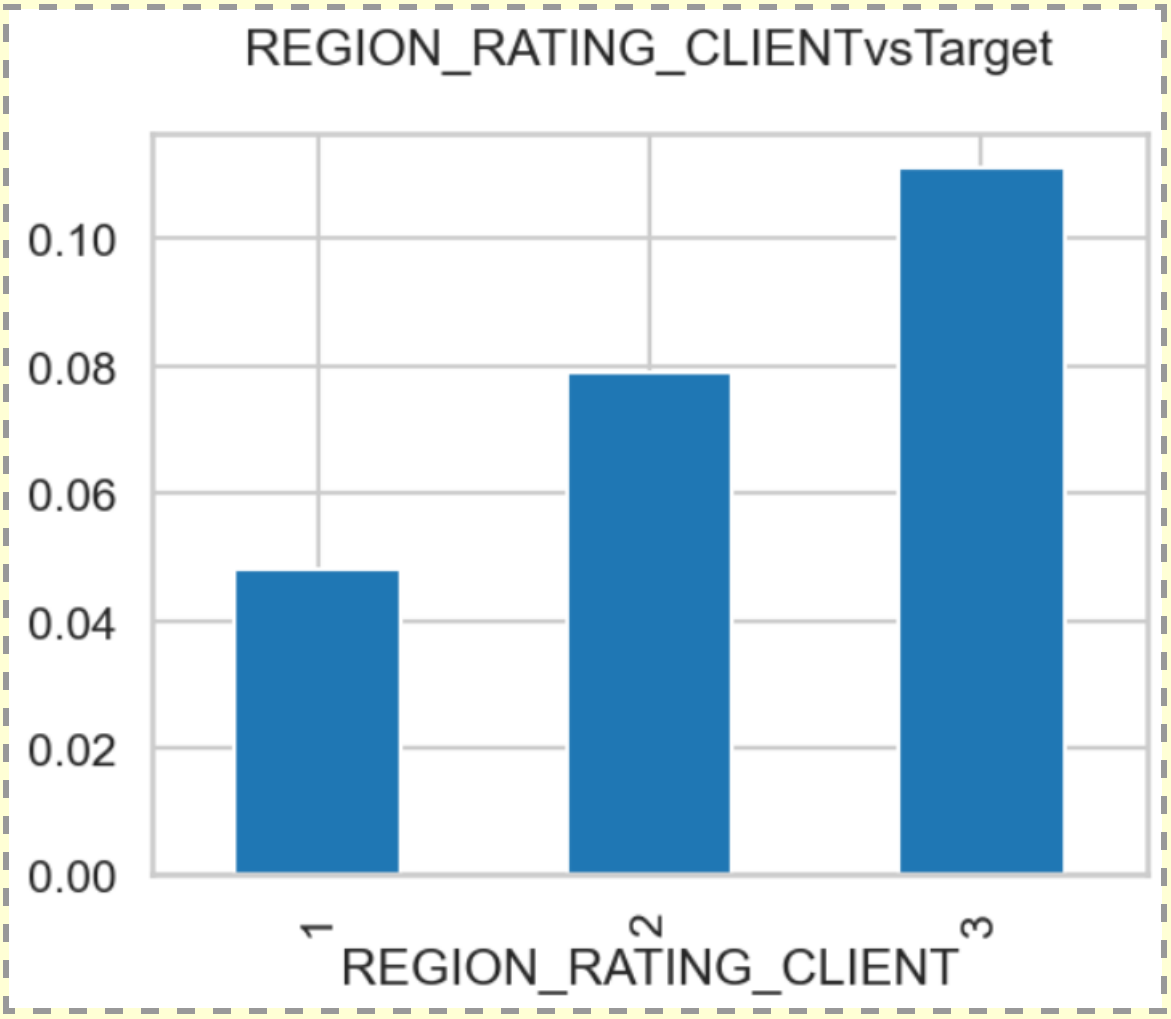
- During the analysis of the dataset, I explored the correlation between **loan defaulters** and **Occupation Type**. My findings highlighted that customers employed as low-skill laborers face a higher risk of defaulting, with a rate of approximately 17%.



- During analysis of the dataset, I explored the correlation between **loan defaulters** and **Housing type**. My findings highlighted that customers living with their parents have a higher risk of defaulting i.e. greater than 10%.



- During the analysis of the dataset, I explored the correlation between **loan defaulters** and **Region Rating**. My findings highlighted that customers living in Region Rating = 3, exhibit a higher risk of loan defaulting i.e. greater than 10%.



ANALYSING NUMERICAL COLUMNS:

I have analyzed the dataset to understand how individuals who face difficulties with payments (TARGET = ‘1’) compare to those who don't (TARGET = ‘0') across different aspects.
Here's a breakdown of what I observed:

- Fig 1: DAYS_LAST_PHONE_CHANGE - It appears that 75% and the median value for defaulters is less than non-defaulters, this shows that defaulters more often change their phone numbers before submitting the loan application.

- Fig 2: DAYS_ID_PUBLISHED - It appears that the median value for defaulters is less than non-defaulters, this shows that defaulters more often change their legal IDs.

FIG 1:

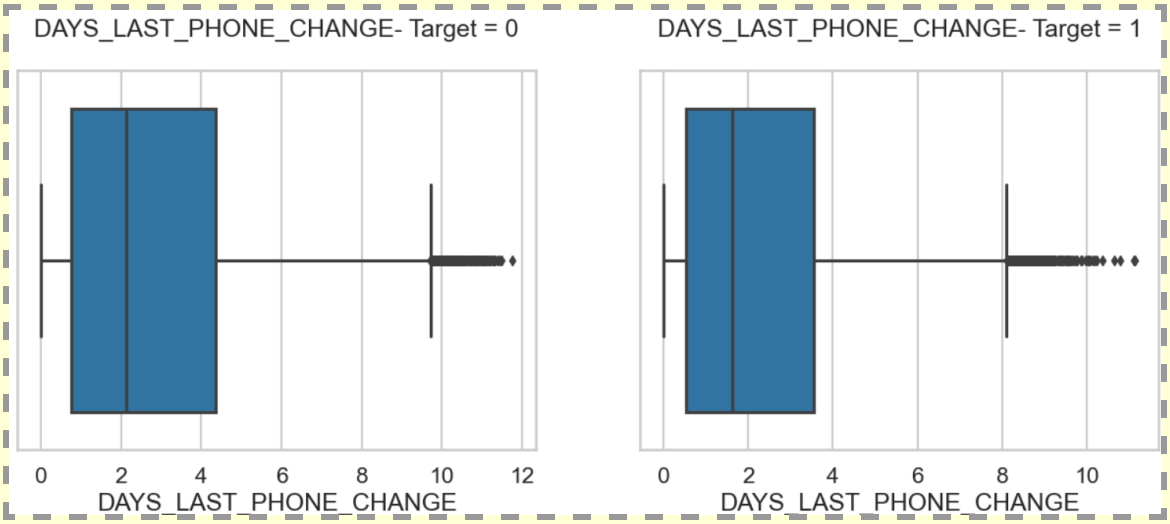
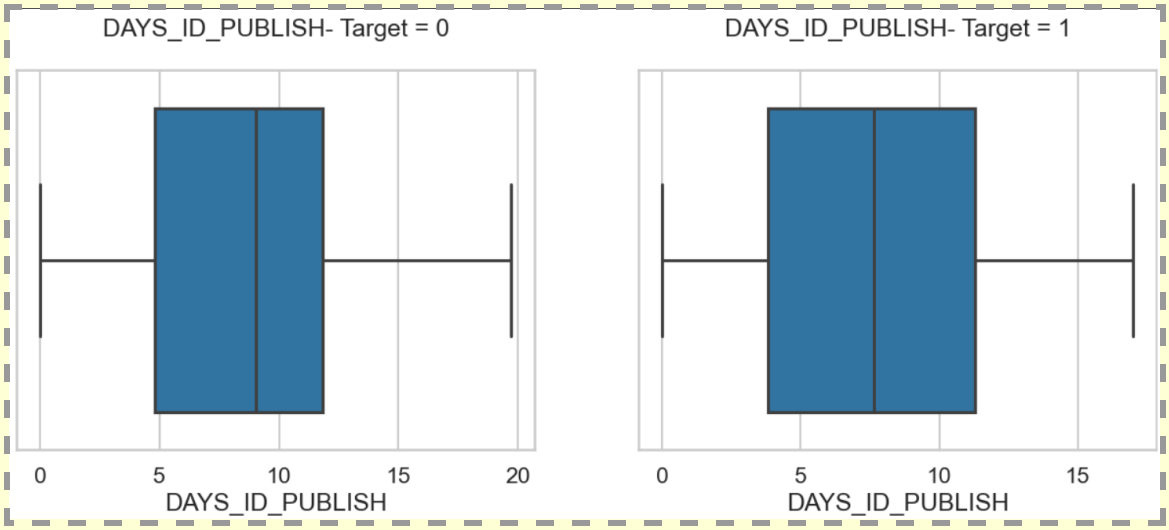


FIG 2:



CORRELATION MATRIX:

I examined the segmented data frames to verify whether the top 10 pairs of correlated variables are consistent across both datasets.

- Fig 1: correlation for variables in target=0
- Fig 2: correlation for variables in target=1

FIG: 1

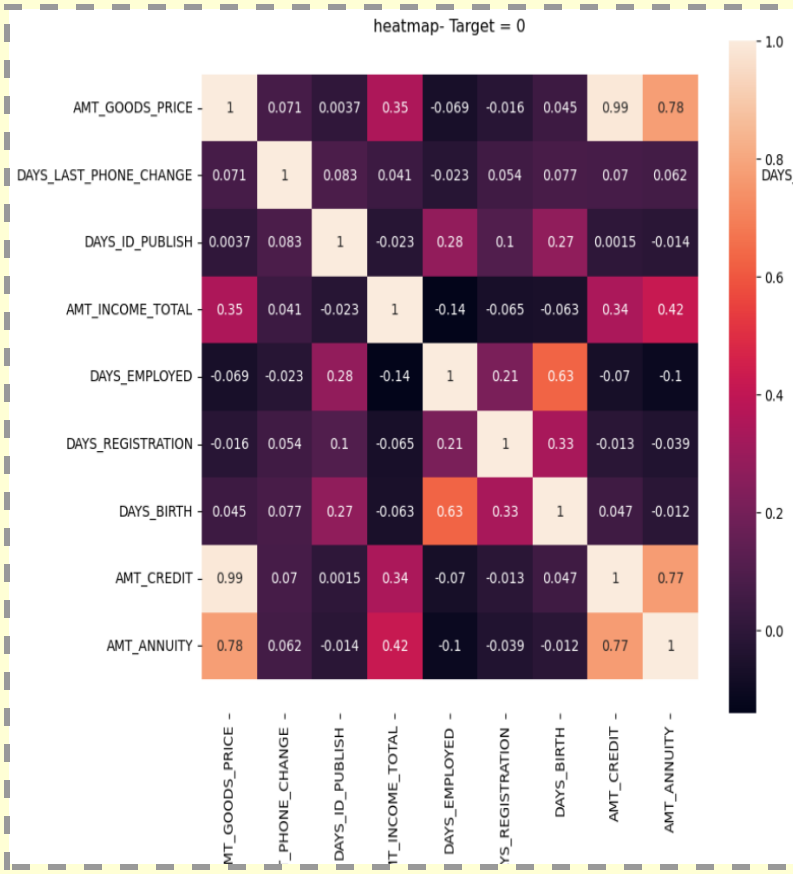
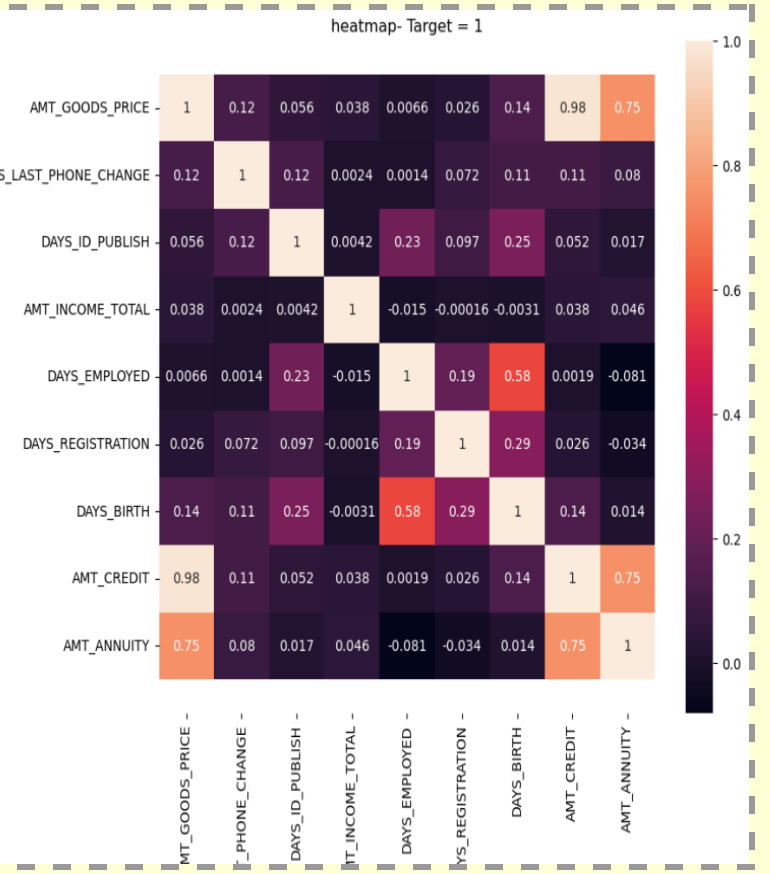


FIG: 2



Then I analyzed both correlation matrix variables to see if they are common across both the data i.e. TARGET=0 & TARGET=1.

FOR TARGET=0

	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	0.99	0.99
337	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95	0.95
277	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88	0.88
440	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86	0.86
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.78	0.78
129	AMT_ANNUITY	AMT_CREDIT	0.77	0.77
207	DAYS_EMPLOYED	DAYS_BIRTH	0.63	0.63
128	AMT_ANNUITY	AMT_INCOME_TOTAL	0.42	0.42
153	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35	0.35

FOR TARGET = 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

Common highly correlated variables pair count across both the datasets is **8**

SUMMARY OF CREDIT EDA ANALYSIS:

In the CREDIT EDA, I analyzed the data through univariate and bivariate analysis and found that below are the top 10 columns that help in predicting the fact that a customer will default in the future or not.

Listing down those column names:

1. Categorical Columns:
   - ★ NAME_EDUCATION_TYPE
   - ★ AGE_GROUP
   - ★ OCCUPATION_TYPE
   - ★ CODE_GENDER
   - ★ NAME_INCOME_TYPE
   - ★ NAME_HOUSING_TYPE
   - ★ REGION_RATING_CLIENT
   - ★ REGION_RATING_CLIENT_W_CITY

2. Continuous/Numerical columns:
   - ★ DAYS_LAST_PHONE_CHANGED
   - ★ DAYS_ID_PUBLISHED