CREDIT EDA CASE STUDY

SREEDEVI GATTU & DISHANT PATEL

AUGUST 2021

CONTENTS

- PROBLEM STATEMENT
- DATA
- STRATEGY/METHODOLOGY FOR ANALYSIS
- DATA – IMBALANCE, HANDLING FOR MISSING/INVALID VALUES, OUTLIERS
- ANALYSIS
  - Demographic – Personal
  - Demographic - Professional
  - Character
  - Capacity
  - Capital
  - Application
- CORRELATION OF VARIABLES
- RECOMMENDATIONS

PROBLEM STATEMENT

What is the THE PROBLEM we trying to address?

A consumer finance company specialises in lending various types of loans to urban customers. As there is insufficient or no credit history, the company does not know whether a customer is reliable or not. Some consumers use it to their advantage by becoming a defaulter.

When the company receives a loan application, it must decide for loan approval **based on the applicant's profile**.

Two types of risks are associated with the bank's decision. If the applicant is

- **likely to repay** the loan, then not approving the loan results in a **loss of business** to the company
- **not likely to repay** the loan, then approving the loan results in a **financial loss** to the company

What is REQUIRED as part of EDA?

Identify patterns present in the data which indicate if a client has difficulty paying their instalments. Identify the **driving factors/variables behind loan default** which are strong indicators of default.

How does this analysis HELP the company?

The company can utilise this knowledge for its portfolio and risk assessment. This will help the company

- To take actions (such as denying the loan, reducing the amount of loan etc.) against risky applicants
- Not to reject applicants capable of repaying the loan

DATA

Two datasets were provided

- **Current Application** dataset contains information about the current loan application of the clients. For each client, there is
  - Personal, professional, assets owned, loan application, load & goods amount information: **121 columns**
  - **TARGET** variable to indicate whether the clients had
    - Payment difficulties TARGET=1
    - No payment difficulties TARGET=0
- **Previous Application** dataset contains information about the previous loan applications of clients.
  - There could be multiple previous applications/entries for a client
  - Each application has information of the status - Approved, Rejected, Cancelled & Unused offer

In both the datasets, unique identifier **SK_ID_CURR** identifies a client.

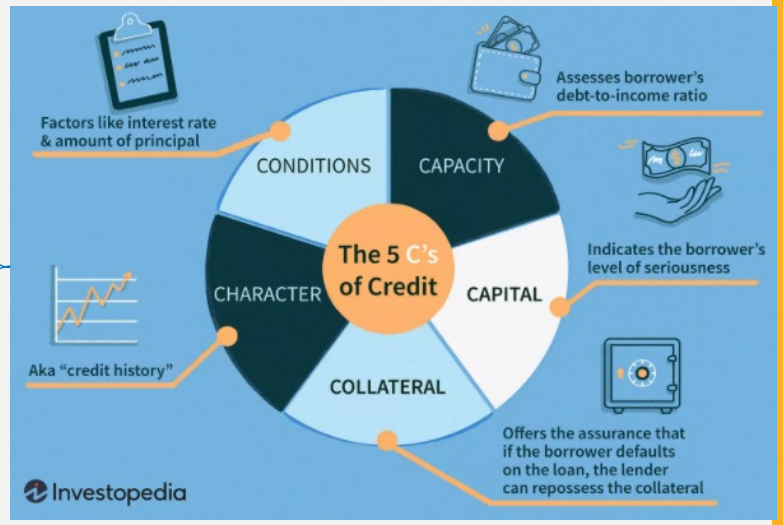
Since the number of columns is very high, the columns will be grouped into multiple categories (see next slide)

ANALYSIS STRATEGY

1. The variables/columns are categorised into following categories

Category	Variables from the dataset
Demographic - Personal	Personal details like Gender, Age, Number of Children etc.
Demographic - Professional	Details related to profession like education, Income, Organization etc.
Character	Previous Credit history, Number of calls to credit bureau
Capacity	Debt to income ratio, Income, Goods, Credit amounts,
Capital	Whether the applicant owns car, realty
Application	Information like contact details, documents provided

2. For each of the categories, following is carried out
   - **Inspect** dataset for data imbalance, missing & invalid values
   - **Analysis** - Univariate, Bivariate
   - **Draw** Observations/Insights
3. Correlation analysis is done on all the variables



Character, Capacity, Capital categories are inspired from the Five C's of Credit system that is used by lenders to gauge the creditworthiness of potential borrowers

DATA - IMBALANCE

In the current application dataset, data corresponding to the clients with TARGET=1 is much lesser (8%) than compared to the data with TARGET=0 (92%).

TARGET	Entries - Count	Entries - %
0	2,82,686	91.93%
1	24,825	8.07%
	307511	100%

There is a **data imbalance** with respect to the TARGET variable in the **ratio 92% (0) and 8% (1).** This means the data corresponding to the **clients with payment difficulties is very less** compared to the **clients without payment difficulties**.

Handling To avoid this imbalance, bias the analysis (towards the Target=0 case), the comparisons of various variables with respect to the Target categories (0 & 1) were made using **percentages**

Note: No duplicate rows were found in the dataset

DATA - MISSING/INVALID

Type of missing/invalid data	Handling	Instances
Columns with data is missing > 50%	Drop the columns	BUILDING related data, OCCUPATION_TYPE etc.
Columns not relevant for analysis	Drop the columns	EXT_SOURCE_*
Columns with similar data	Retain only one & drop the rest	REGION_RATING_CLIENT & REGION_RATING_CLIENT_W_CITY
Columns with similar data	Aggregate (add multiple columns) into a single column & drop the rest	FLAG_DOCUMENT_* FLAG_MOBIL, FLAG_*_PHONE, FLAG_EMAIL OBS_CNT_SOCIAL_CIRCLE DEF_CNT_SOCIAL_CIRCLE AMT_REQ_CREDIT_BUREAU_* → ENQUIRIES_CREDIT_BUREAU
Columns with related information	Derive from other columns	AMT_GOODS_PRICE = 90 % of AMT_CREDIT AMT_ANNUITY = 5% of AMT_CREDIT
Columns with missing data < 50%	Impute with median value	CNT_FAM_MEMBERS, AMT_REQ_CREDIT_BUREAU_* , OBS_CNT_SOCIAL_CIRCLE, DEF_CNT_SOCIAL_CIRCLE
Rows with missing/invalid data < 1%	Drop the rows	CODE_GENDER, NAME_FAMILY_STATUS

DATA - TRANSFORM

Wherever relevant, multiple columns were transformed into more meaningful formats

Type of data	Transformation	Examples
Days relative to the time of application (-ve values)	Converted to years (+ve value)	DAYS_BIRTH → AGE DAYS_REGISTRATION → YEARS_REGISTRATION DAYS_ID_PUBLISH → YEARS_ID_PUBLISH
Categories in numerical format	Converted to categorical values (string)	REGION_RATING_CLIENT*
New columns	Aggregated multiple columns to percentage	DEF_PC_SOCIAL_CIRCLE = DEF_CNT_SOCIAL_CIRCLE /OBS_CNT_SOCIAL_CIRCLE DEBT_TO_INCOME = AMT_ANNUITY/_INCOME_TOTAL

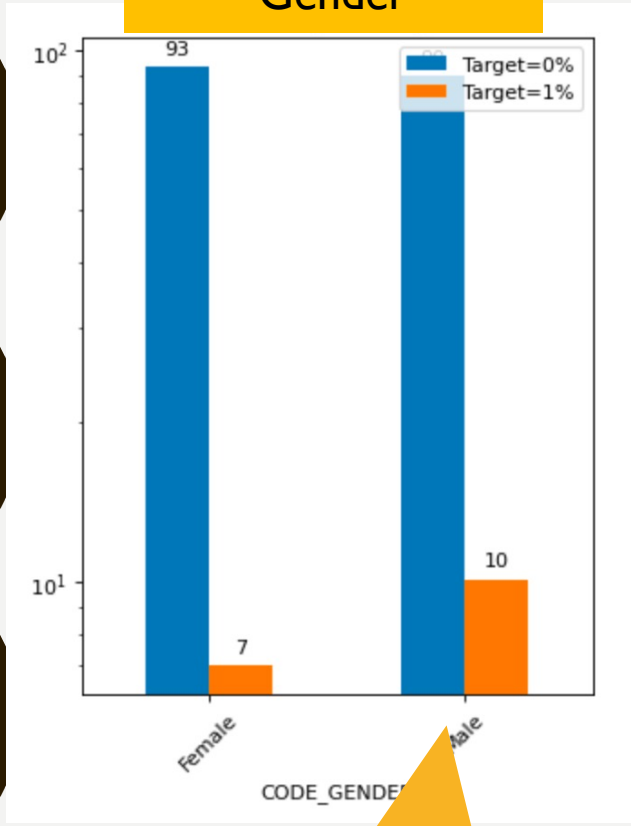
DATA - OUTLIERS

The following outliers were identified by using boxplots

Variable	Outliers
CNT_CHILDREN	> 2.5 till 19
CNT_FAM_MEMBERS	> 5 till 20
REGION_POPULATION_RELATIVE	1 area is heavily populated compared to the rest
AMT_REQ_CREDIT_BUREAU_*	Extremely high value of 262
AMT_INCOME_TOTAL	Extremely high value of 117000000.0
YEARS_REGISTRATION	> 40 years
DEBT_TO_INCOME	> 50
AMT_CREDIT	> 1600000
AMT_ANNUITY	> 80000
AMT_GOODS_PRICE	> 1500000

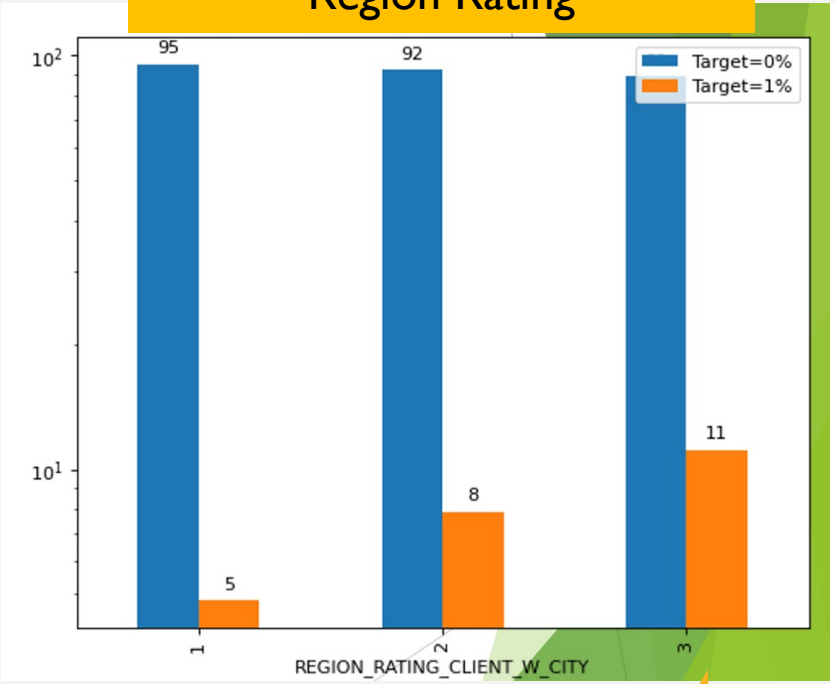
DEMOGRAPHIC *PERSONAL*

Gender



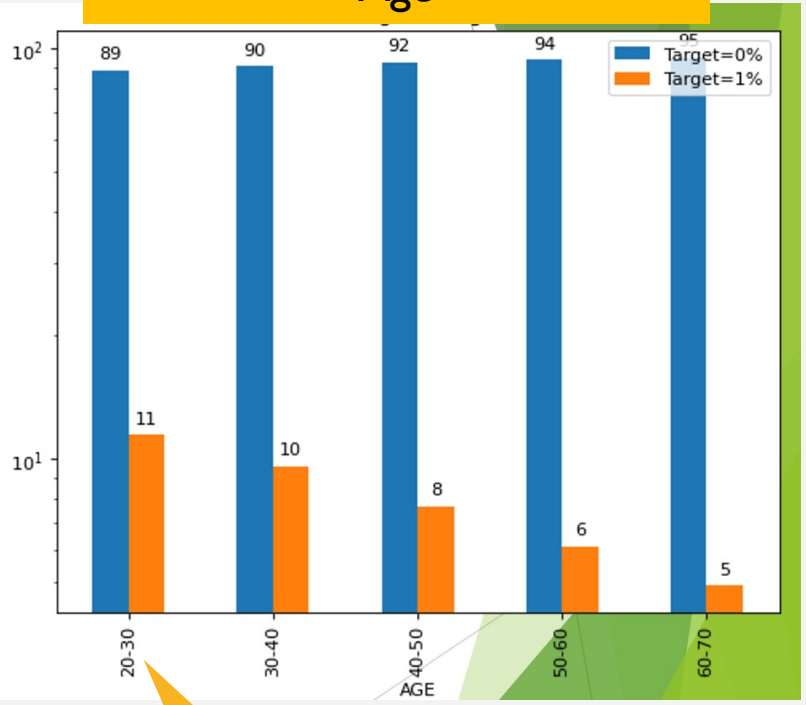
Male clients have difficulties in paying

Region Rating



Clients in region 3 have difficulties in paying

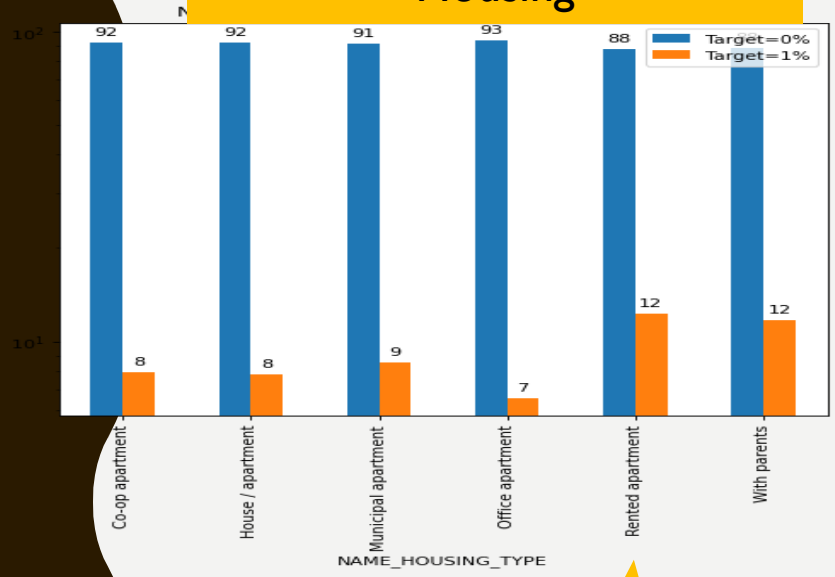
Age



Younger clients have difficulties in paying

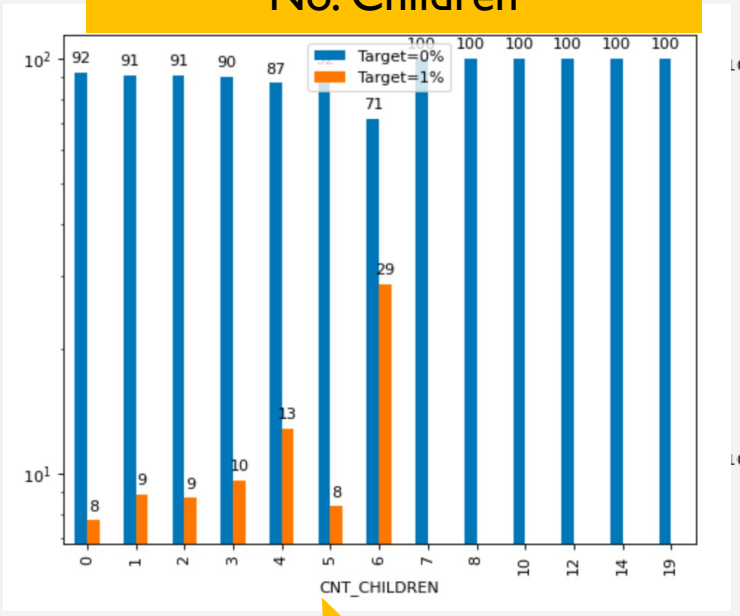
DEMOGRAPHIC *PERSONAL*

Housing



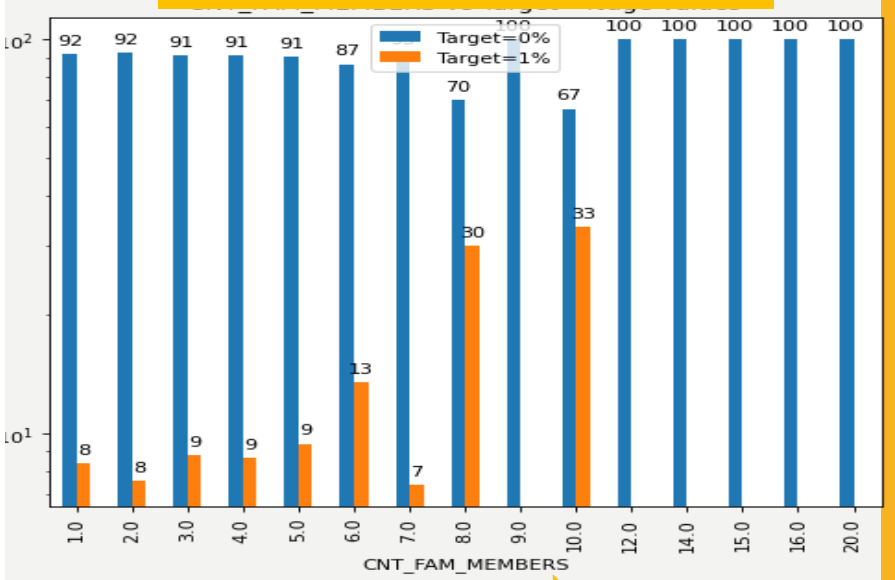
Clients living in Rented apartments or with parents have issues

No. Children



As the number of children increases, issues of payment increases

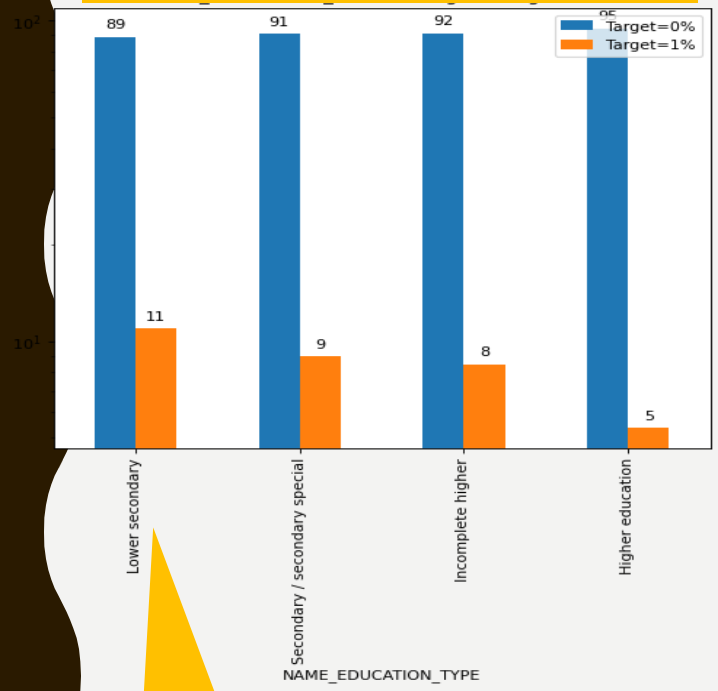
Family Size



As the family member count increases, issues of payment increases

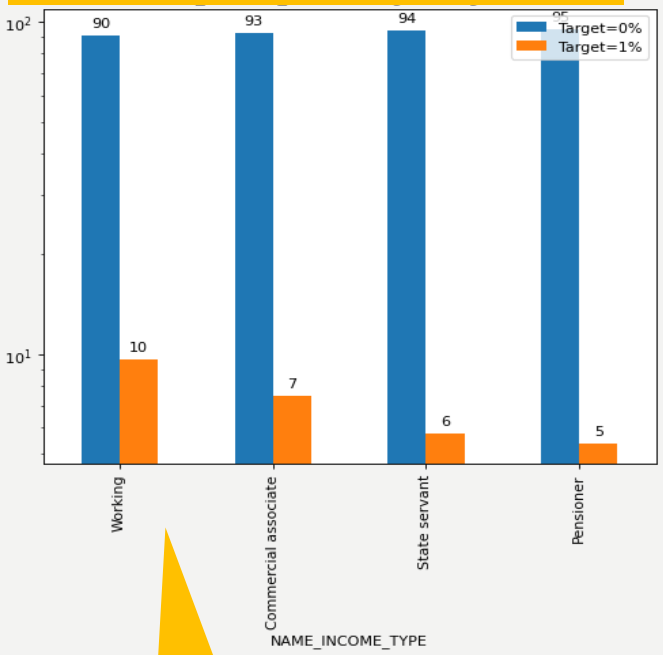
DEMOGRAPHIC *PROFESSIONAL*

Education



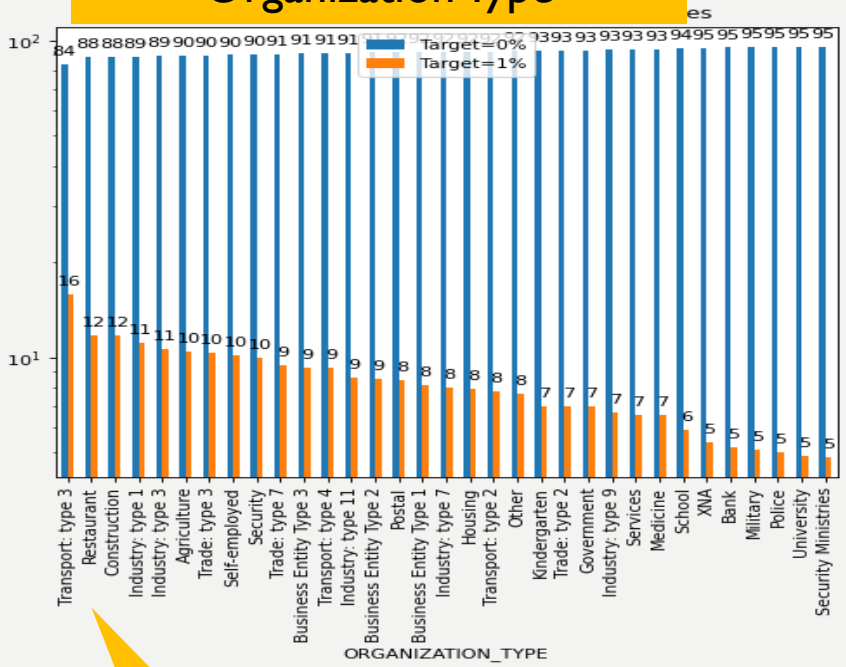
Less educated clients have issues with payment

Income Type



Working class clients have issues with payment

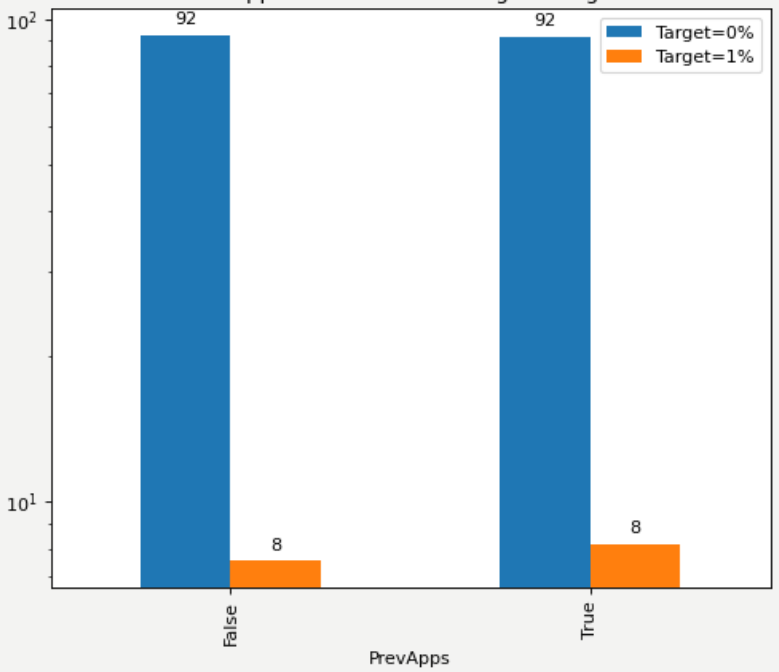
Organization Type



Transport Type 3, Restaurant have issues with payment

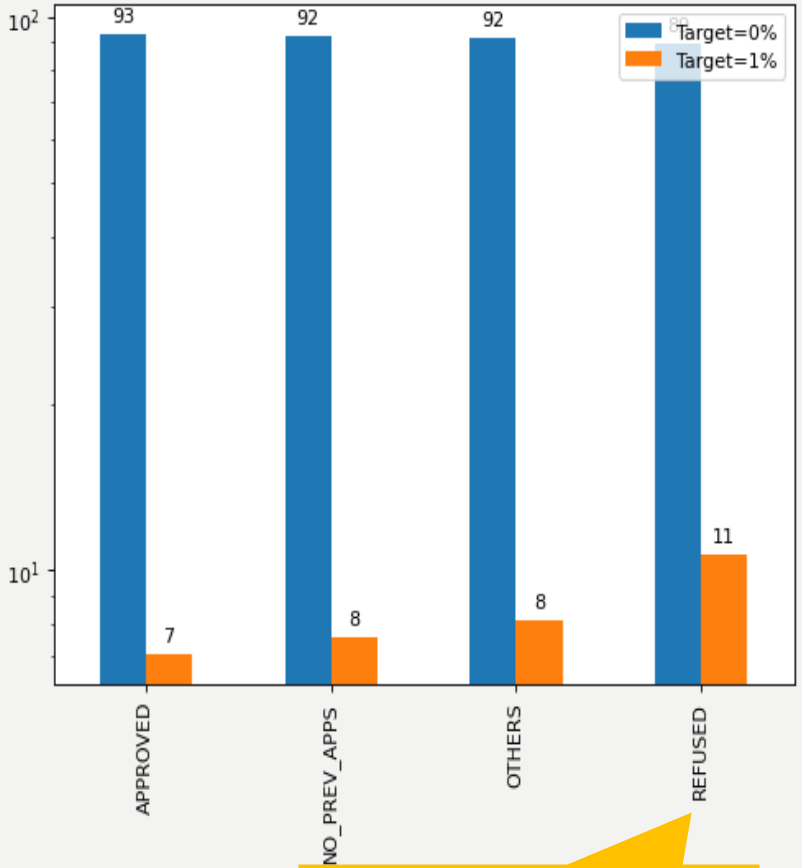
CHARACTER *CREDIT HISTORY*

No. Previous Applications



No difference

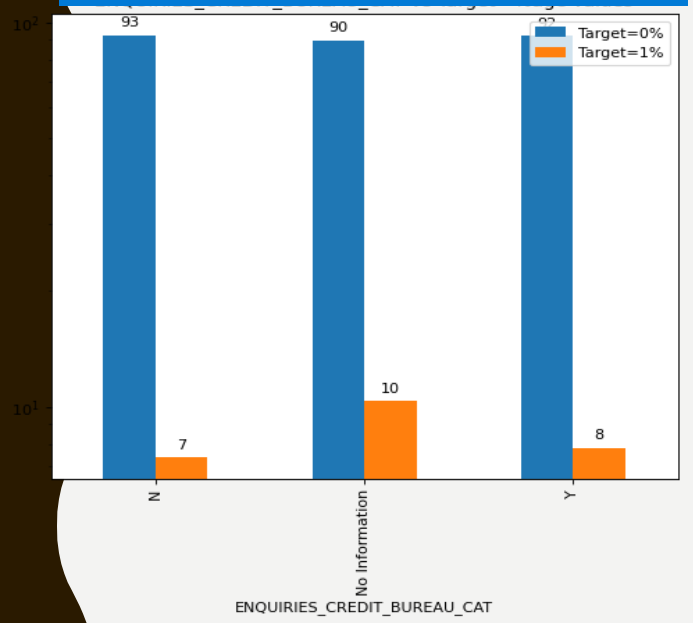
Previous Application Status



If earlier application was refused, then client faces difficulties

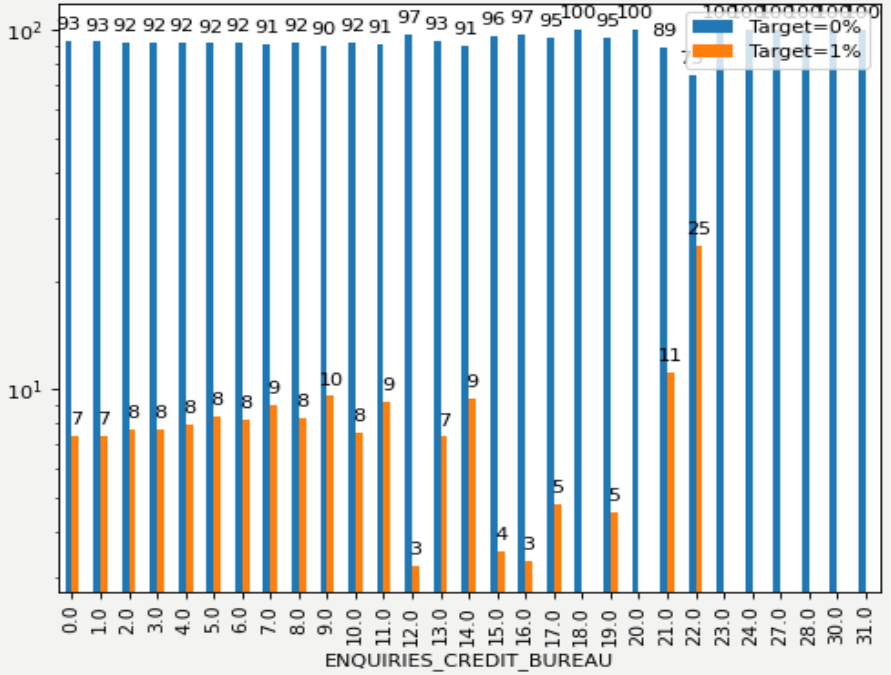
CHARACTER *CREDIT HISTORY/ SOCIAL CIRCLE*

Enquiries to Credit Bureau



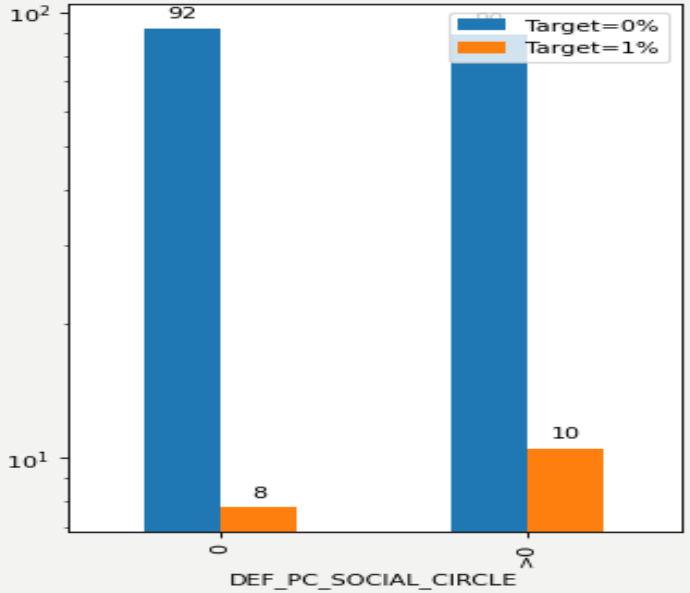
No clear trend

No. Enquiries to Credit Bureau



No clear trend

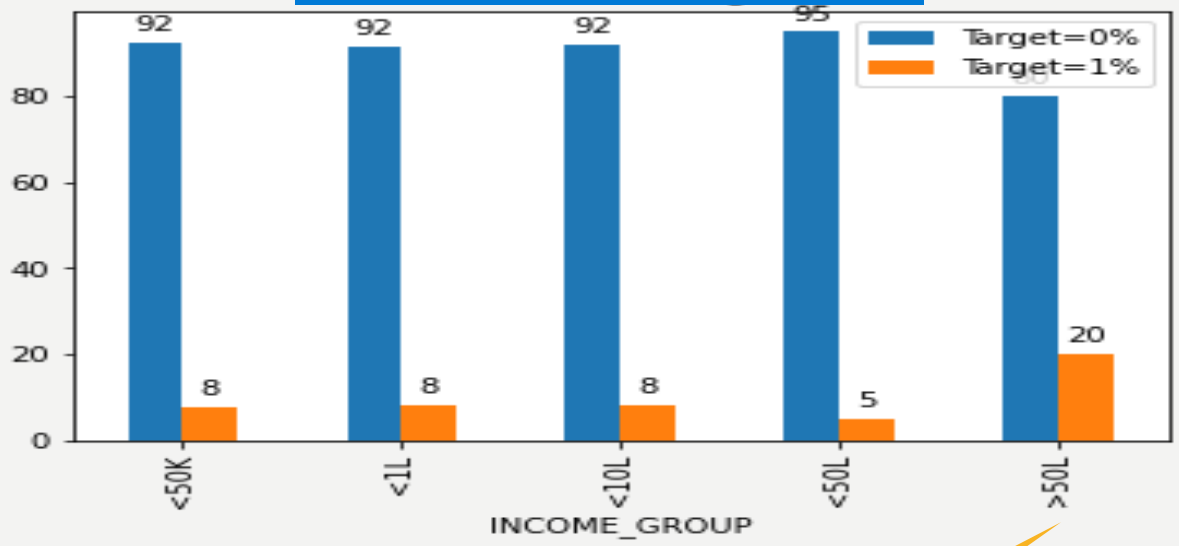
Enquiries to Credit Bureau



Clients with defaulters in social circle faces issues

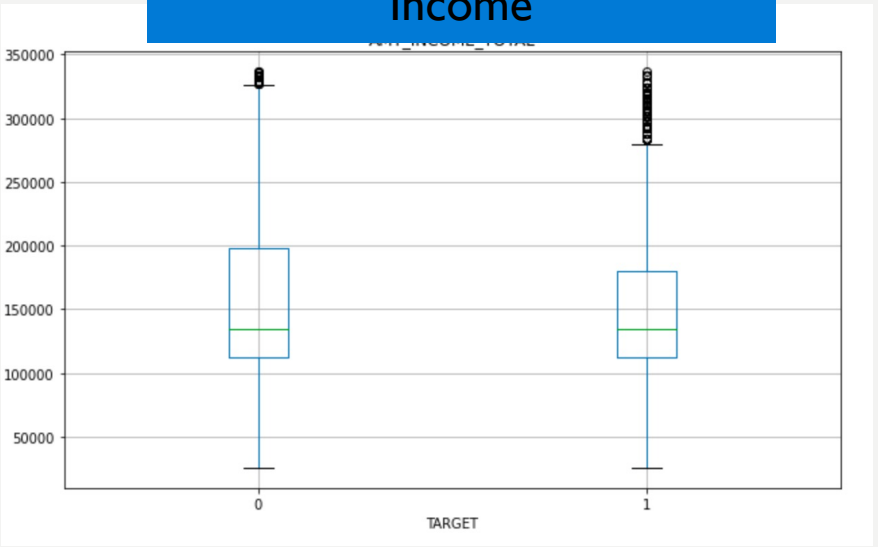
CAPACITY *AMOUNT INCOME*

Income



Income>50 L faces issues

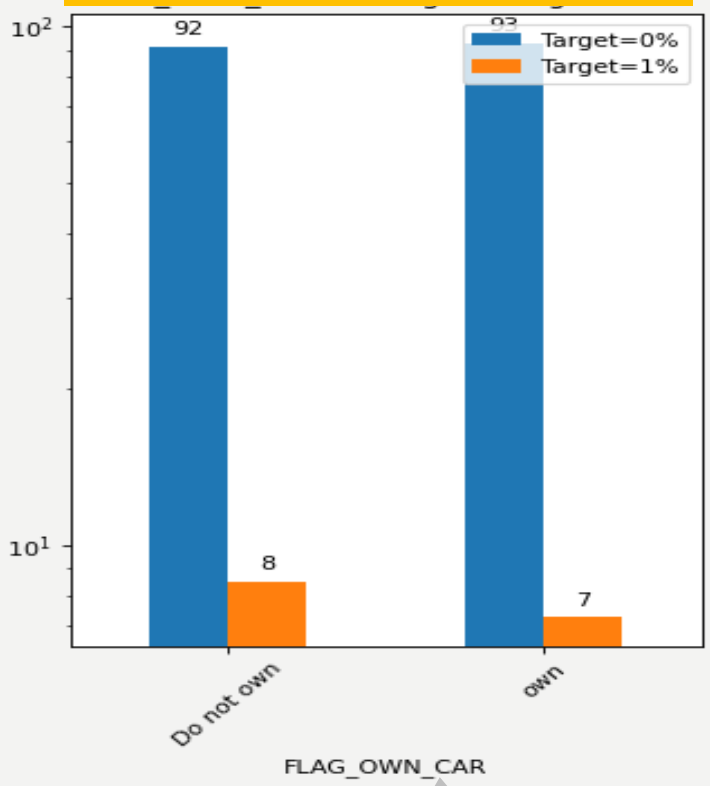
Income



No clear trend

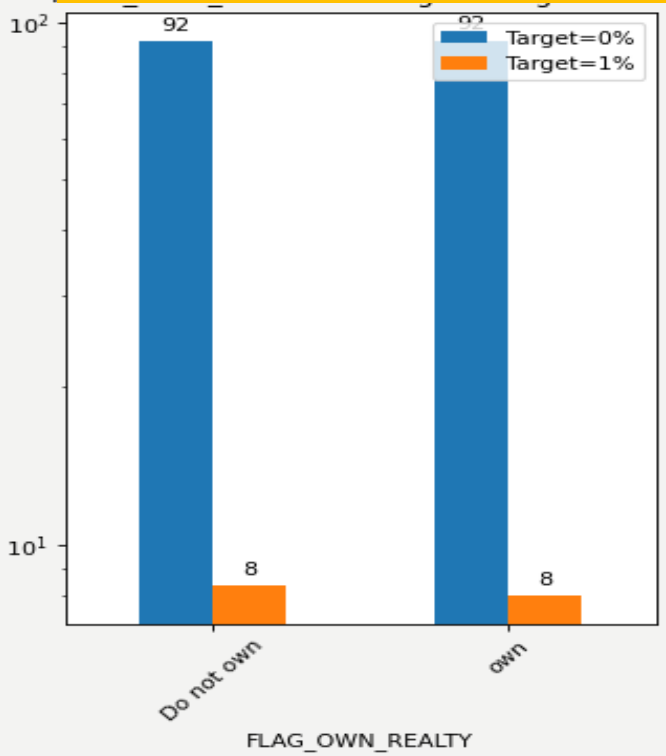
CAPITA *OWN CAR/ REALITY*

Own a Car?



No clear trend

Own Realty?

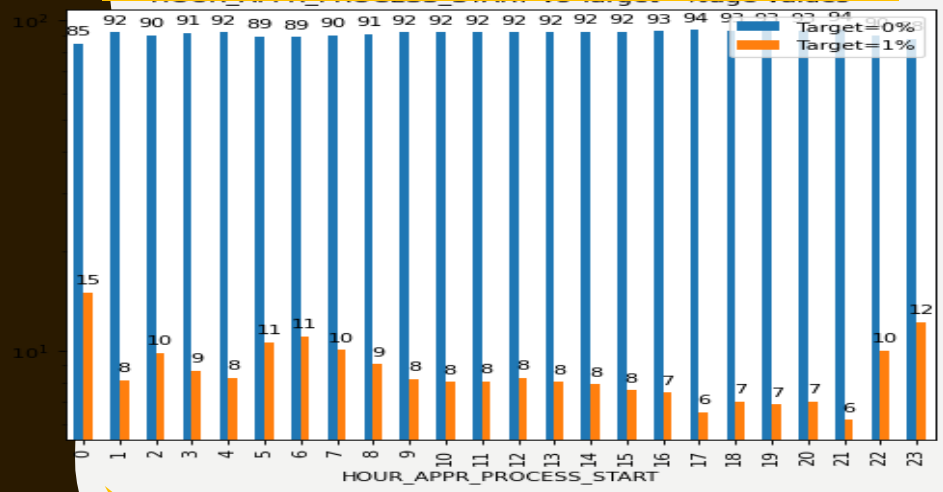


No clear trend

APPLICATION

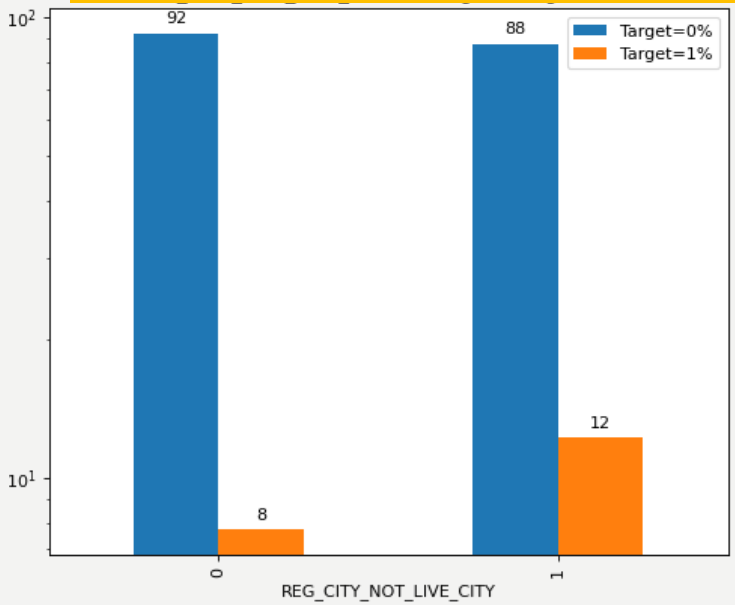
This data set has multiple plots, hence only those plots which show some difference between Target 1 & 0 have been shown

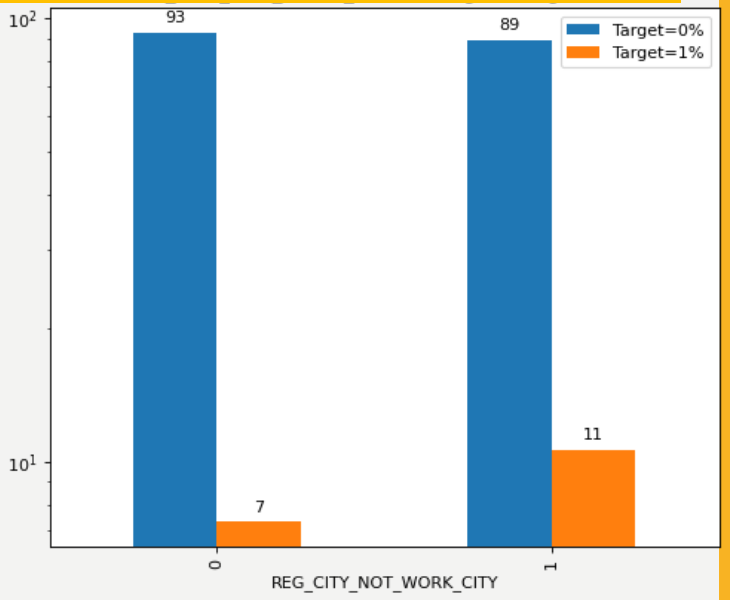
Time when application is filled



Application made in late night hours have issues

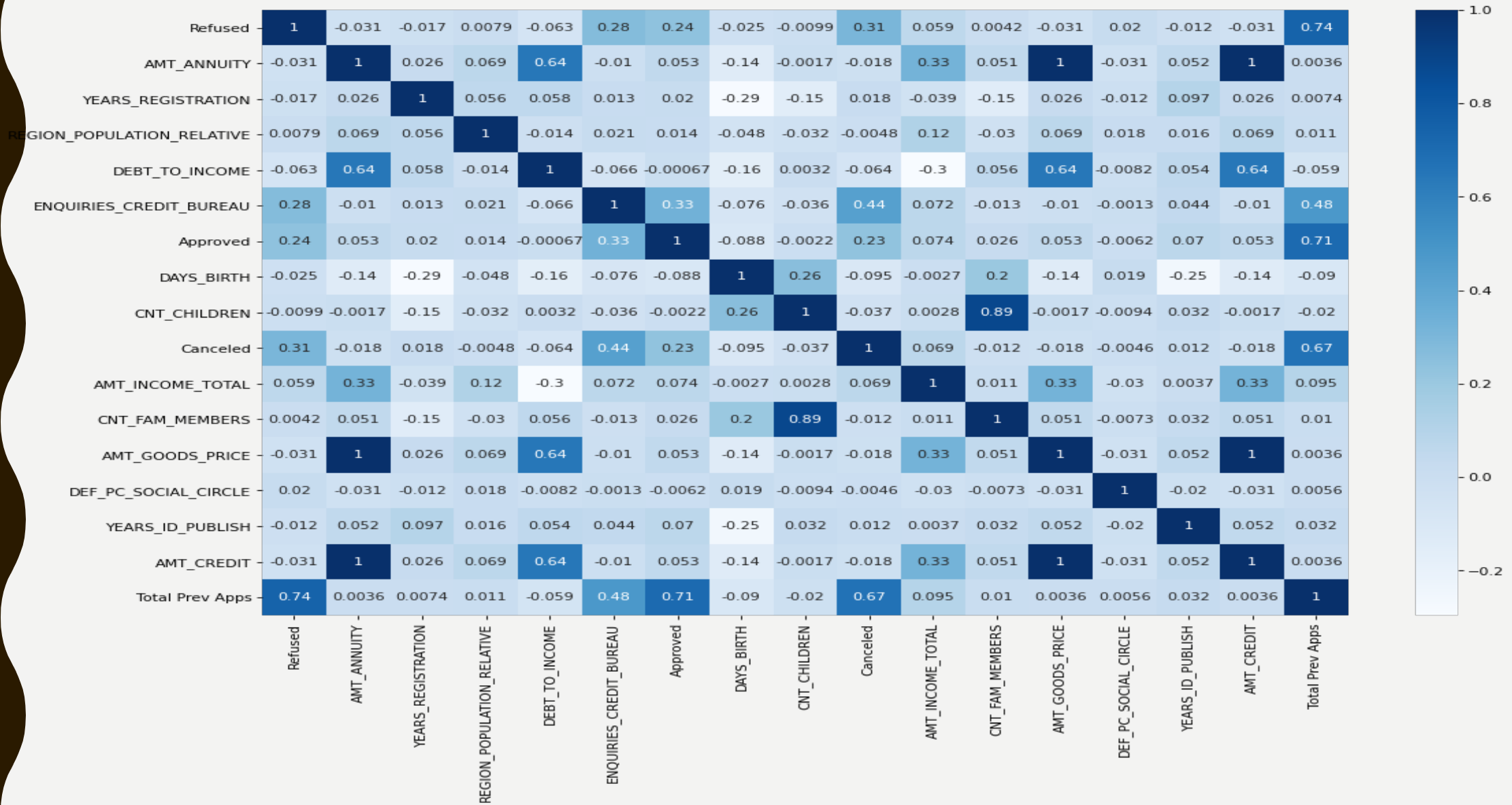
Permanent, Present, Work Address – Same or not?



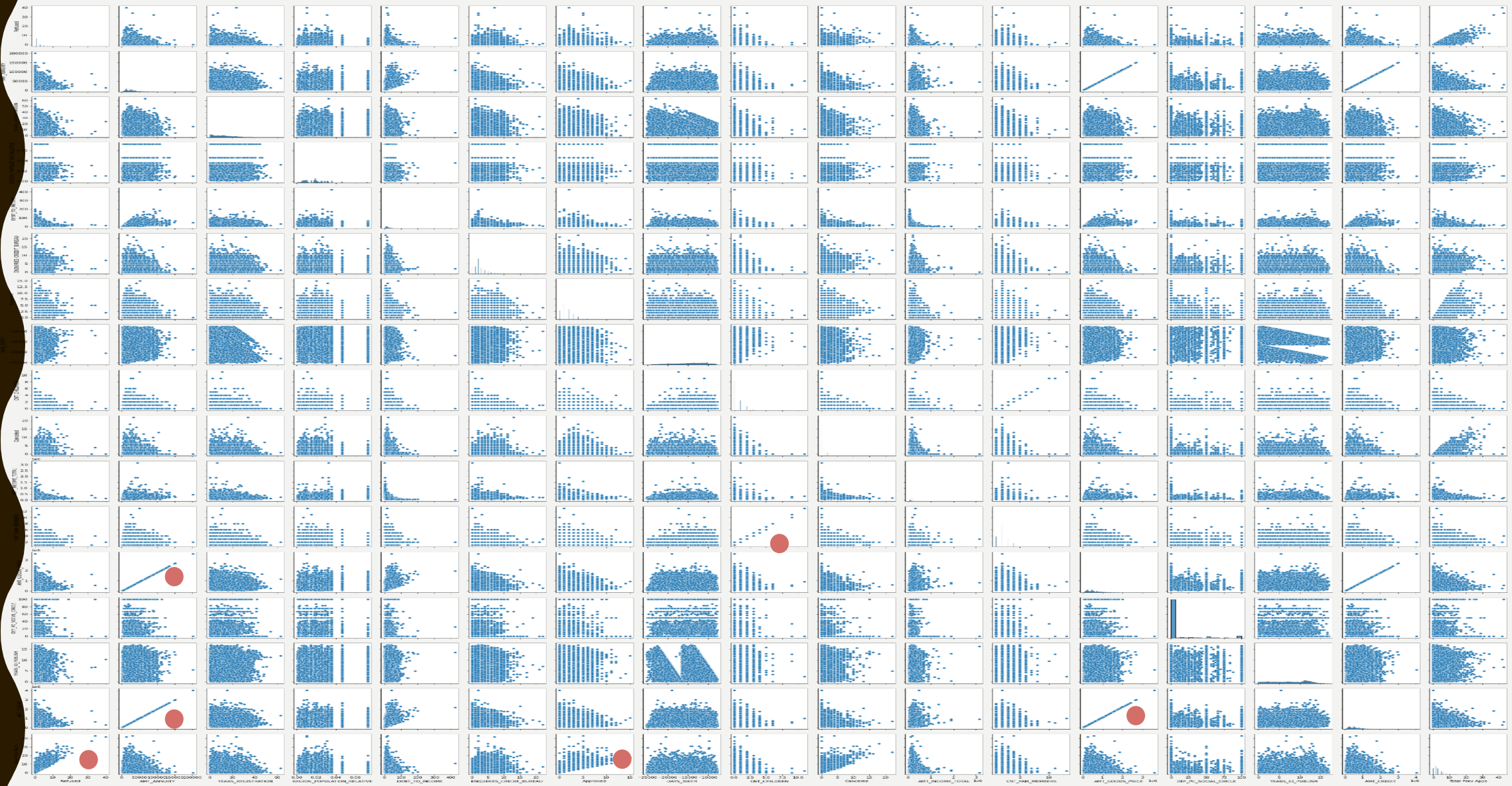


If client permanent address does not match present/work then issues are there – but not prominent

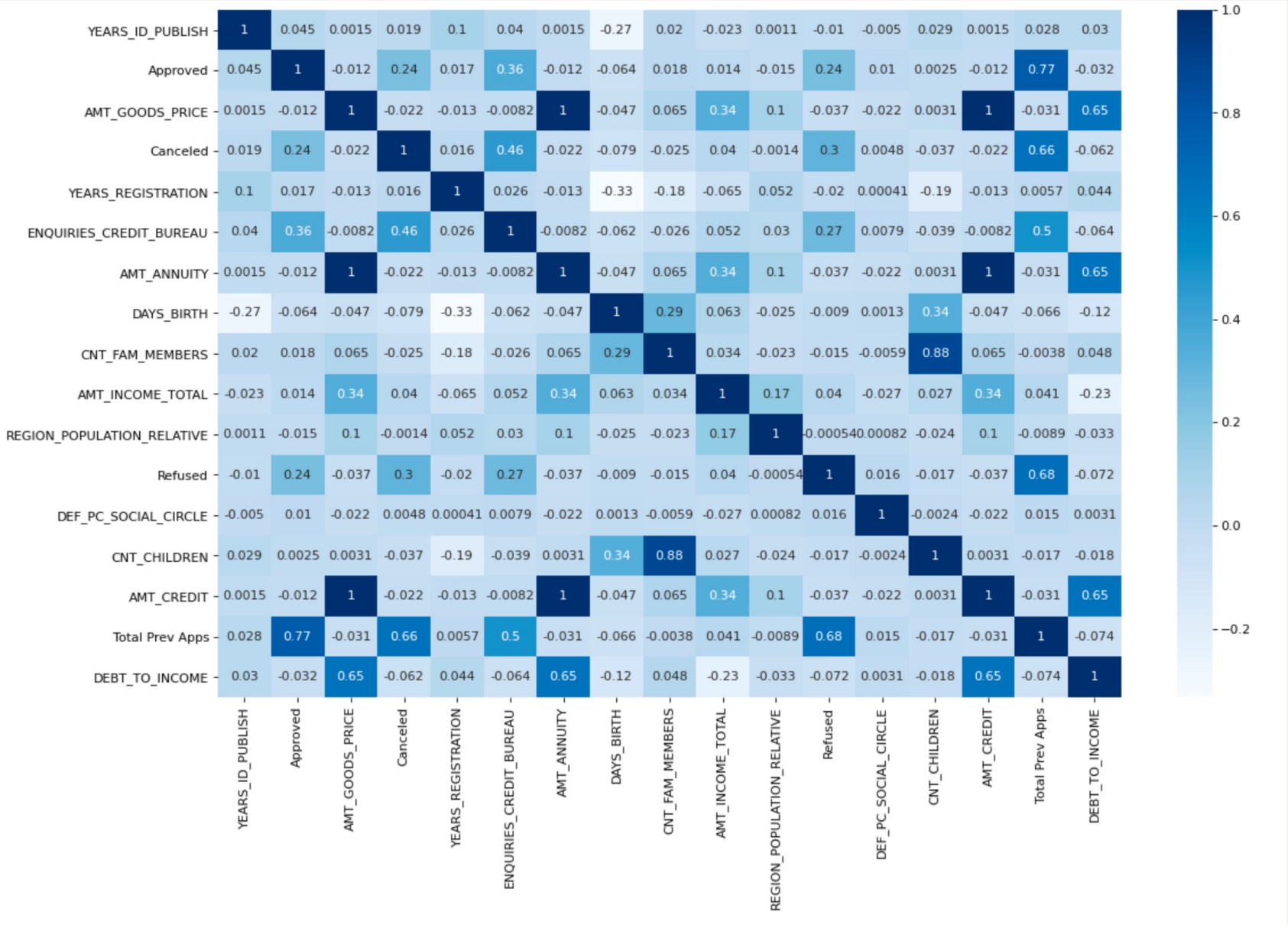
CORELATIONS TARGET =1



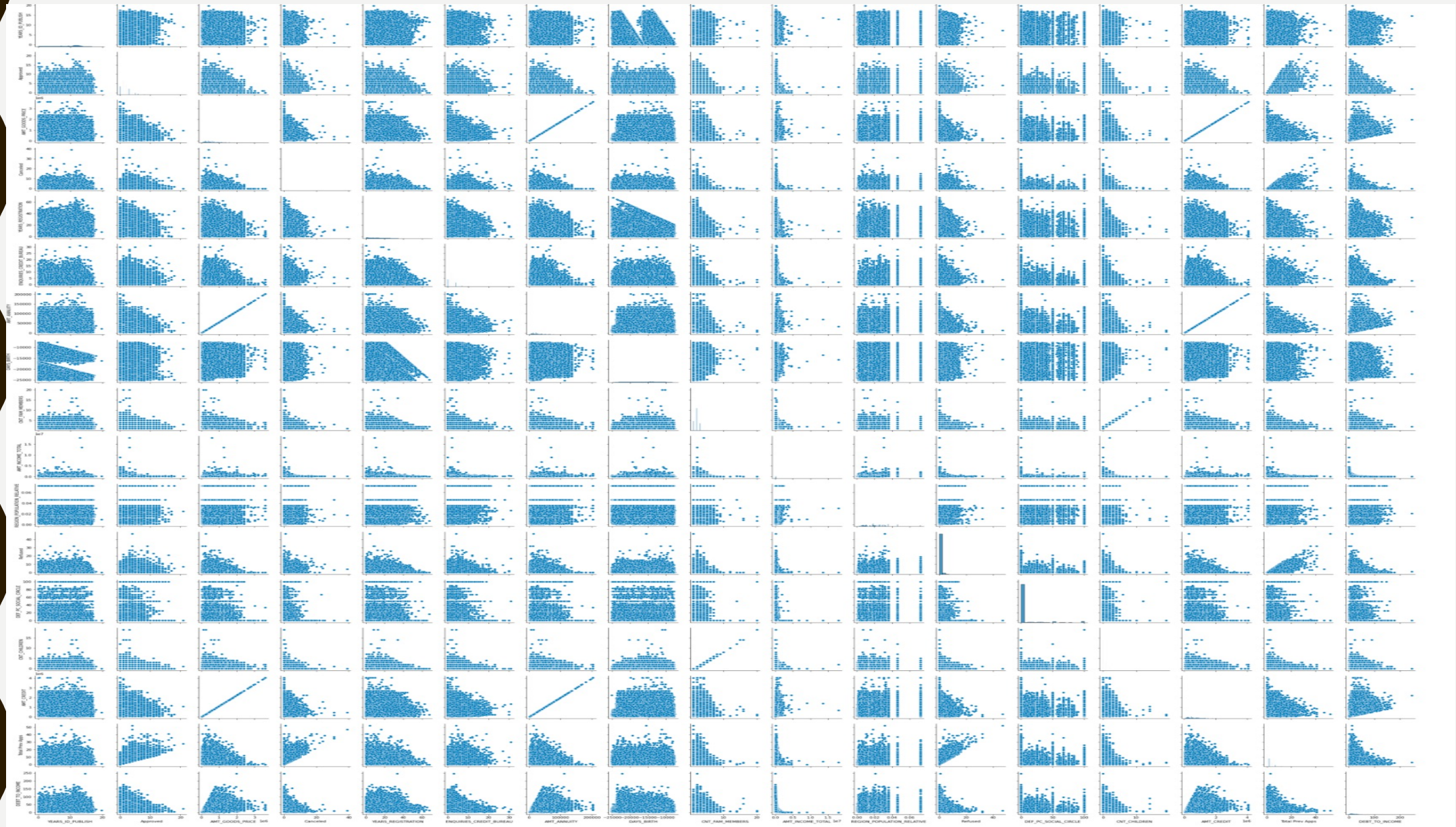
PAIR PLOTS TARGET = 1



CORELATIONS *TARGET=0*



PAIR PLOTS *TARGET=0*



TOP 10 CORELATIONS

TOP 10 CORRELATIONS for TARGET=1

- AMT_CREDIT & AMT_GOODS_PRICE 1
- AMT_CREDIT & AMT_ANNUITY 1
- AMT_GOODS_PRICE & AMT_ANNUITY 1
- CNT_CHILDREN & CNT_FAM_MEMBERS 0.89
- Total Prev Apps & Refused 0.74
- Approved & Total Prev Apps 0.71
- Total Prev Apps & Cancelled 0.67
- Debt to Income & Amt_Credit 0.64
- Debt to Income & Amt_Goods_Price 0.64
- AMT_ANNUITY & DEBT_TO_INCOME 0.64

TOP 10 CORRELATIONS for TARGET=0

- AMT_CREDIT & AMT_GOODS_PRICE 1
- AMT_CREDIT & AMT_ANNUITY 1
- AMT_GOODS_PRICE & AMT_ANNUITY 1
- CNT_CHILDREN & CNT_FAM_MEMBERS 0.88
- Approved & Total Prev Apps 0.77
- Refused & Total Prev Apps 0.68
- Total Prev Apps & Cancelled 0.66
- Debt to Income & Amt_Credit 0.65
- Debt to Income & Amt_Goods_Price 0.65
- AMT_ANNUITY & DEBT_TO_INCOME 0.65

RECOMMENDATIONS

If the profile of a new client falls in the

RISKY group, then caution should be exercised to extending a loan

SAFE group, then loan should be given

DRIVING FACTOR	RISKY Difficulties in paying	SAFE No difficulties in paying
Gender	Male Clients	Female clients
Age	Younger Clients	Older clients
No. Children	More children	Less children
Family size	Larger family size	Smaller family size
Family Status	Civil Married, Single/Not Married	Widow
Housing	Clients in Rented apartment or staying at parents house	-
Region Rating	Higher Region category	Lower Region category
Education	Clients with lesser education	Clients with higher education
Income Type	Working	Pensioner, State Servant
Organization	Transport Type 3	Bank
Previous Applications	With Prev Apps REFUSED	-
Defaulters in social circle	Clients with higher percentage of defaulters in social circle	Clients with lesser %age of defaulters in social circle
Income	Clients from the >50L income group seem to be having difficulties in paying the loan	-

THANK YOU

OBSERVATIONS *CONSOLIDATED*

DRIVING FACTOR	Description	Risky Difficulties in paying	Safe No difficulties in paying
Gender CODE_GENDER	More percentage of male clients have difficulties in paying	Male Clients	Female clients
Age DAYS_BIRTH --> AGE	Lesser age, more difficulty in paying	Younger Clients	Older clients
No. Children CNT_CHILDREN	With increase in number of kids, the ability to pay decreases.	More children	Less children
No. Family members CNT_FAM_MEMBERS	With increase in number of family members, the ability to pay decreases	Larger family size	Smaller family size
Family Status NAME_FAMILY_STATUS	No clear observation	Civil Married Single/Not Married	Widow
Housing HOUSING_TYPE		Clients in Rented apartment or staying at parents house	
Region Rating REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY	Higher the region, the clients find it more difficult to pay. Clients from regions 3 have difficulty in paying off	Higher Region category	Lower Region category

CONTD…

CONTD… OBSERVATIONS *CONSOLIDATED*

DRIVING FACTOR	Description	Risky Difficulties in paying	Safe No difficulties in paying
Education NAME_EDUCATION_TYPE	With increase in education, the ability to pay increases.	Clients with lesser education	Clients with higher education
Income Type NAME_INCOME_TYPE	None	Working	Pensioner, State Servant
Organization ORGANIZATION_TYPE	None	Transport Type 3	Bank
Previous Applications PREV_APPS	Having Prev Apps does not indicate any difficulty or otherwise	With Prev Apps refused	
Defaulter in social circle SOCIAL_CIRCLEDEFAULT%	Clients with defaulters in social circle	Clients with higher percentage of defaulters in social circle	Clients with lesser %age of defaulters in social circle
Income INCOME	This does not affect the client's ability to pay	Clients from the >50L income group seem to be having difficulties in paying the loan	None
Debt-To-Income Ratio DEBT-TO-INCOME	This does not affect the client's ability to pay		
Amounts AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE	Independently these does not affect the client's ability to pay		